Computers and drug discovery : construction and data mining of chemical and biological databases

Kazius, J.

Chapter 1

Introduction

In this first chapter, the disciplines of bioinformatics and chem(o)informatics are introduced. These fields serve to manage and analyse biological and chemical information. Several features between the two overlap, although each discipline also knows many specific methods. We will then discuss how the two disciplines contribute to the general aim of the thesis. This aim is to further develop and employ methods from computer science for a better understanding of the biological and chemical causes of adverse effects in man, such as toxic effects that are introduced by genetic variants or potential drugs.

The title of this thesis overlaps with two multi-disciplinary scientific disciplines: bioinformatics and chem(o)informatics. In this first chapter, we will introduce both disciplines. Many definitions exist for either bio- or chem(o)informatics. In general, bioinformatics encompasses the research, development, or application of computational approaches to manage or analyse biological information. Similarly, cheminformatics, or chemoinformatics, covers the research, development, or application of computational approaches to manage or analyse chemical information. Though these disciplines have grown considerably over the last years to become very broad, the need for new and accurate computational methods grows even faster. In this chapter, we will discuss the general aims and origins of these disciplines, conceptually compare them, and select topics that warrant further research.

General aims of the disciplines

Despite near-identical definitions, the goals of the two fields appear different. The most commonly analysed entities of either field are biological sequences in bioinformatics and chemical compounds in cheminformatics. One of the popular aims of bioinformatics analyses is to elucidate the biological function of newly detected sequences of genes or proteins. Another, related aim is to identify which genes, or genetic variants, affect observed phenotypes or respond to environmental changes. The engineering of proteins towards particular biological effects is, interestingly, only a minor field in bioinformatics, where little attention is paid to hypothetical sequences. Rather, the many sequences and sequence mutations that occur in a variety of organisms are studied to further understand their impact on biological processes. In contrast, cheminformatics efforts have hardly focused on the biological effects of compounds, such as natural products, that surround us. Even the detailed effects of the thousands of compounds that we digest through our food receive relatively little attention. Instead, the ultimate aim of cheminformatics is to engineer novel chemical structures that show a combination of desirable effects, such as drugs with their effects on health. Cheminformatics-based predictions are therefore often made for virtual, hypothetical compounds that have not (yet) been bought or synthesised.

The aims of the disciplines are nevertheless complementary for several purposes. For one, both can aid our understanding of adverse biological effects like diseases, with the ultimate goal to prevent or cure them. The discovery process of new pharmaceuticals is the clearest example where bioinformatics and cheminformatics are used in concert[1]. Once disease-related functions of one or more proteins have been elucidated by experimental means and the aid of bioinformatics, a search can start for chemicals that bind these proteins and intervene with their biological effects, now with the aid of cheminformatics.

Origins of bioinformatics

Bioinformatics - terminology

The term bioinformatics was first introduced in 1988 by Paulien Hogeweg , who published the term to describe "the study of informatic processes in biotic systems"[2]. Around that time, Hwa Lim also proposed bioinformatics as "a collective term for data compilation, organisation, analysis and dissemination", although no publication exists to confirm this claim. Irrespective of who first defined the word, researchers were already performing bioinformatics, with a particular focus on sequence analysis. For most applications, proteins and genes are concisely represented by their sequence through a string of single letter codes for their nucleotides or amino acids, as depicted in Figure 1.1. Though a protein in its biological environment is much more than its sequence alone as they are mostly large and complex molecules of particular shapes, the value of sequence information in bioinformatics is well-established[3], and relatively cheap to determine.

Bioinformatics - algorithms

Well before the introduction of the term bioinformatics, algorithms by Needleman and Wunsch[4] and software[5] by Staden were already developed in the 1970's to represent, analyse and compare biological sequences. In 1981, Doolittle[6] first recognised the concept that a small part of a protein sequence, called a sequence motif, could have functional importance: motifs can recognise sites of biological importance, such as those that interact with proteins or small molecules. In the same year, Smith and Waterman proposed their algorithm for the comparison and alignment of multiple sequences[7]. Another algorithmic breakthrough in the field of bioinformatics was the development in 1990 of a fast algorithm for the pair-wise alignment of a new sequence to a large database of sequences by Altschul et al.[8]

```
1 ALPAVNTS.GFTPQR
2 ALPGTNASAGFNGQR
3 .LPGTNTTGGFTNQR
```

Figure 1.1.
Alignment of three short hypothetical amino acid sequences. A N-X-S/T-motif is shown in bold in each sequence. This motif finds sites for potential N-glycosylation. Where N identifies only asparagine, X stands for any amino acid and S/T identifies either serine or threonine.

Bioinformatics - databases

Many types of biological data have been determined over the years. Remarkably, the first collection of biological data that was successful enough to become the standard online resource contained a very complex type of biological information. In 1971, the Protein Data Bank was initiated to contain three-dimensional structures of biological macromolecules like peptides or proteins, sometimes in the presence of a small molecule (e.g., an enzyme inhibitor) or a piece of DNA[9]. In the 1980's, DNA sequences began to be stored and publicly distributed via the Entrez databases Nucleotide and GenBank[10]. Protein sequences were

similarly stored in the manually curated Swiss-Prot database that originated in 1991[11]. When the value of such databases for future research was well established, scientific journals started to require online submission of structures or sequences to these databases prior to publication. In the 1990's, disease-related and naturally occurring mutations in genetic sequences were beginning to be deposited in OMIM[12] and dbSNP[13], respectively. From the year 2000, other databases were developed to contain new additional biological data, such as gene functionality[14] and gene and protein expression profiles[15]. All these biological data are currently available at a large scale thanks to the development of the internet and the 'free data' mentality of the researchers in the field.

Numerous efforts in terms of algorithms and databases have broadened the definition of bioinformatics such that it has become an umbrella term for almost any computational method that involves biological information. By now, scientific articles that are related to bioinformatics account for a significant fraction, about 2%, of all papers in the biomedical literature from the PubMed database[16].

Origins of cheminformatics

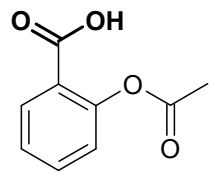
Cheminformatics - terminology

As late as in 1998, Brown was the first to define the term chemoinformatics. He stated: "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."[17]. The word cheminformatics also rapidly appeared to identify the same field[18]. Currently, equal amounts of papers in PubMed use one word or the other (cheminformatics vs chemoinformatics). A Google search shows that cheminformatics is over 50% more popular and therefore we will use this term in the remainder of the thesis. Numerous chemical principles that were later incorporated in cheminformatics methods had been around for a long time before computers were invented. For one, Crum-Brown and Fraser concluded in 1863 that the physiological actions of compounds were related to their chemical constitution[19]. In the 1930's, also before computers were available, Hammett developed parameters that enabled him to mathematically model rates of chemical reactions[20]. As was the case with bioinformatics, researchers were already using computers to store and process chemical information long before the term cheminformatics was coined, even as early as the 1950's, as will be discussed below[18].

Cheminformatics - representation

In cheminformatics, the input examples are mostly chemical structures. Computers can consider compounds as mathematical graphs that consist of atom types (C, N, etc.) which are connected by single, double or triple bonds[21] such that branched and cyclic shapes occur in a chemical structure. Linear, single-line notations were developed to represent chemical structures even before they were used by computers. The oldest of these is the WLN (Wiswesser Line Notation), of which the initial version was developed in 1949[22]. Amongst the other linear notations that followed[18], the SMILES (Simplified Molecular Input Line Entry Specification) notation, which was developed in 1988 by Weininger[23], is now the most widely used linear notation. With the use of computers to mathematically represent compounds as graphs[24], new formats were developed to readily depict structure diagrams of the chemical structure. The example in Figure 1.2 shows a traditional two-dimensional structure diagram of the chemical structure of caffeine (a), and its SMILES notation (b), both without explicit hydrogens. These representation formats could, moreover, contain three-dimensional information to capture a particular conformation of a compound, thus further standardising the representation of chemicals.

a)



b) **O=C(O)**c1ccccc1OC(=O)C

Figure 1.2.
Molecular structure (a) and SMILES notation (b) of aspirin, with its carboxylic acid substructure in bold. Each letter in the SMILES notation represents a heavy atom, where capital and small letters signify aliphatic and aromatic atoms. In a SMILES string, subsequently placed atoms are connected via single/aromatic bonds, as are the two aromatic carbons that are labelled with 1. By following the atoms in a SMILES string, the whole molecular structure can be traversed.

Cheminformatics - algorithms

One of the earliest, ground-breaking algorithmic uses of computers to analyse chemical structures was to perform *substructure searches*, that is, to find compounds that posses a particular predefined substructure through an algorithm from graph theory that was already developed in 1957[25]. A substructure can be any part of a molecule, varying from a single atom to a linked combination of atoms. Halfway in the 1960's, Hansch, et al. used Hammett's parameters to computationally relate the biological activity to the chemical structures of a series of closely related compounds by deriving *(Quantitative) structure-activity relationships ((Q)SARs)*[26, 27], a popular research line for pharmaceutical companies[28]. In 1965, Morgan of the Chemical Abstracts Service (CAS) corrected and broadened an algorithm proposed by Gluck to generate unique machine descriptions of chemical structures, which is useful for cleaning chemical databases and rapid performing *full structure*

searches[29]. In 1969, an automated method was proposed for the comparison of a compound with an arbitrary set of compounds via a process called *similarity searching*[30]. In the 1970's, the first method originated to assign values for biological activities to substructures[31]. In the 1980's, computational advances by Willett et al. in three-dimensional queries enabled so-called *pharmacophore searching*, which searched a database of compounds for those that contained predefined electronic and steric effects in a particular orientation[32]. Also in the 1980's, Klopman and Rosenkranz initiated the first form of *substructure mining* by computing all linear substructures from a set of chemicals that possessed a toxic activity called mutagenicity, that is, reactivity towards human DNA[33]. Over time, these algorithms and many others have been significantly enhanced to more rapidly analyse larger datasets and to consider more detailed chemical information.

Cheminformatics - databases

Many types of chemical data have been determined and stored. In the 1950's, the Chemical Abstracts Service[34] was one of the first to store two-dimensional chemical structures of published chemicals in their commercial database. Another commercially distributed database, called the Cambridge Structural Database, was developed to contain three-dimensional crystal structures of conformations of compounds and it was developed in 1965[35]. As already noted, the Protein Data Bank started in 1970's with structures of peptides and proteins and compounds that were bound to them was made publicly available[9]. Many chemical databases were digitalised in the 1980's such that their chemical structures could be searched by a computer. Examples are databases that store chemical reactions[36], the Beilstein database with facts from the scientific literature[37] and patent databases with (series of) chemicals[38]. In 1997, the National Cancer Institute built and publicly distributed their database with compounds a associated biological anti-tumor data [39], which was the largest freely available database for almost a decade. Only very recently did PubChem become the largest source of chemical structures[40], many of which have data on a number of biological activities. A larger freely available database of around 3 million chemical structures, but without biological activities, was the recently developed ZINC database (ZINC Is Not Commercial)[41].

Progress in algorithms and databases has also extended the term cheminformatics such that it now encompasses numerous computational methods that involve chemical information. As a result, Gasteiger and Engel recently defined it as "the application of informatics methods to solve chemical problems"[1].

Overlaps between bioinformatics and cheminformatics

Overlap exists between the two disciplines. Both cheminformatics and bioinformatics have two sides: a theoretical one that aims to invent or advance underlying algorithms and an

applied side that aims to enhance the interpretation and knowledge discovery of experimentalists by using existing methods.

Both disciplines can involve the use of data from robotised experiments or other sources to design, construct and clean new databases. Databases can be built to facilitate electronic access to their content, and thus facilitate searches. As the amount of data grows, so does the expectation that they can be used to extend upon existing knowledge, for instance, via a process called data mining[42].

During data mining, databases are analysed computationally for the purposes of searching for examples that are similar to a starting example, clustering related examples and/or extracting relationships such as those between simple properties of the studied examples and their complex biological effects. Data mining is particularly useful in the analysis of large databases as they can be difficult to comprehend manually. A general aim of data mining is to find a workable set of potentially novel relationships that are causal (predictive), accurate and understandable. The extracted relationships can be used to make educated guesses for the biological effects of new test cases. Given a set of sequences or compounds, that do and do not possess a particular biological activity, the simplest result of a data mining campaign would be a motif or substructure that is responsible for the activity. On the other hand, a result of a data mining study can also be an abstract mathematical model that can distinguish between active and inactive examples, but that does not provide any insight into the underlying causes.

Cheminformatics or bioinformatics can give or refine hypotheses, but they will never give definitive answers for untested compounds or sequences: experiments are always needed to validate the suggested hypotheses and predictions. This is why data mining algorithms should not only derive a method that is predictive, but they should also derive relationships that are interpretable to chemists/biologists. They are the ones that need to be convinced as to why they should perform the suggested chemical or biological research instead of pursuing their own plans. While theoreticians may prefer the most ‘accurate’ method, experimentalists may prefer intuitive methods that give interpretable and workable relationships, also at a cost in accuracy. One prerequisite for deriving such interpretable relationships is that the examples (compounds or sequences) are represented by descriptors, or features, that are meaningful. In both disciplines, the value of computational methods lies in the derivation of causal relationships between intuitive descriptors and biological effects, that is, between simple causes and complex effects. These relationships can then prioritise resource-intensive experiments and thereby increase efficiency of hypothesis-driven research.

Differences between bioinformatics and cheminformatics

Numerous data mining techniques and statistical tests are useful to both disciplines, but only if these methods are not data-specific. However, most approaches in bioinformatics and cheminformatics are just that: data-specific. This difference results from how examples (such as sequences or chemical structures) are represented by descriptors, often also called properties or patterns. Such a representation is needed before data mining techniques can detect which of these descriptors relate to biological activity. These descriptors should be able to highlight chemical or biological similarities between the examples. The similarity principle[43] states that examples are more likely to exert similar biological effects it they are more similar in terms of their descriptors. In general, data mining methods assume that these descriptors are independent of each other. If, however, the order between the descriptors is important, non-standard techniques are needed. This will be explained in the following paragraphs.

In cheminformatics, compounds can also be represented via simple, computable descriptors, such as molecular weight, the number of rings and the number of atoms of each type. However, such a description is, of course, a considerable simplification compared to the two-dimensional chemical structure that connects atoms via bonds. Such simple descriptors are therefore not always relevant for biological activity. In addition, to describe the effects of compounds experimental values can be determined or even predicted at relatively low cost, such as the common quantification of hydrophobicity, *log*P. Chemical structures can, moreover, be analysed directly by determining which substructures they contain. The presence or absence of a substructure can also be regarded as a descriptor. A catalogue exists of over a thousand types of categorical and/or numerical descriptors[44], many of which are not intuitive to chemists, let alone relevant for biological activity. Each compound can thus be represented as a - possibly arbitrary - set of values for these descriptors. Several of such descriptor values can correlate with each other because the descriptors represent overlapping properties, such as size and weight. Nevertheless, each chemical descriptor has a particular meaning and can thus be determined independently of another descriptor, and no relevant order exists between them. Because no important information is lost when each descriptor value is analysed individually, most general data mining methods can be applied in cheminformatics to determine relationships between these chemical descriptors and biological effects.

In bioinformatics, theoretical descriptors of sequences can be order-free, such as experimentally determined properties or the quantity of each amino acid that is present per sequence. Such simple and global descriptors, however, are neither insightful nor likely to relate to biological effects in practice. Rather, as sequences are ordered stretches of nucleotides or amino acids, the most useful patterns that can be derived from them are ordered groupings of these elements, such as motifs. In a sense, motifs are also used in methods of

sequence alignment, but these techniques also consider the order between motifs. Although motifs can represent an ordered (sub)sequence by themselves, descriptors that describe the presence or absence of motifs are independent of each other. General data mining methods can still be applied to these independent motif descriptors, but they then ignore the order between motifs. As a result, highly specific methods, like those for alignment, are employed more often in bioinformatics analyses because the order between descriptors is also very important.

Despite these differences between the methods for representation and analysis in either discipline, most data mining projects aim to answer the same general questions: what knowledge can we derive from a given dataset? Can we find simple descriptors that relate to biological activities? Can the extraction of such knowledge be automated, and if so, how? The successful extraction of knowledge from example datasets has a second value by encouraging similar searches for simple chemical or biological descriptors that underlie other biological effects.

Focus in bioinformatics: GPCR databases and variants with adverse effects

Of the protein families that exist in the human genome, the superfamily of G protein-coupled receptors (GPCRs) is of particular importance to the pharmaceutical industry[45]. GPCRs function by specifically recognising chemical signals from the extracellular environment of a cell and communicate this information to the inside of this cell. By transducing extracellular signals over cell membranes, GPCRs regulate a plethora of important physiological processes and they are frequently targeted by drug design efforts[46]. Moreover, a particular diversity of harmful and harmless genetic variants occurs in these GPCRs. Due to the importance of the superfamily of GPCRs, molecular biologists and pharmacologists that research these proteins have been supported by public, GPCR-oriented databases. One of these databases, the GPCRDB, has been around for over a decade and contains a categorisation of GPCRs into subfamilies along with their protein sequences, alignments thereof, and information on GPCR ligands[45]. Among several other databases that focus on GPCRs is the olfactory receptor database, which supplements GPCRDB-like information with a knowledge-based classification of these receptors of smell[47]. Moreover, the tinyGRAP database has been set up to encompass the many artificially introduced GPCR mutations that were described in the scientific literature and studied for their effects on GPCR functioning and ligand binding *in vitro*[48]. This effort has recently been supplemented by MuteXt, which automatically extracts information on artificially introduced GPCR mutations from electronic papers and provides its results through the GPCRDB[49]. Numerous natural variants occur throughout the human genome, including at GPCR genes and these have not yet been catalogued in a database, although such a database could further facilitate research into GPCRs. Several genetic variants in GPCRs are responsible for human diseases by

distorting GPCR function. In contrast, many natural variants that have been elucidated from healthy human volunteers have not been linked to adverse effects. Several mutations that have strong adverse effects on human health have been stored in different source databases, but the majority of human health data on GPCR variants is present only in scientific articles. GPCR researchers would therefore be greatly aided by a database on GPCR variants that also captures variants from the scientific literature, especially when several automated error-checks filters out incorrect data. Moreover, such a database could help to understand why and how several genetic GPCR variants cause various adverse effects like diseases while other GPCR variants appear harmless.

Focus in cheminformatics: structure-activity relationships for adverse effects

In the early days of cheminformatics, quantitative structure-activity relationships (QSARs) were derived for series of closely related molecules[26]. The outcome of such calculations would ideally help in the prediction or rather suggestion of new, analogous compounds with an improved activity profile. Today's perspective is largely different. While many biological data have been available for a long time, cheminformaticians have recently been overwhelmed by new, large databases of compounds with available biological activities[40, 50], which resulted from a flood of results from public high-throughput biological screening efforts. A typical example are the thousands of very diverse compounds that are assessed for their inhibition of the so-called hERG ion channel in the heart[40], which causes adverse effects like cardiac arrhythmia[51]. Such large, diverse chemical datasets on adverse biological effects are impossible to analyse by traditional QSAR methods. In fact, the current data explosion further stresses the need to automatically relate biological effects to chemical structure. New cheminformatics methods are therefore needed to better identify chemical causes for various types of biological activities. More specifically, computational methods that aim at extracting structure-activity relationships (SARs) should incorporate more sophisticated properties of atoms and non-linear shapes during the search of biologically relevant substructures.

Aim and content

The previous sections of this chapter have outlined the general purposes of bioinformatics and cheminformatics, their origins, and the overlaps and differences between these fields. The last sections have focussed on the more specific problems that are addressed in this thesis. In short, the goal of the thesis is to understand biological and chemical causes for adverse effects in man. More specifically, we employ methods from computer science to extract the causes of several toxic effects that are introduced by genetic variants or potential drugs. The ideal result would be that such automatically extracted biological and chemical knowledge also proves useful for predictive purposes.

Genetic variants can only have the potential to cause adverse effects like diseases if they affect the genes of important proteins such as G protein-coupled receptors (GPCRs). A large variety of GPCR variants have been determined and only some of them cause diseases while others lack an obvious effect on phenotype. GPCRs are introduced in Chapter 2, as are several natural variants in GPCR genes and their links to disease.

As introduced above, numerous natural variants occur throughout the human genome, including at GPCR genes and these have not yet been catalogued in a database. To raise awareness about natural variants in GPCRs and to facilitate access by GPCR researchers to variants and their effects on health, Chapter 3 describes the development and content of a database on GPCR variants and their effects.

One advantage of examining the variants of one superfamily of related proteins, rather than a set of unrelated proteins, is that additional descriptors can be employed during data mining, such as more detailed family-specific descriptors. Several features of variants that correlated with the adverse effects of GPCR variants on health are revealed in the statistical analysis of Chapter 4. Knowledge of the disease potential of these variants also sheds insight into GPCR domains that are critical for normal receptor activity, as mutations herein often distort GPCR function.

Like Chapter 4, the next part of this thesis concerns the extraction of knowledge for adverse biological effects. The following chapters, however, focus on chemical problems: which chemical descriptors cause adverse biological effects? Which methods can be used to determine such chemical descriptors? To address the latter, different techniques of data mining and the validation of their results are reviewed in Chapter 5 by using examples from toxicity prediction.

Considerable costs can be saved when adverse effects of candidate pharmaceuticals are detected in an early - rather than a late - stage of drug discovery research, as their toxicity may hinder their therapeutic potential. Reliable predictive methods can help prioritise compounds for initial and more extensive toxicity screening to minimize late stage rejections and prevent expensive screens early on. For this reason, the study in Chapter 6 derived substructures that are likely to cause mutagenicity according to statistical findings. Moreover, these substructures were only considered valid when literature supported a mechanism of action.

Fully computational methods can analyse large datasets more rapidly and they are not limited to personal expertise on the investigated biological activity. Current substructure mining methods that have been used for cheminformatics problems consider only substructures that are linear and/or of limited size[52-54]. If, however, the compound graphs consist of only these simple elements, such as C, N, etc., then so do their substructures. Although the resulting substructures are interpretable to chemists, substructure mining methods to date have neglected information that chemists can readily infer from a molecular

structure, like aromaticity[52-54]. Chapter 7 aims to overcome these drawbacks and reports to what extent the substructures found in Chapter 6 could be reproduced by a computational technique that derives all linear, branched and cyclic substructures that occur in chemical structures.

This thesis concludes with general conclusions on the bioinformatics and cheminformatics approaches that are described in this thesis. Furthermore, perspectives for future research are put forward with a final focus on newly available data and the synergy between cheminformatics and bioinformatics approaches.

REFERENCES

1. Gasteiger, J. Introduction. In *Chemoinformatics : A Textbook*, Gasteiger, J.; Engel, T., Eds. John Wiley & Sons: 2004; pp 1-13.
2. Hogeweg, P. MIRROR beyond MIRROR, puddles of LIFE. In *Artificial Life*, Langton, C. G., Ed. Addison Wesley: 1998; pp 297-316.
3. Mount, D. W. What is bioinformatics? In *Bioinformatics: Sequence and Genome Analysis*, Mount, D. W., Ed. 2004; p 7.
4. Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **1970,** 48, 443-53.
5. Staden, R. Sequence data handling by computer. *Nucleic Acids Res* **1977,** 4, 4037-51.
6. Doolittle, R. F. Similar amino acid sequences: chance or common ancestry? *Science* **1981,** 214, 149-59.
7. Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **1981,** 147, 195-7.
8. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **1990,** 215, 403-10.
9. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R., et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* **1977,** 80, 319-24.
10. Keller, C.; Corcoran, M.; Roberts, R. J. Computer programs for handling nucleic acid sequences. *Nucleic Acids Res* **1984,** 12, 379-86.
11. Bairoch, A.; Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **1991,** 19 Suppl, 2247-9.
12. Pearson, P.; Francomano, C.; Foster, P.; Bocchini, C.; Li, P.; McKusick, V. The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res* **1994,** 22, 3470-3.
13. Sherry, S. T.; Ward, M.; Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **1999,** 9, 677-9.
14. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000,** 25, 25-9.
15. Sherlock, G.; Hernandez-Boussard, T.; Kasarskis, A.; Binkley, G.; Matese, J. C.; Dwight, S. S., et al. The Stanford Microarray Database. *Nucleic Acids Res* **2001,** 29, 152-5.
16. Luscombe, N. M.; Greenbaum, D.; Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* **2001,** 40, 346-58.
17. Brown, F. K. Chemoinformatics: What is it and How does it Impact Drug Discovery? *Annu. Rep. Med. Chem.* **1998,** 33, 375-384.
18. Engel, T. Basic overview of chemoinformatics. *J Chem Inf Model* **2006,** 46, 2267-77.
19. Crum-Brown, A.; Fraser, T. R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia. *Trans. Roy. Soc. Edinburgh* **1868–1869,** 25, 151–203.
20. Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935,** 17, 125-136.
21. Bonchev, D.; Rouvray, D. H. *Chemical Graph Theory: Introduction and Fundamentals*. Gordon and Breach Science Publishers: 1990.
22. Wiswesser, W. J. How the WLN began in 1949 and how it might be in 1999. *J. Chem. Inf. Comput. Sci.* **1982,** 22, 88-93.

23. Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988,** 28, 31-36.
24. Balaban, A. T. *Chemical Applications of Graph Theory*. Academic Press: London, 1976.
25. Ray, L. C.; Kirsch, R. A. Finding Chemical Records by Digital Computers. *Science* **1957,** 126, 814-819.
26. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962,** 194, 178-180.
27. Hansch, C.; Fujita, T. F. ρ-σ-π Analysis; Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964,** 86, 1616-1626.
28. Boyd, D. B. Drug Design. In *The Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F.; Schreiner, P. R., Eds. J. Wiley & Sons: Chichester, 1998; pp 795-804.
29. Morgan, H. L. T. he Generation of Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Services. *J. Chem. Doc.* **1965,** 5, 107-113.
30. Armitage, J. E.; Lynch, M. F. Automatic Detection of Structural Similarities among Chemical Compounds. *J. Chem. Soc. (C)* **1967**, 521-528.
31. Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1973,** 17, 533-535.
32. Brint, A. T.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Graphics*. **1987,** 5, 49-56.
33. Klopman, G.; Rosenkranz, H. S. Structural requirements for the mutagenicity of environmental nitroarenes. *Mutat Res* **1984,** 126, 227-38.
34. Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *J. Chem. Inf. Comput. Sci.* **1979,** 19 94-98.
35. Allen, F. H.; Hoy, V. J. Cambridge Structural Database. In *The Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F.; Schreiner, P. R., Eds. J. Wiley & Sons: Chichester, 1998; pp 55-167.
36. Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of AAM and Related Reaction Features in the Reaction Access System (REACCS). In *Chemical Structures*, Warr, W. A., Ed. Springer-Verlag: Berlin, 1988; pp 303-313.
37. Jochum, C. B. Building a Structure-Oriented Numerical Factual Database. In *Chemical Structure;* , Warr, W. A., Ed. Springer-Verlag: Berlin, 1988; pp 187-193.
38. Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy. *J. Chem. Inf. Comput. Sci.* **1981,** 21, 148-150.
39. Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W., et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997,** 275, 343-9.
40. Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2006,** 34, D173-80.
41. Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** 45, 177-182.

42. Witten, I. H.; Frank, E. Data mining and machine learing. In *Data mining: practical machine learning tools and techniques with Java implementations*, Gray, J., Ed. 2000; pp 2-7.
43. Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*. John Wiley & Sons: New York, 1990.
44. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, 2000; p 690
45. Horn, F.; Bettler, E.; Oliveira, L.; Campagne, F.; Cohen, F. E.; Vriend, G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* **2003,** 31, 294-7.
46. Hopkins, A. L.; Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **2002,** 1, 727-730.
47. Crasto, C.; Marenco, L.; Miller, P.; Shepherd, G. Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res* **2002,** 30, 354-60.
48. Beukers, M. W.; Kristiansen, I.; AP, I. J.; Edvardsen, I. TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol Sci* **1999,** 20, 475-7.
49. Horn, F.; Lau, A. L.; Cohen, F. E. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* **2004,** 20, 557-68.
50. Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S., et al. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* **2007,** 18, 18.
51. Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006,** 440, 463-9.
52. Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comput Sci* **2004,** 44, 1402-11.
53. Klopman, G.; Srivastava, S.; Kolossvary, I.; Epand, R. F.; Ahmed, N.; Epand, R. M. Structure-activity study and design of multidrug-resistant reversal compounds by a computer automated structure evaluation methodology. *Cancer Res* **1992,** 52, 4121-9.
54. Malacarne, D.; Pesenti, R.; Paolucci, M.; Parodi, S. Relationship between molecular connectivity and carcinogenic activity: a confirmation with a new software program based on graph theory. *Environ Health Perspect* **1993,** 101, 332-42.