6
Role of secondary interactions in genomic space in specifying phenotypic side effects of drugs

In Chapter 5, we used data of chemical features and drug targets towards prediction of side effects. Chemical properties capture information about chemistry of molecules, whereas data of drug-target relationships brings in an additional layer of information about how drug as an external agent interacts with target, which forms part of the cellular machinery (Figure 6.1). Hence, going a step further, we added another layer of information that depicts the relationships among the targets. Towards this end, we endeavored to investigate contribution of not only targets but also their secondary effects, for prediction of adverse reactions. We constructed 'genomic space', a network depicting relationships among drug targets measured in terms of their sequence similarities. We studied graph theoretical properties (such as degree centrality, page rank, betweenness, and closeness) of this network in order to identify network features critically linked to side effects prediction. We find that degree (which quantifies its secondary neighbors) offers comparable performance with better complexity, as compared to page rank, closeness and betweenness.

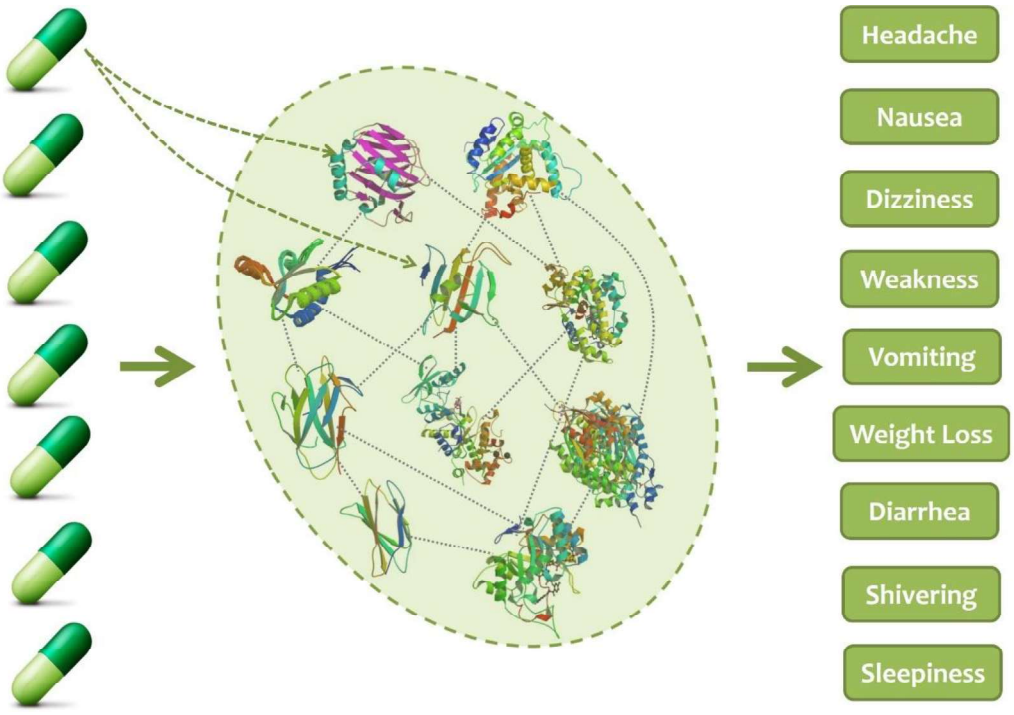

Figure 6.1 : Symbolic depiction of how drugs interact with cellular mechanisms leading to an array of reactions. Cell's pathways are shown as a network of molecular interactions. Starting with a therapeutic target, rational drug discovery aims at identification of chemicals that are specific regulators of such targets. In reality, drugs interact with cellular mechanisms in an unintended manner giving rise to adverse reactions. In this study, we created Genomic Space representing molecular network in cell, to assess the value of information derived from various graph theoretical features towards prediction of side effects.

In addition to the information of target(s) that interact(s) with a given drug, it is important to consider influence of other targets which could be critical for specifying drug side effects. We

reckon that the interaction of drug with target gene's neighborhood is one of the key reasons for side effects. This may be referred to as 'secondary information'. In this context, we constructed genomic space, based on sequence similarity between genes. We used graph theoretical features to enumerate importance of every target to model secondary information. Towards this end, the secondary information was encoded to capture various types of secondary influences using degree centrality, page rank, betweenness centrality, and closeness centrality (See Section 6.3). These centrality measures manifest local as well as global interaction characteristics of a target mediated by other genes.

6.1 GENOMIC SPACE

Genomic space (G) represents the inter-relationship among genes by virtue of sequence similarity (Figure 6.2). A drug may target a set of genes (shown in an oval). These targets are part of the genomic space in which they are related to other genes. Genes were linked to each other based on sequence similarity. Using information of gene sequences, the links were inferred using Smith-Waterman local alignment technique [Smith and Waterman, 1981]. The local alignment was preferred over global alignment, as drug-target interactions are better inferred using information of local regions. This is due to the fact that the chemical affinity between drug and target is known to be specific to small pockets in the target (binding sites).

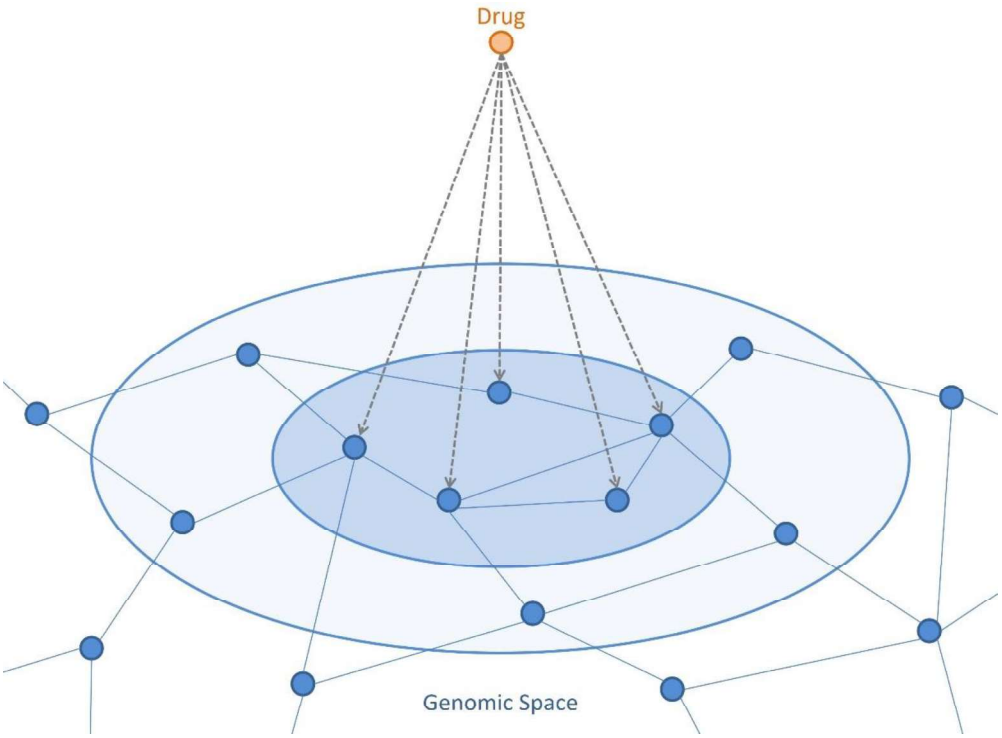

Figure 6.2 : Illustration of Genomic Space, and drug target interactions. The primary targets and their secondary as well as higher order interactions with the cellular milieu are depicted.

The genomic space comprised of all 3520 targets (genes) listed on DrugBank. Among these, 774 targets were associated with 837 drugs for which side effects information was available from SIDER2. For this subset of data, number of shared drugs for every pair of targets is shown in Figure 6.3. While most target pairs have no shared drugs, 4540 pairs of targets have at least one drug shared between them. Number of drugs shared falls exponentially, as shown in the graph, with no target pair sharing more than 33 drugs. Sharing of drugs between two targets suggests potential similarity between them (common binding sites).

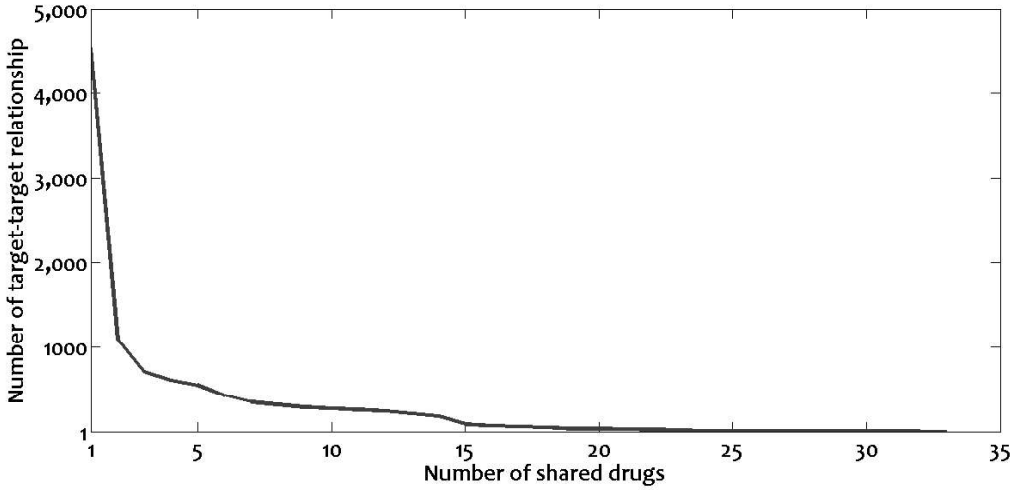

Figure 6.3 : Statistics of 'number of target pairs' in the Genomic Space and number of shared drugs.

Further, to create a target-target network embodying the genomic space, we applied a cut-off on sequence similarity. Figure 6.4 depicts the relationship between 'sequence similarity' and the 'number of shared drugs'. To avoid spurious target pairs that show up with shared drugs despite low sequence similarity, we applied a cut-off of 0.35 to identify meaningful target pairs that form the genomic space. The data suggest that, with this cut-off we are able to identify meaningful target pairs that are interlinked in the genomic space by virtue of similarity.

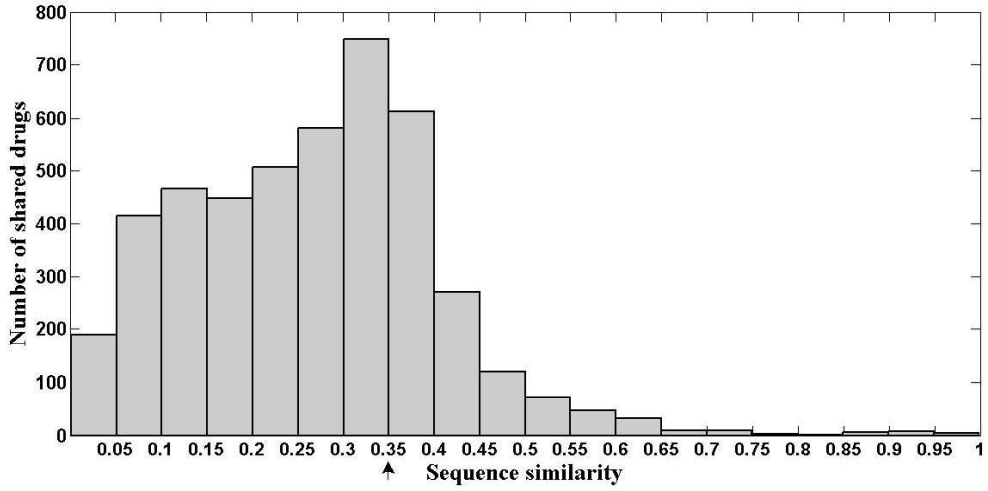

Figure 6.4 : Relation between 'number of drugs shared' by a pair of targets and their 'sequence similarity'. The sequence similarity cut-off used for creating the genomic space (0.35) is indicated with an arrow.

6.2 MEASURING SECONDARY INTERACTIONS USING GRAPH THEORETICAL METRICS

Network features characterize the network and quantify relative importance of nodes with the help of their centrality. In this work, we attempted to quantify the relevance of each target in terms of their network centrality measures. We anticipated that each of these measures may enable us to capture the higher order (secondary, tertiary) interactions, thereby providing a better handle to predict side effects. Formally a network can be represented as sets of nodes and sets of edges as, $G :=< V, E >$. where $|V|$ and $|E|$ are the number of vertices and edges respectively. Following are the definitions of metrics as defined for an undirected and unweighted network.

6.2.1 Degree

It enumerates the immediate or local information in terms of the number of neighbors. In Genomic space, for a given gene, degree represents number of other genes it shares sequence similarity above the cut-off value. Thus, this parameter reflects number of secondary interactions.

$$d_i = \sum_{j \neq i}^{|V|} A_{ij} \quad \text{Here, } j = 1,2,3,...|V| \tag{6.1}$$

Calculating degree centrality (d_i) for all the nodes in a graph has complexity $\Theta(V^2)$ or $\Theta(E)$ in a dense graph.

6.2.2 Page rank

Page rank captures information of important neighbors (targets) in Genomic space. In contrast degree centrality that treats each neighbors identically, it reveals centrality by assigning relative importance to its neighbors according to their degree. Eventually, it yields centrality of a node by tracing entire graph starting from its neighbors. Eigenvector centrality was the first method proposed using this notion. However, it does not consider centrality of isolated vertices and bias due to high degree nodes. Page rank addressed this limitation by modifying eigenvector centrality equation.

$$X_i = \alpha \sum_j A_{ij} \frac{X_j}{d_j} + \beta_j \tag{6.2}$$

Here, d_j represents degree of node j, which also ensures unbiased estimation of centrality. α and β_j are constant terms which have been introduced so as to avoid divergence and to address the centrality problem of isolated node, respectively. In matrix form Eq. (6.2) could be expressed as,

$$X = \alpha A D^{-1} X + \beta 1$$

Here, 1 represents the vector $(1,1,1,\cdots 1)$ and D is the diagonal matrix defined as $D_{ii} = \max(d_i, 1)$. This definition of diagonal matrix ensures that an isolated node would not lead to diverging condition. This equation could be iterated to compute the global connection, which reflects as page rank.

Lets assume $X(0) = 0$, and hence $X(1)$=β. We iterate this until Page rank values in consecutive steps becomes identical.

$$X(2) = \alpha A D^{-1} \beta + \beta$$

$$X(3) = (\alpha A D^{-1})^2 \beta + \alpha A D^{-1} + \beta$$

Repeating this equation for n times, eventually traces all the nodes in the network as shown below.

$$X = \sum_{i=1}^{n} (\alpha A D^{-1})^i \beta$$

$$X = (I - \alpha A D^{-1})^{-1} \beta$$

This equation shows α^{-1} to be the eigenvalue of AD^{-1}. Therefore to avoid divergence, either none of the eigenvalues should be α^{-1} or $\alpha < \frac{1}{k_i}$, where k_i is the highest eigenvalue. Thus the page rank can be expressed as $X = D(D - \alpha A)^{-1} \beta$.

6.2.3 Betweenness centrality

This term measures importance of a node in terms of number of times it is part of all shortest paths in the network. For a gene, it represents its relevance for global information dynamics over the Genomic space.

$$b_i = \sum_{s \neq i \neq t \notin V} \frac{P_{st}(i)}{P_{st}} \tag{6.3}$$

Here, P_{st} is a total number of shortest paths between s to t, and $P_{st}(i)$ are the number of shortest paths that pass through i.

6.2.4 Closeness centrality

It measures the global proximity of a node, by taking inverse sum of all shortest paths from rest of the nodes. It identifies genes that are critical for quick diffusion of information over the network, reflecting its central role in genomic space.

$$c_i = \sum_{j \neq i \in V} \frac{1}{d(j,i)} \tag{6.4}$$

Here, $d(j,i)$ represents length of shortest path between i to j. From computational complexity perspective, both betweenness and closeness centralities involve calculation of shortest paths between all pairs of vertices on a graph, with time complexity of $\Theta(V^3)$ using Floyd–Warshall algorithm [Floyd, 1962]. However, on sparse graphs Johnson's algorithm is more efficient with complexity of $O(V^2 \log V + VE)$ [Johnson, 1977; T. H. Cormen, C. E. Leiserson, R. L. Riverst, 2001].

6.3 REPRESENTATION OF FEATURES OF DRUG PROFILE

A drug may interact with a target via primary, secondary, or higher order interactions. To compare the role of primary interactions and secondary interactions, we considered graph theoretical parameters that enumerate different aspect of effects in the genomic space. Here, genomic space (G), as illustrated in Figure 6.1, refers to the unweighted network of all targets interlinked by virtue of their sequence similarities.

Lets say we have d drugs associated with t targets and s side effects. Each drug $x_{i=1,2,3\cdots d}$ is assigned with a target profile vector and side-effects profile vector having dimension $1 \times t$ and $1 \times s$ respectively. Hence, drug-target matrix D_t and drug-side effect matrix D_s have dimension of $d \times t$ and $d \times s$ respectively. The mathematical formula used for computation of primary as well as secondary interactions (degree centrality, page rank, closeness and betweenness) of a target $(T_{i=1,2,3\cdots 774})$ are described in Section 6.3. The entries in the drug-target binary matrix were replaced by the corresponding network feature. Further, CCA model was used to predict side effects profiles. The predicted profiles were compared with the empirical data to assess the value of each network feature.

6.4 RESULTS

Using the ordinary canonical correlation method (Refer to Chapter 4), we predicted side effects with 10-fold cross-validation, with each set containing 83 drugs. The predictions were made using the largest eigenvector of the drug-target matrix (For justification of use of largest eigenvector, please see Section 5.3). For each of these experiments, AUC, AUPR, Accuracy, F1 score and MCC were computed to assess the performance of network features. When we refer to drug-target matrix with computation of eigenvectors, we are referring to their corresponding P matrices (For more details, please refer to Section 4.2.2). Figure 6.5 and 6.6 depict prediction performance using various network measures encoding different information of secondary interactions. We observed that betweenness, degree and page rank features were better than closeness in terms of capturing relevant secondary information for side effect prediction, except when evaluated with F1 score.

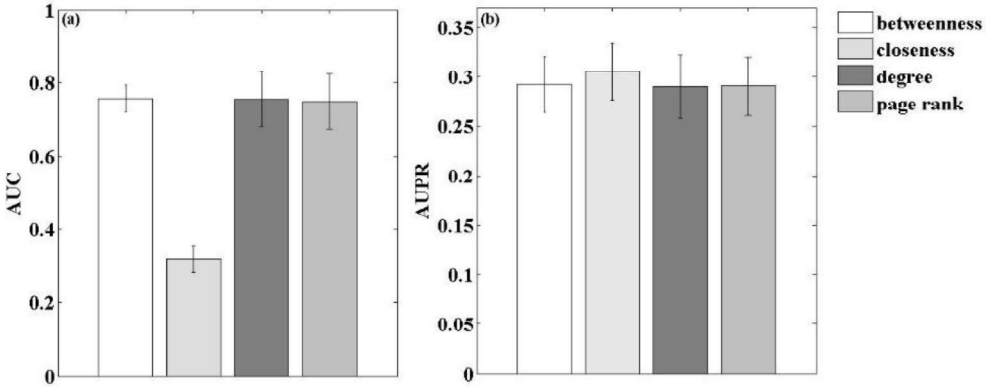

Figure 6.5 : (a) AUC and (b) AUPR measured using the largest eigenvector. The error bars indicate standard error of data from 10-fold cross-validation.

We explored side effect phenomena with ordinary canonical correlation analysis in the context of neighborhood of targets in the genomic space. We find that, among the higher order network properties degree and betweenness are better in capturing relevance to side effects. Computational complexity of degree is $O(n^2)$, whereas the same for betweenness and page rank is far higher, $O(n^3)$. Moreover, computation of page rank involves eigenvector calculations and approximations with large matrices. Due to the prediction performance and lesser computational complexity, degree emerged as the best network feature extracted from the genomic space for modeling secondary information towards prediction of drug side effects.

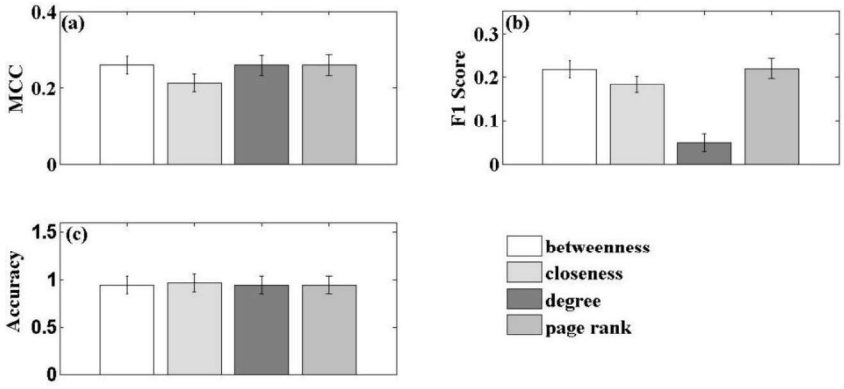

Figure 6.6 : Mathew correlation coefficients (a) F1 score (b) and Accuracy (c) measured using the largest eigenvector. The error bars indicate standard error of data from 10-fold cross validation.

While degree is better when seen in terms of AUC, AUPR, MCC and Accuracy, it is inferior to Page Rank in terms of F1 Score. As a middle ground between computationally intensive parameter of Page Rank and Degree which is inferior in terms of F1 score, we propose that an intermediate parameter that considers not only first order interactions (as in degree) but next level of interactions cold be useful.

…