Chapter 1

Protein Kinases Inhibitors and Domain Families

1.1 Introduction

1.1.1 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 [Santos et al., 2017]). Protein kinases are enzymes that are involved in several cellular pathways. These enzymes catalyse the transfer of γ-phosphate of ATP to the hydroxyl groups of an acceptor molecule which can either be protein substrates, lipids or small molecules. Through this phosphorylation process, the enzymatic activities of the protein kinases are covalently modified which lead to alteration in 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 human following zinc finger proteins, G-protein coupled receptors, immunoglobulins, and the protease enzymes [Roskoski, 2016]. About 2% of the human genome have been shown to encode protein kinases. Manning and colleagues in 2002, identified all sequenced eukaryotic protein kinases by searching all human genome sequence sources including Celera genomic 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 [Manning et al., 2002]. 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 and the kinase domain of eukaryotic proteins are 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 as well as subfamilies. There are 9 groups of protein kinases based on the sequence and structural similarities of the catalytic domain. Classification is also guided by knowledge of the domain structure outside the catalytic domains, known biological functions and evolutionary history of the kinases. The Manning classification (stored in the KinBase database) is an extension of the work by Hanks and Hunter who initially performed a conservation and phylogeny analysis of the catalytic domain of eukaryotic protein to reveal the conserved features of catalytic domains and thus, classified the protein kinase into 5 groups, 44 families and 51 subfamilies [Hanks and Hunter, 1995] while [Manning et al., 2002] further extended this to 9 groups, 134 families and 196 subfamilies. Figure 1.1 below shows the grouping of the human protein kinases.

Other classification schemes for the protein kinases were also developed over the years. For instance, [Miranda-Saavedra and Barton, 2007] used a multilevel hidden Markov model library to classify protein kinases into 12 families. They showed that classification by multilevel HMM library outperformed BLASTP and 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 [Miranda-Saavedra and Barton, 2007] was built using sequences derived from KinBase [Manning et al., 2002] and this was stored in the Kinomer database [Martin et al., 2008]. Another classification considered the diversity of the accessory domains and their organisations. In their approach, [Martin et al., 2010] classification was performed manually and it was based on similarity in the sequences and structural features outside the catalytic domain.

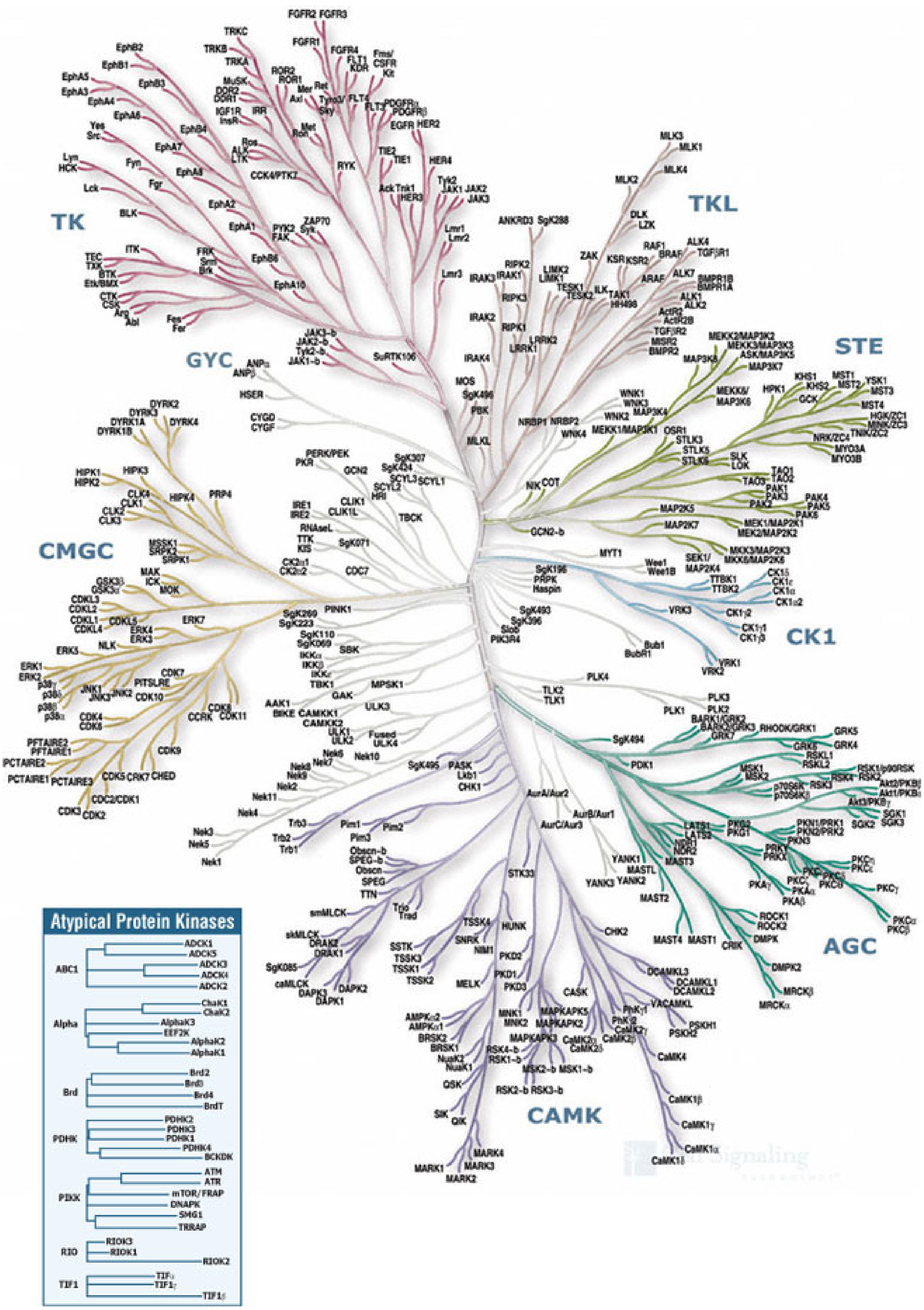


Figure 1.1: The human kinome. Kinome illustration courtesy of Cell Signalling Technology, Inc (www.cellsignal.com) based on [Manning et al., 2002].

1.1.2 Classification of Protein Kinases

The comprehensive work done on the classification of kinases by [Manning et al., 2002] has accomplished several accolades and citations with over 8000 reported by google scholar. Below, the groups identified by [Manning et al., 2002] are described in more detail.

AGC Kinases

These group of kinases include PKA, PKG, and PKC. They are involved in diverse cellular roles such as cell growth and proliferation, cell survival, glucose metabolism and protein synthesis. They are also dysregulated in several diseases such as cancer and neurological disorder, inflammation and viral infection [Rakshambikai et al., 2015]. The Akt isoform possesses the pleckstrein homology domain (PH-domain) at the N-terminal and with this domain, it interacts with PIP3 and PIP2 which leads to the activation of pyruvate dehydrogenase kinase isoenzyme (PDK1). PKC also interacts with DAG and calcium by its N-terminal conserved domains (C1 and C2) which leads to conformational changes and activation of the protein [Duong-Ly and Peterson, 2013].

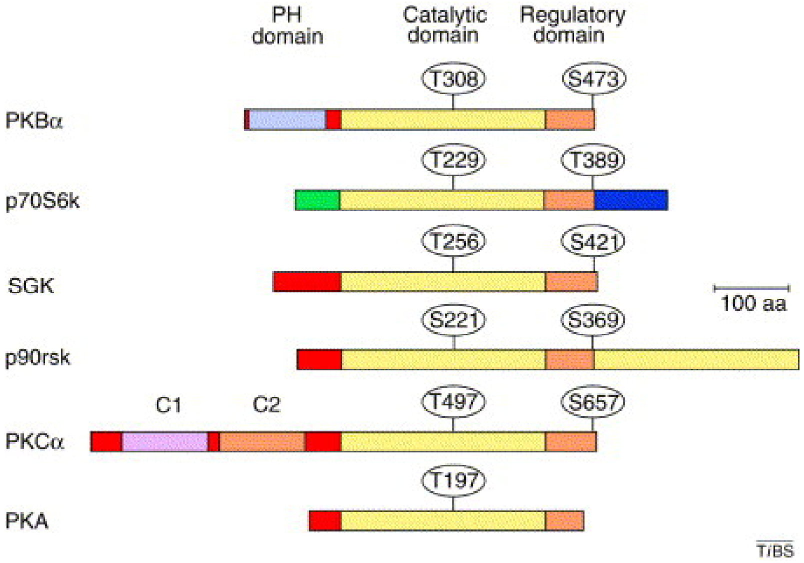


Figure 1.2: The domain structure of AGC kinase family. All members contain Thr/Ser in the activation loop. Figure taken from [Brazil and Hemmings, 2001]

CAMK Kinases

These kinases are involved in calcium signalling and are basically 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 and they are all in multidomain architectures. Each member of this family possess additional unique domains in addition to the conserved kinase domain. For instance, the CASK (Ca^{2+} /calmodulin dependent serine kinases) contains a series of L27 repeats, PDZ, two Src homology 3 (SH3) domain as well as a C-terminal guanylate kinase domain [Rakshambikai et al., 2015]. 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.

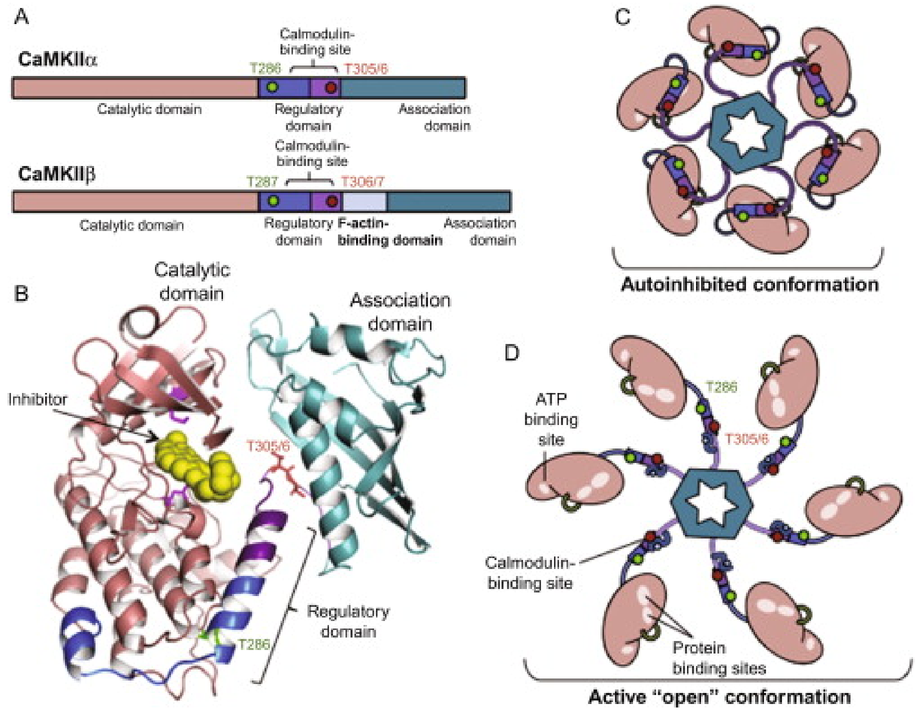


Figure 1.3: Domain organisation and structure of CaMKII. (A) Similar domain organisation in the CaMKIIα and CaMKIIβ with the exception of F-actin binding domain inserted in CaMKIIβ. (B) Structure of autoinhibited CaMKII subunit PDB ID: 3SOA, (C) Cartoon showing the compact inactive holoenzyme (D) Cartoon showing conformational changes associated with CaMKII activation.

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-gamma subfamily which possesses CK1-gamma domain whose function is not yet known.

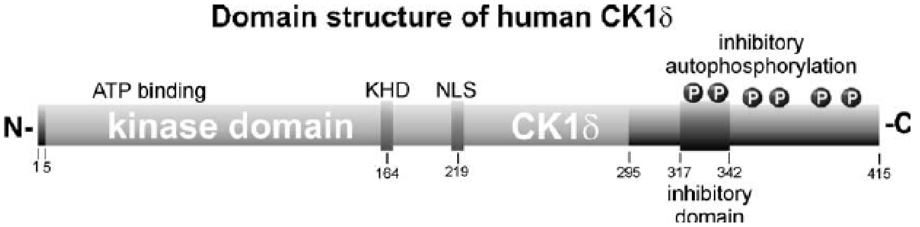


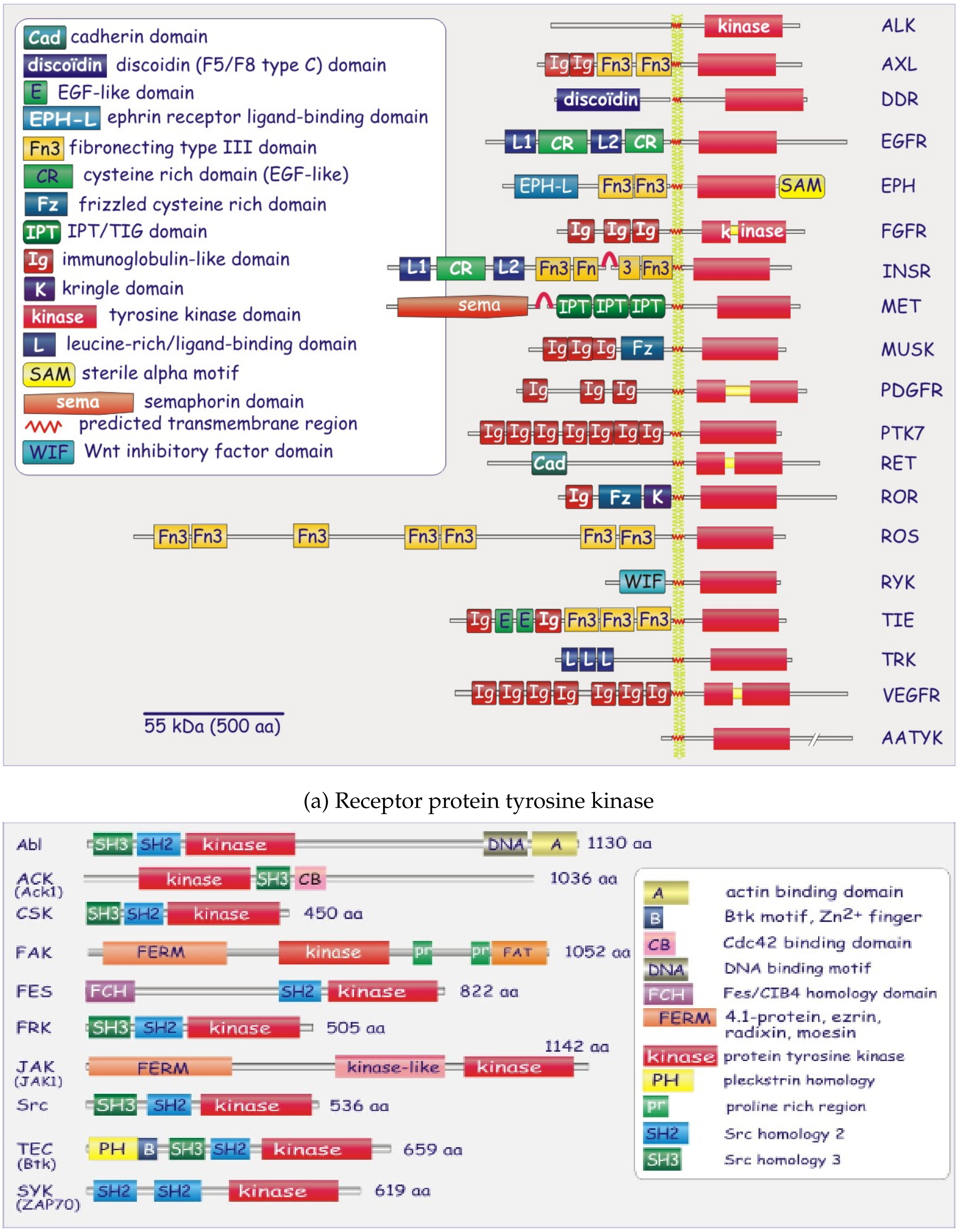
Figure 1.4: Domain structure of human CK1δy. The members share a common conserved kinase domain but differs in their variable N-and-C terminal domains. The regulatory c-terminal domain has multiple inhibitory autophosphorylation sites. The nuclear localization signal(NLS) and kinesin homology domain (KHD) are also located within the kinase domain. Figure obtained from [Knippschild et al., 2005]

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 regulates the progresion through the different phases of the cell cycle in association with their activating partners cyclins. The MAP kinases are amongst the most highly studied signal molecules. The MAP kinase cascade controls proliferation, differentiation, and cell-death across various eukaryotes. The GSK-3 kinases are key metabolic enzyme 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 large families,receptor and non-receptor (cytosolic) kinases. The receptor TKs are subdivided based on the sequence homology and the structure of their extracellular domains into 20 subfamilies. One of the most studied extracellular domain is the Ig-like domain which occurs in most of the members of this subgroup. The extracellular domains acts 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 [Rakshambikai et al., 2015]. In addition to the kinase catalytic domain, they also possess additional domains that are important for enzymatic regulation and substrate recognition. Src families for instance possess additional SH3, SH2 domains.The Abl has an F-actin binding site and a DNA-binding region, Fak possess a ferm domain and a focal adhension-binding domain which are important for mediating protein-protein interation [Duong-Ly and Peterson, 2013, Rakshambikai et al., 2015].



(a) Receptor protein tyrosine kinase

From Cell Biology

(b) Non-receptor protein tyrosine kinase families

Figure 1.5: The multidomain architecture of tyrosine kinases. Figures were taken from [Foreman et al., 2010].

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 nonreceptor kinases. They comprise of 8 major subfamilies 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 it 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 the kinases and they consist entirely of pesudo-kinases that lack certain residues that are critical for phosphate transfer [Manning et al., 2002]. They convert GTP to GMP

Others

These include members that lack sufficient sequence similarity to those given above and display unusual phosphorylation properties using ATP and GTP as phosphate donor. Examples include CK2, IKKs.

Atypical protein kinases (aPks)

The atypical kinases represents group of human kinases that lacks similar sequence identity with the ePKs kinase domain HMM profile but have been shown experimentally to have protein kinase activity. Examples includes PIKK family, A6 family, RIO and Pyruvate dehydrogenase kinase [Manning et al., 2002].

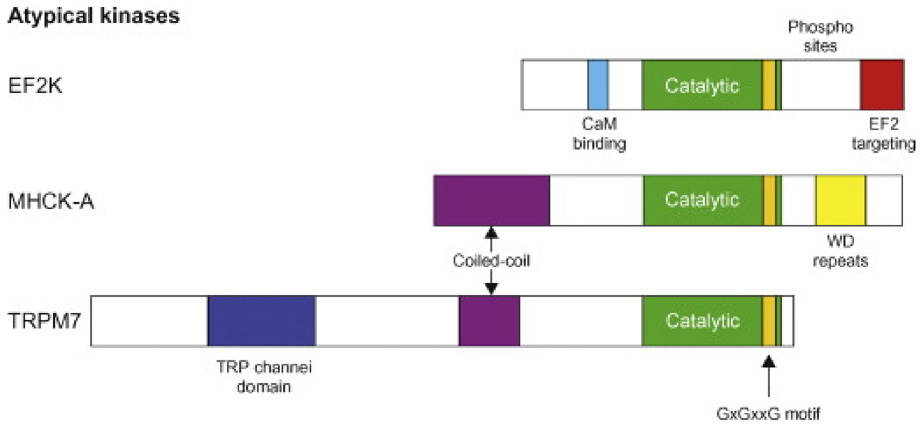


Figure 1.6: Domain organisation of atypical family of protein kinases. In contrast to classical kinases, the GXGXXG motif of atypical kinases is not involved in MgATP binding but likely involved in peptide interaction.

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 reveal some of the major differences between these two. See Figure 1.7.

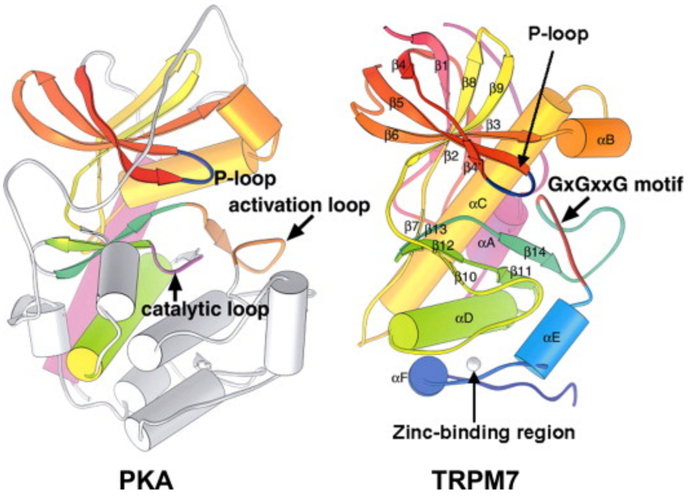


Figure 1.7: Structural comparison of the kinase domains of TRPM7 and PKA. The N-lobe of both PKA and TRPM7 is largely comprised of β-strands while the MgATP binds at the cleft formed from both the N and C-lobes and the binding of Mg in both also involve the conserved P-loop. However, the catalytic loop is not conserved. The GXGXXG motif in TRPM7 contains an extended loop that may play a similar role as the activation loop in classical protein kinase, PKA. Figure obtained from [Wiseman et al., 2010].

1.1.3 Structural Features of Protein Kinase

The protein kinases possess catalytic domains and the 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 above section. Below, structural features of the catalytic domain of the protein kinases are described. As mentioned earlier, the catalytic domain of the kinase spans about 250 amino acids 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. The N-lobe has about 90 amino acids that are organised 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 γ-phosphate of the ATP for catalysis. It is the most flexible part of the N-lobe [Taylor and Kornev, 2011]. 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 the purine is different. For instance, the P-loop does not contact the purine moeity of the ATP while the gly-rich loop connects both β strands that habour 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 [Fabbro et al., 2015]. 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.

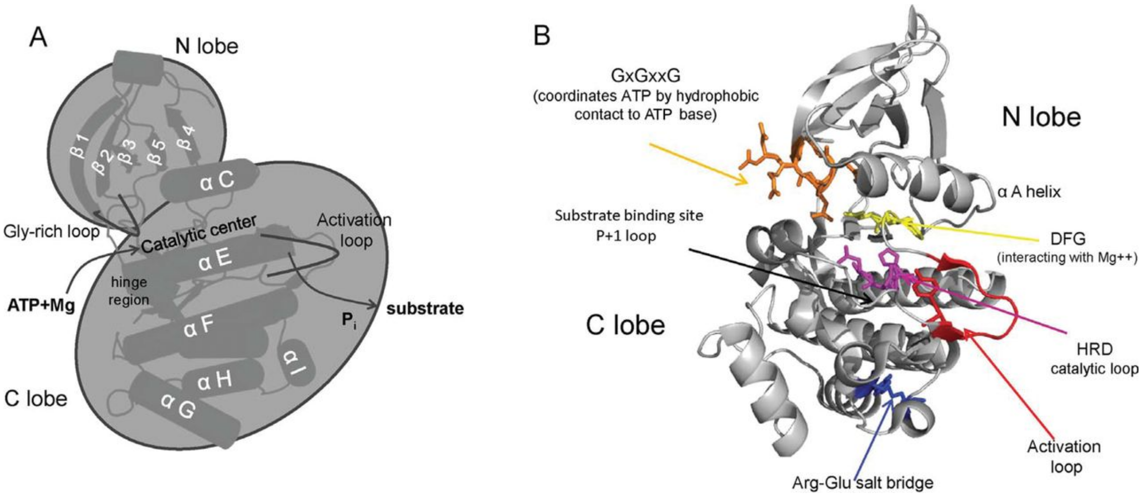


Figure 1.8: Schematic overview of kinase features. (a). General overview of the organisation of the kinase domain (b) Conserved motifs and residues of the catalytic core of the protein kinases. Taken from [Lorenzen and Pawson, 2014].

The C-helix serves as a ”signaling integration motif” as it connects to different parts of the kinase domain (See Figure 1.9a). Its C-terminus is connected to the C-lobe by the αC-β4 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 [Taylor and Kornev, 2011].

The C-lobe varies in size, sequence and topology. It is predominatly 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 well protected from contact with solvent, however, the G-helix is exposed to the solvent. The β-subdomain of the C-lobe comprises 4 short β strands (6-9) and contains much of the catalytic machinery for transfering 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 DFG (magnesium positioning loop) and 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 segment are the most variable part of the kinase core and this is responsible for turning on and off the kinase [Taylor and Kornev, 2011]. Furthermore, the F-segment extends to the GHI-subdomain where substrates and regulatory proteins bind. This part is also responsible for stabilizing the active kinase core and also for its allosteric sites (See Figure 1.9c).

The hinge region of the kinase represents the connecting loop between 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 (LSP) alignment of protein kinase (a method for comparing two protein structures and identifying spatially conserved residues) revealed two hydrophobic motifs called ″spines,″ that connects the N and C-lobe. Structural analysis of the spines give insight into how an active protein kinase is assembled from an inactive protein kinase [Taylor and Kornev, 2011]. 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. The R spine is therefore a hydrophobic spine that links the two lobes.

Using the LSP on the conserved core of the protein kinase, another hydrophobic spine was identified, called the catalytic (C-spine). Like the R-spine, it comprises hydrophobic residues belonging to both lobes. In the N-lobe, the Val^{57} in the β2 and Ala^{70} from the AxK-motif as well as the Leu^{173} in the C-lobe that 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 [Taylor and Kornev, 2011, Fabbro et al., 2015, Roskoski, 2016].

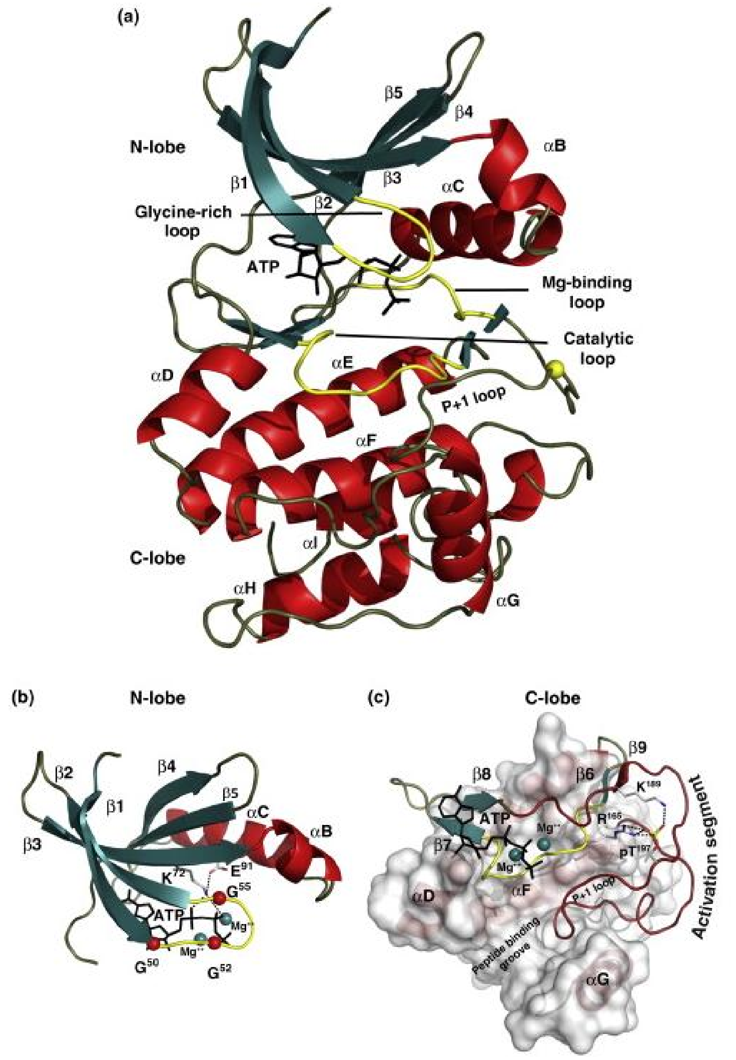


Figure 1.9: The structure of conserved kinase core. (a) Protein kinase with the characteristic bilobal structure. (b) In the N-lobe structure, the glycine rich loop coordinates the ATP phosphate binding while the $\beta 3$ strand couples the phosphates and the C-helix. (c) Catalytic and regulatory machinery outlined against the rigid core of the C-lobe [Taylor and Kornev, 2011]

1.1.4 Active and Inactive Protein Kinases

The structure of protein kinases reveals the conformational variation of active and inactive kinases. One of the most common forms of inactive protein kinases is the assumption of the DFG-Asp of the activation segment in an out-conformation where the aspartate is directed outward whereas the DFG-Phe is directed inward towards the active site [Roskoski, 2016]. Structural elements of the kinases show distinct conformation in the active and inactive state. 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 Figure1.10). The phosphorylation of the residues within the activation loop activates the kinases [Roskoski, 2016].

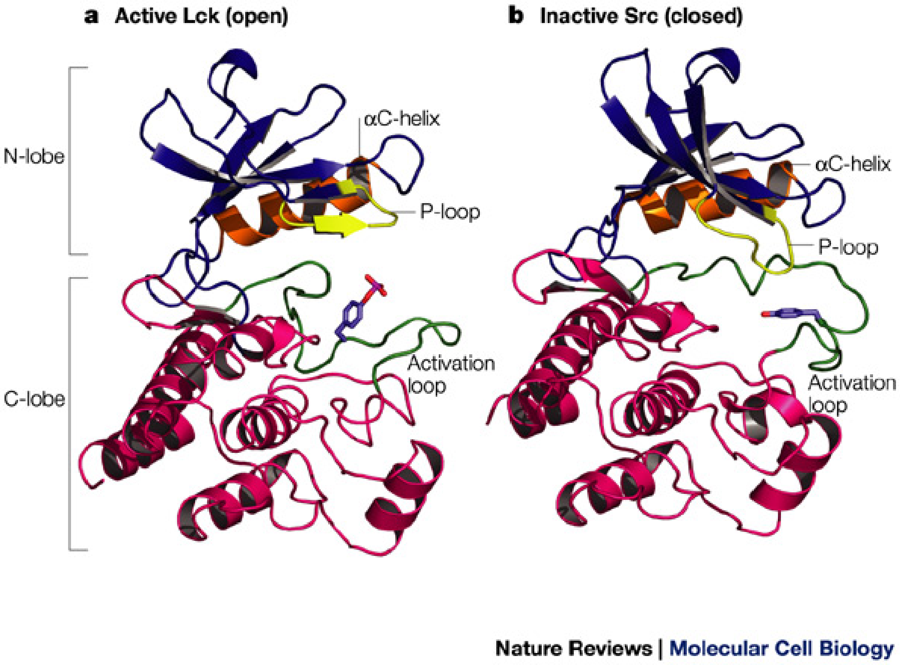


Figure 1.10: The active and inactive conformation of LCK and Src respectively. (a) Active conformation with activation loop adopting an extended conformation while it folds in the inactive c-Src kinase domain (b). Figure taken from [Hantschel and Superti-Furga, 2004].

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 diassembly of the R-spine. The rotation or shift of 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. [Tsai and Nussinov, 2013, Roskoski, 2016].

1.1.5 Kinase Inhibitors

The kinases are quite diverse in their primary sequences. 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; a β sheet containing N-lobe as well as α helix dominated C-terminal (C-lobe) with a connecting hinge region. 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 [Wu et al., 2015b]. 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 nonreversible (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. Example of such a drug in clinical trial is the AVL-292 which is a tyrosine kinase inhibitor which covalently binds to the Bruton tyrosine kinase (BTK)[Robak and Robak, 2012]. Ibrutinib targets BTK as well, while afatinib targets the gefitinib resistant EGFR [Akinleye et al., 2014].

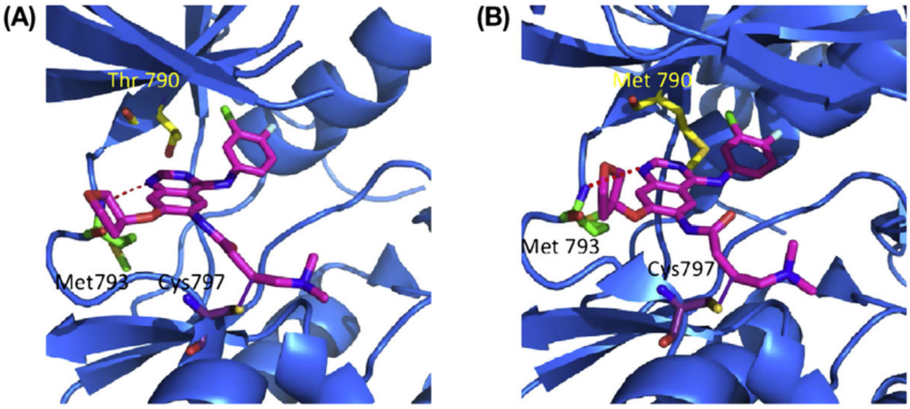


Figure 1.11: Afatinib co-crystal structure with wild-type EGFR (PDB ID: 4G5J) and mutant T790M EGFR (PDB ID:4G5P). Afatinib binds to the kinase domain in its active conformation and forms a hydrogen bond with the backbone NH of Met793 and also forms covalent interaction with the sulphur of Cys797. Figure obtained from [Hossam et al., 2016]

The 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). The 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 [Fabbro et al., 2015]. Classical examples of such approved inhibitors include gefitinib, dasatinib, erlotinib and sunitinib.

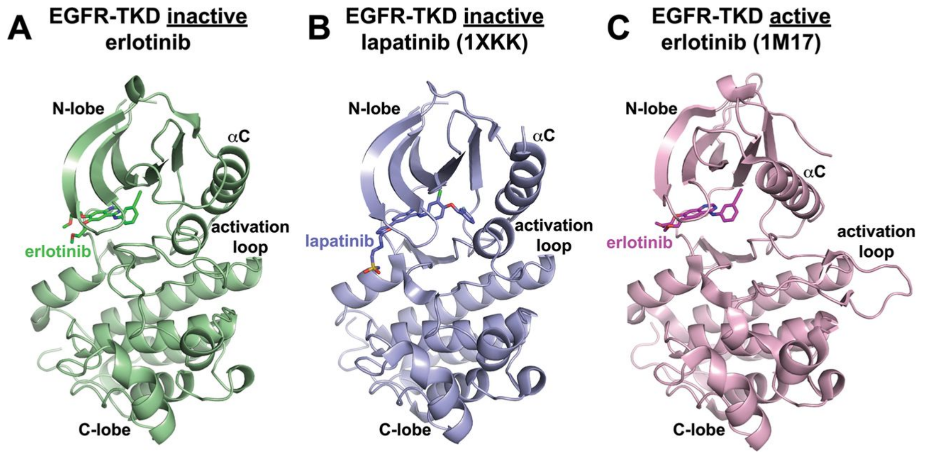


Figure 1.12: Crystal structure of EGFR tyrosine kinase domain (TKD) bound with inhibitors. (A)Erlotinib bound with EGFR-TKD in the inactive state (B) Lapatinib with inactive EGFR-TKD (1XKK) (C) Erlotinib with the active EGFR-TKD (1M17). Figure obtained from [Park et al., 2012]

The type-II inhibitor binds 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 inhibitor to lock the kinase in an inactive conformation [Cowan-Jacob et al., 2009]. The type-II are generally less promiscuous as compared to the type-I inhibitors. Examples of FDA-approved type-II kinase inhibitors include imatinib, nilotinib, and sorafenib [Fabbro et al., 2015]. The type-I and II inhibitors however face competition with the millimolar concentration of ATP *in vivo* as well as lack of selectivity due to the extensive adenosine binding cleft [Lamba and Ghosh, 2012]. There have therefore been efforts directed towards kinase inhibitors with high selectivity, high affinity and less side effect.

The 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 and 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 [Fabbro et al., 2015]. 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 advantage over the type-I and II inhibitors [Cowan-Jacob et al., 2009]. 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 [Fabbro et al., 2015].

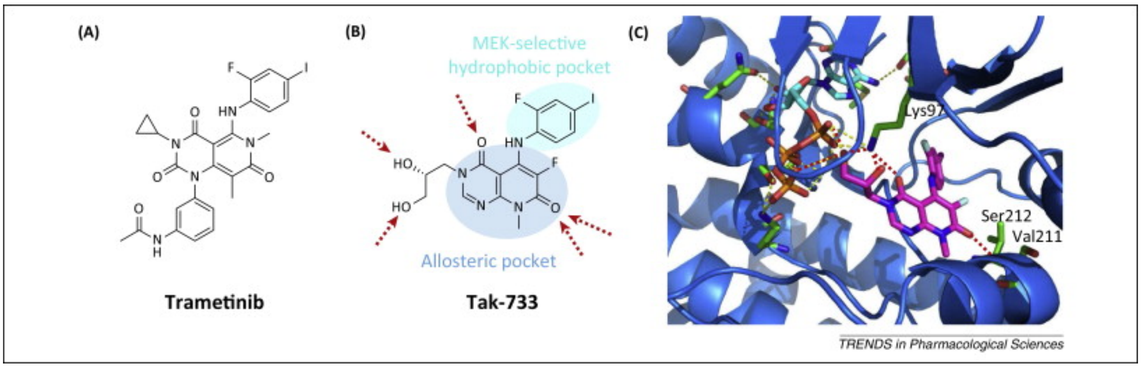

(A)
(B)
(C)
MEK-selective hydrophobic pocket
Allosteric pocket
Trametinib
Tak-733
Lys97
Ser212
Val211
TRENDS in Pharmacological Sciences

Figure 1.13: 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 [Wu et al., 2015b]

.

The 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. This clearly shows that the inhibitor interacts with both the catalytic domain and the PH domain and prevents the activation of the upstream kinase PDK1 [Barnett et al., 2005]. Other types of allosteric protein kinase inhibitors includes the type-5 which are also referred to as bivalent or bi-substrate inhibitor. 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 [Lamba and Ghosh, 2012]. 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 FDA-approval of trametinib as well as several other allosteric inhibitors that are in clinical trials [Wu et al., 2015a].

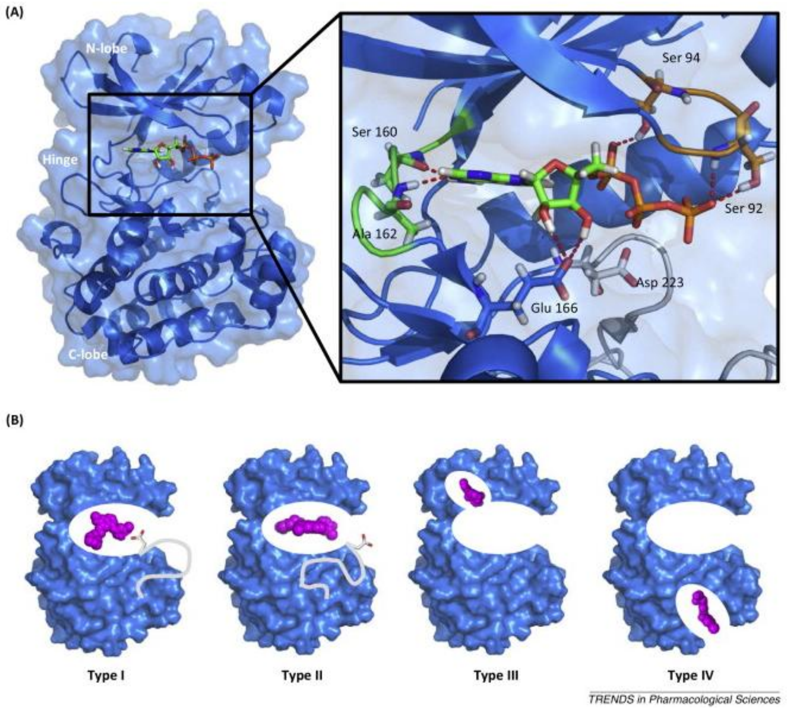


Figure 1.14: Kinase structure and the various types of inhibitor schematics. (A). Co-crystal structure of PDK1 with ATP (adenine and ribose in green, phosphate in orange) Enlarged area shows hydrogen bond in red, hinge and hinge residues in green backbone, P-loop and P-loop residues in brown-orange, Asp residue of the DFG motif and the activation loop in grey [PDB ID:4RRV, 1.41A], (B) The four types of reversible binding mode. Figure taken from [Wu et al., 2015b]

.

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 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. [Wu et al., 2015a].

1.1.6 Understanding the promiscuity of protein kinase inhibitors

Protein kinase inhibitors are generally considered promiscous because of their lack of specificity and ability to interact with several kinases and families, because of the common ATP-binding site the kinase inhibitors interact with. 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 [Hu et al., 2017]). Hu et al classified the protein kinse inhibitors into single and multiple kinase inhibitors based on the numbers of targets the PKI compounds in 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 [Hu et al., 2017].

The promiscuity associated with protein kinase inhibitors can lead to various side effects. This is because the 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 in the study evaluating the promiscuous nature of tyrosine kinase inhibitors, [Giansanti et al., 2014] using proteomics techniques, studied the promiscuity of 4 tyrosine kinase inhibitors (imatinib, dasatinib, bosutinib and nilotinib) in epidermoid carcinoma cells as a model system for skin cancer. 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 signalling pathway. The study of the promiscuity of kinases and kinase inhibitor have been well considered from experimental and computational perpectives providing selective criteria that can be used to minimize the off target propensity associated with protein kinase inhibitors.

Using computational approach, [Huang et al., 2009] mapped all the 518 human kinase sequences onto structural alignment of 116 kinases of known 3D structure. They considered the binding site of protein kinases and compared the key residues of the ATP binding sites rather than the overall kinase residues.The residues were encoded in a 9-bit fingerprint and analysed using a network approach to partition the kinases into clusters of similar fingerprints. This approach reveals the likelihood of finding selective inhibitors targeting the ATP binding site.

Furthermore, databases like the KIDFamMap [Chiu et al., 2012] provides biological insights into the selectivity of kinase inhibitors and the mechanism of binding. The database explores kinase-inhibitor families as well as kinase-inhibitor disease relationship. The database also provides KIDFamMap ″anchors″ which represents conserved interaction between kinase subsites and the moieties of the inhibitors. Thus, this creates a platform for accessing conformation, function and selectivity of kinase and kinase inhibitors.

Objectives

We have previously demonstrated the importance of using CATH-FunFams as a reasonable annotation level for drug-domain interaction and also used network analysis to associate the propensity of side effects with the network properties of these families through our druggable FunFam approach [Moya-García et al., 2017]. In this study, we are focusing on protein kinase inhibitors, a set of molecules that inhibit the activities of kinases. Using a set of publicly available protein kinase inhibitors as well as FDA-approved kinase inhibitor drugs, we studied the Functional Families of kinases and associated these FunFams with protein kinase inhibitors. We subsequently measured some network characteristics to distinguish FunFams containing proteins with side effects from others. We also explored some similarity measures to shed more lights on possible repurposing of protein kinase inhibitors to members of a given FunFam. The outline of the study is summarized below

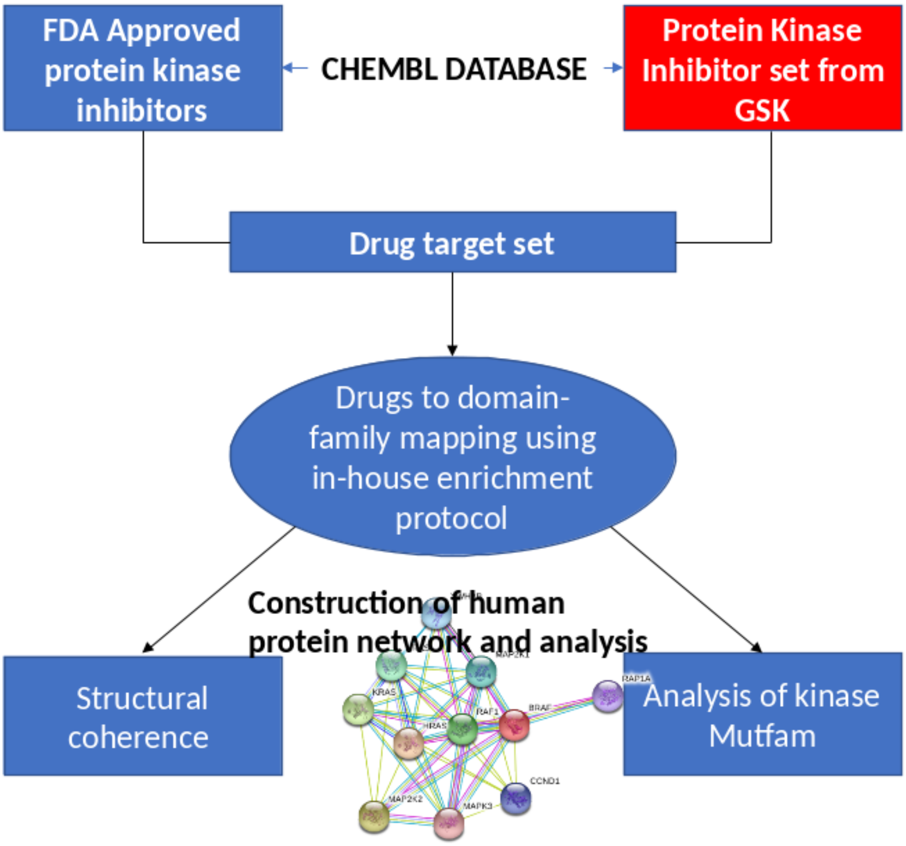

FDA Approved protein kinase inhibitors
CHEMBL DATABASE
Protein Kinase Inhibitor set from GSK
Drug target set
Drugs to domain-family mapping using in-house enrichment protocol
Construction of human protein network and analysis
Structural coherence
Analysis of kinase Mutfam
MAP2K1
RAP1A
KRAS
RAF1
BRAF
HRAS
CCND1
MAP2K2
MAPK3

Figure 1.15: Frame-work of the methodology used in this study

1.2 Materials and Methods

1.2.1 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 [Dranchak et al., 2013, Knapp et al., 2013]. These compounds have been annotated with protein kinase activity [Knapp et al., 2013] and are of various chemotypes and are openly available from the ChEMBL database (ChEMBL release 23) [Gaulton et al., 2016]. The PKIS are active against some known target kinases and can be extended to other new target kinases. We collected a subset of the PKIS that showed an inhibitory activity level above 50% as [Dranchak et al., 2013, Anastassiadis et al., 2011] had reported this threshold as appropriate for considering the inhibition of kinase catalytic activity.

Furthermore, we compiled kinase-inhibitor datasets of the FDA-approved drugs by querying ChEMBL release 23. We defined this drug target set based on the concentration with which the drug affects the protein. We considered a drug as a small molecule with therapeutic application, which has a direct binding to a single protein (ASSAY_TYPE ="B"), with a maximaum phase of development ="4" which indicate that the drug has been approved. We filtered out weak activity by considering drug-target activity stronger than 1μM and a pchembl_value$\geq$6). The ATC-code was used to filter the drug-targets such that only the protein kinase inhibitor was obtained. The ATC code classifies drug into different groups at different levels. The code "L01XE" correspond to antineoplastic drugs which are protein kinase inhibitors.

1.2.2 Network Data and Analysis

The representation of proteins on a network gives a view of the information flow and interactors for biological process. We obtained a functional protein association network for human from the STRING database version 10.0 [Szklarczyk et al., 2014]. We chose STRING database as it is widely used and oftentimes updated. STRING database generate scores to measure reliabilty of an interaction by benchmarking predicted interactions against common set of true positive associations. We filtered this data by applying a cut-off of 0.8 on the combined score of the interaction which correspond to those PPI with high reliabilty. This gave 219,608 physical interactions between 10,430 proteins. We extracted the largest connected subgraph and then computed the node centrality measures.

Centrality Measure

Centrality measures identify important nodes relative to other nodes within the network. Such measures include the degree, closeness centrality, betweenness centrality as well as the PageRank. The degree of a node is the number of connections (edges) it shares with other nodes. The betweenness centrality (BC) is the fraction of the number of shortest paths that pass through each node. The BC measures how often a node occurs on all the shortest paths between two nodes. Therefore, a node with high BC influences the flow of information in the network.

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

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 $v \in s,t$, $\sigma(s,t|v) = 0$

We also measured the topological properties of the targeted kinases and compared them to other human kinases (excluding the targeted kinases) and non-kinase proteins in a given human functional network. We defined ″Hubs″ as the top 20% of nodes with the highest degree (connections) while ″High-BC″ are those top 20% of nodes with the highest betweenness centrality. The ″Bottlenecks″ are defined as those nodes within the ″High-BC″ but excluded from ″Hubs″ i.e. those with low degree connectivity. The ″Hubs & High-BC″ are those set of nodes in the Hubs group with high betweenness centrality measure.

Dispersion Measure in a Functional Protein Network

We transformed the STRING network [Szklarczyk et al., 2014] (all edge weights) into a similarity matrix by taking its adjacency matrix. The adjacency of the full STRING network contains information about the functional association between proteins: the value in row i, column j had the STRING combined score (0-1) between protein i and protein j. This adjacency matrix has the properties of a similarity matrix and reflects the integration of the disparate protein interaction types. Based on this matrix we defined the matrix similarity of a group as the mean STRING combined score which reflects the closeness of these proteins in a protein functional network i.e. proteins with high matrix similarity forms high neighbourhood compared to low similarity which are scattered in a protein functional network.

We also used another similarity measure adapted from [Menche et al., 2015] called ”DS-Score”. The DS-Score measures the mean distance of separation of genes within a cluster. We also measured the largest connectivity component (S) of each module formed by the proteins associated with the same drug. These results were compared against random protein sets.

1.2.3 FunFam Mapping and Enrichment Analysis

Protein kinases have been divided divided into 9 groups by [Manning et al., 2002] as highlighted in the introductory section. We associated all human kinases with our functional families (groups of evolutionarily related, structurally and functionally coherent protein families). We scanned all human kinase sequences from Pfam against the in-house Pfam-FunFams library using HMMer3. Pfam-FunFam data from the Gene3D database was used as Pfam provides sequences that cover the entire kinase catalytic region, whilst CATH divides the kinases into the N and C lobe domains.

The drug targets were mapped to Pfam-FunFams and we evaluated the enrichment of the targets of a drug for Pfam-FunFam and calculated the p-value (Benjamini-Hochberg corrected for multiple testing) to deteremine whether the observed overrepresentation is statistically significant by means binomial test.

All data processing, statistical analysis and results were produced using Python and Networkx, R computing environment and the R library ggplot2.

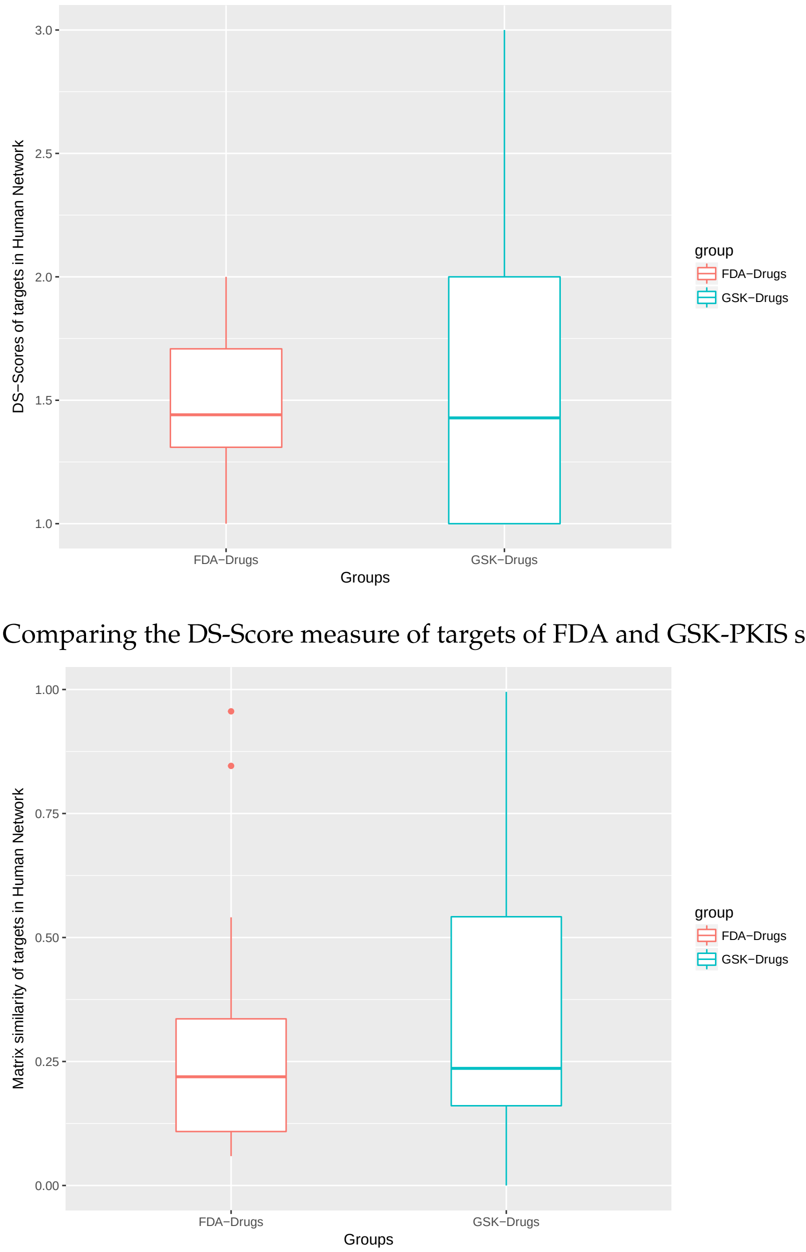
1.3 Results and Discussion

1.3.1 Protein Kinase Inhibitors set

The GSK-PKIS from ChEMBL at 50% published activity level gave a set of 205 protein kinase inhibitors that were distributed across 133 protein kinases. This covers about 60% of the entire PKIS set and is thus a reasonable dataset to consider for a network assessment and characterization of protein kinase inhibition. The FDA-approved drug set with using p-ChEMBL activity level ≥ 6 had 29 approved drugs that interact with 324 targets. The targets were filtered to exclude those that are not kinases reducing the numbers of targets to 305 kinases.

Comparison of the FDA drug-target set and the GSK-PKI drug-target

We had two different drug target sets. Both drug-target sets were selected using different criteria as they were distinct set of drugs. The GSK-PKIS were considered experimental drugs as they have not yet been approved for clinical trials but have been shown to have inhibitory properties against the kinase and could be used as probe molecules for the untargeted kinome [Dranchak et al., 2013]. Hence, we compared the network properties of the targets from both sets. We measure the network properties of these targets to distinguish the network properties of the approved and experimental drug sets. We measured the matrix similarity and the DS-Score for both drug sets.



(a) Comparing the DS-Score measure of targets of FDA and GSK-PKIS sets

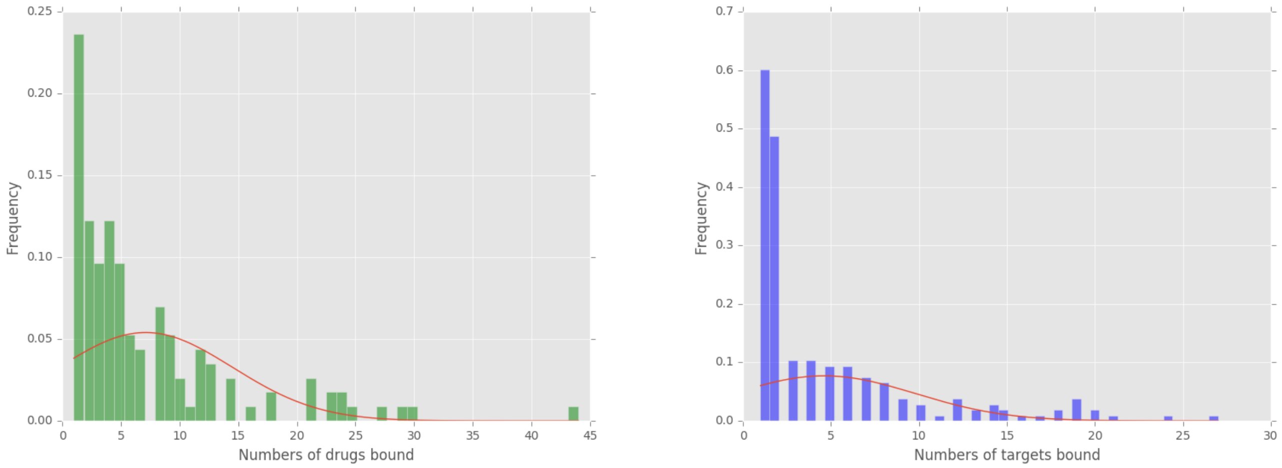
(b) Comparing the Matrix similarity measure of targets of FDA and GSK-PKIS sets

Figure 1.16: Comparison of the FDA and GSK target similarity in a functional protein network

We observed no statistically significant difference between the DS-Score and the Matrix similarity score of the FDA-targets and the GSK-PKIS targets (Mann-Whitney test, pvalue = 0.3699 and 0.07789). This implies that targets from both drug sets (FDA and GSK-PKIS) form similar network communities. Observation of the boxplot in figure 1.16 shows that GSK-set had a wider range of values compared to the FDA-set (interquartile range of both the DS-Score and Matrix similarity score). This may be because the numbers of selected drugs in the GSK is higher compared to those in the FDA and may also indicate that not all the available drugs in the GSK set would have targets that are aggregated based on our network similarity measure. Also, as reported by [Wu et al., 2015b], we found that the tyrosine kinase family are the most targeted family by the the approved drug with a coverage of over 70% represented in our drug-target dataset. We chose the GSK drug set as our ”test set” to analyse the promiscuity of protein kinase inhibitors and possible repurposing of the drugs by associating the targets to the domain families (Pfam-FunFams). We compared the observed results with the reference set (FDA-approved set) as the FDA set comprises effective marketed drugs that target protein kinases with proven therapeutic applications.

1.3.2 Target Promiscuity of Protein Kinase Inhibitors

The PKIS dataset obtained from ChEMBL database was analysed to study the association of kinases (targets) with inhibitors to understand the inherent promiscuity associated with the protein kinase inhibitors set.



(a) Promiscuity of drugs for kinases

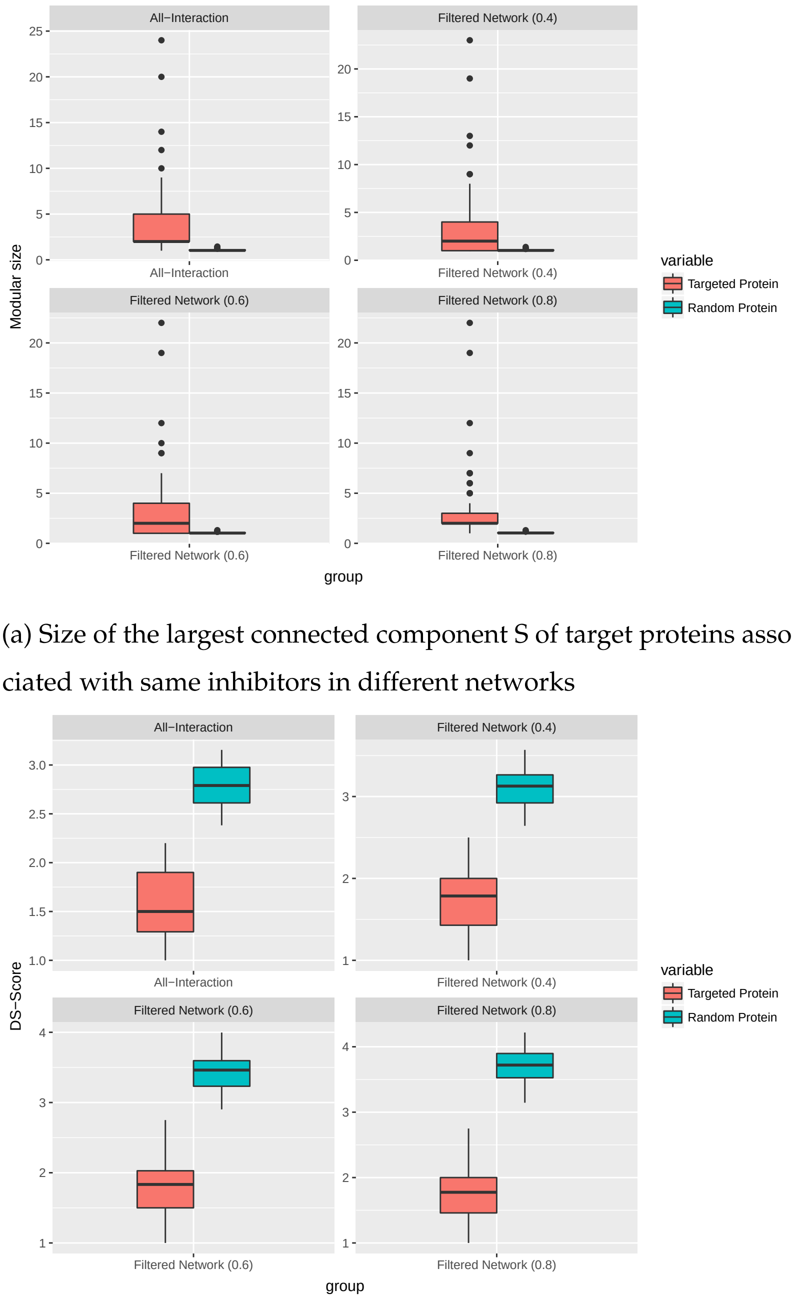
(b) Promiscuity of targets for PKIS

Figure 1.17: Distribution of kinase-drug interaction and drug-kinase interaction for the GSK-PKIS sets

Figure 1.17 shows the inherent promiscuity of the kinases for inhibitors as well as inhibitors for kinases. Protein kinase inhibitors interact with more than one kinase while some kinases could also interact with more than one protein kinase inhibitor, which indicates the non-specificity of these kinase sets. This is not surprising as the majority of the protein kinase inhibitors are the classical type-I inhibitors that binds at the ATP site directly and compete with ATP which is quite universal and important for several cellular functions. We also observed that aside the promiscuity associated with protein kinases, majority of the protein kinases are also essential proteins. We have defined essentiality as genes in organism that are required for its survival. We compiled all human essential genes as obtained from the database of Online GEne Essentiality (OGEE) [Chen et al., 2017]. We found 78% of the kinases are essential genes while 81% of the targeted protein kinases are essential. This suggest that the kinases are highly important considering the functional roles they play in organisms.

1.3.3 Network Analysis of Drug Targets in a Functional Protein Network

We tested our similarity measures using several cut off on the combined edge score of the STRING protein interaction data. This was done to see the effect of human network incompleteness on the ouput of the analysis of our protein functional network.



(a) Size of the largest connected component S of target proteins associated with same inhibitors in different networks

(b) DS-Score of the target proteins compared against random proteins in different networks

Figure 1.18: Network similarity measure compared in different protein interaction network

The result showed that there is no variation in the module size and the DS-Score of the different networks which indicate that the protein aggregation algorithm measure is not distorted by the incompleteness of protein interaction network [Menche et al., 2015]. We hence chose STRING network with an edge weight above 0.8 corresponding to protein interaction network with high confidence, to modelled our protein functional network.

The measures of degree and betweenness centrality are amongst the most profound topological properties of nodes in a protein interaction network. The bottlenecks in a protein functional network represents key connectors and show functional and dynamic properties [Yu et al., 2007]. We hypothesize that bottlenecks could be a good drug target as they are central in the network but associated with less nodes (thus less functional disintegration expected). Table 1.1 shows the comparison of some topological properties of the targeted kinases, all human kinases and all human proteins in a functional network. The proportion is a measured relative to each group as;

$$\frac{\text{Number of Hubs or bottlenecks in targeted set} \times 100}{\text{Total number proteins in targeted set}}$$

Table 1.1: Topological analysis of kinases in a functional protein network

Groups	*Hubs (%)*	*Bottlenecks (%)*	*High-BC (%)*	*Hubs & High-BC (%)*
Targeted kinases	*42.06*	*5.61*	*41.12*	*46.73*
Other-human kinases	*19.59*	*6.76*	*16.44*	*23.20*
All-human proteins	*19.19*	*11.51*	*7.47*	*18.98*

The results therefore show that only a small percentage of the targeted kinases are bottlenecks. The kinases are most likely to be hubs and they are highly central in protein functional network with high betweenness centrality (Mannwhitney test for hubs and centrality between kinases and other proteins in a network; pvalue = 2.54e-13 amd 6.617e-25). However, there is still a lot of debate on this opinion as some studies have suggested bottlenecks to be associated with side effects. For instance, in the studies by [Perez-Lopez et al., 2015], they observed that targets of drugs with side effects are better spreader of pertubation in the interactome indicating that these targets are quite central and they disrupts the interactome more as compared to the drug-targets without side effects or non-target proteins.

1.3.4 Dispersion measure of targets of kinase inhibitor in a protein network

We also investigated the behaviour of kinase inhibitor targets in human protein interaction network and measured the dispersion of the targets in network using dispersion measures such as the matrix similarity as well as the shortest path distance across all the targets of a drug (DS-Score).

We selected all the targeted proteins from ChEMBL 23 and excluded the kinases whilst we grouped all kinase drugs targets as ”Kinase Inhibitor targets”. We also consider less specific interaction between drugs and proteins as a way of classifying ”Off-targets” using a pchembl<6 compared against random protein sets.

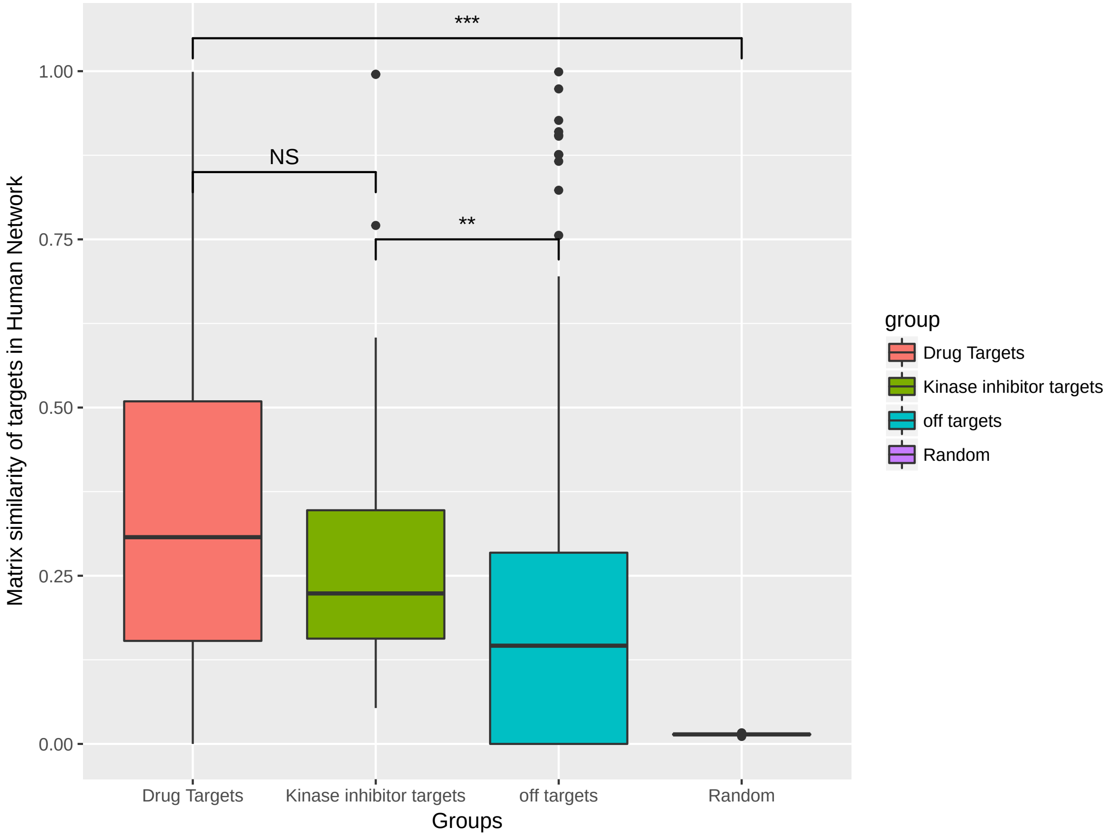


Figure 1.19: Box plot comparing the distribution of the matrix similarity of the various group against random. (NS indicate not statistically significant, while ** indicate statistical significance between pairs)

The result in figure 1.19 show that there is no difference in the observed matrix similarity of the kinase-targets and other drug targets (with pvalue = 0.9113). This observation is consistent with our previous report about drug targets having higher similarities in the protein functional network [Moya-García et al., 2017]. However when we compared the matrix similarities of the drug targets and the kinase inhibitor targets with drug off-targets, we found a statistically significant difference of the observed matrix similarities (pvalue = 2.09e-10 and 4.419e-9 for drug-targets and kinase inhibitor targets respectively).

We also compared the Matrix similarity with the ”DS-Score”. It is expected that the lower the ”DS-Score”, the more aggregated the proteins are in the network. We also used this measure to compare the targets aggregation in a given functional network.

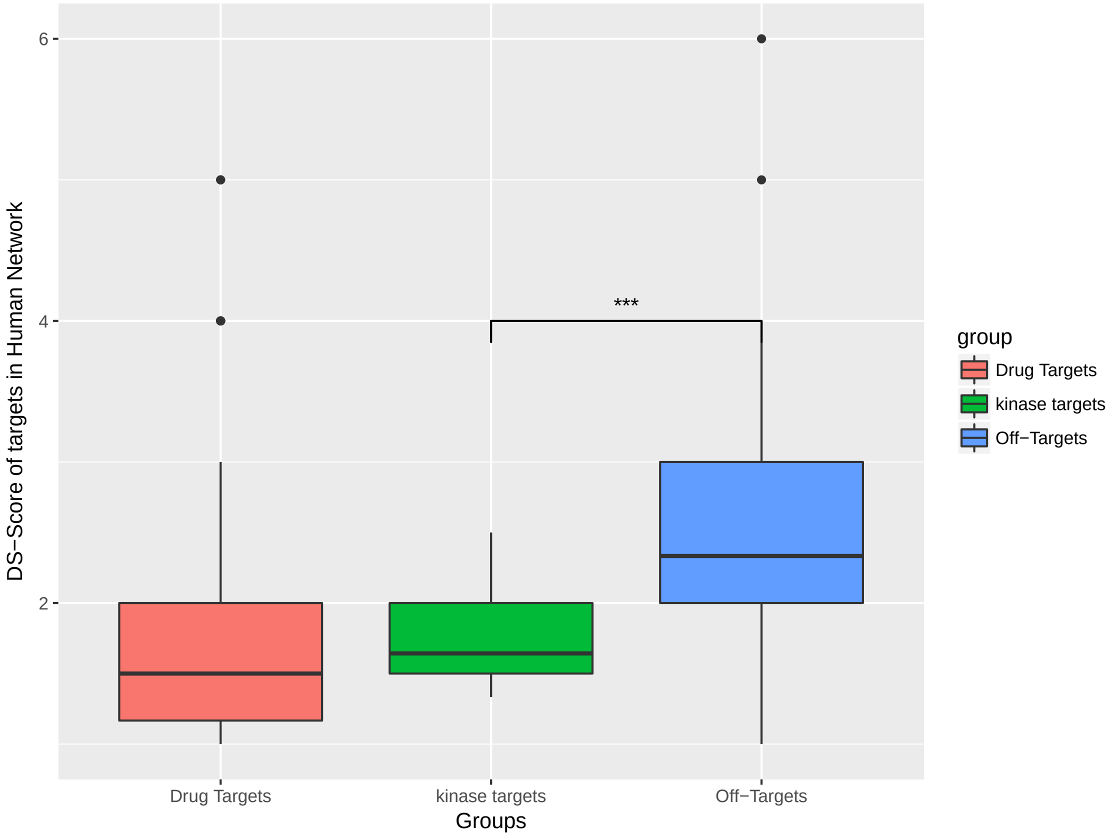


Figure 1.20: Boxplot comparing ”DS-Score” of the drug targets, kinase inhibitor targets and off-targets.

The result (figure 1.20), show that the DS-Score is higher in offtargets as compared to the drug targets as well as the kinase targets and this variation is statistically significant (pvalue = 7.035e-8 and 0.0001542). There is however no statistically significant difference between the DS-Score of the drug targets and kinase inhibitor targets (p=0.1447). We have therefore shown using two different measures ”matrix similarity” and ”DS-Scores” that protein targets are more clustered in protein interaction network compared to offtargets or random proteins.

1.3.5 FunFam Mapping and Enrichment Analysis

About 1300 human protein kinases sequences obtained from Pfam were clustered at 90% sequence identity to remove protein isoforms and obtain a unique non-redundant human kinase set using CD-HIT clustering algorithm [Li and Godzik, 2006]. The 741 non-redundant kinases obtained were distributed amongst 130 Pfam-FunFams which map to 16 of the 35 Pfam clans. We mapped the Pfam-FunFam to the human kinome tree and found that our domain-family classification correspond well to the various groups with no overlaps between the groups identified by [Manning et al., 2002] as shown in figure 1.21 below.

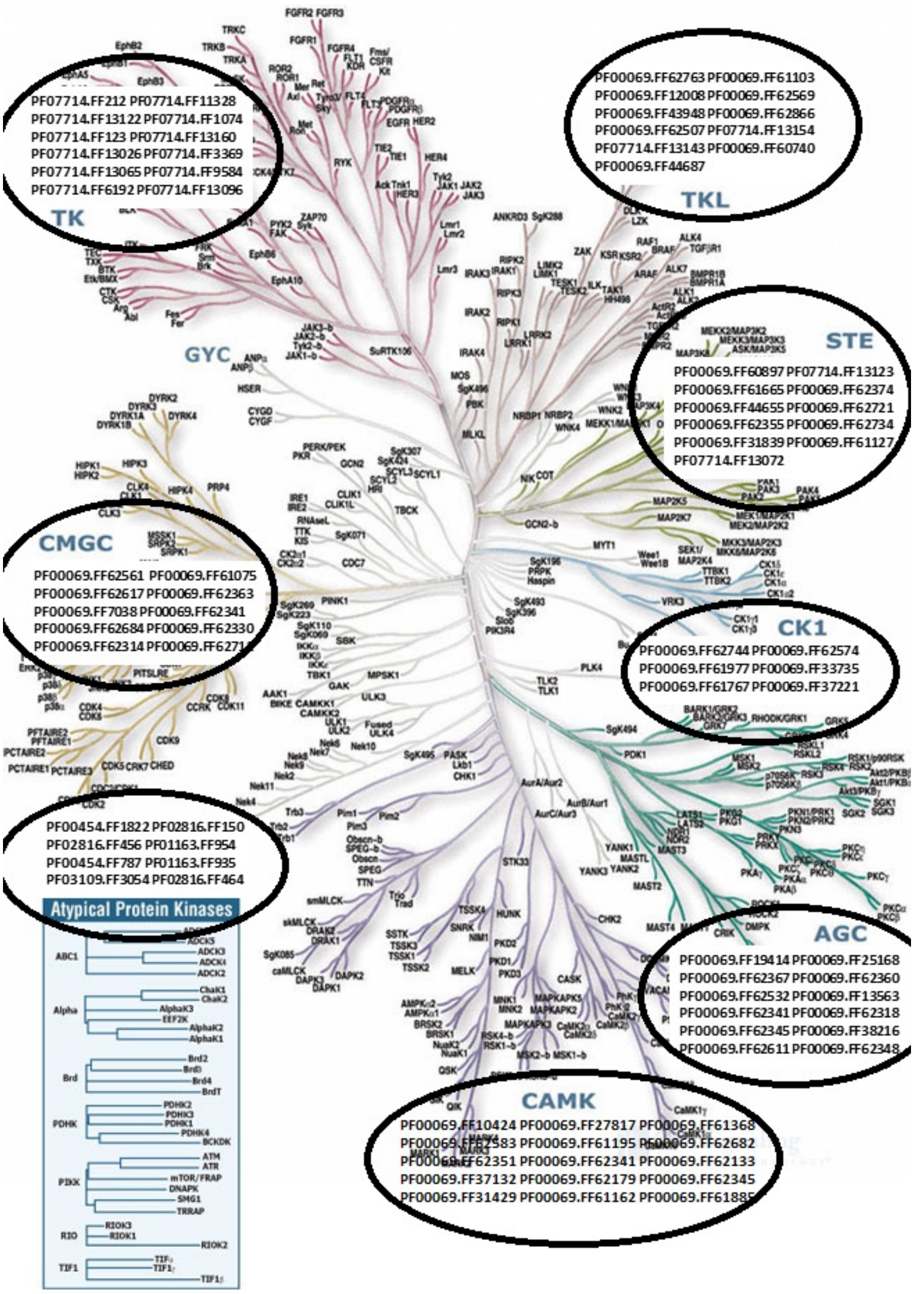

TK
PF07714.FF212 PF07714.FF11328
PF07714.FF13122 PF07714.FF1074
PF07714.FF123 PF07714.FF13160
PF07714.FF13026 PF07714.FF3369
PF07714.FF13065 PF07714.FF9584
PF07714.FF6192 PF07714.FF13096
TKL
PF00069.FF62763 PF00069.FF61103
PF00069.FF12008 PF00069.FF62569
PF00069.FF43948 PF00069.FF62866
PF00069.FF62507 PF07714.FF13154
PF07714.FF13143 PF00069.FF60740
PF00069.FF44687
STE
PF00069.FF60897 PF07714.FF13123
PF00069.FF61665 PF00069.FF62374
PF00069.FF44655 PF00069.FF62721
PF00069.FF62355 PF00069.FF62734
PF00069.FF31839 PF00069.FF61127
PF07714.FF13072
GYC
CMGC
PF00069.FF62561 PF00069.FF61075
PF00069.FF62617 PF00069.FF62363
PF00069.FF7038 PF00069.FF62341
PF00069.FF62684 PF00069.FF62330
PF00069.FF62314 PF00069.FF6271
CK1
PF00069.FF62744 PF00069.FF62574
PF00069.FF61977 PF00069.FF33735
PF00069.FF61767 PF00069.FF37221
PF00454.FF1822 PF02816.FF150
PF02816.FF456 PF01163.FF954
PF00454.FF787 PF01163.FF935
PF03109.FF3054 PF02816.FF464
Atypical Protein Kinases
ABC1
Alpha
Brd
PDHK
PIKK
RIO
TIF1
AGC
PF00069.FF19414 PF00069.FF25168
PF00069.FF62367 PF00069.FF62360
PF00069.FF62532 PF00069.FF13563
PF00069.FF62341 PF00069.FF62318
PF00069.FF62345 PF00069.FF38216
PF00069.FF62611 PF00069.FF62348
CAMK
PF00069.FF10424 PF00069.FF27817 PF00069.FF61368
PF00069.FF62583 PF00069.FF61195 PF00069.FF62682
PF00069.FF62351 PF00069.FF62341 PF00069.FF62133
PF00069.FF37132 PF00069.FF62179 PF00069.FF62345
PF00069.FF31429 PF00069.FF61162 PF00069.FF61885

Figure 1.21: Pfam-FunFam distribution across the human kinome.The kinome tree was adapted from [Manning et al., 2002]. The mapping of Pfam-FunFam to the group "Others" is not shown in this figure

We then mapped the drug-targets of the PKIS obtained from ChEMBL to the Pfam-FunFams at 50% inhibition level with an affinity of 0.1μM and found that they are associated with 37 of the 130 Pfam-FunFams which represents about 30% of the human kinase

Pfam-FunFams obtained. We speculated that this 30% represents the kinases most targeted by the pharmaceutical industry based on their relevance to human disease and those most studied as the current research in kinase therapeutics indicates that only 10-15% of the kinases are being targeted [Li et al., 2016, Elkins et al., 2016]. We also observed this ratio from our kinase-inhibitor set being used in this study. We also compared numbers of shared targets and families in the FDA and GSK drug set as shown in table 1.2.

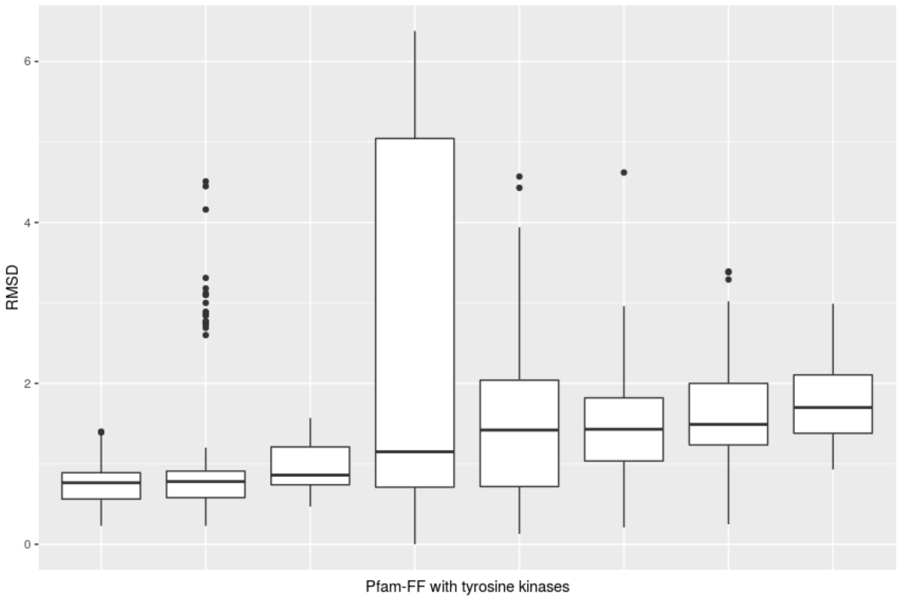
Table 1.2: Pfam-FunFam families shared by the GSK and FDA dataset

Pfam-FunFams	GSK-drugs	FDA-drugs	GSK-targets	FDA-targets	Shared-targets
PF00069.FF62355	1	2	2	16	2
PF00069.FF62314	4	1	3	6	2
PF00069.FF62345	5	1	7	12	5
PF07714.FF212	9	1	3	3	3
PF07714.FF13154	2	2	1	3	1
PF07714.FF13026	17	4	25	36	25
PF07714.FF13065	4	5	8	17	8
PF07714.FF123	23	7	5	5	5

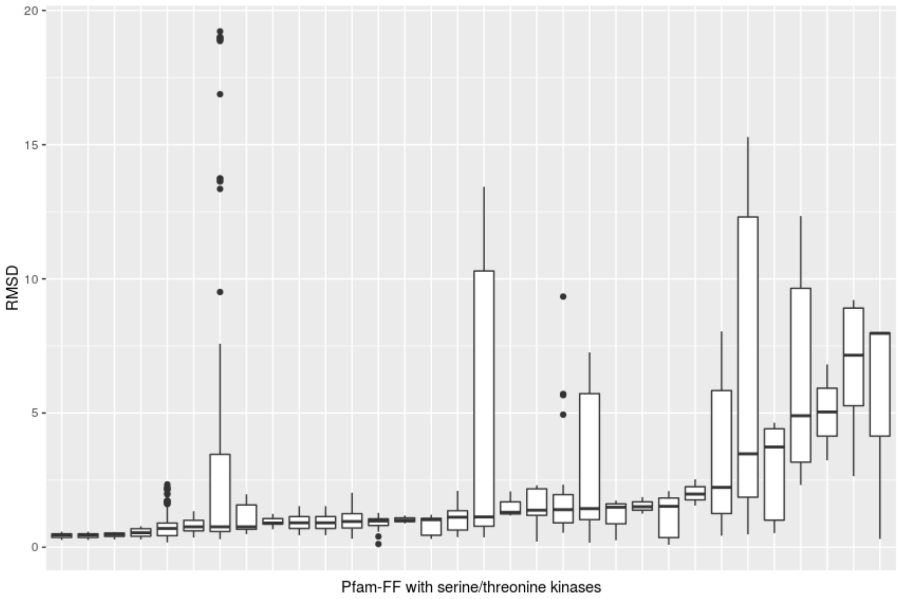
Table 1.2 shows the list of the targets shared by the Pfam-FunFams that are shared by the GSK and FDA approved drugs. Although, the numbers of FDA approved drugs are quite small compared to the experimental GSK-PKIS, we however observed a higher number of targets associated with them. This creates room for possible repurposing of the experimental drugs (GSK) to other targets within the same family.

1.3.6 Structural conservation of human kinase relatives in the Pfam-FunFams

We measured how structurally conserved the kinases are across a given Pfam-FunFam. The human relatives of the Pfam-FunFam were mapped to structure by using the SIFT mapping of UniProt sequences to PDB, while the domain regions were specified using Pfam. The structures were then evaluated for structural conservation by measuring RMSD following the pairwise alignment by the SSAP algorithm [Orengo and Taylor, 1996] and the superposition by ProFit [Martin, 2009]. The groups were divided into the tyrosine-kinase Pfam-FunFams and serine/threonine Pfam-FunFams. The distribution of the RMSD across the two Pfam-FunFam groups is shown in figure 1.22 below.



(a) RMSD across the Tyr Pfam-FunFam



(b) RMSD across the Ser/Thr Pfam-FunFam

Figure 1.22: Structural coherence and conservation of the Pfam-FunFam kinases measured using the SSAP-algorithm

Figure 1.22 suggest significant structural conservation of relatives in these families. As for many comparisons, the observed RMSD was below 2 in both the tyrosine kinases as well as the serine/threonine kinase FunFams. This supports the overall view of structural conservation of the kinases in the literature [Manning et al., 2002, Taylor and Kornev, 2011, Elkins et al., 2016, Roskoski, 2016]. However, in some Pfam-FunFams pairs of relatives have a high RMSD score which tends to distort the overall RMSD measure, indicating lack of structural coherence of relatives in these FunFams.

A deeper insight into one of such families was carried out using the Pfam-FunFam PF07714.FF13122 which gave an overall RMSD below 2A but contains outliers having RMSD as high as 5A. The reason for this could be attributed to the multidomain architecture of the relatives of the Pfam-FunFam as majority of the relative in this protein have additional SH2-domain while others either have an immunoglobulin-domains or no additional domain.

1.3.7 Enrichment test of Pfam-FunFams associated with drug targets

The Published Kinase Inhibitor Sets (PKIS) was identified as the chemical starting point for probing orphan kinases. The study by [Anastassiadis et al., 2011] illustrated the utility of these compounds for developing selective inhibitors against untargeted kinases LOK and SLK. Thus, the use of domain families could help increase the coverage of potential targets of the kinase-inhibitor set as the targeted kinases are about 10-15% . We used our FunFam-target enrichment analysis protocol (see Method section) to test for the overrepresentation of drugs in our FunFams.

This potocol uses multiple testing for correction followed by binomial test to determine the most appropriate FunFam with which a Kinase inhibitor associates. We found that 109 PKIS-drugs were overrepresented in 30 Pfam-FunFam at p-value $\leq$ 0.05.Figure 1.23 shows the distribution of these 30 FunFams and the numbers of drugs they are associated with.

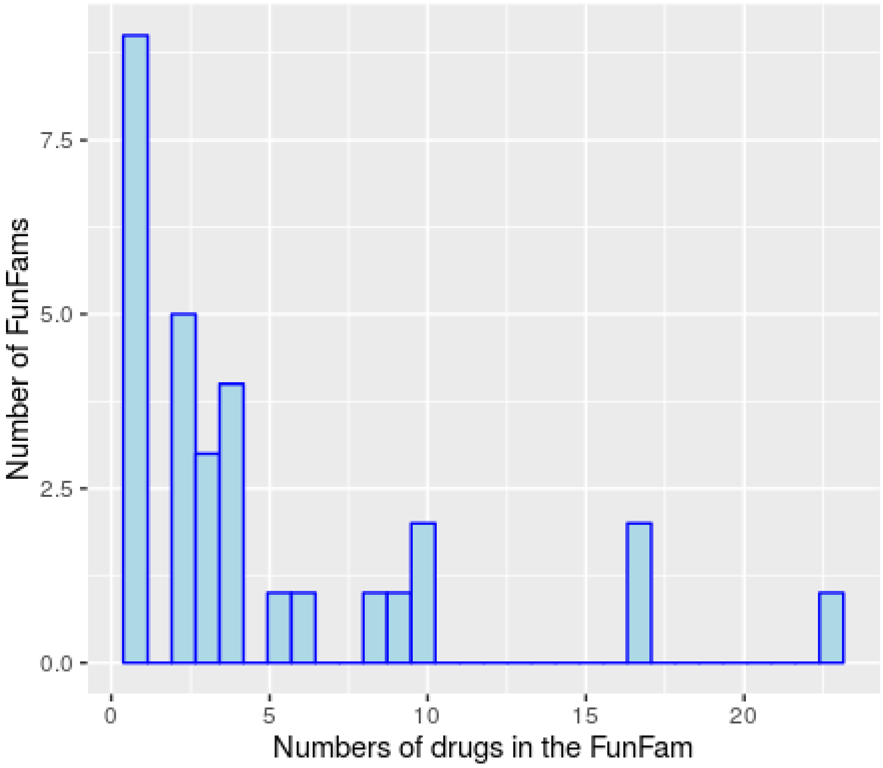


Figure 1.23: Distribution of drugs associated with FunFams

As shown in figure 1.23, we observed that 70% of the enriched Pfam-FunFams were associated with more than 2 kinase inhibitors from the PKIS set. This enriched Pfam-FunFam dataset was used for further analysis of the protein-kinase inhibitor targets on protein-protein interaction network.

The Pfam-FunFam-drug interaction was represented in network for visualization in Cytoscape as shown in figure 1.24.

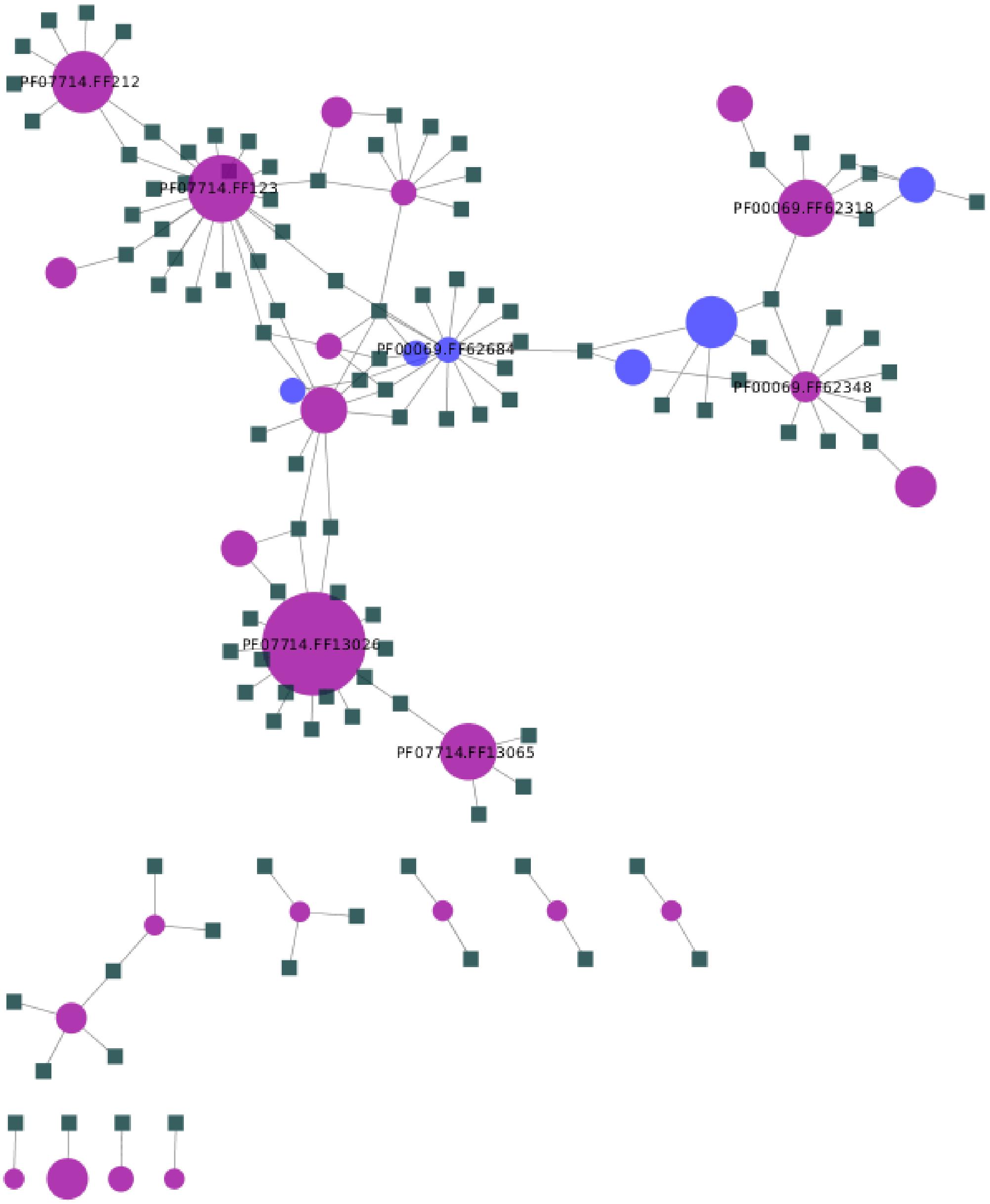


Figure 1.24: Pfam-FumFam-drug interaction network. In this network, the green coloured square node represents the CheMBL protein kinase inhibitors while the purple coloured circle nodes are the Pfam-FunFam whose relative are structurally coherent with a mean RMSD$\leq$ 2 while the blue coloured circle are Pfam-FunFam with RMSD value $\geq$ 2.5. The size of each node reflects the numbers of targets (relatives) in each family. Also labelled are some families with relatives $\geq$ 5 and interacting with at least 5 drugs.

1.3.8 Dispersion of the relatives of Pfam-FunFams in a Protein Functional Network

Following our initial hypothesis that FunFams whose relatives aggregate in a network neighbourhood are likely to be enriched with potential targets and free of off-targets, we assessed the matrix similarity of all the Pfam-FunFam kinases. The matrix similarities have a score ranging from 0-1 with 1 indicating a highly aggregated families, i.e. close proximity in the network and 0 indicate higly dispersed.

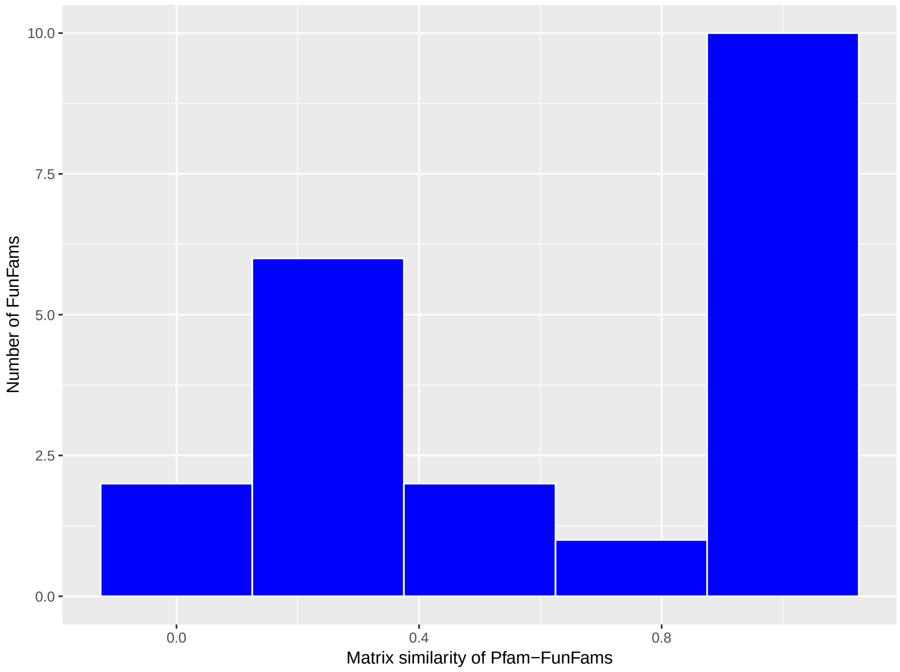
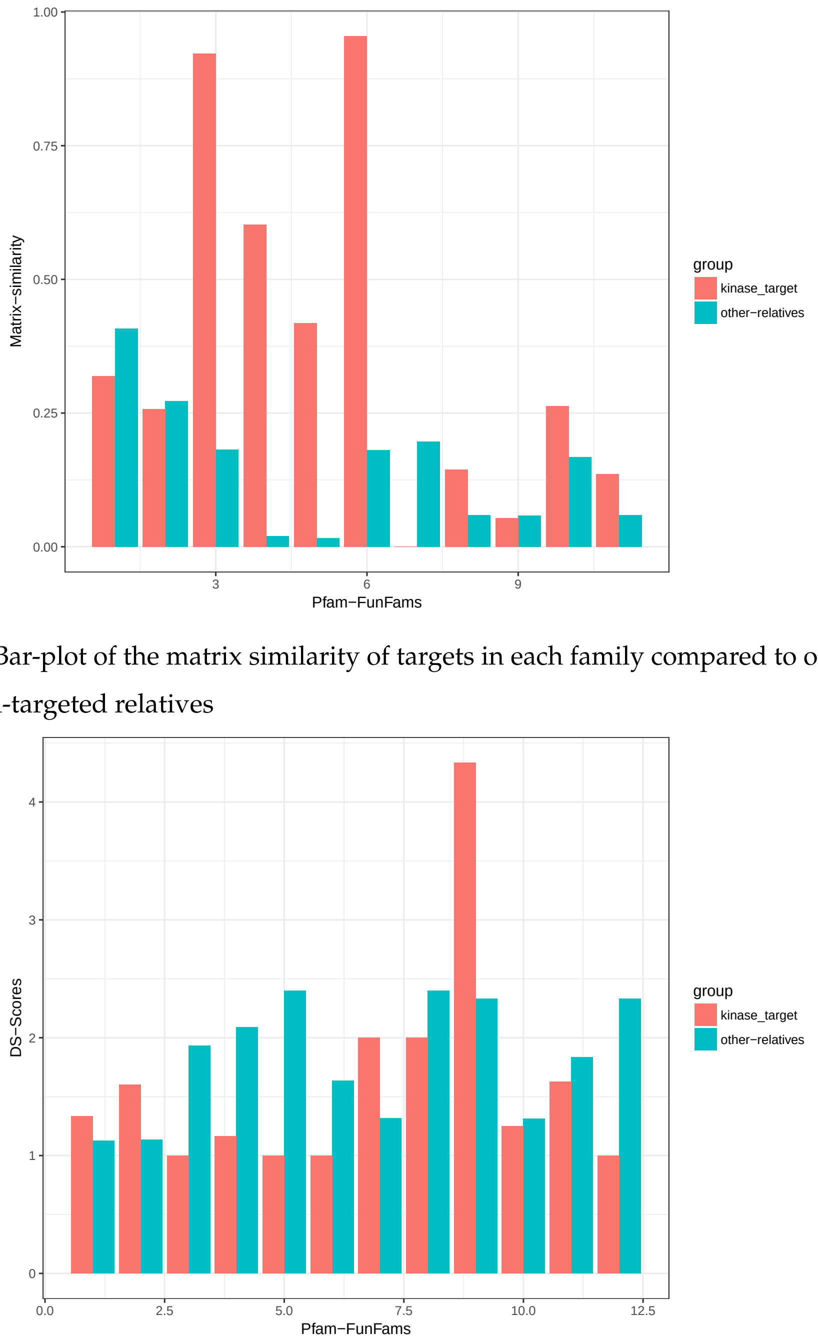


Figure 1.25: Distribution of the matrix similarity measure across the Pfam-FunFam.

The result in figure 1.25 shows the distribution of the mean matrix similarity of the relatives of the 30 enriched Pfam-FunFam in the functional protein network. The mean matrix similarity of the Pfam-FunFam family has varying similarity measure in the protein network with some families showing relatives that are more aggregated than others. Overall, majority of the members of the Pfam-FunFam are aggregated in the human functional protein network.

We then compared the matrix similarity of targeted kinases in each FunFam against other relatives of the same family to compare the network properties of targets and other non-targeted relatives of the same FunFam.



(a) Bar-plot of the matrix similarity of targets in each family compared to other non-targeted relatives

(b) Bar-plot of the DS-Scores of targets in each family compared to other non-targeted relatives

Figure 1.26: Similarity measure across the Pfam-kinase family measured using the matrix similarity and DS-Scores

As shown in figure 1.26, targeted kinases have a higher matrix similarity compared to other relatives of the Pfam-FunFam. However, some families shows non-targeted relatives that are not widely dispersed in the network and could be focused upon for novel targets to consider for repurposing of the FunFam-associated drug. Similar observation was also found when a different measure (DS-Score) was used which indicate that majority of the kinase target are aggregated in the protein functional network with lower DS-Score as we have earlier shown that the lower the DS-Score, the more aggregated a set of proteins are in the network [Menche et al., 2015]

1.3.9 Structural coherence of the binding site of the enriched Pfam-FunFam relatives

The work published by [Elkins et al., 2016] reported some solved structures for the binding of inhibitors from the PKIS with some kinases. Crystal structures of the inhibitor (CHEMBL237571) bound to the lymphocyte-oriented kinase (LOK) in the inactive DFG-out state (PDB-ID:4USD) and active DFG-in state (PDB-ID: 4USE) have been deposited in PDB. Using this example, we examined the binding site conservation across members of the Pfam-FunFam this protein belongs to. Data on the residues involved in binding was obtained from NCBI IBIS resource [Shoemaker et al., 2011].

The target for this inhibitor belongs to the Pfam-FunFam PF00069.62355 (a STE-group kinase) that has about 80 relatives. Using the SIFTS mapping of UniProt-sequences to PDB structures, we were able to identify 15 members (about 18%) of this family with PDB structures. In case of multiple structures of the same kinase, firstly, we find an unbound structure, or if there are no unbound structures then we chose the best resolved structure for the particular kinase. We structurally superposed all the members of this family. This family was found to be structurally coherent as we used ProFit algorithm guided by SSAP alignnment and obtained an average RMSD of 1.11 ± 0.48.

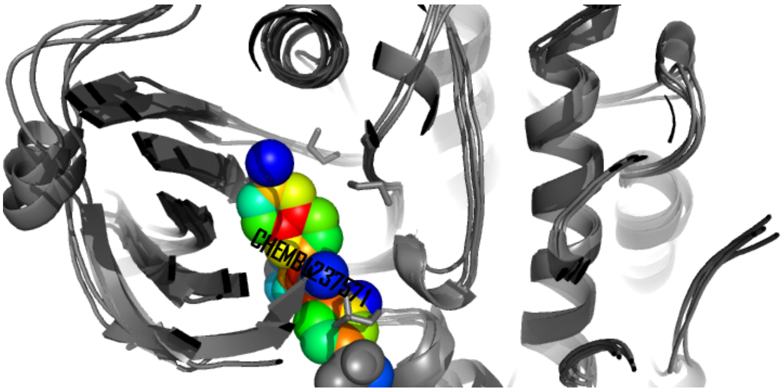


Figure 1.27: Structural alignment and superposition of the relatives in Pfam-FunFam (PF00069.62355) based on the alignment of the binding region. The inhibitor is shown as a rainbow coloured sphere.

1.3.10 Network analysis of MutFams enriched with kinases

We considered the network properties of mutationally enriched FunFams (MutFams). These families are enriched in cancer mutations (taken from the COSMIC database [Bamford et al., 2004]). We compared the network properties of MutFams with the FunFams enriched in non-disease associated neutral variations (mutation data taken from HumVar [Capriotti et al., 2006]). This approach was taken to find out whether the kinases linked to diseases are more dispersed in the network making it hard to target them with drugs.

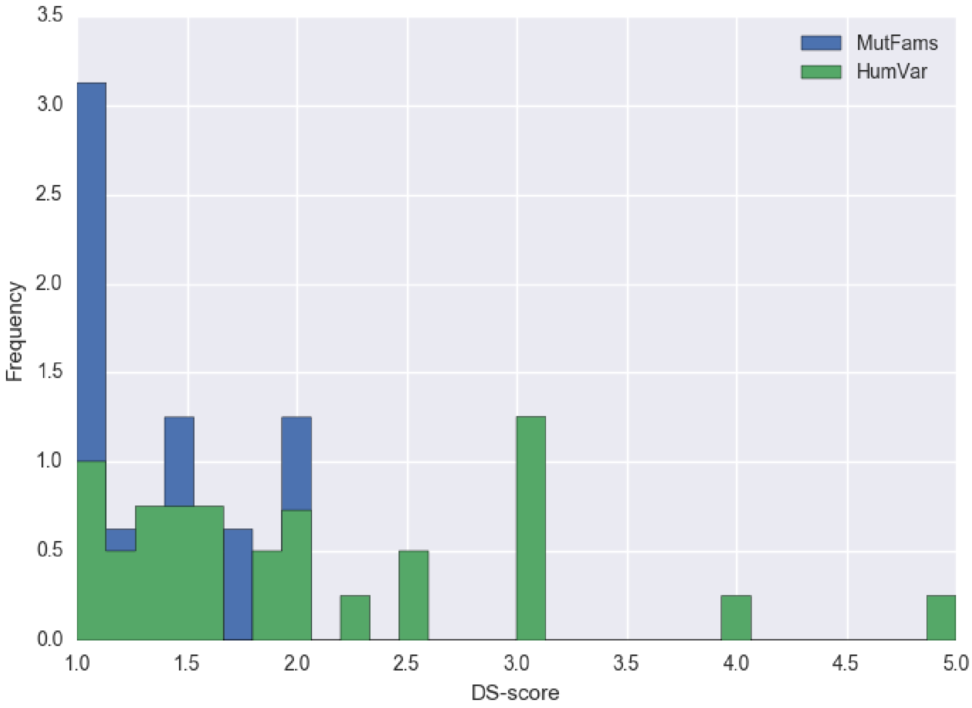


Figure 1.28: DS-measure of the MutFam in comparison with the HUMVAR

Figure 1.28 shows a plot of the distribution of the DS-score between MutFams and HumVar. There is a significant difference between these two sets of genes as the relatives of MutFams are highly clustered in the human functional network compared to the relatives in FunFams ($P = 0.00645$). MutFams therefore provides a reasonable annotation of disease genes and based on this network analysis, these families could be targeted by inhibitors as their relatives are not dispersed in the protein functional network.

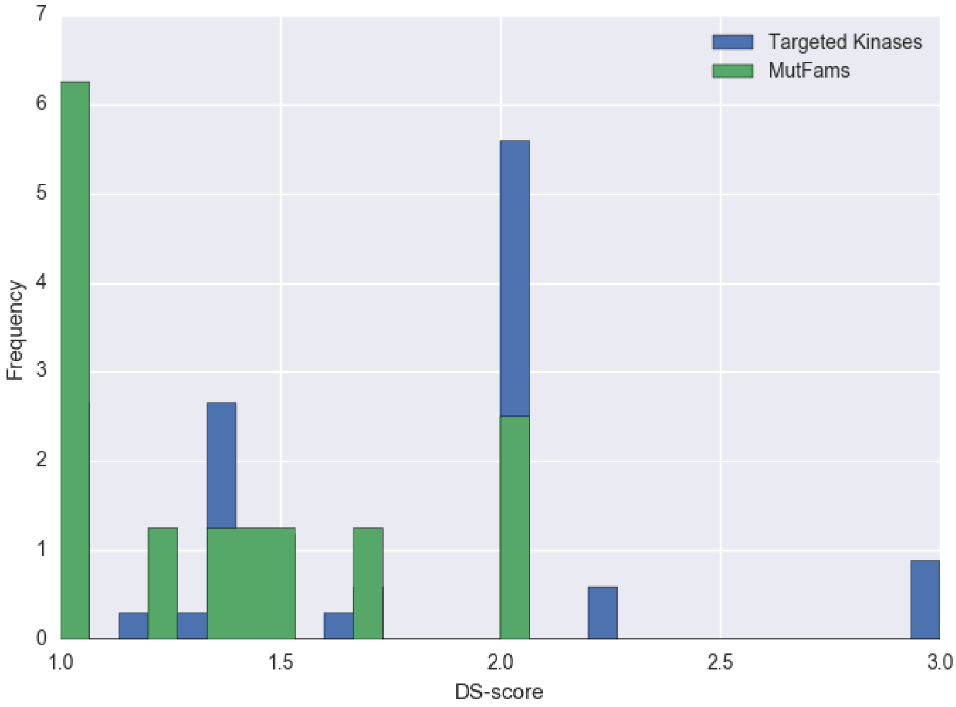


Figure 1.29: DS-measure of the MutFam in comparison with the targeted kinases

The DS-score of the MutFams were compared with the targeted kinases (figure 1.29), these two sets of genes tend to be clustered in the similar way as we observe no difference in the DS-score of the MutFams and the targeted kinases ($P = 0.0191$). This indicate that the targeted kinases share similar network characteristics as the mutated sets of genes that are implicated in human diseases. These MutFams are therefore potential therapeutic targets that could be harnessed and considered for therapeutic purposes.

1.4 Conclusion

In this study, we considered the dispersion of kinases targeted by protein kinases inhibitors (GSK-PKIS) in protein functional network to understand the inherent promiscuity associated with kinase inhibitors. We also associated the PKIS with domain families (Pfam-FunFam) and showed that the domain-families behaved similarly in the network as the drug-targets. We considered the MutFams which are mutationally enriched family and found that they behave in similar manner as the targeted kinases. We have therefore shown that there is inherent promiscuity in kinase inhibitors as well as the kinases being targeted. However, there is possibility to selectively filter the kinase inhibitors by considering the network properties of the kinases as well as the domain family it is associated with. Families with relatives that are more aggregated in the network could be suggested as a better target as compared to those dispersed in the network and this also open opportunities for repurposing of drugs when the domain families is focused on.

1.5 Future Plan

Having considered the network properties of protein kinases inhibitors in a functional protein network, we intend to extend this work by considering the following

- The effect of protein kinase inhibitors on CATH-FunFams as we are currently generating the CATH-FunFams for kinases
- Analysis of the MutFams enriched with kinases for potential drug targets
- Analysis of disease network for the possibility of repurposing some drugs to disease genes using the functional families of such disease associated proteins. Bladder cancer will be used as a prototype and extending this approach to other diseases.

1.6 Appendix

Table 1.3: The list of Pfam-FunFam (PF00069.62355) relatives with structural information

UniProt-ID	PDB-ID	Resolution (Å)
O95819	4u3y	1.45
Q9P289	3ggf	2.35
Q9P286	2f57	1.80
O96013	2j0i	1.60
Q8IVH8	5j5t	2.85
Q9Y6E0	3a7f	1.55
Q13153	1yhv	1.80
Q9UKE5	2x7f	2.80
Q9H2G2	2j51	2.10
Q9NQU5	2c30	1.60
Q99759	2o2v	1.83
Q9Y2U5	5ex0	2.70
Q13177	3pcs	2.86
O00506	2xik	1.97
O94804	2j7t	2.00

Table 1.4: The mutfam classes and their representation in Pfam-FunFams with the similarity measure in the protein functional network

Cancer Types	CATH-FunFam	Pfam-FunFam	%overlap	No of Drugs	matrix-sim	DS-score
LGG	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
LGG	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
BLCA	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
BRCA	3.30.200.20.FF2866					
BRCA	2.30.29.30.FF22238	PF00069.FF25168	14.2	4	0.319	1
BRCA	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
BRCA	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
COAD	3.30.200.20.FF64824					
COAD	1.20.120.330.FF23932					
COAD	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
COAD	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
COAD	1.10.510.10.FF79298	PF00069.FF62569	100		0.584709	1
COAD	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
COAD	1.10.510.10.FF78966	PF07714.FF3369	35.7	1	0.097	
COAD	1.10.510.10.FF79140	PF07714.FF13122	100		0.220333	1
GBM	1.10.510.10.FF79478	PF00069.FF61939	45.5		0.694	1
GBM	3.30.505.10.FF4305				0.509667	
GBM	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
GLI	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
GLI	1.10.510.10.FF79478	PF00069.FF61939	45.5		0.694	1
GLI	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.509667	1
GLI	3.30.505.10.FF4305					
GLI	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
KIRC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
LAML	1.10.510.10.FF78745	PF07714.FF13026	53.1	17	0.184	1.2
LIHC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
LIHC	3.30.60.20.FF5564	PF00069.FF62318	23.9	6	0.3411	1.17
LUAD	3.30.200.20.FF1240					
LUAD	3.30.200.20.FF64824					
LUAD	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
LUAD	1.10.510.10.FF79228	PF00069.FF62599	71.4		0.588333	1.5
LUSC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
PAAD	1.10.510.10.FF78763	PF00069.FF62351	43.5	1	0.155692	1.5
READ	1.10.510.10.FF78531	PF07714.FF13154	47.1		0.846	1.2
READ	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
READ	1.10.510.10.FF78946	PF00069.FF62345	43.8	5	0.4562	1.72
SKCM	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
THCA	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
UCEC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5

Bibliography

[Akinleye et al., 2014] Akinleye, A., Furqan, M., and Adekunle, O. (2014). Ibrutinib and indolent b-cell lymphomas. Clinical Lymphoma Myeloma and Leukemia, 14(4):253–260.

[Anastassiadis et al., 2011] Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H., and Peterson, J. R. (2011). Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nature biotechnology, 29(11):1039–1045.

[Bamford et al., 2004] Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., et al. (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. British journal of cancer, 91(2):355.

[Barnett et al., 2005] Barnett, S. F., Defeo-Jones, D., Sheng, F., Hancock, P. J., Kathleen, M., Jones, R. E., Kahana, J. A., LEANDER, K., MALINOWSKI, J., McAVOY, E. M., et al. (2005). Identification and characterization of pleckstrin-homology-domain-dependent and isoenzyme-specific akt inhibitors. Biochemical Journal, 385(2):399–408.

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

[Capriotti et al., 2006] Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics, 22(22):2729–2734.

[Chen et al., 2017] Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2017). Ogee v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic acids research, 45(D1):D940–D944.

[Chiu et al., 2012] Chiu, Y.-Y., Lin, C.-T., Huang, J.-W., Hsu, K.-C., Tseng, J.-H., You, S.-R., and Yang, J.-M. (2012). Kidfammap: a database of kinase-inhibitor-disease family maps for kinase inhibitor selectivity and binding mechanisms. Nucleic acids research, 41(D1):D430–D440.

[Cowan-Jacob et al., 2009] Cowan-Jacob, S. W., Möbitz, H., and Fabbro, D. (2009). Structural biology contributions to tyrosine kinase drug discovery. Current opinion in cell biology, 21(2):280–287.

[Dranchak et al., 2013] Dranchak, P., MacArthur, R., Guha, R., Zuercher, W. J., Drewry, D. H., Auld, D. S., and Inglese, J. (2013). 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.

[Duong-Ly and Peterson, 2013] Duong-Ly, K. C. and Peterson, J. R. (2013). The human kinome and kinase inhibition. Current protocols in pharmacology, pages 2–9.

[Elkins et al., 2016] Elkins, J. M., Fedele, V., Szklarz, M., Azeez, K. R. A., Salah, E., Mikolajczyk, J., Romanov, S., Sepetov, N., Huang, X.-P., Roth, B. L., et al. (2016). Comprehensive characterization of the published kinase inhibitor set. Nature biotechnology, 34(1):95.

[Fabbro et al., 2015] Fabbro, D., Cowan-Jacob, S. W., and Moebitz, H. (2015). Ten things you should know about protein kinases: Iuphar review 14. British journal of pharmacology, 172(11):2675–2700.

[Foreman et al., 2010] Foreman, J. C., Johansen, T., and Gibb, A. J. (2010). Textbook of receptor pharmacology. CRC press.

[Gaulton et al., 2016] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. (2016). The chembl database in 2017. Nucleic acids research, 45(D1):D945–D954.

[Giansanti et al., 2014] Giansanti, P., Preisinger, C., Huber, K. V., Gridling, M., Superti-Furga, G., Bennett, K. L., and Heck, A. J. (2014). 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.

[Hanks and Hunter, 1995] Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. The FASEB journal, 9(8):576–596.

[Hantschel and Superti-Furga, 2004] Hantschel, O. and Superti-Furga, G. (2004). Regulation of the c-abl and bcr-abl tyrosine kinases. Nature reviews. Molecular cell biology, 5(1):33.

[Hossam et al., 2016] Hossam, M., Lasheen, D. S., and Abouzid, K. A. (2016). Covalent egfr inhibitors: Binding mechanisms, synthetic approaches, and clinical profiles. Archiv der Pharmazie, 349(8):573–593.

[Hu et al., 2017] Hu, Y., Kunimoto, R., and Bajorath, J. (2017). 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.

[Huang et al., 2009] Huang, D., Zhou, T., Lafleur, K., Nevado, C., and Caflisch, A. (2009). Kinase selectivity potential for inhibitors targeting the atp binding site: a network analysis. Bioinformatics, 26(2):198–204.

[Knapp et al., 2013] Knapp, S., Arruda, P., Blagg, J., Burley, S., Drewry, D. H., Edwards, A., Fabbro, D., Gillespie, P., Gray, N. S., Kuster, B., et al. (2013). A public-private partnership to unlock the untargeted kinome. Nature chemical biology, 9(1):3–6.

[Knippschild et al., 2005] Knippschild, U., Wolff, S., Giamas, G., Brockschmidt, C., Wittau, M., Würl, P. U., Eismann, T., and Stöter, M. (2005). 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.

[Lamba and Ghosh, 2012] Lamba, V. and Ghosh, I. (2012). New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. Current pharmaceutical design, 18(20):2936–2945.

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13):1658–1659.

[Li et al., 2016] Li, Y. H., Wang, P. P., Li, X. X., Yu, C. Y., Yang, H., Zhou, J., Xue, W. W., Tan, J., and Zhu, F. (2016). The human kinome targeted by fda approved multi-target drugs and combination products: a comparative study from the drug-target interaction network perspective. PloS one, 11(11):e0165737.

[Lorenzen and Pawson, 2014] Lorenzen, K. and Pawson, T. (2014). Hdx-ms takes centre stage at unravelling kinase dynamics.

[Manning et al., 2002] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science, 298(5600):1912–1934.

[Martin, 2009] Martin, A. (2009). Profit program using mclachlan algorithm. http://www.bioinf.org.uk/software/profit.

[Martin et al., 2008] Martin, D. M., Miranda-Saavedra, D., and Barton, G. J. (2008). Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. Nucleic acids research, 37(suppl_1):D244–D250.

[Martin et al., 2010] Martin, J., Anamika, K., and Srinivasan, N. (2010). Classification of protein kinases on the basis of both kinase and non-kinase regions. PloS one, 5(9):e12460.

[Menche et al., 2015] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. Science, 347(6224):1257601.

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

[Moya-García et al., 2017] Moya-García, A., Adeyelu, T., Kruger, F. A., Dawson, N. L., Lees, J. G., Overington, J. P., Orengo, C., and Ranea, J. A. (2017). Structural and functional view of polypharmacology. Scientific Reports, 7.

[Orengo and Taylor, 1996] Orengo, C. A. and Taylor, W. R. (1996). [36] ssap: sequential structure alignment program for protein structure comparison. Methods in enzymology, 266:617–635.

[Park et al., 2012] Park, J. H., Liu, Y., Lemmon, M. A., and Radhakrishnan, R. (2012). Erlotinib binds both inactive and active conformations of the egfr tyrosine kinase domain. Biochemical Journal, 448(3):417–423.

[Perez-Lopez et al., 2015] Perez-Lopez, Á. R., Szalay, K. Z., Türei, D., Módos, D., Lenti, K., Korcsmáros, T., and Csermely, P. (2015). Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. Scientific reports, 5:10182.

[Rakshambikai et al., 2015] Rakshambikai, R., Manoharan, M., Gnanavel, M., and Srinivasan, N. (2015). Typical and atypical domain combinations in human protein kinases: functions, disease causing mutations and conservation in other primates. RSC Advances, 5(32):25132–25148.

[Robak and Robak, 2012] Robak, T. and Robak, E. (2012). Tyrosine kinase inhibitors as potential drugs for b-cell lymphoid malignancies and autoimmune disorders. Expert opinion on investigational drugs, 21(7):921–947.

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

[Santos et al., 2017] Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., et al. (2017). A comprehensive map of molecular drug targets. Nature reviews Drug discovery, 16(1):19–34.

[Shoemaker et al., 2011] Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T., and Panchenko, A. R. (2011). Ibis (inferred biomolecular interaction server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic acids research, 40(D1):D834–D840.

[Szklarczyk et al., 2014] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research, 43(D1):D447–D452.

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

[Tsai and Nussinov, 2013] Tsai, C.-J. and Nussinov, R. (2013). The molecular basis of targeting protein kinases in cancer therapeutics. In Seminars in cancer biology, volume 23, pages 235–242. Elsevier.

[Wiseman et al., 2010] Wiseman, S. L., Wei, F. Y., and Nairn, A. C. (2010). The ef2k/mhck/trpm7 family of atypical protein kinases. In Handbook of Cell Signaling, 2/e. Elsevier Inc.

[Wu et al., 2015a] Wu, P., Clausen, M. H., and Nielsen, T. E. (2015a). Allosteric small-molecule kinase inhibitors. Pharmacology & therapeutics, 156:59–68.

[Wu et al., 2015b] Wu, P., Nielsen, T. E., and Clausen, M. H. (2015b). Fda-approved small-molecule kinase inhibitors. Trends in pharmacological sciences, 36(7):422–439.

[Yu et al., 2007] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS computational biology, 3(4):e59.