CHAPTER 1

General Introduction

This thesis is about drug discovery and how the computer can assist in this lengthy and costly process. As I will explain in the paragraphs below it still makes good sense today as it did in the past to have a closer look at the chemical structure of drugs, and in particular to elements or fragments in these chemical structures. Therefore, I should like to start my thesis with some thoughts on drugs and drug discovery *per se* and how developments in informatics and computer science offer new opportunities. I will mostly focus on the most important drug targets of today, the so-called G protein-coupled receptors (GPCRs)

1.1 Drug discovery

1.1.1 Drugs

Drug discovery is the process of finding new drugs. It has evolved from early serendipitous discovery from natural sources, such as morphine from poppy seeds, to today's industrial-scale screening projects.[1, 2] Modern drug discovery starts with the identification of a biological target that can be modulated to induce the desired therapeutic effect. To search for potential drugs, compounds are tested for their ability to modulate the target. Technologies such as high-throughput screening (HTS) and combinatorial chemistry facilitate screening of libraries of millions of compounds. However, even though screening capacity reaches millions of compounds and continues to grow, the number of possible molecules that could be synthesized and tested is infinitely larger. This notion has important consequences; apparently, we can only test a tiny sample of what is virtually available. The obvious question then is 'what to test'.

1.1.2 Chemical space

The collection of all possible energetically stable molecules is referred to as 'chemical space'. Chemical space can be traversed by means of chemical transformations, which start from one point in chemical space, i.e. a molecule, and end at another. These transformations are either hypothetical modifications of the molecular structure or synthetic chemical reactions. While 'travelling' from one molecule to another takes place by chemical transformations, a clear account on the dimensionality of chemical space or the axes along which is travelled cannot be given. Usually, a few molecular properties are chosen to serve as dimensions. The choice of properties depends on the purpose of the chemical space representation, for instance, whether the aim is to analyze distribution of a set of molecules in chemical space or whether the aim is to

extrapolate to find molecules with better properties. Defining axes in chemical space facilitates distance measurement between molecules, an important concept in cheminformatics.

Chemical space is infinite; however, by putting limits on the size of the molecules that are considered, estimates of the number of molecules in chemical space can be made. For instance, when restricting to small organic molecules, size estimates range from 10^{20} to more than 10^{60} molecules,[3, 4] a number for which there would not be sufficient matter in the universe to synthesize all. The number of known organic and inorganic molecules fades when considering the number of theoretical possibilities. For instance, the number of molecules registered in the CAS Registry is approx. 64 million inorganic and organic substances (accessed on November 2, 2011).[5]

While it is impossible to synthesize all possible (small) molecules, it is feasible to construct and explore a virtual chemical space with the aid of the computer. Efforts to do so are not new, and started in the field of automated structure elucidation, for instance for the purpose of NMR spectroscopy.[6] However, a systematic enumeration of all possible organic molecules (with constrained size and type) was first carried out by Fink *et al*.[7] The authors constructed a database named GDB with all stable stereoisomers consisting of up to 11 atoms of the types carbon, nitrogen, oxygen, and fluorine, by considering simple rules for valency, chemical stability, and synthetic feasibility. The chemical space that this database represented consisted of 26.4 million molecules, or 110.9 million stereoisomers. Mapping compounds from known 'physical' compound databases onto this chemical space indicated large unexplored areas. Later, the same group reported the subsequent construction of a database of all feasible organic molecules that consisted of up to 13 atoms of the elements carbon, nitrogen, oxygen, sulfur, and chlorine.[8] This database (named GDB-13) contains more than 977 million (virtual) structures, which makes it the largest publicly available small organic molecule database known to date. Note that both databases consist of drug-like compounds only, according to Lipinski's rule of five. This rule states that for a compound to be orally absorbed it should have not more than five hydrogen bond

donors, not more than 10 hydrogen bond acceptors, a molecular weight below 500 Daltons, and a calculated octanol-water partitioning coefficient (logP) not higher than 5.[9]

Even if it were possible to synthesize all molecules in chemical space, this would not be desirable since most of the molecules would not have any wanted activity, and thus be a waste of materials and resources. In addition, molecules could display off-target effects, i.e. side effects, or have undesirable properties such as mutagenicity, reactive groups, poor solubility etc. Lastly, active molecules could be too similar to already known or patented drugs. Instead of considering all possible molecules in chemical space, one should attempt to select only those molecules that possess the desired properties. Such molecules can then be acquired or synthesized, and tested. To find the desired molecules from chemical space, computational techniques are needed to predict the properties of the virtual molecules.

1.1.3 Drug design

The purpose of rational drug design is to find new bioactive compounds with favorable properties from the total chemical space. This often implies knowledge of the target, usually a protein, too to find new ligands (i.e. compounds that bind – 'ligate' – to a drug target). These ligands do not necessarily originate from a design process but can also stem from virtual screening of compound libraries. Virtual screening represents the computational counterpart of high-throughput screening. Instead of the physiological target, a computational model that represents the target is used to identify hits. Two main approaches exist in rational drug design: structure-based approaches and ligand-based approaches. The structure-based approaches depend on the availability of the three-dimensional structure of the target under study, for instance, the X-ray crystal structure or a suitable homology model. Computer-modeling techniques are then applied to find new ligands, or modify existing ones, that fit into the target structure. In contrast, ligand-based approaches do not require information on the target protein, but instead rely upon the availability of a sufficient amount of

ligand data. Combinations of the two approaches are found in the realm of chemogenomics, which will be discussed later. While structure-based approaches typically place a high demand on computational resources, these allow for the discovery of truly new chemistry. Moreover, structure-based approaches become more feasible with the steady growth of computing power (known as "Moore's law" [10]) available to the medicinal chemist. On the other hand, ligand-based virtual screening has been shown to perform often equally well compared to structure-based virtual screening.[11]

Finding the desired molecules in chemical space may be accomplished in two distinct ways: first, using selection or prioritization of molecules, and second, by performing a steered search. With selection, molecules with the desired properties are selected (or molecules with undesirable properties are filtered out) from the total set of possible molecules. Prioritization on the other hand, ranks molecules according to the desired properties and selects a certain number of the top-ranked molecules. Note that in practice not the real chemical space is screened; instead, large virtual libraries, or existing libraries of commercially available or in-house compounds are screened. The second approach to find the desired molecules is exploration of chemical space by performing a steered search. With this approach, a 'walk' is conducted along neighboring molecules in the direction of molecules with properties that are more favorable. Neighboring molecules are molecules that can be transformed into one another by modifications of the chemical structure. The general principle is to move gradually towards better molecules by repeatedly selecting the best neighboring molecule as the next 'step'. The process is comparable to what medicinal chemists do to search for molecules with improved properties by synthesizing a set of derivatives from a starting structure. However, with a steered search in chemical space, the properties of the virtual molecules are predicted by computational methods.

1.2 Structures

1.2.1 Compounds

In the search for new bioactive compounds, only the most promising candidates are qualified to be acquired or synthesized and tested; the possibilities should therefore be explored virtually. Analysis of existing ligands may help in the search for new bioactive compounds, provided that the set of ligands is sufficiently large. Diversity among ligands favors extrapolation to new chemistry. Analysis of multiple sets of ligands may provide insight into selectivity. For both virtual exploration and ligand analysis, compounds are transformed into an abstract representation that can be processed by a computer, for instance, 3D minimized structures or molecular graphs. A graph is a mathematical object used to represent a molecule, consisting of nodes, the atoms, which are connected by edges, the bonds. Changing how molecules are represented as graph allows different types of analyses, as demonstrated in chapters 3, 4, and 5. Manipulating the molecular graph is equivalent to modifying the corresponding molecule, and is employed in chapter 6.

1.2.2 Targets

The target family that is central to this thesis is the superfamily of G protein-coupled receptors, or GPCRs. GPCRs are a class of cell surface receptors that consist of seven transmembrane (TM) helices, connected via extra- and intracellular loops. In the human proteome more than 800 known and putative members exist, which are involved in the transduction of a range of stimuli, including small molecules and proteins, and even photons (light).[12] Approximately half of the drugs that recently entered the marked target a GPCR.[13, 14] GPCR drug discovery and design rely heavily on ligand-based approaches since the number of known crystal structures is limited.[13,15] Apart from the bovine rhodopsin structure that was published a decade ago, a handful of 'druggable' receptors has been crystallized to reveal their 3D architecture (for a recent review see ref. 16 and references therein). Some more structures are expected

to be available soon, but that ensemble would still not surpass 1-2 % of the total number of GPCRs. This small number does not even account for the option of co-crystallizing different ligands, which has been done in a few cases now and which would exponentially increase the number of possible structures. The scarcity of structure information does not necessarily have to be problematic though, since ligand-based approaches may even outperform structure-based approaches, e.g. for retrieving known antagonists from virtual libraries, as was previously shown for amine-binding GPCR ligand screening by Evers *et al.*[11] In addition, the absence of suitable crystal structures is compensated for by the wealth of ligand information that is available for GPCRs. A range of ligand-based approaches have been successfully applied for screening and design of new ligands, such as property-based methods, pharmacophore models, and substructure methods.[17, 18, 19] Two of these methods, pharmacophore- and substructure-based methods, will be discussed later in this thesis.

1.3 Cheminformatics

The scientific discipline that covers the approach presented in this thesis is cheminformatics or chemoinformatics, also known as chemiinformatics, chemical informatics, or molecular informatics.[20] Cheminformatics is the application of informatics methods to research questions in the field of chemistry,[21] in particular related to drug discovery and design.[22] It is a relatively new science that combines a number of other disciplines, including computational and medicinal chemistry, and computer and information science. There is still some disagreement over the name; although the 'chemoinformatics' spelling clearly dominated the early literature, it is now overtaken by 'cheminformatics';[20,23] it is also favored by the Journal of Cheminformatics,[24] hence we will use the 'cheminformatics' spelling here. In 1998, the first formal definition of cheminformatics was given by Brown, stating: "The use of information technology and management has become a critical part of the drug

discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization." [25]

The emergence of cheminformatics is easily understood from the vast growth of data that needs to be handled. This 'data deluge' is most apparent in research programs that search for new biologically active molecules, which explains the prominent role of cheminformatics in the pharmaceutical industry.

The primary focus of cheminformatics is on handling chemical information (e.g., compounds, reactions). This ranges from chemical structure representation, to storage and retrieval of chemical information. Note that structure representations facilitate storage and retrieval through chemical file formats or molecular databases. Examples of molecular structure representations are SMILES strings, connection tables, and fingerprints.[26, 27] A SMILES string is a simple line notation of a molecule, readable by both man and computer. Connection tables are tabular representations of the molecular graph that play an important part in chemical file formats such as the Structure Data File (SDF) format. These help the chemist manage his/her chemical data.

Another example of the merits of cheminformatics is the fingerprint. Fingerprints are (short) binary representations that indicate the presence of features in a molecule. These features could be structural features such as fragments, e.g. ring systems or functional groups, or more fuzzy such as atom types, atom pairs, or substructural features. Comparison of fingerprints expresses the similarity between two molecules. When two fingerprints are different, the corresponding molecules are different; however, when two fingerprints are the same, the corresponding molecules are not necessarily identical, although the two are closely related in terms of molecular features. Fingerprints are used for time-efficient identification, searching, and predictions.

Cheminformatics typically performs data mining or statistical analyses on large sets of chemical data; heavy computational techniques, e.g. quantum mechanical calculations, applied to only a few molecules are the realm of computational chemistry, although the two fields also overlap. From a computer science point-of-view, the difference between cheminformatics and bioinformatics is the type of data that is central to each discipline: cheminformatics works with graphs, i.e. molecules, while bioinformatics works with sequences, such as proteins and DNA. In drug discovery, bioinformatics is mainly restricted to the target side, while cheminformatics is relevant for both the target (e.g. pharmacophore elucidation) and the ligand side (e.g. structure-activity relationships). The so-called chemogenomics or proteochemometrics approach combines information from both the target side and ligand side by studying groups of ligands against groups of targets.[28,29] For this, a blend of bioinformatics and cheminformatics techniques is applied.

1.3.1 Databases & data mining

In recent years, a considerable number of publicly available data sources have emerged in the cheminformatics field. Most notable examples are the PubChem[30] database from the National Center for Biotechnology information (NCBI) and the ChEMBL[31] database from the European Bioinformatics Institute (EBI). These databases store small molecule structures together with additional properties and information such as bioactivity and literature references. Other sources of small molecule data with bioactivity are GLIDA and PDSP.[32, 33] These four databases provided the data for the research described in this thesis.

The wealth of small molecule data creates opportunities for applying data mining methods to analyze the molecules and their properties. Data mining is the process of finding patterns from data. It has been applied to diverse problems in a wide variety of fields, for instance weather forecasting, surveillance, or decision making. As another example, data mining is used to gain insight into shopping behavior of customers, suggesting how to improve marketing and increase revenue. Data mining is also a

fundamental part of cheminformatics research to find relationships between selected properties and features of small molecules. These may be used to make predictions about other small molecules. A special type of data mining is graph mining. Molecules are represented as graphs, which are then subjected to graph mining to find interesting patterns. Patterns that are found are subgraphs that occur above a set detection limit. A subgraph is some part of a graph, analogous to a substructure to a molecular structure. The studies described in this thesis apply frequent subgraph mining to analyze GPCR ligands.

1.3.2 Fragments

Molecular fragments are the parts that constitute a molecule. In the context of drug discovery and design, there are two different meanings of 'fragment': chemical fragments and computational fragments. Chemical fragments denote compounds with low molecular weight, typically between 120 and 250 Da. These are used in fragment-based drug discovery, to be screened against a target to identify weak binders.[34] Fragments that bind with low-affinity are subsequently connected or expanded aiming to produce high-affinity ligands. Computational fragments denote discrete parts of the 2D structure of a molecule, such as ring systems, side chains, functional groups.[35,36,37] Fragments can be used to analyze (sets of) molecules to find interesting structural features, for example, privileged structures in the context of GPCR ligands.[19] In this thesis, scaffold analysis was used to evaluate novelty of compounds. Although fragments are intuitive to the chemist, fragment analysis could miss interesting structural patterns. An exhaustive approach is to consider all possible substructures that occur in a molecule. A substructure is any part of a molecular structure, which includes aforementioned fragments.[38] Although this approach is computationally more demanding, it is unbiased. An example of this approach is the frequent substructure mining described in this thesis. Note that in the literature, the meaning of fragment or substructure is sometimes interchanged.[39]

1.4 Aim and Content

In the previous sections, the concept of chemical space has been presented including the challenges it imposes on drug discovery. The cheminformatics field has been introduced together with the target class, GPCRs, and a short primer on fragment and substructure-based approaches. The main objective of the work presented here is to apply substructure-based approaches to aid GPCR drug discovery and design.

Chapter 2 introduces the concepts of substructure-based approaches and provides a general overview of the field. Some of the algorithms used and applications are also reviewed.

Chapter 3 elaborates more on graph mining and its application to study small molecules. It demonstrates how interesting structural patterns are derived from varying sets of GPCR ligands, resulting in both new and characteristic substructures.

Chapter 4 describes a chemogenomics analysis with potential of receptor de-orphanization. It builds upon the concept of characteristic substructures per GPCRs subfamily introduced in chapter 3. A potential application of this method is receptor de-orphanization. While chapter 3 is a descriptive analysis, the method in chapter 4 also has predictive potential.

Chapter 5 discusses ligand-based virtual screening of a commercial vendor library. It is a practical application of the substructure analysis introduced in chapters 3 and 4. Substructures from a set of high-affinity ligands are used to select new ligands from a compound library. The method is validated with experimental test results.

Chapter 6 describes the implementation of a steered search through chemical space to find new ligands. This surpasses selection from known compounds since truly new chemistry is explored. From a collection of fragments, new molecules are generated by modification and combination of the (sub)structures. The design phase is followed up by a synthesis program, yielding active compounds indeed.

All these chapters outline the broader aim of the thesis, which is to apply informatics and computer science methods to medicinal chemistry, in particular to apply fragment/substructure-based techniques to GPCRs drug discovery.

Finally, in chapter 7, the results presented in the thesis are discussed and a future outlook is presented.

1.5 References

(1) Kubinyi, H. Chance favors the prepared mind--from serendipity to rational drug design. *J. Recept. Signal Transduct. Res* **1999**, *19*, 15-39.

(2) Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Current Opinion in Pharmacology* **2009**, *9*, 580-588.

(3) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 374-380.

(4) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* **1996**, *16*, 3-50.

(5) Leiter, D. P.; Morgan, H. L.; Stobaugh, R. E. Installation and Operation of a Registry for Chemical Compounds. *Journal of Chemical Documentation* **1965**, *5*, 238-242.

(6) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Analytica Chimica Acta* **1995**, *314*, 141-147.

(7) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *Journal of Chemical Information and Modeling* **2007**, *47*, 342-353.

(8) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **2009**, *131*, 8732-8733.

(9) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, *46*, 3-26.

(10) Ethan Mollick Establishing Moore's Law. *IEEE Annals 28*, 62-75.

(11) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual Screening of Biogenic Amine-Binding G-Protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. *J. Med. Chem.* **2005**, *48*, 5448-5465.

(12) Fredriksson, R.; Lagerstrom, M. C.; Lundin, L.-G.; Schioth, H. B. The G-Protein-Coupled Receptors in the Human Genome Form Five Main Families. Phylogenetic Analysis, Paralogon Groups, and Fingerprints. *Mol. Pharmacol.* **2003**, *63*, 1256-1272.

(13) Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002**, *3*, 928-944.

(14) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **2006**, *5*, 993-996.

(15) Ballesteros, J.; Palczewski, K. G protein-coupled receptor drug discovery: Implications from the crystal structure of rhodopsin. *Curr Opin Drug Discov Devel.* **2001**, *4*, 561–574.
(16) Salon, J. A.; Lodowski, D. T.; Palczewski, K. The Significance of G Protein-Coupled Receptor Crystallography for Drug Discovery. *Pharmacological Reviews* **2011**, *63*, 901 -937.
(17) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Property-Based Design of GPCR-Targeted Library. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332-1342.
(18) Chang, L. C. W.; Spanjersberg, R. F.; von Frijtag Drabbe Künzel, J. K.; Mulder-Krieger, T.; van den Hout, G.; Beukers, M. W.; Brussee, J.; IJzerman, A. P. 2,4,6-Trisubstituted Pyrimidines as a New Class of Selective Adenosine A1 Receptor Antagonists. *J. Med. Chem.* **2004**, *47*, 6529-6540.
(19) Bywater, R. P. Privileged Structures in GPCRs. In *GPCRs: From Deorphanization to Lead Structure Identification*; Ernst Schering Foundation symposium proceedings; 2006; Vol. 2, pp 75-92.
(20) Willett, P. A bibliometric analysis of the literature of chemoinformatics. *Perspectives* **2008**, *60*, 4–17.
(21) Gasteiger, J. Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem* **2005**, *384*, 57-64.
(22) Hann, M.; Green, R. Chemoinformatics -- a new name for an old problem? *Current Opinion in Chemical Biology* **1999**, *3*, 379-383.
(23) Molinspiration Cheminformatics or Chemoinformatics ? *Cheminformatics or Chemoinformatics ?* **2010**.
(24) Chemistry Central Journal of Cheminformatics. *Journal of Cheminformatics*.
(25) Brown, F. K. Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery.; Annual Reports in Medicinal Chemistry; Academic Press, 1998; Vol. Volume 33, pp 375-384.
(26) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.
(27) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31-36.
(28) Doddareddy, M. R.; Westen, G. J. P. van; Horst, E. van der; Peironcely, J. E.; Corthals, F.; Ijzerman, A. P.; Emmerich, M.; Jenkins, J. L.; Bender, A. Chemogenomics: Looking at biology through the lens of chemistry. *Statistical Analysis and Data Mining* **2009**, *2*, 149-160.
(29) Westen, G. J. P. van; Wegner, J. K.; IJzerman, A. P.; Vlijmen, H. W. T. van; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.* **2010**, *2*, 16-30.
(30) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway,

M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2008**, *36*, D13–D21.

(31) de Matos, P.; Alcantara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S.; Steinbeck, C. Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.* **2010**, *38*, D249-254.

(32) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. *Nucl. Acids Res.* **2008**, *36*, D907-912.

(33) Roth, B. L.; Lopez, E.; Beischel, S.; Westkaemper, R. B.; Evans, J. M. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacology & Therapeutics* **2004**, *102*, 99-110.

(34) Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Current Opinion in Biotechnology* **2006**, *17*, 643-652.

(35) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.

(36) Xu, J. A New Approach to Finding Natural Chemical Structure Classes. *J. Med. Chem.* **2002**, *45*, 5311-5320.

(37) Nilakantan, R.; Nunn, D. S.; Greenblatt, L.; Walker, G.; Haraki, K.; Mobilio, D. A Family of Ring System-Based Structural Fragments for Use in Structure-Activity Studies: Database Mining and Recursive Partitioning. *J. Chem. Inf. Model.* **2006**, *46*, 1069-1077.

(38) Nijssen, S.; Kok, J. N. Frequent graph mining and its application to molecular databases. In *Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics*; 2004; Vol. 5, pp 4571-4577.

(39) Borgelt, C.; Berthold, M. R.; Patterson, D. E. Molecular Fragment Mining for Drug Discovery. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*; Lecture Notes in Computer Science; Springer Berlin / Heidelberg: Barcelona, Spain, 2005; Vol. 3571, pp 1002-1013.