2
Literature Survey

This chapter is aimed at providing an overview of computational strategies implemented for aiding drug discovery, with a focus on those used for prediction of side effects aligned with the theme of this thesis. We start by discussing challenges in drug discovery process and highlighting the relevance of side effects prediction. We further point out some of the computational strategies used in drug discovery and the concepts of graph theoretical modeling as well as biological space. Machine learning premises used for feature selection and the motivation behind using linear models are also discussed to set the stage for specific research presented in the thesis.

2.1 ASSESSING ADVERSE DRUG REACTIONS

Progress in concepts in chemistry that are critical for understanding drug actions (such as periodic table, Avogadro's hypotheses, acid-base interplay, discovery of aromatic structure) has added to our knowledge significantly in last few centuries. This has provided impetus to pharmacology improving the drug discovery process. Recent advances in molecular biology and genomics, along with computational biology, have opened avenues for future medicine towards curing and controlling diseases [Drews, 2000]. Moving beyond the discovery of therapeutic solutions that were primarily obtained by trial and error, today drug discovery has become a far more rational process. Its progress and speed are heavily aided by methods in computational chemistry such as virtual screening, de-novo design, and lead optimization. Such procedures focus on developing drug-like molecules most suitable for interacting with a known target. Despite their utility, this process has been riddled with challenges involving computational complexity and lack of understanding of biological mechanisms, for finding appropriate ligand conformations, achieving rigorous optimization, molecular alignment, and substructure searching [Yang, 2010]. These problems limit the use of computational chemistry techniques in drug discovery process.

In addition to above-mentioned challenges, drug discovery is a costly and time consuming process as it requires rigorous testing of candidate molecules for their safety. Adverse drug reactions (commonly known as side effects) is one of the major factors needed to be addressed in search of therapeutic molecules. In this thesis, we focused on data-driven investigations of side effects, to address this cornerstone of drug discovery pipeline. While it may be possible to identify drugs having good binding affinity with a target, ensuring that it does not have off-target interactions, giving rise to phenotypic side effects, is an important unsolved problem. Even rationally designed target-specific drugs show undesired reactions, which could be attributed their non-specific interactions. To cite an example, drug combination fenfluramine-phentermine targeted against 5-HT_{2B} (serotonin) receptor as an agonist of weight-loss is reported to cause heart valve problems, pulmonary hypertension and rapid heartbeat [Avron, 2005; Meyer, 2013]. Another example is that of a recreational drug 3,4-Methylenedioxymethamphetamine (that primarily binds to neurotransmitters such as dopamine and serotonin), which is known to generate empathy, euphoria while also causing depression, memory loss, teeth grinding [Meyer, 2013].

Side effects are primarily assessed through animal studies and human trials. Such studies are often time consuming and resource intensive. Furthermore, animal ethics policies, such as Three R's model (replacement, reduction and refinements) implemented in European Union and USA, mandate strict guidelines prohibiting large scale animal studies [W. M. S. Russell, 1959].

With the availability of large amount of data accumulated over decades, data analytics strategies are of immense value towards enhancing drug discovery protocols. Computational strategies coupled with empirical data analysis have tremendous scope for complementing experimental techniques towards addressing the problem of adverse drug reactions.

2.2 COMPUTATIONAL STRATEGIES USED IN DRUG DISCOVERY

In last few decades, various computational strategies have been developed for aiding complex drug discovery process. These methods deal with target structures and their interaction with small molecules. Herein, we summarize some of these methods along with their utility.

2.2.1 Molecular Docking and Dynamics

Molecular docking and dynamics simulate binding of two molecular entities (e.g. protein and drug) to predict their binding affinity and stability of the complex. Molecular docking is used to predict the preferred orientation of a ligand (drug) bound to a target (receptor) to form a stable complex. In a simpler docking strategy, the target is assumed to be static, and ligand conformations are varied to find a stable complex. In a more realistic strategy (flexible docking), both are considered to be flexible to search for a stable conformation. With the availability of better computational capabilities, high resolution docking is performed by simulating molecular interactions at picosecond-level resolutions. The dynamics of atomic interactions in molecules can be simulated using molecular dynamics (MD) [Haile, 1992]. MD captures the mean structural information by computing ensemble of all states. Force field methods, including advances such as SWARM-MD and targeted-MD, are used in MD for finding minimum energy conformation.

Molecular dynamics has been successfully used for drug development on many occasions. For a known inhibitor of HIV integrase, MD simulations were found to be useful in elucidation of a novel binding trench, which was undetectable from static structure. The knowledge of this trench was further used for development of raltegravir, a drug used to treat HIV infection [Summa et al., 2008]. Inhibition of a-glucosidase is a known mechanism for suppressing glucose absorption and lowering postprandial blood glucose level. By screening a library of compounds using MD simulations, four novel inhibitors of a-glucosidase were identified [Parka et al., 2008]. Similarly, in the case of human African trypanosomiasis, MD was found to be useful in identification of inhibitors of RNA Editing Ligase-1 enzyme of *T. brucei*, a parasitic kinetoplastid causing the disease [Durrant et al., 2010].

2.2.2 Quantitative structure-activity relationship (QSAR)

QSAR is used to create a mathematical relationship between quantitative expression of chemical structure and its biological activities [Nantasenamat, Isarankura-Na-Ayudhya, Naenna, and Prachayasittikul, 2009]. It is used widely to find drug-like molecules, their bioactivity and ADMET properties [Norinder and Bergström, 2006]. This methodology mostly relies on linear relationships that are realized with linear regression and partial least square method. It correlates biological activities such as affinities of ligands, inhibition constants, rate constants, with structural features such as lipophilicity, polarizability, electronic and steric property. This technique has also been applied to properties like narcotic, bactericidal, fungicidal, hemolytic and toxic properties.

In one of the studies, regular field approach was applied to quantify chemical properties and PCA model was used for prediction of biological activities. Later, partial least square was applied to achieve better prediction performance. Using this strategy, inhibitors of *E. coli* and *L. casei* were obtained with the help of physicochemical properties of ligands [Kubinyi, 1997b]. Lipophilicity of drug, which represents its absorption and distribution level, has been observed to correlate with certain features (van der Waals volume and hydrogen-bonding capability) of blood-brain barrier penetration of various H_2-receptor antagonists and some CNS-active drugs,

using QSAR studies [Kubinyi, 1997a]. For this purpose, various non-linear, bilinear models were devised to predict the optimum level of lipophilicity. Such studies have serious pharmacological relevance as increased absorption are linked to adverse effects of nervous system.

Other than molecular docking, MD, and QSAR, many other techniques have been developed for drug screening. Techniques such as scoring function and genetic algorithm have been implemented for lead optimization. Scoring functions that are based on force field, empirical data, or consensus scoring, are employed for identification of stable complexes in docking studies [S.-Y. Huang, Grinter, and Zou, 2010]. Genetic algorithms utilize heuristic optimization strategies for predicting minimum energy conformations of drug-target complexes [Jones, Willett1, Glen, Leach3, and Taylor, 1997].

2.3 COMPUTATIONAL MODELS FOR SIDE EFFECTS PREDICTION

Existing strategies presently used for assessing side effects are limited in their abilities. In the presence of large amount of data available on drugs, their molecular targets, gene sequences, molecular interactomes, specifics of pathways and such, mathematical and computational methods acquire significant importance. Advances in computational techniques, that facilitate dealing with large amount of data in search of patterns, are enabling data-driven models towards prediction of side effects.

Among the computational models developed for prediction of side effects, nonlinear models tend to be computationally complex. In the presence of large number of features, they tend to over-fit data leading to non-optimum solutions. Hence, we focused on linear models and studied those reported to work well in drug discovery. As a strategy, linear models could be used to identify a ranked list of features critical for specifying side effects, which could further be embedded in nonlinear models to improve time complexity and to avoid over-fitting.

Linear models optimize associations among variables assuming linear relationships. The data of drug targets and side effects are naturally structured as matrices, and could be modeled using linear framework within the scope of advanced matrix theory. Various such methods have been successfully applied for prediction of drug targets as well as side effects: Linear discriminate analysis, Network diffusion model, Canonical correlation analysis, and Bipartite modeling. Herein we summarize these models along with their applications.

2.3.1 Linear discriminate analysis

Linear discriminate analysis (LDA) is applied for binary classification of data by projecting the dataset in a space where means and standard deviations of these two classes are brought closer and farther from each other, respectively. First proposed by Fisher in 1936 [Fisher, 1936], LDA has been successfully used for classification of disease (thyroid and cancer) associated genes and control genes [Guo, Hastie, and Tibshirani, 2007; Luo, Kim, Dighe, and Kim, 2011]. This method uses intra-class and inter-classes scatter matrix which represent mean and standard deviation of classes, respectively. This strategy is known to lead to situations when scatter matrix of inter classes becomes singular, leading to null parameters. Various approaches have been proposed to tackle this problem [Guo et al., 2007; Li, Haifeng and Jiang, Tao and Zhang, 2006]. For multiclass classification, Lu et.al. have proposed a method for addressing the problem of singularity [Lu and Liang, 2016]. However, LDA cannot be used for simultaneous (multi-label) classification problems such as predicting the entire side effects or target profile, as it assigns data point to only one class. To resolve this problem, one could use LDA by converting the task involving n classes, into a multi-label classification task with n number of binary classifiers. This solution can be logically extended for multiclass classifiers by assigning classes after ranking the prediction scores.

2.3.2 Diffusion model

This model performs multi-label classification by combining local information from nearest data points with global information obtained from network diffusion [Shrager, Hogg, and Huberman, 1987]. The algorithm uses an iterative strategy involving spread of information from nearest data points to global state. Here, the global state corresponds to a cluster of similar data points [Zhou, Bousquet, Lal, Weston, and Schölkopf, 2004]. Various kernel-based diffusion models have been proposed for improving strategy for identification of clusters of similar data points.

Atias et.al. have used diffusion model for predicting side effect profile of drugs [Atias and Sharan, 2011]. Their strategy starts by assigning nearest data points, using chemical structural similarity with drugs, to further arrive at the final clusters through diffusion on side effects similarity network (defined based on shared drugs). Apart from prediction of side effects, diffusion model has also been applied for protein function prediction and identification of disease associated gene clusters [Bersanelli, Mosca, Remondini, Castellani, and Milanesi, 2016; Inoue, Li, and Kurata, 2010; Sun, Ji, and Ye, 2008].

2.3.3 Bipartite graph learning model

This is a supervised method and involves modeling relationships between datasets of two distinct types. Graphically, such a system could be imagined as a bipartite graph. Among the methods that can model homogenous data, such as protein-protein interactions, are kernel CCA [Y Yamanishi, Vert, and Kanehisa, 2004] and distance metric [Vert, Jean-philippe and Yamanishi, 2005]. For relating heterogeneous datasets such as those involving ligand-protein interactions or drug-side effects, we need to invoke bipartite model [Yoshihiro Yamanishi, 2009]. This model assimilates information about two covariance matrices for each type of data and one matrix of relationships across two domains. In case of drug-side effects data, the model would involve covariance within drugs, within side effects, and across drug-side effects. Here, the across domain relationships could be defined using the connectivity (similarities) of side effects. In a study implementing bipartite graph learning model, Yamanishi and colleagues used a strategy for finding unknown drug-target interactions in which target-target relationship was defined using sequence similarity score [Yoshihiro Yamanishi, Araki, Gutteridge, Honda, and Kanehisa, 2008; Yoshihiro Yamanishi, Kotera, Kanehisa, and Goto, 2010].

2.3.4 Canonical correlation analysis

Canonical Correlation Analysis (CCA) finds an optimum linear map between two feature domains by projecting them into a common subspace so as to maximize their correlation. CCA was first implemented for prediction of side effects by Atias and Sharan, using chemical structures and side effects profiles [Atias and Sharan, 2011]. This was a breakthrough study as it involved prediction of entire side effects profile, yielding very good results. Going beyond CCA, Pauwels et.al. proposed sparse canonical correlation analysis (SCCA), an improvisation of CCA [Pauwels, Stoven, and Yamanishi, 2011]. This method brings in sparcity in latent features so as to reduce false positive rate yielding better prediction performance. The authors predicted side effects using chemical substructures as a basis, and concluded that SCCA outperforms neural network, support vector machine as well as CCA [Pauwels et al., 2011]. Yamanishi et.al. compared the performance of CCA and SCCA along with different target profiles comprising of information about G-protein-coupled receptors, enzymes and ion channel proteins [Mizutani, Pauwels, Stoven, Goto, and Yamanishi, 2012]. By comparing performance of target and chemical profiles, it was concluded that the performance of the former is better than the latter. Between these two methods, the performance of SCCA was observed to be slightly better than CCA.

The details of CCA as well its analytical derivation has been presented in Chapter 4. Application of CCA and its improvisations are presented in Chapter 5, Chapter 6, Chapter 7 and Chapter 8.

2.4 SYSTEMS MODELING

In a cellular system, mechanism of drug action is an event involving a host of complex chemical reactions among biomolecules such as genes, proteins, enzymes, ions etc. Therefore, it is imperative to include as many details of these interactions as possible for analyzing side effects. In the context of computational problem of side effects prediction, systems modeling refers to integration of relevant biochemical features of drugs. These features could include, both, the primary data (target profiles, chemical structures, chemical properties) as well as secondary data (similarity of drugs based on shared targets, chemical structures, 3D chemical properties, 2D chemical properties, ADMET). Herein, we describe a few studies involving systems level modeling.

Compounds that are similar by virtue of their chemical features are likely to share biological function as well as side efects [Dunkel, Gunther, Ahmed, Wittig, and Preissner, 2008]. With this premise, Huang et.al. proposed a chemical similarity score by integrating chemical structure and shared targets information, which was further used to rank the potential side effects [Chen, Huang, et al., 2013]. Another study reported integration of chemical structures, targets and features from PPI network. Such an integration of information yielded better prediction performance compared to individual features [L. C. Huang, Wu, and Chen, 2013]. Chen. et.al. investigated cardiac adverse drug reactions by integrating different data such as PPI, gene ontology, and chemical structures. In this work, logistic regression and SVM were used as a tool for predicting adverse reactions. They observed that holistic representation of data offers best performance [L.-C. Huang, Wu, and Chen, 2011]. Yamanishi et.al. proposed a kernel based regression model for integrating chemical similarity and target similarity information to predict side effects, and concluded that the integrated model outperforms reductive models [Yoshihiro Yamanishi, Pauwels, and Kotera, 2012a]. In addition to exploiting information of drug targets and sub-structures, Liu et.al. used pathway associations and treatment indications for predicting adverse drug reactions. In order to refine the procedure to predict side effects, they implemented different machine learning tools and concluded that profile integration is an effective approach [M. Liu et al., 2012]. In Chapter 5, we present a systematic integration of target profile with 3D and 2D chemical properties to compare the relevance of these profiles for predicting side effects.

2.5 GRAPH THEORETICAL MODELING

Since the elements that form the causal basis for expression of adverse reactions are interconnected, it is natural that graph theoretical framework has been used as a modeling premise in endevors involving prediction of side effects. Function of drugs depends on pathways with which the drug is known to interact with. Fukuzaki et.al. proposed a method for identification of pathways that participate in cellular functions under a set of identical conditions, for investigating side effects [Fukuzaki, Seki, Kashima, and Sese, 2009]. They termed such pathways as 'cooperative pathways' and modeled them as a graph where each node represents a target and edges represent metabolic reactions. Further, sub-graphs representing parts of pathways that are associated with a common set of drugs were identified. They concluded that that big pathways associated with large number of shared drugs were more closely linked to side effects.

In another study, through in vitro investigations, Campillos et.al. reported that shared targets among drugs is significantly correlated with sharing side effects [Campillos, Kuhn, Gavin, Jensen, and Bork, 2008]. With examples, the study also reported that sharing of targets is a more robust measure as compared to sharing of chemical sub-structures for prediction of side effects. As opposed to focusing on drug similarity based on their target profiles or chemical profiles, Brouwers et.al. studied side effect similarity in drugs by finding connection between their targets in the network representing protein-protein interactions [Brouwers, Iskar, Zeller, van Noort, and Bork, 2011]. However, they concluded that such a feature was only a marginal factor. Graph theory could also be used for predicting drug-target interactions. Towards this end, Alaimo et.al.

have reported an effective strategy for enumerating domain-domain interactions [Alaimo, Pulvirenti, Giugno, and Ferro, 2013].

2.6 BIOLOGICAL SPACE

Biological space is a hypothetical space intended to represent relevant biological mechanisms that form molecular basis of drug interactions in the cellular milieu. This space could be categorized into different types such as Genomic space, Chemical space, Chemo-genomic space and Pharmacological space etc. Each of these spaces could be constructed by appropriately incorporating details of gene sequences, protein structures, chemical features, and other pharmacologically relevant details. Herein, we present description and applications of these graph theoretical representations that form the basis of drug interactions and hence their side effects.

2.6.1 Genomic space

'Genomic space' represents biologically relevant relations among genes or proteins modeled as a graph. Relationships among genes could be encoded in terms of sequence similarity, number of shared drugs or based on their expression profiles. Experimentally validated protein-protein interactions could also be used to establish links between the elements in genomic space.

In one of the studies, Yamanishi et.al. employed genomic space constructed based on sequence similarity [Yoshihiro Yamanishi et al., 2008]. They proposed various models, such as nearest neighbor, weighted nearest neighbor, and bipartite graph learning, that derive features from the genomic space for associating potential targets to a query drug. Bipartite model uses genomic space for training interactions of a drug with unknown genes, which was observed to yield better prediction performance compared to the other two models. Zhao and Li proposed a strategy that integrates information from genomic space, chemical space as well as therapeutic space for predicting potential targets of a query drug [Zhao and Li, 2010]. Sequence similarity between protein sequences has also been used to create the genomic space. Yaminishi et.al. implemented Smith-Waterman algorithm for this purpose [Yoshihiro Yamanishi et al., 2010]. To assess potential drug-target interaction between a query drug and query target, they integrated features derived from genomic space as well as those from chemical space. Other than predicting target of a drug, it is also important to predict off-target interactions that are central to expression of side effects. Towards this Takarabe et.al. modeled genomic space [Takarabe, Kotera, Nishimura, Goto, and Yamanishi, 2012] that computes gene sequences similarity score using a local kernel alignment technique that was proposed by Saigo et.al. [Saigo, Vert, Ueda, and Akutsu, 2004]. By integrating side effects information with genomic space, they predicted probable off targets of drugs. A drug could be associated with a disease other than targets. Towards this objective, Dai et.al. have proposed computational framework based on matrix factorization [Dai et al., 2015]. They used genomic space defined with the help of gene sequence similarity for assigning a disease feature vector to every drug. Their study suggested that including information from genomic space improves prediction of drug-disease association.

2.6.2 Chemical space

Chemical space represents a graph theoretical model of meaningful associations between drugs. Typically these associations are encoded by drug similarity. In this space, where drugs are nodes, edges may represent number of shared targets, extent of side effects sharing, Pearson's correlation measured between chemical structures of drugs or between their physicochemical properties. The similarities measured using above strategies could be improvised using regular equivalence, for example by using information from the neighborhoods in chemical space [M.E.J Newman, 2010]. Drug-drug relationships could also be enumerated by recruiting kernel modeling with kernels such as radial basis, gaussian as well as hybrid kernels [Re and Valentini, 2012; Vert, Tsuda, and Schölkopf, 2004].

Chemical space is an integral part of bipartite graph learning and network diffusion models. Hence, the judicious creation of chemical space is critical for the model performance. For dug repositioning, chemical similarity is known to play a crucial role in identification of unknown interactions [Noeske et al., 2006; Re and Valentini, 2012]. Structure-based similarity of drugs has been used to predict unknown biological function [Chen, Zeng, Cai, Feng, and Chou, 2012; Hu, Chen, Huang, Cai, and Chou, 2011]. Finding drug combinations is active research field for achieving better treatment of diseases. In this direction, Chen et.al. showed the utility of chemical space constructed based on shared targets, physical interactions and membership of pathways [Chen, Li, et al., 2013]. Chemical space has also been shown to be of value for the prediction of bioactivity of candidate drugs [Glick and Jacoby, 2011; Klekota and Roth, 2008]. This feature has also been shown to be useful for suggesting drug indications for a new disease [G. Huang, Lu, Lu, Zheng, and Cai, 2015].

2.6.3 Chemo-genomic space

Chemo-genomic space is defined by projecting drugs on genomic space. Such a space has been used for therapeutic applications, towards the identification of unknown drug interactions [Rognan, 2007]. Zhao et.al. have proposed a framework for measuring potential interactions of a drug. Towards this end, starting with the knowledge of known targets, they used protein-protein interaction network (genomic space) as a premise to predict potential drug interactions. Yamanishi et.al. have also implemented chemo-genomic space using bipartite graph learning model [Yoshihiro Yamanishi et al., 2008, 2010].

2.6.4 Pharmacological space

Pharmacological space embeds drug information that has bearing on its pharmacological properties. This space could be defined to include details of individual features or to create a systems model by integrating multiple features. Using individual features, drug similarity can be computed with various metrics such as Jacard index, Tannimotto coefficient, Pearson coefficient, kernel functions etc. [Michael Levandowsky and Winter, 1971; Pearson, 1895; T.T.Tanimotto, 1957; van Laarhoven, Nabuurs, and Marchiori, 2011]. Other than this, drug similarities have also been computed using target interactions and Anatomical Therapeutic Chemical code which specifies pharmacological responses [Z. Liu et al., 2015; Yıldırım, Goh, Cusick, Barabási, and Vidal, 2007; Zhao and Li, 2010]. Using integrative strategies, linear combinations of chemical properties as well as embedding chemical features into models such as bipartite model, kernel regression model, have also been used to predict the chemical response of unknown drugs [Chen et al., 2012; Yoshihiro Yamanishi et al., 2008; Yoshihiro Yamanishi, Pauwels, and Kotera, 2012b; Zhao and Li, 2010].

2.7 FEATURE SELECTION

When dealing with large set of features, it is desirable to identify the most informative subset of features to build an effective predictive model. In this thesis, we dealt with an array of drug features of heterogeneous nature, making it necessary to use feature selection strategies. In machine learning, feature selection is implemented with three broad objectives as described below.

Objectives of feature selection

1. Model simplification: Selection of features facilitates use of simple classification model.
2. Complexity reduction: Selection of relevant features is useful for improving training time, especially in methods such as artificial neural network and deep learning.
3. Avoiding over-fitting: Training with large number of features biases the model towards training data rather than adapting their general characteristics. Thus, the model may become oversensitive to small deviations in data. Implementing feature selection by pruning less informative features, can enable the model to avoid this problem.

Feature selection methods are broadly divided into three categories: wrapper, filter, and embedded method. We briefly explain these methods and their application for feature selection in drug discovery, in the following sections.

2.7.1 Wrapper method

Wrapper method uses the classifier over randomly chosen subsets of features to identify most informative subset by comparing their performances. This method is fairly effective and usually identifies the most informative feature set. However, it is computationally intensive as it needs to evaluate a large number features sets. Wrapper methods are divided into two types: Deterministic and Heuristic.

Deterministic: As part of deterministic wrapper methods, algorithms such as sequential forward selection and sequential backward selection are used for reducing complexity. Sequential forward selection method starts by assigning random features to a set, according to some objective function, which is appended with random features. This procedure is continued until an optimal features set is obtained or the performance of successive feature sets does not change significantly. The backward selection works opposite to forward selection. Starting with all features, a feature is eliminated if its inclusion does not change the objective function significantly. 'Plus-q-take away r' algorithm combines these procedures for improving performance. This strategy selects q features and discards r of them, to simultaneously implement forward and backward selection strategy [Ferri, Pudil, Hatef, and Kittler, 1994]. Wrapper method has been applied for selection of an optimum set of genes used for disease diagnosis [Iñaki Inza, Larrañaga, Blanco, and Cerrolaza, 2004; Xiong, Fang, and Zhao, 2001].

Heuristic: Among the heuristic strategies implemented in wrapper, among the most common strategies are genetic algorithm and estimated distribution. Estimated distribution is an evolutionary algorithm [Goldberg, 2006], where typical crossing over and mutation steps are executed by factorization of the joint probability distribution of features [I Inza, Larrañaga, Etxeberria, and Sierra, 2000]. Due to the complex nature of these computations, genetic algorithm are preferred to Estimated distribution method [Jirapech-Umpai and Aitken, 2005; Ruiz, Riquelme, and Aguilar-Ruiz, 2006].

2.7.2 Filter method:

In contrast to wrapper methodology, filter methods are designed without a classifier for feature selection. For this reason, it is known to have inferior prediction performance compared to wrapper method. However, these methods are effective in producing ranking of features with less computational complexity. Thus, it is useful as a preprocessing step for leveraging wrapper when applied over a large set of features. Among the popular filter methods are Mutual information, Markov blanket, and Fast-correlation based method etc. Mutual information procedure is used for measuring the relevance of a feature to know another feature, to further identify the less mutual informative feature set. It is widely used for identifying informative genes from large number of microarray experiments. In one of the studies, genes thus identified were further trained for classification of cancer tissue [Ding and Peng, 2003]. In another study, this method has been extended to implement 'minimum redundancy maximum relevance' [Peng, Long, and Ding, 2005]. Another method is Markov blanket which enhances the predictive ability of a feature by considering behavior of adjacent features. Filter method is limited in not being able to capture joint relevance of features, thereby ignoring them in isolation, which is addressed by Markov blanket [Kohavi and John, 2011]. Gevaert et.al. and Yu et.al. have used it for prognosis of breast cancer, and for classification of tuberculosis from cancers, respectively (Gevaert, De Smet, Timmerman, Moreau, & De Moor, 2006; K. Yu, Wu, Ding, Mu, & Wang, 2016).

But, finding the set of dominant features through Mutual information and Pearson correlation coefficient procedures are highly inefficient for large number of features. Fast

correlation method can be used to reduce complexity, which iteratively chooses the predominant feature based on the rank in terms of correlation with target class and removes features on the basis of Markov blanket criterion, until an optimum set is achieved. Apart from reducing complexity, another advantage of Fast correlation method is that it can produce feature set by considering non-linear correlations [L. Yu and Liu, 2004]. We have implemented feature selection for studies presented in Chapter 8, for finding most independent set of drug features to correlates with a class of side effects. Given the nature of the problem, feature selection method could not applied in this case.

2.7.3 Embedded method

Embedded method is considered to be most effective in terms of computational complexity as it does not rely on random selection of features and also implements a model for feature tuning. LASSO (Least absolute shrinkage and selection operator) is primarily used for implementation of this method [Tibshirani, 1996]. Essentially it uses constraint of type L1-norm to make the objective function intersect a hyperplane, resulting in nullification of insignificant features. Also, using constraint equation one can restrict the number of chosen parameters. Improvisation of this procedure include Bolasso, which follows the bootstrapping of samples [Bach, 2008], and FeaLact which weighs the features according to regression coefficients [Zare, Haffari, Gupta, and Brinkman, 2013]. Bolasso is a hard feature selection technique, whereas FeaLect considers entire features profile for each sample to approve features based on their resultant score. LASSO based techniques are more suitable for binary classification.

Another feature selection methodology, among Embedded methods, is 'greedy hill climbing' which implements an iterative method for improving local optima. Starting with a random choice of features set, at each iteration it adds a feature seeking for better solution. Simulated annealing, Targeted projection pursuit and Greedy backward elimination belong to the class of Embedded methods, and have been applied for classification of gene expression data [Faith, Mintram, and Angelova, 2006]. Although, this type of feature selection has the problem of stopping criterion and convergence with locally optimized solution.

Guided by different procedures of wrapper and embedded methods, we addressed classification of features for set of side effects with PCCA in conjunction with Wrapper method. The detailed description of PCCA technique and its application for identification of crucial features are provided in Chapter 4 and Chapter 8, respectively.

2.7.4 Applications for drug discovery

Feature selection methods are of immense value in reducing the search space. QSAR is an important technique used for drug discovery, and relies on a large number of chemical descriptors such as molecular structure, field parameters, quantum properties etc. This can create several problems such as model complexity and overfitting. Prior features selection is a good remedy for finding most relevant features. QSAR has also been used with wrapper style, in conjunction with partial least square and genetic algorithm methods [Mehmood, Liland, Snipen, and Saebø, 2012; Venkatraman, Dalby, and Yang, 2004]. In another study, classification of active and inactive compounds was performed in conjunction with five different filter methods, such as information gain, mutual information, χ^2-test, odds ratio, and Galavotti-Sebastiani-Simi coefficients. From this study, it was concluded that use of odd-ratio as a preprocessing step improved predictive performance. Although, in some studies it was observed that filtering of features was not useful for improving SVM performance [Ying Liu, 2004]. Feature selection was implemented for applications of biomedical relevance such as prediction of cell cycle response, screening chemicals and gene function discovery, using high content microscopic images with multi-dimensional image features [Shariff, Kangas, Coelho, Quinn, and Murphy, 2010].

2.8 PREDICTION OF ABSENT INTERACTIONS

So far we have discussed the use of drug properties, substructures and target information towards prediction of side effects. Other than these features derived from chemical structures and biological space, the knowledge of 'known side effects' could itself be a used as a descriptor/feature that embodies chemical interactions of a drug. The empirically derived side effects profile of a drug could be treated as incomplete information, to predict the 'remaining side effects'. Interestingly, the problem of 'finding unknown side effects' could be treated as analogous to a 'recommender system' or 'a link prediction problem'. Herein, we describe two link prediction methods which have been used in drug discovery.

Restricted-Boltzmann machine

Restricted Boltzmann Machine (RMB) is a stochastic artificial neural network methods and was first proposed by Paul Samlonaski in 1986 [Smolensky, 1986]. It learns the mapping or weight between the known and hidden data through probability distribution. This method has wide range of applications such as dimension reduction, classification, and clustering. This method has also been extended for applications in deep learning and in particular deep belief network. Wang et.al. applied RBM for finding new usage of existing or disapproved drugs with the knowledge of known target interactions [Y. Wang and Zeng, 2013]. Zhang et.al. combined RBM and integrated neighborhood based method for predicting unknown side effects, starting with the information of known adverse reactions [Zhang et al., 2016].

Perturbation modeling

Real-world networks are known to have patterns of connectivity that are maintained in the presence of perturbations (such as addition and deletion of edges). This property could be exploited to predict unknown or future links. Perturbation methodology is employed for restoring the network in the presence of small changes (perturbations) in the network. Lu et.al have studied link predictability in networks such as collaboration network, *C. elegens*, metabolic network, using perturbation modeling [Lü, Pan, Zhou, Zhang, and Stanley, 2015]. Search for multiple drug therapy is one of the endeavors in modern drug discovery. Towards this, spotting effective drug combinations is important to identify potential drug-drug interactions. Zhang et.al. have reported the application of perturbation modeling for identification of unknown drug-drug interactions [Zhang et al., 2017]. This method could also be applied in other areas of drug discovery, including prediction of unknown side effects.

…