1
Introduction

1.1 MOTIVATION

Progress in modern medicine heavily relies on discovery of new molecular entities. However, arduous nature of research and development required for discovery of new drugs limits its pace. Despite advances in processes involved and improved understanding of biological systems, drug discovery remains an inefficient task, with the high attrition of candidate molecules. Adverse reactions is one of the key factors contributing to rejection of molecules with therapeutic potential. Hence, accurate prediction of phenotypic side effects is an important problem in drug discovery.

A living cell is a complex dynamical system of biomolecules interacting at different hierarchies. In this web of molecular interactions with intricate cross-talks among genes, proteins, metabolites and small molecules, drugs act as agents of molecular control. As a consequence, while drugs are intended for therapeutic effect, they often lead to side effects through unintended interactions with cellular processes. Hence while modeling factors responsible for side effects, it is important to account for the intertwined nature of cellular mechanisms.

Beyond experimental protocols that are traditionally used to assess the effectiveness of candidate molecules and their therapeutic value, computational methods have the potential to transform drug discovery by complementing existing approaches. Availability of large amount of open data on drug targets, their phenotypic side effects, gene and protein sequences, metabolic pathways, molecular interactomes and such, have opened new avenues for data- and hypothesis-driven questions. Accurate prediction of drug side effects is one the most exciting challenges on the interface of pharmaco-informatics. Towards addressing this challenge, the objectives set in this thesis were driven by the idea of creating holistic models using empirical data (DrugBank, SIDER, protein sequences), and devising mathematical as well as computational strategies.

1.2 OBJECTIVES

This thesis was aimed at developing mathematical and computational models to gain insight into mechanisms of side effects and their prediction. Specifically, we set following objectives.

- To build systems-level models aimed at prediction of side effects.
- To generalize canonical correlation analysis for integrating multiple drug features.
- To model genomic space of drug targets to assess secondary contributions.
- Identification of a minimal set of 'known side effects' to predict 'unknown side effects'.
- Identification of most independent drug features responsible for side effects.

1.3 CONTRIBUTIONS

Towards addressing the objectives of the thesis, we integrated data from existing resources such as DrugBank and SIDER for systems-level investigations of side effects. Inspired by Canonical Correlation Analysis (CCA), one of the most successful models implemented for prediction of adverse reactions, we developed an integrative Generalized Canonical Correlation

Analysis (GCCA) model which facilitates consolidation of various drugs features. Using GCCA, we conclude that models implementing chemical profiles have more consistent accuracy than those based on target profiles. Our studies highlight importance of chemical features in driving the accuracy of GCCA.

Further, we constructed a graph theoretical model to account for associations among drug targets, and studied contribution of various network metrics. We conclude that degree (which quantifies influence of neighbors of drug target) offers comparable performance with better complexity, than page rank, closeness, and betweenness metrics.

Knowing the complete side effects profile of a drug is of critical importance to leverage their therapeutic utility. Since most often such comprehensive profile is unavailable, we probed the possibility of using partial information of side effects for predicting the rest of the (unknown) profile. By dividing side effects into seven organ-specific classes, we implemented CCA to obtain subsets of side effects for each class, knowing which the remainder of side effects could be predicted with good accuracy. Thus our results point to 'partial side effects profile' as a possible factor for arriving at the remaining side effects, which is based on the hypothesis that phenotypic effects of a drug are interlinked.

The analytical treatment of GCCA model is heavily simplified by assuming that drug features integrated in it are independent. In reality, drug features are intricately linked with each other with shared mechanisms. Hence, to be able to build an effective computational model, it is extremely important to identify the contribution of individual features towards accurate prediction of side effects. One of the challenges in this direction is to disentangle interdependence of features to identify contribution of individual features that specify side effects. Towards our goal of obtaining features that contribute the most to side effects prediction, we developed a partial canonical correlation analysis (PCCA) model that facilitates enumeration of contribution from individual drug features.

1.4 OUTLINE OF THE THESIS

The main body of this thesis is organized into seven chapters. In Chapter 2, we provide survey of literature around the key themes on which the thesis rests. Chapter 3 describes the strategy implemented for compilation and curation of data, relevant for modeling side effects. Mathematical and technical premise for methods developed in this thesis is provided in Chapter 4. Chapter 5 deals with the objective of generalizing CCA, and presents our investigations with the generalized canonical correlation analysis model. In Chapter 6, we present our studies that enumerate secondary contributions from the neighborhood of drug targets in the 'biological space'. Our studies for identification of best minimal 'known side effects' to predict remaining side effects are presented in Chapter 7. Chapter 8 presents our generalized 'partial canonical correlation' model, and demonstrates its application for identification of drug features specific to side effects classes.

…