Indexing and mining topological patterns for drug discovery

Sayan Ranu
Dept. of Computer Science
University of California, Santa Barbara
CA 93106, USA
sayan@cs.ucsb.edu

Ambuj K. Singh
Dept. of Computer Science
University of California, Santa Barbara
CA 93106, USA
ambuj@cs.ucsb.edu

ABSTRACT

Increased availability of large repositories of chemical compounds has created new challenges and opportunities for the application of data-mining and indexing techniques to problems in chemical informatics. The primary goal in analysis of molecular databases is to identify structural patterns that can predict biological activity. Two of the most popular approaches to representing molecular topologies are graphs and 3D geometries. As a result, the problem of indexing and mining structural patterns map to indexing and mining patterns from graph and 3D geometric databases.

In this tutorial, we will first introduce the problem of drug discovery and how computer science plays a critical role in that process. We will then proceed by introducing the problem of performing subgraph and similarity searches on large graph databases. Due to the NP-hardness of the problems, a number of heuristics have been designed in recent years and the tutorial will present an overview of those techniques. Next, we will introduce the problem of mining frequent subgraph patterns along with some of their limitations that ignited the interest in the problem of mining statistically significant subgraph patterns. After presenting an in-depth survey of the techniques on mining significant subgraph patterns, the tutorial will proceed towards the problem of analyzing 3D geometric structures of molecules. Finally, we will conclude by presenting two open computer science problems that can have a significant impact in the field of drug discovery.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process; I.2.8 [**Problem Solving, Control Methods, and Search**]: Graph and tree search strategies

General Terms

Graph Indexing, Graph Mining, Algorithms

Keywords

top-k queries, graph databases, frequent subgraphs, significant subgraphs, significant geometric patterns

EDBT 2012, March 26–30, 2012, Berlin, Germany.

1. INTRODUCTION

With the recent advent of high-throughput technologies for both compound synthesis and biological screening, there is no shortage of publicly or commercially available data sets that can be used for computational drug discovery applications. Recent estimates of the number of known small molecules, encountered so far in nature, or synthesized by man, is on the order of 10^7. Furthermore, the size of the virtual space of molecules that could be created is reported to be more than 10^{60}. The goal in drug discovery is to analyze this huge chemical space and identify molecules that show a certain desired activity. However, given the magnitude of the chemical space, an exhaustive exploration is not feasible. As a result, the key challenge for computational methods is therefore not to explore the entire chemical space, but rather to be able to identify and analyze interesting regions within this space.

In this tutorial, we will present a survey of the computer science techniques that allow us to better understand the chemical space. Typical querying and mining tasks involve clustering of large molecular libraries, developing index structures for fast answering of top-k queries, mining structural patterns, and predicting biological activity of molecules. Among the various approaches to represent molecular topology in the virtual space, two of the most prominent approaches are graphs and 3D geometry of molecules. In this tutorial, we will survey the state of the art in querying and mining techniques that are based on analyzing graph and geometry based representations of molecules.

The tutorial will first introduce the challenges in the drug discovery process and how computer science plays a critical role in that pipeline. Next, the tutorial will outline the graph and geometry based approaches to represent molecules in the virtual space. After establishing the basic groundwork, the tutorial will present the graph based indexing techniques on performing subgraph containment [2, 4, 8, 18, 19, 21, 24, 28–30] and subgraph similarity searches [4, 17, 25–27]. Next, the tutorial will proceed towards the problem of mining subgraph patterns and their applications in molecular classification. Two graph mining techniques will be presented: frequent subgraph mining [1, 6, 7, 10, 11, 20, 23] and significant subgraph mining [3, 5, 9, 14, 15, 22]. The tutorial will highlight one key weakness in frequent subgraph patterns that resulted in much enthusiasm towards solving the problem of mining statistically significant subgraph patterns. After presenting an in-depth survey on significant subgraph mining with special focus on the techniques *Leap* [22] and *GraphSig* [14], the tutorial will demonstrate the application of significant subgraph patterns on molecular classification. The tutorial will conclude the graph section of the tutorial by introducing the problem of mining probabilistically labeled graph databases [13]. Next, the tutorial will proceed towards presenting the problem of mining geometric patterns. The tutorial will first

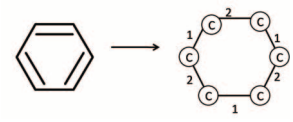

Figure 1: Graph representation of benzene

illustrate the unique challenges faced while dealing with geometric structures and then present two recent techniques on mining frequent [12] and statistically significant geometric patterns [16]. Finally, the tutorial will conclude by discussing two open problems that lie at the intersection of computer science and drug-discovery.

The remainder of the tutorial proposal is organized as follows. Section 2 discusses the material to be covered in the tutorial to provide a sense of both its scope and the depth to which the material will be covered. In Section 3, we discuss logistical issues such as length of the tutorial, difference to previous editions of the tutorial, and intended audience. Finally, we conclude by providing a brief biography of the authors in Section 4.

2. TUTORIAL MATERIAL

In this section, we discuss the material to be covered in the tutorial. The tutorial can be divided primarily into three topics: graph indexing, graph mining, and geometry mining. The following subsections elaborate on each of these individual topics.

2.1 Graph Indexing

A common approach to model 2D structural properties of molecules is in the form of graphs. In the graph representation of chemical compounds, the nodes represent atoms and the edges represent covalent bonds between them. An example is shown in Fig. 1. As a result, top-k queries on molecular substructure containment and molecular similarity map to the problems of subgraph queries, and graph similarity queries. Both problems are known to be NP-hard [4]. As a result, efficient heuristics are required to compute the answer set in a scalable manner.

2.1.1 Subgraph Queries

Substructure queries are one of the most popular and simplest techniques to predict biological activity of molecules. Often, chemists have prior knowledge about the biological activity of certain substructures. For example, the activities of functional groups are well documented. Given this knowledge, chemists are often interested in finding molecules containing a substructure demonstrating a desired activity. Under the graph-based representation of molecules, the problem translates to subgraph containment queries. Due to the NP-hardness of the problem, over the past years, a number of techniques have been developed to efficiently process subgraph containment queries. Interestingly, majority of the techniques fall under the general framework of fragment-based indexing [2, 8, 18, 19, 24, 28, 29]. In this tutorial, we will present two approaches: *fragment-based indexing* and *graph summarization based indexing*.

Taking gIndex [24] as the representative example, we will illustrate the fragment-based indexing approach. The major components of this approach are identifying a set of discriminative fragments, constructing an inverted index on the fragments, and a filtering step to construct a candidate set on which subgraph isomorphisms are performed. Existing algorithms under the fragment-based indexing scheme have mostly focused on refining the quality of the discriminative fragments. Thus, the tutorial will emphasize the importance of mining discriminative fragments by analyzing the dependence of the computation cost of subgraph containment with the quality of the discriminative fragments. Finally, the tutorial will illustrate gIndex's method of quantifying the discriminative potential of a fragment to complete the presentation on fragment-based indexing.

Ctree [4] introduced the graph summarization based approach for subgraph indexing which is significantly different from the fragment-based indexing technique. Inspired from the bounding box approach in R-tree, Ctree recursively constructs summarizations of the graph database, called *closures*, in a bottom-up manner and constructs a *closure-tree*. The closures ensure the property that if a query subgraph is not contained in a closure, then the set of graphs under the closure are guaranteed to not contain the query as well. As a result, the number of subgraph containment verifications can be significantly reduced by starting the search operation from the root of the closure-tree, and discarding all children of a closure that do not contain the query subgraph.

2.1.2 Similarity Queries

Similarity queries, as in various other domains, find ubiquitous application in molecular databases. In drug discovery, it is common to assume that molecules with similar structures have similar biological activity. Based on this assumption, chemists are often interested in screening a molecular database to find those molecules that are structurally similar to a known biologically active molecule. In the graph setting, this translates to the problem of graph similarity searches. A number of indexing techniques exist for graph similarity queries [4, 17, 25–27]. However, not all of them use the same distance measure. In this tutorial, we will present the classical *graph edit distance*, and explain how the Ctree index can be utilized for indexing similarity queries as well.

2.2 Graph Mining

Mining substructural patterns from molecular databases is key to understanding the biological activity of molecules. Typically, given a database of molecules where each molecule is tagged as either 'active' or 'inactive', the goal is to mine substructures that can predict the molecular tags. Towards that goal, based on the graph-based representation, the problem of frequent subgraph mining was formulated. Later, the problem of mining statistically significant subgraphs was formulated which allows embedding of more information in the mining process and consequently increase the quality of the mined subgraphs.

2.2.1 Mining Frequent Subgraphs

Given a database of active molecules, the goal behind mining frequent subgraphs is to identify the frequently occurring substructures under the assumption that such substructures are likely to correlate with biological activity. A number of techniques have been designed over the past few years on frequent subgraph mining. Interestingly, the techniques can be grouped into two major design frameworks: the *pattern-growth approach* [1, 11, 23], and the *join-based approach* [6, 7, 10, 20]. Taking gSpan [23] and FSG [10] as the representative examples of the two approaches, we will illustrate their design principles.

2.2.2 Mining Statistically Significant Subgraphs

While frequent subgraphs are effective in identifying substructural properties, they may not provide the best characterization of the datasets. More specifically, the frequent subgraph mining framework only considers the information embedded in the active molecules and completely ignores the inactive set. Consequently, a frequent substructure in the active molecules could also be frequent in the in-

active molecules. Such substructures do not show any correlation with biological activity. A better formulation to mine substructure patterns would therefore be to identify only those substructures whose frequencies are significantly higher in the active set of molecules than in the inactive set. Recently, four techniques have been developed to tackle this problem by mining statistically significant subgraphs. In this tutorial, we will illustrate two of them: *Leap* [22], and *GraphSig* [14].

For efficient pruning of the search space, Leap introduced three heuristics: vertical pruning, horizontal pruning, and frequency descending mining. GraphSig, on the other hand, employs the strategy of converting graphs into feature vectors as a result of which the problem of mining significant subgraphs translates to mining significant feature vectors. After mining significant feature vectors using a technique originally developed in *GraphRank* [5], GraphSig develops a technique to map the significant feature vectors back into the graph space to compute the final answer set of significant subgraphs.

In this tutorial, we will also highlight the application of mining statistically significant subgraphs in molecular classification [15]. Molecular classification is an important problem in drug development where libraries of chemical compounds are screened and molecules with the highest probability of success against a given target are selected. For this purpose, molecules can be characterized using feature vectors that indicate the presence or absence of statistically significant substructures. Once converted to feature vectors, molecules can be classified using any of the existing classifiers.

Finally, the section on graph mining will be concluded by presenting the problem of mining significant subgraphs from probabilistically labeled graph databases. The problem is important since deterministic labeling of molecular activity is expensive, both monetarily and temporally. At the cost of accuracy, molecular activity can be estimated more scalably using any of the existing classifiers. Consequently, an important question arises: *In the absence of sufficient amount of high-quality data, can our knowledge base be expanded by analyzing high-volume, but noisier, datasets?* A recent work [13] investigating this issue reveals that incorporating noisier data, when managed appropriately within a probabilistic framework, enhances the quality of the answer sets.

2.3 Mining Geometric Patterns

Since graphs represent 2D structures of molecules, information on the third dimension is lost in this representation form. The entire information can be retained if molecules are characterized by the 3D geometry of their structures. In this tutorial, we also present techniques that focus on analyzing the geometry of molecules in the 3D space. The underlying geometry of the pharmacophores is responsible for binding between compounds and targets as well as properties of compounds such as blood brain barrier permeability. As a result, the questions asked in the graph setting can be asked in the 3D setting as well. More specifically, we will look at the problems of mining *frequent geometric patterns*, and mining *statistically significant geometric patterns*.

The problem of mining frequent geometric patterns was first formulated by Podolyan et al. [12]. In their approach, the authors model the geometry of molecules as cliques. As a result, the problem of mining frequent geometries translates to the problem of mining frequent cliques and the existing frequent subgraph mining techniques can be employed to identify the frequent geometric patterns. To solve the problem of mining statistically significant geometric patterns, Ranu et al. [16] employed a triangle based characterization of molecules. In their approach, all triangles from a molecular database are first extracted and then clustered. Next, based on a background model derived from the active and inactive molecules in the given database, each of the clusters are analyzed to identify the statistically significant ones. Finally, the centers of the significant clusters are presented as the answer set containing statistically significant geometric patterns.

2.4 Open Problems

After discussing the graph and geometry-based indexing and mining techniques, the tutorial will present two open problems in computer science that can also make a significant impact in the field of drug-discovery.

2.4.1 Diversity-aware top-k queries on graph databases

In molecular databases, it is common to tag molecules with feature vectors representing their inhibition values against various targets as well properties such as toxicity, blood brain barrier permeability etc. Given this setting, chemists are often interested in identifying molecules that maximize a scoring function where the function quantifies a certain desirable property. Therefore, the problem can be formulated as a standard top-k query. However, the standard formulation ignores an important aspect of the answer set: *diversity*.

In traditional scoring functions, the score of a data object is independent of the data objects that have already been included in the answer set. Such a formulation risks increasing the information redundancy in the answer sets. More specifically, if multiple molecules in the answer set are structurally similar to each other, then the information content embedded in the answer set is highly diminished. It is therefore desirable to compute an answer set that is both high scoring and structurally *diverse*. While techniques have been developed for diversity-aware top-k queries, most of them are applicable to text documents. More importantly, no technique exists for diverse-aware top-k queries on graph databases.

2.4.2 Scaffold Hopping

The problem of scaffold hopping assumes the same database setting as described in diversity-aware top-k queries. However, instead of the query being a function, in this problem, the query is a molecule and its corresponding feature vector. The goal is to identify molecules which are as similar as possible in their inhibition values and structurally as dissimilar as possible to the query. Among various applications, the primary utility of scaffold hopping lies in identifying a diverse set of molecules that can be used as seeds to synthesize more molecules with the desired activity.

3. LOGISTICAL DETAILS

In this section, we discuss the logistical aspects of the tutorial.

3.1 Length

The tutorial is not expected to exceed a time duration of one hour and thirty minutes.

3.2 Earlier Presentations

A tutorial covering similar material was presented at the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB), 2011. The audience of ACM BCB was familiar to the problem of drug discovery. Thus, less time was spent motivating the problem. Furthermore, due to the relative unfamiliarity of the ACM BCB audience with the problems of graph mining and indexing, the tutorial only presented high-level details of the material. The current version of the tutorial will spend more time illustrating why computer science is an important part of the drug-

discovery pipeline. Once the motivation behind computation analysis of molecular databases is well established, the tutorial will cover the technical content at an higher depth than in ACM BCB since the EDBT audience is likely to be more exposed to graph mining and indexing techniques. Finally, the tutorial will have added material on two open problems: diversity-aware top-k queries on graph databases and scaffold hopping.

3.3 Intended Audience and Prerequisite Knowledge

The tutorial makes the assumption that the audience is familiar with basic computer science concepts such as graphs, subgraph queries, similarity searches and tree traversal techniques such as breadth-first and depth-first searches. The tutorial does not make any assumption on prior knowledge of bioinformatics, chemoinformatics, or drug discovery. To summarize, the intended audience for the tutorial are computer scientists who have an interest in graph indexing, graph mining, and mining geometric patterns.

The tutorial is of interest to the EDBT audience since it presents recent computer science techniques from top publication venues and highlights how they play an important role in the field of drug discovery. The interdisciplinary nature of the tutorial will help the audience in defining problems with an higher impact on the disciplines of both computer science and chemistry.

4. AUTHOR BIOGRAPHY

In this section, we provide a brief biography of the authors.

- **Sayan Ranu:** Sayan Ranu is a Ph.D. student at the computer science department of University of California, Santa Barbara. His research work is centered on querying and mining molecular databases.
- **Ambuj K. Singh:** Ambuj K. Singh is a professor of computer science at the University of California, Santa Barbara. His research interests are in the areas of data mining, databases, chemoinformatics, and bioinformatics. He received his PhD from the University of Texas at Austin in 1989.

5. REFERENCES

[1] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. *ICDM*, 2002.

[2] J. Cheng, Y. Ke, W. Ng, and A. Lu. Fg-index: towards verification-free query processing on graph databases. In *SIGMOD*, 2007.

[3] M. A. Hasan and M. J. Zaki. Output space sampling for graph patterns. *PVLDB*, 2(1), 2009.

[4] H. He and A. K. Singh. Closure-tree: An index structure for graph queries. In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE, 2006.

[5] H. He and A. K. Singh. GraphRank: Statistical Modeling and Mining of Significant Subgraphs in the Feature Space. In *ICDM*, 2006.

[6] J. Huan, W. Wang, and J. Prins. Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism. In *ICDM*, 2003.

[7] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery*, volume 1910, pages 13–23. 2000.

[8] H. Jiang, H. Wang, P. S. Yu, and S. Zhou. Gstring: A novel approach for efficient search in graph databases. In *ICDE*, 2007.

[9] N. Jin, C. Young, and W. W. 0010. Gaia: graph classification using evolutionary computation. In *SIGMOD*, 2010.

[10] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *ICDM*, 2001.

[11] S. Nijssen and J. N. Kok. The Gaston tool for Frequent Subgraph Mining. In *Proceedings of the International Workshop on Graph-Based Tools*, 2004.

[12] Y. Podolyan and G. Karypis. Common pharmacophore identification using frequent clique detection algorithm. *Journal of Chemical Information and Modeling*, 49(1):13–21, 2009.

[13] S. Ranu, B. T. Calhoun, A. K. Singh, and S. J. Swamidass. Probabilistic substructure mining from small-molecule screens. *Molecular Informatics*, 30(9):809–815, 2011.

[14] S. Ranu and A. K. Singh. Graphsig: A scalable approach to mining significant subgraphs in large graph databases. In *ICDE*, 2009.

[15] S. Ranu and A. K. Singh. Mining statistically significant molecular substructures for efficient molecular classification. *J. Chem. Inf. Model.*, 49:2537–2550, 2009.

[16] S. Ranu and A. K. Singh. Novel method for pharmacophore analysis by examining the joint pharmacophore space. *Journal of Chemical Information and Modeling*, 51(5):1106–1121, 2011.

[17] H. Shang, X. Lin, Y. Zhang, J. X. Yu, and W. W. 0011. Connected substructure similarity search. In *SIGMOD*, 2010.

[18] H. Shang, Y. Zhang, X. Lin, and J. X. Yu. Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. *VLDB*, 2008.

[19] D. Shasha, J. T.-L. Wang, and R. Giugno. Algorithmics and applications of tree and graph searching. In *PODS*, 2002.

[20] N. Vanetik and E. Gudes. Mining Frequent Labeled and Partially Labeled Graph Patterns. In *ICDE*, 2004.

[21] D. W. Williams, J. Huan, and W. Wang. Graph database indexing using structured graph decomposition. In *ICDE*, 2007.

[22] X. Yan, H. Cheng, J. Han, and P. S. Yu. Mining Significant Graph Patterns by Scalable Leap Search. In *SIGMOD*, 2008.

[23] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, 2002.

[24] X. Yan, P. S. Yu, and J. Han. Graph indexing: a frequent structure-based approach. In *SIGMOD*, 2004.

[25] X. Yan, P. S. Yu, and J. Han. Substructure similarity search in graph databases. In *SIGMOD*, 2005.

[26] X. Yan, F. Zhu, J. Han, and P. S. Yu. Searching substructures with superimposed distance. In *ICDE*, 2006.

[27] Z. Zeng, A. K. H. Tung, J. Wang, J. Feng, and L. Zhou. Comparing stars: On approximating graph edit distance. *PVLDB*, 2(1), 2009.

[28] S. Zhang, M. Hu, and J. Yang. Treepi: A novel graph indexing method. In *ICDE*, 2007.

[29] P. Zhao, J. X. Yu, and P. S. Yu. Graph indexing: Tree + delta >= graph. In *VLDB*, 2007.

[30] L. Zou, L. C. 0002, J. X. Yu, and Y. Lu. A novel spectral coding in a large graph database. In *EDBT*, 2008.