Improving Drug Discovery Process by Identifying Frequent Toxic Substructures in Chemical Compounds - A Graph Mining Approach

B.Chandra 1, Shalini Bhaskar 1
1 Indian Institute of Technology, Delhi, Hauz Khas, New Delhi-110016

Abstract - Discovery of drug using computer modeling is one of the major challenges in contemporary medicine. Developing new therapeutic drugs is an expensive and time consuming process. Toxicity, caused by substructures that are carcinogenic in nature, is one of the important aspects that need to be explored during drug discovery. The use of in silico methods at an early stage of drug discovery can greatly reduce the time to test the drugs in laboratories. Identifying frequent substructures related to carcinogenic toxicity from a database of chemical compounds can be of great help in reducing the search space during drug discovery process. Graph mining algorithms can be used in order to identify substructures that occur frequently in a given compound database.

Keywords: graph mining algorithms, drug discovery, substructures, carcinogenic

1 Introduction

New drug discovery is a challenging task since a single compound is declared safe for research on human beings among thousands of new compounds. There has been a phenomenal increase in the size of the chemical compound database in the recent years. This has changed the entire course of discovering and developing new drugs. Databases available for chemical compounds are being used as the initial point for screening candidate molecules and hence facilitate the pharmaceutical industry to manufacture more than one hundred thousand new compounds every year [23]. The use of efficient computer modeling techniques can be employed for drug discovery to reduce the search space. Compounds that are identified as promising are further investigated in the process of development wherein apart from the other parameters that include absorption, distribution, metabolism and excretion (ADME); their prospective toxicity is evaluated. This is a complex and costly process that often requires years before the compounds can be tested on human beings [9]. Further, about ninety percent of the initially chosen drug candidates fail to reach the market due to their toxic properties [26]. A compound can be declared toxic if it contains carcinogenic substructures, substructures that can cause mutations, etc. This fact underlines the significance of determining probable toxic effects at an early stage in the process of development. Toxicity tests decide whether a candidate molecule is expected to create toxic effects in human beings. It generally involves using animal models at a pre-clinical phase. There is a greater need for competent in silico techniques to predict the toxicity of chemical compounds with the increase in number of biological targets and the demand for drug screening campaigns. Graph mining algorithms have been efficiently used for finding frequent substructures from database of chemical compounds that can cause toxicity.

Identification of frequent substructures that cause toxic effect in chemical compounds can greatly help the new drug discovery process since such substructures could be avoided in the new drugs. Graph mining facilitates the process of drug discovery by uncovering the chemical and biological characteristics of the drug. Thus, pharmaceutical industry can hope to reduce both the time and cost involved in the drug discovery process.

Graphs can be used for representing chemical compounds with atoms represented by the vertices in the graph and bond between the atoms as edges in the graph. Identification of frequent substructures is one of the predominant tasks performed by graph mining. Various graph mining algorithms have been discussed in the literature. AGM [16], a vertex growth method, follows apriori principle [1] for candidate generation procedure. FFSM [14] uses adjacency matrix for storing each isomorphic graphs. Code for each isomorph is generated by reading adjacency matrix. Subdue [15], an approximate mining algorithm, identifies frequent substructures and at the same time compresses graph dataset. ‘gSpan’ [28] uses DFS tree for representing graph. A graph can have more than one DFS tree with DFS code associated with each DFS tree. Graph mining algorithms have been used in the past for finding frequent substructures that produce toxic effect [6].

In this paper, graph mining has been efficiently used for finding frequent substructures in datasets belonging to
carcinogenic compounds [10][11][12] and dataset of estrogen receptors [2][4][8] taken from DSSTox. A database of substructures identified as carcinogenic and frequently occurring in different chemical compound databases has been prepared. This database can act as a repository to be referred for testing toxicity in terms of carcinogenicity in a new compound to be tested on human subjects.

The paper has been divided into the following sections. Section 2 discusses the gSpan algorithm. Section 3 gives analysis of the results obtained by applying graph mining algorithm on different datasets taken from Distributed Structure-Searchable Toxicity.

2  Details of gSpan algorithm

In order to identify frequent substructures from the graph database, support of the identified candidate substructures is computed. Support of a given substructure is defined as the number of distinct (not isomorphic) substructures in a database of graphs. ‘gSpan’ algorithm [28] starts by generating different DFS (Depth First Search) trees followed by identifying their DFS codes. For generating DFS tree from a given graph, the graph is traversed in depth first search [7] fashion. For a given graph, there can be more than one DFS trees. The concept of forward and backward edge is used by gSpan in DFS tree construction where forward edge set is defined as the set of edges from the graph that forms part of the DFS tree and backward edge set is the set of edges of the graph which are not in the DFS tree. For each DFS tree, DFS code is generated based on the lexicographical ordering defined on the vertex and edge labels. Among the different DFS code representations of a graph, the one with the minimum DFS code generated on the basis of lexicographic order is the canonical representation of the graph. A DFS tree is represented as a sequence of edges, called the DFS code, with each edge in the DFS tree being represented in terms of \( l_i \) and \( l_j \), the labels of vertices \( v_i \) and \( v_j \) respectively; \( e_{i,j} \) is the edge label where \( i \) and \( j \) denote the vertex identification numbers. The structure of a chemical compound is expressed in terms of a graph with nodes of the graph representing the atoms and edges of the graph representing the bonds.

Consider the graph given in Figure 1(a) for which there can be more than one DFS tree that are isomorphic to it and are given in Figure 1(b), 1(c) and 1(d). The lexicographical ordering among the different DFS codes is identified and finally a search tree based on this lexicographic ordering is constructed.

In the first stage, frequent 1-edge substructures (which represent two atoms in a chemical compounds) that satisfy minimum support criteria are identified. Minimum support is the support threshold defined for identifying frequently occurring substructures in the chemical compound database.

If the support of a given substructure is greater than the minimum support, the substructure is identified as frequent. Using the concept of forward and backward extensions the substructures identified frequent at the previous level are extended to generate candidate substructures at the next level (see Fig.2 for chemical compound).
The process of recursively exploring substructures stops when either the support of the substructure is less than minimum support or its code is not a minimum code. ‘gSpan’ algorithm has been applied on the various datasets and relevant frequent substructures have been identified.

3 Results

The results obtained by applying graph mining algorithm on various datasets taken from Distributed Structure-Searchable Toxicity (DSSTox), a Public Database Network from the U.S. Environmental Protection Agency has been analyzed. The database uses a standard chemical structure annotation and is aimed at toxicology studies.

Frequent substructures related to toxicity have been identified using gSpan algorithm. All evaluations were carried out on Intel Pentium® 4 CPU 2.93GHz having 512 MB RAM. Results are compiled using gcc 3.0 compiler on Red Hat Linux 9 (2.4.20-8). Three datasets CPDBAS (Carcinogenic Potency Database Summary Tables - All Species) [10][11][12], NCTRER (FDA’s National Center for Toxicological Research Estrogen Receptor Binding Database File) [2][4][8] and DBPCAN (EPA Water Disinfection By-Products with Carcinogenicity Estimates Database File [27] are used. Brief description of each of the datasets along with the relevant frequent substructures identified is shown in the following subsections.

(a) CPDBAS Dataset

The CPDB Summary, Tables list the summarized results for experiments on 1547 substances in the Carcinogenic Potency Database (CPDB). These Summary Tables report the strongest evidence of carcinogenicity for each chemical, in each sex/species and represent one of many possible summarizations of the data in the CPDB. Since, the substructures present in most of the chemical compounds are required; the minimum support is taken to be of a high value. By applying gSpan algorithm, frequent substructures obtained from this dataset at minimum support of 99.7% belongs to the family of propyl- and methyl- hydrazines, byphenyl and nitroso(proply)amines. These frequent substructures are given in Figure 3. The toxic effect associated with the frequent substructures is discussed as follows:

![Figure 3 Relevant frequent substructures identified from CPDBAS dataset](image.png)
Methylhydrazine is associated with many diseases, mainly cancer [17]. Inflammation of pancreas and infertility in males are some of the diseases related with PAPA-NONOate that contains propylhydrazine as its substructure [25]. Nitroso(propyl)amines and nitrosamines interact with cytochrome CYP2A6 and can cause diseases related to liver such as hepatitis B and C, cirrhosis, liver cancer and other types of cancer [13]. Diseases like tumor of liver and methemoglobinemia are associated with the decreased activity of hepatic cytochrome CYP1A2 enzyme, caused by biphenyl molecule [19]. Many biphenyl derivatives are associated with bioaccumulation [3], endocrine disruption [5] and neurotoxicity [24] because of their symmetry, hydrophobicity and ease of conjunction with halogen atoms. These identified substructures should be avoided in new drug design.

(b) Dataset NCTRER

Database of experimental estrogen receptor (ER) binding results was generated by researchers within FDA’s National Center for Toxicological Research (NCTR) for the purpose of developing improved QSAR model for predicting ER binding affinities. The NCTRER database provides activity classifications for a total of 232 chemical compounds selected apriori based on structural characteristics and tested in a well validated and standardized in vitro rat uterine cytosol ER competitive-binding assay [2] [4][8]. By applying gSpan algorithm frequent substructures obtained from this dataset at minimum support of 0.7% belongs to the family of acyclic alcohols and 4-(propan-2-yl) phenol. Frequent substructures of significance identified from NCTRER dataset are given in Figure 4. The frequent substructures identified from the NCTRER dataset embrace well known hormonal effects, mainly due to their affinity for the estrogen receptors, for example, 17 beta estradiol (an active metabolic product of testosterone) matches with 4-(propan-2-yl) phenol substructure [22], diethylstilbestrol and bisphenols analogues on the other hand matches with acyclic alcohols. If the chemical compounds having these substructures do not show any carcinogenic effect, then, these chemical compounds can be further investigated for other properties and can be helpful in new drug design for removing estrogen deficiency.

(c) Dataset DBPCAN (EPA Water Disinfection By-Products with Carcinogenicity Estimates Database File)

The DBPCAN database consists largely of comparatively simple aliphatic organic structures, almost half halogenated, placed into one of 18 general chemical functional classes. The water disinfection by-products database contains calculated estimates of carcinogenic potential for 209 chemicals. The DBP’s are grouped by a semiquantitative ranking scale of high (H), high-moderate (HM), moderate (M), low-moderate (LM), marginal (mar), and low (L) levels of concern for possible carcinogenicity. By applying gSpan algorithm number of frequent substructures obtained from this dataset at minimum support of 99.7% belongs to the family of halogens and haloalkanes. Frequent substructures identified from DBPCAN dataset are given in Figure 5.

Almost all substructures obtained are associated with some kind of toxic effect. Bromomethane was used for pest control by farmers before it was related to diseases associated with people working in agricultural fields and heart diseases [21]. Hence, the chemical compounds with these substructures should not be used in drugs. A drug with the name Remoxipride was withdrawn from the market due to the presence of bromomethane substructure that leads to aplastic anemia [18]. Presence of bromine is not desired by therapeutic chemists of good repute [20].

If a repository of frequent substructures identified in different toxic compound databases can be made and preserved, it will be of great help in new drug design. When a new drug is identified, it is checked for the existence of these toxic substructures in it. If any of these structures exists in the chemical compound, it is not explored further on other aspects as it is not suitable for human beings.
4 Conclusions

The paper presents an application of graph mining algorithms for identifying frequent substructures from the database of chemical compounds. The analysis of the results on different datasets taken from Distributed Structure-Searchable Toxicity (DSSTox) has been done and presented. The scope of graph mining algorithms in identifying toxicity from chemical compounds has been investigated with the objective of reducing the search space being explored in drug discovery process.

5 References


