ANALYSING DRUG TARGETS USING LIGAND SIMILARITY

by

Hakime Öztürk

B.S., Computer Engineering, Dokuz Eylül University, 2012

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2014

ANALYSING DRUG TARGETS USING LIGAND SIMILARITY

APPROVED BY:

Assist. Prof. Arzucan Özgür .....
(Thesis Supervisor)

Assoc. Prof. Elif Özkırımlı Ölmez .....
(Thesis Co-supervisor)

Assoc. Prof. A. Taylan Cemgil .....

Prof. Özlem Keskin .....

Suzan Üsküdarlı, Ph.D. .....

DATE OF APPROVAL:

to my family

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my supervisor Asst. Prof. Arzucan Özgür and co-supervisor Assoc. Prof. Elif Özkırımlı Ölmez who guided me throughout this journey by offering their continuous advices and encouragements. It was a privilege for me to work with them and to be benefited from their wisdom. I thank them for their endless support, patience and the effort they put into guiding me in the scientific field.

I am very grateful to my thesis committee members; Assoc. Prof. Taylan Cemgil, Prof. Özlem Keskin and Suzan Üsküdarlı, PhD, for spending their time on my thesis and their valuable recommendations.

I would like to thank the members of KB407; Begüm Alaybeyoğlu and Utku Deniz, who kindly welcomed me to their lab -even risking their time to be bothered with my non-ending questions!-, for their kind support and insightful comments and suggestions on my work.

I am very thankful to Jale Günbak and Ş. Betül Bilgin whose friendship helped me to turn these two years a memorable experience. I also owe special thanks to Neslihan Kurnaz who never gave up on me and has proved that the distance did not matter by always being there for me.

I would like to express my gratitude to my aunt, cousins and my dearest grandfather and grandmother who made me feel like I was home. I also thank to my little sunshines Kaan and Rüya who filled me with joy and laughter and refreshed my soul.

I owe an important debt to my grandfather, Mehmet Öztürk, whom I value so much, for always supporting me and being my cake chef!

Thank you is never enough to explain how grateful I am to my family for their

life-long support and guidance. I thank to my dad, whose generosity and graciousness will always be an inspiration for me, for his endless faith in me. I thank to my mother, my best friend, my mentor, literally the one I needed her to be, for everything she have done for me ever since I was born. Finally I thank to my little brother who lighten up my world with his heart of gold and always managed to find a way to make me smile. Without his friendship, it would be hard to wait for Tardis alone!

TUBİTAK-BİDEB 2210 scholarship program is gratefully acknowledged.

ABSTRACT

ANALYSING DRUG TARGETS USING LIGAND SIMILARITY

Analysis of the interactions between target proteins and drugs is crucial not only for drug discovery, but also for a better understanding of the possible evolutionary pressure that the drugs exert on the proteins. Based on the hypothesis that similar proteins bind to similar ligands, ligand similarity is utilized with two different approaches. We first introduce ligand-centric network models to analyse the relationships of protein family members via the drugs that they bind to. We build three different types of networks in which the proteins are represented as nodes, and two proteins are connected by an edge with a weight that depends on the number of shared identical or similar ligands. As a test case, we focus on β-lactamases and Penicillin-Binding Proteins. The use of ligand sharing information to cluster proteins results in modules comprising proteins both with sequence and functional similarity. Consideration of ligand similarity not only enhances the clustering of the target proteins, but also highlights some interactions that were not detected in the identical ligand network. In the second part, we follow a machine learning approach for predicting protein-ligand interactions using Support Vector Machines (SVM) where we focus on comparing different ligand similarity kernels. For this task, a larger data set of GPCR and ion channels is examined. Among the 16 different ligand kernels we experiment with, LINGO based TF-IDF cosine similarity achieves a 0.009 better AUC score than the widely used 2D Fingerprint Tanimoto model on the GPCR data set.

ÖZET

İLAÇ HEDEFLERİNİN LİGAND BENZERLİĞİ YOLUYLA ANALİZİ

Protein ve ilaçlar arasındaki ilişkinin analizi, yalnızca yeni ilaçların keşfi konusunda değil, proteinlerin ilaçlar üzerinde oluşturabileceği olası evrimsel baskının daha iyi anlaşılması açısından da büyük önem taşımaktadır. Benzer proteinlerin benzer ligandlara bağlanması esasına dayalı olarak tasarladığımız bu çalışmada, ligand benzerliği iki farklı yaklaşımla ele alınmıştır. İlk olarak, protein aileleri üyeleri arasındaki ilişkiyi bağlandıkları ligandlar yolu ile inceleyen ligand-merkezli ağ modelleri tanıtılmıştır. Proteinlerin ağın düğümleri olarak temsil edildiği üç farklı ağ modelinde, iki protein düğümü, ağırlığı proteinlerin ortak olarak bağlandıkları ligandların sayısına veya bağlandıkları ligandların benzerliğine bağlı olarak değişen bir kenar ile bağlanır. Bu kısımda β-laktamaz ve Penisilin-Bağlayan Protein aileleri üzerine yoğunlaşılmıştır. Ligand paylaşım bilgisinin kullanımıyla oluşturulan grupların hem amino-asit dizilimi hem de fonksiyonel benzerlikleri olan proteinleri biraraya topladığı gözlenmiştir. Ligand benzerlik bilgisinin kullanımı proteinlerin gruplanması işlemini iyileştirmekle kalmayıp aynı zamanda ortak ligand ağlarının bulamadığı bazı etkileşimleri vurgulamıştır. İkinci kısımda, protein-ligand ilişkisi tahminlemede makine öğrenmesi yaklaşımını izleyerek Destek Vektör Makinelerinin kullanıldığı farklı ligand benzerlik çekirdek fonksiyonlarının karşılaştırılmasına odaklanılmıştır. Bu modelde daha büyük bir veri kümesi olarak GPCR ve iyon kanalları aileleri incelenmiştir. Test ettiğimiz 16 farklı ligand çekirdek fonksiyonu arasında GPCR veri kümesinde, SMILES karakter dizisini kullanan LINGO bazlı TF-IDF kosinüs benzerliği, 2D parmakizi Tanimoto modelinden daha iyi bir performans üretmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$K(a, b)$	Kernel function for a and b
V_s	Feature vector for of string s
V_s	TF-IDF feature vector of string s
$\otimes$	Tensor product
$\times$	Multiply
β	Beta
$\Phi_{ligand}(d)$	Feature vector of ligand d
$\Phi_{ligand}(d)_T$	Transpose of feature vector of ligand d
$\Phi_{protein}(t)$	Feature vector of protein t
$\Phi_{protein}(t)_T$	Transpose of feature vector of protein t
$\theta(S_1)$	Frequencies of all possible substrings with length $q \geq 2$ in string S_1

LIST OF ACRONYMS/ABBREVIATIONS

1D	One Dimensional
2D	Two Dimensional
2RG	Ertapenem
3D	Three Dimensional
34D	Thirty-four Dimensional
38D	Thirty-eight Dimensional
ACA	6-aminohexanoic acid
ACY	Acetic acid
AIX	Ampicillin, open form
API	Application Programming Interface
AUC	Area Under Curve
BindingDB	Binding Database
BlaR	β-lactamase Regulatory Protein
CB4	Pinacol
CBD	City Block Distance
ChEMBL	Chemical database of European Molecular Biology Laboratory
CLCS	Combination of LCS algortihms
CO	Cobalt
COBALT	Constraint-based Multiple Alignment Tool
CSW	Cysteine sulfinic acid
DRW	Doripenem, open form
DWZ	(2S,3R,4S)-4-[(3S,5S)-5-(dimethylcarbamoyl)pyrrolidin-3-yl]sulfanyl-2-[(1S,2R)-1-formyl-2-hydroxypropyl]-3-methyl-3,4-dihydro-2H-pyrrole-5-carboxylic acid
EC	Enzyme Commision
EPE	Ethanesulfonic Acid
ESBL	Extended Spectrum β-lactamases
GPCR	G-protein Coupled Receptor

IC_{50}	The half maximal inhibitory concentration
ID	Identification
IDE	Integrated Development Environment
IDF	Inverted Document Frequency
IM2	Imipenem
KCX	Lysine NZ-Carboxylic acid
KEGG	Kyoto Encylopedia of Genes and Genomes
Ki	binding affinity of the inhibitor
LCS	Longest Common Subsequence
LINGO	Set of defined length of overlapping substrings in SMILES string
MA4	Cyclohexyl-Hexyl-Beta-D-Maltoside
MCLCS	Maximal Consecutive Longest Common Subsequence
MCODE	Molecular Complex Detection
MER	Meropenem, bound form
MES	2-(n-morpholino)-Ethanesulfonic acid
MN	Manganese
MW	Molecular Weight
NCLS	Normalized Longest Common Subsequence
OCS	Cysteinesulfonic acid
PBP	Penicillin Binding Protein
PCZ	Cefotaxime product, open form
PDB	Protein Data Bank
PFAM	Protein Family
PNM	Penicillin G
PNV	Penicillin V
ROC	Receiver Operating Characteristic
SMIfp	SMILES Fingerprint
SMILES	Simplified Molecular Input Entry Specification
SUC	Sucrose
SVM	Support Vector Machne

TBE	Tazobactam intermediate
TF	Term Frequency
UniProt	Universal Protein Resource
ZN	Zinc

1. INTRODUCTION

Identification of potential interactions between target proteins and drugs carries a big importance in drug discovery. Increase in the diversity of target proteins due to selective pressure and evolutionary process results in the need to discover new active compounds. However, the high cost of novel drug discovery has led to repurposing of existing compounds. Efficient prediction of target-compound interactions using computational methods accelerates research efforts in this area.

In this study, we first analyse existing protein-ligand interactions on network models and then expand the existing interaction data set by protein-ligand interaction prediction. While conducting these complementary tasks, we mainly focus on ligand similarity to identify its impact on protein-ligand interactions.

The aim of our first task is to analyse the relationships among members of a protein family. Unlike most previous studies that use sequence similarity to classify proteins, our approach is based on creating ligand-centric networks of proteins. We introduce three types of networks, where the nodes represent proteins, and the edges correspond to the sharing of identical or chemically similar ligands. Among the drug targets, β-lactamase and Penicillin Binding Protein (PBP) families lie at the heart of the mechanism that enable bacteria to gain β-lactam resistance, which currently is one of the major threats to public health. Identifying the relations among the proteins in these families and their ligands is crucial for a better understanding of bacterial evolution that results in antibiotic resistance. Therefore, as our case study we apply our methods to this dataset. We first provide an exhaustive study of the β-lactamase and PBP families and their ligands annotated in Protein Data Bank (PDB). The use of ligand sharing information to cluster proteins resulted in modules comprising proteins with not only sequence similarity but also functional similarity, even though no structural information was provided for proteins. Consideration of ligand similarity not only enhanced the clustering of the target proteins, but also highlighted some interactions that were insignificant in the identical ligand network. Analysing the β-

lactamases and PBPs using ligand-centric network models enabled the identification of novel clusters, which can be used to guide drug design efforts.

Our second task seeks to answer two main questions: (i) how much does ligand similarity affect protein-ligand interaction prediction, and (ii) whether a simplified molecular input line entry specification (SMILES) representation based similarity kernel can perform better than the widely used 2D fingerprint model with the Support Vector Machines (SVM) machine learning algorithm. We utilize a target-ligand model in which the corresponding compound space is screened against a family of proteins, namely GPCR and ion channels [3]. Proteins and ligands are represented in kernel space and SVM classification is used. In a previous study, using 2D fingerprint Tanimoto similarity as the ligand kernel, different target similarity kernels; mismatch, local alignment, dirac, musltitask and hierarchy, were tested and hierarchy kernel which considers the systematic classification of proteins, was found to produce the best performance among the others. In this work, we select the hierarchy kernel as our protein similarity kernel and focus on 16 different ligand kernels. We utilize string kernels including edit distance [4], normalized longest common subsequence (NLCS) [2], and a model that combines different longest common subsequence (LCS) [2] algorithms as well as the SMILES specialized algorithms such as, a method based on representing SMILES strings as a set of overlapping substrings with predefined length, namely LINGO [5], SMILES fingerprint (SMIfp) [6], and SMILES based substring kernel [7]. We propose some modifications to these SMILES based algorithms in which for LINGO model we use different parameter settings (substring length $q = 3, 4, 5$), and propose a weighted model. For SMIfp we modify the definition of scalar fingerprint generation and hold tests using different similarity metrics (Euclid, CBD, Tanimoto). SMILES based substring kernel is experimented with modified representation of SMILES strings. Finally, we propose LINGO-based term frequency -inverted document frequency (TF-IDF) cosine similarity ligand kernel. On the GPCR data set, the LINGO-based TF-IDF cosine similarity method that we propose performed better than the 2D fingerprint Tanimoto similarity model in terms of ROC-AUC (Receiver Operating Characteristic - Area Under the Curve). For ion channels data, the ligand kernels we tested failed to achieve higher scores than the original ligand kernel, with the NLCS algorithm with a close

performance.

The following sections of this study include detailed information about the biological and theoretical background of the research, the adopted protocol with computational methods, analysis of the results and conclusion. Background information about β-lactamase and PBP families, ligand and target similarity kernels, use of ligand similarity concept in protein clustering and protein-ligand interaction prediction tasks are given in Chapter 2. Chapter 3 includes information about the collected data, the computational approaches that were used throughout the study, and the design of the experiments. In Chapter 4 the results of the study are reported and discussed. Chapter 5 covers a small summary of the research including the fundamental aim, methodology and the important findings and conclusion. Possible strategies to follow for future work are also included. Supplementary information is presented in the Appendix.

2. THEORY

2.1. β-lactam Antibiotics and Resistance Mechanism

2.1.1. β-lactam Antibiotics

β-lactam antibiotics, which constitute 60% of the worldwide antibiotic usage, are the most effective and commonly used agents in the treatment of infectious diseases [1]. The door for the the discovery of the very first antibiotic, penicillin, was opened in 1921 by Alexander Fleming's discovery of a protein, which he called lysozyme, which was found to be the reason of lysis and decomposition of cell-wall in Gram-positive bacteria [8]. Later in 1928, Fleming's experiments on *Staphylococcus* resulted in observation of lysis, which he did previously observe with lysozyme protein, and led to discovery of penicillin [8]. However, it was not until 1940s when the penicillin was introduced as a clinical agent.

Today β-lactam antibiotics comprise several subclasses such as penicillins, cephalosporins, carbapenems, and monobactams. The drugs classified under β-lactam antibiotics are defined by a four-membered β-lactam ring. The structures of penicillins, cephalosporins, carbapenems and monobactams are shown in Figure 2.1 in which β-lactam rings are indicated in orange. They function by inhibiting the cell wall synthesis in bacteria by penicillin-binding proteins (PBPs), which reside in the cell wall and are responsible for maintaining cell wall [9].

Resistance to β-lactam antibiotics was observed even before the introduction of the penicillin to medical use [10,11]. Evolution of resistance in bacteria is an inevitable response that enhances the overall fitness of the organism [12, 13]. As a result of the evolutionary process and selective pressure, emergence of antibiotic resistant bacteria is a natural outcome. There are four known ways of resistance to β-lactam antibiotics: (i) production of β-lactamase enzymes that hydrolyse the β-lactam ring of the antibiotic, (ii) penicillin binding proteins that maintain the peptidoglycan structure in bacterial

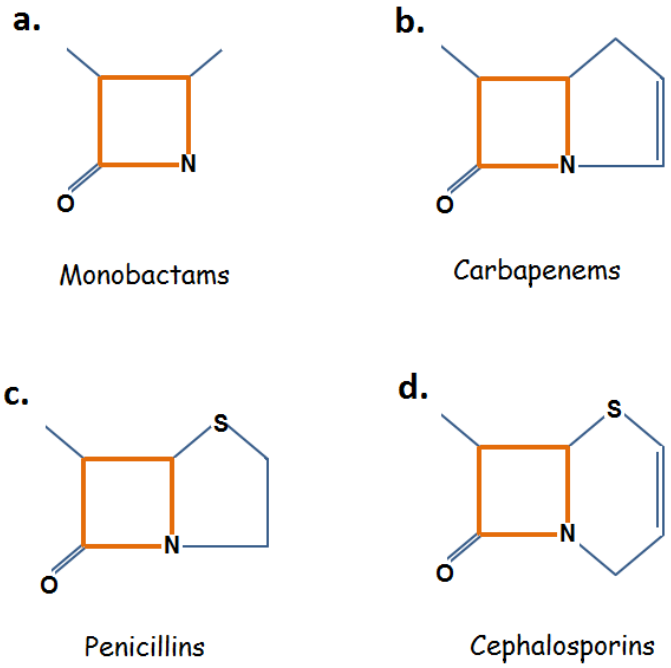

Figure 2.1. Structure of some β-lactam antibiotics. (a) Monobactams. (b) Carbapenems (c) Penicillins (d) Cephalosporins.

cell wall, (iii) alteration of porin channels, and (iv) initiation of efflux exporter proteins [14, 15].

2.1.2. β-lactamases

Bacterial β-lactamases are members of an enzyme family (EC 3.5.2.6) that deactivate the effect of the β-lactam antibiotics by attacking their β-lactam rings. In order to dissolve the β-lactam ring, most of the β-lactamases use serine ester hydrolysis mechanism, while a few use zinc ion mechanism [1]. Figure 2.2 illustrates the action of serine ester mechanism.

The first β-lactamase was observed in *Staphylococcus aureus* strains in the 1940s before the use of penicllin as an antimicrobial agent [16]. After its introduction in 1944, benzylpenicillin's activity against *Staphylococcus aureus* was reported to be decreased dramatically within five years as a result of genetic transfer and selection bacteria [16]. The newly introduced β-lactams against Gram-negative bacteria also followed the similar fate as a result of quickly developed resistance mechanism [1]. In 1980s, Extended-Spectrum β-lactam antibiotics, such as ceftazidime and cefotaxime, and β-lactam inhibitors, such as sulbactam and clavulanic acid, were introduced as a response to widespread β-lactamase producing pathogens [17]. However, the emergence of the Extended-Spectrum Beta-Lactamases (ESBLs) with resistance to cephalosporins quickly followed [18, 19]. As a result of the dramatic increase in the diversity of the β-lactamases and the emergence of ESBLs; many different classification schemes were published to organize the β-lactamase family.

The first broadly accepted classification of β-lactamases was proposed by Richmond and Sykes in late 1960s, in which they decided according to the enzymes's hydrolysis rate of penicllin and inhibition rate by cloxacillin and/or p-chloromercuribenzoate [1]. However, this scheme required major revisions, therefore it was replaced by Bush in 1989 [20]. Today, there are two globally accepted classification schemes for β-lactamases, where the first one is based on amino-acid sequence classification and the second one is based on functionality. β-lactamases were divided into four classes (Class

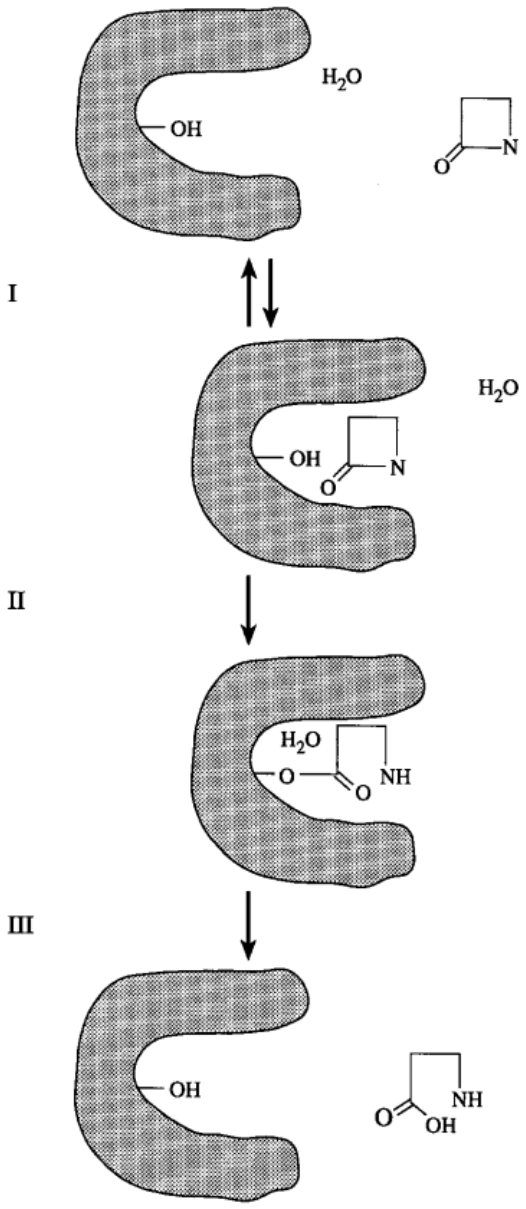

Figure 2.2. Hydrolysis of a β-lactam ring by β-lactamase using serine ester mechanism. -OH group of the active site serine is also shown [1].

A–D) based on their sequence similarity by Ambler in 1980 [21]. Classes A, C and D function by the serine ester hydrolysis mechanism, whereas class B β-lactamases, also known as Metallo β-lactamases, use zinc ion to destroy the β-lactam ring. The classification scheme by functionality resulted in three major groups: Group 1 cephalosphorinases (Class C), Group 2 serine β-lactamases (Class A and Class D), and Group 3 Metallo β-lactamases (Class B), each of which is also divided into several different subgroups [22, 23]. The functionality based classes of the β-lactamases were determined according to their hydrolysis rates by some pre-defined drugs such as EDTA, and bezylpenicillin. Some widely known β-lactamases and their functional and sequence-based classes are given in Table 2.1.

Table 2.1. Classification schemes for β-lactamases [23].

Bush-Jacoby Group (2009)	**Bush-Jacoby-Medeiros Group (1995)**	**Molecular Class (Ambler)**	**Representative Enzymes**
1	1	C	AmpC, P99, ACT-1, CMY-2, FOX-1, MIR-1
1e	–	C	GC1, CMY-37
2a	2a	A	PC1
2b	2b	A	TEM-1, TEM-2, SHV-1
2be	2be	A	TEM-3, SHV-2, CTX-M-15, PER-1, VEB-1
2br	2br	A	TEM-30, SHV-10
2ber	–	A	TEM-50
2c	2c	A	PSE-1, CARB-3
2ce	–	A	RTG-4
2d	2d	D	OXA-1, OXA-10
2de	–	D	OXA-11, OXA-15
2df	–	D	OXA-23, OXA-48
2e	2e	A	CepA
2f	2f	A	KPC-2, IMI-1, SME-1
3a	3	B	MP-1, VIM-1, CcrA, IND-1
3b	3	B	CphA, Sfh-1

By the end of 2009, over 890 unique protein sequences of β-lactamases were reported by Jacoby and Bush (http://www.lahey.org/Studies/) [23]. Today, a search through UniProt with EC classification number 3.5.2.6 returns more than 4900 hits.

2.1.3. Penicillin Binding Proteins

Peptidoglycan is responsible for forming the bacterial cell wall. Penicillin-binding proteins (PBPs), take part in the polymerization of the glycan strands, which form the main substance of the peptidoglycan, and link gylcan chains together [24]. PBPs, which are found in all bacteria in varying amounts and are located in the bacterial membrane, have the ability of covalently bind to penicillin [25, 26].

PBPs are classified into two groups according to their molecular weights (MW) as low MW PBPs and high MW PBPs, both of which are also divided into subgroups namely A, B, and C based on sequence similarity [27]. High MW PBPs are also referred to as multimodular PBPs, whereas the PBPs that are not able to synthesize peptidoglycan are called monofunctional [28]. PBPs and β-lactamases tend to cluster together instead of forming clusters of their own when sequential similarity is considered [14]. PBPs are reported to be ancestors of the β-lactamases, and most of the members of both of these families harbor active-site serine [14].

2.2. Targets and Target Similarity

Proteins are usually classified mostly according to their amino-acid sequences and structures. Besides these popular ways of representation, motifs and pharmacological properties are also used to represent and classify proteins.

A large volume of work has been published to design similarity kernels for proteins, varying from the kernels using sequence representation to kernels based on 3D structures [29–33]. REFS Sequence based similarity allows use of any type of string similarity kernel such as edit distance. Among the sequence-based kernels, the most widely used ones are global alignment (Needleman-Wunsh) and local alignment (Smith-Waterman) [34, 35]. Global alignment finds the best alignment of given sequences by comparing their complete length, whereas local alignment finds the best alignment of string pieces of maximum possible length. The spectrum kernel, computes a similarity value by counting the number of common k-mers in target sequences [29]. The Mis-

match kernel is the extended version of the spectrum kernel which allows mismatches. These kernels ignore position information, therefore kernels that consider position information such as weighted degree (WD) kernel [36], the WD kernel with shifts [37] and oligo kernel [38] were also proposed.

2.2.1. Hierarchy Kernel

In this study as a target kernel we use a family hierarchy dependent kernel, which considers the number of shared ancestors of two targets [3]. Considering hierarchy system actually leads to use of functional similarity of the proteins since these hierarchies are organized using protein-ligand interaction information. For instance, enzymes are divided into hierarchies by EC (Enzyme Commission system) according to the reactions they catalyze. EC is succeeded by at most 4 numbers all of which are separated by a period. To give an example, EC 3.5 represents the hydrolases that are active on carbon-nitrogen bonds and peptide bonds. It is a subclass of EC 3 (hydrolases) and it is superior of 3.5.2 which represents cyclic amides in EC 3.5.

The GPCR family is divided into four classes: Class A (the rhodopsin family), Class B (the sectretin family), Class C (the metabotropic family) and Class D (the rest of the receptors) all of which are also divided into subgroups. Ion channels are also divided into subclasses by KEGG according to different features which can be useful for the hierarchy as means of similarity.

Equation 2.1 below describes the hierarchy kernel where $\theta_h(p)$ denotes to a feature vector where each of the rows corresponds to a node in the hierarchy. A row is set to 1 if a node is placed under the hierarchy of p or 0 otherwise.

$$K_{hierarchy}(p, p') = \langle \theta_h(p), \theta_h(p') \rangle \tag{2.1}$$

2.3. Ligands and Ligand Similarity

Ligand is a small molecule that binds to a protein by forming covalent or noncovalent bonds [39]. If the bond established between protein and ligand is covalent, then it results in irreversible binding. However, reversible binding is more common.

Ligands are described using convenient properties, which are called descriptors. These descriptors are mostly classified according to their dimensionality, 1D, 2D and 3D, all of which allow use of varying similarity metrics to calculate the distance between two ligands [40]. Figure 2.3 illustrates the three different representations of a sample ligand, tazobactam intermediate (TBE). In Figure 2.3a, 1D representation of the ligand, which is a string named SMILES, is shown. Figure 2.3b is 2D model of TBE which is a graph created by the atoms and bonds. Finally, Figure 2.3c illustrates the 3D form of TBE.

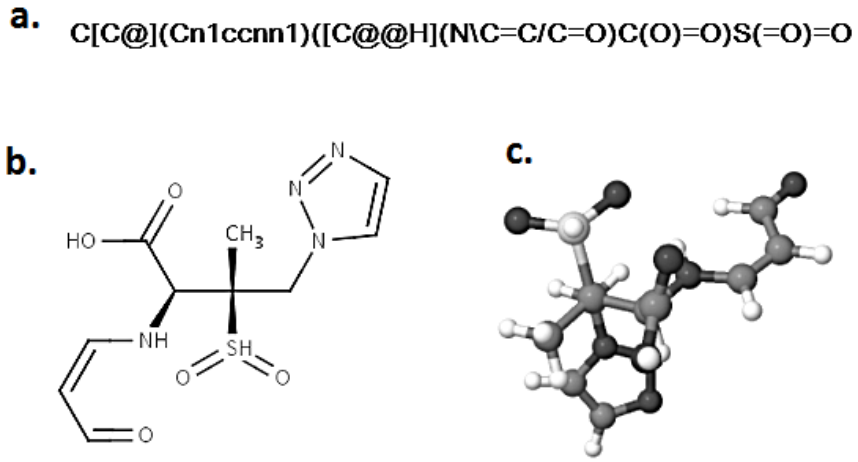

Figure 2.3. Three representations for Tazobactam Intermediate (TBE). (a) 1D SMILES string. (b) 2D model as a graph of atoms and bonds. (c) 3D object.

2.3.1. 1D based similarity

1D descriptors, which are simple way of representing ligands, convey the information of global properties such as molecular weight, atom and bond counts.

The most popular use of 1D representation is the Simplified Molecular Input Entry Specification (SMILES), which is a way of describing molecular structures in the form of strings. SMILES was introduced by Arthur Weininger in 1988, then unique SMILES format was developed the following year by Weininger *et al.* [41, 42]. SMILES strings convey information about molecular structures by representing atoms and bonds with some specific symbols. The atoms are represented with their corresponding symbols and bonds are represented with some special characters where '-', '=', '#' and ':' denotes the single, double, triple and aromatic bonds respectively [43]. Parantheses in SMILES string indicate branches, while lower case letters are used to indicate aromatic rings. More detailed description of the SMILES string is given in the Daylight website (www.daylight.com/dayhtml/doc/theory/theory.smiles.html).

Since SMILES format is based on string representation, different ordering of strings can cause ambiguities even though arrangement of the string does not affect the structure of the molecule. Therefore, to indicate the uniqueness of a SMILES string, Canonical SMILES is used, which allows an algorithm generated string. There are different canonicalization algorithms developed by Daylight Chemical Information Systems, Chemistry Development Kit, ChemAxon [44–46].

Ligand similarity comparisons on 1D representation model can make use of any string similarity algorithm. In this section we will provide some of the widely used string similarity methods with the ones specialized to measure the similarity of SMILES strings. Before introducing these methods, let us remind three of the most popular distance and similarity metrics which will be referred to in these methods.

Euclidean Distance is a way of computing distance between points a and b defined in $\mathbb{R}^n$ [47].

$$EuclidDistance(a, b) = \sqrt{\sum_{i=1}^{n} |a_i - b_i|^2} \tag{2.2}$$

City Block Distance (CBD) or Manhattan distance computes the distance between points a and b in a $\mathbb{R}^n$ space [48].

$$CBD(a,b) = \sum_{i=1}^{n} |a_i - b_i| \tag{2.3}$$

Tanimoto coefficient (Tc) for two binary vectors, X and Y is given by,

$$Tanimoto(X,Y) = \frac{z}{(x+y-z)} \tag{2.4}$$

where x represents the number of bits set to 1 in X, y represents the number of bits set to 1 in Y, and z represents the number of bits set to 1 in both [49].

2.3.1.1. LINGO. LINGO refers to q-character substrings of a SMILES text [5]. A SMILES string of length n can be represented with $(n-(q-1))$ q-length substrings (LINGOs). LINGO profiles are generated from Canonical SMILES which undergo two main modifications before the LINGO creation process: First, all ring numbers in the SMILES are set to '0', and then 'Cl' and 'Br' atoms are replaced with 'L' and 'R' respectively.

For example, a SMILES string 'ccc1c2NcCl' becomes 'ccc0c0NcL' and the following LINGOs are generated for $q = 4$ with the corresponding frequencies: 'ccc0':1, 'cc0c':1, 'c0c0':1, '0c0N':1, 'c0Nc':1, and '0NcL':1. To calculate the similarity between two compounds, LINGO profiles for each molecule are generated and then the Tanimoto coefficient is used,

$$LINGOsim = \frac{\sum_{i=1}^{m} 1 - \frac{|N_{S_1,i} - N_{S_2,i}|}{|N_{S_1,i} + N_{S_2,i}|}}{m} \tag{2.5}$$

where m is the number of LINGOs created from the SMILES string of any of the compounds while $N_{S_1,i}$ represents the number of LINGOs of type i in compound S_1

and $N_{S_2,i}$ represents the number of LINGOs of type i in compound S_2 [5] .

2.3.1.2. SMILES Fingerprint (SMIfp) Kernel. SMILES Fingerprint (SMIfp) is introduced by Schwartz *et al.* as a method to perform ligand-based virtual screening. SMIfp is based on representing SMILES strings in a 34-dimensional space where each of the dimensions correspond to the frequency of a different symbol in that string [6]. More than 32 million compounds in PubChem are analyzed to identify the most-frequent symbols to form the best-representative scalar fingerprint and as a result, 34 relevant symbols are selected. Table 2.2 depicts the selected symbols with their definitions. Once SMILES strings are converted to scalar fingerprints, City Block Distance (CBD) is used to measure their similarities.

2.3.1.3. SMILES representation-based string Kernel. The idea behind the SMILES representation-based string kernel is to compare the substrings of two strings [7]. Given two strings S_1 and S_2, $\theta(S_1)$ and $\theta(S_2)$ respectively denote the frequencies of all the possible substrings with at least $q = 2$ character length. The string kernel is defined as the inner product of these frequencies (Equation 2.6).

$$K(S_1, S_2) = \langle \theta(S_1), \theta(S_2) \rangle \tag{2.6}$$

To illustrate the model better, let us provide an example for two SMILES strings S_1 = CCCOC and S_2 = CCC1. For S_1, the following substrings are generated with the corresponding frequencies:

'CC':2,
'CO':1,
'OC':1,
'CCC':1,
'CCO':1,
'COC':1,

Table 2.2. SMIfp symbol table [6].

no	**symbol**	**definition**
1	C	nonaromatic carbon atoms
2	c	aromatic carbon atoms
3	N	nonaromatic nitrogen atoms
4	n	aromatic nitrogen atoms
5	O	nonaromatic oxygen atoms
6	o	aromatic oxygen atoms
7	S	nonaromatic sulfur atoms
8	s	aromatic sulfur atoms
9	F	fluorine atoms
10	Cl	chlorine atoms
11	Br	bromine atoms
12	I	iodine atoms
13	P	nonaromatic phosphorus atoms
14	p	aromatic phosphorus atoms
15	B	boron atoms
16	'X'	any other character
17	–	single bonds
18	=	double bonds
19	#	triple bonds
20	[	Nonorganic elements, charges, isotopes, protonation states
21	-	negative charges
22	+	positive charges
23	H	explicit hydrogen atoms
24	(	acyclic branching points
25	1	nonfused ring systems
26	2	bicyclic systems
27	3	tricyclic systems
28	4	tetracyclic systems
29	5	pentacyclic systems
30	6	hexacyclic systems
31	7	heptacyclic systems
32	8	octacyclic systems
33	9	nonacyclic systems
34	%	higher order ring systems

'CCCO':1,
'CCOC':1,
'CCCOC':1

For S_2, the following substrings are generated with the corresponding frequencies:

'CC':2,
'C1':1,
'CCC':1,
'CC1':1,
'CCC1':1,

The inner product of S_1 and S_2 is performed between the common substrings of the strings which equals $K(S_1, S_2) = 2 \times 2 + 1 \times 1 = 5$. In the same manner, the similarity of S_2 with itself equals to $K(S_2, S_2) = 2\times2+1\times1+1\times1+1\times1+1\times1 = 8$ [7].

2.3.1.4. Edit Distance. Edit distance is one of the most widely used measures to make comparisons between strings. Given two strings S_1 and S_2, edit distance is defined by the number of minimum edit operations required to convert S_1 into S_2 [50]. There are three main operations allowed which are insertion, deletion and substitution. For instance, between two strings S_1 = 'night' and S_2 ='delight', $edit(S_1, S_2) = 3$. We perform three operations: replacing 'n' with 'l' and inserting two characters, 'd' and 'e' .

2.3.1.5. Normalized Longest Common Subsequence (NLCS). The Longest Common Subsequence (LCS) algorithm finds the common subsequence with the maximum possible length of two strings [51]. The algorithm does not require the characters in the common subsequence to be consecutive. For example, given two SMILES strings S_1 = C(COCCOCCO) and S_2 =C(CS(=O) the longest common subsequence lcs,

S_1 = **C(CO**CCOCCO**)**

$S_2 =$ **C(C**S(=**O)**
$lcs =$**C(CO)**

Normalized LCS, is modified in a way such that the algorithm considers the lengths of both strings. Given two strings S_1 and S_2 the NLCS is [2],

$$NLCS(S_1, S_2) = \frac{length(LCS(S_1, S_2))^2}{length(S_1) \times length(S_2)} \tag{2.7}$$

2.3.1.6. Combination of LCS Models (CLCS). Islam and Inkpen proposed a method combining three algorithms each of which modifies the LCS algorithm in its own way [2]. The first algorithm is Normalized LCS (NLCS), which is described in the previous section. It normalizes the LCS algorithm by considering the lengths of the strings. The second algorithm, which is described in Figure 2.4, is Maximal Consecutive Longest Common Subsequence starting from the character **1** $MCLCS_1$.

input: r_i, s_j /* r_i and s_j are two input strings where $|r_i| = t$, $|s_j| = n$ and $t \leq n$ */
output: r_i /* r_i is the maximal consecutive LCS starting at character 1 */

$t \leftarrow |r_i|$, $n \leftarrow |s_j|$
while $|r_i| \leq 0$ **do**
  if $r_i \cap s_j$ **then**
    return r_i
  else
    $r_i \leftarrow r_i \backslash c_t$ /* remove the right most character from r_i /*
  end if
end while

Figure 2.4. $MCLCS_1$ algorithm. Maximal consecutive LCS starting at any character 1 [2].

The last one is Maximal Consecutive Longest Common Subsequence starting from the character **n** $MCLCS_n$ [2]. Figure 2.5 depicts the algortihm of $MCLCS_n$.

input: r_i, s_j /* r_i and s_j are two input strings where $|r_i| = t$, $|s_j| = n$ and $t \leq n$ */
output: x /* x is the maximal consecutive LCS starting at any character n */
$t \leftarrow |r_i|$, $n \leftarrow |s_j|$

while $|r_i| \leq 0$ **do**
  determine all n-grams from r_i where $n = 1...|r_i|$ and
  /* $\bar{r_i}$ is the set of n-grams */
  if $x \in s_j$ where $(x|x \in \bar{r_i}, x = Max(\bar{r_i}))$ **then**
    /* i is the number of n-grams and $Max(\bar{r_i})$ returns the maximum length n-gram from $\bar{r_i}$ */
    return x
  else
    $r_i \leftarrow r_i \backslash c_t$ remove the right most character from r_i
  end if
end while

Figure 2.5. $MCLCS_n$ algorithm. Maximal consecutive LCS starting at any character n [2].

Unlike LCS algorithm, both of these algorithms require common subsequences to be successive. Before combining these methods, $MCLCS_1$ and $MCLCS_n$ are also normalized, becoming $NMCLCS_1$ and $NMCLCS_n$, respectively. Given two strings S_1 and S_2, $NMCLCS_1$ and $NMCLCS_n$ are calculated as [2],

$$NMCLCS_1(S_1, S_2) = \frac{length(MCLCS_1(S_1, S_2))^2}{length(S_1) \times length(S_2)} \tag{2.8}$$

$$NMCLCS_n(S_1, S_2) = \frac{length(MCLCS_n(S_1, S_2))^2}{length(S_1) \times length(S_2)} \tag{2.9}$$

In order to compute the similarity between S_1 and S_2, the weighted sum of these three algoritms are taken as follows:

$$\begin{aligned} v_1 &= NLCS(S_1, S_2) \\ v_2 &= NMCLCS_1(S_1, S_2) \\ v_3 &= NMCLCS_n(S_1, S_2) \\ Similarity(S_1, S_2) &= v_1 \times w_1 + v_2 \times w_2 + v_3 \times w_3 \end{aligned} \tag{2.10}$$

where w_1, w_2, w_3 are the weights. The original method gives each algorithm the same weight ($w_1 = w_2 = w_3 = 0.33$).

2.3.1.7. TF-IDF based Cosine Similarity. Term Frequency - Inverse Document Frequency (TF-IDF) weighting is one of the most popular methods in Information Retrieval to measure string similarity. One of the advantages of this method is that it assigns more weight to the strings that share some exclusive terms. Term Frequency (TF) reflects the number of times a term occur in the document [52]. Inverse Document Frequency (IDF), on the other hand, assigns higher weights to the rare terms of the document collection and it is described as [53],

$$idf(t, D) = log \frac{N}{|d \in D : t \in d|} \tag{2.11}$$

where t, D and N denote the term, document corpus and number of documents in the corpus respectively [53]. Tf-idf weighting is computed as follows,

$$tfidf = tf \times idf \tag{2.12}$$

In order to compute the similarity between two strings using this method, each string has to be converted into a feature vector V_s. The number of features of V_s is equal to the number of terms in the corpus. Each feature contains the tf-idf score of the corresponding term in the string. Similarity of two strings is determined according to the cosine angle between their vectors.

$$CosineSimilarity(S1, S2) = \frac{\sum_{i=1}^{m} V_{S1,i} V_{S2,i}}{\|V_{S1}\| \|V_{S2}\|} \tag{2.13}$$

V_{S1} and V_{S2} are feature vectors and m denotes the length of the vectors in Equation 2.13 [54].

2.3.2. 2D based similarity

In the 2D representation of a molecule, a graph model with atoms and the connecting bonds is shown. The 2D based similarity of a ligand is calculated using this graph model and it is often referred to as fingerprint. A fingerprint is a bit vector (0 - 1), which is constructed by using depth-first search from each vertex of our atom-bond graph of a molecule. For each path defined in the molecule, the graph is converted into a hash value which is used to create the fingerprint vector [55]. The most commonly used method that calculates the similarity between fingerprints is the Tanimoto coefficient.

2.3.3. 3D based similarity

3D representation denotes the three dimensional coordinates for each atom of the molecule. The coordinates of each atom allows the calculation of distance between two molecules. Besides atomic coordinates, 3D pharmacophores, shapes, potentials, and spectra are also used as descriptors.

2.4. Use of Ligand Similarity for Protein Clustering

Prior studies have classified proteins based on their sequence, structural and functional similarities. For instance β-lactamases are divided into four classes according to their sequence similarities, GPCRs are separated into four classes by a model which considers both sequence and functional similarity. In this study, we propose a ligand-based clustering model, where proteins are grouped by considering the shared identical or similar ligands without requiring any information on the structures of the proteins.

The first attempt to cluster proteins on a network model was proposed by Yildirim *et al.* in which they create a network called target-protein network by connecting proteins (nodes) if they have at least one common ligand [56]. However, this study did not consider the similarity of the different compounds by which the proteins are targeted. Using ligand similarity to characterize the relationship among biomolecules has gained the attention of the researchers in the recent years. A study of the relationship between alpha helical proteins and their ligands showed that proteins with at least 45% sequence identity tend to bind to similar ligands [57]. In a more exhaustive study that included 87 protein super families, it was observed that sequence similarity can be as low as 30% for proteins to interact with similar ligands [58]. Keiser *et al.* used ligand chemical similarity information to cluster a subset of activity classes in the '2006.1 MDDR' database and showed that using only ligand information, biologically related proteins grouped together [59]. It was found that when two proteins bind to the same ligand, it is likely that the ligands of one of these proteins will bind to the other protein as well. This information is used in predicting the structure of protein-ligand complexes [60]. Using ligand similarity rather than using sequence and structure similarity allowed the clustering of some sequentially different proteins [61]. Cheng *et al.* used both protein and ligand similarity representing the compounds and the targets that they bind to as nodes in a bipartite network. The binding affinity or the inhibitory activity was used for calculating edge weights [62,63]. The network based model they proposed was used to predict compound-protein interactions without the use of structural information of the components.

2.5. Use of Ligand Similarity for Protein-Ligand Interaction Prediction

The prediction of the interactions between protein and ligands is a significant task which allows us to discover new drugs for proteins or novel proteins for existing drugs. There have been two basic generally accepted approaches to drug discovery, the one which is called ligand-based and the one which is called structure-based or docking. Ligand-based approaches follow an experimental process where the known ligands of a target are screened against a large compound library [64]. Structure-based approaches, on the other hand, utilize the structure of the binding site of a target in order to decide the best suitable ligand by screening a large database [64]. Ligand based approaches lack applicability in cases in which the target is orphan, whereas structure-based approaches are unable to make predictions when 3D structure of the target is not available. There are also other approaches such as literature mining where interacting genes and compounds are extracted from the related articles [65].

With chemogenomics, which overcomes the problems of previous strategies, a new era has begun in the field of drug discovery [66]. Chemogenomics is an interdisciplinary field that aims to shed a light to basic questions such as how to define ligand similarity, what properties make two ligands similar or whether ligands can be predicted for a given target [67]. All chemogenomic approaches are built on these two basic assumptions: (i) "chemically similary compounds also should share target proteins", and (ii) "targets that share similar binding sites should also share ligands" [40]. Therefore, chemogenomics has three main components: (i) set of compounds (ii) set of targets (iii) reliable interaction information [40]. Among these components, we described the ligand and target spaces in the previous sections. Target-ligand space, however, is the combination of these three components such that targets are the rows and ligands are the columns of a matrix and each cell points to an interaction (e.g. IC_{50}, Ki).

Chemogenomics is divided into three approaches by Ronan [40], the first of which is called ligand-based chemogenomics and aims to design specific compound libraries for protein families or sub-families based on the the idea that of similar ligands have similar biological activities. The second approach, target-based chemogenomics, concentrates

on target binding site similarity by which the unknown ligands are inferred using the known ligands of similar targets. Finally, target-ligand based chemogenomics utilizes target-ligand space to make predictions for a single target by using ligand-binding information of targets.

Many different machine learning approaches based on ligand-target model have been proposed as a response to the drug-target prediction issue [68–71]. Most of these models made use of kernel functions to calculate similarity. Bleakley *et al.* developed a methodology that combines bipartite graphs with local supervised models [72]. Gaussian interaction kernel, introduced by Laarhoven *et al.*, was built on binary vectors indicating interaction status (present/absent) of compounds and targets [73]. Jacob *et al.* used different target kernels for SVM classification while Geppert *et al.* presented a comparative study of different SVM models for the task of ligand-protein prediction [3, 74]. Among the presented studies, the ones which require the use of a ligand kernel utlize 2D fingerprint representation with Tanimoto similarity.

The focus of our study is to analyze the effect of ligand similarity on this task. We select the study of Jacob *et al.* as a base, and replicate their study by replacing their original ligand kernel with the ones we select. In Jacob *et al.*'s study, each ligand-target pair is represented by a vector of features such that,

$$\Phi(d,t) = \Phi_{ligand}(d) \otimes \Phi_{protein}(t) \tag{2.14}$$

where $\Phi(d,t)$ is the tensor product of $\Phi_{ligand}(d)$, which denotes the feature vector of a compound d and $\Phi_{protein}(t)$, which represents the feature vector of target protein t [3]. Factorization of Equation 2.14 into the Equation 2.15,

$$\begin{aligned}&\big(\Phi_{ligand}(d) \otimes \Phi_{protein}(t)\big)^T \big(\Phi_{ligand}(d') \otimes \Phi_{protein}(t')\big) \\ &\quad = \Phi_{ligand}(d)^T \Phi_{ligand}(d') \times \Phi_{protein}(t)^T \Phi_{protein}(t')\end{aligned} \tag{2.15}$$

allows the use of the inner product of the vector models of ligands and targets [3]. This

representation can be easily converted into kernel space by [3],

$$\begin{aligned} K_{ligand}(d, d') &= \Phi_{ligand}(d)^T \Phi_{ligand}(d') \\ K_{protein}(t, t') &= \Phi_{protein}(t)^T \Phi_{protein}(t') \\ K((c, t), (c', t')) &= K_{ligand}(d, d') \times K_{protein}(t, t') \end{aligned} \tag{2.16}$$

Therefore, this form can be integrated into any problem solving methodology which uses kernel models such as SVM. Two-Class SVM, introduced by Vapnik, aims to find a maximum margin hyperplane that separates positive instances from negatives in order to solve the classification problem. For cases where data is not linearly separable, SVM aims to,

$$\begin{aligned} min \quad & \frac{1}{2}\|w\|^2 + C\sum_t \xi^t \quad \text{subject to,} \\ & r^t(w^T x^t + w_0) \geq 1 - \xi_t \end{aligned} \tag{2.17}$$

where the sample $X = \{x_t, r_t\}$, ξ^t is the slack variable and C is the penalty factor of misclassification [75].

.

3. MATERIALS AND METHODS

3.1. Network Models

3.1.1. Data collection

We collected our data set of protein-ligand interactions from the Protein Data Bank (PDB) database (http://www.uniprot.org/uniprot, accessed On October 27, 2013). We selected the proteins based on their Enzyme Commission (EC) and Protein Family (PFAM) numbers. The β-lactamase base was obtained by selecting EC 3.5.2.6 that denotes the β-lactamase family, PF13354 that denotes the β-lactamase enzyme family, and PF00144 that represents the β-lactamase domain. The PBP family proteins were obtained by selecting EC 3.4.16.4 that represents DD-transpeptidase family, PF00905 that refers to the transpeptidase family, and PF00768 that denotes the peptidase s11 family. The idea behind combining different classification schemes was to be able to detect entries which may not be reported in one classification scheme, but might be reported in another one. For instance, extended spectrum β-lactamase GES-5 (Q09HD0) was reported under the PF13354 classification, whereas EC 3.5.2.6 did not contain information about it.

Identification system in PDB is different than the one in UniProt. PDB assigns unique identifiers (IDs) to each entry including different complexes of the same protein. For instance, UniProt accession number P62593 for β-lactamase TEM is represented as 62 PDB chains. Therefore, we first mapped all the PDB IDs to the corresponding UniProt accession numbers. The final data set used in this study, consists of unique UniProt acession numbers, as shown in Table 3.1. Throughout this thesis proteins are referred to by their six character UniProt accession numbers (e.g. P00811) whereas ligands are referred to by their PDB IDs (e.g. IM2).

A total of 146 proteins with unique UniProt accession numbers were retrieved. We then filtered out the data set using the following criteria: (i) Ions (ZN, CO, MN

Table 3.1. Distribution of the proteins in the data set according to EC and PFAM classifications.

ID	num. of proteins	num. of ligands
EC 3.5.2.6	60	222
EC 3.4.16.4	16	61
PF00905	36	72
PF00768	12	60
PF13354	37	136
PF00144	50	203
TOTAL	111	304

etc.) were removed; (ii) Ligands which are reported in PDB to interact with more than 3000 targets were removed. Thus, the proteins that do not bind to any ligand or the proteins that only bind to ligands that are filtered by the criteria described above were not included in the protein data set in our study. As a result, 111 unique proteins represented in the PDB by more than 2000 structures were included in our database.

3.1.2. Sequence Alignment

Multiple sequence alignment was performed on the 111 proteins in the data set using COBALT [76] to provide an insight about the distribution of the proteins in the data set according to their amino acid similarities.

3.1.3. Ligand Similarity

The ligand molecules are defined using the chemical hashed fingerprint model, which conveys the information of the 2D structure in bit strings (0 and 1). Then to calculate the similarity between two fingerprints, Tanimoto coefficient is used. We used ChemAxon (http://www.chemaxon.com/), which provides JChem 6.0 interface for .NET development environment, to create fingerprints from the SMILES repre-

sentations of the ligands and then, to calculate the Tanimoto similarity between the pairs. PDB also uses ChemAxon to advance the chemical structure search options (http://www.rcsb.org/pdb/search/advSearch.do).

3.1.4. Ligand Clustering

Ligand clustering was performed using ChemMine [77] with the average linkage hierarchical clustering model.

3.1.5. Protein-ligand binding network construction

The networks presented in this study were visualized and analysed using Cytoscape (Version 2.8.3; http://www.cytoscape.org/) [78]. We proposed three different undirected network models, namely unweighted identity, weighted identity, and similarity networks to represent protein-ligand binding information. In all three networks, the target proteins were represented as the nodes and the ligands were represented as the edges. Figure 3.1 depicts the creation of the network models with four proteins A, B, C, D shaped as circles and five ligands $lg1, lg2, lg3, lg4, lg5$ shaped as diamonds. For each protein, ligands that it binds to are given together. A sample Tanimoto similarity matrix is also provided for the ligand pairs.

- In the unweighted identity network, A and B, which share two identical ligands, are connected with an edge of weight 1.
- In the weighted identity network, A and B are connected with an edge of weight 2, since they have two common ligands, lg1 and lg5.
- In the similarity network, the nodes that bind to ligands whose pairwise similarities exceed the 0.7 cut-off value are connected. Therefore, we have two new connections in the similarity network: C and D are connected with an edge of weight 0.8 (lg2–lg4), and A and C are connected with an edge of weight 0.75 (lg2–lg3). Since A and B also share chemically similar ligands along with the identical ones, the weight of the edge connecting them becomes 2.8.

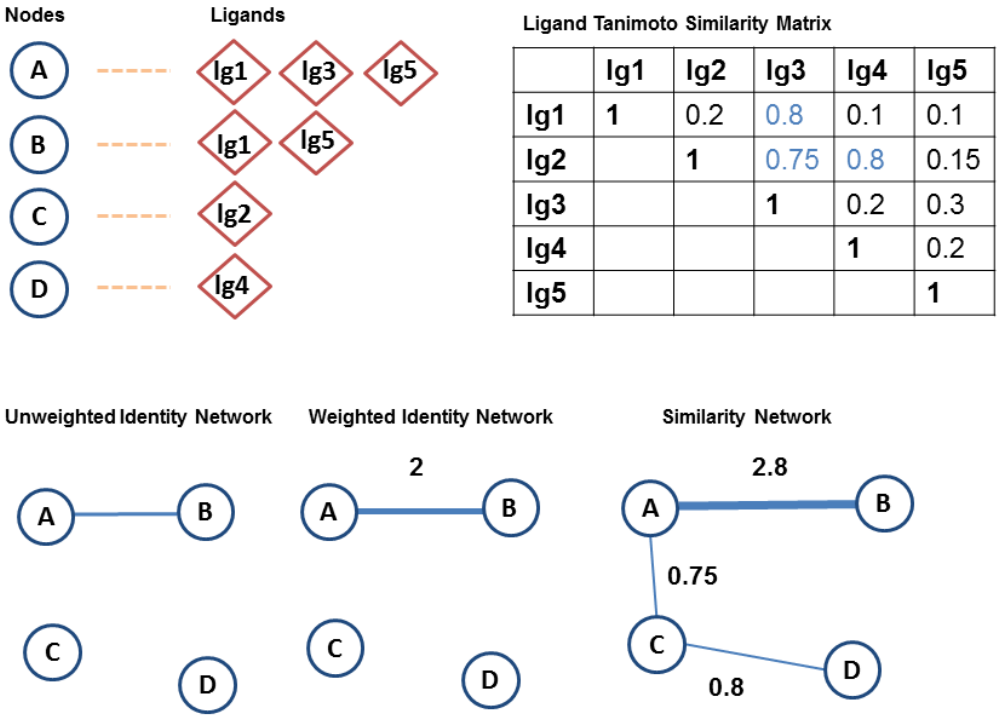

Figure 3.1. Generation of unweighted identity network, weighted identity network and similarity network for a sample system. Ligand binding information of 4 proteins and 5 ligands is used to construct networks.

3.1.5.1. Unweighted Identity Network. Unweighted identity network is the model where we illustrated the interactions in the simplest form. In this model, two nodes with identical ligand(s) are connected. Two nodes, i.e. two proteins, are connected if they share an identical ligand. The edges are unweighted. In other words, even if two proteins share more than one ligand, they are still connected with a single edge of weight 1.

3.1.5.2. Weighted Identity Network. In the weighted identity network, the number of identical ligands shared by two nodes is taken into account. For instance, if two nodes have five common ligands, the weight of the edge connecting these proteins is set to 5. Unlike the unweighted identity network, the weighted identity network provides information on the number of shared ligands.

3.1.5.3. Similarity Network. The similarity network is a weighted network model, where the chemical similarities between the ligand pairs are considered. This model enables us to link two nodes that do not have any common ligands, but bind to ligands whose chemical similarity is above some pre-determined threshold. Matter revealed that similarity cut-off value of 0.85 resulted in complete enclosure of the biological classes in his compound data set of IC93 [79]. It was previously shown that compounds with similarity values higher than 0.7 also had similar biological activity [57,80]. Therefore, in this study, the similarity threshold was selected as the Tanimoto coefficient of 0.7. In other words, ligands with similarity value higher than 0.7 contributed to the edge weights. The weight of the edge between two nodes (i.e. proteins) X and Y was computed by taking the sum of the pairwise similarity scores among their ligands (Equation 3.1).

$$weight = \sum_{i=1}^{n} \sum_{j=1}^{m} (Tanimoto(X_i, Y_j) > 0.7) \tag{3.1}$$

X_i represents the i^{th} ligand that X binds to, and Y_j represents the j^{th} ligand that

Y binds to. Higher weight suggests stronger relationships between the corresponding nodes. An advantage of the similarity network is that it is able to contribute more nodes to the network, which could not be achieved by using the previous models. Moreover, the similarity based model aims to discover some hidden relationships or emphasize the existing ones using the ligand chemical similarity feature.

3.1.6. Network Parameters

3.1.6.1. Centrality. Degree centrality is measured according to the number of connections a node has. As the number of nodes that a node connects to increase, the centrality degree of that node becomes higher. Betweenness centrality measures how many times a node acts as bridge in the shortest path between two other nodes. As a part of the network analysis, we also measured these centrality metrics on our network models. The betweenness and degree centrality metrics were computed using the CytoHubba (Version 1.6) plugin for Cytoscape [81].

3.1.6.2. Community. MCODE (Molecular Complex Detection) (Version 1.32) was used in this study to detect densely connected modules in our networks [82]. MCODE uses a vertex-weighting metric which measures the cliquishness of a node's neighborhood. The haircut model was used with a value of three for the K-core to detect the clusters. The degree cut-off value, which indicates the minimum number of links for each node to be calculated, was set to two. We only considered the communities of four or more nodes.

3.1.7. Pair Scores

The pair score indicates the weight of the edge between two nodes according to the model it is calculated in. In the unweighted identity network, since the weights of all edges are equal to 1, the scores of all possible pairs are the same. For the weighted identity and similarity networks, we calculate the weight of the edge between each protein pair. In the weighted identity network we consider the number of identical

ligands two proteins share, whereas in the similarity network we also consider the chemical similarity of the ligands. As illustrated in Figure 3.1, in the weighted identity network the weight of the edge between nodes A and B is equal to 2, which indicates the score for the A–B pair is 2. In the similarity network, since the ligand pairs whose similarities are above 0.7 also contribute to the weight, the weight of the edge between nodes A and B becomes 2.8. Therefore, the score for the A–B pair is now equal to 2.8.

The pair scores help us to infer the protein pairs that are strongly associated based on their ligands. The complete list of the pairs in the weighted identity and similarity models are provided as a supplementary file.

3.2. Machine Learning

3.2.1. Data collection

The complete data set, which is used by Jacob *et al.*, is downloaded from http://bioinformatics.oxfordjournals.org/content/24/19/2149/suppl/DC1. Their data set of ligand interaction information for protein families, namely, GPCR and ion channels are provided by KEGG BRITE Database. The data is composed of the compounds of each target of the regarding protein family, all of which are formed by hierarchical subfamilies. For the experiments, for each hierarchy level only one protein is selected as a representative. Orphan targets and compounds without molecular description are filtered. Then, negative ligand-target pairs are produced by selecting compounds randomly from the compounds of the other targets. The reason behind this procedure is that some of these negative pairs may actually be positive since the targets have not been experimented for those particular compounds. Each data point consists of these three main components, (i) target ID, (ii) compound ID, and (iii) label for interaction status. Table 3.2 provides a simple illustration of the data points, for instance 1.2.7 is the family hierarchy ID for the protein selected as representative for 1.2.7 family in the GPCR data set.

As a result, for enzymes 2436 data points (1218 known pairs and 1218 generated false

Table 3.2. Illustration of sample data points that are generated for each protein family.

compound ID	**target ID**	**label**
D02066	1.2.7	true
C07533	1.2.7	true
C00029	1.2.7	false
D02066	1.2.9	true

pairs), formed by 675 proteins and 524 ligands; for GPCR 798 points, formed by 100 proteins and 219 compounds; and for ion channels 2330 data points, formed by interactions among 114 proteins and 462 ligands, are generated. We should report that the data points in the online provided data set are different than the reported version, thus we use the data illustrated below in Table 3.3. In this study we use the GPCR and ion channels data sets.

Table 3.3. Distribution of the proteins and ligands in the data set and the number of data points generated.

	GPCR	ion channels
proteins	100	114
ligands	219	462
data points	884	2330

3.2.2. Target Kernels

The hierarchy kernel is selected to measure protein similarity since it outperformed the other target kernels, namely Dirac, Mismatch, Local Alignment, and Multitask when tested with the 2D Tanimoto ligand kernel [3].

3.2.3. Ligand Kernels

In this study, instead of using the 2D Tanimoto similarity on compounds, we experiment with several different similarity kernels based on SMILES representation of compound data. The compounds in the data set provided as mol files were converted into unique SMILES strings via ChemAxon library JChem 6.0.2.215 .NET API in Visual Studio 2010 using C# programming language [46].

All of the kernels presented below were developed using the Java programming language on NETBeans IDE 7.4. The LCS, Combined LCS and Edit distance algorithms introduced in the Theory Chapter are used as similarity kernels without any modification.

3.2.3.1. LINGOsim Kernel. LINGOsim, introduced in the theory section, is based on LINGO length $q = 4$. Ligand kernels are created for the values of $q = 3, 4, 5$. We also introduce weighted LINGOsim, described in Equation 3.2, which assigns weight values for each LINGO in the whole data set.

$$weight_i = \frac{TF_i}{N} \tag{3.2}$$

where $TF(i)$ denotes the term frequency of LINGO of type i and N represents the number of unique LINGOs created from the whole compound SMILES database. We then rearrange the original LINGOsim equation as,

$$T_c = \frac{\sum_{i=1}^{m}(1 - \frac{|N_{S_1,i} - N_{S_2,i}|}{|N_{S_1,i} + N_{S_2,i}|})weight_i}{m} \tag{3.3}$$

where m is the number of LINGOs created from the SMILES string of any of the compounds while $N_{S_1,i}$ represents the number of LINGOs of type i in compound S_1, $N_{S_2,i}$ represents the number of LINGOs of type i in compound S_2 and $weight_i$ refers to the weight of LINGO of type i.

3.2.3.2. LINGO based TF-IDF cosine similarity. We propose a model where we combine LINGO representation with the TF-IDF weighting-scheme. TF-IDF has originally been developed in the Information Retrieval domain for weighting the words in the documents. Words are selected as terms of a document corpus and each document is treated as a collection of words (terms). In our model, we treat each SMILES string as a set of LINGOs and LINGOs are the terms of our model. LINGO length q is selected as 4 as it is in the original algorithm.

Let us assume we have the following compounds with the given SMILES in our ligand data set, $SMILES_1$:CCOC1c, $SMILES_2$:NCCOC1, and $SMILES_3$:cc1N. We first modify SMILES strings, then generate LINGO sets for each of them such that,

$$LINGOSet(SMILES_1) = CCOC, COC0, OC0c,$$
$$LINGOSet(SMILES_2) = NCCO, CCOC, COC0,$$
$$LINGOSet(SMILES_3) = cc0N.$$

Thus, we have a sample system that contains 5 unique LINGOs and 3 compounds. Table 3.4 below illustrates the frequency scores for all LINGOs(terms) for the corresponding SMILES string (document) which is utilized as set of LINGOs. For each SMILES string, occurences of the LINGOs are counted to find the term frequency.

Table 3.5 depicts the inverse document frequencies (IDFs) of all the LINGOs. IDF weighting-scheme allows the model to assign importance to the rare LINGOs. After term frequencies and inverted term frequencies of all the LINGOs are calculated, TF-IDF cosine similarity is applied to calculate the similarity between two compounds.

Table 3.4. TFs of each LINGO for the corresponding SMILES.

terms	$SMILES_1$	$SMILES_2$	$SMILES_3$
CCOC	1	1	0
COC0	1	1	0
OC0c	1	0	0
NCCO	0	1	0
CC0N	0	0	1

Table 3.5. IDFs of each LINGO are depicted.

terms	idf
CCOC	$log(3/2)$
COC0	$log(3/2)$
OC0c	$log(3/1)$
NCCO	$log(3/1)$
CC0N	$log(3/1)$

3.2.3.3. SMIfp Kernel. The original SMIfp method converts each SMILES string into a 34-dimensional scalar fingerprint, and then performs City Block Distance (CBD) on these vectors to find their similarity. In this study, we use Euclid and Tanimoto distances in addition to CBD in order to measure the effect of the distance metric on the model.

We then present a 38-dimensional model which is modified according to our SMILES database. First, we extract the number of occurrences for each character in our compound SMILES database for each of the protein families. Then, we compare the frequent characters with the original 34 character list of SMIfp. We find out that the characters '@', '\', '/', '.' which are ignored by the SMIfp method, are among the most frequent characters. In addition, the '%' character, which is listed as frequent, is a rather rare character. '@' and '@@' characters are called as chiral specification that indicates the arrangement of the neighbour atoms is anti-clockwise and clock-wise respectively. '/' and '\' are the directional bonds. Therefore, we replace '%' with '.', and we add four more characters, '@', '@@', '\', and '/.' We then, use Euclid, CBD and Tanimoto to measure the similarity between two 38-dimensional scalar fingerprints.

3.2.3.4. SMILES based substring similarity kernel. We use SMILES based similarity method of Cao *et al.* that utilizes common substring frequencies. We also tested a modified version of this kernel, where we replace ring numbers [1-9] with 0. We refer to this model as modified substring similarity kernel.

3.2.4. Experiment Setup

Jacob *et al.* follow two procedures for each target using the described data points created for each protein family data set [3].

- For each target protein p, the first procedure generates k-folds where $k = min(n, 10)$ for each data point where n is the number of data points related to p. Then it then utilizes the data points of the other target proteins as training samples along

with some of the data points of p itself.

For instance, let us assume our target protein is 1.2.9 (Adrenergic receptor, beta 3) from the GPCR family, thus data points created for this target look like as it is depicted in Table 3.6. Each line represents a data point in which label indicates the presence of an interaction between the given target and drug. Folds are named as 'train' and 'test' to indicate the role of the data point in the classification. As it is shown in Table 3.6, target 1.2.9 is trained with other proteins for the first fold, whereas in the second fold it is used as a test sample.

Table 3.6. Sample data points created for target 1.2.9 (Adrenergic receptor, beta 3).

compound ID	**target ID**	**label**	**folds**
D02066	1.2.7	true	train train train
C07533	1.2.7	true	train train train
C00029	1.2.7	false	train train train
D02066	1.2.9	true	train test train

Briefly on each fold, SVM classifier is trained with the data points marked as train and is tested on the unused data points of p. The idea behind this experiment is to quantify the impact of considering ligands of other proteins for each protein tested on the performance of the classifier.

- The second procedure uses data points of target p only for testing and trains the SVM on the remaining data points that belong to other proteins. By this way, it aims to measure the success of prediction.

In our study, we hold the first experiment using the provided data points to test our ligand kernels.

4. RESULTS

4.1. Ligand-centric β-lactamase superfamily networks

In this section we provide a detailed examination of the β-lactamases and PBPs in the Protein Data Bank database with the ligands that they bind to. We also discuss the three different network models constructed using ligand-binding information.

4.1.1. Database

4.1.1.1. Proteins. Our protein data set which was collected from PDB contains 111 protein structures with unique UniProt accession numbers, all of which bind to at least one ligand from the ligand data set. 43 of these proteins bind to a single ligand. The distribution of the β-lactamases based on the Ambler classification of β-lactamases is as follows: 28 Class A β-lactamases, 17 Class B (Metallo) β-lactamases, 10 Class D β-lactamases and 10 Class C β-lactamases. Our data set also contains 37 PBPs. There are 9 proteins that do not fit into any of these groups including the β-lactamases with no Ambler class definition. β-lactamase ampC (P00811) has the highest number of ligands (61 ligands). It is followed by SHV-1 (P0AD64), β-lactamase CTX-M-9a (Q9L5C8), β-lactamase blaA (P0C5C1), DD-carboxipeptidase from Actinomadura sp. organism (P39045), and DD-carboxipeptidase from Streptomyces sp. organism (P15555) all of which bind to more than 15 ligands.

4.1.1.2. Ligands. Our ligand data set includes 304 ligands with unique PDB identifiers (IDs), 222 of which only bind to a single protein in the data set. IM2 (Imipenem) interacts with the highest number of proteins (12 proteins). It is followed by KCX (Lysine NZ-Carboxylic acid) with 11 protein interactions, and PNM (Penicillin G) and MES (2-(n-morpholino)-Ethanesulfonic acid) with 10 protein interactions. The molecular weights of the ligands mostly vary between 30 and 800. The mean of the molecular weights of the ligands is 304 and the median is 301. There are 15 ligands

whose molecular weights are smaller than 100. The distribution of the ligand molecular weights is provided in Figure 4.1a.

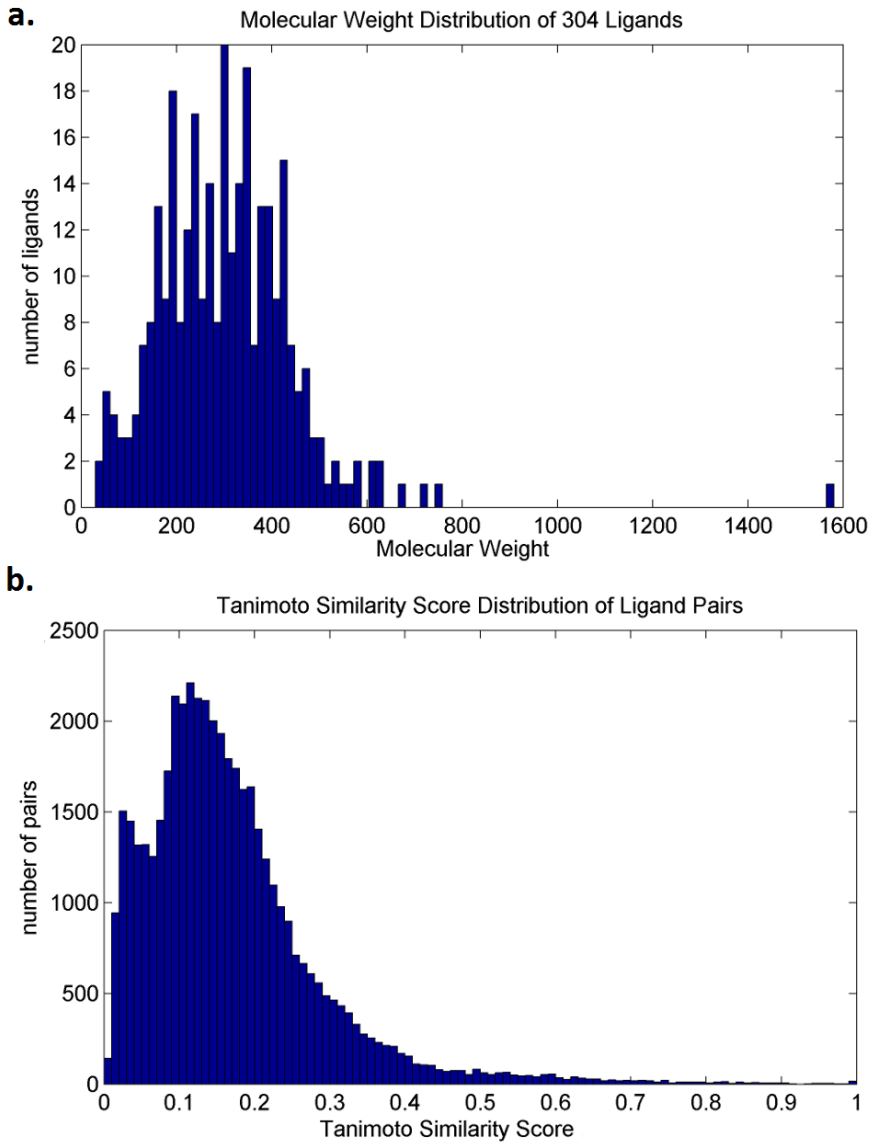

Figure 4.1. MW and Tanimoto similarity score distribution of the 204 ligands in our data set. (a) Distribution of the molecular weights. (b) Distribution of Tanimoto chemical similarity scores for the 47056 pairs.

The distribution of the pairwise Tanimoto similarity amongst 304 ligands is shown in Figure 4.1b. These data correspond to all possible unique pairs of ligands, excluding ligand A–ligand A pairs, where A represent any ligand in the data set, which always

yield a similarity score of 1. 255 out of 46056 ligand pairs have Tanimoto similarity scores above 0.7. The majority of the ligand pairs have similarity scores between 0.1 and 0.2. The mean score is 0.169 and the median is 0.146.

Since only 82 out of 304 ligands bind to more than one protein in our dataset, the weighted identity network is created by 82 ligands. However, with the use of ligand similarity, the similarity network is constructed by 152 ligands which almost double the number of contributor ligands in the weighted identity network. Besides, the ligands added to the network model are mostly β-lactamase and PBP oriented ones.

4.1.1.3. Comparison of Sequential and Functional Similarities. The multiple sequence alignment of 111 protein sequences, which is performed using COBALT, shows similarity with the Ambler's classification scheme of β-lactamases (Figure 4.2a). PBPs are sequentially closer to the Class D β-lactamases. Average hierarchical clustering based on pairwise ligand similarity for 304 ligands is performed using ChemMine (Figure 4.2b) and each branch is colored according to the protein class that a ligand binds to. A total of 249 ligands bind to the same class of proteins, and four ligands only bind to transesterase protein, two ligands only bind to d-amino acid amidase, and two ligands bind to unknown β-lactamase like proteins. The rest of the ligands bind to more than one different classes of proteins which are left colourless. The phylogenetic trees for both of the representations are visualized using Interactive tree of life (ITOL) [83].

The majority of the ligands bind to either Class A β-lactamases or PBPs. There are 78 ligands that only bind to Class A β-lactamases and 75 ligands that only bind to PBPs. The number of ligands that only bind to Class C β-lactamases is 53, whereas the number of ligands that only bind to Class B is relatively less, 30. Only five ligands are identified which only bind to Class D β-lactamases whereas, 17 ligands are found to bind to Class D β-lactamases along with other classes. We can infer that most of the ligands that bind to Class D β-lactamases, also bind to other classes of proteins. The ligand similarity tree seems more diverse, yet we can capture some small clusters of ligands which reflect a protein class such as PBPs, Class C β-lactamases, and Class A

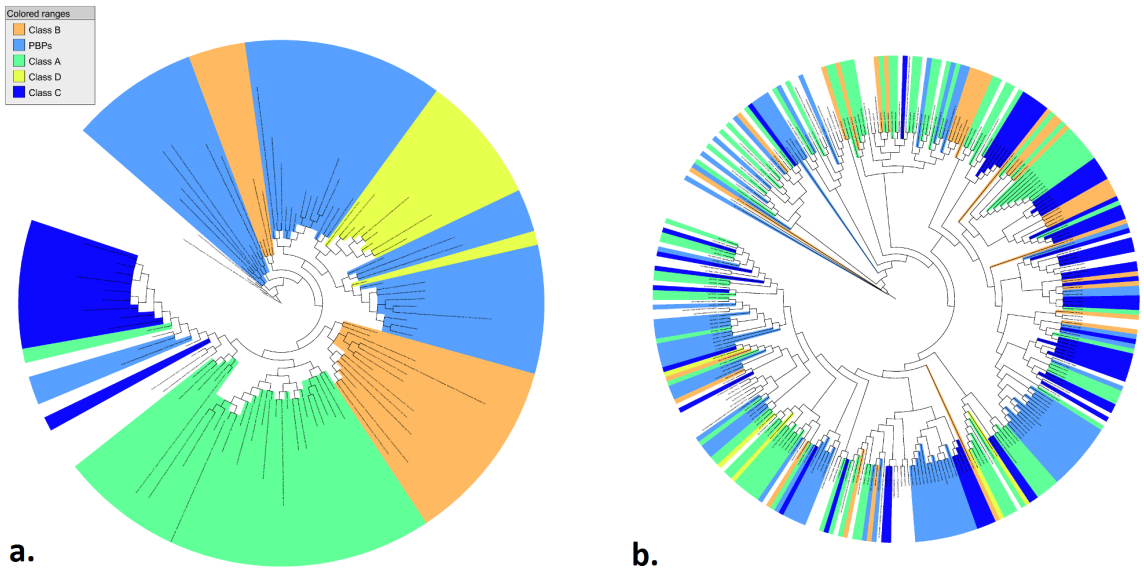

Figure 4.2. Clustering of proteins and compounds. (a) Multiple sequence alignment of 111 protein sequences. (b) The hierarchical clustering of the 304 ligands. (Blue: PBPs, Green: Class A, Dark Blue: Class C, Yellow: Class D, Orange: Class B).

β-lactamases. The diversity of the proteins according to their ligand chemical similarity suggests that clustering of proteins based on ligand sharing can lead interesting results.

4.1.2. Protein-ligand binding networks

It is important to detect the densely connected communities and to identify the central nodes using centrality metrics such as betweenness and degree centralities for a better understanding of biological networks. In this study, where the network nodes represent proteins and the edges represent shared or chemically similar ligands, identifying central nodes and communities yielded important clues on a ligand centric classification of β-lactamases. We now explain the similarities and differences between the three networks.

4.1.2.1. Unweighted Identity protein-ligand binding network. The unweighted network connects target proteins that share a common ligand regardless of the number of shared

ligands between them. The unweighted network comprises 99 proteins. 12 of the 111 proteins in our initial protein data set are excluded, since they do not share any ligands with other proteins. In the unweighted network three densely connected clusters are detected (Table 4.1). These clusters are ranked according to their MCODE scores, calculated by multiplying the density of the cluster by the number of the members. 46 of the 99 proteins are placed within a cluster.

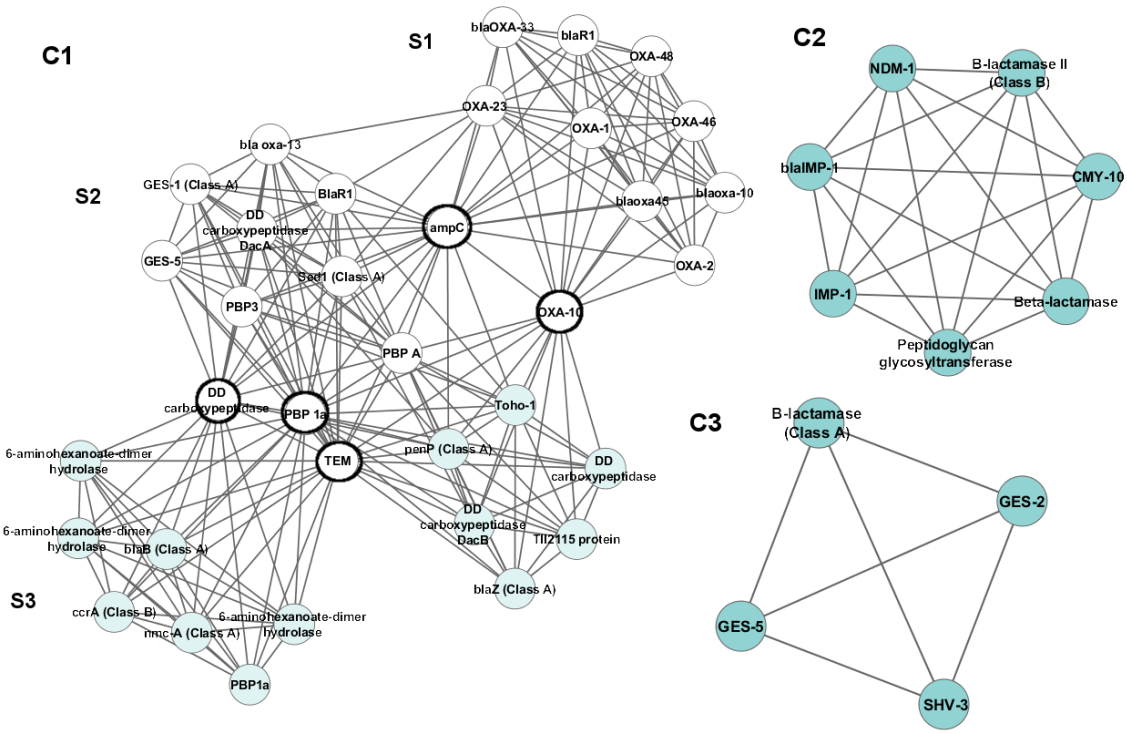

Figure 4.3. Communities in the unweighted identity network. Nodes are colored according to their MCODE scores. From blue to white, the scores of the nodes increase. (The same colouring scheme is used for the other community figures.).

Cluster 1 has the highest MCODE score and contains 35 proteins. It includes the top-three highest scoring nodes in terms of degree centrality in the unweighted network, namely ampC, TEM and PBP 1a (Figure 4.3). It is not suprising that both ampC and TEM are listed as the central nodes since they are among the most studied β-lactamases. Therefore, this knowledge suggests that PBP 1a is also experimented with different ligands. The majority of the proteins in the cluster are Group 2 β-

Table 4.1. Communities in the Unweighted Identity Network.

	Num	Names
Cluster 1		
Class A	9	blaB (P52664), penP (P00808), TEM (P62593), blaZ (P00807), SED-1 (Q93PQ0),GES-1 (Q9KJY7), GES-5 (Q09HD0), Toho-1 (Q47066) , nmc-A (Q7ATJ4)
Class B	1	ccrA (P25910)
Class C	1	ampC (P00811)
Class D	10	OXA-10 (P14489), OXA-1 (P13661), OXA-2 (P0A1V8), OXA-23 (Q9L4P2), OXA-46 (Q8GRH0), OXA-48 (Q6XEC0), blaOXA-10 (Q7BNC2), blaOXA-13 (Q51400), blaOXA-33 (Q8RLA6), blaOXA-45 (Q7WZC7)
PBP	10	(2 ×) PBP1a (Q9RET4, G1C794), PBP3 (Q51504), (2 ×) DD carboxipeptidase (P15555, P39045), DD carboxipeptidase DacB (P24228), PBP A (P71586), DD carboxipeptidase DacA (P0AEB2), (2 ×) BlaR1 (Q7WU28, P18357)
Others	4	TII2115 protein (Q8DH45), (3 ×) 6-aminohexanoate-dimer hydrolase (Q59710, P07061, P07062)
Cluster 2		
Class B	4	NDM-1 (C7C422), IMP-1 (P52699), blaIMP-1 (Q79MP6), β-lactamase II (P04190)
Class C	1	CMY-10 (Q99QC1)
PBP	1	Peptidoglycan glycosltransferase (C8WPP1)
Others	1	β-lactamase (D1C5R0)
Cluster 3		
Class A	4	GES-2 (Q93F76), GES-5 (A0EL75), SHV-3 (P30896), β-lactamase (Q8EMP8)

lactamases and PBPs. The cluster is composed of three sub clusters where the first subgroup (S1) contains only Class D β-lactamases except for BlaR1 (Q7WU28), which has been shown to have high structural similarity with Class-D β-lactamases [84]. The S1 subgroup is connected by the KCX ligand.

The second subgroup (S2) of the cluster is dominated by Class A β-lactamases and PBPs, whereas the third subgroup (S3) does not define a theme, but contains three

6-aminohexanoate-dimer hydrolases which belong to the β-lactamase family according to their PFAM identification and which show similarities with Class A and Class C β-lactamases [85, 86]. S2 is connected mostly by the IM2 and PNM ligands, whereas S3 is constructed by the MES ligand edges.

AmpC, TEM, PBP1a, OXA-10 and DD carboxipeptidase, indicated by bold black circles in Figure 4.3, have the highest betweenness centralities and act as bridges between the subgroups of Cluster 1. Betweenness centrality is the indicator of how much the nodes affect the flow of information through the network. In ligand sharing network, it can be said that the nodes listed as important according to betweenness cenrality have the ability of binding even the distant nodes. Further investigation of these nodes may give clues about ligand-mediated evolution.

Cluster 2 contains seven proteins; four Class B β-lactamases as well as the Class C β-lactamase CMY-10, β-lactamase with no Ambler class identification (D1C5R0) and a peptidoglycan glycosltransferase protein (C8WPP1). ACY (Acetic Acid) is the only ligand that connects this cluster. Cluster 2 is fully connected since ACY only binds to these proteins. Except for CMY-10, D1C5R0 and C8WPP1, all other proteins in the network bind to other ligands besides ACY. Acetic acid is a rather non-specific ligand when compared to the other ligands.

Coordinates for β-lactamase with no Ambler class definition was deposited in 2012 under the name of "beta-lactamase from Sphaerobacter thermophilus dsm 20745" and publication for its structure is not yet available. The sequence is 31.5%[1] similar to a Class A β-lactamase. The fact that it is clustered with Class B β-lactamases in the ligand-based network, might indicate that it is a Class B β-lactamase.

Cluster 3. All four proteins in this cluster are Class A β-lactamases. Class A β-lactamase (Q8EMP8) so called, OIH-1, was described as not only the first example of antibiotic resistance that evolved in deep-sea, but also as the most highly halo-tolerant enzyme discovered [87]. All edges in this network are formed by the EPE

[1]Protein sequence similarity search through PDBe (http://www.ebi.ac.uk/)

(Ethanesulfonic Acid) ligand.

4.1.2.2. Weighted Identity protein-ligand binding network. Weighted identity network is the weighted version of the unweighted network model, where the number of shared ligands is also considered. The weighted identity network consists of 99 nodes. The network is constructed by number of 481 interactions between protein pairs. Six densely connected clusters are detected that include 53 out of the 99 nodes (Table 4.2). It is observed that with the use of weighting based on the number of shared ligands; the communities identified by MCODE have changed. Instead of having one big cluster (Cluster 1 in the unweighted identity network), we now have several smaller communities. Indeed, the first three clusters of the weighted identity network correspond to the three sub clusters of Cluster 1 of the unweighted network (Figure 4.4).

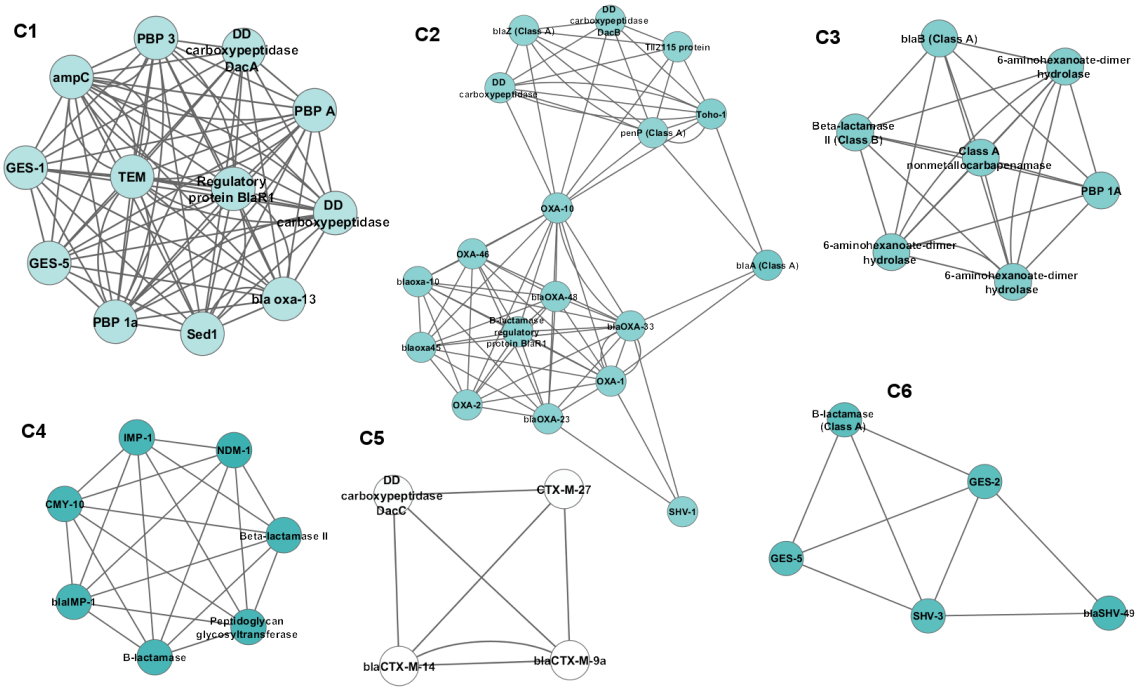

Figure 4.4. Communities in the weighted identity network. Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, and Cluster 6.

Cluster 1. There are 12 members in Cluster 1 half of which belong to PBPs. Cluster 1 of the weighted identity network comprises 12 proteins from the second

Table 4.2. Communities in the Weighted Identity Network.

	Num	Names
Cluster 1		
Class A	4	TEM (P62593), SED-1 (Q93PQ0), GES-1 (Q9KJY7), GES-5 (Q09HD0)
Class C	1	ampC (P00811)
Class D	1	blaOXA-13 (Q51400)
PBP	6	PBP1a (G1C794), PBP3 (Q51504), PBP A (P71586), DD carboxipeptidase (P39045), DD carboxipeptidase DacA (P0AEB2), BlaR1 (P18357)
Cluster 2		
Class A	5	SHV-1 (P0AD64), blaZ (P00807), penP (P00808), blaA (P0C5C1), Toho-1 (Q47066)
Class D	9	OXA-1 (P13661), OXA-2 (P0A1V8), OXA-10 (P14489), blaOXA-10 (7BNC2), blaOXA-23 (Q9L4P2), blaOXA-33 (Q8RLA6), blaOXA-45 (Q7WZC7), OXA-46 (Q8GRH0), blaOXA-48 (Q6XEC0)
PBP	3	DD carboxipeptidase DacB (P24228), DD carboxipeptidase (P15555), BlaR1 (Q7WU28)
Others	1	TII2115 protein (Q8DH45)
Cluster 3		
Class A	2	blaB (P52664), nmc-A (Q7ATJ4)
Class B	1	β-lactamase II (P25910)
PBP	1	PBP 1A (Q9RET4)
Others	3	(3 $\times$) 6-aminohexanoate-dimer hydrolase (Q59710, P07061, P07062)
Cluster 4		
Class B	4	NDM-1, β-lactamase II (P04190), IMP-1 (P52699), blaIMP-1 (Q79MP6)
Class C	1	CMY-10 (Q99QC1)
PBP	1	Peptidoglycan glycosyltransferase (C8WPP1)
Others	1	β-lactamase (D1C5R0)
Cluster 5		
Class A	3	blaCTX-M-9a (Q9L5C8), blaCTX-M-14 (Q9L5C7), CTX-M-27 (Q840M4)
PBP	1	D carboxipeptidase DacC (P08506)
Cluster 6		
Class A	5	GES-2 (Q93F76), GES-5 (A0EL75), SHV-3 (P30896), blaSHV-49 (Q5VCA8), β-lactamase (Q8EMP8)

subgroup (S2) of Cluster 1 of the unweighted identity network.

Cluster 2 is formed by 18 nodes. It includes the first subgroup (S1) of Cluster 1 of the unweighted network, as well as six members from S2 of Cluster 1 of the unweighted network, which are displayed as blue nodes in Figure 4.3, and BlaA and SHV-1. Cluster 2 has two subunits, one is dominated by Class A β-lactamases which are connected by PNM and the second one is dominated by Class D β-lactamases connected by KCX. The Tll2115 protein in the first subunit is defined as PBP-A and described to be highly related to Class A β-lactamases, mostly TEM-1 [88].

OXA-10 and BlaA act as bridges where they connect two subunits in this cluster. OXA-10 is connected to the first subunit with PNM edges and to the second subunit with KCX edges. BlaA, on the other hand, is connected to the first subunit with PCZ (Cefotaxime product, open form) ligand edges and to the second subunit with DRW (Doripenem, open form) ligand edges.

Cluster 3 is formed by the third subgroup (S3) of Cluster 1 of the unweighted identity network. It includes seven nodes: Three 6-aminohexanoate-dimer hydrolases, Blab and Nmc-A from Class A β-lactamases, PBP 1a and CcrA from Class B β-lactamases. MES and ACA (6-aminohexanoic acid) ligands connect this community.

Cluster 4 is the exact replicate of the Cluster 2 of the unweighted identity network, which consists of seven proteins and contains only ACY ligand edges.

Cluster 5 consists of four proteins, three of which belong to the CTX-M family and the fourth is DD carboxipeptidase DacC, which is also known as PBP 6. In the article which defines PBP [89], it was stated that active site configuration and the topography of its domain shows similarities to Ambler Class A. SUC (Sucrose) and CB4 (Pinacol) ligands connect this cluster.

Cluster 6 is constructed by the addition of SHV-49 to Cluster 3 of the unweighted identity network. Ligands, which connect this cluster, are EPE, MA4 (Cyclohexyl-

Hexyl-Beta-D-Maltoside) and TBE (Tazobactam intermediate).

4.1.2.3. Similarity protein-ligand binding network. Our aim for constructing a weighted similarity network is both to observe the contribution of ligand chemical similarity to the existing protein interactions and to identify possible relationships between proteins, even if they do not share any ligands, but bind to ligands that have high chemical similarity. Two ligands are considered similar if their Tanimoto coefficient of chemical similarity is above 0.7. To determine the strength of the relationship between two nodes, we sum up the pairwise similarities of their ligands that pass the threshold.

The similarity network utilizes the assumption that 'chemically similar ligands should target same proteins'. Therefore, weight of the edges we observe between two nodes (proteins) is the indicator of how much those nodes (proteins) are similar to each other in terms of biological activity. As the weight of the edge among two nodes increase, it is more likely that they are worth for further investigation.

The similarity network contains 100 nodes. Using ligand similarity enabled the inclusion of β-lactamase BlaC (A5U493), which only binds to DWZ ((2S,3R,4S)-4-[(3S,5S)-5-(dimethylcarbamoyl)pyrrolidin-3-yl]sulfanyl-2-[(1S,2R)-1-formyl-2-hydroxy–propyl]-3-methyl-3,4-dihydro-2H-pyrrole-5-carboxylic acid), to the network. The high similarity of DWZ with DRW and 2RG (Ertapenem) ligands, which are already in the network, resulted in the addition of a new node to the network. The networkcontains 1447 interactions, which is three times the number of interactions in the weighted identity network. 69 proteins are placed in five clusters (Figure 4.5 and Table 4.3). The use of similarity also leads to an increase in the number of proteins placed within clusters in the similarity network.

Cluster 1 is the highest scoring cluster and 17 out of 34 proteins in this cluster are PBPs and Group 2 Beta-lactamases. This community is formed by two subgroups, where the first one contains mostly PBPs and Class A β-lactamases, as well as ampC, OXA-10, Tll2115 and NDM-1. The relationships established within this subgroup are

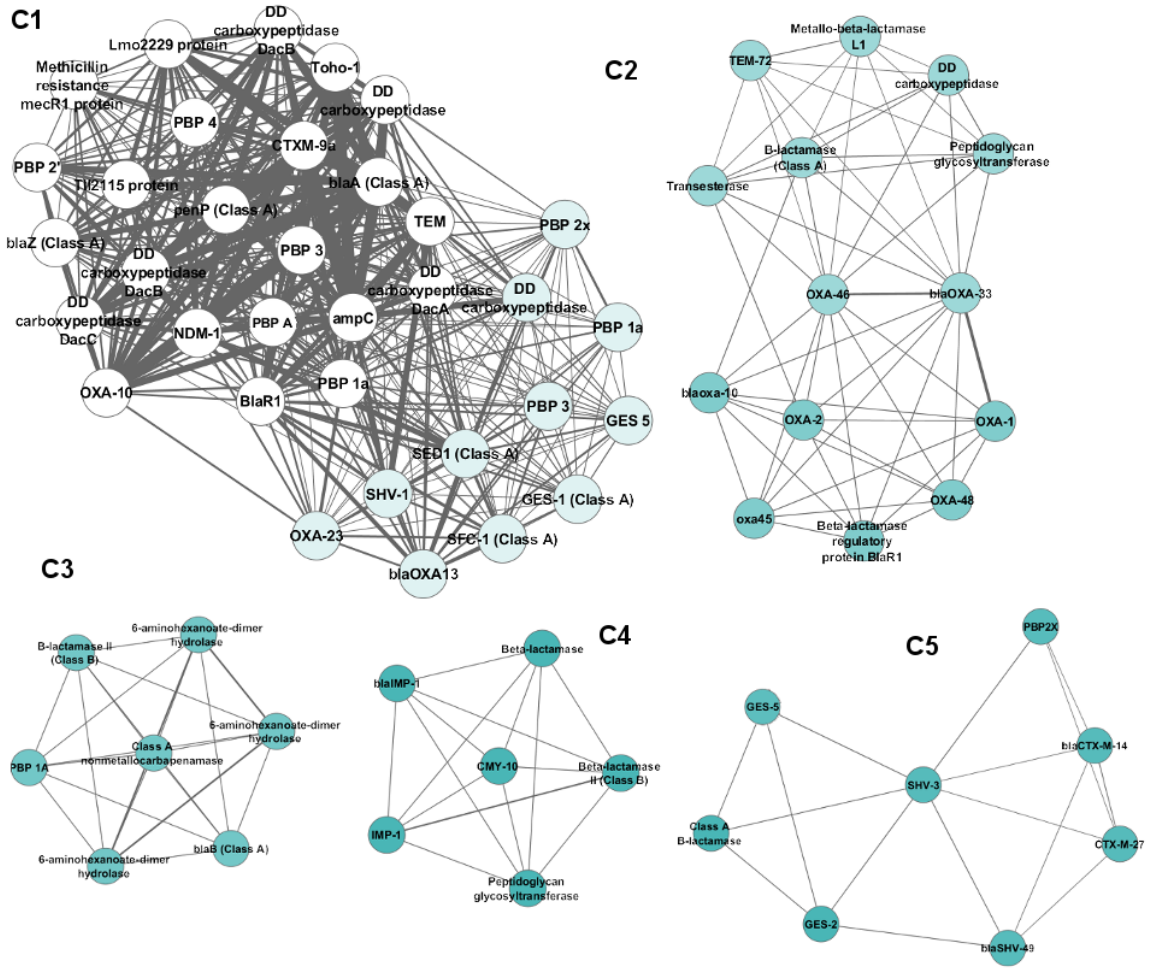

Figure 4.5. Communities in the similarity network. Cluster 1, Cluster 2, Cluster 3, Cluster 4, and Cluster 5.

Table 4.3. Communities in the Similarity Network.

	Num	Names
Cluster 1		
Class A	11	SHV-1 (P0AD64), SED-1 (Q93PQ0), SFC-1 (Q6JP75), GES-1 (Q9KJY7), GES-5 (Q09HD0), TEM (P62593), blaA (P0C5C1), penP (P00808), CTX-M-9a (Q9L5C8), blaZ (P00807), Toho-1 (Q47066)
Class B	1	NDM-1 (C7C422)
Class C	1	ampC (P00811)
Class D	3	OXA-10 (P14489), blaOXA-13 (Q51400), OXA-23 (Q9L4P2)
PBP	17	PBP2X (P14677), (2 ×) PBP1a (G1C794, Q04707), PBP2" (Q93IC2), PBP4 (Q5HI26), (2 ×) PBP3 (G3XD46, Q51504), PBP A (P71586), (2 ×) DD carboxipeptidase (P39045, P15555), DD carboxipeptidase DacA (P0AEB2), (2 ×) DD carboxipeptidase DacB (P24228, P45161), DD carboxipeptidase DacC (P08506), BlaR1 (P18357), mecR1 (P0A0B0), Lmo2229 protein (Q8Y547)
Others	1	TII2115 protein (Q8DH45)
Cluster 2		
Class A	2	TEM-72 (Q9R429), β-lactamase (P94458)
Class B	1	Metallo L1 (P52700)
Class D	7	OXA-1 (P13661), OXA-2 (P0A1V8), blaOXA-10 (Q7BNC2), blaOXA-33 (Q8RLA6), blaOXA-45 (Q7WZC7), OXA-46 (Q8GRH0), blaOXA-48 (Q6XEC0)
PBP	3	DD carboxipeptidase (Q6MHT0), BlaR1 (Q7WU28), Peptidoglycan glycosyltransferase (C8W8H7)
Others	1	Transesterase (Q9Y7D1)
Cluster 3		
Class A	2	blaB (P52664), nmc-A (Q7ATJ4)
Class B	1	β-lactamase II (P25910)
PBP	1	PBP1A (Q9RET4)
Others	3	(3 ×) 6-aminohexanoate-dimer hydrolase (Q59710, P07061, P07062)
Cluster4		
Class B	3	IMP-1 (P52699), blaIMP-1 (Q79MP6), β-lactamase II (P04190)
Class C	1	CMY-10 (Q99QC1)
PBP	1	Peptidoglycan glycosyltransferase (C8WPP1)
Others	1	β-lactamase (D1C5R0)
Cluster 5		
Class A	7	GES-2 (Q93F76), GES-5 (A0EL75), SHV-3 (P30896), blaSHV-49 (Q5VCA8), bla-CTX-M-14 (Q9L5C7), CTX-M-27 (Q840M4), β-lactamase (Q8EMP8)
PBP	1	PBP 2X (P59676)

strong judging by the edge weights. Therefore, since these proteins share chemically similar ligands, they may be expected to express similar biological activities. PBPs in this subgroup do not reflect any classification scheme, since there are both high MW and low MW PBPs included. The first subgroup is connected with thicker edges when compared to the second and smaller subgroup. There are many different ligands contribute to the first subgroup, but PNM is the most frequent one.

The second subgroup of Cluster 1 contains four PBPs and the rest is Group 2 β-lactamases. PBPs in this subgroup belong to high MW PBP class. The interaction between SFC-1, blaOXA-13 and SED-1, where all three are connected to each other with thick edges is conspicuous when compared to the other edges within this subgroup. The edges within the second subgroup are mostly formed by interactions with two ligands. The most frequent ligands in this subgroup are IM2 and MER (Meropenem, bound form).

Cluster 2 is formed by two subgroups, where OXA-46 and blaOXA-33 act as bridges with the highest betweenness centrality in this cluster. While one of the subgroups contains five OXA β-lactamases and BlaR1 (Q7WU28), the other subgroup contains β-lactamases from Class B, Class A, PBP and a transesterase protein.

Both Cluster 1 and Cluster 2 contain BlaR1 proteins in which BlaR1 (P18357) in Cluster 1 is grouped with PBPs and other proteins, while BlaR1 (Q7WU28) in Cluster 2 is grouped with OXA β-lactamases. The reason they are separated is that they bind to different types of ligands.

Cluster 3 of the similarity network is the exact replica of Cluster 3 of the weighted identity network and contains seven nodes.

Cluster 4 contains six nodes. It is similar to Cluster 4 of the weighted identity network, except for the NDM-1 protein which is placed in Cluster 1 in the similarity network. Apart from losing NDM-1, we capture a thicker edge between IMP-1 and β-lactamase II this time due to their sharing of chemically similar ligands, which are

OCS (Cysteinesulfonic acid) and CSW (Cysteine sulfinic acid).

Cluster 5. It is formed by eight nodes where the only protein which is not a Class A β-lactamase is PBP2x. PBP2x is a member of high MW class B PBPs which are sequentially close to Class A β-lactamases [14]. Cluster 5 is expanded from Cluster 6 of the weighted identity network with the addition of CTX-M-14, CTX-M-27 and PBP2x.

4.1.2.4. Overall Discussion of the Network Models. Firstly, moving from the unweighted identity network to the similarity network, we observed an enormous increase not only in the number of connections established but also in the number of nodes placed in the clusters. The inclusion of ligand similarity information in protein-ligand interaction network has shown to provide useful clues on detecting densely connected clusters. For example, we observe that the weighted identity network is a denser version of the unweighted identity network, where a single cluster (Cluster 1) of the unweighted identity network is divided into three clusters (Cluster 1, Cluster 2, and Cluster 3) in the weighted identity network. With the similarity network, we were able to capture strong relationships between protein pairs when ligand similarity is considered compared to the two identity networks. Use of ligand similarity also led to increase in the number of active ligands in the network which are also more β-lactamase and PBP oriented ones.

Analysing the three network models reveals the evolution of a specific cluster when it is expanded moving from the unweighted identity network towards the similarity network (Figure 4.6a). Cluster 3 of the unweighted identity network, which is formed by four Class A β-lactamases, is observed as Cluster 6 in the weighted identity network with the addition of blaSHV-49. In the similarity network, CTX-M-27, CTX-M-14 and PBP2x are included into this cluster. Out of five connections that bind these three proteins to the Cluster 6 of the weighted identity network, four are made with the use of ligand similarity information, namely similarity of SUC and MA4. We also observe that one of the clusters is quite similar in the three network models (Figure 4.6b). In

the unweighted identity and weighted identity networks, all of the edges are ACY. In the similarity network, besides ACY we also observe OCS and CSW ligands. Except for the loss of NDM-1 protein from this cluster in the similarity network, the core six proteins remain together in all three models. Due to the use of ligand similarity, NDM-1 establishes strong relationships with other proteins such as BlaC, OXA-10, and DD carboxipeptidase DacB, which forced it to change its cluster. Grouping of NDM-1 with other proteins when ligand similarity is used highlights the impact of ligand chemical similarity.

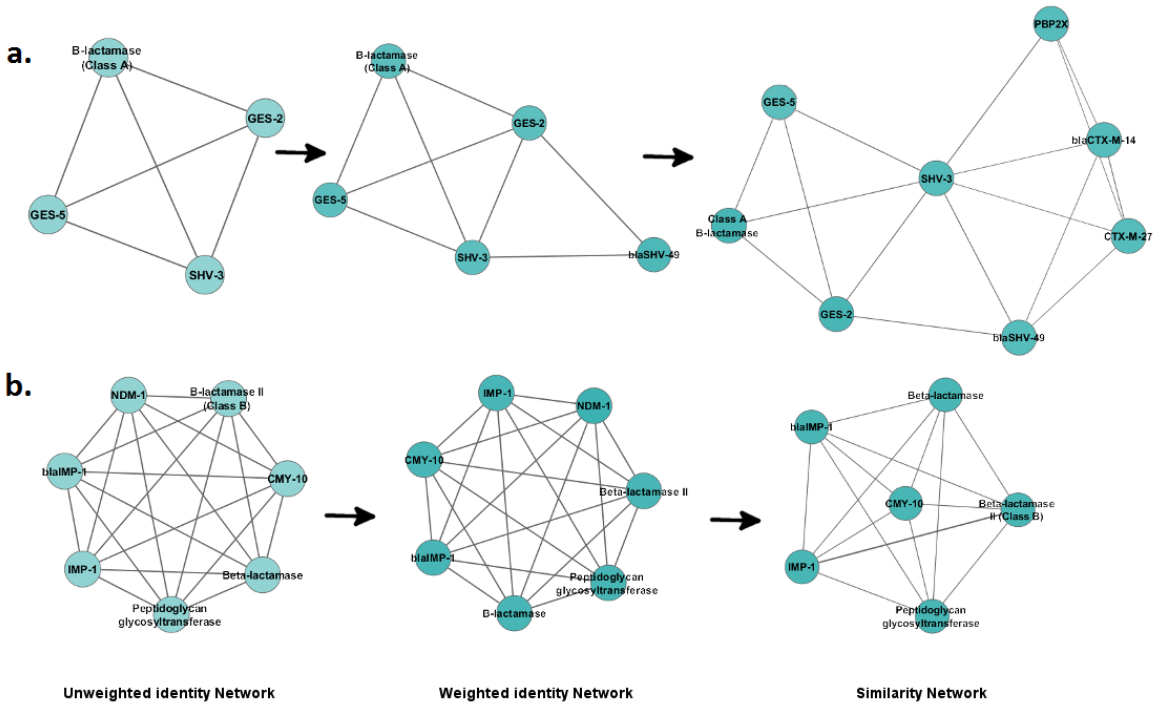

Figure 4.6. Evolution of a cluster during the change of the network models. (a) The cluster gains new nodes as the network model changes from the unweighted identity to similarity. (b) The cluster loses NDM-1 in the similarity network model.

4.1.3. Protein pairs with high scores

Top scoring pairs describe the protein pairs with highest edge weights. Top scoring pairs can be calculated only for the weighted identity and similarity networks, where the edges are weighted. This feature enables the observation of pairs that have

strong connections.

In the weighted identity network, there are 416 pairs, the highest score of which is 6 and the lowest score is 1. This means that the highest scoring pair shares six identical ligands, whereas the lowest scoring pair only shares a single identical ligand. The average score is 1.15. We selected the value of 3 as a threshold for a protein pair to be accepted as high scoring for the weighted identity network, since there are nearly 40 pairs with a score of 2. There are 10 pairs whose score is equal to or above 3 (Table 4.4). In the similarity network, 748 pairs are reported in which the highest score is 11.42 and the lowest score is 0.71. The average score is 1.7. We selected the pairs with a score value higher than 7 (Table 4.5). When we compare the top-pairs in the models, half of the top-scoring pairs in the weighted identity network consist of proteins from the same classes such as SHV-3 – SHV-1, and OXA-1 – OXA-33. However, in the similarity network there are only two pairs whose members are from the same class.

Table 4.4. Top pairs in the weighted identity network.

UniProt acc. num.	**Name**	**Class**	**UniProt acc. num.**	**Name**	**Class**	**Weight**
P00811	ampC	C	Q9L5C8	CTX-M-9a	A	6
P00811	ampC	C	Q47066	Toho-1	A	5
P62593	TEM	A	P00811	ampC	C	4
P62593	TEM	A	G1C794	PBP 1a	PBP	3
Q47066	Toho-1	A	P15555	DD carbox.	PBP	3
P62593	TEM	A	Q9L5C8	CTX-M-9a	A	3
P30896	SHV-3	A	P0AD64	SHV-1	A	3
P13661	OXA-1	D	Q8RLA6	OXA-33	D	3
P0AD64	SHV-1	A	Q5VCA8	blaSHV-49	A	3
P08506	PBP 6	PBP	Q54113	PBP 2'	PBP	3

The first two pairs in both of the models are the same, namely ampC – CTX-M-9a and ampC – Toho-1. It is not surprising that ampC took the lead in both models, since it is one of the most studied β-lactamases. It is reported to have almost

Table 4.5. Top pairs in the similarity network.

UniProt acc. num.	Name	Class	UniProt acc. num.	Name	Class	Weight
P00811	ampC	C	Q9L5C8	CTX-M-9a	A	11.42
P00811	ampC	C	Q47066	Toho-1	A	9.32
P0C5C1	β-lactamase blaA	A	C7C422	NDM-1	B	8.94
P00811	ampC	C	P0C5C1	β-lactamase blaA	A	8.65
Q47066	Toho-1	A	P0C5C1	β-lactamase blaA	A	8.55
Q9L5C8	CTX-M-9a	A	P0C5C1	β-lactamase blaA	A	7.96
P0C5C1	β-lactamase blaA	A	P15555	DD carbox.	PBP	7.86
P14489	OXA-10	D	P0C5C1	β-lactamase blaA	A	7.57
Q9L5C8	CTX-M-9a	A	P08506	DacC (PBP6)	PBP	7.57
Q9L5C8	CTX-M-9a	A	P24228	DacB (PBP4)	PBP	7.41
P0C5C1	β-lactamase blaA	A	P24228	DacB (PBP4)	PBP	7.27
P62593	TEM	A	P00811	ampC	C	7.26

62000 bioactivities of which nearly 390 are known to have Ki and IC50 values in ChEMBL (ChEMBL2026)[2] . Furthermore, CTXM-9a is also one of the best binders with high ligand count. For ampC and Toho-1, apart from their sharing of five common ligands, we can infer that the other ligands they bind to are chemically similar. Toho-1 is a CTX-M type Group 2 β-lactamase and 7 ligands are reported in ChEMBL (ChEMBL16976753) [23].

The other pairs in the Tables 4.4 and 4.5 are different from each other. Use of similarity highlighted interactions with β-lactamase BlaA and more PBPs in the similarity network model. β-lactamase BlaA, also known as BlaC, is encoded by Mycobacterium tuberculosis which is a dangerous pathogen that causes tuberculosis, and kills 1.3 million people every year according to the report of WHO in 2013 [90]. BlaC is a Class A and Group 2 β-lactamase and reported to interact with 20 ligands in PDB. It connects to 18 different proteins via 8 ligands in the weighted identity network in PDB. In the similarity network, it is observed that BlaC interacts with 37 proteins via 16 ligands. BlaC shares similar ligands with NDM-1, ampC, Toho-1, CTX-M-9a

[2]The IDs start with 'ChEMBL' denotes ChEMBL identifiers.

and DD carboxipeptidase. A recent study showed that BlaC is irreversibly inhibited by NXL104 (Avibactam), a β-lactamase inhibitor, and clavulanic acid [91]. Therefore, considering the strong relationships between BlaC and the proteins listed above, we suggest that these proteins might also interact with NXL104. A previous study that reported the inhibition of Class A (CTX-M-15: Q9EXV5) and Class C (ampC: P24735) β-lactamases by NXL104 [92] also strengthens our suggestion.

OXA-10 and BlaC pairing is another interesting relationship identified through the similarity network, although they are not reported to share an identical ligand in PDB, the similarity of their ligands is 7.57, which is considerably high. OXA-10 is a Class D β-lactamase and belongs to Group 2. OXA-10 was reported to connect to meropenem while meropenem-clavulanate was found effective against BlaC in another study [93–95]. These interactions were not listed in PDB; however with the help of our similarity model we were able to capture them.

We also observe that CTX-M-9a seems more similar to Penicillin-binding proteins than it is to TEM when ligand similarity is taken into account. From the top-pairs list we observe that CTX-M-9a shares ligands with PBP4 (P24228) and PBP6 (P08506). Considering CTX-M-9a as a bridge, we can observe a relationship between PBP6 and PBP4, which are actually reported in PDB to have one common ligand, AIC (Ampicillin). They also have a similarity score of 4.34 which means they bind to some other chemically similar ligands, namely PNV–AIC, PNM–AIC, PNM–AIX and AIX–AIC (PNV: Penicillin V, PNM: Penicillin G, AIX: Ampicillin, open form). Further research in DrugBank revealed that they both bind to Ertapenem (DB00303) [96] which strengthens the relationship between PBP4 and PBP6. PBP6 is reported to bind to AIX, whereas no interaction for PBP4–AIX is defined. Thus, we suggest that PBP4 might also bind to AIX, where AIX was already reported to bind other PBP4s (Q8Y547, A8E0K8) [97, 98]. Similarly, PBP4 is reported to bind PNV and PNM, whereas no interactions involving these ligands are reported for PBP6. Therefore, we also suggest that the PNV and PNM binding to PBP6 can be investigated.

Another interesting point is that SHV-1 is also no longer in the top list of the

similarity network, even though it is the second best binder with the number of 26 ligand interactions. Instead, NDM-1, which is a Class B β-lactamase and a global threat [99, 100], makes its way to the top list with BlaA (BlaC) pairing where the interaction is completely built on ligand similarities. Although NDM-1 has half of the ligand interactions SHV-1 has, our similarity model highlights its relationship with BlaC. This is a nice indicator of the importance of ligand similarity over considering only ligand sharing information as we did in the weighted identity network.

We provide the pair scoring tables for top 100 pairs of the weighted identity and similarity models in the Appendix.

4.2. Comparative study of SMILES-based ligand kernels for protein-ligand interaction prediction

In this section, results of the SVM classification which is presented in Section 3.2.4 for different ligand similarity kernels on the GPCR and ion channels data sets are presented. In the first protocol, for each target protein, the data points regarding the remaining proteins and a small portion of the current protein's data points are trained in the SVM classifier and the rest of the data points of the target protein are used for testing. As a target kernel, the hierarchy kernel is used. This original setup does not change except for the ligand similarity kernels.

All of the 16 different ligand kernels we present here utilize the SMILES string descriptors of the compounds. Among these kernels, the string similarity ones; edit distance, NLCS, combination of NLCS algorithms (CLCS) as well as the SMILES-specialized ones; LINGOsim, SMIfp and SMILES based substring kernels are tested. The original LINGOsim algorithm uses substring length $q = 4$ to create LINGOs from the SMILES string and then, applies Tanimoto similarity coefficient to measure the similarity between two LINGO sets. In this study, we perform LINGOsim on the compounds based on the substring (LINGO) lengths $q = 3, 4, 5$. We also propose a weighted LINGOsim model where we assign weights to each LINGO. SMIfp is another SMILES oriented compound similarity algorithm which converts SMILES strings into

34D scalar fingerprints where each dimension holds the frequency of a pre-determined symbol in that SMILES. SMIfp is designed for the virtual screening task and CBD is used to extract similar compounds. In this study not only we tested SMIfp with CBD, but we also apply two other metrics, Euclid Distance and Tanimoto coefficient. We also modify SMIfp in a way such that we can expand the dimensions of the fingerprint with the symbols frequent in our compound data set. Therefore, we propose a 38D SMIfp method which is also tested with Euclid distance, CBD, and Tanimoto similarity coefficient. SMILES based substring kernel is tested with both original and modified SMILES strings. Finally, we propose a novel similarity kernel which we call LINGO-based TF-IDF cosine similarity. It threats each SMILES string as a document and each LINGO in the SMILES as a term.

We set the original experiment results for the hierarchy kernel with 2D fingerprint ligand similarity as baseline. The classifier results are given as area under the ROC curve (AUC). A ROC curve depicts the performance of a classifier in a two-dimensional space where these dimensions represent True-Positive and False-Positive rates. Calculating the area under ROC reduces the dimension to a scalar value which represents the success of the classifier [101]. Table 4.6 depicts the AUC results for each protein family. We should report that the AUC results for the 2D fingerprint Tanimoto kernel are given as 0.926 ± 0.015 for GPCR and 0.925 ± 0.012 for ion channels [3]. The comparisons made here are based on our AUC results with 2D fingerprint Tanimoto kernel. Inconsistency of the data points which are provided online with the ones reported in [3] (explained in Section 3.2.1) might be the reason of the difference between the AUC scores.

For the GPCR data set, LINGO-based TF-IDF cosine similarity produced the best score by outperforming the original ligand kernel with a slight difference. SMILES-based substring with modified SMILES input (Modified Substring) and NCLS kernels produced the second and third best AUCs. We observed that the similarity distance CBD selected by the original SMIfp algorithm was not a good choice, but Tanimoto similarity coefficient improved the results in both of the SMIfp models with 34D and 38D. Unfortunately, increasing fingerprint dimension from 34D to 38D in SMIfp did

Table 4.6. AUC for the ligand kernels on GPCR and ion channels data sets.

Kernel Type	**GPCR**	**ion channels**
2D fingerprint Tanimoto	**0.810 ± 0.026**	**0.862 ± 0.015**
Edit distance	0.767 ± 0.028	0.713 ± 0.016
Substring	0.771 ± 0.027	0.827 ± 0.017
Modified substring	0.790 ± 0.027	0.826 ± 0.016
LINGOsim (q=3)	0.690 ± 0.027	0.712 ± 0.016
LINGOsim (q=4)	0.715 ± 0.027	0.732 ± 0.017
LINGOsim (q=5)	0.768 ± 0.022	0.748 ± 0.017
LINGOsim weighted	0.626 ± 0.026	0.599 ± 0.021
SMIfp 34D Euclid	0.714 ± 0.028	0.809 ± 0.017
SMIfp 34D CBD	0.715 ± 0.027	0.810 ± 0.017
SMIfp 34D Tanimoto	0.773 ± 0.029	0.820 ± 0.016
SMIfp 38D Euclid	0.707 ± 0.029	0.796 ± 0.018
SMIfp 38D CBD	0.696 ± 0.029	0.796 ± 0.018
SMIfp 38D Tanimoto	0.782 ± 0.028	0.821 ± 0.017
NLCS	0.784 ± 0.029	0.853 ± 0.014
CLCS	0.766 ± 0.028	0.852 ± 0.015
LINGO based TF-IDF	**0.819 ± 0.024**	0.845 ± 0.014

not enhance the results significantly as it was expected. However, when Tanimoto coefficient was applied, 38D SMIfp performed better than 34D SMIfp. When we compare LINGOsim kernels with different subsequence length q, selecting $q = 5$ significantly improved the AUC. Weighted LINGOsim model produced the worst performance among all ligand kernels. The reason behind this fail could be that assigning term frequencies to LINGOs was not appropriate.

For ion data set, none of the ligand kernels outperformed the original kernel; however NLCS, CLCS and LINGO-based TF-IDF kernel produced the relatively closest scores. We again observed that using Tanimoto similarity coefficient enhanced the results in both of the SMIfp models. When we compared LINGOSim kernels, we saw

that increasing the substring length also improved the AUC. The weighted LINGOsim, however, produced the worst result.

For both data sets, CLCS failed to achieve better performance than NLCS even though it was a more complex model. CLCS combines three different algorithms that utilize LCS structure and assigns weight to each algorithm. Increasing the weight of the LCS component might improve performance of CLCS.

Our results showed that SMILES string can be used to measure the similarity of the compounds without requiring any other descriptor. Furthermore, when applied an efficient similarity method, SMILES based kernels obtained promising results (LINGO-based TF-IDF cosine similarity).

5. CONCLUSION

5.1. Conclusions

With this work we have examined protein-ligand interactions through the lens of ligand similarity using both network models and machine learning. First, we introduced a method for clustering proteins using a ligand-centric network model where proteins were represented as nodes and the ligands they share were used to create edges to connect them. We proposed three network models; unweighted identity, weighted identity and similarity. The unweighted identity and weighted identity networks connected only proteins that share identical ligands, whereas in the similarity network, chemical similarity of the ligands was also considered. By constructing different network models, we had the opportunity to observe how the clusters are affected when the networks include ligand information on: (i) number of identical ligands, and (ii) chemical similarity of the ligands.

We apply our method on β-lactamase and Penicillin-Binding-Protein family of proteins. These β-lactam binding proteins were selected because there is enormous evolutionary pressure on them as novel β-lactam type drugs are developed to combat antimicrobial resistance. With the use of chemical similarity not only denser clusters were observed, but also some clusters were expanded. We have shown that new scientific hypotheses, which deserve further investigations, can be generated by analysing the top scoring pairs in the weighted identity and similarity networks. For example, chemical similarity use highlighted some relationships which might seem insignificant otherwise such as relationships of β-lactamase BlaA (BlaC) with NDM-1, ampC, Toho-1, CTX-M-9a, DD carboxipeptidase and OXA-10. Considering the knowledge of inhibition of BlaC by Avibactam (NXL104) inhibitor, we suggested the use of NXL104 inhibitor for these listed proteins. We also observed a relationship between PBP4 and PBP6, and suggested ampicillin, might bind to PBP4 and penicillin V and penicillin G might bind to PBP6 with high affinity.

In ligand based clustering functionally similar proteins tended to group together, where in most cases Group 2 proteins and PBPs were placed within the same cluster. Ligand based clustering was also consistent with sequential similarities. For example, BlaR1 was clustered together with Class D β-lactamases. BlaR1 protein is a high MW class C PBP. This class of proteins are reported to be sequentially similar to Class D β-lactamases [14].

Second, we presented a comparative study of ligand similarity kernels on the task of protein-ligand interaction prediction using SVM. We chose 2D fingerprint Tanimoto similarity kernel as base kernel and utilized 16 different ligand similarity kernels, LINGO with different parameter settings (substring length $q = 3, 4, 5$, weighted model), SMIfp with different distance metrics (Euclid, CBD, Tanimoto), SMILES based substring kernel and its modified version, LINGO-based TF-IDF cosine similarity, edit distance, Normalized Longest Common Subsequence, combination of LCS algorithms all of which are based on SMILES string. Among these kernels, SMIfp, LINGOsim and SMILES-based substring models are SMILES similarity targeting kernels while the rest of them are string kernels.

The original SMIfp algorithm is based on representing SMILES string on 34D vector where each dimension reflects the frequency of a pre-determined symbol in that SMILES. Then two vectors are compared using CBD. In our study, we tested 34D SMIfp model using Euclid and Tanimoto similarity coefficient along with CBD. We observed that when compared with CBD using Tanimoto coefficient significantly improves the AUC results, although it failed to outperform the base kernel. We then utilize LINGOsim kernel whose substring length q was set as 4 by definition. We held our experiments for $q = 3, 4, 5$ using LINGOsim. AUC results showed that selecting q as 5 will be wiser decision. Comparison of CLCS with NLCS showed that, even though CLCS combined three different LCS approaches it failed to improve the LCS model. The method we propose by combining the original LINGO representation ($q = 4$) with TF-IDF cosine similarity outperformed the base kernel on the GPCR data set, although it failed on the ion channels data set.

Our study proved that efficient selection of the ligand kernel directly affects the success of the classification algorithm in the task of target-ligand prediction. We also observed that SMILES string is adequate when used with the suitable similarity kernel, such as LINGO-based TF-IDF.

5.2. Future Studies

The proposed network models are applicable for all protein families interacting with small compounds. We presented β-lactamase and PBP families as a case study. The available data for these families in the PDB is limited, but even in this condition our method obtained promising results. When applied to a richer data set of interactions, we believe that more interesting results may be produced. We first aim to curate publicly available interaction data from databases such as ChEMBL and BindingDB. Then with the help of text mining and protein-ligand interaction methods we will focus expanding this data set.

The method we propose named LINGO-based TF-IDF cosine similarity produced promising results on the GPCR data set. In addition, setting substring length $q = 5$ improved the results of LINGOsim when compared with $q = 3, 4$. Therefore, modifying LINGOs as substring length q=5 in the TF-IDF model can give better results.

APPENDIX A: PAIR SCORE TABLES

Table A.1. Top 100 pairs of the weighted identity network according to the scores.

no	Protein ID	Ligand ID	Protein ID	Score
1	P00811	CB4,SM2,SUC,DMS,GF4,GF1	Q9L5C8	6
2	P00811	BZB,CAZ,CLS,CEO,PCZ	Q47066	5
3	P62593	IM2,CB4,105,SM2	P00811	4
4	P62593	MES,IM2,PNM	G1C794	3
5	Q47066	PNM,CEF,CEP	P15555	3
6	P62593	CB4,SM2,NBF	Q9L5C8	3
7	P30896	EPE,MPD,MA4	P0AD64	3
8	P13661	KCX,DRW,1S6	Q8RLA6	3
9	P0AD64	MA4,TBE,TSL	Q5VCA8	3
10	P08506	DAL,AMV,FGA	Q54113	3
11	P62593	MES,IM2	P39045	2
12	P39045	MES,IM2	G1C794	2
13	P07062	MES,ACA	P07061	2
14	P07062	MES,ACA	Q59710	2
15	P07061	MES,ACA	Q59710	2
16	P62593	FOS,PNM	P00807	2
17	P62593	IM2,PNM	P71586	2
18	P62593	IM2,EPE	P18357	2
19	P00811	IM2,BSF	P39045	2
20	P00811	IM2,PCZ	P71586	2
21	P00811	IM2,CAZ	P18357	2
22	P00811	IM2,CAZ	Q51504	2
23	Q51400	IM2,MER	Q93PQ0	2
24	Q51400	IM2,MER	P18357	2
25	Q93PQ0	IM2,MER	P18357	2
26	P71586	IM2,PNM	G1C794	2
27	P18357	IM2,CAZ	Q51504	2
28	Q51504	IM2,AZR	G1C794	2
29	P00807	PNM,CED	P00808	2
30	P00808	PNM,PCZ	Q47066	2

Table A.1. Top 100 pairs of the weighted identity network according to the scores (cont.).

no	Protein ID	Ligand ID	Protein ID	Score
31	P00808	PNM,PCZ	P71586	2
32	Q47066	PNM,PCZ	P71586	2
33	Q47066	PNM,AZR	G1C794	2
34	P62593	CB4,EPE	P0AD64	2
35	P00811	CB4,SUC	Q9L5C7	2
36	Q9L5C8	CB4,SUC	Q9L5C7	2
37	P0AD64	EPE,MER	P18357	2
38	P0AD64	EPE,TBE	Q93F76	2
39	Q47066	CAZ,AZR	Q51504	2
40	P00811	AXL,PCZ	P0C5C1	2
41	P00811	KCX,DMS	P14489	2
42	P00811	KCX,PEG	Q9L4P2	2
43	P14489	KCX,PG4	Q8RLA6	2
44	P00811	APB,0NG	Q8FGC8	2
45	Q59401	WY4,WY2	P0AD64	2
46	P00808	EOH,PGE	P94458	2
47	Q93PQ0	SFR,FPM	C7C422	2
48	P0C5C1	AIX,CB9	Q8Y547	2
49	P14677	BMG,TEB	Q04707	2
50	P25910	MES	P62593	1
51	P25910	MES	Q7ATJ4	1
52	P25910	MES	P52664	1
53	P25910	MES	P39045	1
54	P25910	MES	Q9RET4	1
55	P25910	MES	G1C794	1
56	P25910	MES	P07062	1
57	P25910	MES	P07061	1
58	P25910	MES	Q59710	1
59	P62593	MES	Q7ATJ4	1
60	P62593	MES	P52664	1
61	P62593	MES	Q9RET4	1
62	P62593	MES	P07062	1
63	P62593	MES	P07061	1
64	P62593	MES	Q59710	1
65	Q7ATJ4	MES	P52664	1
66	Q7ATJ4	MES	P39045	1
67	Q7ATJ4	MES	Q9RET4	1
68	Q7ATJ4	MES	G1C794	1
69	Q7ATJ4	MES	P07062	1
70	Q7ATJ4	MES	P07061	1

Table A.1. Top 100 pairs of the weighted identity network according to the scores (cont.).

no	Protein ID	iigand ID	Protein ID	Score
71	Q7ATJ4	MES	Q59710	1
72	P52664	MES	P39045	1
73	P52664	MES	Q9RET4	1
74	P52664	MES	G1C794	1
75	P52664	MES	P07062	1
76	P52664	MES	P07061	1
77	P52664	MES	Q59710	1
78	P39045	MES	Q9RET4	1
79	P39045	MES	P07062	1
80	P39045	MES	P07061	1
81	P39045	MES	Q59710	1
82	Q9RET4	MES	G1C794	1
83	Q9RET4	MES	P07062	1
84	Q9RET4	MES	P07061	1
85	Q9RET4	MES	Q59710	1
86	G1C794	MES	P07062	1
87	G1C794	MES	P07061	1
88	G1C794	MES	Q59710	1
89	P62593	IM2	Q51400	1
90	P62593	IM2	P0AEB2	1
91	P62593	IM2	Q93PQ0	1
92	P62593	IM2	Q51504	1
93	P62593	IM2	Q9KJY7	1
94	P62593	IM2	Q09HD0	1
95	P00811	IM2	Q51400	1
96	P00811	IM2	P0AEB2	1
97	P00811	IM2	Q93PQ0	1
98	P00811	IM2	G1C794	1
99	P00811	IM2	Q9KJY7	1
100	P00811	IM2	Q09HD0	1

Table A.2. Top 100 pairs of the similarity network according to the scores.

no	Protein ID	Ligand ID	Protein ID	Score
1	P00811	CB4-CB4,CLS-CFX,CLS-CE3,SUC-SUC,AXL-PNN, AXL-PNK, AXL-YPP, SM2-SM2,SM3-SM2, DMS-DMS,SM4-SM2, GF4-GF4,GF1-GF1	Q9L5C8	11,42
2	P00811	BZB-BZB,CAZ-PCZ,CAZ-CAZ,CLS-CLS,KCP-CEP,KCP-CLS, CEO-CEO,AXL-PNM,PCZ-PCZ,PCZ-CAZ	Q47066	9,32
3	P0C5C1	AXL-ZZ7,AXL-PNK,AIX-ZZ7,AIX-PNK,CB9-ZZ7, CB9-PNK,NFF-0RM,7EP-0RM,SFR-FPM,SFR-SFR	C7C422	8,94
4	P00811	CAZ-PCZ,CLS-9EP,KCP-9EP,CEO-CD6,AXL-AXL, AXL-AIX,AXL-CB9,AXL-7EP,MXG-CD6,PCZ-PCZ	P0C5C1	8,65
5	Q47066	CEP-9EP,PNM-XD1,PNM-AXL,PNM-AIX, PNM-CB9,PNM-7EP,CEO-CD6,CLS-9EP,PCZ-PCZ,CAZ-PCZ	P0C5C1	8,55
6	Q9L5C8	PNN-AXL,PNN-AIX,PNN-CB9,PNK-AXL,PNK-AIX, PNK-CB9,WPP-AIX,YPP-AXL,YPP-AIX,YPP-CB9	P0C5C1	7,96
7	P0C5C1	XD1-PNM,XD1-HE0,AXL-PNM,AXL-HE0,AIX-PNM, AIX-HE0,CB9-PNM,CB9-HE0,9EP-CEP,7EP-PNM	P15555	7,86
8	P14489	HOQ-XD1,PNM-XD1,PNM-AXL,PNM-AIX,PNM-CB9, PNM-7EP,ZZ7-AXL,ZZ7-AIX,ZZ7-CB9	P0C5C1	7,57
9	Q9L5C8	SUC-SUC,PNN-AIX,PNN-AIC,PNK-AIX,PNK-AIC, WPP-AIX,WPP-AIC,YPP-AIX,YPP-AIC	P08506	7,57
10	Q9L5C8	PNN-AIC,PNN-PNM,PNN-PNV,PNK-AIC, PNK-PNM,WPP-AIC,YPP-AIC,YPP-PNM,CE3-FXM	P24228	7,41
11	P0C5C1	XD1-PNM,AXL-AIC,AXL-PNM,AIX-AIC, AIX-PNM,CB9-AIC,CB9-PNM,7EP-PNM,SFR-FPM	P24228	7,27
12	P62593	IM2-IM2,PNM-AXL,CB4-CB4,105-105, SM2-SM2,SM2-SM3,SM2-SM4,CXB-CB4	P00811	7,26
13	P14489	PNM-PNN,PNM-PNK, PNM-YPP, ZZ7-PNN, ZZ7-PNK,ZZ7-WPP,ZZ7-YPP,DMS-DMS	Q9L5C8	6,79
14	P0AD64	LN1-MXC,LN1-MXF,17K-MXS,17K-MXC, 17K-MXF,MXF-MXS,MXF-MXC,MXF-MXF	Q8RLA6	6,58
15	Q9L5C8	PNN-ZZ7,PNN-PNK,PNK-ZZ7,PNK-0RM, PNK-PNK,WPP-ZZ7,YPP-ZZ7,YPP-PNK	C7C422	6,58
16	P0C5C1	AXL-CMV,AXL-FMZ,AXL-AIX,AIX-CMV, AIX-FMZ,AIX-AIX,CB9-AIX,7EP-AIX	P45161	6,51
17	P62593	PNM-PNN,PNM-PNK,PNM-YPP, CB4-CB4,SM2-SM2,NBF-NBF,CXB-CB4	Q9L5C8	6,36
18	P14489	PNM-ZZ7,PNM-PNK,ZZ7-ZZ7,ZZ7-0RM, ZZ7-PNK,PG4-P6G,PG4-PEG	C7C422	6,27
19	P0C5C1	AXL-AIX,AXL-CB9,AIX-AIX,AIX-CB9, CB9-AIX,CB9-CB9,7EP-AIX	Q8Y547	6,18
20	Q9L5C8	PNN-AIX,PNK-CMV,PNK-FMZ,PNK-AIX, WPP-AIX,YPP-CMV,YPP-FMZ,YPP-AIX	P45161	6,09
21	P00811	CAZ-PCZ,CLS-CFX,CLS-CED, KCP-CED,AXL-PNM,PEG-PGE,PCZ-PCZ	P00808	6,06
22	P00808	PNM-XD1,PNM-AXL,PNM-AIX, PNM-CB9,PNM-7EP,PCZ-PCZ,CED-9EP	P0C5C1	6,01
23	P0C5C1	AXL-AIX,AXL-AIC,AIX-AIX, AIX-AIC,CB9-AIX,CB9-AIC,7EP-AIX	P08506	5,84
24	Q9L5C8	PNN-HEL,PNN-PNM,PNK-PNM,PNK-HE0, WPP-HEL,YPP-PNM,CE3-CEF	P15555	5,69
25	P39045	EWB-E08,EWB-E07,EWA-A01, BH6-ZA2,ZA3-ZA3,B07-A01	Q7CRA4	5,64
26	P0C5C1	AXL-PG1,AXL-7EP,AIX-PG1,AIX-7EP, CB9-PG1,NFF-7EP,7EP-7EP	Q93IC2	5,61
27	Q9L5C8	PNN-AIX,PNN-CB9,PNK-AIX, PNK-CB9,WPP-AIX,YPP-AIX,YPP-CB9	Q8Y547	5,60

Table A.2. Top 100 pairs of the similarity network according to the scores (cont.).

no	Protein ID	Ligand ID	Protein ID	Score
28	P00811	MOX-MX1,AXL-PNM,AXL-ZZ7,KCX-KCX,DMS-DMS,PEG-PG4	P14489	5,50
29	P00808	CFX-CLS,PNM-PNM,PCZ-PCZ,PCZ-CAZ,CED-CEP,CED-CLS	Q47066	5,34
30	Q47066	CEF-CTJ,PNM-CB9,PCZ-CAZ,AZR-AZR,AZR-PFV,CAZ-CAZ	Q51504	5,30
31	C7C422	ZZ7-AIC,ZZ7-PNM,PNK-AIC,PNK-PNM,FPM-FPM,SFR-FPM	P24228	5,23
32	P00811	CEO-NCF,CEO-REC,IM2-IM2,MXG-REC,BSF-BSF,BSG-BSF	P39045	5,22
33	P62593	IM2-MER,CB4-CB4,CB4-CZ6,CXB-CB4,CXB-CZ6,EPE-EPE	P0AD64	5,18
34	P0C5C1	XD1-PNM,AXL-PNM,AIX-PNM,CB9-PNM,7EP-PNM,PCZ-PCZ	P71586	5,15
35	P00807	CED-9EP,PNM-XD1,PNM-AXL,PNM-AIX,PNM-CB9,PNM-7EP	P0C5C1	5,01
36	Q47066	CEF-CE3,PNM-PNN,PNM-PNK,PNM-YPP,CLS-CFX,CLS-CE3	Q9L5C8	4,94
37	P62593	PNM-XD1,PNM-AXL,PNM-AIX,PNM-CB9,PNM-7EP,ALP-XD1	P0C5C1	4,93
38	P14489	PNM-CMV,PNM-FMZ,PNM-AIX,ZZ7-CMV,ZZ7-FMZ,ZZ7-AIX	P45161	4,88
39	C7C422	ZZ7-CMV,ZZ7-FMZ,ZZ7-AIX,PNK-CMV,PNK-FMZ,PNK-AIX	P45161	4,82
40	P0AD64	MA4-MA4,TBI-TBE,TBE-TBE,TSL-TSL,TAU-ESA	Q5VCA8	4,73
41	Q47066	CEF-CEF,CEP-CEP,PNM-PNM,PNM-HE0,CLS-CSC	P15555	4,59
42	P00811	CAZ-CAZ,IM2-IM2,IM2-MER,AXL-PG1,PCZ-CAZ	P18357	4,42
43	P00811	CXU-0WO,AXL-ZZ7,AXL-PNK,PEG-P6G,PEG-PEG	C7C422	4,38
44	P24228	AIC-AIX,AIC-AIC,PNM-AIX,PNM-AIC,PNV-AIC	P08506	4,33
45	P00811	CXU-CXV,CLS-CFX,KCP-HJ2,IM2-IM2,AXL-HJ3	P0AEB2	4,30
46	P15555	DAL-DAL,HEL-AIC,PNM-AIX,PNM-AIC,HE0-AIX	P08506	4,27
47	P15555	HEL-AIC,HEL-PNV,PNM-AIC,PNM-PNM,HE0-PNM	P24228	4,16
48	P0C5C1	XD1-PNM,AXL-PNM,AIX-PNM,CB9-PNM,7EP-PNM	Q8DH45	4,15
49	P0C5C1	XD1-PNM,AXL-PNM,AIX-PNM,CB9-PNM,7EP-PNM	G1C794	4,15
50	P45161	CMV-AIX,CMV-AIC,FMZ-AIX,AIX-AIX,AIX-AIC	P08506	4,09
51	P0C5C1	1RG-MER,AXL-PG1,AIX-PG1,CB9-PG1,PCZ-CAZ	P18357	4,06
52	P30896	EPE-EPE,MA4-MA4,MPD-MPD,MPD-MRD	P0AD64	4,00
53	P24228	AIC-CMV,AIC-AIX,PNM-CMV,PNM-FMZ,PNM-AIX	P45161	3,95
54	P00811	CAZ-PCZ,IM2-IM2,AXL-PNM,PCZ-PCZ	P71586	3,81
55	P00811	CAZ-CAZ,IM2-IM2,AXL-CB9,PCZ-CAZ	Q51504	3,75
56	P08506	AMV-MUR,AMV-AMV,DAL-DAL,FGA-FGA	Q54113	3,72
57	P14489	PNM-AIX,PNM-CB9,ZZ7-AIX,ZZ7-CB9	Q8Y547	3,66
58	C7C422	ZZ7-AIX,ZZ7-CB9,PNK-AIX,PNK-CB9	Q8Y547	3,63
59	P00808	PGE-P6G,PGE-PEG,PNM-ZZ7,PNM-PNK	C7C422	3,61
60	P0C5C1	AXL-CB9,AIX-CB9,CB9-CB9,PCZ-CAZ	Q51504	3,61
61	Q51504	CB9-PNM,IM2-IM2,AZR-AZR,PFV-AZR	G1C794	3,60
62	Q93PQ0	SFR-FPM,SFR-SFR,FPM-FPM,FPM-SFR	C7C422	3,57
63	P62593	IM2-IM2,IM2-MER,PNM-PG1,EPE-EPE	P18357	3,56
64	P18357	PG1-CB9,CAZ-CAZ,IM2-IM2,MER-IM2	Q51504	3,53
65	P00808	CFX-CFX,PNM-PNN,PNM-PNK,PNM-YPP	Q9L5C8	3,52

Table A.2. Top 100 pairs of the similarity network according to the scores (cont.).

no	Protein ID	Ligand ID	Protein ID	Score
66	P14489	PNM-AIC,PNM-PNM,ZZ7-AIC,ZZ7-PNM	P24228	3,51
67	P14489	PNM-AIX,PNM-AIC,ZZ7-AIX,ZZ7-AIC	P08506	3,50
68	Q51400	IM2-IM2,IM2-MER,MER-IM2,MER-MER	Q93PQ0	3,49
69	Q51400	IM2-IM2,IM2-MER,MER-IM2,MER-MER	P18357	3,49
70	Q93PQ0	IM2-IM2,IM2-MER,MER-IM2,MER-MER	P18357	3,49
71	P71586	IM2-IM2,IM2-MER,PNM-PG1,PCZ-CAZ	P18357	3,49
72	C7C422	ZZ7-AIX,ZZ7-AIC,PNK-AIX,PNK-AIC	P08506	3,45
73	P14489	PNM-PNM,PNM-HE0,ZZ7-PNM,ZZ7-HE0	P15555	3,42
74	P08506	AIX-AIX,AIX-CB9,AIC-AIX,AIC-CB9	Q8Y547	3,42
75	P00811	CB4-CB4,CB4-CZ6,SUC-MA4,IM2-MER	P0AD64	3,40
76	P24228	AIC-AIX,AIC-CB9,PNM-AIX,PNM-CB9	Q8Y547	3,40
77	P45161	CMV-AIX,FMZ-AIX,AIX-AIX,AIX-CB9	Q8Y547	3,40
78	C7C422	ZZ7-PNM,ZZ7-HE0,PNK-PNM,PNK-HE0	P15555	3,36
79	P15555	PNM-AIX,PNM-CB9,HE0-AIX,HE0-CB9	Q8Y547	3,29
80	Q9L5C8	PNN-ZZ7,PNK-ZZ7,WPP-ZZ7,YPP-ZZ7	Q5HI26	3,28
81	P00811	CLS-CSC,KCP-CEP,AXL-PNM,AXL-HE0	P15555	3,23
82	P15555	PNM-CMV,PNM-FMZ,PNM-AIX,HE0-AIX	P45161	3,15
83	Q9L5C8	PNN-CB9,PNK-CB9,YPP-CB9,CE3-CTJ	Q51504	3,11
84	P62593	IM2-IM2,PNM-PNM,MES-MES	G1C794	3,00
85	P13661	KCX-KCX,DRW-DRW,1S6-1S6	Q8RLA6	3,00
86	P0AD64	TBI-TBE,TBE-TBE,EPE-EPE	Q93F76	3,00
87	P0AEB2	HJ3-XD1,HJ3-AXL,HJ3-AIX,HJ3-CB9	P0C5C1	2,96
88	Q47066	PNM-PNM,PCZ-PCZ,CAZ-PCZ	P71586	2,93
89	P14489	PNM-PNM,ZZ7-PNM,PG4-PGE	P00808	2,90
90	P71586	IM2-IM2,PNM-CB9,PCZ-CAZ	Q51504	2,83
91	P62593	PNM-PNM,PNM-ZZ7,ALP-HOQ	P14489	2,79
92	Q47066	PNM-PG1,PCZ-CAZ,CAZ-CAZ	P18357	2,75
93	P52700	MCO-X8Z,PEG-P6G,PEG-PEG	C7C422	2,75
94	C7C422	P6G-PG4,PEG-PG4,FMT-FMT	Q9Y7D1	2,75
95	P62593	IM2-IM2,NBF-B07,MES-MES	P39045	2,74
96	P00811	IM2-MER,KCX-KCX,PEG-PEG	Q9L4P2	2,74
97	P0AD64	EPE-EPE,MER-IM2,MER-MER	P18357	2,74
98	Q8RLA6	PG4-P6G,PG4-PEG,1S6-0WO	C7C422	2,72
99	P00811	SUC-SUC,AXL-AIX,AXL-AIC	P08506	2,71
100	P0C5C1	AXL-ZZ7,AIX-ZZ7,CB9-ZZ7	Q5HI26	2,71

REFERENCES

1. Livermore, D. M. and N. Woodford, "The beta-lactamase Threat in Enterobacteriaceae, Pseudomonas and Acinetobacter", *Trends in Microbiology*, Vol. 14, No. 9, pp. 413–20, 2006.

2. Islam, A. and D. Inkpen, "Semantic Text Similarity Using Corpus-based Word Similarity and String Similarity", *ACM Transactions on Knowledge Discovery from Data*, Vol. 2, No. 2, pp. 1–25, 2008.

3. Jacob, L. and J.-P. Vert, "Protein-ligand Interaction Prediction: An Improved Chemogenomics Approach", *Bioinformatics*, Vol. 24, No. 19, pp. 2149–2156, 2008.

4. Wagner, R. A. and M. J. Fischer, "The String-to-String Correction Problem", *Journal of the ACM*, Vol. 21, No. 1, pp. 168–173, 1974.

5. Vidal, D., M. Thormann and M. Pons, "LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities", *Journal of Chemical Information and Modeling*, Vol. 45, No. 2, pp. 386–393, 2005.

6. Schwartz, J., M. Awale and J.-L. Reymond, "SMIfp (SMILES fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules", *Journal of Chemical Information and Modeling*, Vol. 53, No. 8, pp. 1979–1989, 2013.

7. Cao, D. S., J. C. Zhao, Y. N. Yang, C. X. Zhao, J. Yan, S. Liu, Q. N. Hu, Q. S. Xu and Y. Z. Liang, "In Silico Toxicity Prediction by Support Vector Machine and SMILES Representation-based String Kernel", *SAR and QSAR in Environmental Research*, Vol. 23, No. 1-2, pp. 141–53, 2012.

8. Kong, K. F., L. Schneper and K. Mathee, "Beta-lactam Antibiotics: From An-

tibiosis to Resistance and Bacteriology", *APMIS*, Vol. 118, No. 1, pp. 1–36, 2010.

9. Poole, K., "Resistance to Beta-lactam Antibiotics", *Cellular and Molecular Life Sciences CMLS*, Vol. 61, No. 17, pp. 2200–2223, 2004.

10. Davies, J. and D. Davies, "Origins and Evolution of Antibiotic Resistance", *Microbiology and Molecular Biology Reviews*, Vol. 74, No. 3, pp. 417–33, 2010.

11. Abraham, E. P. and E. Chain, "An Enzyme from Bacteria Able to Destroy Penicillin. 1940", *Reviews of Infectious Diseases*, Vol. 10, No. 4, pp. 677–8, 1988.

12. Orencia, M. C., J. S. Yoon, J. E. Ness, W. P. C. Stemmer and R. C. Stevens, "Predicting the Emergence of Antibiotic Resistance by Directed Evolution and Structural Analysis", *Nature Structural and Molecular Biology*, Vol. 8, No. 3, pp. 238–242, 2001.

13. Brown, N. G., J. M. Pennington, W. Huang, T. Ayvaz and T. Palzkill, "Multiple Global Suppressors of Protein Stability Defects Facilitate the Evolution of Extended-Spectrum TEM Beta-Lactamases", *Journal of Molecular Biology*, Vol. 404, No. 5, pp. 832–846, 2010.

14. Massova, I. and S. Mobashery, "Kinship and Diversification of Bacterial Penicillin-Binding Proteins and Beta-lactamases", *Antimicrobial Agents and Chemotherapy*, Vol. 42, No. 1, pp. 1–17, 1998.

15. Fisher, J. F., S. O. Meroueh and S. Mobashery, "Bacterial Resistance to Beta-lactam Antibiotics: Compelling Opportunism, Compelling Opportunity", *Chemical Reviews*, Vol. 105, No. 2, pp. 395–424, 2005.

16. Majiduddin, F. K., I. C. Materon and T. G. Palzkill, "Molecular Analysis of Beta-lactamase Structure and Function", *International Journal of Medical Microbiology*, Vol. 292, No. 2, pp. 127 – 137, 2002.

17. Bradford, P. A., "Extended-Spectrum Beta-lactamases in the 21st Century: Characterization, Epidemiology, and Detection of This Important Resistance Threat", *Clinical Microbioogyl Reviews*, Vol. 14, No. 4, pp. 933–51, 2001.

18. Knothe, H., P. Shah, V. Krcmery, M. Antal and S. Mitsuhashi, "Transferable Resistance to Cefotaxime, Cefoxitin, Cefamandole and Cefuroxime in Clinical Isolates of Klebsiella Pneumoniae and Serratia Marcescens", *Infection*, Vol. 11, No. 6, pp. 315–7, 1983.

19. Paterson, D. L. and R. A. Bonomo, "Extended-spectrum Beta-lactamases: A Clinical Update", *Clinical Microbioogyl Reviews*, Vol. 18, No. 4, pp. 657–86, 2005.

20. Bush, K., "Characterization of Beta-lactamases.", *Antimicrobial Agents and Chemotherapy*, Vol. 33, No. 3, pp. 259–263, 1989.

21. Ambler, R. P., "The Structure of Beta-lactamases", *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, Vol. 289, No. 1036, pp. 321–31, 1980.

22. Bush, K., G. A. Jacoby and A. A. Medeiros, "A Functional Classification Scheme for Beta-lactamases and Its Correlation with Molecular Structure", *Antimicrobial Agents and Chemotherapy*, Vol. 39, No. 6, pp. 1211–33, 1995.

23. Bush, K. and G. A. Jacoby, "Updated Functional Classification of Beta-lactamases", *Antimicrobial Agents and Chemotherapy*, Vol. 54, No. 3, pp. 969–76, 2010.

24. Sauvage, E., F. Kerff, M. Terrak, J. A. Ayala and P. Charlier, "The Penicillin-Binding Proteins: Structure and Role in Peptidoglycan Biosynthesis", *FEMS Microbiology Reviews*, Vol. 32, No. 2, pp. 234–258, 2008.

25. Georgopapadakou, N. H., "Penicillin-binding Proteins and Bacterial Resistance to Beta-lactams", *Antimicrobial Agents and Chemotherapy*, Vol. 37, No. 10, pp.

2045–53, 1993.

26. Kelly, J., O. Dideberg, P. Charlier, J. Wery, M. Libert, P. Moews, Knox, C. Duez, C. Fraipont, B. Joris and a. et, "On the Origin of Bacterial Resistance to Penicillin: Comparison of a Beta-lactamase and a Penicillin Target", *Science*, Vol. 231, No. 4744, pp. 1429–1431, 1986.

27. Ghuysen, J. M., "Serine Beta-lactamases and Penicillin-Binding Proteins", *Annual Review of Microbiology*, Vol. 45, pp. 37–67, 1991.

28. Ghuysen, J. M., "Penicillin-Binding Proteins. Wall Peptidoglycan Assembly and Resistance to Penicillin: Facts, Doubts and Hopes", *International Journal of Antimicrobial Agents*, Vol. 8, No. 1, pp. 45–60, 1997.

29. Leslie, C., E. Eskin and W. S. Noble, "The Spectrum Kernel: A String Kernel for SVM Protein Classification", *Pac Symp Biocomput*, pp. 564–75, 2002.

30. Lanckriet, G. R., M. Deng, N. Cristianini, M. I. Jordan and W. S. Noble, "Kernel-based Data Fusion and Its Application to Protein Function Prediction in Yeast", *Proceedings of the Pacific Symposium on Biocomputing 2004*, pp. 300–11, 2004.

31. Ben-Hur, A. and D. Brutlag, "Remote Homology Detection: A Motif Based Approach", *Bioinformatics*, Vol. 19, No. suppl 1, pp. 26–33, 2003.

32. Kin, T., T. Kato and K. Tsuda, *Kernel Methods in Computational Biology*, MIT Press, Cambridge, MA; USA, 2004.

33. Kuksa, P. P., "2D Similarity Kernels for Biological Sequence Classification", *Proceedings of the 11th International Workshop on Data Mining in Bioinformatics*, BIOKDD '12, pp. 15–20, ACM, Beijing, China, 2012.

34. Needleman, S. B. and C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *Journal of*

Molecular Biology, Vol. 48, No. 3, pp. 443 – 453, 1970.

35. Smith, T. and M. Waterman, "Identification of Common Molecular Subsequences", *Journal of Molecular Biology*, Vol. 147, No. 1, pp. 195 – 197, 1981.

36. Rätsch, G. and S. Sonnenburg, "13 Accurate Splice Site Detection for Caenorhabditis Elegans", *Kernel Methods in Computational Biology*, p. 277, 2004.

37. Ratsch, G., S. Sonnenburg and B. Scholkopf, "RASE: Recognition of Alternatively Spliced Exons in C.elegans", *Bioinformatics*, Vol. 21, No. suppl 1, pp. 369–377, 2005.

38. Meinicke, P., M. Tech, B. Morgenstern and R. Merkl, "Oligo Kernels for Datamining on Biological Sequences: A Case Study on Prokaryotic Translation Initiation Sites.", *BMC Bioinformatics*, Vol. 5, p. 169, 2004.

39. Liu, T., Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, "BindingDB: A Web-accessible Database of Experimentally Determined Protein–Ligand Binding Affinities", *Nucleic Acids Research*, Vol. 35, No. suppl 1, pp. 198–201, 2007.

40. Rognan, D., "Chemogenomic Approaches to Rational Drug Design", *British Journal of Pharmacology*, Vol. 152, No. 1, pp. 38–52, 2007.

41. Weininger, D., "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules", *Journal of Chemical Information and Computer Sciences*, Vol. 28, No. 1, pp. 31–36, 1988.

42. Weininger, D., A. Weininger and J. L. Weininger, "SMILES. 2. Algorithm for Generation of Unique SMILES Notation", *Journal of Chemical Information and Computer Sciences*, Vol. 29, No. 2, pp. 97–101, 1989.

43. Cao, D.-S., J.-C. Zhao, Y.-N. Yang, C.-X. Zhao, J. Yan, S. Liu, Q.-N. Hu, Q.-S. Xu and Y.-Z. Liang, "In Silico Toxicity Prediction by Support Vector Machine and

SMILES Representation-based String Kernel", *SAR and QSAR in Environmental Research*, Vol. 23, No. 1-2, pp. 141–153, 2012.

44. Taitz, Y., "Daylight Chemical Information Systems, Inc.", `http://www.daylight.com/`, 2014, accessed at May 2014.

45. Steinbeck, C., Y. Han, S. Kuhn, O. Horlacher, E. Luttmann and E. Willighagen, "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics", *Journal of Chemical Information and Computer Science*, Vol. 43, No. 2, pp. 493–500, 2003.

46. ChemAxon, "Chemaxon JChem .NET API 6.0.2.215", `http://www.chemaxon.com/products/jchem-base/`, 2013, accessed at October 2013.

47. Deza, M. and E. Deza, "Metrics on Normed Structures", *Encyclopedia of Distances*, pp. 89–99, Springer Berlin Heidelberg, Hershey, USA, 2013.

48. Krause, E. F., *Taxicab Geometry: : An Adventure in Non-Euclidean Geometry*, Courier Dover Publications, 1987.

49. Rogers, D. J. and T. T. Tanimoto, "A Computer Program for Classifying Plants", *Science*, Vol. 132, No. 3434, pp. 1115–8, 1960.

50. Levenshtein, V. I., "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710, 1966.

51. Hirschberg, D. S., "A Linear Space Algorithm for Computing Maximal Common Subsequences", *Communications of ACM*, Vol. 18, No. 6, pp. 341–343, 1975.

52. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 309–317, 1957.

53. Jones, K. S., "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, pp. 11–21, 1972.

54. Bilenko, M. and R. J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures", *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 39–48, ACM, Washington, DC, U.S.A, 2003.

55. Swamidass, S. J., J. Chen, J. Bruand, P. Phung, L. Ralaivola and P. Baldi, "Kernels for Small Molecules and the Prediction of Mutagenicity, Toxicity and Anti-cancer Activity", *Bioinformatics*, Vol. 21, No. suppl 1, pp. 359–368, 2005.

56. Yildirim, M. A., K. I. Goh, M. E. Cusick, A. L. Barabasi and M. Vidal, "Drug-Target Network", *Nature Biotechnology*, Vol. 25, No. 10, pp. 1119–26, 2007.

57. Mitchell, J. B., "The Relationship Between the Sequence Identities of Alpha Helical Proteins in the PDB and the Molecular Similarities of Their Ligands", *Journal of Chemical Information and Computer Science*, Vol. 41, No. 6, pp. 1617–22, 2001.

58. Nobeli, I., R. V. Spriggs, R. A. George and J. M. Thornton, "A Ligand-centric Analysis of the Diversity and Evolution of Protein-Ligand Relationships in E.coli", *Journal of Molecular Biology*, Vol. 347, No. 2, pp. 415–36, 2005.

59. Keiser, M. J., B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, "Relating Protein Pharmacology by Ligand Chemistry", *Nature Biotechnology*, Vol. 25, No. 2, pp. 197–206, 2007.

60. Kalinina, O. V., O. Wichmann, G. Apic and R. B. Russell, "ProtChemSI: A Network of Protein-Chemical Structural Interactions", *Nucleic Acids Research*, Vol. 40, No. Database issue, pp. 549–53, 2012.

61. Lin, H., M. F. Sassano, B. L. Roth and B. K. Shoichet, "A Pharmacological Organization of G Protein-Coupled Receptors", *Nature Methods*, Vol. 10, No. 2,

pp. 140–6, 2013.

62. Cheng, F., C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang and Y. Tang, "Prediction of Drug-Target Interactions and Drug Repositioning via Network-based Inference", *PLoS Computational Biology*, Vol. 8, No. 5, p. 1002503, 2012.

63. Cheng, F., Y. Zhou, W. Li, G. Liu and Y. Tang, "Prediction of Chemical-Protein Interactions Network with Weighted Network-based Inference Method", *PLoS One*, Vol. 7, No. 7, p. 41064, 2012.

64. Gohlke, H., M. Hendlich and G. Klebe, "Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions", *Journal of Molecular Biology*, Vol. 295, No. 2, pp. 337 – 356, 2000.

65. Zhu, S., Y. Okuno, G. Tsujimoto and H. Mamitsuka, "A Probabilistic Model for Mining Implicit 'Chemical Compound–Gene' Relations From Literature", *Bioinformatics*, Vol. 21, No. suppl 2, pp. 245–251, 2005.

66. Kubinyi, H., "Chemogenomics in Drug Discovery", S. Jaroch and H. Weinmann (Editors), *Chemical Genomics*, Vol. 58 of *Ernst Schering Research Foundation Workshop*, pp. 1–19, Springer Berlin Heidelberg, 2006.

67. Klabunde, T., "Chemogenomic Approaches to Drug Discovery: Similar Receptors Bind Similar Ligands", *British Journal of Pharmacology*, Vol. 152, No. 1, pp. 5–7, 2007.

68. Campillos, M., M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, "Drug Target Identification Using Side-Effect Similarity", *Science*, Vol. 321, No. 5886, pp. 263–266, 2008.

69. Yamanishi, Y., M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, "Prediction of Drug–Target Interaction Networks from the Integration of Chemical and

Genomic Spaces", *Bioinformatics*, Vol. 24, No. 13, pp. 232–240, 2008.

70. He, Z., J. Zhang, X.-H. Shi, L.-L. Hu, X. Kong, Y.-D. Cai and K.-C. Chou, "Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features", *PLoS ONE*, Vol. 5, No. 3, p. 9603, 2010.

71. Fernandez, M., A. Sarai and S. Ahmad, "Recognition of Drug-target Interaction Patterns Using Genetic Algorithm-optimized Bayesian-regularized Neural Networks and Support Vector Machines", *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, SMC'09, pp. 498–503, IEEE Press, San Antonio, TX, USA, 2009.

72. Bleakley, K. and Y. Yamanishi, "Supervised Prediction of Drug-Target Interactions Using Bipartite Local Models", *Bioinformatics*, Vol. 25, No. 18, pp. 2397–2403, 2009.

73. van Laarhoven, T., S. B. Nabuurs and E. Marchiori, "Gaussian Interaction Profile Kernels for Predicting Drug–Target Interaction", *Bioinformatics*, 2011.

74. Geppert, H., J. Humrich, D. Stumpfe, T. Gartner and J. Bajorath, "Ligand Prediction from Protein Sequence and Small Molecule Information using Support Vector Machines and Fingerprint Descriptors", *Journal of Chemical Information and Modeling*, Vol. 49, No. 4, pp. 767–79, 2009.

75. Alpaydin, E., *Introduction to Machine Learning*, The MIT Press, Cambridge, UK, 2nd edn., 2010.

76. Papadopoulos, J. S. and R. Agarwala, "COBALT: Constraint-Based Alignment Tool for Multiple Protein Sequences", *Bioinformatics*, Vol. 23, No. 9, pp. 1073–9, 2007.

77. Backman, T. W., Y. Cao and T. Girke, "ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules", *Nucleic Acids Research*, Vol. 39, No.

Web Server issue, pp. 486–91, 2011.

78. Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", *Genome Research*, Vol. 13, No. 11, pp. 2498–504, 2003.

79. Matter, H., "Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-dimensional and Three-dimensional Molecular Descriptors", *Journal of Medicinal Chemistry*, Vol. 40, No. 8, pp. 1219–29, 1997.

80. Xue, L., F. L. Stahura, J. W. Godden and J. Bajorath, "Mini-fingerprints Detect Similar Activity of Receptor Ligands Previously Recognized Only by Three-dimensional Pharmacophore-based Methods", *Journal of Chemical Information and Computer Science*, Vol. 41, No. 2, pp. 394–401, 2001.

81. Lin, C. Y., C. H. Chin, H. H. Wu, S. H. Chen, C. W. Ho and M. T. Ko, "Hubba: Hub Objects Analyzer–A Framework of Interactome Hubs Identification for Network Biology", *Nucleic Acids Research*, Vol. 36, No. Web Server issue, pp. 438–43, 2008.

82. Bader, G. D. and C. W. Hogue, "An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks", *BMC Bioinformatics*, Vol. 4, p. 2, 2003.

83. Letunic, I. and P. Bork, "Interactive Tree Of Life (iTOL): An Online Tool For Phylogenetic Tree Display and Annotation", *Bioinformatics*, Vol. 23, No. 1, pp. 127–8, 2007.

84. Borbulevych, O., M. Kumarasiri, B. Wilson, L. I. Llarrull, M. Lee, D. Hesek, Q. Shi, J. Peng, B. M. Baker and S. Mobashery, "Lysine Nzeta-decarboxylation Switch and Activation of the Beta-lactam Sensor Domain of BlaR1 Protein of Methicillin-Resistant Staphylococcus Aureus", *The Journal of Biological Chem-*

istry, Vol. 286, No. 36, pp. 31466–31472, 2011.

85. Kawashima, Y., T. Ohki, N. Shibata, Y. Higuchi, Y. Wakitani, Y. Matsuura, Y. Nakata, M. Takeo, D.-i. Kato and S. Negoro, "Molecular Design of a Nylon-6 Byproduct-degrading Enzyme from a Carboxylesterase with a Beta-lactamase Fold", *FEBS Journal*, Vol. 276, No. 9, pp. 2547–2556, 2009.

86. Ohki, T., N. Shibata, Y. Higuchi, Y. Kawashima, M. Takeo, D.-i. Kato and S. Negoro, "Two Alternative Modes for Optimizing Nylon-6 byproduct Hydrolytic Activity From a Carboxylesterase With a Beta-lactamase Fold: X-ray Crystallographic Analysis of Directly Evolved 6-aminohexanoate-dimer Hydrolase", *Protein Science*, Vol. 18, No. 8, pp. 1662–1673, 2009.

87. Toth, M., C. Smith, H. Frase, S. Mobashery and S. Vakulenko, "An Antibiotic-Resistance Enzyme from A Deep-sea Bacterium", *Journal of the American Chemical Society*, Vol. 132, No. 2, pp. 816–23, 2010.

88. Urbach, C., C. Evrard, V. Pudzaitis, J. Fastrez, P. Soumillion and J. P. Declercq, "Structure of PBP-A from Thermosynechococcus Elongatus, a Penicillin-Binding Protein Closely Related to Class A Beta-lactamases", *Journal of Molecular Biology*, Vol. 386, No. 1, pp. 109–120, 2009.

89. Chen, Y., W. Zhang, Q. Shi, D. Hesek, M. Lee, S. Mobashery and B. K. Shoichet, "Crystal Structures of Penicillin-Binding Protein 6 from Escherichia coli", *Journal of the American Chemical Society*, Vol. 131, No. 40, pp. 14345–14354, 2009.

90. Sweden, E. E. T., "WHO Publishes Global Tuberculosis Report 2013", *Euro Surveillance*, Vol. 18, No. 43, 2013.

91. Xu, H., S. Hazra and J. S. Blanchard, "NXL104 Irreversibly Inhibits the Beta-lactamase from Mycobacterium Tuberculosis", *Biochemistry*, Vol. 51, No. 22, pp. 4551–7, 2012.

92. Lahiri, S. D., S. Mangani, T. Durand-Reville, M. Benvenuti, F. De Luca, G. Sanyal and J. D. Docquier, "Structural insight into potent broad-spectrum inhibition with reversible recyclization mechanism: avibactam in complex with CTX-M-15 and Pseudomonas aeruginosa AmpC beta-lactamases", *Antimicrobial Agents and Chemotherapy*, Vol. 57, No. 6, pp. 2496–505, 2013.

93. Pryka, R. D. and G. M. Haig, "Meropenem: A New Carbapenem Antimicrobial", *Annals of Pharmacotherapy*, Vol. 28, No. 9, pp. 1045–54, 1994.

94. Bonfiglio, G., G. Russo and G. Nicoletti, "Recent Developments in Carbapenems", *Expert Opin Investig Drugs*, Vol. 11, No. 4, pp. 529–44, 2002.

95. Hugonnet, J. E., L. W. Tremblay, H. I. Boshoff, r. Barry, C. E. and J. S. Blanchard, "Meropenem-clavulanate is Effective Against Extensively Drug-Resistant Mycobacterium Tuberculosis", *Science*, Vol. 323, No. 5918, pp. 1215–8, 2009.

96. Kohler, J., K. L. Dorso, K. Young, G. G. Hammond, H. Rosen, H. Kropp and L. L. Silver, "In Vitro Activities of the Potent, Broad-spectrum Carbapenem MK-0826 (L-749,345) Against Broad-spectrum Beta-lactamase-and Extended-spectrum Beta-lactamase-producing Klebsiella Pneumoniae and Escherichia Coli Clinical Isolates", *Antimicrobial Agents and Chemotherapy*, Vol. 43, No. 5, pp. 1170–6, 1999.

97. Jeong, J. H., Y. S. Kim, C. Rojviriya, S. C. Ha, B. S. Kang and Y. G. Kim, "Crystal Structures of Bifunctional Penicillin-Binding Protein 4 from Listeria Monocytogenes", *Antimicrobial Agents and Chemotherapy*, Vol. 57, No. 8, pp. 3507–12, 2013.

98. Kawai, F., T. B. Clarke, D. I. Roper, G. J. Han, K. Y. Hwang, S. Unzai, E. Obayashi, S. Y. Park and J. R. Tame, "Crystal Structures of Penicillin-Binding Proteins 4 and 5 from Haemophilus Influenzae", *Journal of Molecular Biology*, Vol. 396, No. 3, pp. 634–45, 2010.

99. Yong, D., M. A. Toleman, C. G. Giske, H. S. Cho, K. Sundman, K. Lee and T. R. Walsh, "Characterization of a New Metallo-Beta-lactamase Gene, bla(NDM-1), and a Novel Erythromycin Esterase Gene Carried on a Unique Genetic Structure in Klebsiella Pneumoniae Sequence Type 14 from India", *Antimicrobial Agents and Chemotherapy*, Vol. 53, No. 12, pp. 5046–54, 2009.

100. King, D. T., L. J. Worrall, R. Gruninger and N. C. Strynadka, "New Delhi Metallo-Beta-lactamase: Structural Insights into Beta-lactam Recognition and Inhibition", *Journal of the American Chemical Society*, Vol. 134, No. 28, pp. 11362–5, 2012.

101. Fawcett, T., "An Introduction to {ROC} Analysis", *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 861–874, 2006.