Computational Approaches for Predicting Drug Targets

Tolulope Tosin Adeyelu

A thesis submitted for the degree of

Doctor of Philosophy

January 2020

Institute of Structural and Molecular Biology

University College London

I, Tolulope Tosin Adeyelu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_ _

Tolulope Tosin Adeyelu

January, 2020

Abstract

This thesis reports the development of several computational approaches to predict human disease proteins and to assess their value as drug targets, using in-house domain functional families (CATH FunFams). CATH-FunFams comprise evolutionary related protein domains with high structural and functional similarity. External resources were used to identify proteins associated with disease and their genetic variations. These were then mapped to the CATH-FunFams together with information on drugs bound to any relatives within the FunFam. A number of novel approaches were then used to predict the proteins likely to be driving disease and to assess whether drugs could be repurposed within the FunFams for targeting these putative driver proteins.

The first work chapter of this thesis reports the mapping of drugs to CATH-FunFams to identify druggable FunFams based on statistical overrepresentation of drug targets within the FunFam. 81 druggable CATH-FunFams were identified and the dispersion of their relatives on a human protein interaction network was analysed to assess their propensity to be associated with side effects. In the second work chapter, putative drug targets for bladder cancer were identified using a novel computational protocol that expands a set of known bladder cancer genes with genes highly expressed in bladder cancer and highly associated with known bladder cancer genes in a human protein interaction network. 35 new bladder cancer targets were identified in druggable FunFams, for some of which FDA approved drugs could be repurposed from other protein domains in the FunFam.

In the final work chapter, protein kinases and kinase inhibitors were analysed. These are an important class of human drug targets. A novel classification proto-

col was applied to give a comprehensive classification of the kinases which was benchmarked and compared with other widely used kinase classifications. Drug information from ChEMBL was mapped to the Kinase-FunFams and analyses of protein network characteristics of the kinase relatives in each FunFam used to identify those families likely to be associated with side effects.

Acknowledgements

So it's finally here! The journey of the past 4 years has several intriguing stories. I would love to give thanks to God Almighty for the grace to complete this great feat. I can't but thank my supervisor whose role on the achievement of my Ph.D. cannot be overemphasised. She is both a mentor and a motivator. Thank you so much Prof. Christine Orengo for taking out time to accept, correct and mentor me all the way. You are such a rare gem. Thank you for being so accommodating, I couldn't have asked for another.

A big thank you to my thesis committee members (Prof. Snezana Djordjevic and the thesis Chair, Prof. Andrew Martins) for their commitment towards the progress on my Ph.D. journey. As a saying goes, "if I have come this far, it is because I stood on the shoulders of those who have go ahead of me". This therefore makes me to acknowledge wonderful scientist who, out of their very busy schedule have found time to share knowledge and helped on this journey. A special thanks to Dr. Moya-Garcia Aurelio, who coached and collaborated with me on drug polypharmacology research and network biology. A sincere appreciation goes to all Post-docs in Orengo group for driving my passion in the area of protein domains classifications and function, and responding always to my questions everytime. A big thank you to Ian Silitoe, Jon Lees, Paul Ashford, Nathalie, Sayoni Das, Nicola Bordin. I will also like to appreciate everyone in the Orengo group for great moments shared together in the group. Thank you Harry Scholes, Millie Pang, Su datt Lam, Vaishali and Joseph Bonello.

I will also like to appreciate the members of the RCCG Faith Chapel who have been an umbrella and a family since I arrived the City of London in 2014. Your

prayers and encouragement have kept me going. A sincere appreciation to Daddy and Mummy Feyibunmi. To my London family; the Ilegbusi, you guys know you rock my world and you are one of the reason for the achievement of this great feat. Thank you Mummy and Daddy Ilegbusi, John and Ike. To all my lovely friend turned family, I can't but mention Folarin, Olawole, Tolu. Thank you so much guys for every moment shared together. I want to also appreciate the Adekunle Ajasin University Management team and the department of Biochemistry for being part of my success story.

A heartfelt gratitude to my parent, Mr and Mrs Samuel Adeyelu, for their encouraging words and prayers throughout the journey. To my siblings Adedamola and Olaide, thank you guys for being the best. To my adorable gift, my wife of inestimable value, Obiageli Jane Adeyelu, you have being a great strength to me on this journey and I can't but thank you for those times I left you alone just to complete this, it is now an added feather and always remember, we did this together. Thank you for believing in me.

Lastly, I can't but appreciate the Federal Government of Nigeria who provided funding for my Ph.D. program through the Presidential Scholarship for Innovation and Development (PRESSID) program. To the amazing world of science in general, discovery and innovations are the heart of creativity. In that wise, there is more to discover and we keep pressing forward.

Impact Statement

The thesis describes the development and application of computational protocols for predicting drug targets and identifying druggable domain families. Proteins are one of the most targeted molecules, and because they mostly function by interaction with other proteins, a comprehensive network of protein interactions was analysed to reveal network properties that could be used to identify drug targets and to characterise the side effects associated with drug targets.

In the first work chapter of this thesis, druggable domain families were identified based on overrepresentation of drug targets. This revealed a subset of domain families whose relatives can be targeted by the pharmaceutical industry. This work was published in Scientific Reports. One major impact of this study is reporting how drugs currently approved and marketed for targeting a particular domain can be repurposed to other relatives within the same druggable domain family. Drug repurposing is of considerable interest to some pharmaceutical companies to re-channel approved drugs to other orphan targets.

Another key area with likely impact that has been addressed in this thesis is the issue of side effects associated with drugs. Side effects arising from drug usage is one of the major causes of death and management are quite costly.

To show the application of this study to diseases lacking drugs, the second work chapter of this thesis reports the repurposing of FDA approved drugs in bladder cancer. This study therefore provides predicted targets that can be validated experimentally.

Contents

List of Figures

List of Tables

List of Abbreviations

ADR	Adverse Drug Reaction
ATC	Anatomical Therapeutic Code
BC	Betweenness Centrality
BLAST	Basic Local Alignment Search Tool
CATH	Class, Architecture, Topology, Homology
CGC	Cancer Genome Census
COSMIC	Catalogue of Somatic Mutation
DEG	Differentially Expressed Genes
DIAMOnD	Disease Module Detection Algorithm
EC	Enzyme Commission
FDA	Food and Drug Administration
FunFams	Functional Families
GeMMA	Genome Modelling and Model Annotation
GO	Gene Ontology
Hi-DEG	Highly Differentially Expressed Genes
HPA	Human Protein Atlas

KEGG	Kyoto Encyclopedia of Genes and Genome
MCODE	Molecular COmplex DEtection
MDA	Multi-domain Architecture
MIBC	Muscle Invasive Bladder Cancer
MutFams	Mutationally Enriched Domain Family
NMIBC	Non-Muscle Invasive Bladder Cancer
PDB	Protein DataBank
PPI	Protein-Protein Interaction
RMSD	Root Mean Square Difference
RNA-seq	RNA sequencing
RO5	Rule-of-five
SDP	Specificity Determining Position
SE	Side effects
SSAP	Sequential Structural Alignment Program
TCGA	The Cancer Genome Atlas
UniProtKB	UniProt Knowledgebase
WGCNA	Weighted Gene Coexpression Network Analysis

Chapter 1

Introduction To Thesis

1.1 Introduction

The drug discovery process is a time consuming challenge that can take up to 15 years with several stages including target identification and optimisation, lead identification and optimisation as well as the drug testing in both preclinical and clinical phases. One of the major groups of biomolecules that are often targeted by drugs are proteins. Drugs elicit their response through either activation or inhibition once in complex with proteins. Proteins are complex molecules that typically interact with proteins, DNA and other biomolecules.

The vast amount of experimental and predicted protein interaction data has given opportunities to analyse pathways associated with various human diseases as summarised in section 1.4. Hence, it is now possible to combine drugs with protein network data and predict the effect of such associations in perturbing the network. Modelling a protein interaction network is not without its challenges, which include the assumption that protein networks are static graphs where partners interact with each other, as opposed to the known dynamic nature of protein interactions. Another limitation of protein networks is the incompleteness of the interaction network. Although these limitations abound, the usefulness of protein interaction networks in the study of human diseases, is now well established. Hence, this thesis reports the

analysis of protein networks data as a means of identifying pathways and biological processes relevant to human diseases and of identifying new drug targets for human diseases.

1.2 Overview of protein interaction networks

The basic functional unit of life, the cell, comprises complex biological systems, whose normal function involves interplay among multiple bio-molecular entities. Proteins are vital components of this system and act as molecular machines, sensors, transporters as well as structural elements. The simplified approach of isolating proteins and studying them as single entities does not take into account the multiple interactions associated with most of the reactions that proteins undergo [1]. Proteins do not function in isolation but interact with other proteins and molecules such as DNA, RNA, carbohydrates. Furthermore, various studies have shown that distinct biological functions can only rarely be assigned to one molecule, thus emphasizing the importance of studying protein-protein interactions using network approaches [2, 3].

Protein-protein interactions are often identified as physical contacts between protein molecules and described by the protein interfaces involved. This implies that protein-protein interaction (PPI) interfaces should be intentional and not accidental and should result in specific selected biomolecular functions. The physical contacts between proteins can either cause static or transient effects. For example, ATP synthase is a molecular machine comprised of static macromolecular complexes whereas the activation of gene expression by the binding of transcription factors and activators on the DNA promoter region of the gene is an example of protein interactions which only occur transiently [4].

Protein-protein interaction networks are important tools for analysing and understanding hidden and known protein functions. It is important to consider protein partners to fully elucidate functional role, as some proteins may elicit their responses based on downstream interaction with other proteins. These interactions have a variety of roles and the perturbation of such roles may lead to phenotypic

changes that can sometimes have a disease outcome [5].

1.2.1 Experimental and computational approaches to constructing protein interaction networks

Due to recent advances in high-throughput technologies, there are increasing large scale data available in various databases to aid the study and understanding of protein-protein interactions [6]. These repositories include data on protein-protein interactions from either experimental studies or prediction methods. As mentioned above, interactions among proteins can occur through obligate complex formation or transient physical contacts [7]. Experimental methods for detecting direct interactions or functional associations of proteins include yeast two-hybrid (Y2H), co-immunoprecipitation, protein complexes determined using affinity purification-mass spectrometry. Experimental analyses to detect indirect associations include gene co-expression and synthetic lethality. Other methods are based on the characterisation of protein interactions using structural methods such as X-ray crystallography, NMR spectroscopy, fluorescence, atomic force, and electron microscopy.

The Y2H method (described in figure 1.1) is one of the *in-vivo* methods that detects a physical interaction between proteins [8]. The key to the success of Y2H lies in the modular nature of most eukaryotic organisms. The two proteins of interest are attached to different domains of a transcription factor. If these proteins are in close proximity to each other, they will bring the DNA-binding domain and the activation domain together, hence forming a functional transcription unit. Although Y2H helps in the recognition of interacting proteins, the method is prone to false negative and false positive interactions; a limitation, leading to the generation of noisy data. Also, it assumes that interacting proteins are localised within the nucleus which means that proteins which are less likely to be found in the nucleus are less likely to activate the reporter gene.

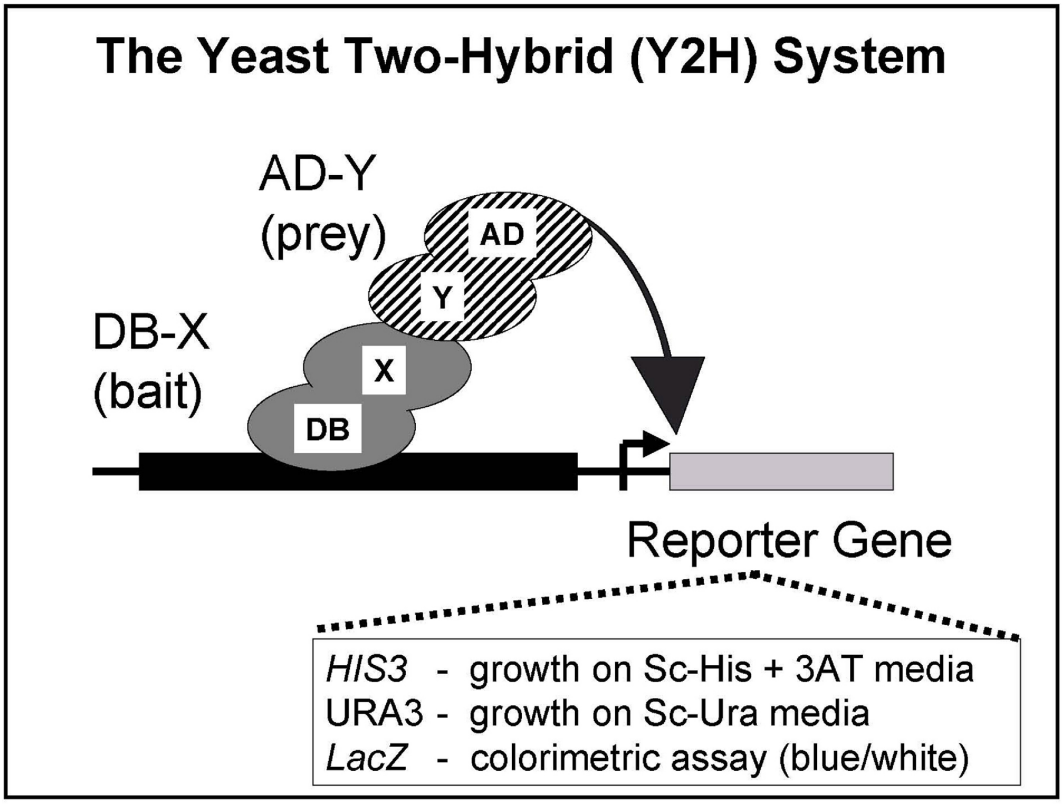

Figure 1.1: The yeast two-hybrid system. A functional transcription factor is reconstituted when X and Y interact. The transcription factor activates reporter genes such as HIS3, URAS3 and lacZ. Figure taken from [9].

Tag-tapping is another *in-vitro* method for identifying a protein and its partners in a high-throughput manner. The procedure involves attaching an epitope tag to the protein of interest and performing a two-stage purification process. The TAP tag is made up of two immunoglobulins G (IgG) binding domains; Protein A and a calmodulin-binding peptide (CBP) and these two parts are separated by a short peptide which is a specific site for TEV protease. The TAP-tagged protein is initially isolated from cell lysate by passing it through a IgG coated bead where it attaches itself through the Protein A domain. The TAP-tagged protein and its interactors are released by incubating in TEV protease. The protein that remains attached to the target protein is examined and identified using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) followed by mass spectrometry analysis [8].

Affinity chromatography can also help in detecting protein interactions; since it

is highly responsive, it can detect even very weak interactions [8]. There is however a tendency for this method to pick up other proteins that are not involved in the same cellular process, which leads to the identification of false positive interactors. Affinity chromatography also involves the use of SDS-PAGE and mass spectroscopy to generate high throughput data [10]. One of the weaknesses of mass spectrometry, is the possibility that complexes may be missed and tagging may disrupt complex formation and dissociate weakly associated components.

X-ray crystallography gives a complete high resolution three dimensional structure of the protein of interest. Protein interaction complexes can be revealed by examining the X-ray crystal structures for interactions that provide mechanistic insights into protein function. While this method reveal the full atom coordinates of the protein, which is highly important to understand the protein function, X-ray crystallography is costly and time consuming [11].

Electron microscopy (EM) has aided the detection of protein complexes. Since proteins are macromolecular complexes and dynamic in nature, cryo-EM helps in dealing with sample heterogeneity and inherent flexibility [12]. Several developments in EM such as the invention of direct electron detection cameras, automated data collections and powerful image processing algorithms have expanded the use of EM in the detection of biocomplexes for a range of sizes from about 150kDa to several hundred of megadaltons.

The increasing amount of available biological evidence and the power of mathematical models means that computational prediction is gaining importance to help increase the coverage of interactions as well as to prune the noisy data from experimental analyses and thereby improve data reliability. Various approaches that have been used for computational prediction of protein interactions are mentioned below.

Gene Fusion

Gene fusion is also known as the Rosetta Stone method, and it is based on the assumption that when two genes are fused together in one species, they tend to also interact in another species even though they might be distinct proteins [8]. The sequences of the proteins in the different species being considered are compared to

detect fusion events that might have occurred, and hence used as a means of predicting the likelihood of the proteins interacting. (See figure 1.2 below for illustration of gene fusion event).

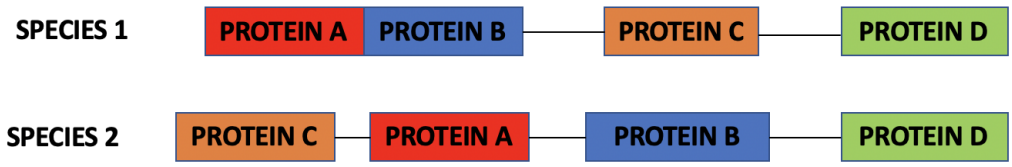

Figure 1.2: Protein A and B have been shown to undergo gene fusion in species 1 which is an indication that they are likely to interact together in species 2.

Gene fusion approach has been exploited in the chimeric protein-protein interaction method (chiPPI) that uses domain-domain co-occurrence score to estimate the likelihood of a protein interaction. The chiPPI method has been used to study fusion proteins in a given network and distinguish between tumor suppressors (TSG) and oncogenes as the TSG are lost in a fusion network while the oncogenes are brought in close proximity to other neighbours in protein network.

Orthology based approach

Orthology involves the transfer of functional annotations from one species to another based on the fact that the proteins are highly similar. Orthology based approaches assume that if proteins A and B interact in a given species, the orthologs of A and B in another species are also likely to interact with each other. This strategy led to the creation of HomoMint [13] where orthologs of human proteins found in other species were used in inferring interacting pairs in human. DIOPT (DRSC Integrative Orthology Prediction Tool) [14] has also been designed as a tool for inferring interactions based on orthology. It combines nine orthology prediction tools and gives a confidence score to infer interactions based on the number of agreements amongst the various methods.

Phylogenetic profiling method

This method relies on the evolutionary history of proteins to predict associations of interacting partners. One of the most interesting methods within this field is the

mirror tree method by Pazos and Valencia [15] in which it is assumed that proteins showing similarity in their molecular phylogenetic protein trees exhibit similarity in their interaction [16]. This relies on the premise that co-evolution between interacting proteins is reflected in the similarity of distance matrices between their corresponding phylogenetic tree. The mirror tree, for instance, computes the Pearson correlation coefficient between two distance matrices derived from the phylogenetic trees and uses this to evaluate the extent of co-evolutionary relationship between two proteins. The schematic representation of the mirror tree method is shown in figure 1.3 below.

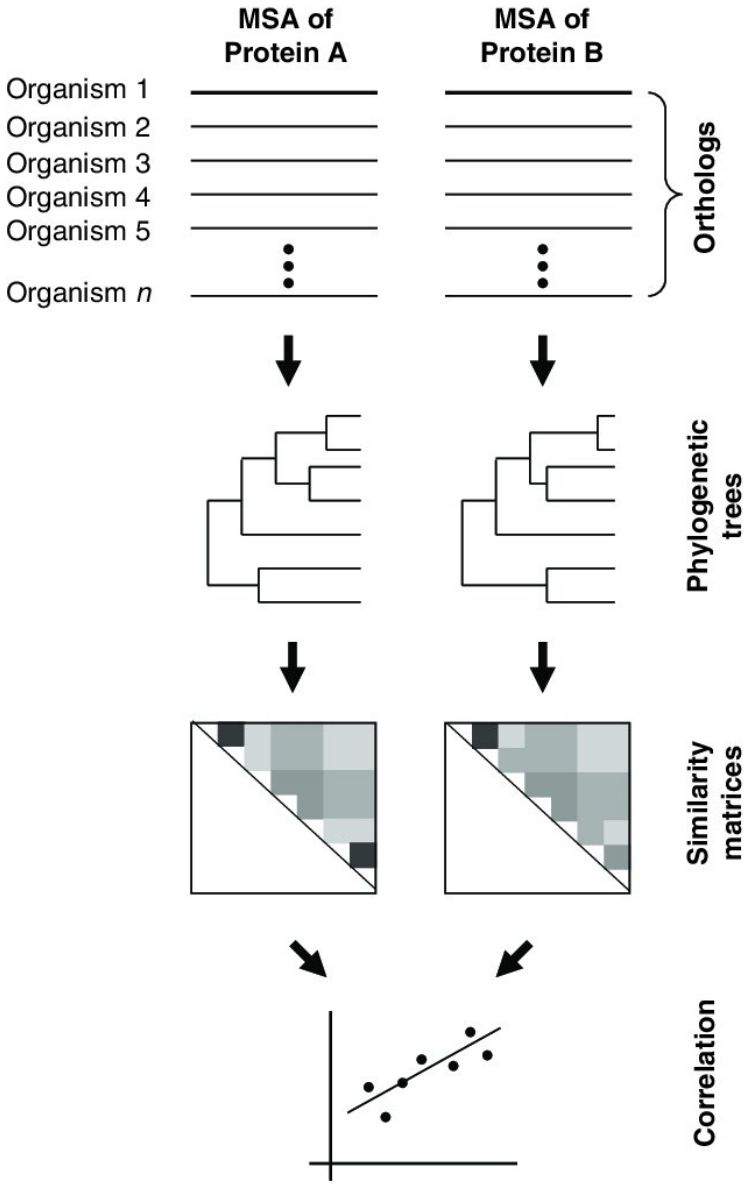

Figure 1.3: Schematic representation of the mirror-tree method. Multiple Sequence Alignments (MSA) of proteins A and B, constructed from orthologs of A and B, respectively, from a common set of species, are used to generate the corresponding phylogenetic trees and distance matrices. Comparison of the corresponding distance matrix based on the linear correlation determines the degree of co-evolution between protein A and B. Proteins A and B are predicted to interact if the degree of coevolution, measured by the correlation score, is high. Figure taken from [17]

Structure-based approaches

It is also possible to use a structure-based approach to predict protein-protein interactions, based on the idea that if proteins A and B interact together, then protein A' and B' with similar known or predicted structures to A and B may also pos-

sibly interact. Structure-based approaches have been widely exploited by several available databases such as Struct2Net [18] and InterPreTS [19]. To predict protein interacting partners, InterPreTS usually thread sequences over protein complexes in the protein databank (PDB) and chooses the best match. Matches are predicted based on empirical potentials by scoring the amino-acids involved in atomic contacts at the interface of the complex and comparing the scores to a background of sequences that are unlikely to be involved in interactions [19]. Machine learning is then used to analyse these patterns and predict whether two proteins are interacting. The method does not rely on any external information such as the gene expression or cellular localisation and hence remains an independent approach for structural prediction of protein interactions.

Supervised learning approaches

Machine learning approaches have been used to facilitate integration of multiple proteomic and genomic features. Various machine learning algorithms such as: Decision Tree and Random Forest (RF), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and kernel methods have been applied for the prediction of protein-protein interactions [20]. Experimentally determined interactions are analysed to find patterns that distinguish the sequences of interacting protein pairs from non-interacting pairs [20]. These predictions are based on protein information such as physicochemical properties of the protein, structural information, evolutionary information, domain information etc. For example, protein domains can be identified within sequences and matching pairs of domains found to be enriched among known interacting proteins pairs have been used in the prediction of new interactions.

Guo *et al* [21] for instance, combined a support vector machine (SVM) with auto covariance to predict protein interactions from protein sequences. Auto covariance takes account of the interactions between amino acids in the protein sequence that are far apart and takes into consideration neighbouring effects of residues in order to discover patterns in the entire sequences. Other methods integrate semantic similarity such as Gene Ontology (GO) annotations with SVM for PPI prediction

[22, 23].

In a related approach, Chen *et al* [24] used hybrid feature representation that combines protein sequence properties and gene ontology information as well as interaction network topology to predict protein interactions, a method available as PPI-MetaGo. Physicochemical properties of the amino acids exploited include: hydrophobicity, hydrophilicity, polarity and solvent accessible surface area. Functional annotations were also adopted from partitioning of the directed acyclic graph (DAG) from GO, while network based features exploited topological properties of the network. This method was benchmarked against experimental PPIs, with the hybrid feature method having a higher performance than each individual prediction method.

1.2.2 Network representation of protein interactions

Network analysis of biological systems provides a framework through which the complexity of associations in relation to pathology and physiology can be studied [25]. Graph representations (see figure 1.4) are typically used to provide an understanding of the binary nature of protein interactions. A biological network such as a protein-protein interaction network can be represented as an undirected graph G = (V, E) where G is the graph, V is the node (protein) while E is the edge (interaction) between two nodes. The interaction represented can either be physical or functional depending on the context in which the interaction is measured.

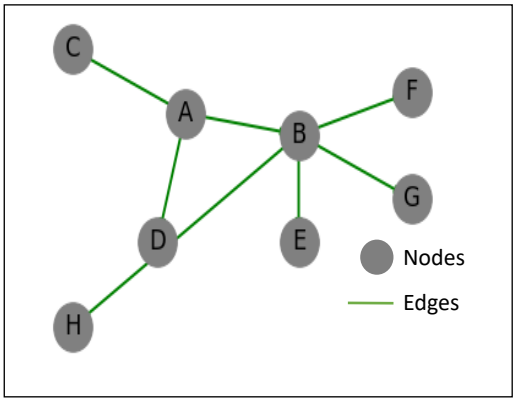

Figure 1.4: Graphical representation of a protein interaction network where each letter represents a protein linked together by the green edges.

As well as undirected graphs, the representation of interactions in a network can also include directed, mixed, weighted or unweighted, acyclic graphs, trees and Boolean networks. The visualization of such networks can be facilitated using several existing software tools like Cytoscape [26], Pajek [27], visANT [28], SNAVI [29] and AVIS [30] amongst others. The network or graph is a mathematical object and therefore its characteristics can be described by considering some mathematical properties. These can be applied to the nodes, edges or the entire network (global topological properties) as well as sub-networks (modules) in the protein-protein interaction networks.

1.2.3 Graph theory and general characteristics of networks

Connectivity

One of the most basic properties considered for a network is the connectivity of each node in the network, which is also referred to as the measure of its degree (k). The degree (as shown in figure 1.5) is the number of edges or connections the node has with other nodes. The degree distribution P(k) of a network is then defined to be the fraction of nodes in the network with degree k [31].

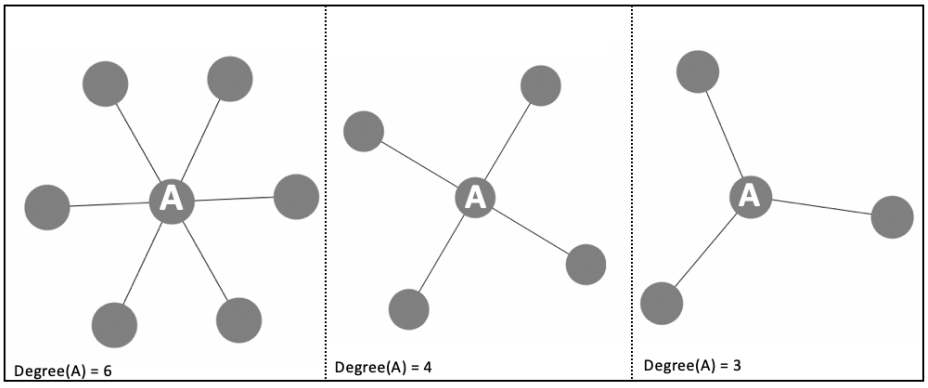

Figure 1.5: Node A showing different degrees as measured by the numbers of interacting nodes.

Biological networks are considered to be scale-free and follow a power law distribution [32]. In this type of network, most nodes have a lower degree while a significantly few nodes have a higher degree than the mean degree of the network. This indicates that the probability P(k) of finding a highly-connected node is lower

than the probability of finding sparsely connected nodes. A plot of the numbers of interactions for a given protein against the probability of observing a given protein with such a number of interactions will give a downward sloping straight line in a double logarithmic plot and this term is referred to as the power law distribution in graph theory. Figure 1.6 shows the power law distribution characteristics of a typical biological network.

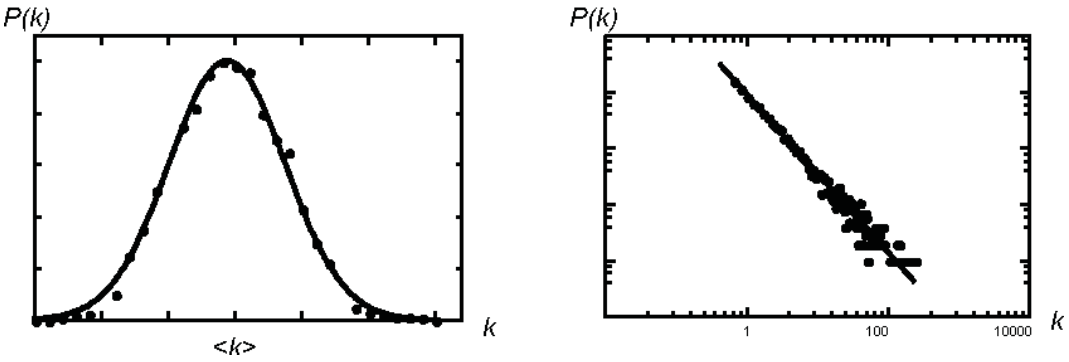

Figure 1.6: Power law distribution characteristics of real and random networks. The random network on the left follows a normal distribution (Gaussian distribution) where most nodes are clustered around the centre as compared to the powerlaw distribution in biological network on the right. Figure taken from [33]

Centrality

The measure of centrality identifies nodes that are important in a given network. Some of the measures of centralities used are given below.

Degree centrality (DC)

The degree centrality (DC) provides a measure of the relative connectivity of node i in a network, as follows;

$$DC_i = \sum_{j=1}^{n} \mathbf{A}_{ij} \tag{1.1}$$

DC is the degree centrality, i and j are nodes in the network, A represents the adjacency of the network. Nodes in the network with higher degree centrality are referred to as hubs and they tend to play important roles in the network.

Betweenness centrality (BC)

The betweenness centrality (BC) considers the number of shortest paths that contain

any given protein and it is a measure of how often a node is present in all sets of the shortest paths [34].

$$BC_a = \sum_{ij} \frac{n_{ij}^a}{g_{ij}}, \tag{1.2}$$

BC_a is the betweenness centrality of node 'a'. n_{ij}^a is the number of shortest paths between nodes i and j that pass through node 'a'. g_{ij} is the total number of shortest path lengths between nodes i and j. A typical small network and the measurement of its betweenness centrality is shown in figure 1.7 below.

In considering the robustness of a network i.e. its ability to withstand perturbation, a damage to a node with a high BC has been seen to be more relevant than the removal of nodes with high degree (k) [35].

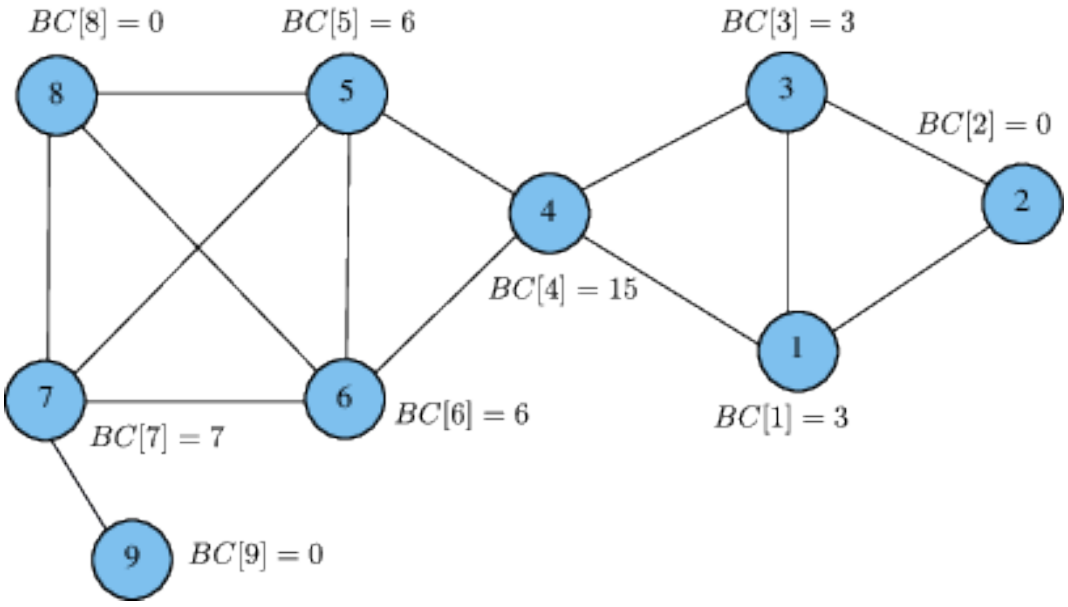

Figure 1.7: Example of betweenness centrality (BC) scores for a small network. Node 4 has a high BC score because it lies in the path of the shortest path of many other pairs of nodes. Node 9 has a BC score of 0 because it does not lie on the path of any pair of the remaining nodes. Figure taken from [36].

Path-length

Another network property that can be assessed is the average path-length or diameter of the network which is expected to be smaller than the size of the network. For example, if the network size i.e. the number of nodes is "n", the average path

length or diameter can be in the order of "log n" or less [32, 37]. This shows that biological networks tend to exhibit small world characteristics [37]. In such networks, most nodes have few links or connections whilst some have high links and are thus referred to as the hubs. Hub proteins are known to be essential to the network since their removal results in more phenotypic responses as compared to the non-hub proteins [38]. The hub proteins have also been shown to be evolutionarily conserved and play a highly dynamic role in the network [38, 39].

Other centrality measures

Other centrality properties that can be measured include the *closeness centrality* of a network, which measures the average shortest path from one protein to another. The *eigenvector centrality* [40], gives an assessment of the closeness of the highly connected nodes. Eignenvector centrality (C) is based on the premise that each node's centrality is the sum of the centrality value of the nodes that are connected to it by an edge $\mathbf{A}_{ij}$ [41].

Eigenvector is mathematically represented as:

$$c_i = \varepsilon^{-1} \sum_{j=1}^{n} \mathbf{A}_{ij} c_j, \tag{1.3}$$

Where ε is the eigenvalue of the eigenvector c; i and j are the nodes in a graph.

The changes in some of the network properties mentioned above can be used to assess the impact of perturbations on the network. The scale-free concept of networks for example, has been linked to the robustness of the network[31]. This is associated with the hypothesis that a random removal of nodes may not influence the network as compared to targeting hubs which disrupt the network. This was first described by Jeong and co-workers in the centrality-lethality rule [42]. They showed that random mutations in the yeast genome did not affect the topology of the network, however, computational elimination of highly connected proteins increased the network diameter (average path length) and the mutations affecting these hubs were associated with diseases [42].

In contrast, Yu and Das [43] showed, using high throughput data, that disease

genes are usually non-essential and occupy peripheral nodes in the human interactome. However, this observation was not reproducible using literature curated data or predicted protein interactions. Kar and colleagues reported that hub proteins were associated with both essential and non-essential genes within an organism but the essential genes are mainly involved in diseases [39].

Hub proteins can be further classified as ”party hubs” and ”date hubs” where the former interact with partners simultaneously while the latter interact with partners at different times and different locations (see figure 1.8). More recently, a new class of hubs was identified by Paci *et al.*, called ”Fight-club hubs” which are characterised by causing a negative correlation with their first neighbours [44]. These sorts of genes are upregulated while their neighbours are downregulated.

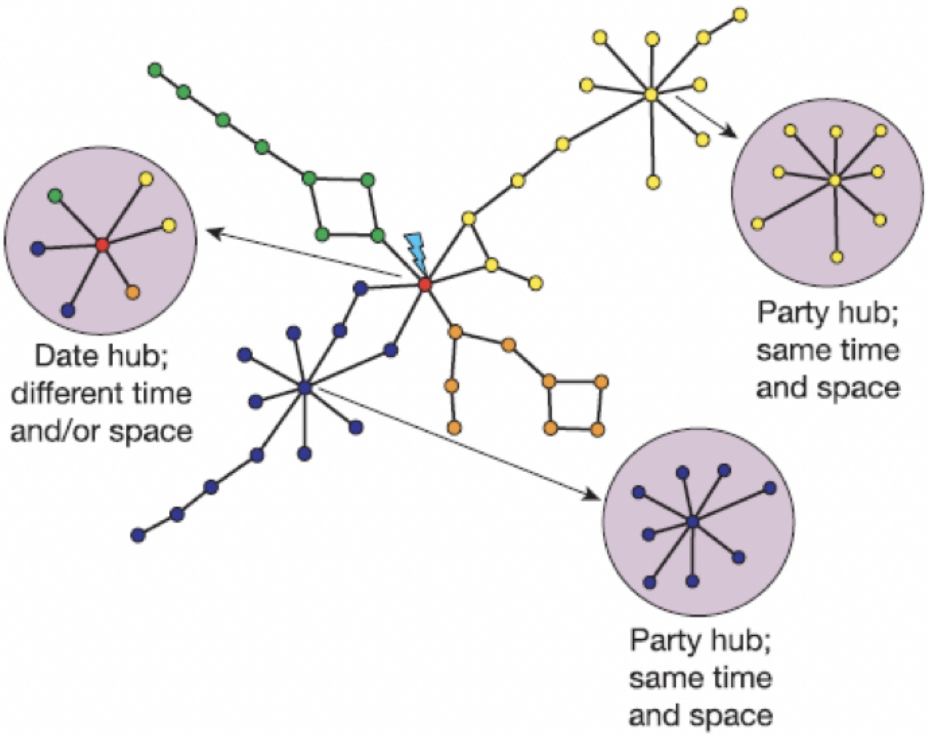

Figure 1.8: Typical network with examples of date and party hubs. Proteins are coloured according to the similarity in their mRNA expression patterns. Date hubs interact with different proteins at different times and different locations whereas party hubs interacts with partners at the same time and same location. Figure taken from [45]

1.3 Identification of Network Modules

The interaction of proteins can either be transient or permanent. Interacting proteins can belong to protein complexes or modules which are associated with a particular biological role. The network can therefore be clustered into sub-networks or modules (also called sub-graphs), which are the densely-populated regions in the network separated by more lowly populated areas. The complexity of a typical protein network can generate ambiguity which can be resolved by examining it in smaller and more manageable modules. However, rather than a simple division of the networks, modules reflect functional associations and biological processes and can also be used in predicting functions for uncharacterised proteins or genes which belong to a module containing proteins with known functions [46].

There are several types of clustering approaches that can be used for the modularisation of protein networks. Pizzuti and Rombo highlight several algorithms and tools that are of interest [47]. They classify clustering approaches into two types; topology-free approaches, and graph based. The topology free approach does not take into consideration the network topology but simply relies on the measurement of the distance between proteins whereas the graph-based approach takes into consideration the topology of the network and is widely used. The graph-based approach can be classified into five categories i.e. Local neighbourhood Density search (LD), Cost-based Local search (CL), Flow Simulation (FS) and Link Clustering (LC) described in more details below.

1.3.1 Local neighbourhood density search

The local neighbourhood density search method (LD) is based on the detection of dense subgraphs within a network. Examples of some local density search methods include; MCODE [48] which detects dense and connected regions by weighting the nodes based on the local neighbourhood densities. It involves three steps; node weighting, complex prediction and optional post processing. The MCODE algorithm first selects the top weighted node (seed node) based on the local neighborhood density and incrementally adds neighbouring nodes provided they are within a

given threshold defined by the user. In the post-processing operations, MCODE removes less dense subgraphs which can be added back by the user, if needed. However, MCODE cannot guarantee the modules are highly interconnected as highly weighted nodes may not necessarily mean high connectivity within the cluster.

Another approach similar in principle to MCODE is DPClus [49] which also uses node weight and cluster property. The cluster property describes how compactly connected a node must be before it can be added to a cluster. As a rule of thumb, the more edges a node is connected to, the more compact they are and the more likely they would be part of a complex. DPClus accesses the minimum density value and the minimum clustering property to determine the insertion of neighbours into the cluster.

SWEMODE (Semantic WEight for MODule elucidation) by [50] also uses the node weighting and clustering property. However, the definition of weight in this case is linked to functional similarity between nodes which is obtained based on Gene Ontology (GO) terms (Gene Ontology is discussed further in section). Lubovac and colleagues [51] showed that modules obtained by incorporating GO semantic similarity into the network topology seem to be advantageous over using only topology information. Other LD-algorithms include CFINDER [52] which extracts cliques (maximal complete subgraphs) in the network, and uses criteria for the minimum numbers of nodes expected in the clique. A clique-clique overlap matrix is built to allow identification of cliques with common nodes and as such CFINDER generates overlapping modules.

1.3.2 Cost-based local search

One of the widely used module detection methods based on cost-based local search is MODULAND [53]. MODULAND is an integrative method for determining network modules (”hills”) of a community landscape. MODULAND examines the regions where nodes influence each other more than the rest of the network. The influence function is estimated by considering the impact of removal of nodes on the links and the entire network and this influence is based on the measurement of, for example, density or any other network weighting approach that can be explored.

The performance of this method in generating non-overlapping modules has been compared to graph clustering method by Lancichinetti *et al* [54], and consistency in the identification of non-overlapping modules was observed.

1.3.3 Flow simulation

The flow simulation (FS) approach mimics the information spread on a network using random walk or biological knowledge of neighbouring proteins to detect clusters in the network. One of the widely used flow simulation approaches is the Markov Cluster Algorithm (MCL) [55] which simulates random walk within the graph by alternative expansion and inflation operations and then separates the graph into different segments. MCL is a fast, scalable approach and the inflation parameters influence the numbers of clusters observed from the network.

1.3.4 Link Clustering

The link clustering method is based on sets of edges rather than sets of nodes and detects modules based on the strength of the edges. Ahn and colleagues [46] proposed an agglomerative link clustering that clusters edges into topologically related communities since networks tend to comprise communities with overlaps in which the nodes belong to more than one group. This is one of the major advantages of the link clustering approach since it allows the automatic clustering of nodes into multiple communities without the necessity of performing multiple clustering of the set of edges. The performance of this method is dependent on the edge similarity measure. Link clustering was shown to outperform three widely used methods: Clique percolation, Greedy modularity optimisation and Infomap [46]. The application of this algorithm to 11 different networks revealed that the link communities are fundamental building blocks that reveal biologically meaningful overlaps as well as the hierarchical organisation of networks [46].

1.4 Protein Networks application to Human Diseases

Biological networks are useful tools to model the complexity associated with genotype-phenotype relationship underpinning diseases [56]. They can help with the understanding of disease network properties, identification of disease sub-

networks and network based classification of diseases. Protein networks may also help with unraveling disease progression which may in turn lead to the identification of novel disease genes and disease pathways as well as targets for drug discovery [2].

There are several databases containing human disease information such as the Online Mendelian Inheritance in Man (OMIM) [57], Comparative Toxicogenomics Database (CTD) [58], Malacards [59], and DisGeNet [56] which can be used to identify disease genes in a network to understand association of the known disease genes with other genes.

Genome-wide association studies (GWAS) [60] give information on genome regions containing genes likely to be associated with disease and can be used to identify genetic variations associated with diseases. Analyses of genetic variations such as non-synonymous single nucleotide polymorphism (nSNPs) have been carried out in a wide range of diseases including cancer, diabetes, Parkinson, Crohn's diseases. Human genes are made up of coding sequences (exons) and non-coding units (which are around 20 times larger) called introns. At the level of the amino acid sequence, disease causing mutation often lead to changes in the physicochemical properties such as charge, hydrophobicity and geometry of amino-acids. Similar observation was also found in structural analysis of disease causing variants which lead to changes in hydrogen bonding and salt-bridges formation when compared to silent/harmless mutations [61].

Several methods have been developed to understand how disease genes behave in a biological network. One technique that has been used to uncover the causal path linking perturbed causal genes to other affected genes in the network is the network propagation technique [62]. This tests whether a given genetic perturbation might affect the expression of a specific gene or gene of interest [63]. It uses this influence function to predict candidate genes that are associated with the known disease genes. Other studies have included patient's clinical data into sub-networks to identify the most frequently perturbed sets of genes. This approach has been adopted in HOTNET and HOTNET2 algorithm [64]. The HOTNET method anal-

yses the local network topology and computes the influence of mutated genes on the network based on a diffusion process from the *source of heat* (mutated genes) within the network to the surroundings. This has helped to identify cancer driver genes and the pathways within the network associated with the disease [64].

Several studies have revealed that disease genes possess distinctive network topological properties [65, 2, 5, 56]. For example, it has been shown that cancer genes tend to be central in a biological network. This may not be the case for other disease genes and has been suggested that the observations in cancer are biased by the fact that some genes have been more extensively studied than others. Research by Goh and colleagues showed that if essential genes are excluded from the analysis of Mendelian diseases, disease genes do not show the tendency to occupy the hub positions in the interactome [65], suggesting no precise conclusion can easily be drawn.

1.4.1 Tissue specificity of Diseases

The actions of genes are dependent on the cellular or tissue specific location in the organism [66]. Disease genes tend to be expressed in tissue-specific patterns and tend to have higher mutation rates over evolutionary time. It is hoped that the analyses of tissue specificities, disease pathologies and gene-disease associations in biological networks will give a clearer understanding of disease mechanisms.

Projects such as ENCODE (ENCyclopaedia Of DNA Element) and The Cancer Genome Atlas (TCGA) provide comprehensive genomic profiles for cell lines and cancer respectively. The Human Protein Atlas Map (HPA) is another resource comprising data on the mapping of proteins using various OMICs methods including antibody-based imaging, proteomics, transcriptomics and system biology approaches [67] . The Human Protein Atlas Map has three atlases: the tissue atlas which contains information on expression profiles of human tissues at the mRNA and protein level; the cell atlas provides information on the spatial distribution of proteins within cells, while the pathology atlas provides information on several types of cancer obtained from about 8000 patients [67, 68]. Other useful information is the GTEx project [69], which provides information on $>$50 non-disease

tissue expression profiles across 900 postmortem samples. The analysis of such data could aid understanding how genetic variations affect normal gene expression in human tissues to enable identification of genetic variations leading to human diseases.

TissGDB [66] combines multiple tissue specific expression resources including the Human Protein Atlas (HPA), Tissue-specific Gene Expression and Regulation (TiGER) and Genotype-Tissue Expression (GTEx). TissGDB currently contains 2461 curated genes across 22 tissue types and 28 cancer types from TCGA. It provides seven categories of annotations: TissGeneSummary, TissGeneExp, TissGene-miRNA, TissGeneMut, TissGeneNet, TissGeneProg, and TissGeneClin.

Network approaches have been used to identify the mechanism of tissue specific interactions [70] and [71]. Barshir and colleagues mapped expression profiles across 16 tissues, to show that genes involved in hereditary diseases are widely expressed in all tissues yet enigmatically cause disease phenotype only in a few. However, two phenotypic observations were made (i) many of the disease causing genes have elevated expression levels in their disease tissue (ii) disease causing genes have a higher tendency for tissue specific interactions in their disease tissue. This therefore means it is possible to identify, predict and prioritise disease genes by annotating the protein networks with this tissue specificity information.

Kitsak and colleagues showed that disease gene expression patterns in selected tissues cannot alone explain the observed tissue specificity. Nevertheless, it is expected that disease associated genes should be highly enriched in the affected tissue compared to non-diseased associated genes [72]. To analyse this, they combined expression patterns with network analysis and found that disease genes expressed in the specific tissue tend to be localised in the same neighbourhood of the interactome. By contrast, genes expressed in different tissues are segregated in different network neighbourhoods. Overall, these results suggest that the integration of gene expression, disease manifestations, molecular network connectivity and tissue specific data can help with the prediction of novel disease candidates.

1.4.2 Analysis of Disease Modules

The identification of modules in a network can also help to identify coordinated biological functions that are not well captured in established canonical pathway annotations. Proteins involved in diseases with similar phenotypes have also been shown to interact together in disease modules [31]. Several other studies have shown that diseased genes are not randomly scattered in the network but agglomerate in specific clusters thereby suggesting specific disease modules for each disease [31, 5]. This therefore means that other proteins in the same modules which are not currently associated with the disease can be potential candidates of the disease.

The disease module hypothesis postulates that "the cellular components associated with disease segregate in the same neighbourhood of the human interactome" [5]. In other words, there is high likelihood of disease associated proteins to interact with each other and be clustered in the same neighbourhood of the interactome, forming disease modules with properties indicating molecular determinants of such disease. Disease modules can therefore help to identify novel disease genes, biomarkers and therapeutic targets [2].

The closer the phenotypic manifestation of two diseases in terms of the tissue location, system effects and drug responses, the higher the expectation of overlap in the protein network of the modules that are associated with the two diseases [72]. Therefore, one of the most valuable ways of understanding molecular mechanisms associated with a disease is to consider its modular representation in a human protein interactome.

Several tools have been used to analyse and identify disease modules in network. The Module-Explorer package of the NetworkAnalyst algorithm [73] is one such tool and uses random walk to identify modules of frequently visited nodes and generate edge weighted networks. In this case, the weights are derived from quantitative node information such as gene expression attributes that may be associated with the node. However, if the modules to be identified are strictly disease associated modules, other algorithms may perform better since the disease associated proteins do not reside necessarily in particularly dense local communities.

Therefore, disease associated proteins may be better predicted using connectivity significance (a measure of the number of connections from a candidate protein to other known disease "seed" proteins which should be greater than statistically expected by chance). This approach has been implemented in the Disease Module Detection algorithm (DIAMOnD) [5] which has been shown to take into consideration the incompleteness of the human network.

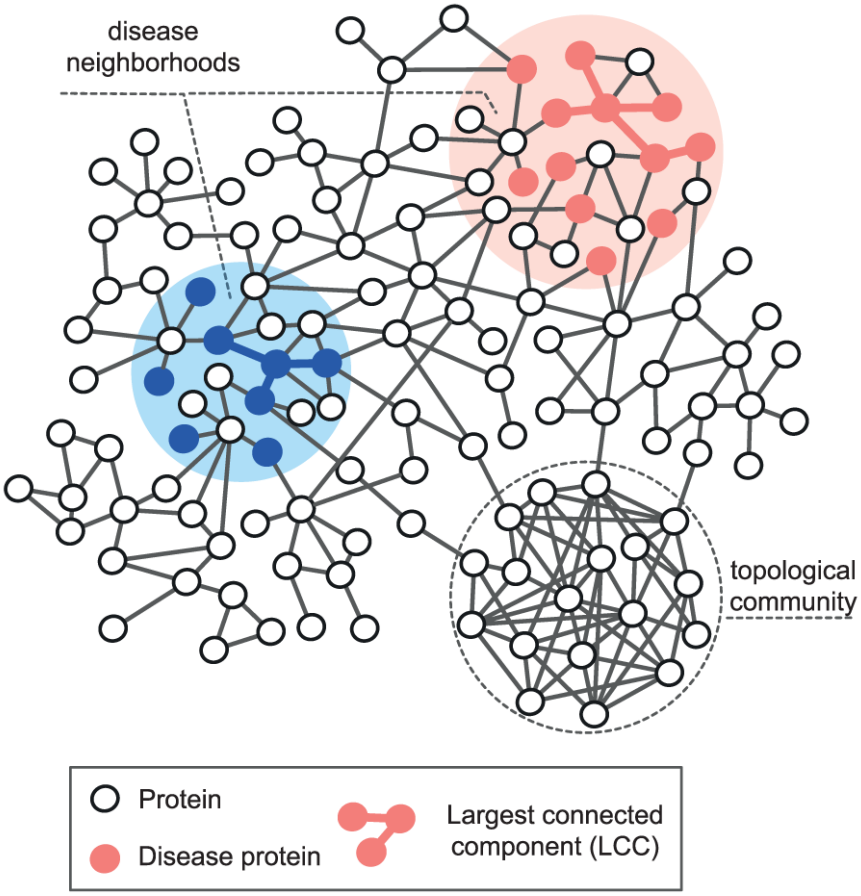

Figure 1.9: Proteins associated with the same disease tend to be localised in the similar neighbourhood (disease module) in an interactome. The blue nodes are the predicted disease genes, while the red nodes are the known disease genes. Figure taken from [5].

DIAMOnD [5] starts from a set of known disease genes as shown in figure 1.9 and iteratively adds nodes from neighbours that are highly connected to the disease genes to the disease module. The union of the putative disease genes identified using

DIAMOnD and the known disease genes forms the disease module. This method of extraction of disease modules has been carried out in networks and disease contexts by several groups [5, 74, 75].

Analysing disease modules may also help in understanding pathobiological relationships between diseases. A study by Menche *et al.* revealed that overlapping disease network modules showed significant co-expression patterns, symptom similarity and co-morbidity as compared to modules in separate neighbourhoods [5]. The analysis of modules in protein interaction networks can also help with identification of novel targets for drugs if the protein network is annotated with drug information.

1.5 Resources used in this thesis for protein and network annotation

1.5.1 Resources with information on protein interaction networks

There are several publicly available databases reporting protein-protein interactions. These use a controlled vocabulary and adopt a common format for ease of use. Protein interaction databases used in this research are described below.

STRING database

The STRING database [76] (`https://string-db.org/`) is one of the most cited protein interaction databases. STRING includes direct (physical) interactions as well as functional associations. STRING incorporates available experimental protein interaction data and also prediction of protein interactions using co-expression analyses, automated text-mining, and the computational transfer of interaction knowledge based on gene orthology. STRING consists of protein interactions across multiple species and provides an annotation confidence score ranging from 0-1000 based on the estimated likelihood that the given interaction is biologically meaningful. This score can be used as a way to filter out low scoring edges which tend to be noise.

Pathway Commons

Pathway information is important as it captures the molecular knowledge of a biological process. Pathway Commons [77] (`https://www.pathwaycommons.org/`) is a collection of publicly available pathway data for several organisms. Currently, the database holds information from nine sources (BioGRID, Cancer Cell Map, HPRD, HumanCyc, IMID, IntAct, MINT, NCI/Nature PID, Reactome). It currently gives integrated data on 1477 pathways and 687,883 interactions. Pathway Commons is regularly updated and also well cited for pathway analysis.

1.5.2 Resources with information on drug target identification

ChEMBL-database

The ChEMBL database (`https://www.ebi.ac.uk/chembl/`) is a resource providing data on protein compound associations and holds information on the bioactivity of small molecules and bio-therapeutics. These activities have been carefully manually curated by consulting peer-reviewed publications [78]. ChEMBL tends to provide a broad coverage of a diverse set of targets, organisms and bioactivities. ChEMBL also provides information about the status of the drug/drug-like compounds (whether approved or experimental) with a score of 4 indicating FDA approved drugs. The current version of ChEMBL, (ChEMBL-24), contains 12,091 curated targets for 2,275,906 compounds from 69,861 publications.

ChEMBL provides structured annotation that can be used to obtain a drug of interest. The Anatomical Therapeutic Code (ATC code) is the drug naming system controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC). It provides classification for drugs based on the active ingredients as well as the organ or system through which they elicit therapeutic or pharmacological effects, along with the drug chemical properties.

DrugBank

DrugBank (`https://www.drugbank.ca/`) holds comprehensive and molecular information about drugs, their mechanism of action, interactions and their targets. Its latest release (version 5.1.3) contains 13,336 drug entries including 2593

approved drugs, 1288 approved biotech (protein/peptide) drugs, 130 nutraceuticals as well as over 6,304 experimental drugs. The database comprises 5,175 non-redundant proteins which includes enzymes, transporters, carriers [79].

1.5.3 Resources with information on protein domain classifications

Eukaryotic proteins are typically made up of one or more domains. Domains represent distinct structural and/or functional units of a protein. Evolutionary related domains tend to have related functions and can be used to provide functional annotations for the whole protein sequence. Whole proteins can be decomposed into domains either based on sequence or structure or both. Pfam [80] is a sequence based classification of protein domains while CATH [81] is a structure based classification and is used for the analysis reported in this thesis.

PFAM

Pfam classifies whole proteins into domains using sequence information. A Pfam entry comprises a seed alignment that forms the basis of the hidden Markov model (HMM) using HMMER software [82]. The profile HMM is then queried against a sequence database (pfamseq) and all matches above a certain threshold are re-aligned back to generate a full alignment. The pfamseq database is derived from UniProtKB [83].

Each entry in Pfam is tagged with one of six types: family, domain, motif, repeat, coiled coil or disordered which indicates the class of the functional unit represented by the entry. The current version (Pfam version 32) has a total of 17,929 entries with 77.2% of UniProtKB having at least one Pfam domain [80]. Pfam contains two types of family: Pfam A which is the high quality manually curated family and the automated derived Pfam B. Relatives in a Pfam family are thought to be evolutionary related and share some degree of functional similarity.

CATH

CATH classifies protein domain structures using manual curation guided by various classification and prediction algorithms, including structural comparison and

hidden-Markov models (HMM). Domains are classified into Class (C-level), Architecture (A-level), Topology (T-level) and Homology (H-level). The Class (C) is the first level of the hierarchy based on the content of the secondary structure such as mostly alpha-helical (Class-1), mostly beta-sheet (Class-2), and domains with significant amounts of both alpha and beta secondary structure (Class-3). Class-4 represents domains with very little secondary structure [84]. Class is subclassified into architecture (A-level) where protein domains sharing similarity in their arrangement in 3D space are grouped together. The next level is the topology or fold group (T-level) which takes into account the connectivity of the secondary structures. Finally, the H-level classifies domains within the same fold group into homologous superfamilies where there is evidence of an evolutionary relationship (based on similarities in their structure, sequence and/or functions).

In the current version (v4.2) of CATH there are over 95 million domain sequences and about 400,000 domains of known structures classified into $\sim$6000 superfamilies. The majority of the CATH superfamilies ($<$90%) are small in population and the domain relatives share similar structures and functions. However, there are a few ($<$5%) superfamilies that are very highly populated and account for about 50% of all the domains in CATH and these superfamilies exhibit large structural and functional diversity. All superfamilies are sub-classified into functional families (FunFams) in order to understand how protein functions are modulated by sequence and structural changes [85].

Functional Families (FunFams)

Functional families comprise groups of homologous sequences that share very similar functions and structures. Functional subclassification is achieved through hierarchical agglomerative clustering of the sequences in the superfamily using the GeMMA algorithm [86]. This generates a clustering tree which is then partitioned using the FunFHMMER algorithm [85] as described below.

Firstly, GEMMA clusters sequences having 90% sequence identity (S90) using CD-HIT [87] and then builds a multiple sequence alignment for each cluster using MAFFT [88]. The sequence profiles derived from the alignments of pairs of

clusters are then compared against each other using the COMPASS algorithm [89]. Clusters having similarity in sequence profiles above a given threshold are merged and alignments are generated for the merged clusters. Clustering continues until a single cluster is left creating a hierarchical clustering tree in a bottom-up format which is built from leaf nodes to the root.

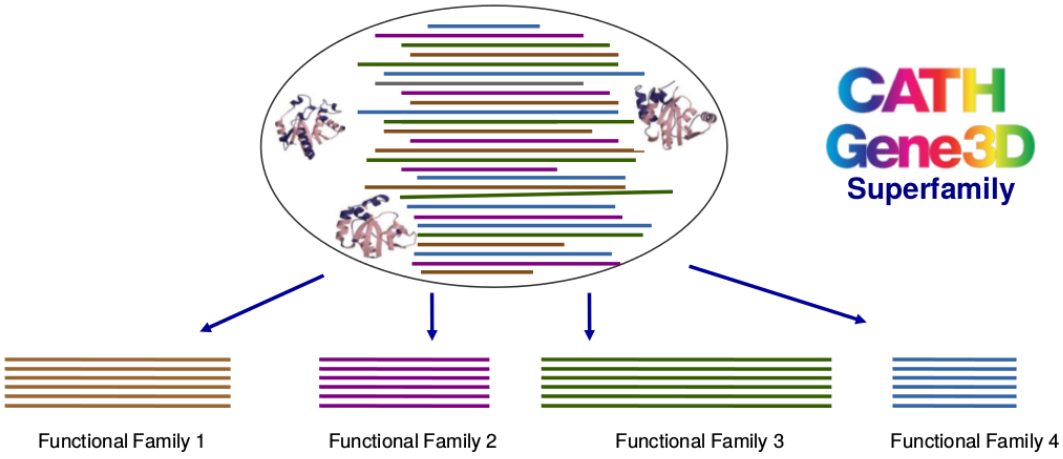

Figure 1.10: Schematic representation of sub-classification of domain sequence and structural relatives of a CATH superfamily into functional families (FunFams). Figure taken from [90].

To determine how to partition the tree, the FunFHMMER algorithm [85] identifies highly conserved positions and specificity determining positions (SDPs) that distinguish clusters from each other, based on their likely functions and therefore ensures functional coherence of relatives within a given cluster. Residues conserved across the multiple sequence alignments of both clusters are likely to be important for structural stability and folding whereas residues that are differentially conserved between the two clusters are likely to be specificity determining residues (SDPs) important for function. FunFams have been shown to be functionally pure and the performance in the CAFA (Critical Assessment of Functional Annotation) assessment protocol, gave an independent validation to the utilisation of CATH-FunFams to provide functional annotations to uncharacterised sequences [91]. In CAFA3, FunFHMMER was ranked the top method for predicting molecular function and the second best for predicting biological processes [92].

1.5.4 Sequence profiling tools

BLAST

Sequence search algorithms are used to search for evolutionary related sequences that share some degree of similarity. Basic Local Alignment Search Tool (BLAST) [93] is one of the most popular methods for performing sequence similarity searches. BLAST uses an heuristic method to find short matches between two sequences and attempts to start an alignment from these matches. BLAST can be applied in different modes depending on the application. For example, BLASTP [93] and PSI-BLAST [94] are frequently used for protein sequence comparison [95]. BLASTP performs a local protein-protein sequence comparison, PSI-BLAST builds profiles by collecting sequence matches from a large sequence database like UniProt through BLASTP. These sequences are then used to build a Position-Specific-Scoring-Matrix (PSSM) which is subsequently used to search the databases again for matches to more distant homologues. PSI-BLAST iterates three times or until no more matches are obtained within a given similarity cutoff.

HMMER

The requirement for a fast and sensitive sequence search method led to the development of HMMER software suite [96]. This contains several programs that carry out protein sequence similarity searches based on probabilistic methods called 'Profile Hidden Markov Model' (Profile HMM). Related sequences are first aligned to build a profile from a multiple sequence alignment (MSA) using 'hmmbuild' program. The profile is then used in searching large databases (such as UniProtKB) to find families or domains present in the sequences.

CD-HIT

The CD-HIT [87, 97] method is based on short word filtering and is a greedy incremental clustering program. Similarities between sequences are estimated by common word counting using word indexing and counting tables. The method first sorts sequences in order of decreasing length. The longest sequence is taken as the representative of the first cluster and the remaining sequences are compared. If a query

sequence is similar to the representative in the cluster (based on chosen criteria), the query sequence is added to the cluster. Otherwise, a new cluster is generated for which the query sequence becomes the representative sequence.

1.5.5 Structure comparison approaches

Several approaches have been used for comparing protein structures. The structural clustering approaches used in this work are discussed below.

SSAP

SSAP (Sequential Structural Alignment Program) by Orengo and Taylor [98] uses a double dynamic programming algorithm to compare the internal geometry between proteins. SSAP first compares the structural environment of residues. Equivalent pairs of the residues are selected based on the secondary structure, local conformation of the residues as well as the solvent accessibility. The structural environments are based on the Cβ atoms of the residues. The 'view' from each Cβ i.e. the distances to all other residues in the protein is represented as a vector.

Comparison of vectors is performed to score a 2D score matrix. Dynamic programming is then employed to find the optimal path through the matrix which is then added to a 2D summary scoring matrix. Another layer of dynamic programming is performed on the summary scoring matrix to determine the optimal path which gives the equivalent residues between the two proteins [98]. Figure 1.11 illustrates the SSAP algorithm.

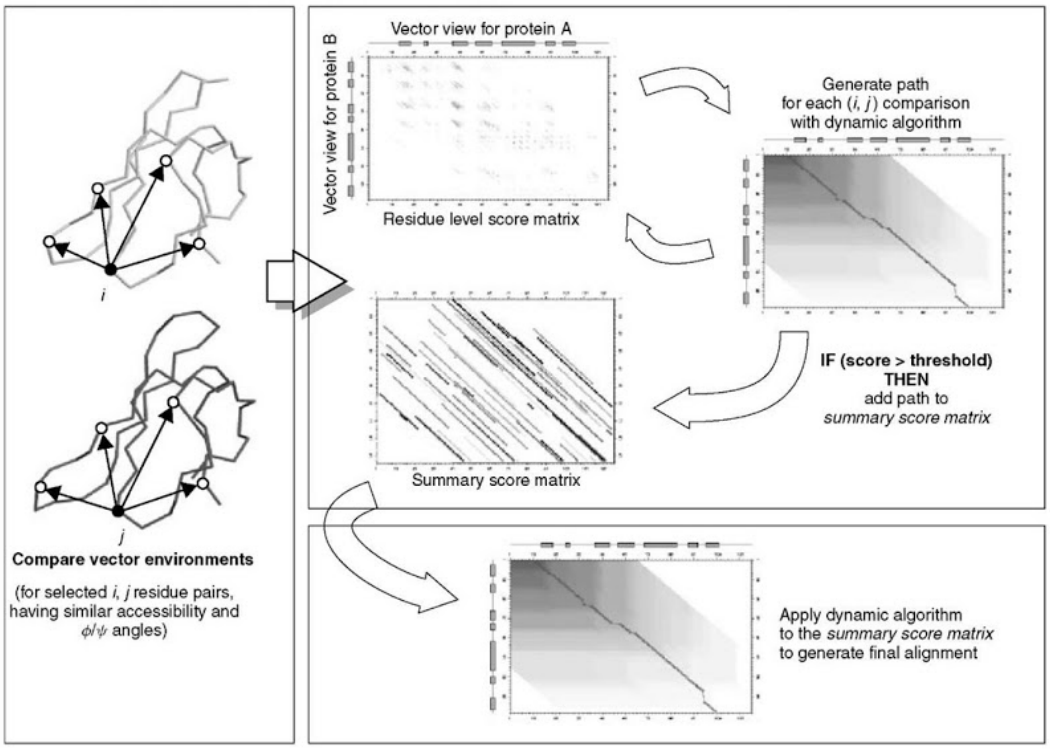

Figure 1.11: SSAP algorithm. Figures taken from [98]

ProFit

ProFit, developed by Andrew Martin [99], is a least squares fitting program that identifies the optimal superposition between protein structures given an alignment of proteins. It then calculates Root Mean Square Difference (RMSD) value as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{1.4}$$

where the mean is carried out over ”N” pairs of equivalent atoms and d_i is the euclidean distance between two atoms in the i-th pair. The lower the value, the higher the structural coherence between the structures. ProFit is incorporated in the cath-superpose tool used to compare domain structures within a given family. In cath-superpose, the alignment from SSAP is used as input in ProFit which then generates the RMSD values.

1.5.6 Resources providing functional annotation and pathway information

Several resources are used in this thesis to functionally annotate genes and provide information on pathways associated with clusters of genes. Enrichment studies are carried out to associate a given cluster of genes to a pathway or GO-term.

Gene Ontology

Gene Ontology provides a structured controlled vocabulary for gene products to be classified based on their function and cellular location [100]. Gene Ontology takes the form of a directed acyclic graph in which a functional term (child node) is sub-classified under one or more other general categorical terms called the parent terms. The branches within Gene Ontology are therefore a set of parent terms and all of its progeny. The Gene Ontology is made of three components/categories: Biological Process, Molecular Function and Cellular Component. A typical example is shown in figure 1.12 below.

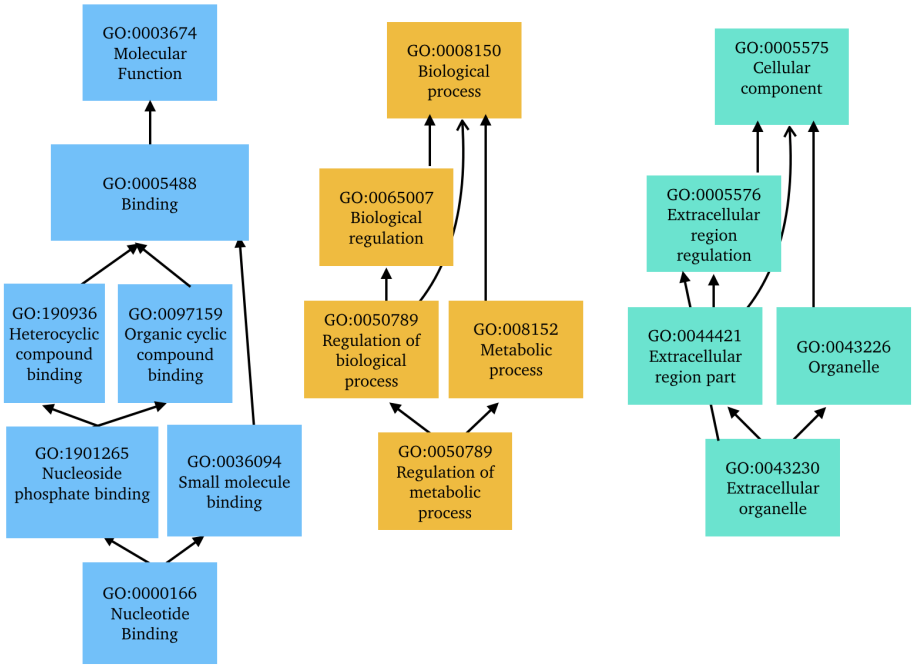

Figure 1.12: The three categories of GO terms: Molecular Function (MF); Biological Process (BP) and Cellular Component (CC). The dark black arrow shows 'is a' relationship.

Biological process refers to the biological objectives to which the genes or gene products contribute. Molecular function is defined as the biochemical activity carried out by the gene or gene products. Cellular components on the other hand refers to the place where the processes are carried out in the cell. There is one-to-many relationship between genes (or gene products) and the Gene Ontology terms associated with them which indicates the diversity of functions associated with each gene.

Kyoto Encyclopedia of Genes and Genome (KEGG) Pathway

The KEGG database is a comprehensive resource for assisting with biological interpretation of genes and gene products [101]. It provides annotation for genes such as Gene Ontology terms, pathways, diseases and drugs. Mapping of pathways to a gene set is done through the KEGG Object (KO) or KEGG identifier system in which each unique function or reaction is assigned to a given KO identifier and placed in an appropriate KEGG pathway map, BRITE hierarchy or KEGG-module based on experimental validation [102] and each gene is assigned to a given KO. KEGG has been widely used for interpretation of different data types including genomes, transcriptomes, metabolomes and metagenomes, for a variety of organisms.

1.6 Overview of Thesis

This thesis aims to predict disease genes and identify novel drug targets. Several novel computational protocols were used based on in-house structurally conserved domain functional families (CATH-FunFams).

The first work chapter of this thesis reports the mapping of drugs to CATH-FunFams to identify druggable CATH-FunFams based on statistical overrepresentation of drug targets within the CATH-FunFams. 81 druggable CATH-FunFams were identified and their propensity to be associated with side effects was predicted.

In the second work chapter, novel drug targets for bladder cancer were identified using a novel computational protocol that expands a set of known bladder cancer genes with genes highly expressed in bladder cancer and found to be linked

to the known bladder cancer genes by protein network analyses. 35 new druggable targets were identified with FDA-approved drugs available for repurposing subject to experimental validation.

In the final work chapter, a comprehensive classification and analysis of protein kinases was carried out to determine druggable kinase families (Kinase-FunFams) and their likelihood of being associated with side effects.

Chapter 2

Domain based approaches to drug polypharmacology

2.1 Introduction

The concept of one drug targeting multiple sites is known as polypharmacology and is gaining importance in the drug discovery process of the pharmaceutical industry. There is also considerable interest in repurposing clinically approved drugs furthermore to meet therapeutic requirements in diseases different from those they were initially designed for [103, 104, 105]. This concept is often referred to as drug repositioning or refocusing.

It has been found retrospectively, that most of the approved drugs elicit their therapeutic effects through a complex polypharmacological pathway [106]. This concept has been considered in targeting many complex diseases such as cancer and Central Nervous System (CNS) disease which have a wide target network comprising multiple cellular pathways. Drugs designed to target multiple proteins include the kinase inhibitors which have been considered of high efficacy in cancer therapy and about 37 FDA approved drugs are ATP directed protein kinase inhibitors used in the treatment of malignancies [107].

Other protein families successfully targeted through the concept of polyphar-

macology include the poly(ADP-ribose) polymerases (PARP) which are involved in the ADP-ribosylation of target proteins resulting in the regulation of several cellular mechanisms such as DNA repair, protein degradation and apoptosis [108]. Another prominent target family is the GPCR family. This family is of interest as a target in the treatment of some CNS diseases. CNS drugs such as anti-psychotics and antidepressants elicit their effects via a complex pattern of biological activities from multiple receptors [104]. Although the intrinsic promiscuity of a drug is partially responsible for its unintended side effects, these studies suggest that FDA approved drugs can be utilized for large scale repurposing [109, 110].

Studying protein networks is therefore valuable for the prediction of polypharmacological effects as networks can provide information on disease associated protein modules in which multiple proteins may share common domains which could be a target of the same drug [39]. Targeting protein modules in disease networks could help in selecting drugs with fewer side effects and the discovery of treatments for new diseases [39]. In selecting targets of polypharmacological drug, the approach should be directed towards identifying multiple targets within a disease module sharing a common domain and avoidance of promiscuous effects arising from off-targets. One of the key aims in the field of polypharmacology is, therefore, the development of pipelines to predict off-target activities and thus inform a better and safer approach in multitarget therapeutics.

Furthermore, Moya-Garcia and Ranea have also shown that drugs target domains in a more specific way than they target proteins partly because drug binding sites tend to be contained within conserved domains [111]. In this chapter, CATH-FunFams will be used to analyse the involvement of different domains in protein-protein interactions associated with disease networks. Domain information from CATH-FunFams will also be used to explore the importance of domains as a tool for understanding and exploiting the multi-target nature of drugs and their value in polypharmacology since domains represent the targeted entity in target identification during the drug discovery process.

2.1.1 The druggable Genome

For several years, researchers have been interested in mining proteins implicated in diseases whose modification through drug therapy can aid the treatment of such diseases. The druggable genome therefore represents a subset of the 30,000 proteins coded by the human genome that are able to bind drug/drug-like molecules [112]. Based on the review of pharmacological profiles, Drews identified 483 targets and further estimated there could be between 5000-10000 potential targets based on the estimation of the numbers of disease related genes [113]. However, this analysis did not focus on the properties of drugs that bound to those targets, and there are suggestions that focusing on ligand binding domains might even increase this number more than the 10,000 estimated [114].

Hopkins and Groom analysed the sites on the proteins binding with an endogenous small molecule. They identified some proteins targeted by experimental drugs (i.e. drugs that have not yet been approved but exhibit drug-like potential) and eliminated targets that lack activities based on the Lipinski Rule-of-Five (RO5)[115]. The RO5 was developed to set druggability guidelines for new molecular entities that can be considered as drugs. The RO5 revealed that poor permeation or absorption of compounds are likely to occur when there are more than five hydrogen-bond donors, the molecular mass is more than 0.5kDa, high lipophilicity (expressed as cLogP > 5); and the addition of nitrogen and oxygen atoms greater than 10. Their analysis of the sequences of the drug-binding domain as obtained from InterPro, revealed about 130 protein families that are associated with known drug targets.

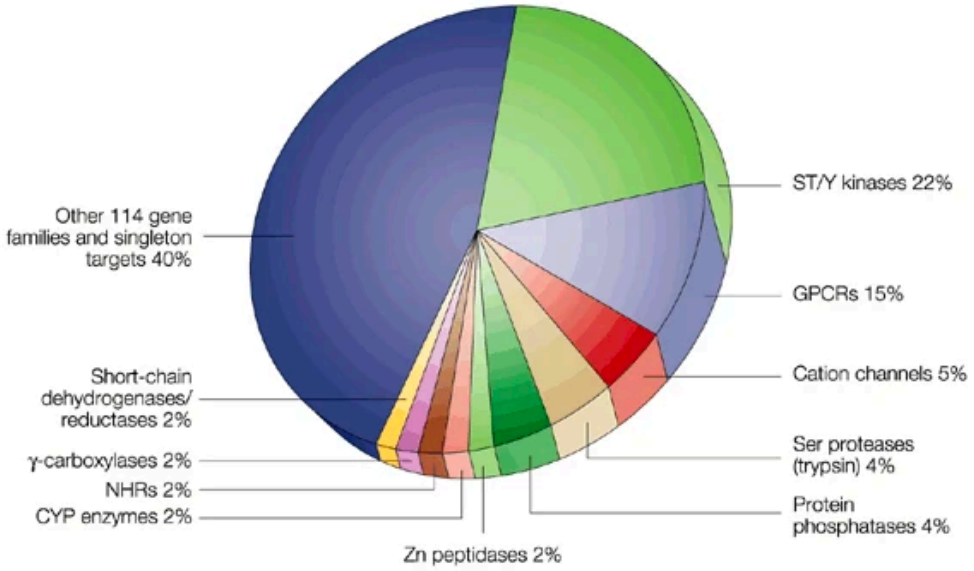

Figure 2.1: Gene-family distribution of the druggable genome as identified by Hopkins and Groom. Figure taken from [112].

Relatives in protein families tend to share sequence and functional similarities and this is generally indicative of conserved binding sites and conserved domain architecture of members of a given family. Using this approach of transferring drug binding information across families, about 3000 genes were predicted as druggable genes coding for proteins that bind drug-like molecules [112].

It is of note that the druggability of proteins does not indicate that they are drug targets. Drug targets have been identified as those druggable proteins that are linked to diseases and currently, there are less than 700 such drug targets targeted by FDA approved drugs [116]. Santos and colleagues also analysed all the human and pathogen derived biomolecules upon which about 1500 drug molecules act. They revealed the privileged target families which have a known long history in drug discovery processes. G-coupled protein receptors (GCPRs), nuclear receptors, protein kinases and ion channels were shown as the privileged families associated with the therapeutic effects for more than 70% of small molecules [116].

2.1.2 Assessing druggability

The term druggability goes beyond a strict adhesion to the Lipinski rule of five as these criteria are largely applied to oral drugs [117] i.e. the absence of compliance to the RO5 does not make a target undruggable. Druggability reflects the ability of a protein to bind a drug and drug-like compounds. Since this is a complex process, several approaches have been used in determining a druggable protein. Some of these approaches employ structural methods as well as machine learning in which several features are used to distinguish druggable targets from non-druggable ones.

Structure based approaches in target druggability assessment rely on the following procedure (i) Identification of potential binding sites; (ii) Exclusion of pockets based on physico-chemical properties; (iii) Using a reference set to then label the outcome and access performance [118]. Such physicochemical properties are used in training and also for reporting the quantitative assessment of druggability of a given target. The methods of assessing druggability can also involve calculating the energetics of protein-ligand binding through the use of docking procedures or molecular simulations. Although, these techniques are computationally expensive, they do not require any training and can be used in detecting unexplored targets.

2.1.3 Drug side effects

Targeting proteins by drug molecules, can sometimes lead to side effects. These are generally defined as a non-therapeutic, undesired, phenotypic response as a result of treatment with drugs. Side effects are often ranked as one of the leading causes of death and potentially a great threat to the pharmaceutical industry and drug development process. They are observed to occur as a result of the interaction of small molecules and the complex biological systems. Experimental profiling of side effects remains a challenge mainly because it is costly, tedious and time consuming. Computational techniques are therefore relied upon to help in predicting side effects associated with the drugs.

Duran-Frigola and Aloy, (2013) carried out enrichment studies to reveal features associated with side effects, using data obtained from biomedical resources

[119]. Although the prediction of side effects of drug is not straight forward as it involves an interplay between biological and chemical entities. Their studies revealed features associated with drug targets and off-target pathways, molecular function and biological processes account for about half of the cause of side effects. Their studies also revealed molecular fingerprints, scaffolds and other chemical entities accounts for roughly 6% of the side effects associated with drug compounds. This information was made available in the IntSide database where catalogs of the various proteins, cellular processes and chemical features that might be associated with undesired responses from chemical treatment [120] are made available.

SIDER is another side effect resource containing information on side effects extracted from the package insert of the drug using a Natural Language Processing approach [121]. Its current version (SIDER-4) has 1430 drugs, 5880 adverse drug reactions (ADRs) and 140,064 drug-ADR pairs.

There have also been studies on similarity in drug side-effects using protein interaction networks. In their approach, Brouwer and colleagues showed that the percentage of drug pairs with significant similarity in side effects was larger in those drug pairs sharing common targets in a protein network than those of drug pairs with non-overlapping proteins in the protein interaction network [122]. Their studies also revealed that side effects for a number of drugs can be explained by the subnetwork being targeted by the drugs which invariably means that common side effects are found in drug targeting similar subnetwork.

2.1.4 Systems polypharmacology

Systems pharmacology addresses the potential limitation of viewing drug actions from the perspective of a single magic bullet hitting a specific target. Thus, it exploits the multifaceted effect of drug binding to multiple targets which are involved in several biological processes and functions [111]. System pharmacology therefore combines several data including clinical observations and molecular pathways to gain insight into how drugs act and the possible side effects associated with drugs.

The concept of polypharmacology is often initially recognised as an unintended phenomenon as most drugs were designed without taking this into consid-

eration [105, 104]. However, drugs are currently intentionally developed in case of multifactorial diseases, to interact with multiple targets in order to give therapeutic advantage towards the treatment of the disease condition[123]. There are several pieces of evidence indicating that complex pathologies are polygenic in nature, in which extended networks of proteins are involved in driving the disease.

There is therefore considerable interest in polypharmacology. One area of interest concerns the negative side effects caused by the unintentional and unwanted interactions with off-targets (adverse polypharmacology). The second involves the positive synergistic effects of a drug hitting multiple targets (beneficial polypharmacology) while other areas of interest involve the redirection of drugs towards another valuable hit or lead for which it was not designed but has beneficial effects (drug repositioning). There are several computational approaches available to study the polypharmacological nature of a drug for targets [124]. Statistical analyses or data modelling approaches use machine learning to learn features of the ligands and targets and predict other possible targets. Structure-based approaches use methods such as docking, protein-ligand interaction, pharmacophore analysis, calculation of free-energy as well as binding site analysis and comparison to predict novel targets for a ligand. The third approach is a ligand-based method that carries out 2D and 3D analysis of the ligand and targets, calculating similarities between the ligands and targets to aid the prediction of novel targets for a given ligand. For effective outcomes, these three approaches can be combined together and the predicted targets can then be validated experimentally.

2.1.5 Objectives of chapter

Most drug targets in humans are proteins that are composed of more than one domain. Domains within families are evolutionary conserved and frequently combine to form different proteins in which some have different overall functions. Domains are involved in protein-protein interactions and also mediate interactions between drugs and targets. Previous studies have shown that drug binding sites are contained within protein domains [125, 126] and that protein domains mediate the drug-target interactions which indicates that protein domains are a major factor in the polyphar-

macology of FDA-approved and not yet approved (experimental drugs) [111]. This, therefore means that drug-binding proteins can be grouped based on the associated domain families aiding the redefinition of the druggable genome (a set of genes considered to be important pharmacological targets) [112, 127].

The in-house CATH domain resource classifies functionally similar homologues into functional families called the CATH-FunFams. Therefore, it is possible to hypothesise that CATH-FunFam domains contain the binding sites of drugs and mediate the interaction between proteins and drugs. A calculation of the statistical overrepresentation of drug-targets was carried out amongst the relatives of CATH-FunFams in such a way that if the targets of a drug belong to a CATH-FunFam, the CATH-FunFam was deduced to contain the binding sites for the drug and thus the drug was associated with the CATH-FunFam. This follows the similarity principle that drugs with similar structures have the same target whether in proteins or CATH-FunFams and this philosophy was used to classify druggable CATH-FunFams. This chapter describes an analysis of the association between drugs and CATH-FunFams and harnesses the study of these structurally and functionally coherent families for drug-target identification and possibly drug repurposing. Network dispersion of the relatives of the druggable CATH-FunFams was also carried out and was used to reveal the likelihood of druggable CATH-FunFams being associated with side effects.

2.2 Materials and Methods

2.2.1 Drug-proteins dataset

Human drug targets were obtained by querying ChEMBL release 21 [78] a database that links chemical and biological targets. FDA approved drugs was selected from the ChEMBL database which provides a scoring scheme from 1-4 where 4 indicates "Approved drugs". Filtering of the drug-protein interaction data was such that all weak interactions of activity less than 1μM were removed. ChEMBL gives a value of the activity of the drug as the half-maximal response potential on a negative logarithm scale where activity includes IC_{50}, EC_{50}, Ki, Kd. For instance, an IC_{50} of $1n$M would be given as $pChEMBL$ of 9. Therefore, the drug off-targets were

classified as those with low affinity of $pChEMBL < 6$ while drug targets were those with $pChEMBL \geq 6$.

2.2.2 Identifying CATH-FunFams with overrepresentation of drug targets (Druggable CATH FunFams)

Protein domain information from the CATH-FunFams v4.1 from CATH-Gene3D v12.0 [128] was used for this analysis. The targets T $[T_1, ...T_n]$ of drug d were evaluated to determine whether they are significantly overrepresented among the relatives of a CATH-FunFam P $[P_1, ...P_n]$. This simply means that CATH-FunFams were tested for enrichment of targets of drug d. A list of all drug targets was obtained and each was annotated according to whether the drug target was a member of a CATH-FunFam. From this list, the expected probability value that any drug target is a relative of the CATH-FunFam was compiled.

The overrepresentation of the targets of a drug among relatives of a CATH-FunFam depends on the expected probability that a protein belongs to the CATH-FunFam. This probability is defined for each CATH-FunFam as the fraction of drug targets that belong to the CATH-FunFam. For example, let's assume there are 10000 drug targets and 200 of them are relatives of CATH-FunFams, the expected value is 0.02 i.e. 2%. If 100 proteins are the targets of a drug, it means that 2 of them are expected to be CATH-FunFam relatives but if more than 2 of the drug targets were relatives of the CATH-FunFam, the targets of the drug are said to be overrepresented in the CATH-FunFam.

A p-value (Benjamin-Hochberg correction of multiple testing) was calculated to determine whether each observed overrepresentation is statistically significant by means of a binomial test. The binomial test evaluates the statistical significance of deviations from the binomial distribution of observations that fall into two categories: (i) the protein is a relative of the CATH-FunFam under consideration, or (ii) the protein is not a relative of the CATH-FunFam under consideration. The binomial distribution is the discrete probability distribution of the number of successes in a sequence of independent yes/no experiments each one with defined success probability. In this case the sequence of independent experiments is T, the targets

of drug d; a success is that a protein from T is a relative of the CATH-FunFam under evaluation. Each individual success has a probability value P_{FF} which is the expected probability that a protein is a relative of FF

$$P_{FF} = \frac{n_{FF}}{N} \tag{2.1}$$

where n_{FF} is the number of relatives of the CATH-FunFam FF and N is the total number of proteins that are relatives of all CATH-FunFams (i.e. all human proteins).

The null hypothesis is that the proteins in T are sampled from the same general population as the proteins in CATH-FunFams P, and thus the probability of observing a target of d as a relative of FF is the same as observing any protein as a relative of FF i.e. P_{FF}. Therefore, the p-value of the binomial test indicates if observing proteins from T in the test list P is likely to happen by chance. A confidence level of 0.95 was used in this study, hence, if p-value $<$0.05, the null hypothesis is rejected and it is considered that the probability of observing the targets of d among the relatives of FF is different from the probability of observing any set of proteins among the relatives of FF. Therefore, the p-value reported indicates if observing the targets of d among the relatives of FF is likely to happen by chance. For p-values $<$ 0.05, the corresponding drug-CATH functional family association was considered to be statistically significant and not likely to happen by chance.

For example, let's consider FF with P relatives and d with T targets, then:

- Success: number of targets from T that are in P
- Trials: number of targets of drug d in T
- Probability of success under the null hypothesis: $P_{FF} = \frac{n_{FF}}{N}$
- Overrepresentation threshold: Trials x P_{FF}

The tables below illustrate the overrepresentation of the targets of a drug with a confidence interval of 0.95 across four CATH-FunFams in two cases. For all cases, there are 25 total targets.

Table 2.1: Drug targets are only enriched in one CATH-FunFam.

FunFam	n_{FF}	P_{FF}	Success	Trials	Overrep. Threshold	p-value
FF_1	6	0.24	0	7	1.68	0.28
FF_2	3	0.12	1	7	0.84	0.59
FF_3	7	0.28	6	7	1.96	0.003
FF_4	9	0.36	0	7	2.52	0.054

From the Table 2.1 above it is observed that the targets of the drug are overrepresented on FF2 and FF3 but the overrepresentation is significant only in FF3.

Table 2.2: Drug targets are not enriched in any CATH-FunFam.

FunFam	n_{FF}	P_{FF}	Success	Trials	Overrep. Threshold	p-value
FF_1	6	0.24	1	5	1.2	1
FF_2	3	0.12	2	5	0.6	0.11
FF_3	7	0.28	1	5	1.4	1
FF_4	9	0.36	1	5	1.8	0.66

The targets of the drug shown in Table 2.2 are overrepresented in FF2 but with no statistical significance.

2.2.3 CD-Hit and SSAP

Structural comparisons of relatives across each druggable CATH-FunFam were performed to determine the structural coherence of the FunFam. This was done using the SSAP algorithm [98]. Since SSAP is computationally expensive, representatives of the CATH-FunFam were compared. Relatives were clustered using CD HIT [87] which applies a greedy incremental clustering algorithm method to cluster protein sequences. CD-HIT was also carried out to correct the bias within the data set of the structural representatives within the members of the druggable CATH-FunFams since some CATH-FunFams contain many nearly identical relatives.

Relatives were clustered at 60% sequence identity. At this threshold, relatives share significant structural and functional similarity. Structural similarity between the representatives was measured using the SSAP alignment to superpose the relatives using the ProFit algorithm [99]. An RMSD value was measured from the

superposition and normalised using the following equation:

$$NormalisedRMSD = \frac{(MaxL1, MaxL2) \times RMSD}{Number\,of\,aligned\,residues} \quad (2.2)$$

MaxL1 or MaxL2 represents the length of the longest domain between domain1 and domain2.

2.2.4 Ligand binding site conservation in the druggable CATH-FunFam

Druggable cavities in domain relatives in a CATH-FunFam were detected using the Fpocket method [129]. Fpocket is a fast protein pocket prediction algorithm that helps identify cavities on the surface of a protein and rank them according to their ability to bind drugs and drug-like molecules.

The relatives of druggable CATH-FunFam were examined to determine whether they have similar binding pockets and similar amino acid residues when interacting with the bound drug. Again, the structural domains from different CATH-FunFams were pairwise structurally aligned using SSAP. SSAP scores were used to construct a distance matrix and a maximum spanning tree was then used to derive a multiple superposition of the structural relatives. Data on residues involved in the binding the drugs of interest were extracted from the NCBI IBIS resource [130] using the PDB IDs of the drug-target complex.

2.2.5 Protein interaction data

Human protein interaction data was obtained from the STRING database version 10.0 [131] and filtered based on the confidence score. The STRING database provides a scoring scheme from 0-1000 in increasing order of confidence (reliability) of the interaction. The experimental score of 800 or above was used to limit the number of false positive interactions. The resulting human protein network from STRING version 10.0 contains 13,460 nodes (proteins) and 141,296 edges (interactions). The edges in this case were unweighted and undirected.

2.2.6 Transforming the protein network

A kernel transformation was carried out to create a similarity matrix on the STRING network and this approach was used to measure the dispersion or separation of each protein in the network. Various transformations were explored but the adjacency matrix was found to be the most effective. The matrix constructed was used to investigate how close the proteins are in the network. The higher the matrix similarity measure, the closer the nodes are in the network. The STRING network (v10) was downloaded and the full network was transformed into an adjacency matrix. The value in row (i), column (j) had a STRING combined score (0-1000) between protein (i) and protein (j). Protein interaction data from the STRING database was chosen because it is widely used and frequently updated.

2.2.7 Network centrality measures

The network centrality is a measure of the importance of a certain node in the network topology. Central nodes are important nodes around which the network revolves. Drug targets have been shown to exhibit differential behaviour on a molecular network occupying central positions and connecting functional modules [132]. Amongst the different measures of centrality, betweenness centrality best captures the ability of important nodes to be 'between' functional modules and also captures the essentiality of a protein in a biological system [133]. The betweenness centrality (BC) represents a measure of the total number of non-redundant shortest paths going through a certain node. Nodes with high BC are said to be central as they control the communications amongst other nodes within the network. The network betweenness centrality (BC) of the targets was measured using the NetworkX package in python.

A random set of drug targets was generated by randomly selecting sets of 1000 proteins, 1000 times from the set of non-druggable proteins. The non-druggable set is defined as a group of proteins excluded from the list of the selected human drug targets in ChEMBL. The average betweenness centrality of the random (non-druggable) targets was measured and compared to the average betweenness central-

ity for the drug targets. The betweenness centrality of the druggable CATH-FunFam relatives was also compared to the non-druggable FunFam relatives. In this case, the non-druggable FunFams are other randomly chosen CATH-FunFams that do not contain any of the targets identified from ChEMBL database. For the CATH-FunFams, the median BC of each FunFam was considered as a representative value of the BC of the relatives of the CATH-FunFam. This assumption was applied to reduce the bias that might be associated by using the mean of the CATH-FunFam.

Hubs and bottlenecks in protein network

The proportion of hubs, hub-bottlenecks and non-hub bottlenecks in the network was measured and compared with non-druggable random proteins. The hubs and bottlenecks of the network were calculated following the method described by Gerstein *et al* which classifies hubs as proteins with 20% of the degree distribution (i.e. those proteins having the highest numbers of connecting neighbours). The bottleneck nodes on the other hand represent a set of proteins described in terms of the betweenness i.e. those proteins that are in the top 20% of interactions with a highest node betweenness centrality measure [133].

Side-effect data and topological characteristics of the protein network

Side effect data was collected from IntSide [120] (a database that integrates the biological and chemical information associated with drugs and uses this information to understand of the molecular mechanisms underlying drug side effects. IntSide catalogues the side effects associated with various drugs. Side effect data was included for all drugs involved with protein interactors (either targets or off-targets). The CATH-FunFams which are more likely to be associated with side effects were determined.

To do this, a logistic regression model of the probability of the CATH-FunFams being free of drug target proteins associated with side-effect protein was built based on the protein network median matrix similarity of these domain families. According to this statistical model, the probability that a CATH-FunFam, that has relatives completely dispersed in the protein interaction network (i.e. matrix similarity = 0), and having side effects was compared with the matrix similarity of other CATH-

FunFams, and the number of side effects associated with them. The threshold of matrix similarity, for which there is a high probability that 50% of the relatives of the CATH-FunFam are not associated with side effect, was obtained. The general equation of logistic regression takes the form:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{2.3}$$

Where π(x) is the probability of the presence of a side effect protein in a CATH-FunFam given its median similarity (x).

2.3 Results and discussion

2.3.1 Drug-Enrichment Analysis

There are 17,229 CATH-FunFams containing 77,082 human proteins in CATH-v4.1. The median number of relatives per CATH-FunFams is 3 but a few of them are highly populated such as the MHC class 1 antigen FunFam (3.30.500.10.FF3475) containing 14% human proteins amongst its relatives. A drug-target dataset was compiled by querying ChEMBL for approved drugs and the human proteins to which they bind directly at high affinity.

A set of 787 human proteins capable of binding drugs was identified and are distributed in 875 CATH-FunFams. For each drug, the statistical overrepresentation of their targets in each of the CATH-FunFams was computed. This gave a mapping of 359 statistically significant associations (Benjamin-Hochberg false discovery rate p-val$<$0.001) between 245 drugs and 81 CATH-FunFams which are therefore called the druggable CATH-FunFams.

2.3.2 Proportion of known druggable classes in the druggable CATH-FunFams

The distribution of protein functional classes in the druggable CATH-FunFams was compared with the druggable genome as defined by Hopkins and Groom [112]. The functional roles of the proteins in the druggable CATH-FunFams were obtained by extracting the functional terms from their UniProt keywords.

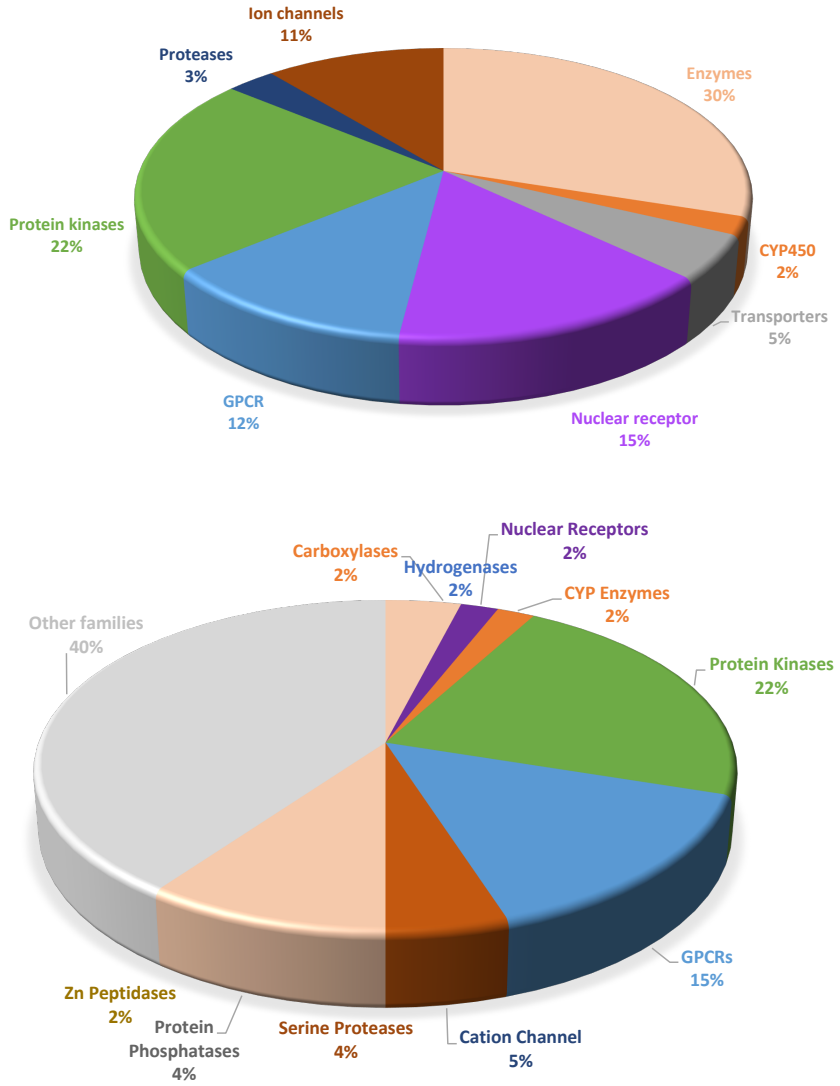

Figure 2.2: Distribution of the protein classes found in the druggable CATH-FunFams shown in 2.2a, compared with distribution of protein families in the druggable genome compiled by Hopkins and Groom (2002) shown in 2.2b

Figure 2.2a shows the proportion and distribution of protein classes associated with the druggable CATH-FunFams. This can be compared with the study shown in 2.2b carried out by Hopkins and Groom, who defined protein families in the druggable genome and estimated that these families represented less than 10% of the whole human genome. As in the Hopkins and Groom study, it was found that certain classes of proteins such as the kinases, ion-channels and enzymes are enriched with druggable CATH-FunFams. However, in contrast to Hopkins and Groom, a high proportion of GPCRs was not found. This may be because of the smaller cov-

erage of transmembrane GPCRs in CATH-FunFams. Rhodopsin-like GPCR, ion channels, protein kinases and nuclear receptors were categorised as privileged families by Santos and colleagues because they account for close to 50% of the drug targets with most of the others being small families with drug targets potential [116].

2.3.3 Structural similarities of the relatives in the druggable CATH-FunFams (CD-HIT and SSAP)

To understand how structurally coherent relatives of the druggable CATH-FunFams are, structural similarity assessment was carried out. This was done by comparing the relatives selected as representatives of 60% sequence identity clusters within the CATH-FunFams. Structural representatives were compared using the structural comparison algorithm SSAP [98]. The alignment generated from SSAP was used to calculate the Root Mean Square Difference (RMSD) by using the alignment as input to the ProFit structure superposition algorithm [99].

The RSMD scores measure the dissimilarity between protein structure, hence, the lower the score, the more structurally coherent the proteins within the families. A normalised RMSD was subsequently generated and the results are shown in figure 2.3 below.

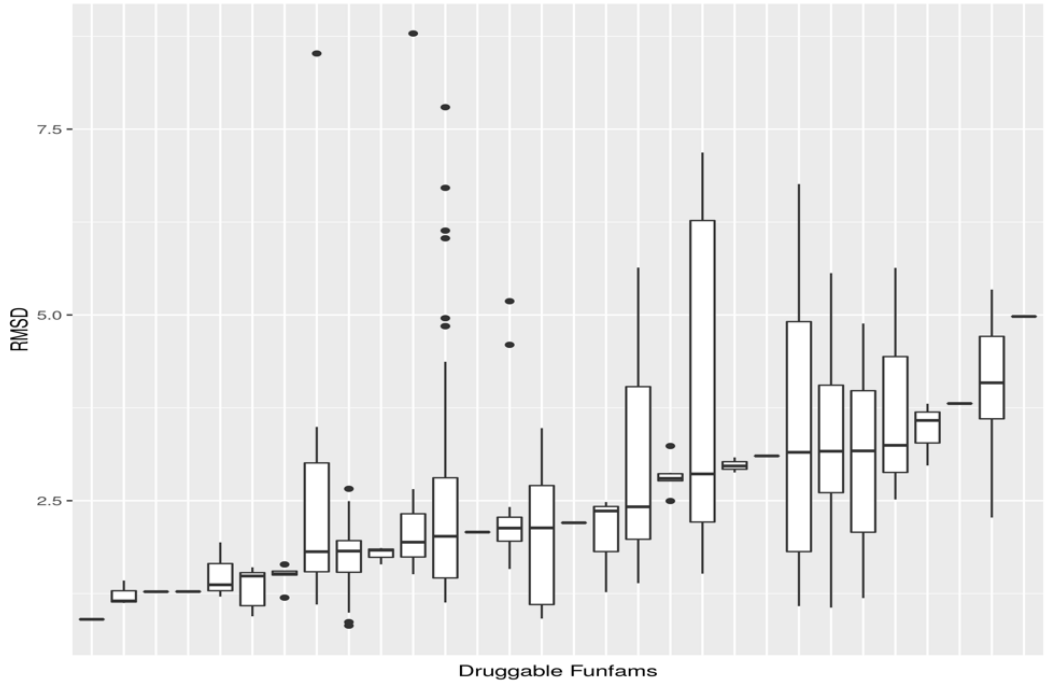

Figure 2.3: The normalised RMSD for 30 druggable CATH-FunFams with two or more structures.

Figure 2.3 shows the normalised RMSD values obtained for each of the druggable CATH-FunFams. The median RMSD value is indicated by the black horizontal line. 75% of the druggable CATH-FunFam have a median value below 3$\mathring{A}$ reflecting the structural coherence of these families. Subsequently, conservation of binding sites in members within the same CATH-FunFams was examined.

2.3.4 Structural superposition and conservation of drug binding sites in CATH-FunFams

The relatives within the 81 druggable FunFams were assessed for similarity in drug binding sites. This was done by first examining 57 CATH-FunFams that had crystal structures in Protein Data Bank (PDB), for their enrichment in drug cavities. Comparison was made with a set of 100 random non-druggable CATH FunFams of which 63 had structures in the PDB. 75% of the 57 druggable CATH-FunFams with structural information available have druggable cavities whereas only 66% of the random non-druggable CATH-FunFams have cavities capable of binding drugs. Thus, druggable CATH-FunFams have a greater proportion of cavities able to bind

drugs and drug-like molecules (p-val<0.0001, Fisher exact test).

The conservation of drug binding in human proteins that are associated with CATH-FunFams was analysed using selected examples of drug-target complexes present in Protein Data Bank (PDB). Six examples of complexes between drugs and CATH-FunFams are shown in figure 2.4.

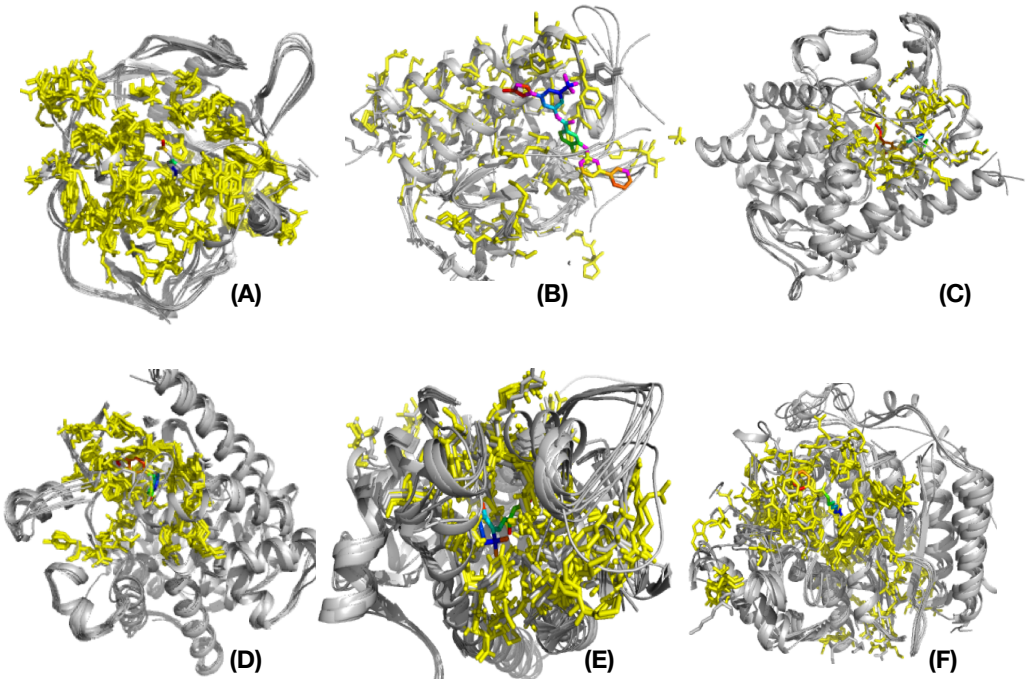

Figure 2.4: Conservation of the drug binding site within CATH-FunFams. Structural alignment of the human CATH-FunFam domains associated with: A) acetazolamide (CATH ID: 3.10.200.10-FF1430; carbonic anhydrase II), B) nilotinib (CATH ID: 1.10.510.10-FF78758; ABL kinase), C) Sildenafil (CATH ID: 1.10.510.10-FF78946; Myosin light chain kinase), D) tadalafil (CATH ID: 1.10.1300.10-FF1260; 3'5'-cyclic nucleotide phosphodiesterase), E) Tretinoin (CATH ID: 1.10.565.10-FF5060; Retinoid X Receptor) and F) vorinostat (CATH ID: 3.40.800.20-FF2855; Histone deacetylase domain) and the drug complex. The protein domain is shaded grey except for the ligand binding residues, which have been mapped across the domains, coloured yellow. The drug molecules are coloured in rainbow.

The domains within the CATH-FunFams are associated with the drugs: acetazolamide (3ML5), nilotinib (3CS9), sildenafil (1UDT), tadalafil (1UDU), tretinoin (2LBD) and vorinostat (4LXZ). The human protein domains from each CATH-

FunFam were aligned pair-wise using SSAP [98] and then superposed. The drug binding residues were obtained from IBIS and mapped onto members of each CATH-FunFam. The relatives of each CATH-FunFam were found to be highly conserved in their amino acid residue types binding the drug and structural location of the drug binding site. The mean RMSD for the aligned domains across all six CATH-FunFams is $1.169 \pm 0.812\mathring{A}$ illustrating considerable structural coherence.

2.3.5 Aggregation of drug targets in the human protein functional network

Drug targets tend to be centrally located and aggregate in the protein network [134, 126]. To measure the aggregation of targets in a protein functional network, the STRING protein functional network was used to measure the network distance between targets of a drug. STRING provides combined scores between proteins and this was used in deriving a similarity matrix between proteins. The similarity matrix values reveals how connected any two proteins are in the protein functional network i.e. the higher this value, the more strongly connected the proteins in a functional network.

Drug targets and off-targets were separated based on the affinity reported in ChEMBL, classifying those with high affinity ($pChEMBL \geq 6$) as targets and those with low affinity ($pChEMBL < 6$) as off-targets. The dispersion of the targets and off targets in the network was also calculated based on the matrix similarity.

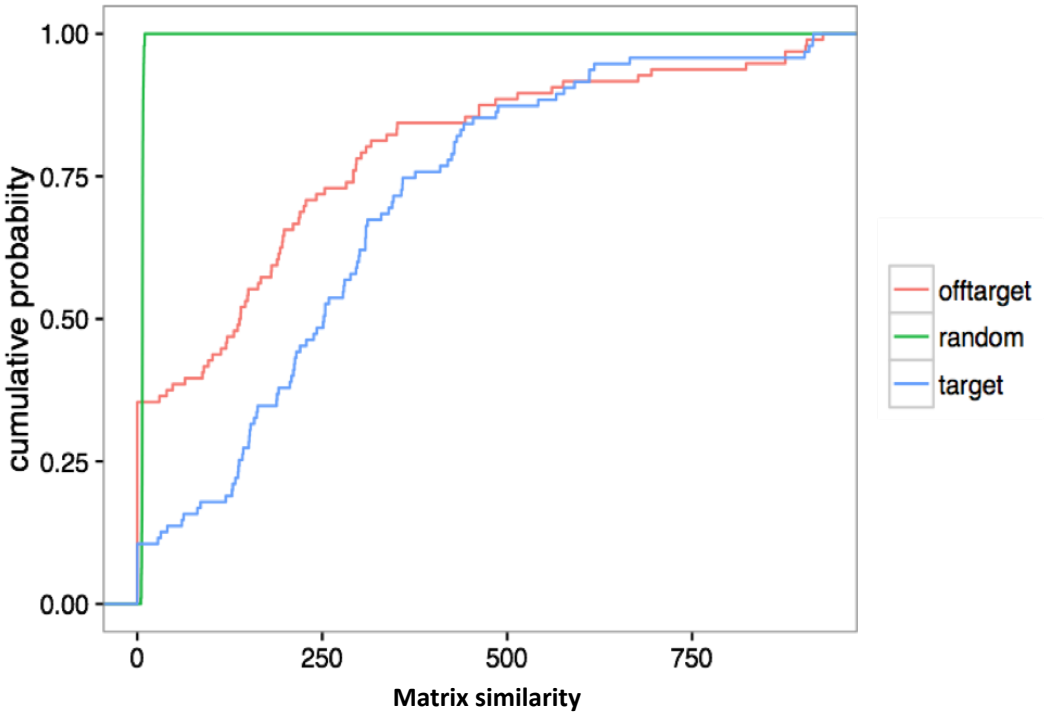

Figure 2.5: The drug neighbourhood in a protein functional network. Cumulative distribution of the matrix similarity of random proteins (green line), off-targets (red line) and targets (blue line) in a human protein functional network

Drug targets have a higher matrix similarity than off-targets and both the drug targets and the off-targets have higher matrix similarity than expected by chance. This indicates that the drug targets tend to aggregate in functional sub-networks forming modules. Since modules in a functional network imply proteins involved in the same biological process or function, it is expected that the interaction of the drugs and their targets will result in alteration of one or more biological functions.

The ability of drug targets to form modules was also measured by using the network distance based metrics developed by Menche *et al.* [5]. The DS-score measures the mean distance of the genes within a given cluster. The lower the score, the more clustered the set of genes are relative to others. The DS-score is marginally significantly lower (p-val = 0.01008) for the drug targets than the off-targets which suggest that the drug targets may tend to be clustered together in the networks. It should be noted that there is high likelihood that this significance may not hold if the DS-score is subsampled multiple times. However, from this current analysis,

these two measures i.e. DS-score as well as the matrix similarity score thus suggest that the drug targets are more clustered within the network than random.

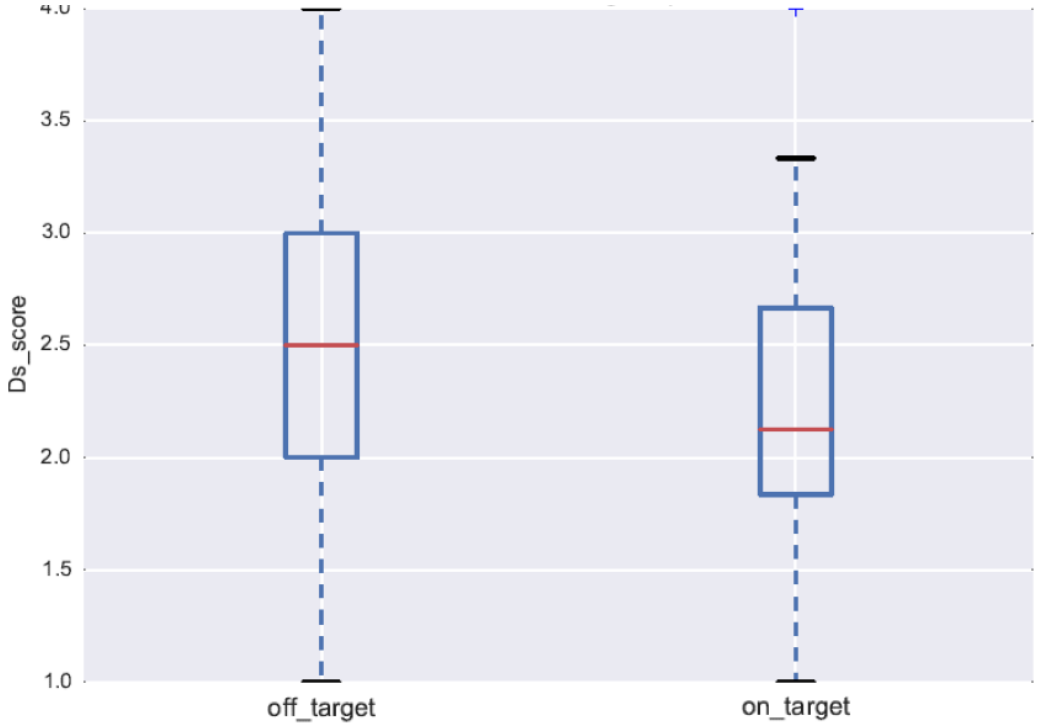

Figure 2.6: DS-score measure of off-targets and targets in a human protein interaction network

2.3.6 Topological characteristics of proteins with side effects

The topological characteristics of proteins with side effect data collected from IntSide [120] was analysed in a human protein network to determine network characteristics associated with side effects and compared with drug targets. The betweenness centrality captures the ability of a node to be important and 'between' functional module and it also reveals essential nodes in a network [133].

Analysis of the betweenness centrality of drug targets and proteins associated with side effects, revealed that proteins associated with side effects have a higher betweenness centrality in the network as compared to the drug targets and non-targets. Using a different network but a similar approach, Wang *et al.*[126] also showed a positive correlation between the numbers of side effects and the betweenness of the drug targets.

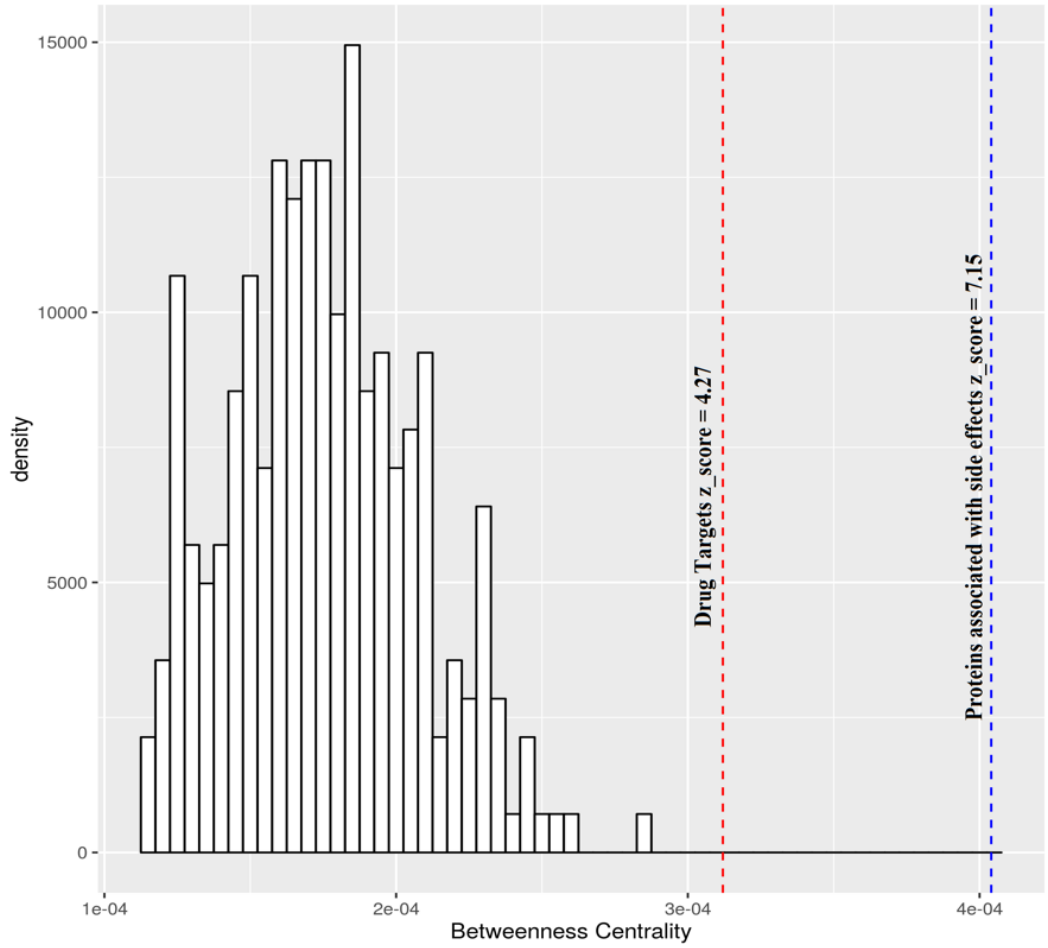

Figure 2.7: Betweenness centrality of drug targets. The mean betweenness centrality of drug targets (red line) and proteins associated with side effects from IntSide (blue line) in a functional network is compared with the distribution of the mean betweenness centralities of random protein sets.

To further establish the association between the side effects of proteins and network topological characteristics, the proportion of side effect proteins associated with hubs, bottlenecks and non-hub bottlenecks was measured. About 40% of target proteins in hubs are associated with drug related side effects while 16% of proteins found to be hub- bottlenecks are associated with drug related side effects. However, the proportion of target proteins associated with side effects for non-hub bottlenecks was lower, with a value of 8%. This suggests that the non-hub bottlenecks although essential proteins within the networks, can be considered as interesting drug targets as they would possibly lead to lesser side effects. In summary, this analysis suggested that network characteristics can be used in screening for possible side effects

associated with a given drug, by considering its possible targets in the network.

2.3.7 Proximity of druggable CATH-FunFam relatives in the human protein network

The previous section show that drug targets tend to be clustered in the human protein functional network. This suggests relatives from druggable FunFams may also be clustered together and thus expected to have a relatively high matrix similarity. To test if this holds true, a similarity matrix was constructed (see methods), with similarity scores ranging from 0-1000 and used to measured how clustered relatives from druggable CATH-FunFams are.

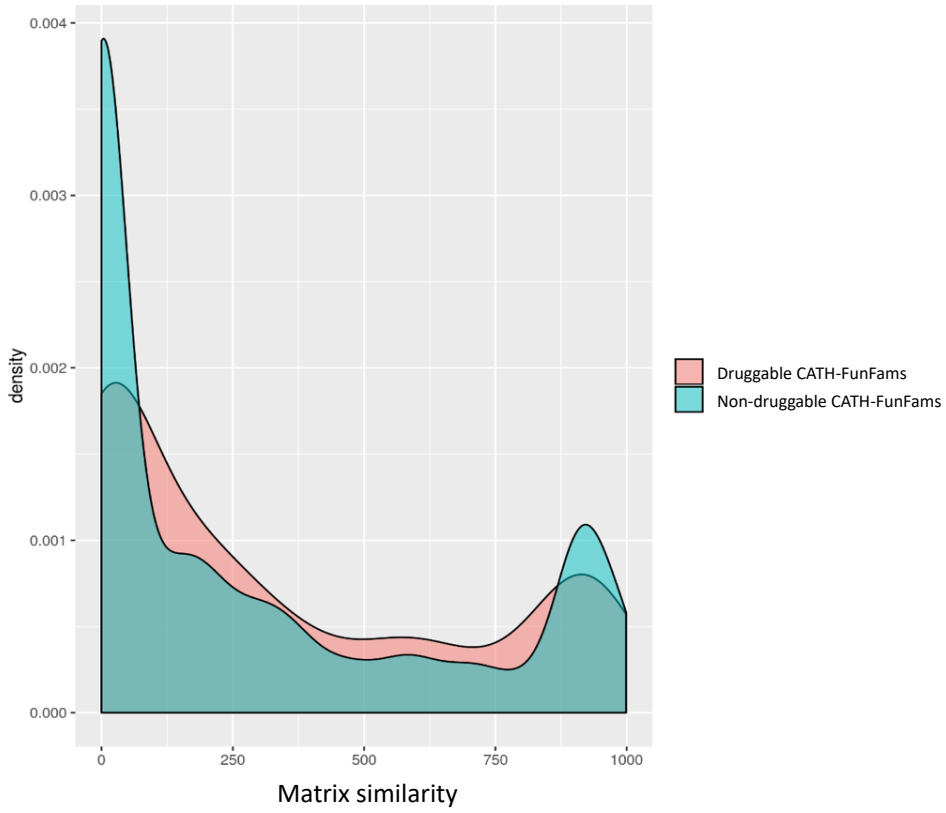

Figure 2.8: Density plots of the proximity of relatives from druggable CATH-FunFams in comparison with relatives from non-druggable CATH-FunFams

Figure 2.8 shows the matrix similarity distribution (0-1000) for pairs of relatives in the druggable CATH-FunFams. The druggable CATH-FunFams tend to be more clustered in a protein functional network and have higher matrix similarity scores than the non-druggable FunFams (median matrix similarity 540 and

230 respectively; Mann-Whitney Wilcoxon test p-val <0.01). CATH-FunFams are designed to cluster relatives sharing similar functional determining position and it links sequence patterns to function [135]. The observation in this analysis shows that this does not translate into proximity in a functional network. However, it is safe to assume that protein domains in the druggable CATH-FunFams have been recurrently targeted in drug design because of their lower association with side effects. Generally, CATH-FunFams whose relatives are much dispersed in the network and associated with side effects are likely to be involved in more generic functions.

Non-druggable CATH-FunFams have mostly very low similarity scores but there were some interesting peaks (those with matrix similarity $\geq$ 500). The relatives of these Non-druaggable CATH-FunFams were obtained and analysed using the DAVID Functional tool [136]. Diverse functional clusters were found: 15% nucleotide binding proteins, 4% show EGFR-like activity, 26% indicate metallic binding activity while others are varying functional clusters of calcium binding, proteolytic activity, RNA-mediated gene silencing, etc. (Table 2.3).

Table 2.3: DAVID's functional annotation tool for terms with high enrichment score

Term	Count	%	P-Value
Nucleotide-binding	439	15.5453	9.586E-40
Extracellular matrix	87	3.0807	2.754E-16
Epidermal growth factor-like domain	105	3.7181	3.074E-28
Focal adhesion	89	3.1515	4.894E-14
Potassium channel	39	1.3810	8.553E-15
calcium-binding	29	1.0269	7.046E-17
Metal-binding	746	26.4164	1.361E-38
Zinc-finger	364	12.8895	1.230E-17
Cadherin 2	43	1.5226	7.037E-10
DNA damage	89	3.1515	4.356E-9
Laminin G domain	33	1.1685	4.338E-12
Fibronectin type-III	37	1.3101	6.095E-11

This suggests that the CATH-FunFams that were classified as non-druggable might include some orphan druggable classes with potential drug–ligand activity but currently lacking drug binding data or other experimental information confirming this at present.

2.3.8 Topological Features of CATH-FunFam relatives

The betweenness centrality of relatives from the druggable CATH-FunFams was measured and compared with random non-druggable FunFams. The relatives within the druggable CATH-FunFams were examined if they were more likely to form bottlenecks or hubs in the network than relatives from random CATH-FunFams. Proteins with high betweenness centrality are considered to be bottlenecks while those with high degree are hubs (See method section 2.2.7).

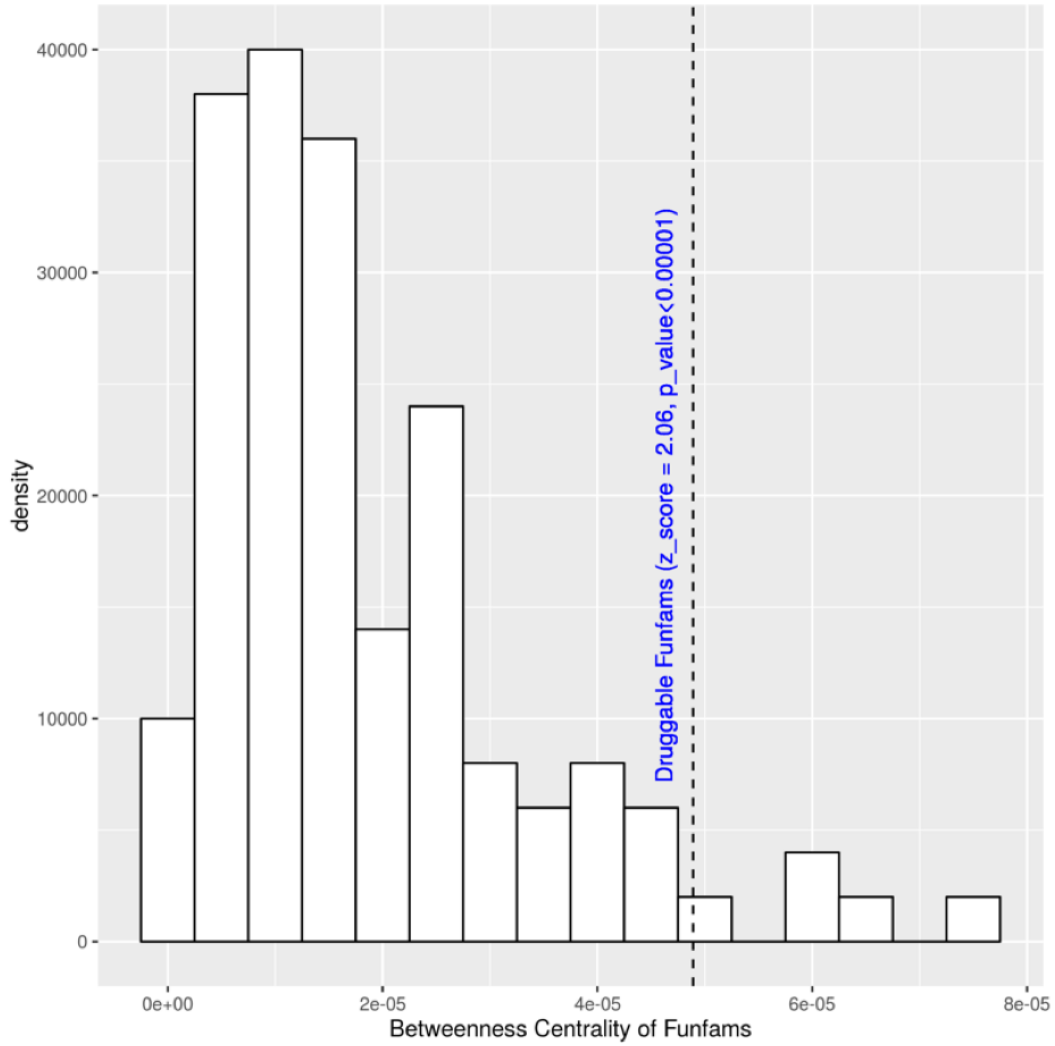

Figure 2.9: The mean betweenness centrality of druggable CATH-FunFams (dashed line) is compared with the distribution of the median betweenness centrality of random sets of non-druggable CATH-FunFams in a protein functional network.

The mean betweenness centrality of the druggable CATH-FunFams (dashed

line) was compared with the distribution of the median betweenness centralities of random sets of non-druggable CATH-FunFams in the protein functional network. As with the drug targets, the relatives of the druggable CATH-FunFams also exhibit a high betweenness centrality in the human protein network.

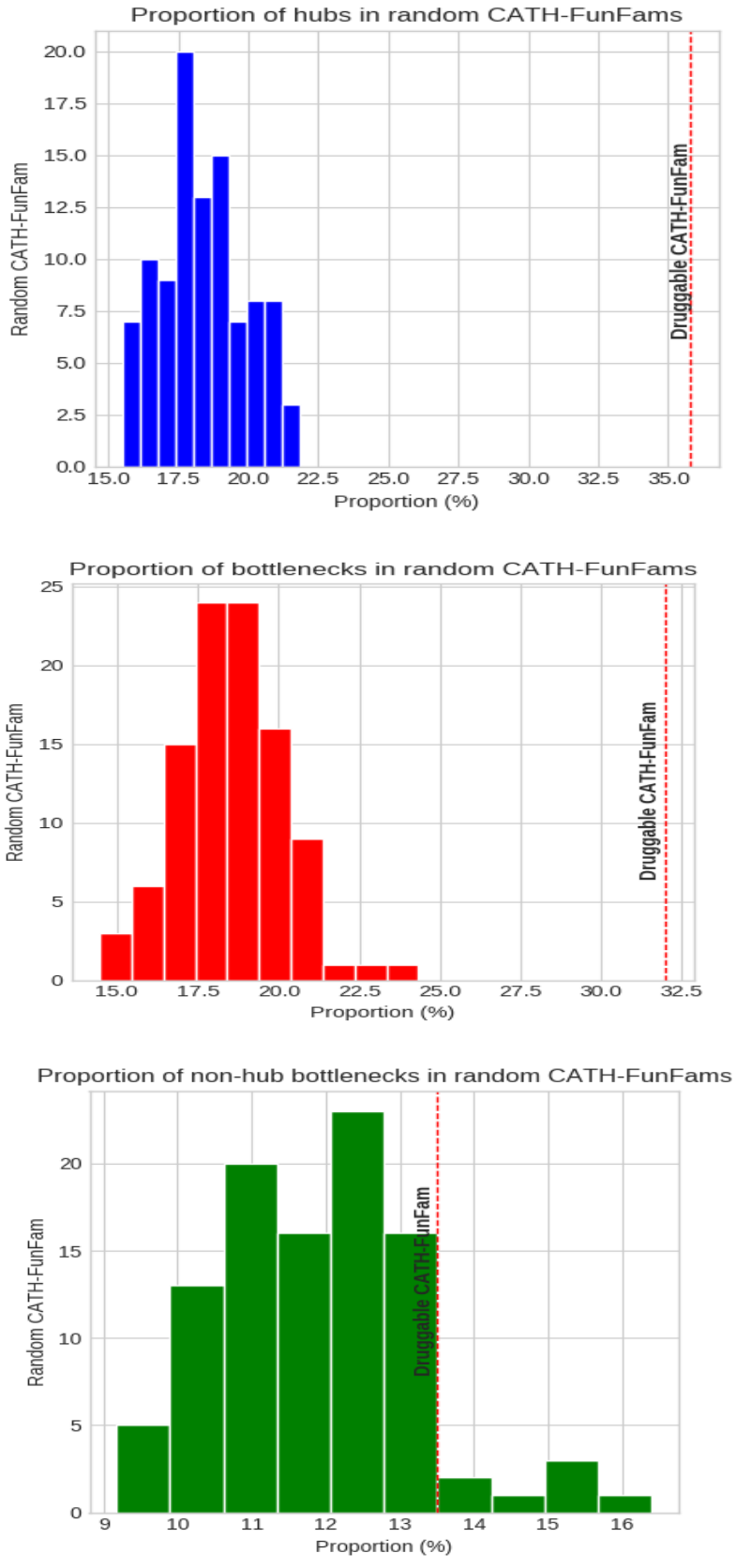

Figure 2.10: Proportion of relatives in druggable CATH-FunFams compared to the random CATH-FunFams for various network topological characteristics, the red dotted vertical line indicates the proportion associated with the druggable CATH-FunFams.

From the plot shown in figure 2.10, a higher proportion of human relatives in druggable CATH-FunFams are hubs and bottlenecks, than for the random CATH-FunFams. The proportion of relatives in druggable CATH-FunFams non-hub bottlenecks was also compared to random proteins. Such proteins are considered essential but are likely to have reduced side effects because they are less highly connected [133]. Figure 2.10c shows that the druggable CATH-FunFams were enriched with human relatives which are non-hub bottlenecks and hence may be a good source of drug-targets with less potential side effects.

2.3.9 Side effects associated with druggable FunFams

A logistic regression model was applied to determine the probability of a CATH-FunFam being free of side effects given the median similarity of its human proteins in protein network. The result showed that for a CATH-FunFam with its relative completely dispersed within the network, the probability it does not contain proteins with side effects is 31%. Relatives in druggable CATH-FunFams with median similarity $>$0.48 have a higher probability greater than 50% of their relatives not being associated with side effects and this was significant at p-val $<$0.05. Therefore relatives in druggable CATH-FunFams, having a median similarity score above this threshold (0.48), cluster together in a functional interaction network and are less likely to be associated with side effects. The list of druggable FunFams and their probability of being free of side effects has been provided in appendix A.

2.4 Chapter summary

This chapter has provided some fundamental support to the idea that domains mediate drug-target binding. 81 CATH-FunFams are druggable as determined by calculating overrepresentation of drug targets within the CATH-FunFam. The functional categories of CATH-FunFams agrees with those of drug targets as identified by other groups. Relatives of the druggable FunFams are central in a human functional network and highly connected in protein network forming drug neighbourhood. By building a regression model, druggable FunFams less likely to be associated with side effects were identified.

In summary, structurally coherent druggable CATH-FunFams, can be used as a proxy for inheriting drugs across relatives. This identification of CATH-FunFams as a reasonable annotation level for drug-target interactions opens a new research direction in target identification, with potential application in drug repurposing.

2.5 Limitations and Future work

This work has identified druggable domain families based on enrichment (over-representation) of drug targets within the CATH-FunFams. The identification of drug targets using computational approaches can aid and hasten drug development processes. Although, drug design by pharmaceutical companies considers the influence of many factors which amongst others include the profitability of the drug, the process of drug repurposing could help rechannel already approved drugs to other relatives of the same CATH-FunFam as it is assumed, they share the similar drug binding site. The identification of druggable CATH-FunFams should be extended beyond the enrichment approach that was carried out in this study. One of the approaches that might be considered is the use of protein-small molecule docking and molecular dynamics to characterise the binding of drugs to the relatives of the druggable CATH-FunFams.

A caveat to consider in this study regards the fact that drugs were associated with the entire protein i.e. all domains in the protein. Hence, domains that are not involved in protein-drug interactions are assumed to be, and this may not be entirely true. One possible approach to this, in the future, may be the separation of drugs binding to single domain proteins and exploring how their characteristics varied compared to those binding to multidomain binding proteins. Another caveat is that the assessment of drug side effects was carried out by measuring dispersion of proteins in a given protein functional network. The static nature and the incompleteness of the human protein-protein interaction network is a further limitation to be considered. However, ongoing experimental and computational work in this direction will aid the improvement of coverage and better deduction of inference using this approach.

Chapter 3

Exploiting Protein Family and Protein Network Data to Identify Novel Drug Targets for Bladder Cancer

3.1 Introduction

Bladder cancer is one of the most common forms of cancer in western countries with men having a higher ratio than women in the range of 3:1 [137]. The incidence of the disease increases with age with a higher proportion found in older individuals above 65 years [138]. The most common form of bladder cancer generally affects the epithelium (urothelium) covering the inner surface of the bladder hence called urothelial carcinoma, it is also referred to as transitional cell carcinoma. Other forms of bladder cancer include the squamous-carcinoma and adenocarcinoma, although they are rare compared to urothelial carcinoma. Environmental pollutants and tobacco smoking have been attributed as risk factors associated with the outcome of the disease [139, 138], however less than 10% of bladder cancers have been attributed to these factors. Evidence from genetic studies supports ge-

netic predisposition to bladder cancer as analysis of polymorphisms in detoxifying genes for carcinogens, such as N-acetyltransferase2 (NAT-2) and glutathione S-transferase (*GSTM1, GSTT1*), was found to increase the risk of developing bladder cancer [137, 140].

Bladder cancer, in common with other forms of cancers, is driven by multi-step accumulation of genetic and epigenetic alterations that lead to uncontrollable cell growth, dysregulation and reduced apoptotic processes as well as other cancer hallmarks [141]. Several studies have identified high rates of mutation in the tumor protein 53 (*P53*) and single nucleotide and structural variations in other genes, such as *FGFR3*, as being predictive of bladder carcinoma [141]. Other genes that are often considered with high mutation rates in bladder cancer include *CREBBP, MLL3, ATM, NF1, FBXW7*. Genes involved in bladder cancer sometimes show co-occurrence of mutations as found in; *P53 and RB1*, *STAG2 and FGFR3*, *MLL2 and NFE2L2*, *KDM6A and FGFR3*, and *ERBB3 and ERBB4*, while others show patterns of mutual exclusivity in bladder cancer as found for *P53 and RAS* and *RB1 and FGFR3* [142].

3.1.1 Bladder cancer stage and grade

Histopathological studies are often undertaken to characterise the stage and grade of a tumour. The initial assessment of the bladder tumor stage can be done by carrying out palpation, imaging and cytoscopic test. The pathology of the tissue can be examined using transurethral resection of the bladder tumor (TURBT), in which a resectoscope is used to remove the lesion, or to take a biopsy. TURBT is considered as a diagnostic and therapeutic procedure as it can be used to remove the lesion or take the biopsy in the bladder for further examination [138]. Depending on the depth and spread of the invasion of bladder tissues, bladder cancer can be categorised by the tumour node and metastasis (TNM) classification system as shown in figure 3.1 below.

Grading, on the other hand, refers to the extent to which the cells are differentiated. Although the stage of the bladder cancer is important for deciding the form of treatment to adopt, the grading also indicates the aggressiveness of cancer. There

are several systems that are used in grading bladder cancer; the 1973 WHO method as well as the recent 2004 WHO/ISUP grading system, is often used as also shown in figure 3.1.

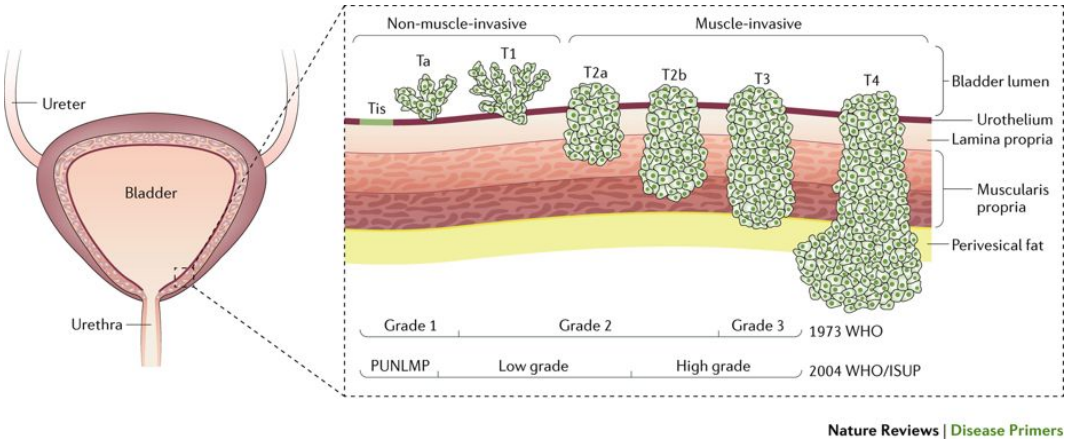

Figure 3.1: The types and stages of bladder cancer. Tis, Ta, and T1 types are confined to the mucosa. Stage 2 (T2a/b) has invaded the muscle layers either superficially or deeply. T3 has invaded into the perivesical layers while T4 has invaded surrounding glands such as the prostate, uterus, bowel. Figure is taken from: [138]

3.1.2 Molecular subtypes of bladder cancer

Broadly speaking, bladder cancer is classified into two types: muscle invasive bladder cancer (MIBC) and non-muscle invasive bladder cancer (NMIBC). About 75% of the newly diagnosed cases have NMIBC while 25% have MIBC or the metastatic stage [138]. The use of large-scale expression and sequencing data from the Cancer Genome Atlas (TCGA) has been fundamental in grouping bladder cancer into sub-types based on the shared RNA expression patterns or other alterations.

Non-muscle invasive bladder cancer (NMIBC)

Two common genetic changes in the NMIBC are the deletion of chromosome 9 and a point mutation in the fibroblast growth factor receptor 3 (*FGFR3*) [143]. Most cases of NMIBC are characterised by mutation of *FGFR3* which leads to activation of the RAS-Mitogen activated receptor protein kinase (*MAPK*) pathways [144]. Activation of *FGFR3* also occurs through chromosomal translocations to form fusion proteins with *TACC3* (transforming acid coiled-coil containing protein 3) and sometimes *BAIAP2L1* (brain specific angiogenesis inhibitor 1-associated protein 2 -like

1) which are potent activating oncogenes. Activation of the *RAS-MAPK* pathway contributes to more than 80% of the cases of NMIBC. The inactivation of tumour suppressor gene *TSC1* contributes to about 15% of the cases of NMIBC [145]. The tuberous sclerosis complex 1 (*TSC1*)-*TSC2* complex controls the mTOR branch of the *PI3K* pathway. Hence, loss of *TSC*-genes in NMIBC leads to the upregulation of *mTOR* which then becomes a major factor to consider in NMIBC.

Other tumour suppressor genes whose inactivation has been thought to lead to NMIBC include *STAG2*, as well as chromatin-modifying genes such as *KDM6A, CREBBP, EP300* and *ARID1A*. This is due to the significantly higher frequency of mutation of these genes in NMIBC compared with other cancer types [138].

Muscle invasive bladder cancer (MIBC)

Analysis of the genome of MIBC has shown some similarities with other types of cancers. One such is the loss of the key tumour suppressor genes leading to the escape from the cell cycle check points and dysregulation of several signalling pathways. *TP53* and *RB1* were found to be frequently mutated and the regulators of their pathways such as *MDM2* and *E3F3* are also altered. Mutation of other genes, encoding the components of the *PI3K* pathways, including *TSC1*, *AKT1*, *PIK3CA* are disrupted in MIBC. *FGFR3* activating point mutation is less frequent in MIBC than NMIBC. The switching to isoforms of *FGFR3* and *FGFR1* also gain prevalence in MIBC [146]. *RAS* mutation and inactivation of the NOTCH pathway genes also contributes to the *MAPK* pathway activation. Upstream activation of *HER2* is also found in some cases of MIBC. Some of the altered genes in the KEGG pathway for bladder cancer are shown in the figure 3.2 below.

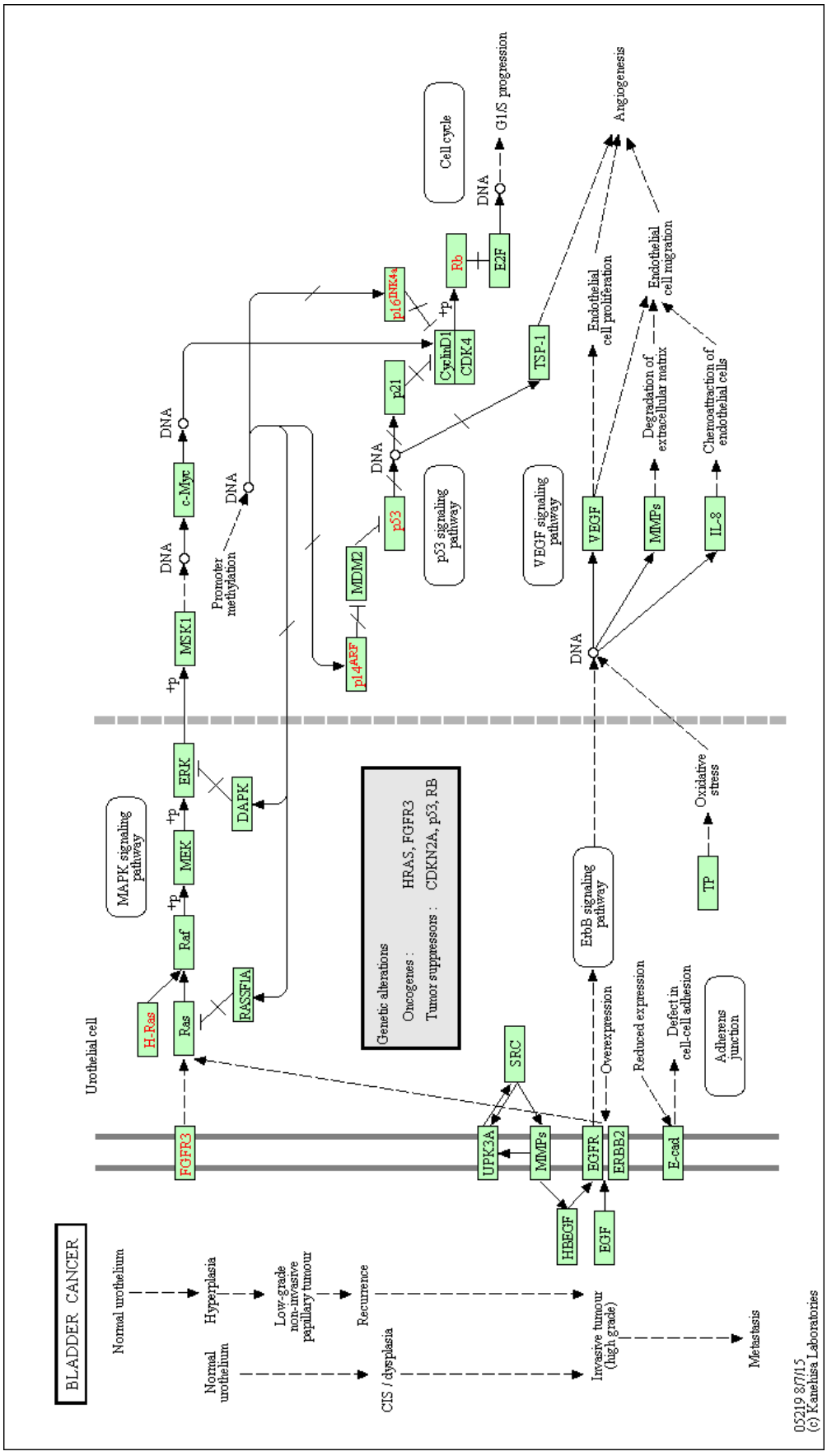

Figure 3.2: KEGG pathway of key genes involved in bladder cancer. Figure obtained from [147]. Highlighted in green are genes that are implicated in various stages of bladder cancer while those in red are the highly mutated genes in bladder cancer.

3.1.3 Current therapeutic approaches for bladder cancer

Traditional treatment of bladder cancer involves surgery, chemotherapy and radiotherapy. The surgical approach involves the removal of the tumour around the bladder a process referred to as TURBT, which involves the insertion of resectoscope through the bladder urethra. This is often followed by the administration of drugs and sometimes radiotherapy or both. Some of the drug administration techniques currently employed in the management of bladder are listed in table 3.1 below.

Table 3.1: Drugs administered for the treatment of bladder cancer. This table was adapted from (`https://www.cancer.net/cancer-types/bladder-cancer/types-treatment`.)

Treatment options	Description	Types	Adverse effect
Chemotherapy	Chemotherapy involves drug administration to help in destroying the cancerous cells	1. Intravesical chemotherapy involves local drug delivery to the organ of interest. Commonly used drugs include: mitomycin, thiotepa. Others are cisplatin, doxorubicin, gemcitabine, valrubicin. 2. Systemic chemotherapy, on the other hand, is also used for the treatment of bladder cancer and this involves passing the chemotherapeutic agent into the blood stream or through oral administering. Common examples include combination therapy of cisplatin and gemcitabine, carboplatin and gemcitabine, MVAC (Methothrexate, vinblastine, doxotubicine and cisplatin) as well as dose dense (DD)-MVAC.	Common side effects include nausea,and vomiting, hair loss, appetite loss and diarrhea.
Immunotherapy	This uses biologic agent to trigger the body's immune response.	1.One of the earliest standard immunotherapeutic drugs is Bacillus Calmette-Guerin (BCG). This is another example of intravesical therapy as the BCG is placed directly on the bladder. 2. Interferons are another immunotherapeutic approach and can sometimes be combined with BCG. 3. Immune checkpoint inhibitors, on the other hand, are monoclonal antibodies that are currently considered because bladder cancer has shown over expression of immune checkpoint protein PD-1. Currently FDA approved drugs include Atezolizumab, nivolumab, avelumab, durvalumab, pembrolizumab	Possible side effects include fatigue, nausea, loss of appetite, fever, urinary tract infections, rash, diarrhea, and constipation.

3.1.4 Targeted therapy for the treatment of bladder cancer

As highlighted in table 3.1 above, the currently approved therapeutic treatments of bladder cancer are delivered either through immunotherapy or chemotherapy. In addition, these options have only helped in increasing the median survival outcome

of patients with metastatic bladder cancer to only about 15 months [148]. However, the cost of bio-engineering antibodies and their bulkiness, which renders them less soluble limiting their excretion from the kidney of the patient, alongside the toxicity of chemotherapy necessitates research into small molecules that could be used for targeted therapy in the treatment of bladder cancer.

Targeted therapy was introduced in the early 90's, as a new approach for treating cancers. The aim of this therapy is to interfere with cellular processes that are driving oncogenic transformation of normal tissues into cancerous ones. This might involve inhibiting oncogenes or activating tumour suppressor genes. Several signalling pathways such as: *FGFR, PI3K/AKT/mTOR, EGFR2* have been suggested as targets to improve the outcome of patients with bladder cancer and some encouraging preliminary results have been reported in the review by Ismaili and colleagues [149] which describes some currently used monoclonal antibodies and small molecules for targets such as *VEGFR, EGFR, mTOR* and *HDAC*. Other therapeutic targets currently being studied include the cell cycle regulation genes, heat shock proteins as well as genes involved with the immune system.

Fibroblast growth factor as a target for bladder cancer

The high frequency of *FGFR* mutation in bladder carcinoma indicates its significance in driving bladder cancer and a potential interest as a therapeutic target. *FGFR3* has been described as the most commonly altered receptor with activating mutations and amplification through fusion with *TACC3* protein (*FGFR3-TACC3*). Antiproliferative activity of FGFR inhibitors has been shown in several pre-clinical studies [150, 151]. There are also several small molecule inhibitors of *FGFR* available in clinical trials for the treatment of bladder carcinoma [152]. Although most of these inhibitors are in the phase 1 and 2 clinical phase, the encouraging results might help in the progression into further phases. An example of such small molecule is BGJ398, a FGFR-1-3 antagonist that is currently in trials in patients who showed relapse after undergoing chemotherapy, with positive responses obtained in such patients [153]. Others include TKI258 and ENMD2076 which are in phase II clinical trials [154].

***PI3K/AKT/mTOR* signalling pathway as a target for bladder cancer**

PI3K/AKT/mTOR pathway is known as the cell survival pathway and is disrupted in the majority of cancers. This pathway is also involved in cell motility and metabolism. As a result of the influence in cancer, several drugs have been designed to target this pathway. One such is buparlisib which is being investigated as a second line treatment of patients with metastatic urothelial carcinoma [137]. There have been limited successes achieved by targeting the PI3K/AKT/mTOR pathways and this may be due to the cross talk or feedback activation of alternative signalling pathways such as the *MAPK* and *JAK-STAT* pathways. Several compounds are currently undergoing trials for the treatment of bladder carcinoma with the aim of achieving positive responses by focusing on patient stratification i.e. the selection of patients that show high expression profiles for genes implicated in bladder cancer such as HER2 relative to other genes.

Epidermal growth factor 2 (*EGFR2/HER2*) as target for bladder cancer

The overexpression and amplification of the *EGFR2/HER2* genes in bladder cancer have been reported [152]. Also, several *EGFR2/HER2* mutations have been suggested to be involved in differentiating the bladder cancer into subtypes. Hence, there are treatments currently available to target patients having urothelial carcinoma with over-expression and amplication of *HER2*. One such is the combination therapy with trastuzumab, paclitaxel, carboplatin and gemcitabine [155]. These patients were reported as having a remarkable response rate and an overall survival of about 14 months. Successful outcomes of targeting of this gene for bladder cancer will rely on clear identification of the subtypes associated with *EGFR2/HER2*.

Other targets available for treatment of bladder cancer include cell cycle regulation genes such as the Aurora kinases, polo-like kinase-1, cyclin-dependent kinase4 (*CDK4*); heat shock proteins as well as those involved with the immune system such as the checkpoint *PD-1* pathway, anti-cytotoxic T lymphocyte associated antigen , and *IL-2/T* lymphoctye receptor fusion protein targeting p53 epitope.

3.1.5 Techniques used in identifying disease proteins

Various studies have aimed to improve the diagnosis and treatment of bladder cancer by assessing the impact of mutations in the protein sequence and likely modification of functions for proteins implicated in bladder cancer [156, 152, 149]. However, there are challenges in analysing disease driving proteins as most current approaches have difficulties filtering out passenger mutations. WGCNA [157] and MutFams [158] are methods which use different approaches to help with the identification of driver mutations in disease conditions.

Network-based approaches for finding cancer driver genes

One of the approaches currently in use to detect causal genes for a given disease is the analysis of protein associations and complexes (modules) driving disease conditions. The construction of gene co-expression associations helps to determine connections in a human protein network. Weighted Gene Co-expression Network Analysis (WGCNA) method generates co-expression networks of genes and identifies modules of interest. WGCNA has been used in the study of many complex diseases such as breast cancer ([159]), schizophrenia ([160]) and osteosarcoma ([161]).

Many of the driver gene prediction methods using biological networks apply guilt by association in which phenotypically similar genes are expected to be co-located in a given network. Measurement of local properties in the network such as the shortest path length between known disease genes and neighbouring genes have been used to predict cancer-associated genes [162]. However, using local network properties can be challenging because of the incompleteness of the human interactome data. Other approaches measure global network topology and explore the overall network using algorithms such as random walk with restart (RWR), kernel diffusion, network propagation and also transformation into a probabilistic model, to predict putative genes for a given disease condition [163].

Methods applying these approaches include; MUFFINN, a pathway-centric method that identifies driver genes by analysing the mutation information of the genes and neighbours in a functional protein network [164]. NetSig, similarly uncovers driver genes by considering mutations in neighbours of a disease gene [165].

DIAMOnD is another disease driver detecting algorithm that assesses the neighbourhood of disease proteins and identifies disease modules that include the known disease protein. It ranks the predicted disease proteins based on the connectivity to the disease proteins [166]. (Review on DIAMOnD in section 1.4.2).

Mutationally enriched domain families (MutFams)

CATH-FunFams are evolutionarily coherent domain families in which relatives have very similar structures and functions. MutFams are mutationally enriched CATH-FunFams. MutFam genes have been shown to be significantly enriched with known cancer driver genes from the Cancer Genome Census (CGC) and were used in this study to identify putative bladder cancer genes [158]. Details about MutFams are provided in the method section.

3.1.6 Objectives of the chapter

The aim of this chapter is to study the molecular and pathway mechanisms that drive bladder cancer and to identify novel targets for therapeutic purposes using the in-house CATH druggable domain families. A gene co-expression network was first constructed using expression data from TCGA following an established WGCNA protocol [157]. Known bladder cancer genes from the Cancer Genome Census (CGC) were combined with putative driver genes from COSMIC and from mutationally enriched domain families (MutFams). This seed set of driver genes was then extended with highly expressed genes from the modules of the co-expression based network. Further expansion of the seed set was carried out by running a diffusion (DIAMOnD) algorithm on a comprehensive human protein network built by combining the Pathway Commons and the gene co-expression network.

Gene enrichment analyses were carried out on this expanded set of putative disease genes using Gene Ontology (GO-terms), cancer-hallmarks signature and KEGG pathway analyses. Drugs associated with these genes were identified by querying the ChEMBL database for approved drugs. Using a similar strategy to that described in the previous chapter on ”Domain based approaches to drug polypharmacology” which explored the association of drugs with functional domain families (called druggable FunFams), the putative bladder cancer proteins were mapped to

these druggable FunFams and the network characteristics of the druggable family relatives were assessed in a protein network. Finally, modules enriched in known cancer genes and targets from druggable CATH-FunFams with few side effects were identified to be of interest for follow up studies and analysis. In summary, this study was able to combine co-expression and protein interaction network analyses to identify putative bladder cancer genes and drug targets.

3.2 Materials and Methods

3.2.1 Study Design

The figure below summarises the workflow employed for the study.

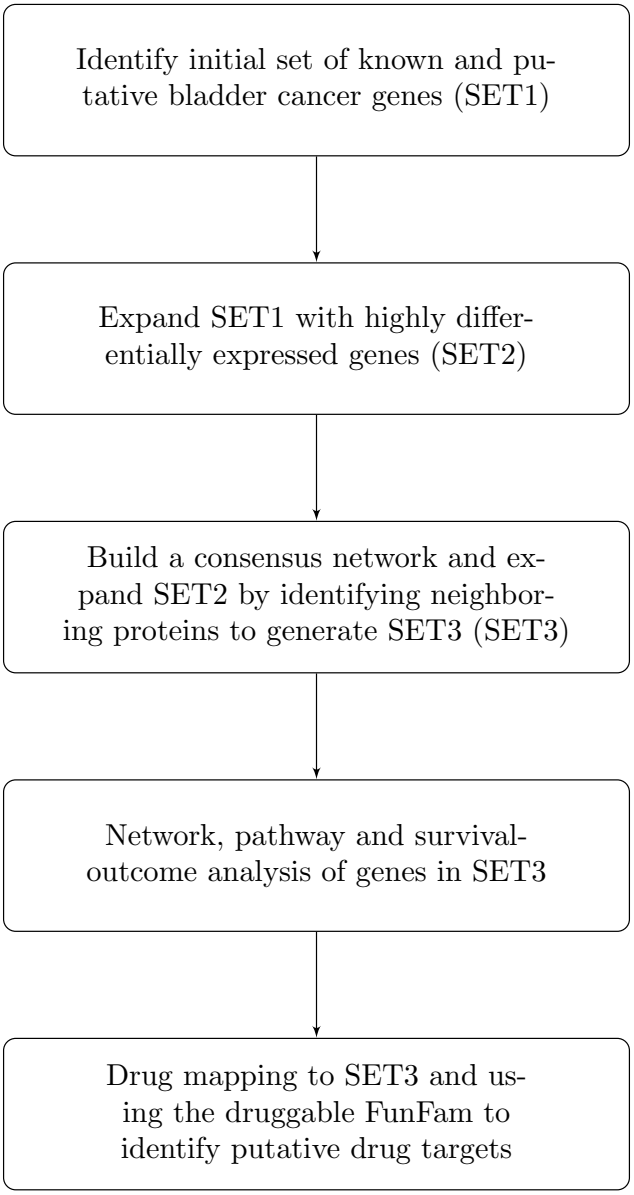

Figure 3.3: Protocols for identifying putative bladder cancer driver proteins

3.2.2 Identification of known and putative driver genes from public resources

A set of 105 known bladder cancer genes (**SET1**) was compiled from the Cancer Genome Census (CGC) [167], Catalogue of Somatic Mutation in Cancer (COSMIC) [168] and the in-house mutation enriched domain families (MutFams). Each data source is described below

Cancer Genome Census (CGC)

The CGC curates genes that are highly annotated with mis-sense mutations and for which there is evidence that the mutations are causally implicated in driving the oncogenesis. CGC genes specifically associated with bladder cancer were selected. CGC genes are classified into two tiers (tier 1 and 2) [168]. Those classified as tier 1 have comprehensive evidence of a mutation that changes the activity of the gene product in a way that promotes oncogenic transformation. Tier 2 have less extensive evidence but have a strong indication of a role in cancer. Both tier 1 and tier 2 genes were considered for this study.

Catalogue of somatic mutations (COSMIC)

COSMIC is one of the largest repositories of data for exploring the impact of somatic mutations in cancer. Missense mutations associated with bladder cancer were obtained from COSMIC-version 84. This was done by searching for keywords such as "UROTHELIAL" or "BLADDER". COSMIC provides numbers of observed mutations in each sample for each gene. For each gene, the mutation ratio (MR) was calculated as :

$$\frac{\text{Number of observed mutations}}{\text{Numbers of samples tested}} \times 100(\%) \tag{3.1}$$

The obtained data was filtered and only those genes with an MR above 3% were added to **SET1**.

Domain families enriched in cancer mutations (MutFams)

Predicted putative cancer driver genes were extracted from CATH domain functional families enriched in cancer mutations, termed MutFams [158]. Bladder

cancer MutFams were identified by analysing mutations (non-synonymous SNVs) found in whole-exome sequencing studies of bladder cancer tumours, obtained from the COSMIC database (v71) [168].

The MutFam algorithm tests for statistically significant enrichment of mutations found within CATH-FunFam domain boundaries compared to the gene as a whole.

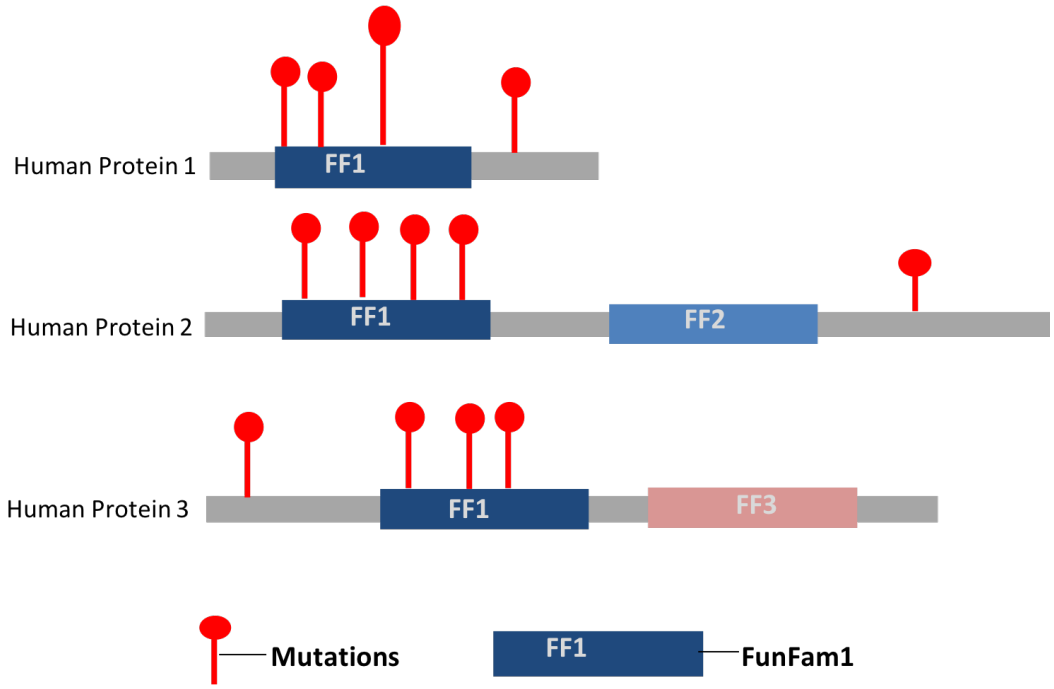

Figure 3.4: Mutfam approach to finding mutationally enriched domain families

Figure 3.4 shows the method for identifying a mutationally enriched family (MutFam). The three human proteins are paralogs (Human protein1, Human protein 2 and Human protein 3) comprising domains belonging to CATH-FunFam-1, 2 and 3 (FF1, 2, 3). Functional family 1 (FF1) is enriched in cancer associated mutations compared to the remaining domain regions of the proteins in which these FunFam domains occur.

The enrichment is based on the numbers of observed mutations(Mf), which includes all mutations in all domains within a CATH-FunFam divided by the expected number of mutations. The expected mutations count (Me) is calculated as the total number of mutations observed in the proteins containing the FunFam; taking into

account the fraction of amino acids within the CATH-FunFam compared to the total length of the protein.

$$enrichment\,factor(Ef) = \frac{\text{observed mutations (Mf)}}{\text{expected mutations (Me)}} \tag{3.2}$$

Me = fraction of amino-acids × number of domains in the FunFam

A permutation test is subsequently carried out to determine whether the observed number of mutations is greater than expected by chance i.e. the observed mutations were compared against randomly estimated mutation counts. The Benjamini-Hochberg correction for multiple testing was applied to the permutation derived p-value at FDR $\geq$5%. Only MutFam proteins expressed in bladder cancer were selected. Tissue expression data was obtained from Human Protein Atlas [67]. The top quartile (by mutation count) of mutated genes from each MutFam was selected, resulting in 40 genes that were then added to the known bladder cancer genes from TCGA and COSMIC giving SET1.

3.2.3 Bladder cancer RNA-seq data

RNA-sequencing data from the Cancer Genome Atlas (TCGA) repository was downloaded from the Genomic Data Commons (GDC) data portal [169]. The "getFirehorseData" method from the RTCGAToolbox R package [170] was used for obtaining different cohorts within the TCGA data. The most recent data for bladder cancer was obtained (dataset="BLCA" and run_date ="20160128", level = "3"). TCGAbiolinks [171], another R package was used in processing and visualisation of the gene expression data from TCGA as described below.

Identification of differentially expressed genes (DEGs)

A total of 427 samples was obtained from the TCGA database. 408 of these samples were from bladder cancer patients while 19 samples were from healthy patients. The gene expression data was analysed to identify genes differentially expressed in bladder cancer using the "TCGAanalyze_DEA" method from the TCGAbiolinks R package [171]. This applies the EdgeR quantile-adjusted conditional maximum likelihood (qCML) to detect differentially expressed genes. The p-values generated

were corrected using Benjamini-Hochberg (BH) multiple test at FDR of 5%.

The differentially expressed genes (**DEGs**) were then filtered by fold change (FC) to obtain those genes with $\log_2$FC above 4 and with a corrected p-value $\leq$0.01. These genes were called the highly differentially expressed genes (**Hi-DEGs**).

3.2.4 Building a bladder cancer gene co-expression network

In building the gene co-expression network, the genes are the nodes while the edges connect co-expressed genes. The RNA-seq expression data for bladder cancer obtained from TCGA was used in the construction of the co-expression network using the Weighted Gene Co-expression Network Analysis (WGCNA) algorithm [157]. The 19 normal samples (patients) and expression estimates with counts in less than 20% of cases were excluded as a quality control step to obtain only genes that strongly associated with bladder cancer.

Genes from the 408 samples were ranked based on the fold-change value and the top 5000 were chosen in the construction of the co-expression network. This filtering step has been routinely applied in studies using WGCNA, because of the high computational demand when considering all genes and has been shown to be optimal in previous studies [172, 173]. To build the co-expression network, an adjacency matrix was first constructed where the connection between the genes (x_i, x_j) was captured using the bi-weight mid-correlation values between their expression counts. The similarity between node i and j is thus:

$$S_{ij} = |bicor(x_i, x_j)| \tag{3.3}$$

The bi-weighted mid correlation measure was chosen as a measure of similarity over Pearson correlation as it is more robust to outliers. Furthermore, the similarity (S_{ij}) was transformed into a weighted adjacency matrix a_{ij} by raising S_{ij} value to a power ($\beta \geq$1).

$$a_{ij} = S_{ij}^{\beta} \tag{3.4}$$

The co-expression network constructed through this method gave rise to a scale-free network topology for a value of β=8 as shown below. Since many studies have

suggested that biological networks tend to be scale free [32, 31], this value of β was selected for building the co-expression network used in this study.

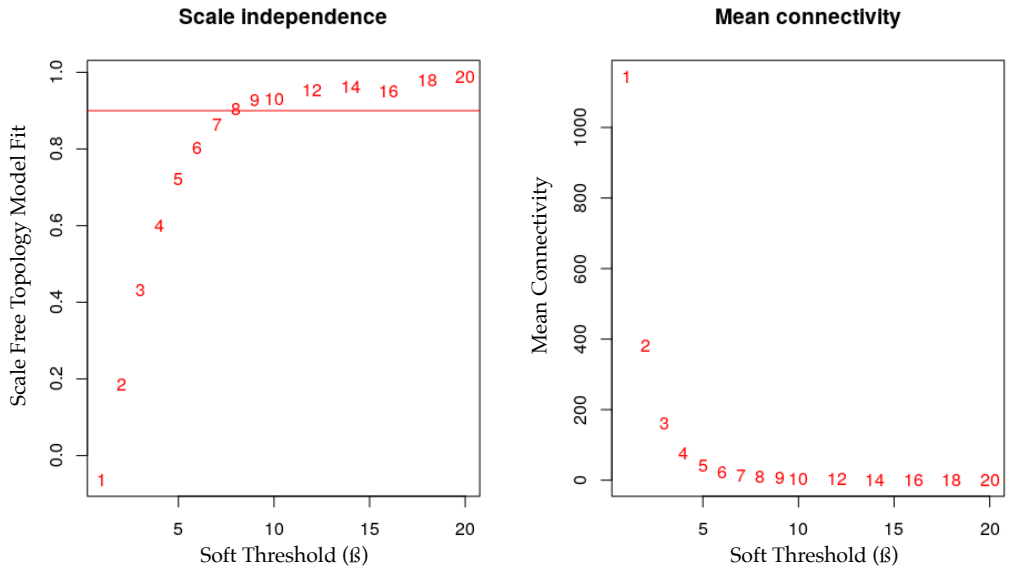

Figure 3.5: Analysis of the scale-free fit index and mean connectivity for various soft-thresholding powers (β shown in red numbers) following the WGCNA approach.

Module detection in the co-expression network and identification of modules enriched with putative bladder cancer proteins

WGCNA uses the "TOMsimilarity" library to calculate the topological overlap matrix (TOM) from the weighted gene co-expression network. TOM uses unsupervised hierarchical clustering to define modules based on the dissimilarities (DISS) measured between clusters i.e.

$$DISS = 1 - TOM. \tag{3.5}$$

TOM similarity is a co-expression measure that is not limited to gene pairs but considers gene relationships across the whole weighted gene network. Genes within the same module are densely inter-connected.

The WGCNA package provides a cutree-Dynamic function that allows presetting the "minimum module size" expected. The minimum module size expected was given a value of 30 as this threshold has previously been reported to define an

optimum number of genes belonging to a module [157, 174]. WGCNA has a graphical display in which each module is presented using a colour scheme. All genes that are not significantly co-expressed were grouped into an additional module, which was not considered for analysis. The script for generating this co-expressed network is publicly available in (`https://github.com/toluadeyelu/bladder_project`).

The over-representation of the initial driver set in each module was measured as a probability determined by comparing the fraction of the **SET1** genes belonging to the modules with those obtained by random sampling. The p-value was calculated to determine whether a given module is over-represented, by means of a binomial test, and corrected for multiple testing using Benjamini-Hochberg (BH). Only those modules with p-value $\leq$0.05 were considered enriched modules and the highly differentially expressed genes were extracted and added to SET1 to create **SET2**.

3.2.5 Construction of a consensus protein-protein interaction (PPI) network

To derive a more comprehensive protein network than the co-expression network, a human signalling network was generated using information from the Pathway Commons database [77] (downloaded 15-01-2018). Pathway Commons integrates protein interaction data from 9 different databases and robustly captures biological information at the molecular level. It comprises information on physical interactions, complexes, regulatory, phosphorylation as well as expression data. A robust human signalling network was built by considering all interactions.

A kernel-based approached was used to extend the pathway commons network by combining the gene co-expression network. This was done because using kernels allow the transformation of a functional association in a protein network into a functional similarity score which can be exploited and analysed.

Kernel methods for combining the PPI network and the co-expression network

Kernels use linear classifiers to solve a non-linear problem by mapping the original non-linear observation into a higher dimensional space called the *feature space* and obtaining the dot product [175].

$$\mathrm{K(x,y)} = \phi(x).\phi(y) \tag{3.6}$$

In this study, the Commute Time kernel was applied to generate a consensus network by integrating the kernel obtained from the PPI from pathway commons with that of the kernel for the gene co-expression network described in section 3.2.4. The Commute Time kernel has been shown to be a robust method for integrating biological data [176, 175].

The Commute Time kernel measures the topology of the network by quantifying the closeness between the nodes. In other words, it counts the number of steps it takes to randomly walk from node ”A” to node ”B” and back to node ”A” across a connected network. This, therefore, assumes that two nodes are similar if they have short paths connecting them. One of the assumptions made in using the Commute Time kernel is that the network is one connected component, thus, the largest connected component of each network is used. The Commute Time kernel is robust as it is parameter-free and hence no additional tuning is required.

Any mathematical operation performed on a kernel still gives a kernel. Hence, the PPI-kernel and the co-expression-kernel were added to generate a consensus kernel which was then transformed to give the consensus network used in this study. This gave a more robust network compared to the individual network from either gene coexpression or the Pathway Common network.

3.2.6 Extending SET2 by identifying neighbours in the consensus network using a diffusion method (DIAMOnD)

To study the pathology and pathways of diseases, the neighbourhood of the disease gene/proteins in a network are often also taken into consideration. This idea has

been exploited by several studies [166, 74]. For example, the DIAMOnD algorithm developed by the Barabasi group [166] has been used to find disease associated genes by considering the neighbourhood of the disease genes [74].

DIAMOnD identifies neighbours by considering the connectivity patterns around the disease genes based on the *connectivity significance*, a scoring scheme that also searches for distantly connected proteins. This is in contrast to the *connectivity density* which is based on the local topology of the network which other module algorithms use. Applying this method to the consensus network, returned additional putative bladder cancer associated proteins, giving the final set of putative bladder cancer proteins **SET 3**.

3.2.7 Pathway analysis of the bladder cancer associated proteins

Gene enrichment analysis was performed on SET3 proteins using ClusterProfiler [177]. ClusterProfiler incorporates several biological databases such as the Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG) [147] as well as the cancer hallmarks signatures from the Molecular Signature Database (MSigDB) [178]. In this analysis, a hypergeometric test was used to determine which terms and pathways were more significantly associated with the bladder cancer associated proteins (SET3) than expected by chance.

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{3.7}$$

N-represents the total number of genes in the background distribution, M-represents the number of genes within that distribution that are annotated with a given term or pathway, n-is the size of the gene list being considered (i.e. SET3) while k is the number of genes within the gene list that are annotated to the term. The terms allocated to the gene-list were adjusted for multiple hypothesis testing and the q-value was also calculated for the control of false discovery rate (FDR).

3.2.8 Network analysis of putative bladder cancer proteins

The network properties of the putative bladder cancer proteins was analysed using network topology measures such as degree and betweenness centrality. Hubs and bottlenecks were identified as the top 20% of the proteins in the consensus network ranked by their degree and betweenness centrality respectively. The proportion of hubs and bottlenecks in the putative bladder cancer proteins was compared to 10000 random sets of equal numbers of proteins.

3.2.9 Mapping drugs to putative bladder cancer associated proteins

Drugs targeting the putative bladder cancer associated proteins **(SET3)** were obtained from ChEMBL version-23 database [179], a database holding information on bioactive molecules and their activity data. The drug-targets selection criteria were such that weak interactions were excluded. Drug-target sets were filtered to obtain those interacting with an affinity above 1μM and a p-*ChEMBL* value $\geq$6. The p-*CheMBL* value is the -log(concentration/affinity). The value ranges from 1-9 and a value $\geq$6 is considered to be of high affinity. Only drugs which had direct binding to the target in question were selected and with a maximum phase of development (4) (where 4 signifies FDA-approved from a range of 1-4).

3.2.10 Mapping putative bladder cancer associated proteins to druggable CATH-FunFams

CATH-Functional Families (FunFams) from version 4.2. of the CATH database [81] was used. Previously, 81 CATH-FunFams highly enriched in known drug targets (named druggable CATH-FunFams) were identified and shown to have value for drug repurposing [180], as the high structural similarity of relatives supports inheritance of drug binding affinities. Druggable FunFams have at least one relative with drug information from drug databases such as ChEMBL. The bladder cancer associated proteins were mapped to druggable CATH-FunFams to assign clinically approved drugs to these putative cancer targets, through inheritance of drugs be-

tween relatives within the FunFam.

To determine the likelihood of side effects arising from targeting a particular CATH-FunFam, an established in-house method that performs regression analyses was used to assess the association of known side effects for drugs bound to relatives in a given CATH-FunFam with the dispersion of the relatives in a human protein network (See Chapter two). This allowed the annotation of each druggable FunFam with information on the likelihood of side effects. Further filtering of drug targets was done by identifying those putative cancer targets with high bladder tissue expression using expression profile from Human Protein Atlas map [67].

3.2.11 Partitioning of consensus network into modules using MCODE clustering algorithm

The MCODE clustering algorithm [48] was used for the detection of clusters in the consensus network. MCODE uses graph density to find protein complexes and subnetworks of highly interacting proteins. The MCODE clustering algorithm operates in three stages: vertex weighting, complex prediction and optimal post processing. The algorithm first assigns a weight to each node based on the local connectivity around the node, then starting from the top-weighted node, it recursively moves outwards to generate cluster vertices within a given threshold. In this study, the default parameters for MCODE were used to identify protein modules in the consensus network.

3.2.12 Survival outcome measurement

The survival outcome of TCGA bladder cancer dataset was analysed using the TCGAanalyze_SurvivalKM function of the RTCGA-Toolbox R-program. Patients were divided into high and low groups based on their upper and lower quartile gene expression profiles and the prognostic correlation was measured using TCGA data survival data. The survival outcome was compared between patients based on low and high expression values for a given gene. Genes whose survival outcome is statistically significant ($p<0.05$) were identified.

3.3 Results and discussion

3.3.1 Generating a set of known and putative bladder cancer proteins from public and in-house resources (SET-1)

A set of known and putative bladder cancer proteins was obtained from publicly available resources (Cancer Genome Census (CGC) with 12 tier1 genes and 2 tier2 genes) and Catalogue of Somatic Mutation (COSMIC)) and from the in-house mutationally enriched domain families (MutFams). In the latter case, proteins were added to the set provided at least one domain mapped to a CATH MutFam. In this study, there was no segregation between muscle-invasive bladder cancer and non-invasive cancer so that all the mutation and expression data associated with bladder cancer were collated and analysed.

Table 3.2: Numbers of putative bladder cancer proteins from public and in-house resources

Dataset	Numbers of proteins
CGC	14
COSMIC	63
MutFams	40
Total **(SET1)**	105

Table 3.2 shows the numbers of proteins retrieved from public resources and from the MutFams. A set of 105 known and putative bladder cancer proteins were obtained, as some proteins were found in more than one resource.

3.3.2 Expanding the known and putative cancer set with genes differentially expressed in bladder cancer

Identifying highly differentially expressed genes (DEGs) in bladder cancer to extend the SET 1

A total of 191 highly differentially expressed genes (DEGs) were identified in the bladder cancer gene expression data from TCGA at a fold-change ≥ 4 and p-value ≤ 0.05. Figure 3.6 below shows the plot of the highly differentially expressed genes from the TCGA RNA-seq bladder cancer data. 72% of the DEGs were up-regulated

genes while 28% were down-regulated.

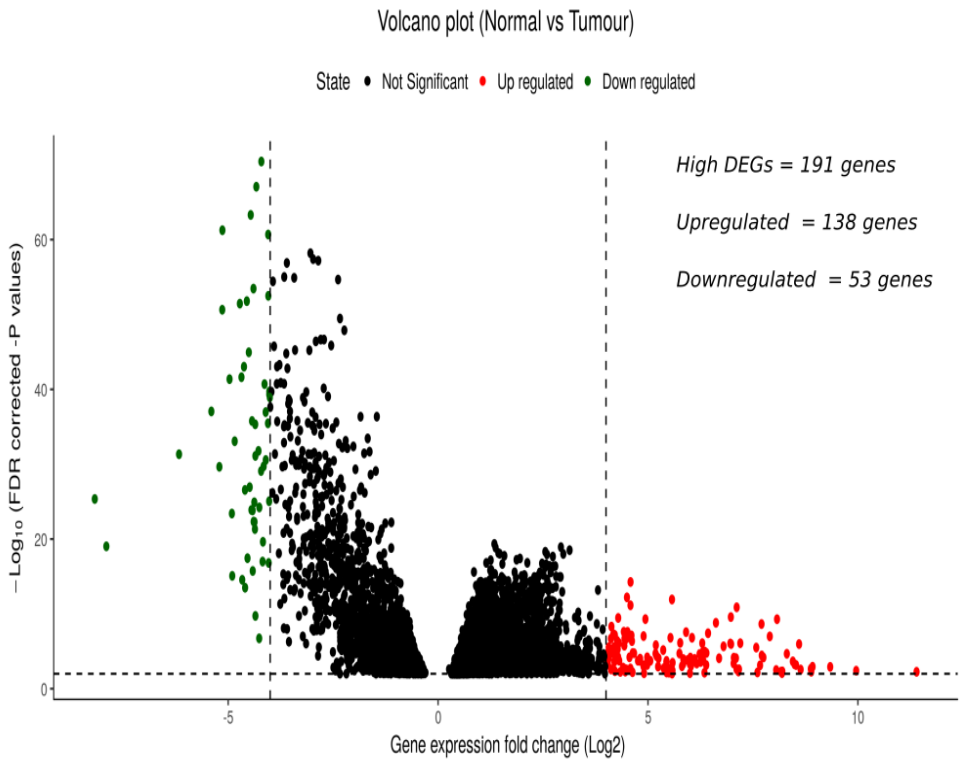

Figure 3.6: Volcano plot of the differentially expressed genes between bladder cancer and normal. The down-regulated genes are shown in green whilst the up-regulated genes are shown in red

3.3.3 Identifying modules in the gene co-expression network enriched with SET 1

Module detection in the co-expression network

Nine modules identified using the WGCNA module approach are summarised in Table 3.3 below.

Table 3.3: Modules detected using hierarchical clustering of the gene co-expression network. In bold are the enriched modules that were subsequently selected.

Modules	#Genes in Modules	#SET1 Genes	#DEGs in Modules	Summarised GO-term	P-value for enrichment of GO-terms
Mod1	67	13	2	Epithelial cell morphogenesis	2..23 X 10^{-5}
Mod2	206	8	5	Extracellular matrix organisation	0.125
Mod3	149	15	1	Protein modification	0.105
Mod4	127	32	30	Muscle cell differentiation	$<$0.0001
Mod5	39	1	0	Complement receptor mediated pathway	0.742
Mod6	115	2	3	Chromatin remodelling	0.206
Mod7	63	0	3	Renal system process	0.234
Mod8	222	16	0	Histone modification	0.084
Mod9	138	23	5	Transcription regulation	0.014

Each module was annotated with summary GO-biological process terms obtained using REVIGO [181]. Modules were found to be associated with specific GO-biological processes such as chromatin remodelling and modification, muscle cell differentiation, renal system process, histone modification which have been reported in other studies of bladder cancer [182, 156]. Other generic terms include protein modification in module 3 and complement receptor mediated pathway in module 5.

Identification of gene co-expression modules enriched with initial seed set genes (SET1)

Three of the nine modules (**Mod1, Mod4 and Mod9**) were found to be significantly enriched with putative bladder cancer genes from SET1, associated with epithelial morphogenesis, muscle cell differentiation as well as transcription regulation processes, respectively.

All highly differentially expressed genes (Hi-DEGs) within these modules were added to the set of known and putative genes in SET1 expanding the set of 105 genes in SET1 to 123 genes in SET2. This meant that SET2 contained both mutated genes from CGC, COSMIC and MutFams as well as highly differentially expressed genes (Hi-DEGs) in bladder cancer obtained from TCGA. Only the Hi-DEGs were selected in order to minimise noise as these are more confidently associated with bladder cancer.

3.3.4 Expanding the set of putative bladder cancer proteins by searching for neighbours of SET2 in a comprehensive consensus protein network

Integrating the protein interaction network and the gene co-expression network.

In order to extend the set of putative bladder cancer driver genes, the neigbours of the genes in a comprehensive human protein network were searched for. The coexpression network analysed in the previous section is a subset of the complete human protein network comprising 4,669 proteins with 970,390 interactions. This

was therefore combined with a comprehensive human protein interaction network from Pathway Commons, comprising 16,850 human proteins and 325,616 interactions.

The two networks were first converted into individual kernels using the Commute Time kernel method (See methods section 3.2.4) and then combined to obtain a consensus network. This was then filtered to reduce noise, by removing the low scoring edges, giving a network with 17,853 proteins and 727,786 interactions.

Identifying putative bladder cancer proteins in the region of SET2 proteins in the consensus protein network

The DIAMOnD algorithm [166] explores the disease neighbourhood of the SET2 genes, finding genes closer to the disease genes by virtue of their connectivity thereby retrieving other proteins likely to play major roles in the disease. DIAMOnD was applied to the consensus network and the top-ranked 200 proteins based on their significantly high connectivity to the disease proteins, added to the set of putative bladder cancer proteins, giving 323 proteins in SET3 (see Table 3.4). Network and pathway enrichment studies were performed to evaluate the validity of this set of putative bladder cancer proteins.

Table 3.4: Summary table of the number of putative bladder cancer associated proteins at each step of this study

Seed sets	Description	Numbers of genes
SET 1	Mutation data from CGC, COSMIC and MutFams	105
SET 2	SET 1 + Highly differentially expressed genes	123
SET 3	SET 2 + Neighbouring genes based on the DIAMOnD diffusion method	323

3.3.5 Network and pathway analysis of the putative bladder cancer-associated proteins

Gene-Ontology, pathway and cancer hallmarks analysis of the putative bladder cancer proteins

Pathway and biological process analysis was performed on the 323 putative bladder cancer proteins by using ClusterProfiler [177] (see Methods) to explore enrichment of genes in GO terms, KEGG pathways and cancer signature hallmarks.

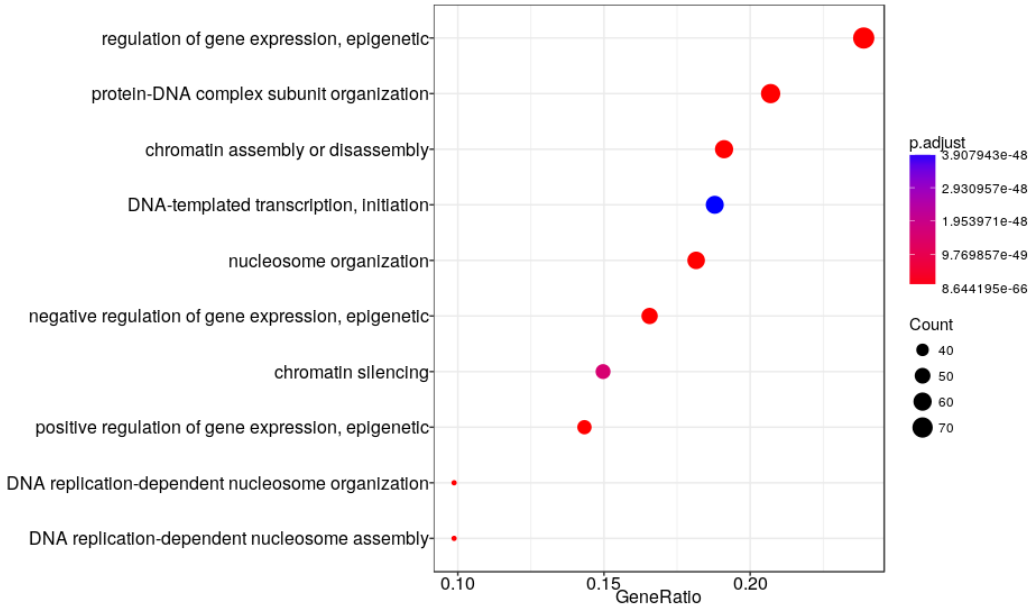

Figure 3.7: Enriched GO-biological processes identified for the SET3 putative bladder cancer associated proteins.

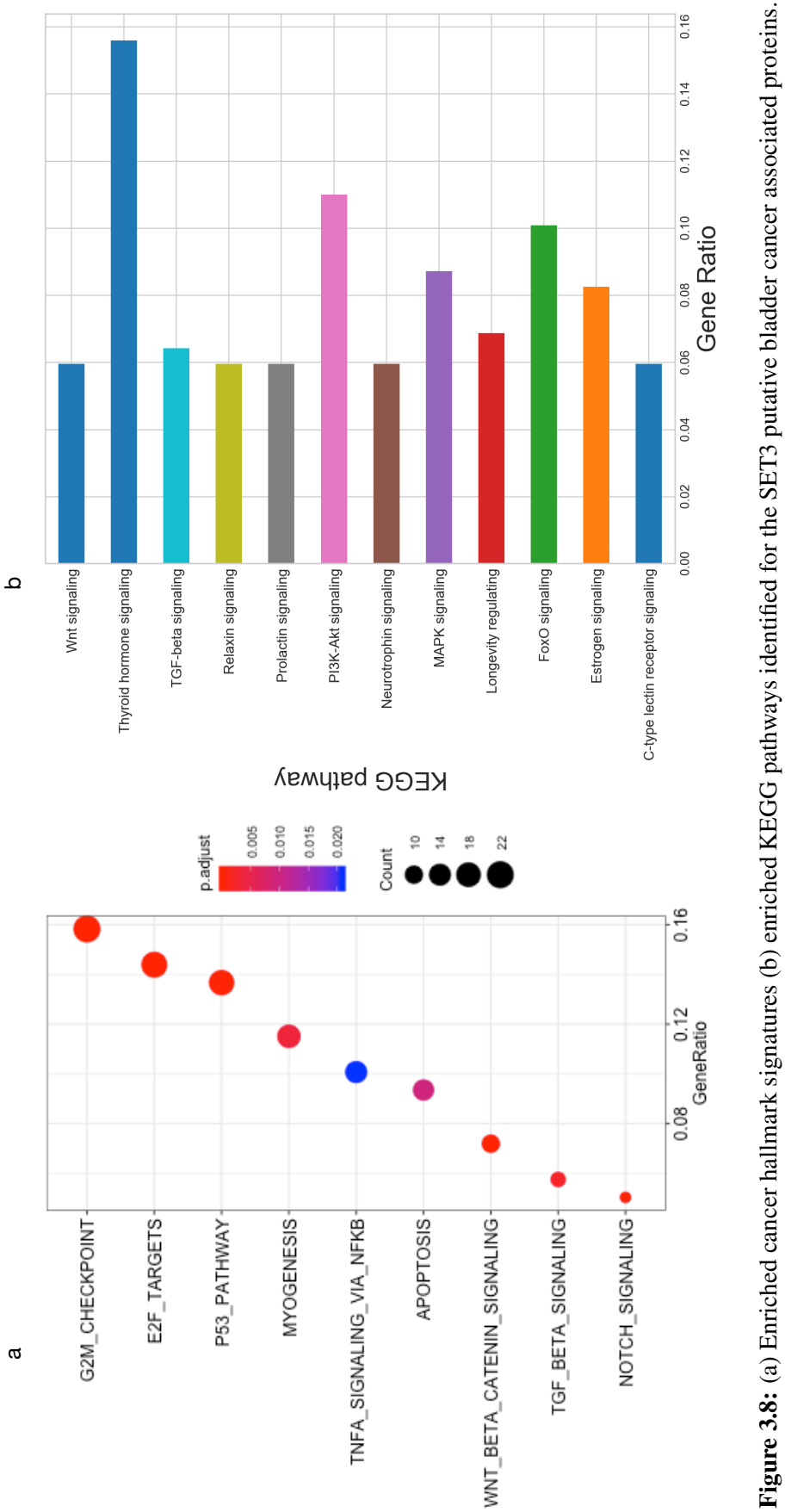

Figure 3.8: (a) Enriched cancer hallmark signatures (b) enriched KEGG pathways identified for the SET3 putative bladder cancer associated proteins.

From the analysis of the SET3 genes, three processes driving bladder cancer were identified with evidence from three independent enrichment analyses (see Table 3.5). Cell cycle is a hallmark for several forms of cancers and chromatin remodelling is known to play a fundamental role in the regulation of transcriptional processes. Lack of DNA repair mechanisms, which could occur as a result of the remodelled chromatin, increases the chance of genomic instability, mutation, cell senescence and cell death [183].

The observed hallmark signatures G2M checkpoints and E2F targets, are also associated with deregulation of the cell cycle, observed generally in cancers. The checkpoints are critical to ensuring maintenance of the genomic stability and deregulation is associated with tumorigenesis [184]. Beyond deregulation of the checkpoints, the MAPK signalling pathway has also been known to contribute to cell proliferation, differentiation and development [185], and activation of oncogenic transformation of bladder tissues.

Table 3.5: Processes associated with oncogenic transformation of bladder cancer, identified by enrichment studies.

Summarised terms	**GO-annotations**	**Hallmark Signatures**	**KEGG pathway**	**Common genes**
Cell cycle/ Mitotic division	ATP-dependent chromatin remodelling, Nucleosome organisation	G2M checkpoints, E2F targets	MAPK signalling process	TP53, RB1, CDKN2A, HRAS, MYC, ERBB3, JUN, HDAC5,HDAC2, FOS
Activating invasion and metastasis	Intracellular receptor signaling pathway, Hormone-mediated signalling	Myogenesis, WNT-catenin signalling, P53 pathways, Notch signalling	WNT-signalling, PI3K-Akt signalling	PPARD, RXRA, CTNNB1
Steroid hormone related processes	Steroid hormone mediated signaling pathway		Sphingolipid signaling, Estrogen signaling, Thyroid hormone signaling	THRB, ESR1, CTNNB1, NCOA3, PGR, NCOA1, RXRG, HDAC1

Furthermore, the putative bladder cancer genes showed enrichment in processes associated with the activation of invasion and metastasis. These are important for the transformation of non-muscle invasion bladder cancer to muscle invasive bladder cancer. P53 is a known tumour suppressor gene, its mutation is associated with cell migration and invasion. Wnt signalling has also been shown to support tumour metastasis in a highly tissue-specific way [186]. The observation of myogenesis, a process also described as the invasion of the muscle occurs in the advanced stage of the cancer and further strengthens the identification of metastasis and invasion as one of the hallmarks of bladder cancer.

Analysis of SET3 also showed enrichment of hormone related processes. There has been a lot of debate about the influence of hormones in bladder carcinogenesis. For example, Tryfonidis *et al* (2015), showed that excessive levels of androgen, observed as a result of aromatase inhibitor given to a post-menopausal patients lacking counterbalancing hormones (oestrogen and progesterone), might have been involved in driving the development of bladder cancer [187]. This analysis identifies many hormone-related genes, which may suggest that bladder cancer is a hormone-dependent malignancy.

Network characteristics of the SET3 proteins

Further analysis of SET3 proteins in the consensus protein interaction network indicated that they are important proteins from a network perspective as they tend to be mostly hubs. 61.60% of the SET3 proteins are hubs and have high connectivity with other proteins in the network. This proportion is statistically significant compared to random proteins in the network (p-value=2.984 x 10^{-47}). The betweenness centrality was also found to be statistically significant (p-value=8.347 x10^{-20}). Hubs and proteins with high betweenness centrality (BC) represent groups of proteins that are highly essential in a protein interaction network. Drug targets and disease proteins have previously been shown to have hub-ness and betweenness in protein interaction networks [133, 188].

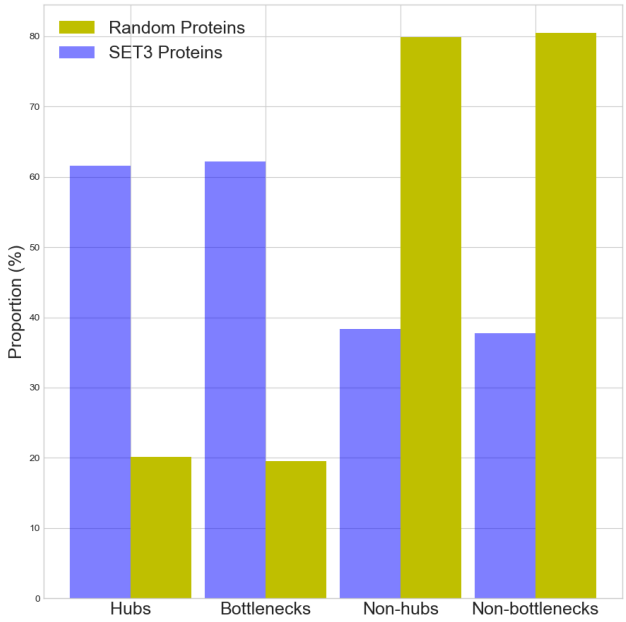

Figure 3.9: Network topological characteristics of the putative bladder cancer associated proteins in the consensus protein network. The SET3 proteins show high centrality measures compared to random.

SET3 proteins linked with survival outcome

Analysis of the information on survival prognostic outcome for genes in the TCGA dataset revealed that 35 SET3 proteins are statistically significantly associated with prognostic predictions in patients with bladder cancer. 11 of these genes including CASQ2, FHL1, ACTC1, SLC2A4, FLNC as well as P2RX1 were highly differentially expressed in bladder cancer. 13 genes associated with prognostic prediction were found to be hub in protein interaction network and shown in the table 3.6 below.

Table 3.6: Survival genes and their expression count

Genes	**logFC**	**Genes**	**logFC**
FBXW7	-1.89	PDZRN4	-4.49
WDR77	0.45	DES	-4.04
HIST1H4C	-1.06	SMAD3	-1.44
CASQ2	-4.45	PPARG	-1.63
FHL1	-4.02	HIST1H4H	2.11
KAT2B	-1.64	ELF3	-1.36
HIST2H2AC	1.37		

3.3.6 Identifying drug targets in the set of putative bladder cancer proteins

Identifying drug targets by direct mapping of drugs to the putative bladder cancer proteins

FDA-approved drugs from the ChEMBL database were mapped to the putative bladder cancer proteins (SET3). These drugs have already been approved for the treatment of other diseases and therefore can be considered for repurposing in the treatment of bladder cancer.

28 proteins from the 323 SET3 bladder cancer proteins have FDA approved drugs in ChEMBL. The targets had high affinity for the drugs (pChEMBL value $\geq$ 6). 95 drugs were associated with these 28 targets, with some proteins binding more than one drug (see figure 3.10). 35 of these drugs, associated with 21 SET3 proteins, are antineoplastic drugs inhibiting cell growth, as depicted by their anatomic therapeutic code (ATC). These approved drugs have not yet been considered for the treatment of bladder cancer and are small molecules designed to inhibit cell growth and block cell proliferation. They are currently used in treatment of other cancers including breast cancer, prostate cancer, hepatocarcinoma and it is therefore reasonable to assume that they could be refocused for bladder cancer treatment.

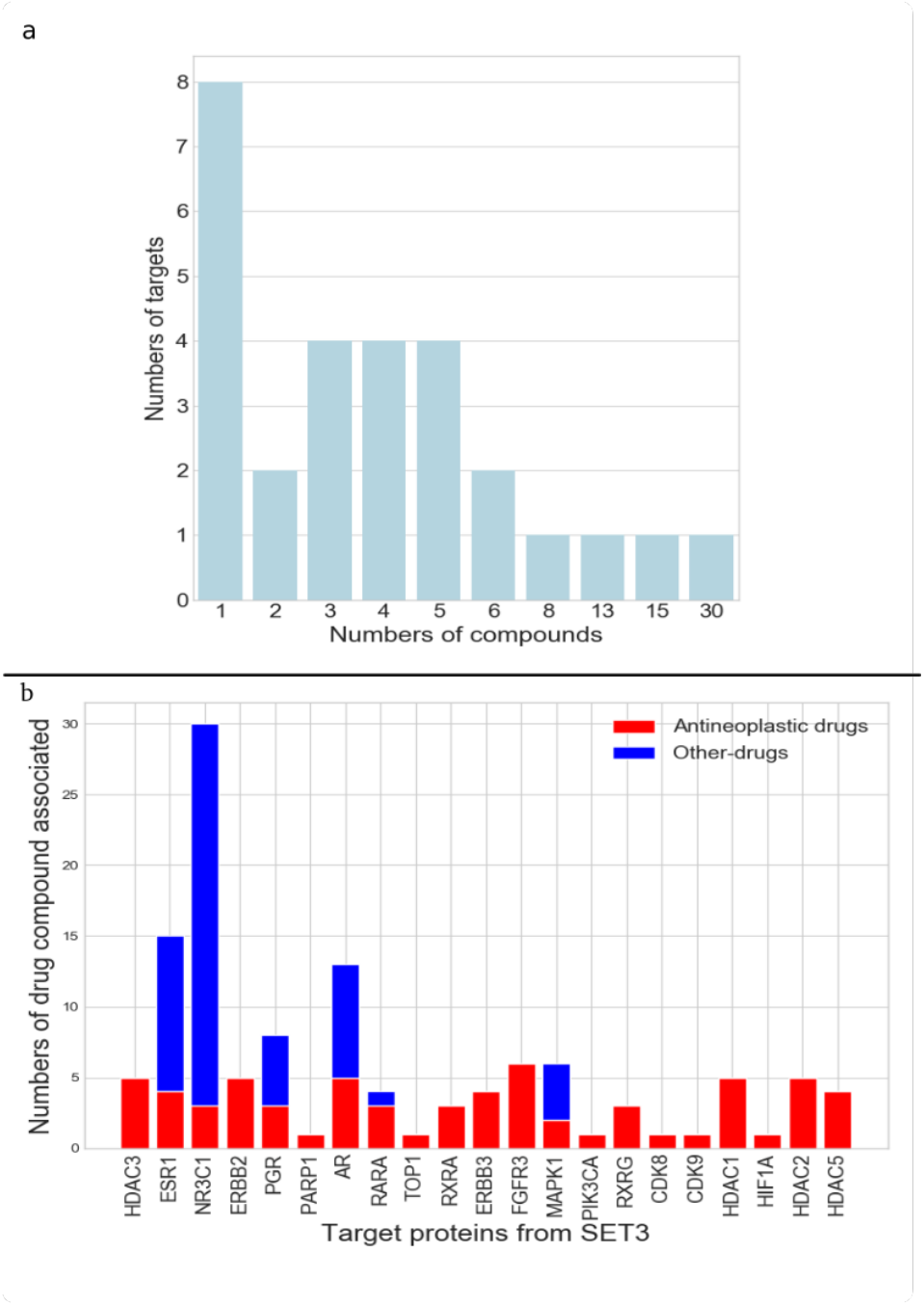

Figure 3.10: Number of compounds associated with putative drug targets in SET3.

Identifying further drug targets by mapping the putative bladder cancer proteins to druggable CATH-FunFams

To identify further drug targets in the putative bladder cancer protein set, we determined whether drugs could be inherited from other members of the CATH-FunFam to which the putative cancer target belonged (see Methods). 35 of the 323 putative

bladder cancer proteins could be mapped to 24 of the druggable CATH-FunFams in which other relatives bind clinically approved drugs. 28 of these proteins had already been associated with a drug (see section above).

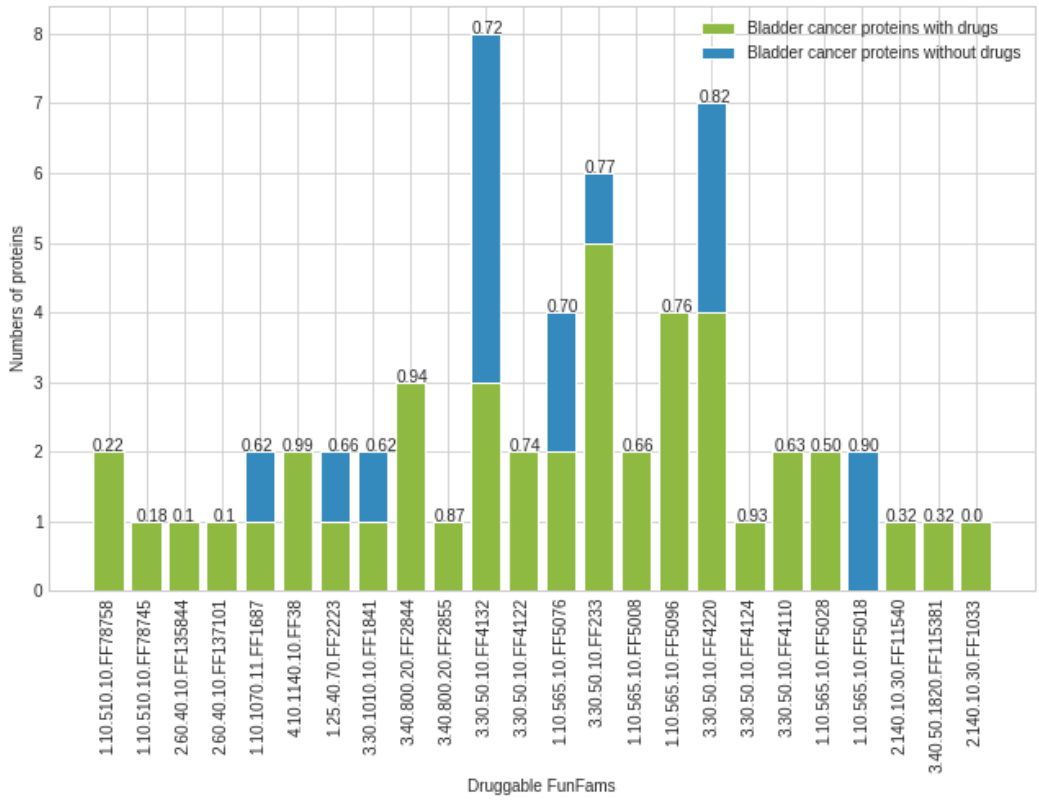

Figure 3.11: Number of relatives in druggable CATH-FunFams (having SET3 putative bladder cancer proteins), that are currently targeted by drugs (in green) and the number of untargeted SET3 proteins (in blue). Each druggable CATH-FunFam has been annotated by the median network similarity measure of the CATH-FunFam (range 0-1) where high values indicate significant likelihood of being free of side effects.

67% (17) of the druggable CATH-FunFams comprising 23 putative bladder cancer drug targets have median similarity above 0.48 and hence predicted to be less likely associated with side effects based. One family, identified as being particularly free of side effects is 3.30.50.10.FF4220, a nuclear receptor family. This family contains proteins such as RARB, RXRB and RXRB which show high expression in bladder cancer and are currently drug targets in other diseases (breast and prostate cancer). These proteins are currently targeted by tamoxifen for breast cancer, which could be potentially harnessed for the treatment of bladder cancer.

Druggable CATH-FunFam relatives are structurally similar

The 323 putative bladder cancer proteins were mapped to 24 of the druggable CATH-FunFams. The relatives within the CATH-FunFams were structurally compared against each other using the SSAP algorithm. The distribution of the RMSD scores confirms that they are structurally similar with an average RMSD score $<$ 2, for the majority of the CATH-FunFams and therefore it can be hypothesized that relatives within each CATH-FunFam possess the same drug binding residues.

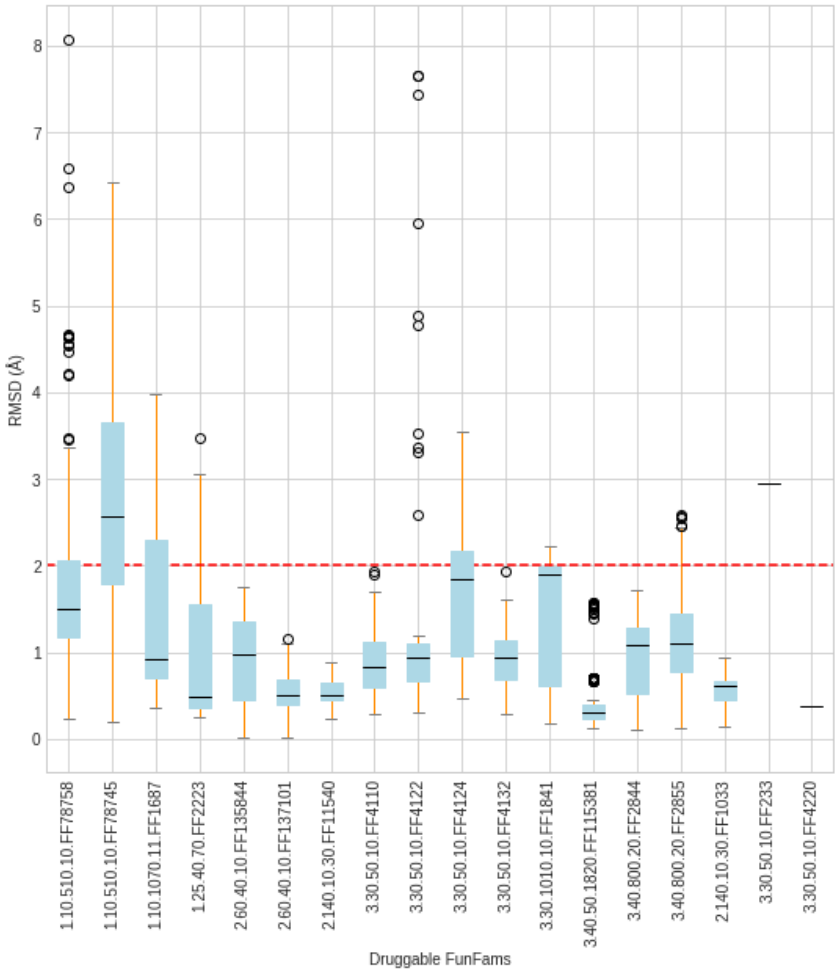

Figure 3.12: SSAP score distribution across relatives of the druggable CATH-FunFams

Table 3.7: Druggable CATH-FunFams associated with the putative bladder cancer genes

Druggable CATH-FunFams	Putative bladder cancer drivers with drugs	Putative bladder cancer drivers without drug	Probability of SE free
1.10.510.10.FF78758	*ERBB2, ERBB3*	–	0.22
1.10.510.10.FF78745	*FGFR3*	–	0.18
2.60.40.10.FF135844	*FGFR3*	–	0.10
2.60.40.10.FF137101	*FGFR3*	–	0.10
1.10.1070.11.FF1687	*PIK3CA*	*PIK3C2A*	0.62
4.10.1140.10.FF38	*ERBB2, ERBB3*	–	0.99
1.25.40.70.FF2223	*PIK3CA*	*PIK3C2A*	0.66
3.30.1010.10.FF1841	*PIK3CA*	*PIK3C2A*	0.62
3.40.800.20.FF2844	*HDAC1, HDAC2, HDAC3*	–	0.94
3.40.800.20.FF2855	*HDAC5*	–	0.87
3.30.50.10.FF4132	*NR3C1, PGR, VDR*	*NR1H2, NR4A1, THRA, NR5A1, NR1H3*	0.72
3.30.50.10.FF4122	–	*PPARA, PPARG*	0.74
1.10.565.10.FF5076	*RARA, VDR*	*NR1H2, NR1H3*	0.70
3.30.50.10.FF233	*PPARA, PPARG, RARA, RXRA, RXRG*	*THRB*	0.77
1.10.565.10.FF5096	*AR, ESR1, PGR, NR3C1*	–	0.62
3.30.50.10.FF4220	*ESR1, RARA, RXRA, RXRG*	*NR4A1, NR5A1, NR1H3*	0.82
3.30.50.10.FF4124	*ESR1*	–	0.93
3.30.50.10.FF4110	*NR3C1, PGR*	–	0.63
1.10.565.10.FF5028	*RXRA, RXRG*	–	0.50
2.140.10.30.FF11540	*DPP9*	–	0.32
3.40.50.1820.FF115381	*DPP9*	–	0.32
2.140.10.30.FF1033	*DPP9*	–	0.00
1.10.565.10.FF5018	–	*THRA, THRB*	0.90
1.10.565.10.FF5008	*PPARA, PPARG*	–	0.66

3.3.7 Identifying modules enriched in putative cancer drivers and druggable targets in the consensus network

The putative bladder cancer proteins were used as seed set for the MCODE clustering algorithm to find modules within the human consensus protein network enriched in druggable targets from SET3. 16 modules were found containing at least two SET3 proteins (see Table A in appendix). Modules ranged in size from 26 to 294 proteins, comprising between 1 to 4 predicted druggable targets.

The modules contain between 3 and 11 known cancer genes from CGC and

between 1 and 26 genes of our putative bladder cancer set (SET3). As in section 3.3.5 above, analysis of the significantly enriched GO-biological processes within the modules revealed a number of cancer related GO-terms including chromatin remodelling ($p<0.0001$), histone modification ($p<0.0001$), translation initiation ($p<0.001$). These enriched pathways and GO-terms were synonymous with those observed from the SET3 putative bladder proteins.

Three of the eighteen modules (Module 11, 12 and 16) were found to contain at least one known bladder cancer protein from CGC (HIFA1, RXRA and ERBB3 respectively). These were analysed in further detail.

Module-11 contains one known bladder cancer protein (CDK1NA), 8 known cancer genes in other types of cancer. This module has 7 putative bladder cancer proteins from SET3 of which one is the putative bladder cancer drug target HIF-1A, a hypoxia inducible factor protein. The genes within this module are found to be involved in histone modification. Hypoxia, as mediated by HIF1A, has been shown to trigger the coordination of chromatin regulating genes [189]. HIF1A is not a hub in the module (although it is connected to 5 other genes) and belongs to druggable CATH-FunFam 3.30.50.10.FF223 which has a median similarity score of 0.66. It has initially been shown that those druggable CATH-FunFam with a median similarity score greater than 0.48 have a less likelihood of being less associated with side effect [180]. This suggests that HIF1A could be considered as a target for bladder cancer therapeutic strategy. The drug topotecan has been previously associated with HIF1A in solid tumours [190] and can be considered for testing experimentally for bladder cancer.

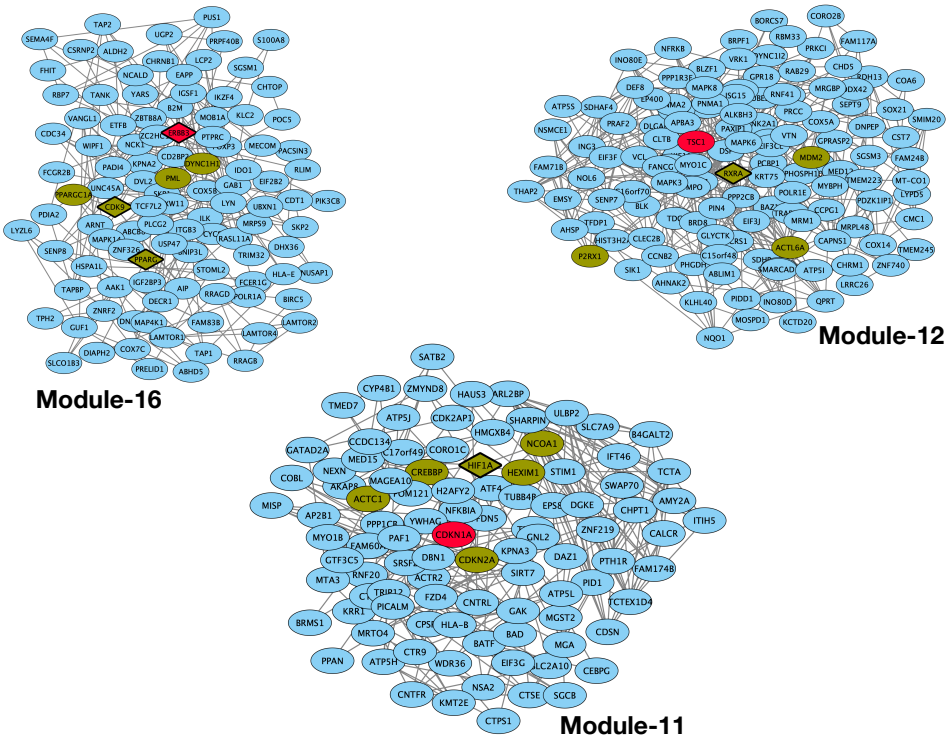

Figure 3.13: Modules clustered by MCODE clustering algorithm. The green coloured nodes are the putative bladder cancer genes while the red nodes are the known bladder cancer genes from Cancer Genome Census (CGC). The diamond shaped nodes are the drug target.

Module-12 contains one known bladder cancer protein (TSC1) and 9 proteins known drivers in other types of cancers. The module contains 5 proteins from SET3 and one putative bladder cancer drug target RXRA. TSC1 is involved in the activation of the mTOR branch of the PI3K pathway. Module-12 is also enriched in genes involved in chromatin remodelling, some of which have also been implicated in prostate cancer. The mutation of RXRA has been cited as one of the activators of peroxisome proliferator-activator receptors (PPARs) which switch on genes driving cell proliferations. RXRA also belongs to druggable CATH-FunFam 3.30.50.10.FF223 with median similarity score of 0.66. FDA approved drugs such as tretinoin and bexarotene have been shown to modulate RXRA expression level [191] and these drugs could be repurposed in treatment of bladder cancer.

Module-16 has in total 6 proteins from SET3. Two of these are putative bladder cancer drug targets identified (PPARG, CDK9) with one known bladder cancer

driver (ERBB3). This module is enriched in proteins associated with mTOR signalling pathway. mTOR signalling is known to be affected in most cancers and alteration of this pathway occurs in about 72% of bladder cancers [192]. The FDA approved drugs vandetanib and bosutinib that bind to ERBB3 are in trials for the treatment of prostate cancer [193], suggesting their possible suitability for bladder cancer once approved. ERBB3 belongs to the CATH-FunFam 1.10.510.10.FF78758 which has a median similarity of 0.22, a low similarity score has been predicted to be associated with a higher propensity of having side effect from the targeting this family. The other two remaining drug targets within this module (PPARG and CDK9) are found in druggable CATH-FunFams 1.10.565.10.FF5008 and 1.10.510.10.FF78743 respectively. These druggable families have a median similarity of 0.66 and 0.19 respectively. This suggests that PPARG is probably the best drug target in this module as it is less likely to be associated with side effects.

3.4 Chapter summary

In this study, an extended set of putative driver proteins was predicted for bladder cancer by combining known cancer drivers with putative drivers from CATH-MutFams and using this initial seed set (SET1) to search for additional drivers, located in the same modules as the seed set in a gene co-expression network generated using WGCNA. Only genes highly differentially expressed in bladder cancer were selected. This expanded set of putative bladder cancer genes (SET2) was further expanded (SET3) by searching for neighbours in a comprehensive human protein network, using a diffusion algorithm (DIAMOnD). This second network was constructed by combining the gene co-expression network with the human network from Pathway Commons.

GO enrichment and pathway analysis of the final set of 323 proteins in SET3 revealed molecular signalling pathways associated with chromatin modification and myogenesis, a phenomenon associated with muscle invasive bladder cancer. Also identified in this study are cancer hallmark signatures, although not limited to bladder cancer, such as G2M-checkpoint, Epidermal-to-mesenchymal-transition, P53,

Apoptosis as well as Notch-signalling, where mutations in the genes associated with these pathways are linked to driving oncogenic transformations associated in bladder cancer.

Currently there are only 10 approved drugs for bladder cancer. None of these are targeted small molecules. They include chemotherapy and immunotherapeutic drugs. The focus of this current study is on harnessing approved small molecules obtained from ChEMBL database used in treatment of other diseases. 28 of the 323 putative cancer drivers were found to be associated with drugs in the ChEMBL database. Using the druggable CATH-FunFams, this list of possible drug targets was expanded to 35. These FDA approved drugs are currently used to treat other diseases but can potentially be repurposed for targeting these putative drivers, subject to experimental validation.

3.5 Limitations and Future work

This study has considered genes implicated in bladder cancer using several data. One possible limitation may be the choice of the data used in the exploration of driver genes in bladder cancer. Bladder cancer heterogeneity and possible differences in stages and grades of the cancer was not considered in this study. This may have impaired the prediction of the drug targets for bladder cancer as different targets may affect different stages/grades of bladder cancer. One approach in dealing with this could be the classification of the genes involved in bladder cancers into those involved in various stages and use of this data to reclassify the drug targets.

Since, WGCNA, the tool for building the co-expression network, was only used to obtain modules significantly enriched with known bladder cancer genes from the heterogeneous TCGA samples, further analysis with WGCNA maybe help in assessing modules that reflect the subcategories of bladder cancer. In the future, the predicted FDA approved drugs and targets that are not currently in use for bladder cancer will require experimental validation before repositioning the drugs for bladder cancer.

Chapter 4

Protein Kinase Domain Families and their inhibitors

4.1 Introduction

4.1.1 Overview of Protein Kinases

Protein kinases are are implicated in several diseases and are of immense interest to the pharmaceutical industry (56% of all human proteins drug targets are protein kinases [116]). Protein kinases are enzymes that are involved in several cellular pathways. They catalyse the transfer of γ-phosphate of ATP to the hydroxyl groups of acceptor molecules which can either be protein substrates, lipids or small molecules. Through this phosphorylation process, the protein targets are covalently modified which leads to regulation of biological processes such as the control of metabolism, transcription processes, cell division and movement, programmed cell death and several other signal transduction events in the cell.

Protein kinases are the second largest enzyme family and the fifth largest family of genes in humans following zinc finger proteins, G-protein coupled receptors, immunoglobulins, and the proteases [194]. About 2% of the protein encoding part of the human genome has been shown to encode protein kinases. Manning et al (2002), identified all sequenced eukaryotic protein kinases by searching all human

genome sequence sources including the Celera Genomics databases, Incytes EST, Genbank cDNAs and expressed sequence tags (ESTs) using hidden Markov Model (HMM) profiling of the known kinase sequences to identify related protein kinase domains [195]. Overall, they identified 518 human protein kinase genes of which 478 were classifed as eukaryotic protein kinases (ePKs) while 40 were Atypical protein kinases (aPK) which lack sequence similarity to the eukaryotic kinase domain but have been reported to have kinase activity.

The catalytic domain of eukaryotic proteins kinases is highly conserved both in sequence and structure. Protein kinase activity requires the binding of a peptide substrate, which is to be phosphorylated, and the ATP to the catalytic domain. Protein kinases can be broadly classified as either tyrosine kinases or serine/threonine kinases based on the specificity of the substrate they phosphorylate and can then be divided into groups, families and subfamilies. There are 9 groups of protein kinases based on the sequence and structural similarities of the catalytic domain [195]. Classification is also guided by knowledge of the domain structure outside the catalytic domain, known biological functions and evolutionary history of the kinases. The Manning classification (KinBase database) is an extension of the work by Hanks and Hunter [196] who initially performed a conservation and phylogeny analysis of the catalytic domain of the eukaryotic proteins to reveal the conserved features of catalytic domains and thus, classified the protein kinases into 5 groups, 44 families and 51 subfamilies [196]. Manning and colleagues further extended this to 9 groups, 134 families and 196 subfamilies. Figure 4.1 below shows the grouping of the human protein kinases.

Other classification schemes for the protein kinases have also been developed over the years. For instance, Saavedra and Barton used a multilevel hidden Markov model library to classify protein kinases into 12 families for 21 eukaryotic genomes [197]. They reported that classification by a multilevel HMM library outperformed BLASTP and the single HMM classification used by KinBase. Multilevel HMM classification involves building subfamily HMMs rather than a single HMM for the entire family. The classification by Saavedra and Barton was built using sequences

derived from KinBase [195] and this is displayed in the Kinomer database [198]. Another classification by Martin *et al* considered the composition of the kinase accessory domains and the organisations of these domains. In their approach, the classification was performed manually and used an alignment-free method to detect the similarity between sequences by assessing short amino acid sequence patterns and structural features outside the catalytic domain [199]. Using this approach, they were able to detect outliers called ”hybrid kinases” that had sequences in the catalytic domains matching with a particular subfamily but sequences outside with catalytic domain matching with a different subfamily [200]. The classical classification approach using only the kinase sequences would not have been adequate to capture this. This classification scheme stores kinase families in the KinG database [201].

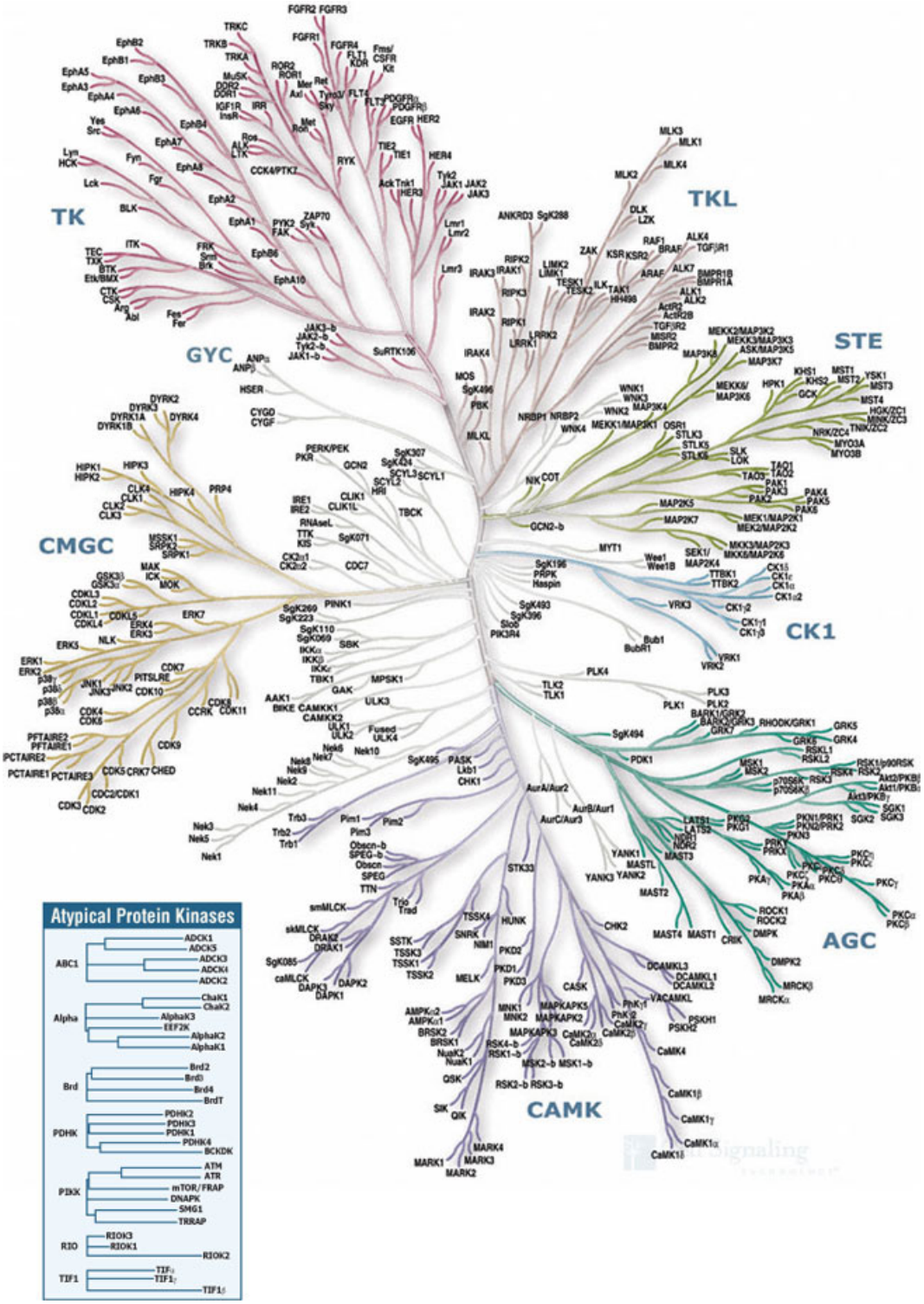

Figure 4.1: The human kinome. Kinome illustration courtesy of Cell Signalling Technology, Inc (www.cellsignal.com) based on [195].

4.1.2 KinBase classification of Protein Kinases

The comprehensive work done on the classification of kinases by Manning *et al* is widely used and has been cited over 7000 times (Google scholar). Below, the groups identified by Manning *et al*. (2002) are described in more detail.

AGC Kinases

This group of kinases includes PKA, PKG, and PKC which are involved in diverse cellular roles such as cell growth and proliferation, cell survival, glucose metabolism and protein synthesis. They are dysregulated in several diseases such as cancer and neurological disorders, inflammation and viral infection [200]. The Akt isoform possesses the Pleckstrin homology domain (PH-domain) at the N-terminus which interacts with PIP3 and PIP2 leading to the activation of pyruvate dehydrogenase kinase isoenzyme-1 (PDK1) as shown in figure 4.2. PKC also interacts with diacyl-glycerol (DAG) and calcium by its N-terminal conserved domains (C1 and C2) which leads to conformational changes and activation of the protein [202].

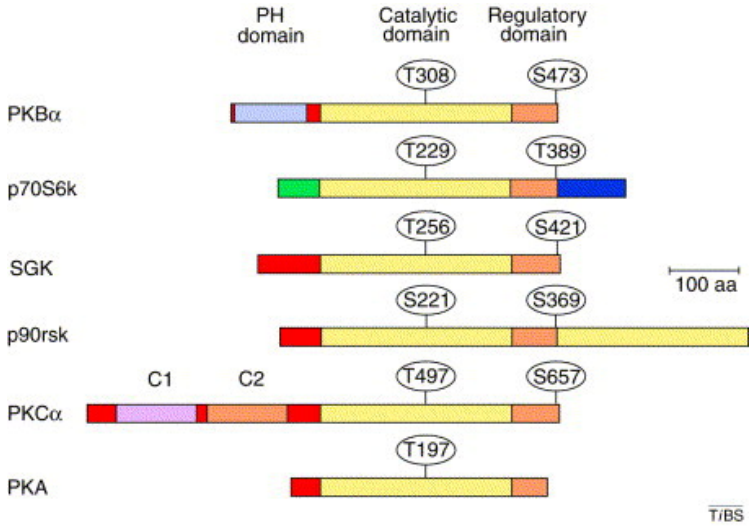

Figure 4.2: The domain structure of AGC kinase family. All members contain Thr/Ser in the activation loop. Figure taken from [203]

CAMK Kinases

These kinases are involved in calcium signalling and are normally autoinhibited. The binding of Ca^{2+}/calmodulin complexes relieves this autoinhibition. Members of this group include MLCK, RAD53, PKD, CAMK2, Trio, CAMKL, DCAMKL,

CASK, and DAPK subfamilies all of which are found in multidomain proteins. Each member of this family possesses additional unique domains in addition to the conserved kinase domain. For instance, the Ca^{2+}/calmodulin dependent serine kinases (CASK) contains several interacting domains; two L27 (LIN-2 and LIN-7) interacting domains, a PDZ (PSD-95-Dig-Z01), Src homology 3 (SH3) domain and a C-terminal guanylate kinase domain [200]. The PKD kinase also possesses a PH domain as found in the Akt family and this is important for the regulation of its enzymatic activity. Figure 4.3 shows the structural arrangement and multidomain architecture of CAMK2.

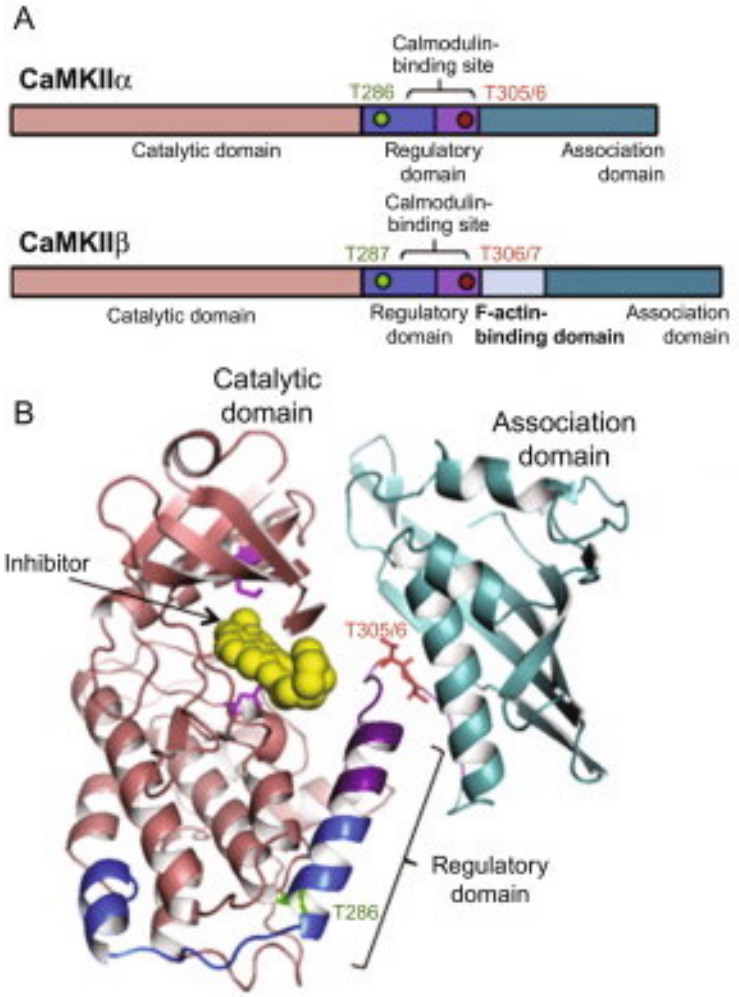

Figure 4.3: Domain organisation and structure of CaMKII. (A) There is a similar domain organisation in the CaMKIIα and CaMKIIβ with the exception of an F-actin binding domain inserted into CaMKIIβ. (B) Structure of CaMKII subunit PDB ID: 3SOA. Figure taken from [204]

CK1 group

The cell kinase 1 (CK1) members are quite ubiquitous in their phosphorylation events as they have a wide range of substrates. They are Ser/Thr kinases and are constitutively expressed. The kinases in this group are single domain proteins i.e. they do not possess additional non-catalytic domains apart from the CK1-γ subfamily which possesses a CK1-γ domain whose function is not yet known as shown in figure 4.4.

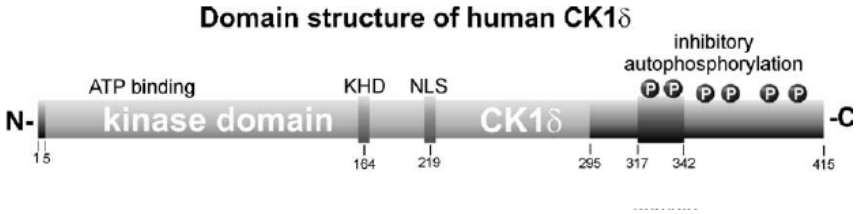

Figure 4.4: Domain structure of human CK1δ. The members of this subfamily share a common conserved kinase domain but differ in their variable N-and-C terminal domains. The regulatory c-terminal domain has multiple inhibitory autophosphorylation sites. The nuclear localization signal (NLS) and kinesin homology domain (KHD) are also located within the kinase domain. Figure obtained from [205]

CMGC group

Members of this group possess single domains like the CK1 group. They include dual specificity tyrosine regulated kinases, dual specificity yak-related kinases (DRYK), cyclin-dependent kinases (CDKs), MAPK, GSK-3, CDK-like kinases. CDKs regulate the progression through the different phases of the cell cycle in association with their cyclins activating partner. The MAP kinases are amongst the most highly studied signal molecules. The MAP kinase cascade controls cell proliferation, differentiation, and death across various eukaryotes. The GSK-3 kinases are key metabolic enzymes in glycogen metabolism and play a role in the *Wnt* pathway which is important in embryonic development.

Tyrosine Kinase group (TK-group)

These kinases catalyse the phosphorylation of tyrosine residues and are heavily implicated in cancer. The tyrosine kinases are divided into 2 families:receptor and non-receptor (cytosolic) kinases. The receptor TKs are subdivided based on the

sequence similarity and the structure of their extracellular domains into 20 subfamilies. One of the most studied extracellular domains is the Ig-like domain which occurs in most of the members of this subgroup. The extracellular domains act as the ligand binding sites for several growth receptors. The non-receptor kinases are subdivided into 10 subfamilies which include Src, Abl, Ack, Csk, Fak, Fes, Frk/Fyn, Tec and Syk [200]. In addition to the kinase catalytic domain, they also possess additional domains that are important for enzymatic regulation and substrate recognition. The Src family for instance possesses additional SH3 and SH2 domains.The Abl subfamily has an F-actin binding site and a DNA-binding region; FAK possess a FERM domain and a focal adhesion-binding domain which are important for mediating protein-protein interation [202, 200]. The multidomain architectures of receptor and non-receptor kinases are shown in figure 4.5.

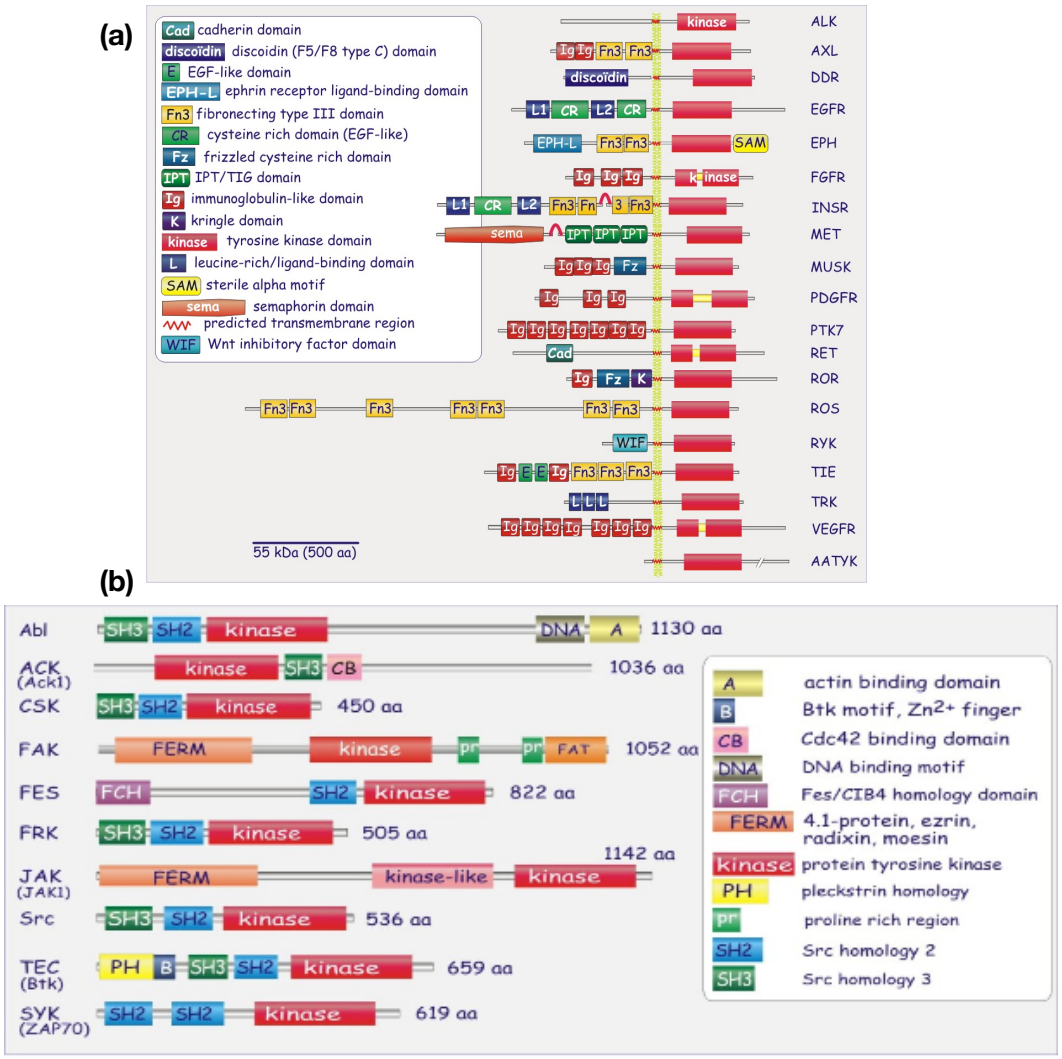

Figure 4.5: The multidomain architecture of tyrosine kinases (a) Receptor tyrosine kinases; (b)Non receptor tyrosine kinases. Figures were taken from [206].

Tyrosine kinase-like group (TKL group)

The members of this group have close sequence similarity to tyrosine kinases, however, they are mostly serine/threonine kinases and lack the TK-specific motifs. They are mostly diverse with members including receptor and non-receptor kinases. They comprise 8 major families which include IRAK, STKR, RIPK, RAF, LRRK, MLK, MLKL, and LISK.

STE-group

The members of this group are classified into three major families. They include STE20 (MAPK4), STE11 (MAPK3) and STE7 (MAP2K). STE stands for ”Sterile”

and was originally identified in yeast. The STE kinases sequentially activate each other to then activate the MAPK family.

RGC-group

The receptor guanylate cyclase represents the smallest group of kinases and consist entirely of pesudo-kinases that lack certain residues that are critical for phosphate transfer [195]. They convert GTP to GMP.

Others

These include members that lack sufficient sequence homology to any of the groups given above and display unusual phosphorylation properties, using ATP and GTP as phosphate donors. Examples include CK2, IKKs.

Atypical protein kinases (aPks)

The atypical kinases represents a group of human kinases that lack sequence similarity with the eukaryotic protein kinase (ePKs) domain HMM profiles, but have been shown experimentally to have protein kinase activity. Examples include PIKK family, A6 family, RIO and Pyruvate dehydrogenase kinases [195]. Domain organisation in atypical kinases is shown in figure 4.6.

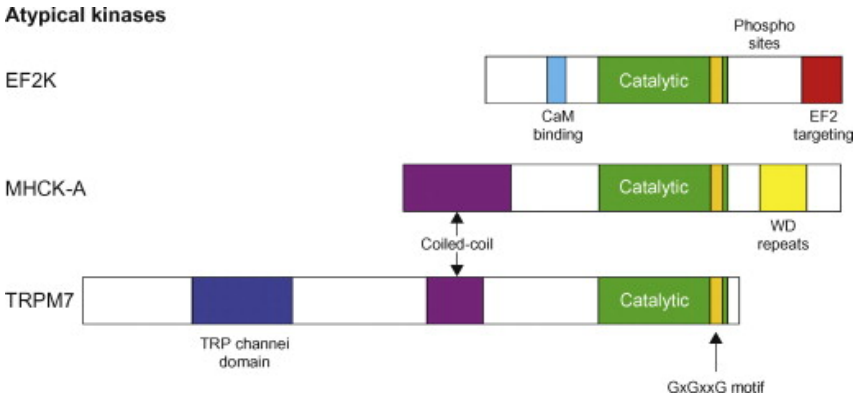

Figure 4.6: Domain organisation of the atypical family of protein kinases. In contrast to classical kinases (EF2K), the GXGXXG motif of atypical kinases(TRPM7) is not involved in MgATP binding but is likely to be involved in peptide interaction. Figure taken from [207]

The crystal structure of the catalytic domain of TRPM7 provides insights into the enzymatic function of the atypical kinases. The comparison of the structure of TRPM7 and PKA catalytic domain reveals some of the major differences between

these two kinases. See Figure 4.7.

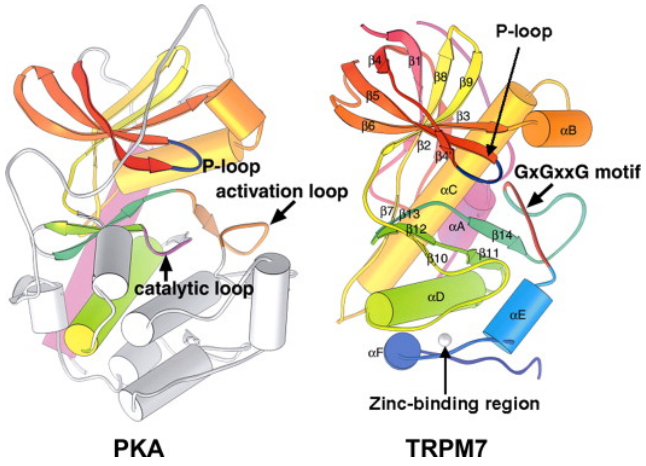

Figure 4.7: Structural comparison of the kinase domains of TRPM7 and PKA. The N-lobe of both PKA and TRPM7 is largely comprised of β-strands and the MgATP binds at the cleft formed from both the N and C-lobes and the binding of Mg in both also involves the conserved P-loop. However, the catalytic loop is not conserved. The GXGXXG motif in TRPM7 is in an extended loop that may play a similar role to the activation loop in classical protein kinase, PKA. Figure obtained from [207].

4.1.3 Structural Features of Protein Kinases

Structural features of the catalytic domain of the protein kinases are described below. Protein kinases possess catalytic domains and non-kinase domains that are responsible for the regulation, scaffolding and substrate specificity. Some of these additional non-kinase domains have been mentioned in the section above. As mentioned earlier, the catalytic domain of the kinase spans about 250 residues and is highly conserved. It has two dissimilar lobes (the N-lobe and the C-lobe) joined by a peptide coil called the linker (See figure 4.8a). The N-lobe has about 90 amino acids that fold into 5 β-strands and one helix (C-alpha-helix). This lobe contains the nucleotide binding site that recognises and binds ATP. The C-lobe is the larger lobe and is mainly alpha-helical.

In the N-lobe, there are highly conserved sequence motifs that are embedded within the first three stands. The first is the GXGXXG motif (Gly-rich loop) which is between $\beta 1$ and $\beta 2$. This loop folds over the nucleotide and positions the ATP γ-phosphate for catalysis. It is the most flexible part of the N-lobe [208]. Another

important loop is the P-loop also called the Walker-A motif (GXXGKT/S). Both the glycine rich motif and the P-loop bind to the nucleotide bound phosphate. However, their interaction with purines is different. For instance, the P-loop does not contact the purine moiety of the ATP while the Gly-rich loop connects the β strands that harbour the adenine ring; the Gly-rich loop is also followed by a conserved Val within the $\beta 2$ strand that makes hydrophobic contact with the base of the ATP [209]. The third important motif is the AxK motif which is found in the $\beta 3$ strand. The lysine from this motif couples the α and β-phosphate of the ATP to the C-helix.

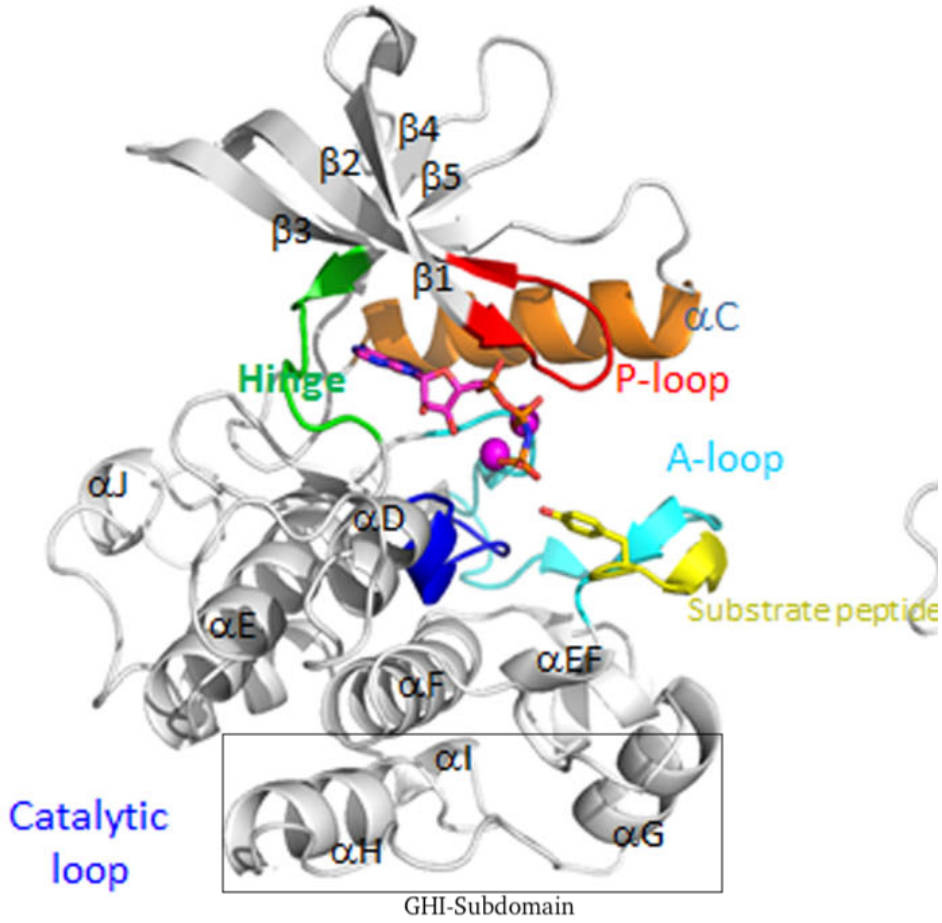

Figure 4.8: The structure of the conserved kinase core showing the bilobal characteristics of kinases. The figure was adapted from [209]

The C-helix in the N-lobe serves as a ”signaling integration motif” as it connects to different parts of the kinase domain. Its C-terminus is connected to the C-lobe by the αC-$\beta 4$ loop whereas the N-terminus interfaces with the activation

loop. Correct positioning of the C-helix is required for activation of the kinase. The distance between the N-terminus and the activation loop of the C-helix is a measure of the open and closed conformation, essential for the catalysis [208].

The C-lobe varies in size, sequence and topology. It is predominantly alpha-helical but also contains a few beta strands. It contains the substrate binding groove, activation loop and the catalytic residues. This helical subdomain forms the core of the kinase and the protein/peptide binding surface. The backbone amide of the core helices (D, E, F and H) are not solvent accessible with the exception of the G-helix. The β-subdomain of the C-lobe comprises 4 short β strands (6-9) and contains much of the catalytic machinery for transferring the associated phosphate from the ATP to the protein substrate. The substrate binding site is formed by hydrophobic residues contributed by the helical core. The activation segment is marked by a conserved Asp-Phe-Gly (DFG) (magnesium positioning loop) and Ala-Pro-Glu (APE) motif.

The activation loop extends from the DFG motif to the aspartate at the beginning of the F-helix. The length and sequences of the activation loop are the most variable part of the kinase core and this is responsible for turning on and off the kinase [208]. Furthermore, the F-segment extends to the GHI-subdomain (an extension of G-helix through I-helix) as shown in figure 4.8 where substrates and regulatory proteins bind. This part is also responsible for stabilizing the active kinase core and contains allosteric sites.

The hinge region of the kinase is the loop connecting the N and C-lobe. It contains several conserved residues which provide the catalytic machinery and make up part of the ATP binding pocket. The local spatial pattern alignment (LSP) is a method for comparing two protein structures and identifying spatially conserved residues [208]. LSP revealed two hydrophobic motifs called ”spines”, that connect the N and C-lobes. Structural analysis of the spines gives insight into how an active protein kinase is assembled from an inactive protein kinase [208, 210]. Two spines are observed to be involved in regulation of protein kinases (R-spine and C-spine). The R-spine comprises four non-consecutive hydrophobic residues; two from the

N-lobe (Leu^{106} from β4 and Leu^{91} from C-helix) and the other two from the C-lobe (Phe^{185} from the activation loop and Tyr^{164} from the catalytic loop) as shown in figure 4.9. The R spine is therefore a hydrophobic spine that links the two lobes.

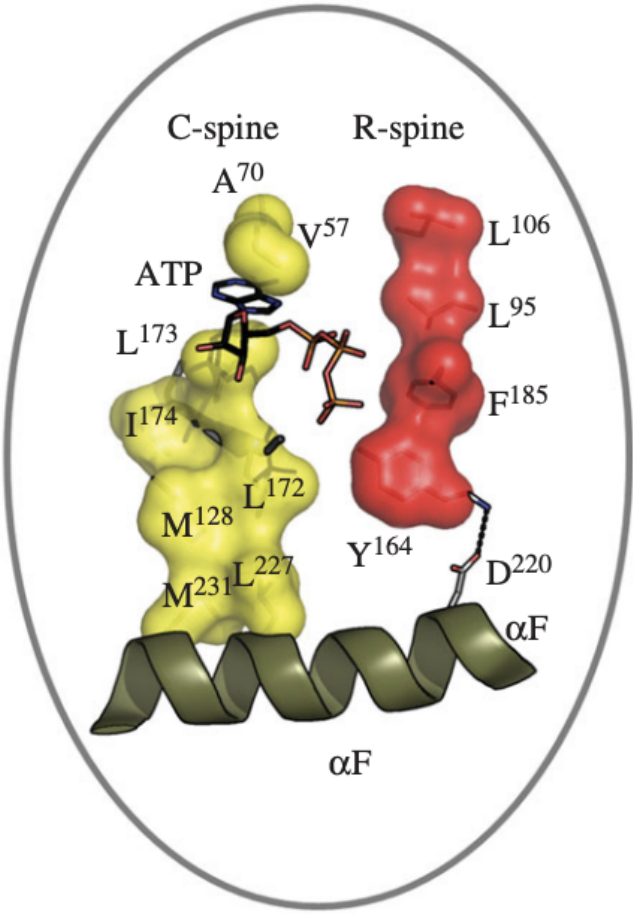

Figure 4.9: The catalytic spine (C-spine; yellow) and regulatory spine (R-spine; red) of cAMP-dependent protein kinase (PKA). The figure was adapted from [210]

Using LSP on the conserved core of the protein kinase, another hydrophobic spine was identified, called the catalytic spine (C-spine). Like the R-spine, it comprises hydrophobic residues belonging to both lobes. In the N-lobe, Val^{57} in β2 and Ala^{70} from the AxK-motif as well as Leu^{173} in the C-lobe docks directly onto the adenine ring of the ATP forming the C-spine. Both spines are anchored to the hydrophobic αF-helix. Once the R-spine is assembled, and the C-helix is correctly oriented, then the kinase is primed for catalysis. The binding of ATP completes the C-spine and commits the kinase for catalysis [194, 208, 209]. From a structural per-

spective, Jacob *et al* [211] compared the 426 available structures corresponding to 71 distinct human protein kinases based on 2 structural elements (the activation segment and the C-helix) and clustered the kinases into three conformations indicating the catalytically active or inactive state of the kinase [211].

4.1.4 Active and Inactive Protein Kinases

Analysis of the structural elements of the kinases show distinct conformations in the active and inactive states. The activation loop for instance is usually in an extended conformation in its active state whereas it is disordered with the loop collapsed to block the substrate binding, in the inactive state (see Figure 4.10). The structures of protein kinases has revealed the conformational variation of active and inactive kinases. One of the most common forms of inactive protein kinases is the positioning of the aspartate of the DFG-motif of the activation segment in an "out-conformation" whilst the phenylalanine of the DFG-motif is directed inward towards the active site [194]. The phosphorylation of the residues within the activation loop activates the kinases [194].

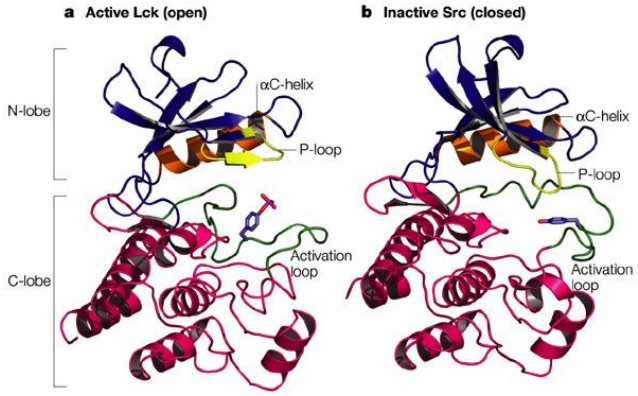

Figure 4.10: The active and inactive conformation of LCK and Src respectively. (a) Active conformation with activation loop adopting an extended conformation while (b) loop is folded back in the inactive c-Src kinase domain. Figure taken from [212].

Furthermore, the presence of a salt bridge between the β3-lysine and the αC-glutamate, together with the formation of the R- and C-spine, are the hallmarks of an active kinase domain while inactivation involves the disassembly of the R-spine. The rotation or movement of the αC-helix also causes a switch from an inactive to

an active kinase as the αC adopts an "in-conformation" in its active state and an out-conformation in its inactive state. [213, 194].

4.1.5 Kinase Inhibitors

Kinases display remarkable diversity in their primary sequences, substrate specificity, structure and the pathways associated with them [214]. However, they share a great degree of similarity in their 3D structure most especially in their catalytic site where the ATP-binding cavity is found. ATP binds in the cleft between the N and C lobes and therefore most kinase inhibitors interact with this region to perturb the binding of ATP [215]. There are several kinds of inhibitors that are being exploited to target protein kinases. These inhibitors differ in their mode of binding and the mechanism of action exhibited upon binding. The kinase inhibitors can either bind covalently or reversibly.

The non-reversible (covalent) inhibitors bind irreversibly with the reactive nucleophilic cysteine or lysine residue close to the ATP-binding site resulting in the blockage of ATP binding and leading to irreversible inhibition. An example of such a drug in clinical trials is AVL-292 which is a tyrosine kinase inhibitor which covalently binds to Bruton tyrosine kinase (BTK) [216]. Ibrutinib targets BTK as well, while Afatinib targets the Gefitinib resistant EGFR as shown in figure 4.11 [217].

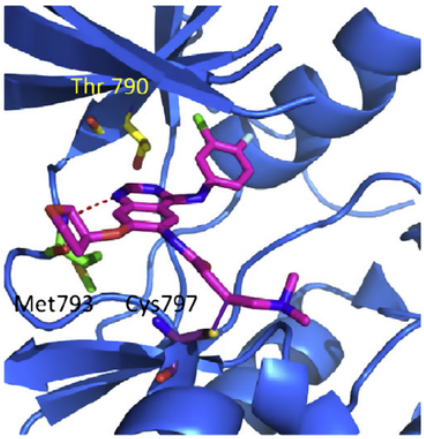

Figure 4.11: Afatinib co-crystal structure with wild-type EGFR (PDB ID: 4G5J). Afatinib binds to the kinase domain in its active conformation and forms a hydrogen bond with the backbone NH of Met793 and also forms covalent interaction with the sulphur of Cys797. Figure obtained from [218]

Reversible (non-covalent) inhibitors on the other hand can be classified into several types, based on their interaction with the binding pocket and the DFG motif (hinge region). Type-I inhibitors are ATP-competitors that bind to the active form of the enzyme with the aspartate residue of the DFG motif facing the active site of the kinase (DFG-in conformation). The conserved Phe of the DFG-motif is buried within the hydrophobic pocket of the groove between the N and C-lobes. Most of the compounds that target this active conformation have been selected using enzymatic assays that select ATP mimetics with the highest inhibitory activity for the kinase [209].

Classical examples of such approved inhibitors include gefitinib, dasatinib, erlotinib and sunitinib. Figure 4.12 shows type-I protein kinase inhibitors (Erlotinib and Lapatinib) interacting with the enzyme in its active and inactive state.

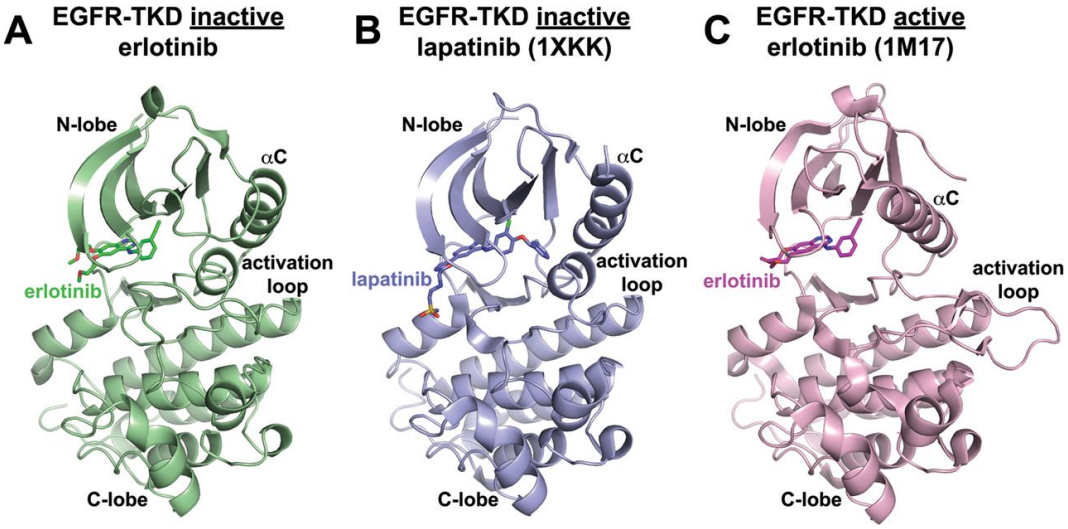

Figure 4.12: Crystal structure of EGFR tyrosine kinase domain (TKD) bound with inhibitors. (A) Erlotinib bound with EGFR-TKD in the inactive state (B) Lapatinib with inactive EGFR-TKD (1XKK) (C) Erlotinib with the active EGFR-TKD (1M17). Figure obtained from [219]

Type-II inhibitors bind to the inactive form of the enzyme with the aspartate residue of the DFG motif protruding outward from the ATP-binding site of the kinase. The transition from the DFG-in to DFG-out conformation exposes the hydrophobic pocket adjacent to the ATP-binding site and this is utilized by the type-II inhibitors to lock the kinase in an inactive conformation [220]. Type-II are generally

less promiscuous compared with the type-I inhibitors.

Examples of FDA-approved type-II kinase inhibitors include imatinib, nilotinib, and sorafenib [209]. The type-I and II inhibitors however face competition with the millimolar concentration of ATP *in vivo* as well as a lack of selectivity due to the extensive adenosine binding cleft [221]. There have therefore been efforts directed towards kinase inhibitors with high selectivity, high affinity and less side effects.

Type-III inhibitors are a heterogeneous group of kinase inhibitors that bind to allosteric or remote sites on the kinase. These inhibitors mostly do not bind at the ATP-binding sites and have no physical contact with the hinge. They have been shown to exhibit the highest form of selectivity by exploiting the binding and regulatory sites that are specific to a particular kinase [209]. The combinations of the structural elements in the kinases such as the C-helix's DFG-in and out state, A-loop, G-loop, C-terminal elements as well as regulatory domains can be exploited to design selective inhibitors with clear advantages over the type-I and II inhibitors [220]. Examples of approved type-III inhibitors include cobimetinib, trametinib, selumetinib, binimetinib and rapamycin. Type-III inhibitor of MEK1 binds to the adjacent pocket to the ATP-site which is referred to as the "allosteric back pocket-DFG-in" in the presence of ATP and "allosteric back pocket-DFG-out" in the absence of ATP as shown in figure 4.13 [209].

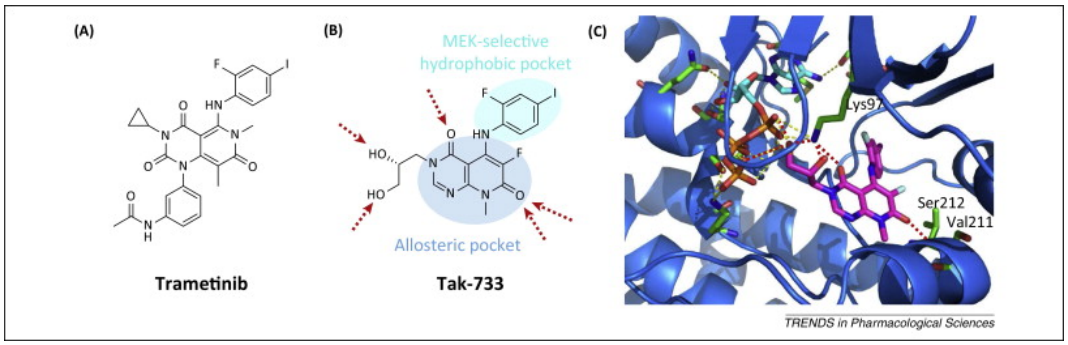

Figure 4.13: MEK kinase inibitor binding mode (A). The chemical structure of trametinib (B) The binding mode of trametinib with MEK1 (C). Tak-733 co-crystallized with MEK1 (PDB ID: 3PPI) ATP is shown in cyan and Tak-733 in magenta. Figure taken from [215]

.

Type-IV allosteric inhibitors bind at allosteric sites that are distant from the ATP-binding site. A unique example is the AktI-1/2 targeted inhibitor that inhibits Akt isoforms 1 and 2 kinases. These inhibitors have no effect against PH-domain mutants which suggest that the PH domain is required to exert their activity and that the inhibitor interacts with both the catalytic domain and the PH domain and prevents the activation of the upstream kinase PDK1 [222].

Other types of allosteric protein kinase inhibitors include the type-5 which are also referred to as bivalent or bi-substrate inhibitors. The bivalent inhibitors tend to have high affinity and more selectivity for targeted therapy. The design of such inhibitors involve the use of an appropriate linker to couple the allosteric site inhibitor with the kinase active site binding agent to achieve improved selectivity from the non-ATP directed inhibitor [221]. Another example of kinase inhibitors is the hybrid-type having both type I and II features. The field of allosteric kinase inhibition is a rapidly evolving field with the recent FDA-approval of trametinib as well as several other allosteric inhibitors that are in clinical trials [223]. A schematic overview of the interaction between the various types of inhibitors and the kinases is shown in figure 4.14

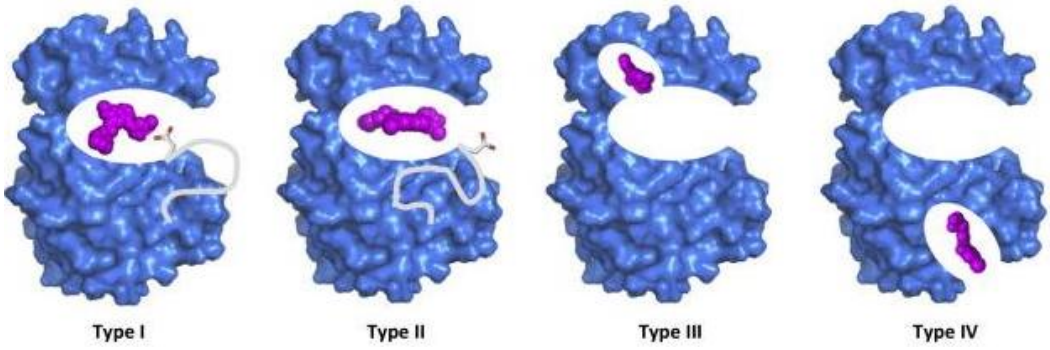

Figure 4.14: Schematic overview of the four types of reversible binding mode of kinase inhibitors. Figure taken from [215]

.

Allosteric inhibition offers some advantages such as high selectivity and ability to overcome drug resistance as most drug resistance to small molecule kinase inhibitors occurs frequently around the hinge region. However, there has been debate about their efficacy since mutation-related resistance may also occur at the allosteric

sites as they are not as essential for kinase function as the ATP binding sites. Also, as a result of the hydrophobic properties of most allosteric pockets, the allosteric inhibitors are lipophilic compounds and this may result in poor bioavailability, and poor solubility. Another major challenge is the limited numbers of structures for allosteric-inhibitor-bound kinases to help in the comparison of the induced changes associated with the on/off-bound state of the enzymes. This may be due to the fact that these sites are involved in protein-protein and protein-peptide interactions and the transient nature of such interactions creates difficulty in solving the structures [223].

4.1.6 Understanding the promiscuity of protein kinase inhibitors

Promiscuity is defined as the ability of a compound to specifically interact with more than one target (the target of interest for which it was designed) [224]. Protein kinase inhibitors are generally considered promiscuous because of their lack of specificity and their ability to interact with several kinases and kinase families, due to the common ATP-binding site that kinase inhibitors interact with. Hu et al classified the protein kinse inhibitors into single and multiple kinase inhibitors by simply counting the numbers of targets the PKI compounds in the ChEMBL database were active against. Furthermore, they also assessed the promiscuity of a kinase for several structurally diverse compounds and found that many kinases recognise structurally diverse compounds [224].

The promiscuity associated with protein kinase inhibitors can lead to various side effects. This is because many developed kinase inhibitors are not target specific and can combine with several potential targets eliciting downstream responses which are associated with side effects. For instance, Giansanti *et al*. evaluated the promiscuous nature of 4 tyrosine kinase inhibitors (imatinib, dasatinib, bosutinib and nilotinib), in epidermoid carcinoma cells using proteomics techniques, as a model system for skin cancer [225]. They observed that over 25 tyrosine kinases had affinity for the drugs with imatinib and nilotinib displaying more specificity while the other two showed larger downstream effects on the phosphotyrosine sig-

nalling pathway. The promiscuity of kinases has been well studied from both experimental and computational perspectives, providing selective criteria that can be used to minimise the off target effects.

Using a computational approach, Huang *et al*. mapped all the 518 human kinase sequences onto a multiple structural alignment of 116 kinases of known 3D structure [226]. They considered the ATP binding sites and encoded the residues in a 9-bit fingerprint (physico-chemical characteristics of residues) into network. Network analysis was used to partition kinases into clusters with similar fingerprints thus enabling more selective targeting of protein kinases.

Databases like the KIDFamMap [227] provide biological insights into the selectivity of kinase inhibitors and the mechanism of binding. The database provides information on kinase-inhibitor families as well as kinase-inhibitor disease relationships. The database also provides KIDFamMap ”anchors” which represent conserved interactions between kinase subsites and the chemical entities of the inhibitors. The KIDFamMap database creates a platform for accessing the conformation, function and selectivity of kinase inhibitors.

4.2 Objectives of chapter

This chapter focuses on protein kinases and their inhibitors and reports a classification of kinases into functional families (FunFams) in order to examine the potential side effects of drugs targeting particular families. This new classification approach developed will help in grouping kinase sequences that do not map to the canonical KinBase classification. The method considers the multidomain architecture of proteins which was not considered in previous classification methods.

Kinase sequences were grouped into functional families (CATH-FunFams) using a newly developed in-house sequence-based method, ”GARDENER”, developed by Dr. Nicola Bordin in the Orengo group. This approach was then benchmarked using multiple criteria to ascertain the performance of the novel classification protocol.

Subsequently, using a set of publicly available protein kinase inhibitors as well

as FDA-approved kinase inhibitor drugs, the association of kinase CATH-FunFams with protein kinase inhibitors was assessed. Network characteristics were examined to determine which kinase CATH-FunFams were likely to have side effects and also shed light on possible repurposing of protein kinase inhibitors to other members of a given family. The outline for the study is summarized in figure 4.15 below.

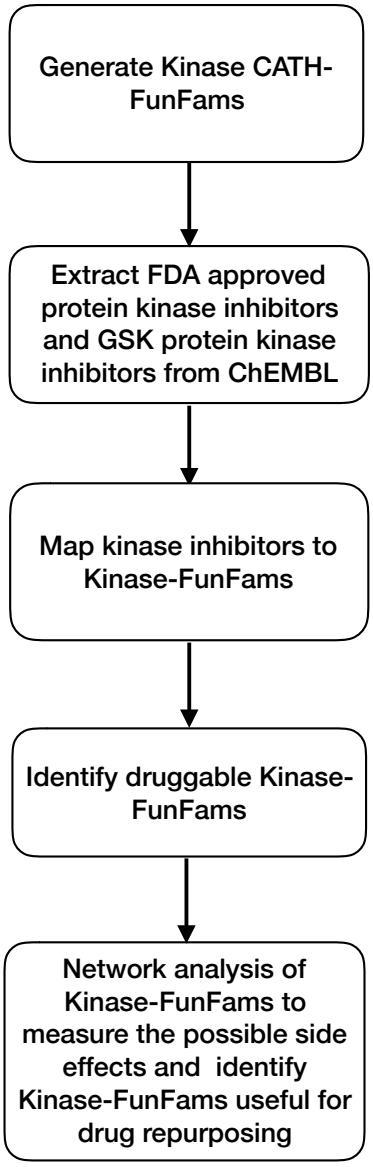

Figure 4.15: Framework of the methodology used in this study

4.3 Materials and Methods

4.3.1 Generating Kinase CATH-FunFams

Updating the protein kinase family in CATH-Gene3D

In CATH-Gene3D version 4.2, the kinases are classified into two separate domain superfamilies based on the distinct structural regions of a typical protein kinase (N- and -C domains). These are represented as the CATH-Superfamilies 3.30.200.20 (N-domain) and 1.10.510.10 (C-domain) respectively. As the majority of the protein kinase inhibitors act at the hinge region between these two domains, it is important to generate a class of Kinase-FunFams incorporating both domains, that can be used for studying kinase-drug inhibitory mechanisms.

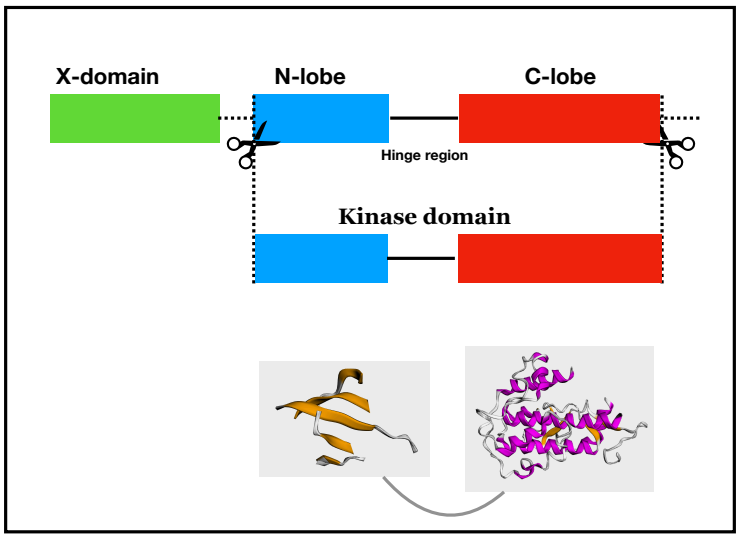

Figure 4.16: Schematic representation of obtaining kinase sequences from CATH/Gene3D resource. The blue box represents the 3.30.200.20 (N-domain) while the red box represents the 1.10.510.10 (C-domain) of the kinases. Other accessory domains represented as the green box.

Since, the most cited kinase classification, the KinBase classification by Manning *et al* [195] has not been updated since 2002, a method that can handle all available kinase sequences and classify them into functionally coherent families is required. To this end, a kinase functional family (Kinase-FunFam) generating protocol was adopted based on a novel in-house approach (GARDENER) that can handle large superfamilies, such as the kinase superfamily, which comprises more

than 300,000 sequences. The protocol adopted is outlined below.

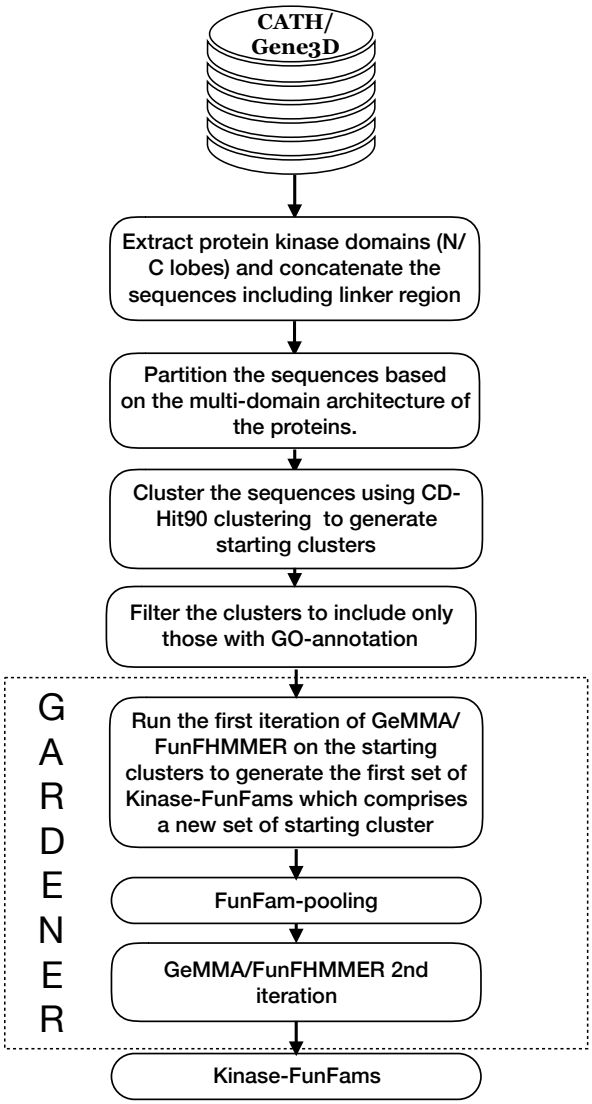

Figure 4.17: Kinase-FunFams generation protocol

Obtaining kinase sequences

Kinase sequences from CATH/Gene3D version 16 [128] with CATH-superfamily '1.10.510.10' and '3.30.200.20' were obtained and concatenated to give the entire kinase unit allowing an extra 20 residues connecting the domains that represents the kinase hinge region (See figure 4.16. The multidomain architectures (MDA) of the proteins (as in the domain partners and order and repetition of the kinase domains) were used as a guide in the partitioning protocol for the GARDENER algorithm.

Running the GARDENER algorithm

The GARDENER algorithm performs iterations of GeMMA/FunFHMMER (See introduction section 1.5.3). Within each MDA partition, the sequences were clustered into 90% sequence identity clusters (s90) using CD-HIT [87]. These are the starting clusters for GeMMA [86]. A list of experimentally-derived GO terms was obtained from the EBI-Protein GO API. Clusters without an associated GO annotation were discarded. Thereafter, the first iteration of GeMMA was performed, obtaining the input tree for FunFHMMER [135]. The resulting FunFams from this iteration were renamed, pooled and fed into GeMMA and FunFHMMER for a second iteration of tree building and cutting, resulting in the final Kinase-FunFams (See figure 4.17).

4.3.2 Benchmarking approaches to validate Kinase-FunFams

Two benchmarking strategies were adopted to ascertain whether the clustering approaches generated functionally relevant kinase families.

Functional purity of Kinase classification based on EC-annotation

The purity of the Kinase-FunFams was analysed by determining whether relatives in each Kinase-FunFam had the same or similar Enzyme Classification (EC) numbers. This approach of using functional information based on the EC numbers have been used in the assessment of the functional purity of FunFams [135]. The Enzyme Classification is a numerical classification scheme for enzymes based on their chemical reaction [228]. The EC-number is a 4 digit number with each digit representing either the class or subclass. The first digit describes the general type of reaction the enzyme undergoes. There are 6 classes of enzymes (oxidoreductase, transferases, hydrolases, lyases, isomerases and ligases) represented by EC:1-6 respectively. The second digit is the subclass which describes the type of bond breakage or formation taking place; the third digit represents the sub-subclass which provides further information on the chemical group involved in the enzymatic reaction while the fourth level indicates the substrate specificity of the enzyme [228].

The enzyme classification number (EC-class) of members in each FunFam was

obtained and compared both at the 3-digit (EC3) and 4-digit (EC4) level. The numbers of different EC numbers identified in a FunFam gives a measure of the functional purity of the Kinase-FunFam.

Comparison of Kinase-FunFams and KinBase

The Kinase-FunFams were also benchmarked using the KinBase (Manning et al) kinase family classification which has been curated and widely used [195, 198, 229]. Since KinBase only provides the GeneID's and FASTA file of the kinase proteins, the sequences were mapped to UniProt using BLAST to obtain the UniProt ID, identifying matches with 100% sequence identity and E-value less than $1e^{-3}$. The sequences within each KinBase were then scanned against the Kinase-FunFam database to give a mapping between the classifications. The EC-annotation at both level 3 and level 4 was compared between Kinase-FunFams and KinBase subfamily classification.

4.3.3 Protein Kinase Inhibitor Dataset

The Published Kinase Inhibitor Set (PKIS) is a collection of 367 compounds that have been made available by GSK to the research community [230, 231]. These compounds have been annotated with protein kinase activity [231] and are of various chemotypes and are openly available from the ChEMBL database (release 23) [179]. The PKIS are active against some known target kinases and can be extended to other new target kinases. PKIS subsets that showed an inhibitory activity level above 50% were selected, since Dranchak *et al.* (2013) and Anastassiadis *et al.* (2011) had reported this threshold as appropriate for considering the inhibition of kinase catalytic activity [230, 232].

An FDA-approved kinase-inhibitor drug dataset was also extracted from ChEMBL release 23. A drug was considered as a small molecule with therapeutic application, with direct binding to a single protein (ASSAY-TYPE ="B"), having a maximum phase of development ="4" which indicates that the drug has been approved. Those with weak activity were filtered out by only considering drug-target activity stronger than 1μM and a pChEMBL_value$\geq$6. The pChEMBL value is the measure of the half-maximal potency/affinity on a negative logarithmic scale. The

Anatomical Therapeutic Code (ATC-code) was used to select drugs that are protein kinase inhibitors. The ATC code classifies drugs into different groups at different levels. The code ”L01XE” corresponds to antineoplastic drugs which are protein kinase inhibitors.

4.3.4 Network Data and Analysis

Human protein association network was obtained from the STRING database version 11 [76]. The STRING database was chosen as it is widely used and frequently updated. It provides scores indicating the reliability of an interaction, benchmarked against common sets of true positive associations. The protein interaction data was filtered by applying a cut-off of 0.8 on the combined score of the interaction which, corresponds to protein-protein interactions (PPI) with high reliability. This gave 365,045 physical interactions between 14,711 proteins. The largest connected subgraph was extracted and node centrality measures were computed by measuring the betweenness centrality (B_C) as shown in equation 4.1 below.

$$B_C(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{4.1}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s, t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node, v, other than s, t. If s = t, $\sigma(s,t) = 1$, and if v $\in s,t$, $\sigma(s,t|v) = 0$

Measuring dispersion of Kinase-FunFam relatives in a Protein Network

The similarity measure adapted from Menche *et al* [5] called the ”DS-score” was used to measure the dispersion of Kinase-FunFam relatives on the human protein network. The DS-score measures the mean distance of separation of genes on the network, representing the diameter of the targets for a drug in an interactome. These results were compared against random protein sets.

Calculating the side effects associated with the Kinase-FunFams

Druggable Kinase-FunFams were identified using the same enrichment protocol previously described in chapter two. Side effect data was obtained from SIDER [121], a known resource containing side effects extracted from drug labels via text

mining and mapped to drug IDs from ChEMBL database. To avoid drugs whose side effects are not well characterised, only those drugs that have at least five side effects in SIDER was considered as suggested by [119, 233]. The network properties of the relatives of druggable Kinase-FunFams were measured in a comprehensive human protein network and druggable Kinase-FunFams with low likelihood of being associated with side effects (i.e. relatives were not highly dispersed in protein network) were identified.

4.4 Result and discussion

4.4.1 Compiling the CATH-Gene3D kinase sequence dataset

291,200 kinase sequences have been classified into two separate domain families in CATH-Gene3D. The N-domain superfamily (CATH superfamily 3.30.200.20) in which relatives have an average of 90 amino acid residues and the C-domain superfamily (CATH superfamily 1.10.510.10) in which relatives have an average of 140 amino acid residues (see figure 4.18).

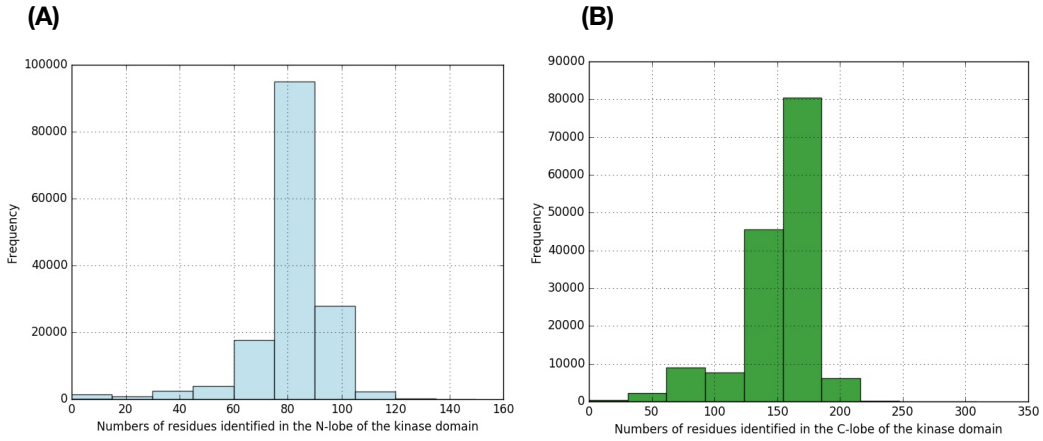

Figure 4.18: Distribution of the numbers of residues in the N- and- C kinase domains in CATH-Gene3D

The two domains were concatenated to represent the entire kinase functional unit. On average, the kinase functional unit had $\sim$ 250 residues, see Figure 4.19.

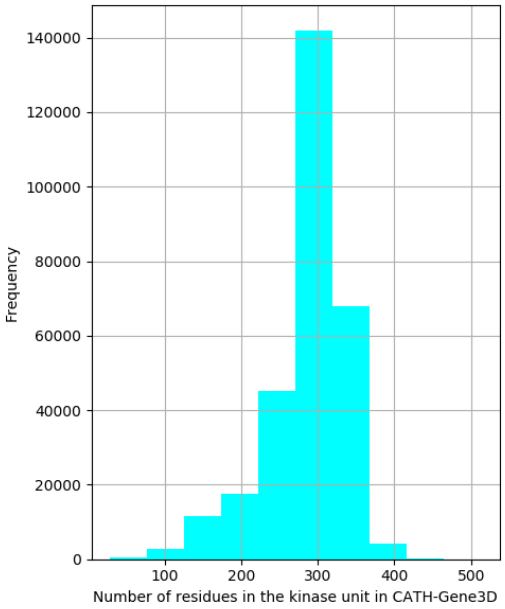

Figure 4.19: Distribution of residues in the kinase unit

4.4.2 Classification of Kinase-FunFams

The 291,200 kinase sequences were clustered into families using a novel method 'GARDENER', that can handle large superfamilies (i.e. >100,000 sequences) developed by Nicolas Bordin in the Orengo group. Kinase sequences were first partitioned according to their multidomain architecture (MDA). 187 MDAs were retrieved. As also reported by Srinivasan and colleague [199], it was observed that most kinases are multidomain proteins and there is considerable domain shuffling observed in their architectures. It is however worth noting that the majority of the kinase sequences (53%) belong to the canonical single domain N-C architecture (3.30.200.20-1.10.510.10), represented in the Table 4.1 as Kinase-MDA-1.

Table 4.1: The top 10 most populated multidomain architectures (MDAs) identified for Kinase-FunFam in the CATH-Gene3D version 16 sequences. 'N' represents the N-domain (3.30.200.20), 'C' represents the C-domain (1.10.510.10) while X is any other accessory domain associated with the kinase domain.

MDA-groups	Numbers of sequences	% of sequences	MDA arrangement
Kinase-MDA-1	154,501	53.1	N-C
Kinase-MDA-2	6,605	2.2	N-C-C
Kinase-MDA-3	6,185	2.1	X-N-C
Kinase-MDA-4	4,498	1.5	X-X-N-C
Kinase-MDA-5	4,152	1.4	X-N-C
Kinase-MDA-6	3,320	1.1	N-C-N-C
Kinase-MDA-7	3,310	1.1	N-C-C-N
Kinase-MDA-8	3,201	1.1	X-X-N-C
Kinase-MDA-9	2,754	0.9	X-X-X-N-C
Kinase-MDA-10	2,595	0.9	X-N-C

Kinase sequences were clustered using CD-HIT at 90% sequence similarity resulting in 11,959 (s90) starting clusters across all MDAs. Clusters without experimental GO annotation were discarded.

Following the GARDENER protocol, 1955 Kinase-FunFams were generated. Several measures were used to analyse the Kinase-FunFams. The Diversity of Position (DOP) score calculates the information content of the multiple sequence alignment (MSA). A DOP score above 70 is considered to be a good indicator of high diversity in the sequences [85]. Highly conserved residues in FunFams with a high DOPs are typically associated with stability or function. About 80% of the Kinase-FunFams have a high DOP score ($\geq$ 70).

4.4.3 Assessing the quality of Kinase-FunFam classification using Enzyme Numbers

1377 Kinase-FunFams having more than one relative (i.e. 70% of the Kinase-FunFams) were assessed for EC-purity. Only experimental EC-terms were used for the assessment. The similarity in Enzyme Classification (EC) number was considered at both the 3-digit and 4-digit EC-level. The EC-terms of sequences within Kinase-FunFams were obtained from UniProtKb, and the numbers of unique EC-terms per family was analysed. 26% of Kinase-FunFams have EC-annotations for

relatives within them comprising a total of 12,894 sequences. This set was analysed to gauge the quality of the functional purity of the Kinase-FunFams classification.

Figure 4.20 below shows the percentage of relatives having the most common EC-term of the Kinase-FunFam. It can be seen that the percentages ranges from 50-100% with a large proportion (70.5%) of the Kinase-FunFams having a single EC-annotation.

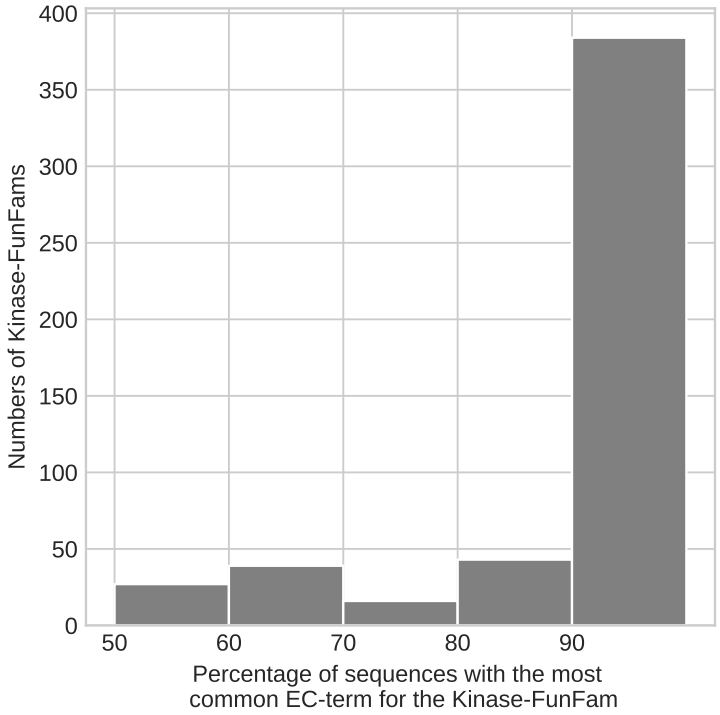

Figure 4.20: Percentage of sequences in Kinase-FunFams with the most common EC terms

Examining the Kinase-FunFam classification at the EC3 levels shows four different EC3 terms (”2.7.11”, ”2.7.10”, ”2.7.12”, ”4.6.1”) (See figure 4.21) representing serine/threonine protein kinases, tyrosine protein kinases, dual-specificity protein kinases and guanylate cyclase kinases respectively. There are also three Kinase-FunFams with relatives having guanylate cyclase activity. These kinases are known to use GTP as the phosphate donor rather than ATP, as observed in majority of the other kinases and they are involved in the cyclisation of GTP to cyclic GMP.

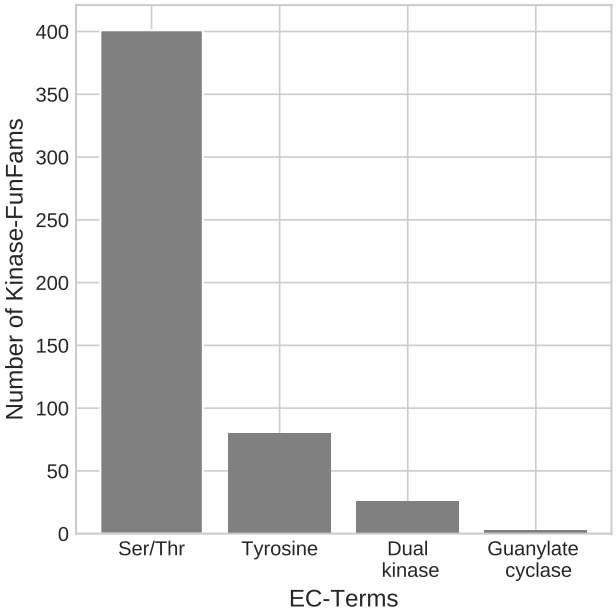

Figure 4.21: Numbers of Kinase-FunFams assigned to Ser/Thr, Tyr, Dual-specificity kinases and Guanylate cyclase respectively.

Figure 4.22 shows that there are 19 different EC4 terms amongst the Kinase-FunFams with the majority (80%) of the Kinase-FunFams having EC-term ”2.7.11.1”-the non specific serine/threonine protein kinases.

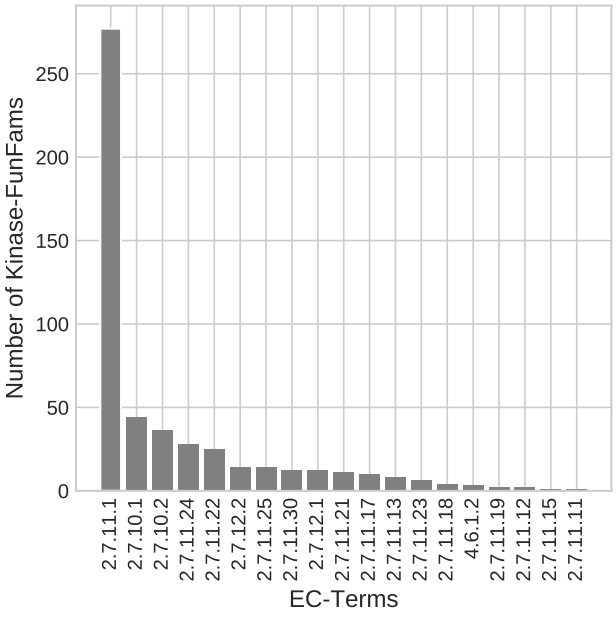

Figure 4.22: Different EC-terms (level 4) found in the Kinase-FunFams.

The 12,894 sequences in Kinase-FunFams with EC annotations were dis-

tributed across 510 Kinase-FunFams. The overall EC-purity Kinase-FunFams at both EC level 4 and level 3 is shown in figure 4.23 below.

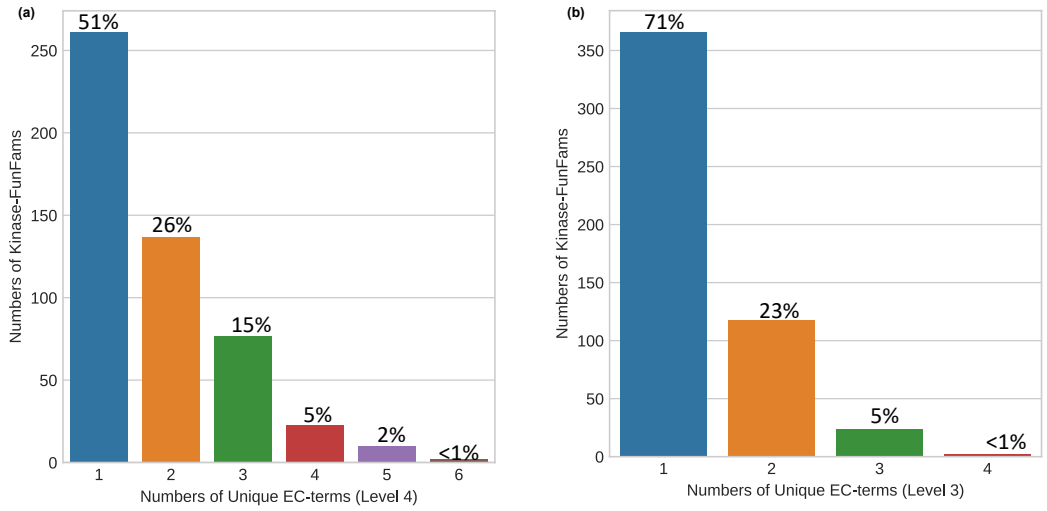

Figure 4.23: Distribution of the numbers of families having one or more EC (level 4) and EC (level 3) in Kinase-FunFams.

Based on the EC-annotations, it can be seen that more than three quarter of the Kinase-FunFams are reasonably pure (<3 EC terms) whilst relatively few have more than 3 different EC terms within them.

4.4.4 Kinase-FunFams compared with KinBase classification of kinase sequences

The quality of the functional subclassification was also assessed by comparing the EC-purity in the Kinase-FunFams to one of the most cited kinase classification resources, KinBase. KinBase groups the kinase domains into 9 groups comprising 8 typical kinases (AGC, CAMK, CGMC, STE, TK, TKL, RGC, Others) and the Atypical kinases [195]. The groups are associated with broad substrate specificity. KinBase then splits each group into families and then subfamilies. Subfamilies comprise relatives with higher functional similarity and are generally specific for each phylum.

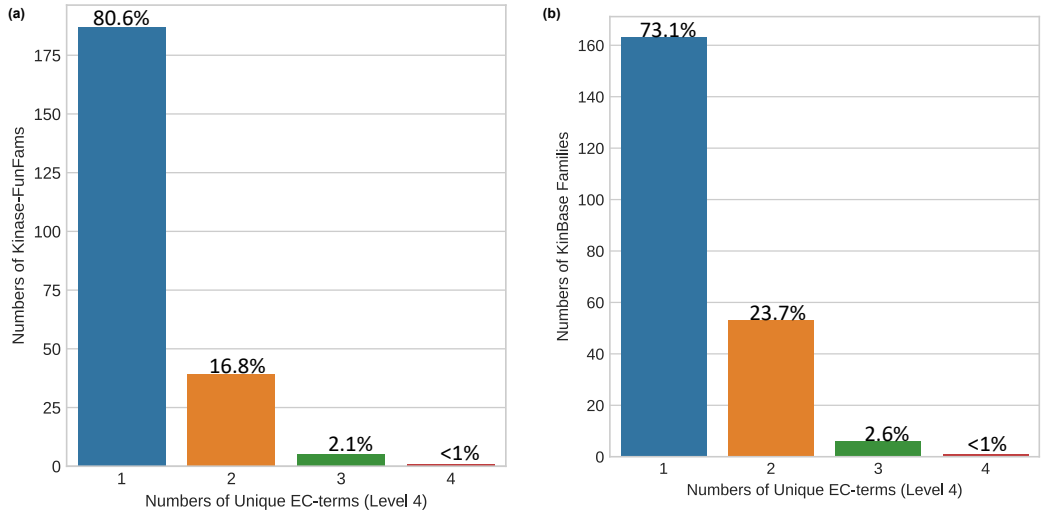

Figure 4.24: Comparison of the numbers of families having a particular EC level 4 purity (a) Kinase-FunFams classification and (b) KinBase classification. The percentage of sequences in each category of purity are shown above the column.

The 7379 sequences in KinBase had 6591 sequences with EC-annotations and these were distributed in 223 KinBase subfamilies and 232 Kinase-FunFams. The functional EC-purity of Kinase-FunFams was compared with KinBase subfamilies at both EC3 and EC4 levels. The analysis revealed similarity in the purity levels. Kinase-FunFams however showed a higher EC-purity, having the majority (80.3%) of the families with one EC4 term compared to 73.1% for the KinBase classification.

At level 3, both Kinase-FunFams and KinBase classifications had the majority of families classified with one EC3-annotation; 97.5% of relatives in Kinase-FunFams with one EC3-annotation compared to 91.9% in KinBase classification (See figure 4.25). Kinase-FunFams comprise tenfold more sequences than KinBase i.e. a total of 69,228 sequences compared to 7,379 for KinBase.

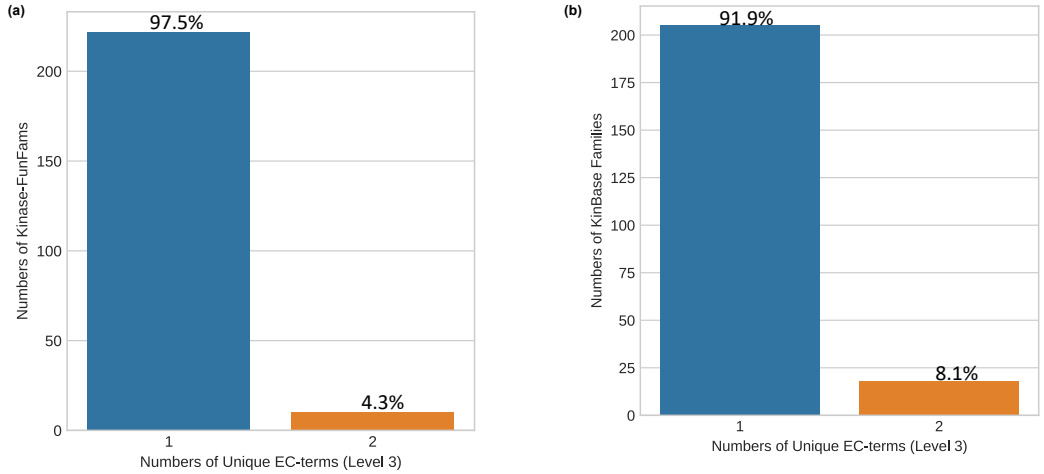

Figure 4.25: Comparison of the numbers of families having a particular EC level 3 purity (a) Kinase-FunFams classification and (b) KinBase classification

4.4.5 Mapping protein kinase inhibitors set to Kinase-FunFams

Extracting GlaxoSmithKline Protein Kinase Inhibitor Sets (GSK-PKIS) from ChEMBL at 50% activity level gave a set of 205 protein kinase inhibitors (PKIs) which could be mapped to 133 protein kinases, out of which 116 protein kinases were found in 62 Kinase-FunFams. This covers about 60% of the entire PKIS set and is thus a reasonable dataset to consider for a network assessment and characterization of side effects. Extracting the FDA-approved drug set using p-ChEMBL activity level ≥ 6 gave 29 approved drugs that interact with 324 targets including kinases and non-kinases. The targets were filtered to exclude those that are not kinases, reducing the numbers of kinase-targets to 305 kinases, out of which 250 were found in 129 Kinase-FunFams.

Comparison of the network properties for the FDA drug-target set and the GSK-PKI drug-target set

The FDA approved set and the GSK-PKIs have different characteristics. GSK-PKIS are considered as experimental drugs as they have not yet been approved for clinical trials, but have shown inhibitory activities against kinase panels from experimental studies and as such could be used to probe molecules for kinases in the untargeted kinome. On the other hand, FDA approved drugs have been clinically tested and approved for use in different diseases. The network properties of both sets of targets

were analysed using the DS-score. The DS-score measures the distances between the targets of these drug sets within a human protein network.

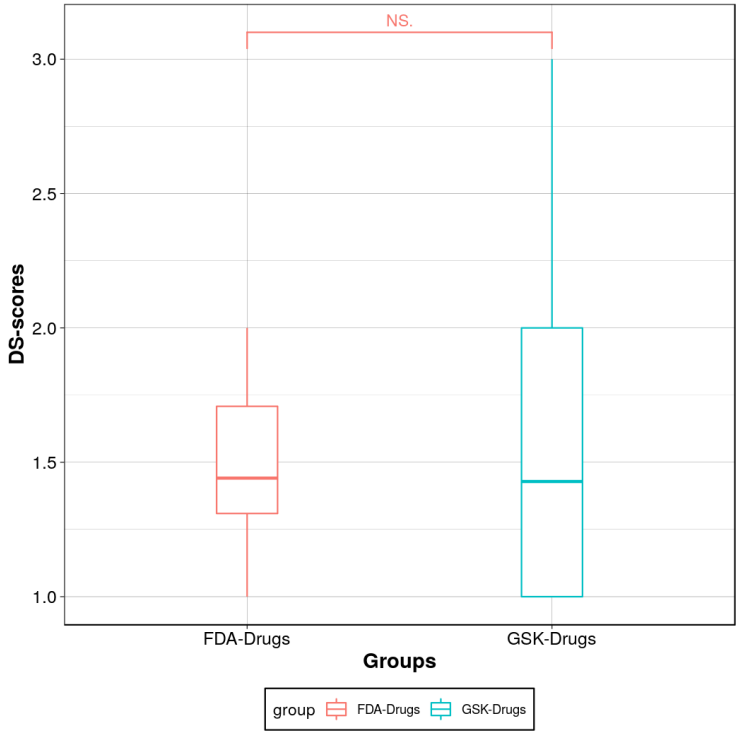

Figure 4.26: Comparison of the network proximity of targets in the FDA and GSK target sets in a human protein network

There was no statistically significant difference observed between the average DS-score of the FDA-targets and the GSK-PKIS targets (Mann-Whitney test, P-value = 0.07789). This implies that targets from both drug sets (FDA and GSK-PKIS) form similar network communities (See figure 4.26).

4.4.6 Identifying druggable Kinase-FunFams

The FDA approved dataset (29 drugs and 250 targets) and the GSK-PKI set (205 drugs and 116 targets) were combined to analyse the promiscuity of the drugs and detect druggable Kinase-FunFam with low side effects. This gave a total of 234 drugs and 270 target kinases, since some kinases are targeted by both FDA and GSK PKIs. The combined dataset comprises multi-target inhibitors with varying degrees of interaction with the kinases. It can be seen from figure 4.27 below that many kinase inhibitors are associated with more than one target.

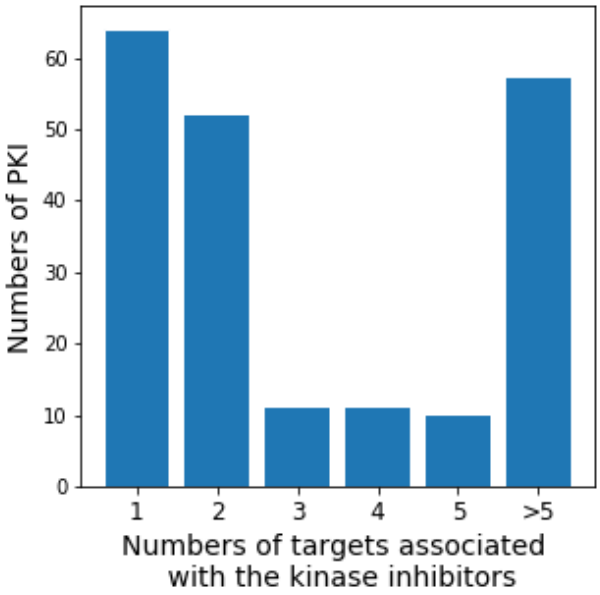

Figure 4.27: The numbers of kinase inhibitors and the numbers of targets they are associated with.

The 270 target kinases are distributed across 135 Kinase-FunFams. There are between 1-13 targeted kinases within each Kinase-FunFam (See figure 4.28 below). 60 of the 135 Kinase-FunFams were found to be statistically overrepresented with drug targets and hence identified as druggable Kinase-FunFams.

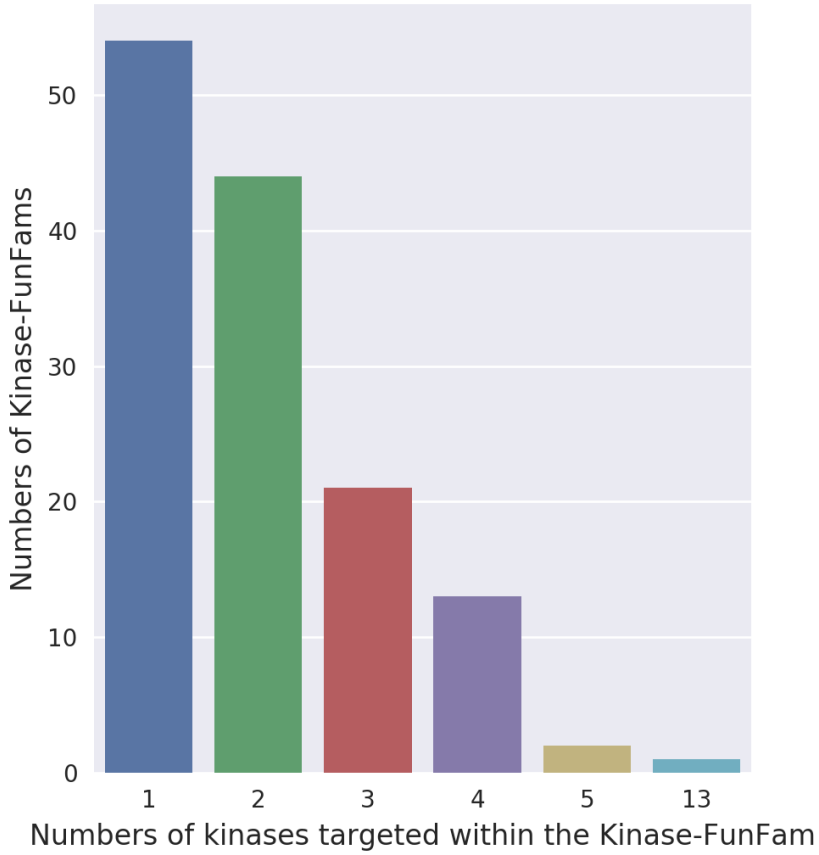

Figure 4.28: The number of kinases in Kinase-FunFams currently targeted by kinase inhibitors. It can be seen from figure 4.28 that 54 Kinase-FunFams have 1 targeted kinase, while 1 Kinase-FunFam has 13 targeted kinases.

The relatives within the Kinase-FunFam with 13 drug targets were further analysed to reveal the diseases associated with the targets. This Kinase-FunFam is a receptor tyrosine kinase (Ephrin receptor) with relatives involved in various diseases including colorectal cancer, prostate cancer, cataract and lymphatic malformation.

4.4.7 Dispersion of the Kinase-FunFams in the human protein interaction network

The dispersion of the relatives of druggable Kinase-FunFams in the human protein interaction network, was assessed to determine their propensity to be associated with side effects. Figure 4.29 shows the mean dispersion score (DS-score) for the

druggable Kinase-FunFams compared to random Kinase-FunFams

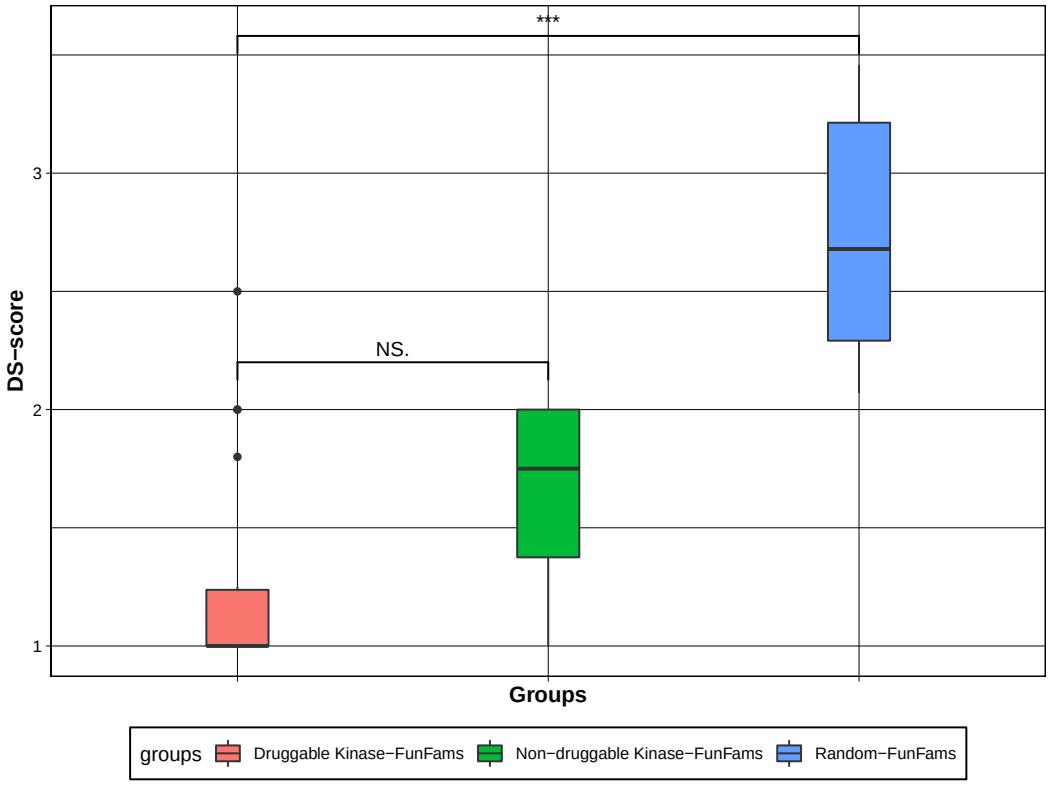

Figure 4.29: Boxplot comparing the DS-score of druggable Kinase-FunFams, other Kinase-FunFams and random-FunFams in a human protein interaction network.

There is a statistically significant difference (MannWhitney test, P-value=2.25×10^{-7}) between the average DS-score of druggable Kinase-FunFams compared to random-FunFams. Relatives of other Kinase-FunFams not enriched with drug targets were also found to be closely connected in the human interaction network.

Further analysis was carried out to associate side effects with the various Kinase-FunFams. This study was based on earlier observation that druggable FunFams whose relatives are well connected in the network are less likely to be associated with side effects. 16 FDA-approved protein kinase inhibitors, found to have at least 5 side effects reported in the SIDER database, were used in this analysis. This filtering criteria has been previously used [119, 180] and is done to avoid side effects that are not well characterised. The number of side effects associated

with druggable Kinase-FunFams ranges from 6-24 as extracted from SIDER. Druggable Kinase-FunFams with less than 2 human proteins in the protein network were discarded. Network characteristics of the relatives in these Kinase-FunFams were analysed to determine the propensity of these Kinase-FunFams to be associated with side effects (see Table 4.2.

Previous analyses reported in chapter 2 have shown a range of DS-scores from 1-3.5 across all druggable FunFams. The analysis revealed that relatives within Kinase-FunFams tend to be close in the human protein network. As in the previous analysis, many of the new Kinase-FunFams have a DS-score of 1. This result supports other studies that show that kinases form more connected interactions than non-kinase proteins in a protein-protein interaction network [234]. However, 16 of 29 FDA approved protein kinase inhibitors are reported in the SIDER database to have between 5-24 side effects per drug.

Further network analysis were carried out to determine possible causes of side effects. The analysis revealed that on a global network scale, the relatives of Kinase-FunFams are hubs and bottlenecks. Kinase targets were found more likely to be hubs (24%) and bottlenecks (32%) than random proteins in the network (MannWhitney test, P-value = 2.85×10^{-14} and 8.92×10^{-11} for bottlenecks and hubs). The association of drug side effects with some Kinase-FunFams may, therefore, be due to kinase inhibitors being associated with kinases that are central in the human protein interaction network. This work supports other studies that showed that side effects are associated with proteins in network with high degree and high betweenness centrality [126, 235, 236].

Another possible reason for the side effects associated with protein kinases might be the high structural similarity of relatives across the whole superfamily. Analysis of the SSAP scores of the individual kinase domains (1.10.510.10 and 3.30.200.20 CATH-superfamily) show that 87.4% of the relatives of the C-lobe superfamily have SSAP scores $\geq$70 (out of 100), and 91.6% of the relatives of the N-terminal lobe domain have SSAP score $\geq$70. This compares with an average of only 68.8% of all the other CATH-superfamily with SSAP score above 70.

Table 4.2: Network properties of Kinase-FunFams.
PKI (Protein Kinase Inhibitors), KFF(Kinase-FunFams), Props (Proportions).

Kinase-FunFams	No. of human kinases	No. of PKIs (Median Side effect)	Promixity measures (DS-score)	No. of hubs (%)	No. of bottlenecks (%)
KFF-233	6	6 (5)	1.8	0	2(33%)
MDA-6-KFF_147	4	5 (24)	1	0	1(25%)
MDA-1-KFF-4310	4	11 (9)	2	0	0
KFF-255	4	5 (24)	1	0	1(25%)
KFF-782	4	11 (15)	1	4 (100%)	4(67%)
KFF-223	6	10	1.17	3 (50%)	5(83%)
KFF-193	7	13 (15)	1	0	3(43%)
KFF-390	5	39 (16)	1.2	4(80%)	5(100%)
KFF-133	4	20 (15)	1	4 (100%)	3(75%)
MDA-29-KFF-49	3	5	1	3(100%)	3(100%)
KFF-715	4	16 (10)	2.5	1 (25%)	0
KFF-241	3	8 (15)	1.33	1(33%)	1(33%)
KFF-269	3	5 (4)	1	2(67%)	2(67%)
KFF-172	14	64 (15)	1	1(7%)	1(7%)
KFF-115	4	3 (24)	1	3(75%)	3(75%)
KFF-260	4	12 (16)	1.25	0	1(25%)
KFF-62	4	20 (6)	1	2 (50%)	3(75%)
KFF-49	3	11 (11)	2	0	0
KFF-856	4	11 (15)	1	4(100%)	4(100%)
KFF-203	4	1	1	2(50%)	2(50%)

4.5 Chapter summary

A novel classification of kinases was generated based on the in-house protein domain classification method 'GARDENER'. 1955 Kinase-FunFams were identified comprising 69,228 sequences. Due to time constraints, the final stage of GARDENER could not be implemented. This involves scanning all sequences in the CATH kinase superfamily against the Kinase-FunFams HMM to expand the numbers of sequences in the Kinase-FunFams. This is expected to increase the numbers of sequences in Kinase-FunFams to about 150,000 sequences or more.

Our classification method comprises ten fold more sequences compared to the widely used KinBase approach that contains 7400 sequences, divided into 579 subfamilies. The EC-purity, as observed by the proportion of EC4 and EC3 annotation, was observed to be higher in the Kinase-FunFams (80.6%) compared to the KinBase classification (73.1%). Although the EC-purity of the Kinase-FunFams with single EC4 annotation reduced to 51% when all the sequences of within the Kinase-FunFams was analysed which suggest further optimisation of the splitting of the trees generated from 'GARDENER' algorithm to hopefully improve the purity of Kinase-FunFams.

234 protein kinase inhibitors could be mapped allowing the identification of 60 druggable Kinase-FunFams. Network analysis revealed that relatives within Kinase-FunFams tend to be more closely associated together than relatives from random FunFams. On a global network scale, the relatives of the Kinase-FunFams are shown to be mainly hubs and bottlenecks in human protein interaction network, a hallmark of side effect proteins as the targeting of such proteins disrupts several biological pathways.

Further studies will involve repurposing approved protein kinase inhibitors to those relatives within Kinase-FunFams without drugs but are involved in known human diseases. This analysis of the kinase-family will hopefully help in obtaining kinase-FunFams that not yet associated with drugs and can also help in revealing pseudokinases.

Chapter 5

Conclusion

This thesis presented computational strategies to identify novel drug targets. The study leverages the availability of protein domain information, and protein interaction data, to detect putative new targets associated with low side effects. Using domain information from the CATH-database, methods were developed to identify druggable domain families, which were then characterised for their side effect propensity using network approaches. In particular, the drug target neighbourhood, as described by the closeness of targets in a given human network was used in this study to assess the likelihood of drug side effects.

In chapter 2, 81 druggable domain functional families (druggable CATH-FunFams) were identified based on the statistical overrepresentation of drug targets in those domain families. Following this, several network based analysis were performed using a well curated, comprehensive human protein network, including measuring the betweenness centrality, degree and network aggregation scores such as the ”DS-score” and the ”Matrix similarity score”. The analyses revealed that drug targets as well as other relatives in druggable FunFams exhibit similar network properties as they tend to be centralised and less dispersed in the protein network when compared to relatives from randomly selected FunFams. It was found that druggable CATH-FunFams whose relatives were dispersed in protein interaction networks tend to be more associated with side effects when compared to those

druggable CATH-FunFams with relatives closely connected in the network.

In chapter 3, bladder cancer was chosen as a disease prototype for the detection of novel drug targets and to explore the repurposing of approved drugs. As with other types of cancer, bladder cancer is a genetically heterogeneous disease with mutation events that lead to uncontrollable growth and development. Putative bladder driver genes were identified using protein network based strategies that detected putative drivers based on their proximity to a set of known drivers and predicted bladder cancer drivers predicted by family based analysis. This revealed 323 putative bladder cancer associated proteins that are enriched in hallmark signatures, pathways and GO terms associated with bladder cancer processes including; cell cycle related processes, activation of invasion and metastasis, as well as steroid hormone processes. Drug mapping revealed that 28 of the 323 putative bladder cancer proteins are associated with FDA approved drugs from ChEMBL database that are currently being used to treat other diseases. This set was extended to 35 targets which could be associated with drugs by using the druggable CATH-FunFams to repurpose approved drugs from relatives in the FunFams.

Chapter 4 presented the generation of a comprehensive classification of kinases using a novel in-house protein domain classification method called ”GARDENER”. This set of Kinase-FunFams was benchmarked against another known and widely used kinase classification, KinBase which has 579 subfamilies generated from 7400 sequences. GARDENER gave 1955 Kinase-FunFams comprising 69,200 sequences from UniProt. Druggable Kinase-FunFams were identified based on overrepresentation of drug targets within the families, and the network dispersion of relatives in the Kinase-FunFams was measured to assess the association with drug side effects. This allowed identification of some novel kinase targets associated with low side effects.

The work performed in this thesis can be extended in several direction. For example, the current implementation has only considered drugs from the ChEMBL database as it is comprehensive and provides annotation for drug-target mapping, with various associated scores such as the published activity and the pChEMBL

value. However, other resources such as DrugBank, PharmGKB, Therapeutic Target Database should be considered and aggregated to give a broader perspective of drug/target interactions.

The identification of druggable families in this study considered only domain families that are statistically overrepresented with drug targets. This approach might have led to omission of some potentially druggable families as it will preferentially identify families whose relatives are currently frequently targeted. One approach that might be used in overcoming this limitation is the use of machine learning to extract possible druggable features of targets and drugs, and using such information to distinguish druggable from non-druggable domain families and thereby predict families whose relatives share these druggable features.

Furthermore, the side effect analysis in this study did not focus on the dosage of the drug compound analysed. This might give additional information as side effects of drugs are sometimes dose dependent. Overall, the predicted novel drug targets from this study could be tested. One such test might be to carry out a molecular dynamic simulation of the binding of repurposed drugs to predicted targets to find out the energetics and binding properties of the drugs-to-ligand. Recently, efforts have shown the value of this techniques being used to study the dynamic interaction of drugs with proteins and characterisation of the structural changes that might be induced by such binding [237, 238]. Also, in the case of bladder cancer, bladder cancer cell lines can be used to test the effect of repurposed drugs. Binding assays can be carried out and subsequently other more time consuming and expensive experimental protocols such as X-ray crystallography or Nuclear Magnetic Resonance (NMR) can be performed to help assess the binding of repurposed drugs.

In conclusion, several possible targets have been identified in this research. These targets are hypothetical because they have been identified purely using computational approaches. Furthermore, pharmaceutical companies may not have considered these areas of repurposing as there may be limited available data on clinical validity. Thus, to take this research forward, experimental analysis is needed to validate the predicted targets.

Appendix A

Table A.1: Drug to CATH-FunFam mapping. The names and codes of association between drugs and CATH-FunFams. Druggable Genome class represented by the type of protein family categorised based on the druggable genome by Hopkins and Groom. Number of Side Effects (SE) is the number of preferred MedDRA terms associated with the drug in the SIDER database. Similarity score is the measured mean STRING combined score of the relatives of the CATH-FunFam. Probability of SE free is the probability that the CATH-FunFam do not contain a relative associated with side effects according to the computed logistic regression model.

Druggable FunFams	**CATH-FunFam Name**	**Druggable Genome**	**Similarity Score**	**Drugs**	**Number of SE**	**Prob of SE free**
1.10.1300.10.FF1262	cAMP-specific 3,5-cyclic phosphodiesterase 4D	Others	0.24	MILRINONE	20	0.401
1.10.510.10.FF78745	Receptor protein-tyrosine kinase	Protein Kinase	0.19	SORAFENIB	165	0.381
1.10.510.10.FF78758	Receptor protein-tyrosine kinase	Protein Kinase	0.22	ERLOTINIB	134	0.393
1.10.510.10.FF78946	Calcium/calmodulin-dependent protein kinase kinase, putative	Protein Kinase	0.15	SUNITINIB	264	0.365
1.10.510.10.FF79093	Mitogen-activated protein kinase kinase kinase	Protein Kinase	0.16	SUNITINIB	264	0.369
1.10.510.10.FF79298	Puntreceptor serine/threonine kinase	Protein Kinase	0.56	DASATINIB	236	0.534
1.10.565.10.FF5008	Peroxisome proliferator-activated receptor gamma	Nuclear Receptor	0.66	ROSIGLITAZONE	69	0.576
1.10.565.10.FF5028	Retinoic acid receptor RXR-beta	Nuclear Receptor	0.5	BEXAROTENE	257	0.509

Table A.1 continued from previous page

Druggable FunFams	CATH-FunFam Name	Druggable Genome	Similarity Score	Drugs	Number of SE	Prob of SE free
1.10.565.10.FF5060	Nuclear hormone receptor E75 putative	Nuclear Receptor	0.93	TRETINOIN	50	0.682
1.10.565.10.FF5076	Retinoic acid receptor alpha	Nuclear Receptor	0.71	TAZAROTENE	39	0.596
1.10.565.10.FF5096	Estrogen receptor beta	Nuclear Receptor	0.77	PREDNISOLONE	117	0.620
1.10.630.10.FF29314	Cytochrome P450 monooxygenase	Cytochrome p450	0.38	RITONAVIR	1	0.459
1.10.630.10.FF29337	Cytochrome P450 monooxygenase	Cytochrome p450	0.36	MICONAZOLE	60	0.450
1.10.640.10.FF1593	Prostaglandin G/H synthase 1	Enzymes	0.59	DICLOFENAC	1	0.547
1.20.1070.10.FF45451	Muscarinic acetylcholine receptor M3	GPCR	0.36	MITOXANTRONE	146	0.450
1.20.1070.10.FF45452	Somatostatin receptor type 1	Others	0.61	NALTREXONE	209	0.555
1.20.1070.10.FF45459	Beta-2adrenergic receptor	GPCR	0.36	CLONIDINE	149	0.450
1.20.1070.10.FF45629	Vasopressin V1a receptor	GPCR	0.9	VASOPRESSIN	21	0.670
1.20.1070.10.FF45672	Adenosine receptor A1	GPCR	0.5	SILDENAFIL	280	0.509

Table A.1 continued from previous page

Druggable FunFams	CATH-FunFam Name	Druggable Genome	Similarity Score	Drugs	Number of SE	Prob of SE free
1.20.1070.10.FF45687	G-protein coupled receptor	GPCR	0.43	RAMELTEON	19	0.480
1.20.1250.20.FF179126	Solute carrier family 22 member 2	Transporters	0	CHLORHEXIDINE	50	0.309
1.20.1250.20.FF180577	Solutecarrier family 22 member 6	Others	0	CHLORHEXIDINE	50	0.309
1.20.1560.10.FF26651	Multidrug ABC transporter ATP-binding protein	Transporters	0	CYCLOSPORINE	172	0.309
1.20.58.390.FF2640	Gamma-aminobutyric acid receptor subunit beta	Ion channel	0.66	LINDANE	10	0.576
2.10.50.10.FF7726	Receptor protein-tyrosine kinase	Protein Kinase	0.23	DASATINIB	236	0.397
2.40.100.10.FF7135	Peptidyl-prolyl cis-trans isomerase	Others	0.07	CYCLOSPORINE	172	0.334
2.60.120.260.FF35252	Receptor protein-tyrosine kinase	Protein Kinase	0.23	NILOTINIB	312	0.397
2.60.40.10.FF135943	Receptor protein-tyrosine kinase	Protein Kinase	0.21	NILOTINIB	312	0.389
3.10.200.10.FF1430	Carbonic anhydrase 2	Others	0.22	SULFANILAMIDE	2	0.393
3.10.50.40.FF13974	Peptidyl-prolyl cis-trans isomerase	Enzymes	0.9	TACROLIMUS	421	0.670
3.20.20.70.FF120512	GMPreductase	Others	0.82	MYCOPHENOLIC ACID	194	0.640

Table A.1 continued from previous page

Druggable FunFams	CATH-FunFam Name	Druggable Genome	Similarity Score	Drugs	Number of SE	Prob of SE free
3.30.50.10.FF233	Retinoicacid receptor, gamma	Nuclear Receptor	0.77	BEXAROTENE	257	0.620
3.30.50.10.FF3991	RAR-related orphan nuclear receptor variant 2	Nuclear Receptor	0.91	TRETINOIN	50	0.674
3.30.50.10.FF4110	Glucocorticoid receptor 2	Nuclear Receptor	0.63	DEXAMETHASONE	1	0.563
3.30.50.10.FF4122	Peroxisome proliferator-activated receptor gamma	Nuclear Receptor	0.74	ROSIGLITAZONE	69	0.608
3.30.50.10.FF4124	Estrogen receptor beta 1	Nuclear Receptor	0.93	ESTRADIOL	336	0.682
3.30.50.10.FF4132	Nuclear receptor related protein	Nuclear Receptor	0.72	MIFEPRISTONE	73	0.600
3.30.50.10.FF4220	Nuclear receptor related protein	Nuclear Receptor	0.83	FULVESTRANT	77	0.644
3.30.505.10.FF4228	Non-specific protein-tyrosine kinase	Protein Kinase	0.91	DASATINIB	236	0.674
3.40.50.1820.FF115121	Carboxylesterase 3	Others	0.18	RIVASTIGMINE	2	0.377

Table A.1 continued from previous page

Druggable FunFams	CATH-FunFam Name	Druggable Genome	Similarity Score	Drugs	Number of SE	Prob of SE free
3.40.50.1820.FF115213	Sn1-specific diacyl-glycerol lipase beta	Enzymes	0.17	ORLISTAT	133	0.373
3.40.50.300.FF627204	ABC transporter, transmembrane region	Transporters	0.27	CYCLOSPORINE	172	0.413
3.40.800.20.FF2844	Histone deacetylase 3	Others	0.94	VORINOSTAT	57	0.685
3.40.800.20.FF2855	Histone deacetylase 6	Others	0.87	VORINOSTAT	57	0.659
3.60.20.10.FF15813	Proteasome subunit beta type	Others	0.99	BORTEZOMIB	292	0.703
4.10.1140.10.FF38	Receptor protein-tyrosine kinase	Protein Kinase	1	LAPATINIB	68	0.707

Table A.2: Modules generated using MCODE clustering algorithm and analysis of genes found in each module.

Modules	Genes in Modules	Genes in SET3	Drug targets	Hi-DEG	Cancer-Genes	Mutation in BLCA	Summarised GO-terms	Enriched KEGG-Pathways
1	159	6	ACTB	0	11	0	Translational initiation	hsa04066
2	214	10	HDAC2, HDAC1	1	14	0		
3	85	10	EP300, ESR1, HDAC3	3	8	0	Endoplasmic reticulum unfolded protein response	hsa04919
4	294	26	HDAC5	2	13	0	DNA replication-dependent nucleosome assembly	hsa05322
5	73	13	PARP1	1	5	0	Positive regulation of transcription initiation from RNA polymerase II promoter	hsa03050
6	140	9	TOP1	3	11	0	Mitochondrial ATP synthesis coupled electron transport	hsa00190

Table A.2 continued from previous page

Modules	Genes in Modules	Genes in SET3	Drug targets	Hi-DEG	Cancer-Genes	Mutation in BLCA	Summarised GO-terms	Enriched KEGG-Pathways
7	125	15	TOP2B	0	12	0	Positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway	hsa04110
8	69	15	PIK3C2A, VDR	2	5	0	ATP-dependent chromatin remodeling	hsa05322
9	152	21	AR	0	11	0	Ribonucleoprotein complex export from nucleus	
10	26	5	CDK8	0	3	0		
11	98	7	HIF1A	3	8	1	Histone modification	
12	129	5	RXRA	2	9	1	Covalent chromatin modification	hsa05215
13	32	2	NR3C1	0	6	0		

Table A.2 continued from previous page

Modules	Genes in Modules	Genes in SET3	Drug targets	Hi-DEG	Cancer-Genes	Mutation in BLCA	Summarised GO-terms	Enriched KEGG-Pathways
14	65	8	STAT3	0	9	0	Protein folding in endoplasmic reticulum	hsa04141
15	3	1	THRA	0	0	0	Regulation of lipid metabolic process	
16	106	6	PPARG, ERBB3, CDK9	1	13	1	Antigen processing and presentation of peptide antigen	hsa04150
17	123	1	ERBB2	0	13	0	Positive regulation of cellular protein catabolic process	hsa04210
18	51	4	RXRG, NR1H2, PIK3CA, RARA	0	8	0	Negative regulation of protein exit from endoplasmic reticulum	hsa04370

References

[1] Marc HV Van Regenmortel. Reductionism and complexity in molecular biology: Scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. *EMBO reports*, 5(11):1016–1020, 2004.

[2] Mika Gustafsson, Colm E Nestor, Huan Zhang, Albert-László Barabási, Sergio Baranzini, Sören Brunak, Kian Fan Chung, Howard J Federoff, Anne-Claude Gavin, Richard R Meehan, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome medicine*, 6(10):82, 2014.

[3] Qiaosheng Zhang, Jie Li, Hanqing Xue, Leilei Kong, and Yadong Wang. Network-based methods for identifying critical pathways of complex diseases: a survey. *Molecular BioSystems*, 12(4):1082–1089, 2016.

[4] Wei Liu, Aiping Wu, Matteo Pellegrini, and Xiaofan Wang. Integrative analysis of human protein, function and disease networks. *Scientific reports*, 5:14344, 2015.

[5] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[6] Chandra Sekhar Pedamallu and Linet Ozdamar. A review on protein-protein

interaction network databases. In *Modeling, Dynamics, Optimization and Bioeconomics I*, pages 511–519. Springer, 2014.

[7] James R Perkins, Ilhem Diboun, Benoit H Dessailly, Jon G Lees, and Christine Orengo. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243, 2010.

[8] Srinivasa Rao, K Srinivas, GN Sujini, and GN Kumar. Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, 2014, 2014.

[9] Albertha Walhout and Simon Boulton. Biochemistry and molecular biology. 2006.

[10] Erica Gerace and Danesh Moazed. Affinity purification of protein complexes using tap tags. In *Methods in enzymology*, volume 559, pages 37–52. Elsevier, 2015.

[11] John Nealon, Limcy Philomina, and Liam McGuffin. Predictive and experimental approaches for elucidating protein–protein interactions and quaternary structures. *International journal of molecular sciences*, 18(12):2623, 2017.

[12] Marina Serna. Hands on methods for high resolution cryo-electron microscopy structures of heterogeneous macromolecular complexes. *Frontiers in molecular biosciences*, 6:33, 2019.

[13] Maria Persico, Arnaud Ceol, Caius Gavrila, Robert Hoffmann, Arnaldo Florio, and Gianni Cesareni. *homomint*: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC bioinformatics*, 6(4):S21, 2005.

[14] Yanhui Hu, Ian Flockhart, Arunachalam Vinayagam, Clemens Bergwitz, Bonnie Berger, Norbert Perrimon, and Stephanie E Mohr. An integrative ap-

proach to ortholog prediction for disease-focused and other functional studies. *BMC bioinformatics*, 12(1):357, 2011.

[15] Alfonso Valencia and Florencio Pazos. Computational methods for the prediction of protein interactions. *Current opinion in structural biology*, 12(3):368–373, 2002.

[16] Tetsuya Sato, Yoshihiro Yamanishi, Minoru Kanehisa, Katsuhisa Horimoto, and Hiroyuki Toh. Improvement of the mirrortree method by extracting evolutionary information. *inSequence and Genome Analysis: Method and Applications*, pages 129–139, 2011.

[17] Raja Jothi and Teresa PRzYTYCkA. Computational approaches to predict protein–protein and domain–domain interactions. *Bioinformatics Algorithms: Techniques and Applications*, pages 465–492, 2008.

[18] Rohit Singh, Daniel Park, Jinbo Xu, Raghavendra Hosur, and Bonnie Berger. Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic acids research*, 38(suppl_2):W508–W515, 2010.

[19] Patrick Aloy and Robert B Russell. InterPreTS: Protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162, 2003.

[20] Javad Zahiri, Joseph Hannon Bozorgmehr, and Ali Masoudi-Nejad. Computational prediction of protein–protein interaction networks: algorithms and resources. *Current genomics*, 14(6):397–414, 2013.

[21] Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030, 2008.

[22] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl_1):i38–i46, 2005.

[23] Sanghamitra Bandyopadhyay and Koushik Mallick. A new feature vector based on gene ontology terms for protein-protein interaction prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(4):762–770, 2016.

[24] Kuan-Hsi Chen, Tsai-Feng Wang, and Yuh-Jyh Hu. Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. *BMC bioinformatics*, 20(1):308, 2019.

[25] Plamen Ch Ivanov, Kang KL Liu, and Ronny P Bartsch. Focus on the emerging new fields of network physiology and network medicine. *New journal of physics*, 18(10):100201, 2016.

[26] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[27] Vladimir Batagelj and Andrej Mrvar. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.

[28] Zhenjun Hu, Joseph Mellor, Jie Wu, and Charles DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. *BMC bioinformatics*, 5(1):17, 2004.

[29] Avi Ma'ayan, Sherry L Jenkins, Ryan L Webb, Seth I Berger, Sudarshan P Purushothaman, Noura S Abul-Husn, Jeremy M Posner, Tony Flores, and Ravi Iyengar. SNAVI: Desktop application for analysis and visualization of large-scale signaling networks. *BMC systems biology*, 3(1):10, 2009.

[30] Seth I Berger, Ravi Iyengar, and Avi Ma'ayan. Avis: Ajax viewer of interactive signaling networks. *Bioinformatics*, 23(20):2803–2805, 2007.

[31] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network

medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011.

[32] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101, 2004.

[33] Francesc Comellas. Complex networks: Deterministic models. *NATO security through science series D-Information and Communication Security*, 7:275, 2007.

[34] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[35] Hátylas Azevedo and Carlos Alberto Moreira-Filho. Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. *Scientific reports*, 5:16830, 2015.

[36] Adam McLaughlin and David A Bader. Accelerating GPU betweenness centrality. *Communications of the ACM*, 61(8):85–92, 2018.

[37] Neil R Taylor. Small world network strategies for studying protein structures and binding. *Computational and structural biotechnology journal*, 5(6):e201302006, 2013.

[38] Attila Gursoy, Ozlem Keskin, and Ruth Nussinov. Topological properties of protein interaction networks from a structural perspective, 2008.

[39] Gozde Kar, Guray Kuzu, Ozlem Keskin, and Attila Gursoy. Protein-protein interfaces integrated into interaction networks: implications on drug design. *Current pharmaceutical design*, 18(30):4697–4705, 2012.

[40] Julie L Morrison, Rainer Breitling, Desmond J Higham, and David R Gilbert. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC bioinformatics*, 6(1):233, 2005.

[41] Christian FA Negre, Uriel N Morzan, Heidi P Hendrickson, Rhitankar Pal, George P Lisi, J Patrick Loria, Ivan Rivalta, Junming Ho, and Victor S Batista. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, 115(52):E12201–E12208, 2018.

[42] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[43] Jishnu Das and Haiyuan Yu. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1):92, 2012.

[44] Paola Paci, Teresa Colombo, Giulia Fiscon, Aymone Gurtner, Giulio Pavesi, and Lorenzo Farina. SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Scientific reports*, 7:44797, 2017.

[45] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921, 2007.

[46] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761, 2010.

[47] Clara Pizzuti and Simona E Rombo. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.

[48] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.

[49] Min Li, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics*, 9(1):398, 2008.

[50] Zelmina Lubovac, Jonas Gamalielsson, and Björn Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, 64(4):948–959, 2006.

[51] Zelmina Lubovac. Integrative approach for detection of functional modules from protein-protein interaction networks. In *Protein-Protein Interactions-Computational and Experimental Tools*. IntechOpen, 2012.

[52] Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

[53] István A Kovács, Robin Palotai, Máté S Szalay, and Peter Csermely. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PloS one*, 5(9):e12528, 2010.

[54] Andrea Lancichinetti, Santo Fortunato, and Janos Kertesz. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[55] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.

[56] Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015.

[57] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

[58] Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Michael C Rosenstein, Wiegers, and Thomas C. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, 41(D1):D1104–D1114, 2012.

[59] Noa Rappaport, Noam Nativ, Gil Stelzer, Michal Twik, Yaron Guan-Golan, Tsippi Iny Stein, Iris Bahir, Frida Belinky, C Paul Morrey, Marilyn Safran, et al. MalaCards: an integrated compendium for diseases and their annotation. *Database*, 2013, 2013.

[60] Tim Beck, Robert K Hastings, Sirisha Gollapudi, Robert C Free, and Anthony J Brookes. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European journal of human genetics*, 22(7):949, 2014.

[61] Marharyta Petukh, Tugba G Kucukkal, and Emil Alexov. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Human mutation*, 36(5):524–534, 2015.

[62] Yoo-Ah Kim, Dong-Yeon Cho, and Teresa M Przytycka. Understanding genotype-phenotype effects in cancer via network approaches. *PLoS computational biology*, 12(3):e1004747, 2016.

[63] Yu Qian, Søren Besenbacher, Thomas Mailund, and Mikkel Heide Schierup. Identifying disease associated genes by network propagation. In *BMC systems biology*, volume 8, page S6. BioMed Central, 2014.

[64] Mark D Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, and Benjamin R Raphael. Pan-cancer identification of mutated pathways and protein complexes, 2014.

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

[66] Pora Kim, Aekyung Park, Guangchun Han, Hua Sun, Peilin Jia, and Zhongming Zhao. TissGDB: tissue-specific gene database in cancer. *Nucleic acids research*, 46(D1):1031–1038, 2017.

[67] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248, 2010.

[68] Peter J Thul and Cecilia Lindskog. The human protein atlas: a spatial map of the human proteome. *Protein Science*, 27(1):233–244, 2018.

[69] Latarsha J Carithers and Helen M Moore. The genotype-tissue expression (GTEx) project, 2015.

[70] Ruth Barshir, Omer Shwartz, Ilan Y Smoly, and Esti Yeger-Lotem. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS computational biology*, 10(6):e1003632, 2014.

[71] Esti Yeger-Lotem and Roded Sharan. Human protein interaction networks across tissues and diseases. *Frontiers in genetics*, 6:257, 2015.

[72] Maksim Kitsak, Amitabh Sharma, Jörg Menche, Emre Guney, Susan Dina Ghiassian, Joseph Loscalzo, and Albert-László Barabási. Tissue specificity of human disease module. *Scientific reports*, 6:35241, 2016.

[73] Jianguo Xia, Maia J Benner, and Robert EW Hancock. Networkanalyst-integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic acids research*, 42(W1):W167–W174, 2014.

[74] Amitabh Sharma, Jörg Menche, C Chris Huang, Tatiana Ort, Xiaobo Zhou, Maksim Kitsak, Nidhi Sahni, Derek Thibault, Linh Voung, Feng Guo, et al. A disease module in the interactome explains disease heterogeneity, drug re-

sponse and captures novel pathways and genes in asthma. *Human molecular genetics*, 24(11):3005–3020, 2015.

[75] Benedetta Lombardi, Paul Ashford, Aurelio A Moya-Garcia, Aleksander Rust, Mark Crawford, Sarah V Williams, Margaret A Knowles, Matilda Katan, Christine Orengo, and Jasminka Godovac-Zimmermann. Unique signalling connectivity of FGFR3-TACC3 oncoprotein revealed by quantitative phosphoproteomics and differential network analysis. *Oncotarget*, 8(61):102898, 2017.

[76] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, 2016.

[77] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010.

[78] A Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J Bellis, Jon Chambers, Mark Davies, Felix A Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, et al. The ChEMBL bioactivity database: an update. *Nucleic acids research*, 42(D1):D1083–D1090, 2014.

[79] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2017.

[80] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar,

Alfredo Smart, et al. The Pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.

[81] Ian Sillitoe, Natalie Dawson, Tony E Lewis, Sayoni Das, Jonathan G Lees, Paul Ashford, Adeyelu Tolulope, Harry M Scholes, Ilya Senatorov, Andra Bujan, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic acids research*, 47(D1):D280–D284, 2018.

[82] Erik LL Sonnhammer, Sean R Eddy, Ewan Birney, Alex Bateman, and Richard Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic acids research*, 26(1):320–322, 1998.

[83] UniProt Consortium et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699, 2018.

[84] Ian Sillitoe, Alison L Cuff, Benoit H Dessailly, Natalie L Dawson, Nicholas Furnham, David Lee, Jonathan G Lees, Tony E Lewis, Romain A Studer, Robert Rentzsch, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic acids research*, 41(D1):D490–D498, 2012.

[85] Sayoni Das, Ian Sillitoe, David Lee, Jonathan G Lees, Natalie L Dawson, John Ward, and Christine A Orengo. CATH FunFHMMer web server: protein functional annotations using functional family assignments. *Nucleic acids research*, 43(W1):W148–W153, 2015.

[86] David A Lee, Robert Rentzsch, and Christine Orengo. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic acids research*, 38(3):720–737, 2009.

[87] Weizhong Li and Adam Godzik. CD-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[88] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[89] Ruslan I Sadreyev, David Baker, and Nick V Grishin. Profile–profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Science*, 12(10):2262–2272, 2003.

[90] Sayoni Das and Christine A Orengo. Protein function annotation using protein domain family resources. *Methods*, 93:24–34, 2016.

[91] Yuxiang Jiang, Tal Ronnen Oron, Wyatt T Clark, Asma R Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S Funk, Indika Kahanda, Karin M Verspoor, Asa Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):184, 2016.

[92] The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens, author=Zhou, Naihui and Jiang, Yuxiang and Bergquist, Timothy R and Lee, Alexandra J and Kacsoh, Balint Z and Crocker, Alex W and Lewis, Kimberley A and Georghiou, George and Nguyen, Huy N and Hamid, Md Nafiz and others. *bioRxiv*, page 653105, 2019.

[93] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[94] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[95] Thomas Madden. The BLAST sequence analysis tool. In *The NCBI Hand-*

book. 2nd edition. National Center for Biotechnology Information (US), 2013.

[96] Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. HMMER web server: 2018 update. *Nucleic acids research*, 46(W1):W200–W204, 2018.

[97] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[98] Christine A Orengo and William R Taylor. SSAP: sequential structure alignment program for protein structure comparison. *Methods in enzymology*, 266:617–635, 1996.

[99] A.C.R Martin. ProFit program using McLachlan algorithm. http://www.bioinf.org.uk/software/profit, 2009.

[100] Rachael P Huntley, Tony Sawford, Maria J Martin, and Claire O'Donovan. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *GigaScience*, 3(1):4, 2014.

[101] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[102] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic acids research*, 47(D1):D590–D595, 2018.

[103] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41, 2019.

[104] Andrew Anighoro, Jurgen Bajorath, and Giulio Rastelli. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *Journal of medicinal chemistry*, 57(19):7874–7887, 2014.

[105] Aislyn DW Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297, 2010.

[106] Roger L Chang, Li Xie, Lei Xie, Philip E Bourne, and Bernhard Ø Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS computational biology*, 6(9):e1000938, 2010.

[107] Khushwant S Bhullar, Naiara Orrego Lagarón, Eileen M McGowan, Indu Parmar, Amitabh Jha, Basil P Hubbard, and HP Vasantha Rupasinghe. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer*, 17(1):48, 2018.

[108] Nicola Curtin. PARP inhibitors for anticancer therapy, 2014.

[109] Yvonne Y Li and Steven JM Jones. Drug repositioning for personalized medicine. *Genome medicine*, 4(3):27, 2012.

[110] A Srinivas Reddy and Shuxing Zhang. Polypharmacology: drug discovery for the future. *Expert review of clinical pharmacology*, 6(1):41–47, 2013.

[111] Aurelio A Moya-García and Juan AG Ranea. Insights into polypharmacology from drug-domain associations. *Bioinformatics*, 29(16):1934–1937, 2013.

[112] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews Drug discovery*, 1(9):727, 2002.

[113] Jürgen Drews. Genomic sciences and the medicine of tomorrow. *Nature biotechnology*, 14(11):1516, 1996.

[114] David Bailey, Edward Zanders, and Philip Dean. The end of the beginning for genomic medicine. *nature biotechnology*, 19(3):207, 2001.

[115] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[116] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19, 2017.

[117] Leslie Z Benet, Chelsea M Hosey, Oleg Ursu, and Tudor I Oprea. Bddcs, the rule of 5 and drugability. *Advanced drug delivery reviews*, 101:89–98, 2016.

[118] Eric B Fauman, Brajesh K Rai, and Enoch S Huang. Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Current opinion in chemical biology*, 15(4):463–468, 2011.

[119] Miquel Duran-Frigola and Patrick Aloy. Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chemistry & biology*, 20(4):594–603, 2013.

[120] Teresa Juan-Blanco, Miquel Duran-Frigola, and Patrick Aloy. IntSide: a web server for the chemical and biological examination of drug side effects. *Bioinformatics*, 31(4):612–613, 2014.

[121] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2015.

[122] Lucas Brouwers, Murat Iskar, Georg Zeller, Vera Van Noort, and Peer Bork. Network neighbors of drug targets contribute to drug side-effect similarity. *PloS one*, 6(7):e22187, 2011.

[123] Alan Talevi. Multi-target pharmacology: possibilities and limitations of the "skeleton key approach" from a medicinal chemist perspective. *Frontiers in pharmacology*, 6:205, 2015.

[124] Giulio Rastelli and Luca Pinzi. Computational polypharmacology comes of age. *Frontiers in pharmacology*, 6:157, 2015.

[125] Yasuo Tabei, Edouard Pauwels, Véronique Stoven, Kazuhiro Takemoto, and Yoshihiro Yamanishi. Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*, 28(18):i487–i494, 2012.

[126] Xiujuan Wang, Bram Thijssen, and Haiyuan Yu. Target essentiality and centrality characterize drug side effects. *PLoS computational biology*, 9(7):e1003119, 2013.

[127] Andreas P Russ and Stefan Lampel. The druggable genome: an update. *Drug discovery today*, 10(23-24):1607–1610, 2005.

[128] Su Datt Lam, Natalie L Dawson, Sayoni Das, Ian Sillitoe, Paul Ashford, David Lee, Sonja Lehtinen, Christine A Orengo, and Jonathan G Lees. Gene3D: expanding the utility of domain assignments. *Nucleic acids research*, 44(D1):D404–D409, 2015.

[129] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):168, 2009.

[130] Benjamin A Shoemaker, Dachuan Zhang, Manoj Tyagi, Ratna R Thangudu, Jessica H Fong, Aron Marchler-Bauer, Stephen H Bryant, Thomas Madej, and Anna R Panchenko. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic acids research*, 40(D1):D834–D840, 2011.

[131] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[132] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013.

[133] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[134] Áron R Perez-Lopez, Kristóf Z Szalay, Dénes Türei, Dezső Módos, Katalin Lenti, Tamás Korcsmáros, and Peter Csermely. Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. *Scientific reports*, 5:10182, 2015.

[135] Sayoni Das, David Lee, Ian Sillitoe, Natalie L Dawson, Jonathan G Lees, and Christine A Orengo. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics*, 31(21):3460–3467, 2015.

[136] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):R60, 2003.

[137] Margaret A Knowles and Carolyn D Hurst. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, 15(1):25, 2015.

[138] Oner Sanli, Jakub Dobruch, Margaret A Knowles, Maximilian Burger, Mehrdad Alemozaffar, Matthew E Nielsen, and Yair Lotan. Bladder cancer. *Nature reviews Disease primers*, 3:17022, 2017.

[139] Bogdan Czerniak, Colin Dinney, and David McConkey. Origins of bladder cancer. *Annual Review of Pathology: Mechanisms of Disease*, 11:149–174, 2016.

[140] Cui Yu, Chen Hequn, Liu Longfei, Wang Long, Chen Zhi, Zeng Feng, Chen Jinbo, Li Chao, and Zu Xiongbing. GSTM1 and GSTT1 polymorphisms are associated with increased bladder cancer risk: Evidence from updated meta-analysis. *Oncotarget*, 8(2):3246, 2017.

[141] Jaegil Kim, Rehan Akbani, Chad J Creighton, Seth P Lerner, John N Weinstein, Gad Getz, and David J Kwiatkowski. Invasive bladder cancer: genomic insights and therapeutic promise. *Clinical Cancer Research*, 21(20):4514–4524, 2015.

[142] Elisabeth Remy, Sandra Rebouissou, Claudine Chaouiya, Andrei Zinovyev, François Radvanyi, and Laurence Calzone. A modelling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer research*, pages canres–0602, 2015.

[143] Hong-Tao Li, Christopher E Duymich, Daniel J Weisenberger, and Gangning Liang. Genetic and epigenetic alterations in bladder cancer. *International neurourology journal*, 20(Suppl 2):S84, 2016.

[144] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315, 2014.

[145] Yanan Guo, Yvonne Chekaluk, Jianming Zhang, Jinyan Du, Wu Chin-Lee Gray, Nathanael S, and David J Kwiatkowski. TSC1 involvement in bladder cancer: diverse effects and therapeutic implications. *The Journal of pathology*, 230(1):17–27, 2013.

[146] Erica Di Martino, Darren C Tomlinson, Sarah V Williams, and Margaret A Knowles. A place for precision medicine in bladder cancer: targeting the FGFRs. *Future Oncology*, 12(19):2243–2263, 2016.

[147] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[148] Hans von der Maase, Lisa Sengelov, James T Roberts, Sergio Ricci, Luigi Dogliotti, T Oliver, Malcolm J Moore, Annamaria Zimmermann, and Michael Arning. Long-term survival results of a randomized trial comparing gemcitabine plus cisplatin, with methotrexate, vinblastine, doxorubicin, plus cisplatin in patients with bladder cancer. *Journal of clinical oncology*, 23(21):4602–4608, 2005.

[149] Nabil Ismaili, Mounia Amzerin, and Aude Flechon. Chemotherapy in advanced bladder cancer: current status and future. *Journal of hematology & oncology*, 4(1):35, 2011.

[150] Joseph M Gozgit, Matthew J Wong, Lauren Moran, Scott Wardwell, Qurish K Mohemmad, Narayana I Narasimhan, William C Shakespeare, Frank Wang, Tim Clackson, and Victor M Rivera. Ponatinib (AP24534), a multitargeted pan-FGFR inhibitor with activity in multiple FGFR-amplified or mutated cancer models. *Molecular cancer therapeutics*, pages molcanther–0450, 2012.

[151] V Chell, K Balmanno, AS Little, M Wilson, S Andrews, L Blockley, M Hampson, PR Gavine, and SJ Cook. Tumour cell responses to new fibroblast growth factor receptor tyrosine kinase inhibitors and identification of a gatekeeper mutation in FGFR3 as a mechanism of acquired resistance. *Oncogene*, 32(25):3059, 2013.

[152] Benedito A Carneiro, Joshua J Meeks, Timothy M Kuzel, Mariana Scaranti, Sarki A Abdulkadir, and Francis J Giles. Emerging therapeutic targets in bladder cancer. *Cancer Treatment Reviews*, 41(2):170–178, 2015.

[153] Sumanta K Pal, Jonathan E Rosenberg, Jean H Hoffman-Censits, Raanan Berger, David I Quinn, Matthew D Galsky, Juergen Wolf, Christian Dittrich, Bhumsuk Keam, Jean-Pierre Delord, et al. Efficacy of BGJ398, a fibroblast growth factor receptor 1-3 inhibitor, in patients with previously treated advanced urothelial carcinoma with FGFR3 alterations. *Cancer discovery*, pages CD–18, 2018.

[154] Guang Liang, Zhiguo Liu, Jianzhang Wu, Yuepiao Cai, and Xiaokun Li. Anticancer molecules targeting fibroblast growth factor receptors. *Trends in pharmacological sciences*, 33(10):531–541, 2012.

[155] MH Hussain, GR MacVicar, DP Petrylak, RL Dunn, U Vaishampayan, PN Lara Jr, GS Chatta, DM Nanus, LM Glode, and DL Trump. Trastuzumab, paclitaxel, carboplatin, and gemcitabine in advanced human epidermal growth factor receptor-2/neu-positive urothelial carcinoma: results of a multicenter phase II National Cancer Institute trial. *J Clin Oncol*, 25(16):2218–24, 2007.

[156] Yuping Han, Xuefei Jin, Hui Zhou, and Bin Liu. Identification of key genes associated with bladder cancer using gene expression profiles. *Oncology letters*, 15(1):297–303, 2018.

[157] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[158] Paul Ashford, Camilla SM Pang, Aurelio A Moya-García, Tolulope Adeyelu, and Christine A Orengo. A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. *Scientific reports*, 9(1):263, 2019.

[159] Colin Clarke, Stephen F Madden, Padraig Doolan, Sinead T Aherne, Helena Joyce, Lorraine O'driscoll, William M Gallagher, Bryan T Hennessy, Michael Moriarty, and John Crown. Correlating transcriptional networks to

breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis*, 34(10):2300–2308, 2013.

[160] Mariana Maschietto, Ana C Tahira, Renato Puga, Leandro Lima, Daniel Mariani, Bruna da Silveira Paulsen, Paulo Belmonte-de Abreu, Henrique Vieira, Ana CV Krepischi, and Dirce M Carraro. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. *BMC medical genomics*, 8(1):23, 2015.

[161] B Ning, DL Xu, JH Gao, LL Wang, S Yan, and S Cheng. Identification of pathway-related modules in high-grade osteosarcoma based on topological centrality of network strategy. *Eur Rev Med Pharmacol Sci*, 20(11):2209–2220, 2016.

[162] Xuesong Wu, Haoran Tang, Aoran Guan, Feng Sun, Hui Wang, and Jie Shu. Finding gastric cancer related genes and clinical biomarkers for detection based on gene–gene interaction network. *Mathematical biosciences*, 276:1–7, 2016.

[163] Kivilcim Ozturk, Michelle Dow, Daniel E Carlin, Rafael Bejar, and Hannah Carter. The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 2018.

[164] Ara Cho, Jung Eun Shim, Eiru Kim, Fran Supek, Ben Lehner, and Insuk Lee. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome biology*, 17(1):129, 2016.

[165] Heiko Horn, Michael S Lawrence, Candace R Chouinard, Yashaswi Shrestha, Jessica Xin Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Eejung Kim, and Atanas Kamburov. NetSig: network-based discovery from cancer genomes. *Nature methods*, 15(1):61, 2018.

[166] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic anal-

ysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology*, 11(4):e1004120, 2015.

[167] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177, 2004.

[168] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, and Sari Ward. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(D1):D805–D811, 2014.

[169] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt. The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, 130(4):453–459, 2017.

[170] Mehmet Kemal Samur. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PloS one*, 9(9):e106397, 2014.

[171] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, and Isabella Castiglioni. TCGAbiolinks: an r/bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, 44(8):e71–e71, 2015.

[172] Simone de Jong, Marco PM Boks, Tova F Fuller, Eric Strengman, Esther Janson, Carolien GF de Kovel, Anil PS Ori, Nancy Vi, Flip Mulder, Jan Dirk Blom, et al. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PloS one*, 7(6):e39498, 2012.

[173] Yan Chen, Yining Liu, Min Du, Wengang Zhang, Ling Xu, Xue Gao, Lupei Zhang, Huijiang Gao, Lingyang Xu, and Junya Li. Constructing a comprehensive gene co-expression based interactome in Bos taurus. *PeerJ*, 5:e4107, 2017.

[174] Rong Liu, Yu Cheng, Jing Yu, Qiao Li Lv, and Hong Hao Zhou. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene*, 563(1):56–62, 2015.

[175] Jean-Karim Hériché, Jon G Lees, Ian Morilla, Thomas Walter, Boryana Petrova, M Julia Roberti, M Julius Hossain, Priit Adler, José M Fernández, and Martin Krallinger. Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Molecular biology of the cell*, 25(16):2522–2536, 2014.

[176] Luh Yen, Francois Fouss, Christine Decaestecker, Pascal Francq, and Marco Saerens. Graph nodes clustering based on the commute-time kernel. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1037–1045. Springer, 2007.

[177] Guangchuang Yu, Li Gen Wang, Yanyan Han, and Qing Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[178] A Liberzon, C Birger, H Thorvaldsdottir, M Ghandi, JP Mesirov, and P Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell system*, 1(6):417–425, 2015.

[179] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2016.

[180] Aurelio Moya-García, Tolulope Adeyelu, Felix A Kruger, Natalie L Dawson, Jon G Lees, John P Overington, Christine Orengo, and Juan AG Ranea. Structural and functional view of polypharmacology. *Scientific reports*, 7(1):10102, 2017.

[181] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo

summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, 2011.

[182] Jason E Duex, Kalin E Swain, Garrett M Dancik, Richard D Paucek, Charles Owens, Mair EA Churchill, and Dan Theodorescu. Functional impact of chromatin remodeling gene mutations and predictive signature for therapeutic response in bladder cancer. *Molecular Cancer Research*, 16(1):69–77, 2018.

[183] Pablo García-Sanz, Juan Carlos Triviño, Alba Mota, María Pérez López, Eva Colás, Alejandro Rojo-Sebastián, Ángel García, Sonia Gatius, María Ruiz, Jaime Prat, et al. Chromatin remodelling and dna repair genes are frequently mutated in endometrioid endometrial carcinoma. *International journal of cancer*, 140(7):1551–1563, 2017.

[184] Jacquelyn J Bower, Leah D Vance, Matthew Psioda, Stephanie L Smith-Roe, Dennis A Simpson, Joseph G Ibrahim, Katherine A Hoadley, Charles M Perou, and William K Kaufmann. Patterns of cell cycle checkpoint deregulation associated with intrinsic molecular subtypes of human breast cancer cells. *NPJ Breast Cancer*, 3(1):9, 2017.

[185] Yu Sun, Wen-Zhou Liu, Tao Liu, Xu Feng, Nuo Yang, and Hua-Fu Zhou. Signaling pathway of mapk/erk in cell proliferation, differentiation, migration, senescence and apoptosis. *Journal of Receptors and Signal Transduction*, 35(6):600–604, 2015.

[186] T Zhan, N Rindtorff, and Michael Boutros. Wnt signaling in cancer. *Oncogene*, 36(11):1461, 2017.

[187] Konstantinos Tryfonidis, Dimitrios Zardavas, Benita S Katzenellenbogen, and Martine Piccart. Endocrine treatment in breast cancer: Cure, resistance and beyond. *Cancer treatment reviews*, 50:68–81, 2016.

[188] Erli Pang, Yu Hao, Ying Sun, and Kui Lin. Differential variation patterns

between hubs and bottlenecks in human protein-protein interaction networks. *BMC evolutionary biology*, 16(1):260, 2016.

[189] Michael Batie, Luis Del Peso, and Sonia Rocha. Hypoxia and chromatin: a focus on transcriptional repression mechanisms. *Biomedicines*, 6(2):47, 2018.

[190] Carmen Belli, Dario Trapani, Giulia Viale, Paolo D'Amico, Bruno Achutti Duso, Paolo Della Vigna, Franco Orsi, and Giuseppe Curigliano. Targeting the microenvironment in solid tumors. *Cancer treatment reviews*, 65:22–32, 2018.

[191] Yogita Dheer, Nitin Chitranshi, Veer Gupta, Mojdeh Abbasi, Mehdi Mirzaei, Yuyi You, Roger Chung, Stuart L Graham, and Vivek Gupta. Bexarotene modulates retinoid-x-receptor expression and is protective against neurotoxic endoplasmic reticulum stress response and apoptotic pathway activation. *Molecular neurobiology*, 55(12):9043–9056, 2018.

[192] Targeting the PI3K/AKT/mTOR pathway in bladder cancer, author=Sathe, Anuja and Nawroth, Roman. In *Urothelial Carcinoma*, pages 335–350. Springer, 2018.

[193] Shafaat A Rabbani, Maria-Luisa Valentino, Ani Arakelian, Suhad Ali, and Frank Boschelli. Ski-606 (bosutinib) blocks prostate cancer invasion, growth, and metastasis in vitro and in vivo through regulation of genes involved in cancer growth and skeletal metastasis. *Molecular cancer therapeutics*, 9(5):1147–1157, 2010.

[194] Robert Roskoski. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological research*, 103:26–48, 2016.

[195] Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[196] Steven K Hanks and Tony Hunter. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB journal*, 9(8):576–596, 1995.

[197] Diego Miranda-Saavedra and Geoffrey J Barton. Classification and functional annotation of eukaryotic protein kinases. *Proteins: Structure, Function, and Bioinformatics*, 68(4):893–914, 2007.

[198] David MA Martin, Diego Miranda-Saavedra, and Geoffrey J Barton. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic acids research*, 37(suppl_1):D244–D250, 2008.

[199] Juliette Martin, Krishanpal Anamika, and Narayanaswamy Srinivasan. Classification of protein kinases on the basis of both kinase and non-kinase regions. *PloS one*, 5(9):e12460, 2010.

[200] Ramaswamy Rakshambikai, Malini Manoharan, Mutharasu Gnanavel, and Narayanaswamy Srinivasan. Typical and atypical domain combinations in human protein kinases: functions, disease causing mutations and conservation in other primates. *RSC Advances*, 5(32):25132–25148, 2015.

[201] A Krupa, KR Abhinandan, and Narayanaswamy Srinivasan. King: a database of protein kinases in genomes. *Nucleic acids research*, 32(suppl_1):D153–D155, 2004.

[202] Krisna C Duong-Ly and Jeffrey R Peterson. The human kinome and kinase inhibition. *Current protocols in pharmacology*, pages 2–9, 2013.

[203] Derek P Brazil and Brian A Hemmings. Ten years of protein kinase B signalling: a hard Akt to follow. *Trends in biochemical sciences*, 26(11):657–664, 2001.

[204] Brian C Shonesy, Nidhi Jalan-Sakrikar, Victoria S Cavener, and Roger J Colbran. CaMKII: a molecular substrate for synaptic plasticity and memory. *Prog Mol Biol Transl Sci*, 122:61–87, 2014.

[205] Uwe Knippschild, Sonja Wolff, Georgios Giamas, Claas Brockschmidt, Mathias Wittau, Peter Uwe Würl, Thorsten Eismann, and Martin Stöter. The role of the casein kinase 1 (CK1) family in different signaling pathways linked to cancer development. *Oncology Research and Treatment*, 28(10):508–514, 2005.

[206] John C Foreman, Torben Johansen, and Alasdair J Gibb. *Textbook of receptor pharmacology*. CRC press, 2010.

[207] Shari L Wiseman, Fan Yan Wei, and Angus C Nairn. The EF2K/MHCK/TRPM7 family of atypical protein kinases. In *Handbook of Cell Signaling, 2/e*. Elsevier Inc., 2010.

[208] Susan S Taylor and Alexandr P Kornev. Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 36(2):65–77, 2011.

[209] Doriano Fabbro, Sandra W Cowan-Jacob, and Henrik Moebitz. Ten things you should know about protein kinases: Iuphar review 14. *British journal of pharmacology*, 172(11):2675–2700, 2015.

[210] Susan S Taylor, Malik M Keshwani, Jon M Steichen, and Alexandr P Kornev. Evolution of the eukaryotic protein kinases as dynamic molecular switches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2517–2528, 2012.

[211] Marc D Jacobs, Paul R Caron, and Brian J Hare. Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: Structure of lck/imatinib complex. *Proteins: Structure, Function, and Bioinformatics*, 70(4):1451–1460, 2008.

[212] Oliver Hantschel and Giulio Superti-Furga. Regulation of the c-Abl and Bcr-Abl tyrosine kinases. *Nature reviews. Molecular cell biology*, 5(1):33, 2004.

[213] Chung-Jung Tsai and Ruth Nussinov. The molecular basis of targeting protein kinases in cancer therapeutics. In *Seminars in cancer biology*, volume 23, pages 235–242. Elsevier, 2013.

[214] Sara Cheek, Hong Zhang, and Nick V Grishin. Sequence and structure classification of kinases. *Journal of molecular biology*, 320(4):855–881, 2002.

[215] Peng Wu, Thomas E Nielsen, and Mads H Clausen. FDA-approved small-molecule kinase inhibitors. *Trends in pharmacological sciences*, 36(7):422–439, 2015.

[216] Tadeusz Robak and Ewa Robak. Tyrosine kinase inhibitors as potential drugs for b-cell lymphoid malignancies and autoimmune disorders. *Expert opinion on investigational drugs*, 21(7):921–947, 2012.

[217] Akintunde Akinleye, Muhammad Furqan, and Oluwaseyi Adekunle. Ibrutinib and indolent B-cell lymphomas. *Clinical Lymphoma Myeloma and Leukemia*, 14(4):253–260, 2014.

[218] Monia Hossam, Deena S Lasheen, and Khaled AM Abouzid. Covalent egfr inhibitors: Binding mechanisms, synthetic approaches, and clinical profiles. *Archiv der Pharmazie*, 349(8):573–593, 2016.

[219] Jin H Park, Yingting Liu, Mark A Lemmon, and Ravi Radhakrishnan. Erlotinib binds both inactive and active conformations of the EGFR tyrosine kinase domain. *Biochemical Journal*, 448(3):417–423, 2012.

[220] Sandra W Cowan-Jacob, Henrik Möbitz, and Doriano Fabbro. Structural biology contributions to tyrosine kinase drug discovery. *Current opinion in cell biology*, 21(2):280–287, 2009.

[221] Vandana Lamba and Indraneel Ghosh. New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. *Current pharmaceutical design*, 18(20):2936–2945, 2012.

[222] Stanley F Barnett, Deborah Defeo-Jones, FU Sheng, Paula J Hancock, M Kathleen, Raymond E Jones, Jason A Kahana, Karen Leander, John Malinowski, and Elizabeth M McAvoy. Identification and characterization of pleckstrin-homology-domain-dependent and isoenzyme-specific Akt inhibitors. *Biochemical Journal*, 385(2):399–408, 2005.

[223] Peng Wu, Mads H Clausen, and Thomas E Nielsen. Allosteric small-molecule kinase inhibitors. *Pharmacology & therapeutics*, 156:59–68, 2015.

[224] Ye Hu, Ryo Kunimoto, and Jürgen Bajorath. Mapping of inhibitors and activity data to the human kinome and exploring promiscuity from a ligand and target perspective. *Chemical biology & drug design*, 89(6):834–845, 2017.

[225] Piero Giansanti, Christian Preisinger, Kilian VM Huber, Manuela Gridling, Giulio Superti-Furga, Keiryn L Bennett, and Albert JR Heck. Evaluating the promiscuous nature of tyrosine kinase inhibitors assessed in A431 epidermoid carcinoma cells by both chemical-and phosphoproteomics. *ACS chemical biology*, 9(7):1490–1498, 2014.

[226] Danzhi Huang, Ting Zhou, Karine Lafleur, Cristina Nevado, and Amedeo Caflisch. Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis. *Bioinformatics*, 26(2):198–204, 2009.

[227] Yi-Yuan Chiu, Chih-Ta Lin, Jhang-Wei Huang, Kai-Cheng Hsu, Jen-Hu Tseng, Syuan-Ren You, and Jinn-Moon Yang. KIDFamMap: a database of kinase-inhibitor-disease family maps for kinase inhibitor selectivity and binding mechanisms. *Nucleic acids research*, 41(D1):D430–D440, 2012.

[228] Edwin C Webb et al. *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Number Ed. 6. Academic Press, 1992.

[229] Jonathan M Goldberg, Allison D Griggs, Janet L Smith, Brian J Haas, Jennifer R Wortman, and Qiandong Zeng. Kinannote, a computer program to

identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics*, 29(19):2387–2394, 2013.

[230] Patricia Dranchak, Ryan MacArthur, Rajarshi Guha, William J Zuercher, David H Drewry, Douglas S Auld, and James Inglese. Profile of the GSK published protein kinase inhibitor set across ATP-dependent and-independent luciferases: implications for reporter-gene assays. *PloS one*, 8(3):e57888, 2013.

[231] Stefan Knapp, Paulo Arruda, Julian Blagg, Stephen Burley, David H Drewry, Aled Edwards, Doriano Fabbro, Paul Gillespie, Nathanael S Gray, and Bernhard Kuster. A public-private partnership to unlock the untargeted kinome. *Nature chemical biology*, 9(1):3–6, 2013.

[232] Theonie Anastassiadis, Sean W Deacon, Karthik Devarajan, Haiching Ma, and Jeffrey R Peterson. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1039–1045, 2011.

[233] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nature communications*, 7:10331, 2016.

[234] Jacques Colinge, Adrián César-Razquin, Kilian Huber, Florian P Breitwieser, Peter Májek, and Giulio Superti-Furga. Building and exploring an integrated human kinase network: global organization and medical entry points. *Journal of proteomics*, 107:113–127, 2014.

[235] Emre Guney. Investigating side effect modules in the interactome and their use in drug adverse effect discovery. In *International Workshop on Complex Networks*, pages 239–250. Springer, 2017.

[236] VK MD Aksam, VM Chandrasekaran, and Sundaramurthy Pandurangan. Hub nodes in the network of human mitogen-activated protein kinase (mapk)

pathways: characteristics and potential as drug targets. *Informatics in Medicine Unlocked*, 9:173–180, 2017.

[237] Jodi A Hadden and Juan R Perilla. Molecular dynamics simulations of protein–drug complexes: A computational protocol for investigating the interactions of small-molecule therapeutics with biological targets and biosensors. In *Computational Drug Discovery and Design*, pages 245–270. Springer, 2018.

[238] Saleem Iqbal, Dhanabalan Anantha Krishnan, and Krishnasamy Gunasekaran. Identification of potential pkc inhibitors through pharmacophore designing, 3d-qsar and molecular dynamics simulations targeting alzheimer's disease. *Journal of Biomolecular Structure and Dynamics*, 36(15):4029–4044, 2018.