Chapter 19

TRENDS IN CHEMICAL GRAPH DATA MINING

Nikil Wale
Computer Science & Engineering
University of Minnesota, Twin Cities, US
nwale@cs.umn.edu

Xia Ning
Computer Science & Engineering
University of Minnesota, Twin Cities, US
xning@cs.umn.edu

George Karypis
Computer Science & Engineering
University of Minnesota, Twin Cities, US
karypis@cs.umn.edu

Abstract

Mining chemical compounds *in silico* has drawn increasing attention from both academia and pharmaceutical industry due to its effectiveness in aiding the drug discovery process. Since graphs are the natural representation for chemical compounds, most of the mining algorithms focus on mining chemical graphs. Chemical graph mining approaches have many applications in the drug discovery process that include structure-activity-relationship (SAR) model construction and bioactivity classification, similar compound search and retrieval from chemical compound database, target identification from phenotypic assays, *etc*. Solving such problems *in silico* through studying and mining chemical graphs can provide novel perspective to medicinal chemists, biologist and toxicologist. Moreover, since the large scale chemical graph mining is usually employed at the early stages of drug discovery, it has the potential to speed up the entire drug discovery process. In this chapter, we discuss various problems and algorithms related to mining chemical graphs and describe some of the state-of-the-art chemical graph mining methodologies and their applications.

C.C. Aggarwal and H. Wang (eds.), *Managing and Mining Graph Data*,
Advances in Database Systems 40, DOI 10.1007/978-1-4419-6045-0_19,

Keywords: Chemical Graph, Descriptor Spaces, Classification, Ranked Retrieval, Scaffold Hopping, Target Fishing.

1. Introduction

Labeled graphs (either topological or geometric) have been a promising abstraction to capture the characteristics of datasets arising in many fields such as the world wide web, social networks, biology, and chemistry ([9], [13], [30], [49]). The vertices of these graphs correspond to the entities in the objects and the edges correspond to the relations between them. This graph-based representation can directly capture many of the sequential, topological, geometric, and other relational characteristics of such datasets. For example, in the domain of the world wide web and social networks the entire set of objects and their relations are represented via a single large graph ([13]). In biology, objects to be mined are represented either as a single large graph (e.g., metabolic and signaling pathways) or via separate graphs (e.g., protein structures) ([65], [30], [33]). In chemistry, each object to be mined is represented via a separate graph (e.g., molecular graphs) ([49]).

Graph mining over the above representations has found applications in the domain of web data analysis such as the analysis of XML documents and weblogs, web searches, web document analysis *etc*([9]). Graph mining is also being used in social sciences for the analysis of social networks that help understand social phenomenon and group behavior([13]). In the domain of traditional sciences like biology and chemistry, graph mining has found numerous important applications. For example, in biology graphs can be used to directly model the key topological and geometric characteristics of protein molecules. Vertices in these graphs will correspond to different amino acids. The edges will correspond to the connections of amino acids in the protein's backbone or the non-covalent bonds(i.e., contact points) in the 3D structure. Mining these graph patterns provides important insights into protein structure and function ([22], [3]).

In chemistry, graphs can be used to directly model the key topological and geometric characteristics of chemical structures. Vertices in these graphs correspond to different atoms and the edges correspond to bonds that connect atoms ([29]). Mining on a set of chemical compounds or molecules helps in understanding the key characteristics of a set molecules for a given process (such as toxicity and biological activity) and has become the primary application area of chemical graph mining ([49], [40]). The typical applications performed on chemical structures include mining sub-structures in a given set of ligands ([40]), mining databases to retrieve other relevant compounds, clustering of chemical compounds based on common sub-structures, and predicting compound bioactivity by classification, regression and ranking techniques ([2], [28]).

Most of the mining algorithms operate on the assumption that the properties and biological activity of a chemical compound are related to its structure ([2], [28]). This assumption is widely referred to as the structure-activity-relationship principle or simply SAR. Hansch ([17]) demonstrated that the biological activity of a chemical compound can be mathematically expressed as a function of its physiochemical properties, which led to the development of quantitative methods for modeling structure-activity relationships (QSAR). Since that work, many different approaches have been developed for building such structure-activity-relationship (SAR) models. All of these models are derived using some notion of structural similarity between chemical compounds. The similarity is determined using a similarity function over a descriptor-space representation, and the descriptor-space is most commonly generated from chemical graphs. These models have become an essential tool for predicting biological activity from the structural properties of a molecule.

The rest of this chapter will review some of the current trends in chemical graph mining and modeling. It will highlight some of the techniques that exist and that were recently developed for representing chemical compounds, building classification models, retrieving compounds from databases, and identifying the proteins that the compounds will bind to. The chapter concludes by outlining some of the future research directions in this field.

2. Topological Descriptors for Chemical Compounds

Descriptor-based representations of chemical compounds are used extensively in cheminformatics, as they represent a convenient and computationally efficient way to capture key characteristics of the compounds' structures ([2], [28]). Such representations have extensive applications to similarity search and various structure-driven prediction problems for activity, toxicity, absorption, distribution, metabolism and excretion ([2]). Many of these descriptors are derived by mining structural patterns from a set of molecular graphs of the chemical compounds. Such descriptors include topological descriptors derived directly from the topology of molecular graphs and 2D/3D pharmacophore descriptors that describe the critical atoms/atom groups that are highly likely to be involved in protein-ligand binding ([7], [32], [55], [28]). In the rest of this section we review some of the topological descriptors that are used extensively to represent chemical compounds and analyze their different properties. This includes both a set of time-tested descriptors as well as recently developed descriptors that have shown promising results.

2.1 Hashed Fingerprints (FP)

Hash fingerprints are generally used to encode the 2D structural characteristics of a chemical compound into a fixed bit vector and are used extensively for various tasks in chemical informatics. These fingerprints are typically generated by enumerating all cycles and linear paths up to a given number of bonds and hashing each of these cycles and paths into a fixed bit-string ([7], [4], [51], [20]). The specific bit-string that is generated depends on the number of bonds, the number of bits that are set, the hashing function, and the length of the bit-string. The key property of these fingerprint descriptors is that they encode a very large number of sub-structures into a compact representation. Many variants of these fingerprints exist, some use predefined structural fragments in conjunction with the fingerprints, for example, Unity fingerprints ([51]), others count the number of times a bit position is set, for example, hologram ([20]). However, a recent study has shown that the performance of most of these fingerprints is comparable ([26]).

2.2 Maccs Keys (MK)

Molecular Design Limited (MDL) has created the key based fingerprints Maccs Keys ([32]) based on pattern matching of a chemical compound structure to a pre-defined set of structural fragments. These fragments have been identified by domain experts ([10]) to be important for bioactivity of chemical compounds. The original set of descriptors consists of 166 structural fragments and each such fragment becomes a key and occupies a fixed position in the descriptor space. This approach relies on pre-defined rules to encapsulate the essential molecular descriptors a-priori and does not learn them from the chemical dataset. This descriptor space is notably different from fingerprint based descriptor space. Unlike fingerprints, no *folding* (hashing) is performed on the sub-structures.

2.3 Extended Connectivity Fingerprints (ECFP)

Molecular descriptors and fingerprints based on the extended connectivity concept have been described by several authors ([42], [19]). The earliest concept of such a descriptor-space was described in [59]. Recently, these fingerprints have been popularized by their implementation within Pipeline Pilot ([11]). These fingerprints are generated by first assigning some initial label to each atom and then applying a Morgan type algorithm ([34]) to generate the fingerprints. Morgan's algorithm consists of l iterations. In each iteration, a new label is generated and assigned to each atom by combining the current labels of the neighboring atoms (i.e, connected via a bond). The union of the labels assigned to all the atoms over all the l iterations are used as the descriptors to represent each compound. The key idea behind this descriptor generation algorithm is to capture the topology around each atom in the form of shells whose radius ranges from 1 to l. Thus, these descriptors can capture rather complex topologies. The value for l is a user supplied parameter and typically ranges from two to six.

2.4 Frequent Subgraphs (FS)

A number of methods have been proposed in recent years to mine frequently occurring subgraphs (sub-structures) in a chemical graph database ([37], [61], [27]). Frequent subgraphs of a chemical graph database D are defined as all subgraphs that are present in at least σ ($\sigma \leq |D|$) of compounds of the database, where σ is the absolute minimum frequency requirement (also called absolute minimum support constraint). These frequent subgraphs can be used as descriptors for the compounds in that database. A descriptor space formed out of frequently occurring subgraphs depends on the value of σ. Therefore, the descriptor space can change for a particular problem instance if the value of σ is changed. An advantage of such a descriptor space is that it can create descriptors suitable for a given dataset. Moreover, the substructures mined consist of arbitrary sizes and topologies. A potential disadvantage of this method is that it is unclear how to select a suitable value of σ for a given problem. A very high value will fail to discover important subgraphs whereas a very low value will result in combinatorial explosion of frequent subgraphs.

2.5 Bounded-Size Graph Fragments (GF)

Recently, a new descriptor space, Graph Fragments (GF), has been developed consisting of sub-structures or fragments that exist in a compound library ([55]). Graph Fragments of a chemical graph database D are defined as all connected subgraphs present in every chemical graph of D that has a size of less than or equal to the user supplied parameter l. Therefore, GF descriptor space is a subset of the FS descriptor space generated using a absolute minimum support threshold of 1. However, instead of the minimum support threshold used in generating FS, the user supplied parameter l is used to control the combinatorial complexity of the fragment generation process for GF and put an upper bound on the size of fragments generated. An efficient algorithm to generate the GF descriptors for a library of compounds is described in [55].

2.6 Comparison of Descriptors

A careful analysis of the descriptor spaces described in the previous section illustrate four dimensions along which these schemes compare with each other and represent some of the choices that have been explored in designing fragment-based or fragment-derived descriptors for chemical compounds. Ta-

Table 19.1. Design choices made by the descriptor spaces.

	Previously developed descriptors			
	Generation	Topological Complexity	Precise	Complete Coverage
FP	dynamic	Low	No	Yes
MK	static	Low to High	Yes	Maybe
ECFP	dynamic	Low to High	Maybe	Yes
FS	dynamic	Low to High	Yes	Maybe
GF	dynamic	Low to High	Yes	Yes

FP refers to the hashed fingerprints, MK to Maccs keys, ECFP to extended connectivity fingerprints, FS to frequent subgraphs, and GF to graph fragments.

ble 19.1 summarizes the characteristics of these descriptor spaces along the four dimensions. The first dimension is associated with whether the fragments are determined directly from the dataset at hand or they have been pre-identified by domain experts. The fragments of Maccs keys have been determined a priori whereas all other descriptors are determined directly from the dataset. The advantage of a priori approach is that it can capture domain knowledge. However, due to the fixed set of fragments identified a priori it might not adapt to the characteristics for a particular dataset. The second dimension is associated with the topological complexity of the actual fragments. Schemes like fingerprints use simple topologies consisting of paths and cycles. Descriptors such as extended connectivity fingerprints, frequent subgraphs and graph fragments allow topologies with arbitrary complexity. Topologically complex fragments along with simple ones might enrich the descriptor space. The third dimension is associated with whether or not the fragments are being precisely represented in the descriptor space. Most schemes generate descriptors that are precise in the sense that there is a one-to-one mapping between the fragments and the dimensions of the descriptor space. In contrast, due to the hashing approach, descriptors such as fingerprints and extended connectivity fingerprints lead to imprecise representations (i.e., many fragments can map to the same dimension of the descriptor space). Depending on the number of these many-to-one mappings, these descriptors can lead to representations with varying degree of information loss. Finally, the fourth dimension is associated with the ability of the descriptor space to cover all or nearly all of the dataset. Descriptor spaces created from fingerprints, extended connectivity fingerprints, and graph fragments are guaranteed to contain fragments or hashed fragments from each one of the compounds. On the other hand, descriptor spaces corresponding to Maccs keys and frequent sub-structures may lead to a descriptor-based representation of the dataset in which some of the compounds have no or a very small number of descriptors. A descriptor space that covers all the compounds

Table 19.2. SAR performance of different descriptors.

Datasets	fp	ECFP	MK	FS	GF
NCI1	0.30	0.32	0.29	0.27	**0.33**
NCI109	0.27	**0.32**	0.24	0.26	**0.32**
NCI123	0.25	**0.27**	0.24	0.23	**0.27**
NCI145	0.30	0.35	0.28	0.30	**0.37**
NCI167	0.06	0.06	0.04	0.06	**0.07**
NCI220	**0.33**	0.28	0.26	0.21	0.29
NCI33	0.26	0.31	0.26	0.25	**0.33**
NCI330	0.34	**0.36**	0.31	0.24	**0.36**
NCI41	0.25	**0.36**	0.28	0.30	**0.36**
NCI47	0.26	**0.31**	0.26	0.24	**0.31**
NCI81	0.27	**0.28**	0.25	0.24	**0.28**
NCI83	0.26	**0.31**	0.26	0.25	**0.31**

The numbers correspond to the ROC_{50} values of SVM-based SAR models for twelve screening assays obtained from NCI. The ROC_{50} value is the area under the receiver operating characteristic curve (ROC) up to the first 50 false positives. These values were computed using a 5-fold cross-validation approach. The descriptors being evaluated are: graph fragments (GF) ([55]), extended connectivity fingerprints (ECFP) ([28]), Chemaxon's fingerprints (fp) (Chemaxon Inc.) ([4]), Maccs keys (MK) (MDL Information Systems Inc.) ([32]), and frequent subgraphs (FS) ([8]).

of a dataset has the advantage of encoding some amount of information for every compound.

The qualitative comparison of the descriptors along the lines discussed above is shown in Table 19.1. This table shows that unlike other descriptors, GF descriptors satisfy all the key properties described earlier such as dynamic generation, complex topology, precise representation, and complete coverage. For example, unlike path-based structural descriptors (fp) and extended-connectivity fingerprints, they are guaranteed to have a one-to-one mapping between a fragment and a dimension in the descriptor space. Moreover, unlike fingerprints, they impose no limit on the complexity of the descriptor's structures ([55]) and unlike Maccs Keys, the descriptors are dynamically generated from the dataset at hand. Lastly, unlike FS, which may suffer from partial coverage, this descriptor space is ensured to have 100% coverage by eliminating the minimum support criterion and generating all fragments. Therefore, GF descriptors allow for better representation of the underlying compounds and they are expected to show better performance in the context of SAR based classification and retrieval approaches.

A quantitative comparison in Table 19.2 shows classification results from a recent study ([55]) using the NCI datasets obtained from the PubChem Project ([39]). These results empirically show that the GF descriptor space achieves a performance that is either better or comparable to that achieved by currently used descriptors, indicating that the above mentioned properties are important to capture the compounds' structural characteristics.

3. Classification Algorithms for Chemical Compounds

Numerous approaches have been developed for building classifying models for various classes of interest (e.g., active/inactive, toxic/non-toxic, *etc*). Depending on the class of interest, these models are often called structure-activity-relationship (SAR) or structure-property-relationship (SPR) models. Over the years, these approaches have evolved from the initial regression-based techniques used by Hansch ([17]), to methods that utilize complex statistical model estimation procedures ([24], [28], [42], [2]). Among them, methods based on Support Vector Machines (SVM) ([52]) have recently become very popular as they have been shown to produce highly accurate SAR and SPR models for a wide-range of problems ([14], [57], [25], [24], [55], [15]). Two broad classes of SVM-based methods have been developed. The first operate on the descriptor-space representation of the chemical compounds, whereas the second use various graph kernels that operate directly on the compounds' molecular graphs. However, despite their differences, the absolute performance achieved by these methods is often comparable, and no winning methodology has emerged.

3.1 Approaches based on Descriptors

The descriptor-space based approaches first represent each chemical compound as a high-dimensional (frequency) vector based on the set of descriptors that they contain (e.g., hashed fingerprints, graph fragments, etc) and then utilize various vector-space-based kernel functions to determine the similarity between the various compounds ([8], [49], [55], [57], [14]). Such functions include linear, radial basis function, Tanimoto coefficient, and Min-Max kernel ([49], [55]). The performance of these kernels has been extensively evaluated with each other and the results have showed that the Tanimoto coefficient (also known as the extended Jacquard similarity) and the Min-Max kernels are often among the best performing schemes ([49], [55]). The Tanimoto coefficient is defined as

$$\mathcal{K}_{TC}(X,Y) = \frac{\sum\limits_{i=1}^{M} x_i y_i}{\sum\limits_{i=1}^{M} (x_i^2 + y_i^2 - x_i y_i)}, \qquad (3.1)$$

and the Min-Max kernel is defined as

$$\mathcal{K}_{MM}(X,Y) = \frac{\sum\limits_{i=1}^{M} min(x_i, y_i)}{\sum\limits_{i=1}^{M} max(x_i, y_i)}, \tag{3.2}$$

where the terms x_i and y_i are the values along the i^{th} dimension of the M dimensional X and Y vectors, respectively.

A number of variations of these descriptor-based approaches have also been developed. One of them, which is applicable when the descriptor spaces contain a very large number of dimensions, involves the use of various feature selection techniques to reduce the effective dimensionality of the descriptor space by retaining only those descriptors that are over-represented in some classes ([8], [31], [58]). Another variation, which is designed for descriptor spaces that contain descriptors of different sizes, calculates a different similarity value for the descriptors belonging to each of the different sizes and then combines them to yield a single similarity value ([55]). This approach ensures that each individual size contributes equally to the overall similarity score and that the score is not unnecessarily dominated by the large-size descriptors, which are often more abundant.

3.2 Approaches based on Graph Kernels

The approaches based on graph kernels determine the similarity of two chemical compounds by directly comparing their molecular graphs without having to generate an intermediate descriptor-based representation ([47], [49], [40], [33]). A number of graph kernels have been developed and used in the context of building SAR and SPR models. This includes approaches that measure the similarity between two molecular graphs as the size of their maximum common subgraph ([41]), by using powers of adjacency matrices ([40]), by calculating Markov random walks on the underlying graphs ([40]), and by using weighted substructure matching between two graphs ([33]). For instance, the kernels based on powers of adjacency matrices count shared labelled sequences (paths) between two chemical graphs. Markov random walk kernels also compute the matches generated by walks (paths) on the two chemical compounds. However, as the name suggests, the match is derived by markov random walks on the two graphs. Note that the above two kernels are similar in flavor to path-based descriptor-space similarity described earlier. Weighted substructure matching kernel assigns weights based on the number of embeddings of a common substructure found in the two chemical graphs. In this approach, a substructure of size l is centered around an atom and consists of all atoms and bonds that can be reached by a path of length l via this atom. This kernel is similar in flavor to the extended connectivity fingerprints (ECFP) described earlier. However, in the case of this kernel function, no explicit descriptor-space is generated.

4. Searching Compound Libraries

Searching large databases of chemical compounds, often referred to as *compound libraries*, in order to identify compounds that share the same bioactivity (i.e., they bind to the same protein or class of proteins) with a certain *query* compound is arguably the most widely used operation involving chemical compounds and an essential step towards the iterative optimization of a compound's binding affinity, selectivity, and other pharmaceutically relevant properties. This search is usually performed against different libraries (e.g., corporate library, libraries of commercially available compounds, libraries of patented compounds, etc) and provide key information that can be used to identify other more potent compounds and to guide the synthesis of small-scale libraries around the initial query compounds.

Depending on the initial properties of the query compound and the goal of the iterative optimization process, there are two distinct types of operations that the database search mechanisms needs to support. The first is the standard *rank-retrieval* operation whose goal is to identify compounds that are similar to the query in terms of their bioactivity. The second is the *scaffold-hopping* operation whose goal is to identify compounds that are similar to the query in terms of their bioactivity but their structures are different from that of the query (different scaffolds). This latter operation is used when the query compound has some undesirable properties such as toxicity, bad ADME (absorption, distribution, metabolism and excretion), or may be promiscuous ([18], [45]). Since these properties are often shared by the compounds that have very similar structures, it is important to identify as many chemical compounds as possible that not only show the desired activity for the biomolecular target but also have different structures (come from diverse chemical classes or chemotypes) ([64], [18], [48]). Furthermore, scaffold-hopping is also important from the point of view of un-patented chemical space. Many important lead compounds and drug candidates have already been patented. In order to find new therapies and offer alternative treatments it is important for a pharmaceutical company to discover novel leads significantly different from the existing patented chemical space.

The solution to the ranked-retrieval operation relies on the well known fact that the chemical structure of a compound relates to its activity (SAR). As such, effective solutions can be devised that rank the compounds in the database based on how structurally similar they are to the query. However, for scaffold-hopping, the compounds retrieved must be structurally *sufficiently* similar to possess similar bioactivity but at the same time must be structurally *dissimilar* enough to be a novel chemotype. This is a much harder operation than simple ranked-retrieval as it has the additional constraint of maximizing dissimilarity that runs counter to the relationship between the structure of a compound and its activity.

The rest of this section describes two sets of techniques for performing the ranked-retrieval and scaffold-hopping operations. The first are inspired by advances in automatic relevance feedback mechanism and use techniques such as the automatic query expansion to identify structurally different compounds from the query. The second measure the similarity between the query and a compound by taking into account additional information beyond their structure-based similarities. This *indirect* way of measuring similarity enables the retrieval of compounds that are structurally different from the query but at the same time possess the desired bioactivity. The indirect similarities are derived by analyzing the similarity network formed by the query and the database compounds. These indirect similarity based techniques operate on the descriptor-space representation of the compounds and are independent of the selected descriptor-space.

4.1 Methods Based on Direct Similarity

Many methods have been proposed for ranked-retrieval and scaffold-hopping that directly operate on the underlying descriptor space representation. These *direct similarity* based methods can be divided into two groups. The first contains methods that rely on better designed descriptor-space representations, whereas the second contains methods that are not specific to any descriptor-space representation but utilize different retrieval strategies to improve the overall performance.

Among the first set of methods, 2D descriptors described in Section 2 such as path-based fingerprints (fp), dictionary based keys (MACCS) and more recently Extended Connectivity fingerprints (ECFP) as well as Graph Fragments (GF) have all been successfully applied for the retrieval problem([55]). However, for scaffold-hopping, pharmacophore based descriptors such as ErG ([48]) have been shown to outperform 2D topology based descriptors ([48], [64]). Lastly, descriptors based on 3D structure or conformations of the molecule have also been applied successfully for scaffold-hopping ([64], [45]).

The second set of methods include the turbo search based schemes ([18]) which utilize ideas from automatic relevance feedback mechanism ([1]). The turbo search techniques operate as follows. Given a query q, they start by retrieving the top-k compounds from the database. Let A be the $(k + 1)$-size set that contains q and the top-k compounds. For each compound $c \in A$, all the compounds in the database are ranked in decreasing order based on their similarity to c, leading to $k+1$ ranked lists. These lists are combined to obtain the final similarity of each compound with respect to the initial query. Similar methods based on consensus scoring, rank averaging, and voting have also been investigated ([64]).

4.2 Methods Based on Indirect Similarity

Recently, a set of techniques to improve the scaffold-hopping performance have been introduced that are based on measuring the similarity between the query and a compound by taking into account additional information beyond their descriptor-space-based representation ([54], [56]). These methods are motivated by the observation that if a query compound q is structurally similar to a database compound c_i and c_i is structurally similar to another database compound c_j, then q and c_j could be considered as being similar or related even though they may have zero or very low direct similarity. This *indirect* way of measuring similarity can enable the retrieval of compounds that are structurally different from the query but at the same time, due to associativity, possess the same bioactivity properties with the query.

The set of techniques developed to capture such indirect similarities are inspired by research in the fields of information retrieval and social network analysis. These techniques derive the indirect similarities by analyzing the network formed by a k-nearest-neighbor graph representation of the query and the database compounds. The network linking the database compounds with each other *and* with the query is determined by using a *k-nearest-neighbor* (NG) and a *k-mutual-nearest-neighbor* (MG) graph. Both of these graphs contain a node for each of the compounds as well as a node for the query. However, they differ on the set of edges that they contain. In the k-nearest-neighbor graph there is an edge between a pair of nodes corresponding to compounds c_i and c_j, if c_i is in the k-nearest-neighbor list of c_j or vice-versa. In the k-mutual-nearest-neighbor graph, an edge exists only when c_i is in the k-nearest-neighbor list of c_j *and* c_j is in the k-nearest-neighbor list of c_i. As a result of these definitions, each node in NG will be connected to at least k other nodes (assuming that each compound has a non-zero similarity to at least k other compounds), whereas in MG, each node will be connected to at most k other nodes.

Since the neighbors of each compound in these graphs correspond to some of its most structurally similar compounds and due to the relation between structure and activity (SAR), each pair of adjacent compounds will tend to have similar activity. Thus, these graphs can be considered as network structures for capturing bioactivity relations.

A number of different approaches have been developed for determining the similarity between nodes in social networks that take into account various topological characteristics of the underlying graphs ([50], [13]).For the problem of scaffold-hopping, the similarity between a pair of nodes is determined as a function of the intersection of their adjacency lists ([54], [56]), which takes into account all two-edge paths connecting these nodes. Specifically, the similarity between c_i and c_j with respect to graph G is given by

$$\mathrm{isim}_G(c_i, c_j) = \frac{\mathrm{adj}_G(c_i) \cap \mathrm{adj}_G(c_j)}{\mathrm{adj}_G(c_i) \cup \mathrm{adj}_G(c_j)}, \tag{4.1}$$

where $\mathrm{adj}_G(c_i)$ and $\mathrm{adj}_G(c_j)$ are the adjacency lists of c_i and c_j in G, respectively.

This measure assigns a high similarity value to a pair of compounds if both are very similar to a large set of common compounds. Thus, compounds that are part of reasonably tight clusters (i.e., a set of compounds whose structural similarity is high) will tend to have high indirect similarities as they will most likely have a large number of common neighbors. In such cases, the indirect similarity measure re-enforces the existing high direct similarities between compounds. However, the indirect similarity between a pair of compounds c_i and c_j can also be high even if their direct similarity is low. This can happen when the compounds in $\mathrm{adj}_G(c_i) \cap \mathrm{adj}_G(c_j)$ match different structural descriptors of c_i and c_j. In such cases, the indirect similarity measure is capable of identifying relatively weak structural similarities, making it possible to identify scaffold-hopping compounds.

Given the above graph-based indirect similarity measures, various strategies can be employed to retrieve compounds from the database. Three such strategies are discussed below. The first corresponds to that used by the standard ranked-retrieval method, whereas the other two are inspired by information retrieval methods used for automatic relevance feedback ([1]) and are specifically designed to improve the scaffold-hopping performance.

Best-Sim Retrieval Strategy. This is the most widely used retrieval strategy and it simply returns the compounds that are the most similar to the query. Specifically, if A is the set of compounds that have been retrieved thus far, then the next compound c_{next} that is selected is given by

$$c_{next} = \arg\max_{c_i \in D - A} \{\mathrm{isim}(c_i, q)\}. \tag{4.2}$$

This compound is added to A, removed from the database, and the overall process is repeated until the desired number of compounds has been retrieved ([56]).

Best-Sum Retrieval Strategy. This retrieval strategy incorporates additional information from the set of compounds retrieved thus far (set A). Specifically, the compound selected, c_{next}, is the one that has the highest average similarity to the set $A \cup \{q\}$. That is,

$$c_{next} = \arg\max_{c_i \in D-A} \{\text{isim}(c_i, A \cup \{q\})\}. \tag{4.3}$$

The motivation behind this approach is that due to SAR, the set A will contain a relatively large number of active compounds. Thus, by modifying the similarity between q and a compound c to also include how similar c is to the compounds in the set A, a similarity measure that is re-enforced by A's active compounds is obtained ([56]). This enables the retrieval of active compounds that are similar to the compounds present in A even if their similarity to the query is not very high; thus, enabling scaffold-hopping.

Best-Max Retrieval Strategy. A key characteristic of the retrieval strategy described above is that the final ranking of each compound is computed by taking into account *all* the similarities between the compound and the compounds in the set A. Since the compounds in A will tend to be structurally similar to the query compound, this approach is rather conservative in its attempt to identify active compounds that are structurally different from the query (i.e., scaffold-hops).

To overcome this problem, a retrieval strategy was developed ([56]) that is based on the best-sum approach but instead of selecting the next compound based on its average similarity to the set $A \cup \{q\}$, it selects the compound that is the most similar to *one* of the compounds in $A \cup \{q\}$. That is, the next compound is given by

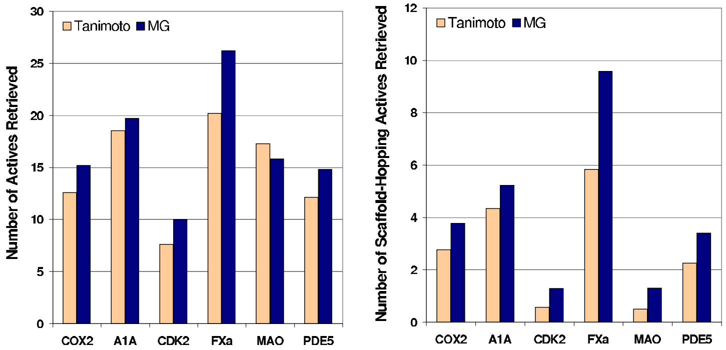
$$c_{next} = \arg\max_{c_i \in D-A} \{ \max_{c_j \in A \cup \{q\}} \text{isim}(c_i, c_j) \}. \tag{4.4}$$

In this approach, if a compound c_j other than q has the highest similarity to some compound c_i in the database, c_i is chosen as c_{next} and added to A irrespective of its similarity to q. Thus, the query-to-compound similarity is not necessarily included in every iteration as in the other schemes, allowing this strategy to identify compounds that are structurally different from the query.

4.3 Performance of Indirect Similarity Methods

The performance of indirect similarity-based retrieval strategies based on the NG as well as MG graph was compared to direct similarity based on Tanimoto coefficient ([56]). The compounds were represented using different descriptor-spaces (GF, ECFP, and ErG). The quantitative results showed that indirect similarity is consistently, and in many cases substantially, better than direct similarity. Figure 19.1 shows a part of the results in [56] which compare MG based indirect similarity to direct Tanimoto coefficient (TM) similarity searching using ECFP descriptors. It can be observed from the figure

Figure 19.1. Performance of indirect similarity measures (MG) as compared to similarity searching using the Tanimoto coefficient (TM).



Tanimoto indicates the performance of similarity searching using the Tanimoto coefficient with extended connectivity descriptors; MG indicates the performance of similarity searching using the indirect similarity approach on the mutual neighbors graph formed using extended connectivity fingerprints.

that indirect similarity outperforms direct similarity for scaffold-hopping active retrieval in all of six datasets that were tested. It can also be observed that indirect similarity outperforms direct similarity for active compound retrieval in all datasets except MAO. Moreover, the relative gains achieved by indirect similarity for the task of identifying active compounds with different scaffolds is much higher, indicating that it performs well in identifying compounds that have similar biomolecule activity even when their direct similarity is low.

5. Identifying Potential Targets for Compounds

Target-based drug discovery, which involves selection of an appropriate target (typically a single protein) implicated in a disease state as the first step, has become the primary approach of drug discovery in pharmaceutical industry ([2], [46]). This was made possible by the advent of High Throughput Screening (HTS) technology in the late 1980s that enabled rapid experimental testing of a large number of chemical compounds against the target of interest. HTS is now routinely utilized to identify the most promising compounds (*hits*) that show desired binding/activity against a given target. Some of these compounds then go through the long and expensive process of optimization, and eventually one of them may go to clinical trials. If clinical trails are successful then the compound becomes a drug. HTS technology ushered in a new era of drug discovery by reducing the time and money taken to find hits that will have a high chance of eventually becoming a drug.

However, the increased number of candidate hits from HTS did not increase the number of actual drugs coming out of the drug discovery pipeline. One of the principal reasons for this failure is that the above approach only focuses on the target of interest, taking a very narrow view of the disease. As such, it may lead to unsatisfactory phenotypic effects such as toxicity, promiscuity, and low efficacy in the later stages of drug discovery ([46]). More recently, research focus is shifting to directly screen molecules to identify desirable phenotypic effects using cell-based assays. This screening evaluates properties such as toxicity, promiscuity and efficacy from the onset rather than in later stages of drug discovery ([23], [46]). Moreover, toxicity and off-target effects are also a focus of early stages of conventional target-based drug discovery ([5]). But from the drug discovery perspective, target identification and subsequent validation has become the rate limiting step in order to tackle the above issues ([12]). Targets must be identified for the hits in phenotypic assay experiments and for secondary pharmacology as the activity of hits against all of its potential targets sheds light on the toxicity and promiscuity of these hits ([5]). Therefore, the identification of all likely targets for a given chemical compound, also called *Target Fishing* ([23]), has become an important problem in drug discovery.

Computational techniques are becoming increasingly popular for target fishing due to large amounts of data from high-throughput screening (HTS), microarrays, and other experiments ([23]). Given a compound, these techniques initially assign a score to each potential target based on some measure of likelihood that the compound binds to the target. These techniques then select as the compound's targets either those targets whose score is above a certain cut-off or a small number of the highest scoring targets. Some of the early target fishing methods utilized approaches based on reverse docking ([5]) and nearest-neighbor classification ([35]). Reverse docking approaches dock a compound against all the targets of interest and identify as the most likely targets those that achieve the best binding affinity score. Note that these approaches are applicable only for proteins with resolved 3D structure and as such their applicability is somewhat limited. The nearest-neighbor approaches rely on the structure-activity-relationship (SAR) principle and identify as the most likely targets for a compound the targets whose nearest neighbors show activity against. In these approaches the solution to the target fishing problem only depends on the underlying descriptor-space representation, the similarity function employed, and the definition of nearest neighbors. However, the performance of these approaches has been recently surpassed by a new set of *model-based* methods that solve the target fishing problem using various machine-learning approaches to learn models for each one of the potential targets based on their known ligands ([36], [25], [53]). These methods are further discussed in the subsequent sections.

5.1 Model-based Methods For Target Fishing

Two different approaches have been employed to build models suitable for target fishing. In the first approach, a separate SAR model is built for every target. For a given test compound, these models are used to obtain a score for each target against this compound. The highest scoring targets are then considered as the most likely targets that this compound will bind to ([36], [53], [23]). This approach is similar to the reverse docking approach described earlier. However, the target scores for a compound are obtained from the models built for each target instead of the docking procedure. The second approach treats target fishing problem as an instance of the multilabel prediction problem and uses category ranking algorithms([6]) to solve this problem ([53]).

Bayesian Models for Target Fishing (Bayesian). This approach utilizes multi-category bayesian models ([36]) wherein a model is built for every target in the database using SAR data available for each target. Compounds that show activity against a target are used as positives for that target and the rest of the compounds are treated as negatives. The input to the algorithm is a training set consisting of a set of chemical compounds and a set of targets. A model is learned for every target given a descriptor-space representation of training chemical compounds ([36]). For a new chemical compound whose targets have to be predicted, an estimator score is computed for each target reflecting the likelihood of activity against this target using the learned models. The target can be ranked according to their estimator scores and the targets that get high scores can be considered as the most likely targets for this compound.

SVM-based Method (SVM rank). This approach for solving the ranking problem builds for each target a one-versus-rest binary SVM classifier ([53]). Given a test chemical compound c, the classifier for each target will then be applied to obtain a prediction score. The ranking of the targets will be obtained by simply sorting the targets based on their prediction scores. If there are N targets in the set of targets $\mathcal{T}$ and $f_i(c)$ is the score obtained for the i^{th} target, then the final ranking $\mathcal{T}^*$ is obtained by

$$\mathcal{T}^* = \underset{\tau_i \in \mathcal{T}}{\operatorname{argsort}} \left\{ f_i(c) \right\}, \tag{5.1}$$

where argsort returns an ordering of the targets in decreasing order of their prediction scores $f_i(c)$. Note that this approach assumes that the prediction scores obtained from the N binary classifiers are directly comparable, which may not necessarily be valid. This is because different classes may be of different sizes and/or less separable from the rest of the dataset, indirectly affecting the nature of the binary model that was learned, and consequently its prediction scores. This SVM-based sorting method is similar to the approach proposed by Kawai and co-workers ([25]).

Cascaded SVM-based Method (Cascade SVM). A limitation of the previous approach is that by building a series of one-vs-rest binary classifiers,

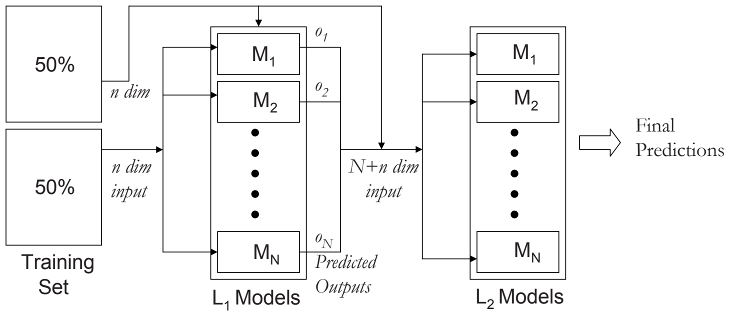


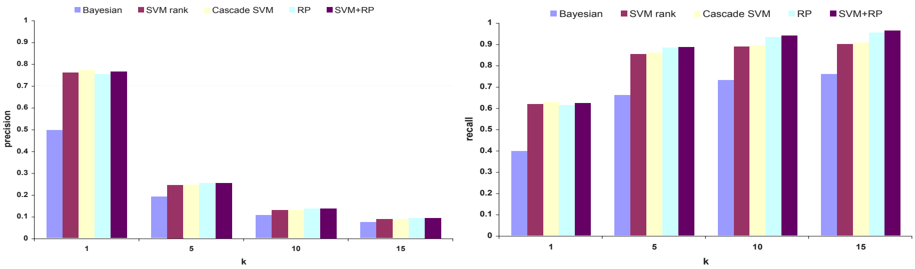
Figure 19.2. Cascaded SVM Classifiers.

it does not explicitly couple the information on the multiple categories that each compound belongs to during model training. As such it cannot capture dependencies that might exist between the different categories. A promising approach that has been explored to capture such dependencies is to formulate it as a cascaded learning problem ([53], [16]). In these approaches, two sets of binary one-vs-rest classification models for each category, referred to as L_1 and L_2, are connected together in a cascaded fashion. The L_1 models are trained on the initial inputs and their outputs are used as input, either by themselves or in conjunction with the initial inputs, to train the L_2 models. This cascaded process is illustrated in Figure 19.2. During prediction time, the L_1 models are first used to obtain predictions which are used as input to the L_2 models which produces the final predictions. Since the L_2 models incorporate information about the predictions produced by the L_1 models, they can potentially capture inter-category dependencies.

A two level SVM based method inspired by the above approach is described in [53]. In this method, both the L_1 and L_2 models consist of N binary one-vs-rest SVM classifiers, one for each target in the set of targets $\mathcal{T}$. The L_1 models correspond exactly to the set of models built by the one-vs-rest method discussed in the previous approach. The representation of each compound in the training set for the L_2 models consists of its descriptor-space based representation and its output from each of the N L_1 models. Thus, each compound c corresponds to an $n + N$ dimensional vector, where n is the dimensionality of the descriptor space. The final ranking $\mathcal{T}^*$ of the targets for a test compound will be obtained by sorting the targets based on their prediction scores from the L_2 models ($f_i^{L_2}(c)$). That is,

$$\mathcal{T}^* = \underset{\tau_i \in \mathcal{T}}{\text{argsort}} \left\{ f_i^{L_2}(c) \right\}, \tag{5.2}$$

Ranking Perceptron Based Method (RP). This approach is based on the online version of the ranking perceptron algorithm proposed to learn a ranking



(a) Precision in Top-k

(b) Recall in Top-k

Figure 19.3. Precision and Recall results

function on a set of categories developed by Crammer and Singer ([6], [53]). This algorithm takes as input a set of objects and the categories that they belong to and learns a function that for a given object c it ranks the different categories based on the likelihood that c binds to the corresponding targets. During the learning phase, the distinction between categories is made only via a binary decision function that takes into account whether a category is part of the object's categories (relevant set) or not (non-relevant set). As a result, even though the output of this algorithm is a total ordering of the categories, the learning is only dependent on the partial orderings induced by the set of relevant and non-relevant categories.

The algorithm employed for target fishing extends the work of Crammer and Singer by introducing margin based updates and extending the online version to a batch setting([53]). It learns a linear model W that corresponds to a $N \times n$ matrix, where N is the number of targets and n is the dimensionality of the descriptor space. Thus, the above method can be directly applied on the descriptor-space representation of the training set of chemical compounds.

Finally, the prediction score for compound c_i and target τ_j is given by $\langle W_j, c_i \rangle$, where W_j is the jth row of W, c_i is the descriptor-space representation of the compound, and $\langle \cdot, \cdot \rangle$ denotes a dot-product operation. Therefore, the predicted ranking for a test chemical compound c is given by

$$\mathcal{T}^* = \underset{\tau_j \in \mathcal{T}}{\operatorname{argsort}} \{\langle W_j, c \rangle\}. \tag{5.3}$$

SVM+Ranking Perceptron-based Method (SVM+RP). A limitation of the above ranking perceptron method over the SVM-based methods is that it is a weaker learner as (i) it learns a linear model, and (ii) it does not provide any guarantees that it will converge to a good solution when the dataset is not linearly separable. In order to partially overcome these limitations a scheme that is similar in nature to the cascaded SVM-based approach previously described was developed in which the L_2 models are replaced by a ranking perceptron ([53]). Specifically, N binary one-vs-rest SVM models are trained, which form the set of L_1 models. Similar to the cascade SVM method, the representation of each compound in the training set for the L_2 models consists of its descriptor-space based representation and its output from each of the N L_1 models. Finally, a ranking model W learned using the ranking perceptron described in the previous section. Since the L_2 model is based on the descriptor-space based representation and the outputs of the L_1 models, the size of W is $N \times (n + N)$.

5.2 Performance of Target Fishing Strategies

An extensive evaluation of the different Target Fishing methods was performed recently ([53]) which primarily used the PubChem ([39]) database to extract target-specific dose-response confirmatory assays. Specifically, the ability of the five methods to identify relevant categories in the top-k ranked categories was assessed in this work. The results were analyzed along this direction because this directly corresponds to the use case scenario where a user may want to look at top-k predicted targets for a test compound and further study or analyze them for toxicity, promiscuity, off-target effects, pathway analysis *etc*([53]). The comparisons utilized precision and recall metric in top-k for each of the five schemes. as shown in Figures 19.3a) and 19.3b). These figures show the actual precision and recall values in top-k by varying k from one to fifteen.

These figures indicate that for identifying one of the correct categories or targets in the top 1 predictions, cascade SVM outperforms all the other schemes in terms of both precision and recall. However, as k increases from one to fifteen, the precision and recall results indicate that the best performing scheme is the SVM+Ranking Perceptron and it outperforms all other schemes for both precision as well as recall. Moreover, these values in figure 19.3b) show that as k increases from one to fifteen, both the ranking perceptron based schemes (RP and SVM+RP) start performing consistently better that others in identifying all the correct categories. The two ranking perceptron based schemes also achieve average precision values that are better than other schemes in the top fifteen (Figure 19.3a)).

6. Future Research Directions

Mining and retrieving chemical data for a single biomolecular target and building SAR models on it has been traditionally used to predict as well as analyze the bioactivity and other properties of chemical compounds and plays a key role in drug discovery. However, in recent years the wide-spread use of High-Throughput Screening (HTS) technologies by the pharmaceutical industry has generated a wealth of protein-ligand activity data for large compound libraries against many biomolecular targets. The data has been systematically collected and stored in centralized databases ([38]). At the same time, the completion of the human genome sequencing project has provided a large number of "druggable" protein targets ([44]) that can be used for therapeutic purposes. Additionally, a large fraction of the protein targets that have or are currently been investigated for therapeutic purposes are confirmed to belong to a small number of gene families ([62]). The combination of these three factors has led to the development of methods that utilize information that goes beyond the traditional single biomolecular target's chemical data analysis. In recent years, the trend has been to integrate chemical data with protein and genetic data (bioinformatics data) and analyze the problem over multiple proteins or different protein families. Consequently, Chemogenomics ([43]), Poly-Pharmacology ([38])and Target Fishing ([23]) have emerged as important problems in drug discovery.

Another new direction that utilizes graph mining is network pharmacology. A fundamental assumption in drug discovery that has been applied widely in the past decades is the "one gene, one drug, on disease" assumption. However, the increasing failure in translating drug candidates into effective therapies raises the challenges to this assumption. Recent studies show that the modulating or effecting an individual gene or gene product has little effects on disease network. For example, under laboratory conditions, many single-gene knockouts by themselves exhibit little or no effects on phenotype and only 19% of genes were found to be essential across a number of model organisms ([63]). This robustness of phenotype can be understood in terms of redundant functions and alternative compensatory signalling routes. In addition, large scale functional genomics studies reveal the importance of polypharmacology, which suggests that is, instead of focusing on drugs that are maximally selective against a single drug target, the focus should be to select the drug candidates that interact with multiple proteins that are essential in the biological network. This new paradigm is refereed to as network pharmacology ([21]).

Graph mining has also been utilized to study the drug-target interaction network. Such networks provide topological information between drug and target interactions that once explored may suggest novel perspective in terms of drug discovery that is not possible by looking at drugs and targets in isolation. Learning from drug-target interaction networks has been focused on predicting drugs for targets that are novel, or that have only a few drugs known (*Target Hopping*). These methods tend to leverage the knowledge of both targets and the drug simultaneously to obtain characteristics of drug-target interaction networks. Many of the learning methods utilize Support Vector Machine (SVM). In this approach, novel kernels have been developed that relate drugs and targets explicitly. For example, Yamanish *et al.*([60]), developed profiles to represent interactions between drugs and targets, and then used kernel regression to the relationship among the interactions. Their framework enables predictions of unknown drug-target interactions.

With the improvement in high throughput technologies in chemistry, genomics, proteomics, and chemical genetics, graph mining is set to play an important role in the understanding of human disease and pursuit of novel therapies for these diseases.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, first edition, 1999.

[2] H.J. Bohm and G. Schneider. *Virtual Screening for Bioactive Molecules.* Wiley-VCH, 2000.

[3] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. Vishwanathan, A. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *BMC Bioinformatics*, 21:47–56, 2005.

[4] Chemaxon. *Screen, Chemaxon Inc.*, 2005.

[5] Y. Z. Chen and C. Y. Ung. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J Mol Graph Model*, 20(3):199–218, 2001.

[6] K. Crammer and Y. Singer. A new family of online algorithms for category ranking. *Journal of Machine Learning Research.*, 3:1025–1058, 2003.

[7] Daylight. *Daylight Toolkit, Daylight Inc, Mission Viejo, CA, USA*, 2008.

[8] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE TKDE.*, 17(8):1036–1050, 2005.

[9] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining*, pages 269–274, 2001.

[10] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Info. Model.*, 42(6):1273–1280, 2002.

[11] ECFP. *Pipeline Pilot, Accelrys Inc: San Diego CA 2008.*, 2006.

[12] Ulrike S Eggert and Timothy J Mitchison. Small molecule screening by imaging. *Curr Opin Chem Biol*, 10(3):232–237, Jun 2006.

[13] F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random walk computation of similarities between nodes of a graph with application to collaborative filtering. *IEEE TKDE*, 19(3):355–369, 2007.

[14] H. Geppert, T. Horvath, T. Gartner, S. Wrobel, and J. Bajorath. Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2d fingerprints and multiple reference compounds. *J. Chem. Inf. Model.*, 48:742–746, 2008.

[15] M. Glick, J. L. Jenkins, J. H. Nettles, H. Hitchings, and J. H. Davies. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.*, 46:193–200, 2006.

[16] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. *PAKDD.*, pages 22–30, 2004.

[17] C. Hansch, P. P. Maolney, T. Fujita, and R. M. Muir. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194:178–180, 1962.

[18] J. Hert, P. Willet, and D. Wilton. New methods for ligand based virtual screening: Use of data fusion and machine learning to enchance the effectiveness of similarity searching. *J. Chem. Info. Model.*, 46:462–470, 2006.

[19] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem*, 2(22):3256–66, 2004.

[20] Hologram. *Hologram Fingerprints, Tripos Inc. 1699 South Hanley Road, St Louis, MO 63144-2913, USA*. http://www.tripos.com, 2003.

[21] Andrew L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4(11):682–690, November 2008.

[22] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J. Comput. Biol.*, 12(6):657–671, 2005.

[23] J. L. Jenkins, A. Bender, and J. W. Davies. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today*, 3(4):413–421, 2006.

[24] R. N. Jorissen and M. K. Gibson. Virtual screening of molecular databases using support vector machines. *J. Chem. Info. Model.*, 45(3):549–561, 2005.

[25] K. Kawai, S. Fujishima, and Y. Takahashi. Predictive activity profiling of drugs by topological-fragment-spectra-based support vector machines. *J. Chem. Info. Model.*, 48(6):1152–1160, 2008.

[26] T. Kogej, O. Engkvist, N. Blomberg, and S. Moresan. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Info. Model.*, 46(3):1201–1213, 2006.

[27] M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. *IEEE TKDE.*, 16(9):1038–1051, 2004.

[28] A. R. Leach and V. J. Gillet. *An Introduction to Chemoinformatics.* Springer, 2003.

[29] Andrew R. Leach. *Molecular Modeling: Principles and Applications.* Prentice Hall, Englewood Cliffs, NJ, second edition, 2001.

[30] W. Liu, W. Lin, A. Davis, F. Jordan, H. Yang, and M. Hwang. A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC Bioinformatics*, 8:121, 2007.

[31] Y. Liu. A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.*, 44:1823–1828, 2004.

[32] MDL. *MDL Information Systems Inc., San Leandro, CA, USA*. http://www.mdl.com, 2004.

[33] S. Menchetti, F. Costa, and P. Frasconi. Weighted decomposition kernels. *Proceedings of the 22nd International Conference in Machine Learning.*, 119:585–592, 2005.

[34] H. L. Morgan. The generation of unique machine description for chemical structures: a technique developed at chemical abstract services. *Journal of Chemical Documentation*, 5:107–113, 1965.

[35] J. Nettles, J. Jenkins, A. Bender, Z. Deng, J. Davies, and M. Glick. Bridging chemical and biological space: "target fishing" using 2d and 3d molecular descriptors. *J Med Chem*, 49:6802–6810, Nov 2006.

[36] Nidhi, M. Glick, J. Davies, and J. Jenkins. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *J Chem Inf Model*, 46:1124–1133, 2006.

[37] S. Nijssen and J. Kok. A quickstart in frequent structure mining can make a difference. *Proceedings of SIGKDD*, pages 647–652, 2004.

[38] G. V. Paolini, R. H. Shapland, W. P. Van Hoorn, J. S. Mason, and A. Hopkins. Global mapping of pharmacological space. *Nature biotechnology*, 24:805–815, 2006.

[39] Pubchem. *The PubChem Project*, 2007.

[40] L. Ralaivola, S. J. Swamidassa, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110, 2005.

[41] J. W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comp. Aided Mol. Des.*, 16(7):521–533, 2002.

[42] D. Rogers, R. Brown, and M. Hahn. Using extended-connectivity fingerprints with laplacian-modified bayesian analysis in high-throughput screening. *J. Biomolecular Screening*, 10(7):682–686, 2005.

[43] D. Rognan. Chemogenomic approaches to rational drug design. *Br J Pharmacol*, 152(1):38–52, Sep 2007.

[44] A. P. Russ and S. Lampel. The druggable genome: an update. *Drug Discov Today*, 10(23-24):1607–10, 2005.

[45] Jamal C. Saeh, Paul D. Lyne, Bryan K. Takasaki, and David A. Cosgrove. Lead hopping using svm and 3d pharmacophore fingerprints. *J. Chem. Info. Model.*, 45:1122–113, 2005.

[46] Frank Sams-Dodd. Target-based drug discovery: is something wrong? *Drug Discov Today*, 10(2):139–147, Jan 2005.

[47] A.J. Smola and R. Kondor. Kernels and regularization on graphs. In *Proceedings COLT and Kernels Workshop*, pages 144–158. M.Warmuth and B. Schelkopf, 2003.

[48] Nikolaus Stiefl, Ian A. Watson, Kunt Baumann, and Andrea Zaliani. Erg: 2d pharmacophore descriptor for scaffold hopping. *J. Chem. Info. Model.*, 46:208–220, 2006.

[49] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(1):359–368, 2005.

[50] B. Teufel and S. Schmidt. Full text retrieval based on syntactic similarities. *Information Systems*, 31(1), 1988.

[51] Unity. *Unity Fingerprints, Tripos Inc. 1699 South Hanley Road, St Louis, MO 63144-2913, USA*. http://www.tripos.com, 2003.

[52] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[53] N. Wale and G. Karypis. Target identification for chemical compounds using target-ligand activity data and ranking based methods. Technical Report TR-08-035, University of Minnesota, 2008. Accepted: Jour. Chem. Inf. Model, Published on the web, September 18, 2009.

[54] N. Wale, G. Karypis, and I. A. Watson. Method for effective virtual screening and scaffold-hopping in chemical compounds. *Comput Syst Bioinformatics Conf*, 6:403–414, 2007.

[55] N. Wale, I. A. Watson, and G. Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, 14:347–375, 2008.

[56] N. Wale, I. A. Watson, and G. Karypis. Indirect similarity based methods for effective scaffold-hopping in chemical compounds. *J. Chem. Info. Model.*, 48(4):730–741, 2008.

[57] A. M. Wassermann, H. Geppert, and J. Bajorath. Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.*, 49:582–592, 2009.

[58] J. Wegner, H. Frohlich, and Andreas Zell. Feature selection for descriptor based classification models. 1. theory and ga-sec algorithm. *J. Chem. Inf. Comput. Sci.*, 44:921–930, 2004.

[59] P. Willett. A screen set generation algorithm. *J. Chem. Inf. Comput. Sci.*, 19:159–162, 1979.

[60] Y. Yamanishi, M. Araki, A. Gutteridge, W. Hondau, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:232–240, 2008.

[61] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. *ICDM*, pages 721–724, 2002.

[62] M. Yildirim, K. Goh, M. Cusick, A. Barabasi, and M. Vidal. Drug-target network. *Nat Biotechnol*, 25(10):1119–1126, Oct 2007.

[63] Brian P. Zambrowicz and Arthur T. Sands. Modeling drug action in the mouse with knockouts and rna interference. *Drug Discovery Today: TARGETS*, 3(5):198 – 207, 2004.

[64] Qiang Zhang and Ingo Muegge. Scaffold hopping through virtual screening using 2d and 3d similarity descriptors: Ranking, voting and consensus scoring. *J. Chem. Info. Model.*, 49:1536–1548, 2006.

[65] Ziding Zhang and Martin G Grigorov. Similarity networks of protein binding sites. *Proteins*, 62(2):470–478, Feb 2006.