promoting access to White Rose research papers

White Rose
Research Online

Universities of Leeds, Sheffield and York
http://eprints.whiterose.ac.uk/

This is an author produced version of a paper published in **Annual Review of Information Science and Technology.**

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/77605

Published paper

Willett, P (2009) *Similarity Methods in Chemoinformatics.* Annual Review of Information Science and Technology, 43. 3 - 71.

White Rose Research Online
eprints@whiterose.ac.uk

Published as: Willett, P. “Similarity methods in chemoinformatics.” *Annual Review of Information Science and Technology*, **43**, 2009, 3-71.

Similarity Methods in Chemoinformatics

PETER WILLETT
University of Sheffield

INTRODUCTION

Many information-processing techniques are applicable across a wide range of academic disciplines with little or no modification. There are, however, some types of processing that are associated with a specific discipline that makes use of a particular type of data or representation and that has particular types of information need. This is becoming more common as increasing amounts of data become available in digital form and as informatics techniques are developed to store, search and rationalise the resulting data archives. An obvious example of such a discipline is molecular biology, where a range of bioinformatics techniques are available for processing biological data, initially protein and nucleic acid sequences but now also phylogenetic trees, metabolic pathway maps and protein three-dimensional (3D) structures *inter alia* (Lesk, 2005; Orengo, Thornton, & Jones, 2002); geography is another such discipline, where geographic information systems provide specialist facilities for the processing of cartographic data of all sorts (Longley, Goodchild, Maguire, & Rhind, 2005; Worboys, 1995). Chemistry, the subject of this review, provides a further example since provision needs to be made for the processing of two-dimensional (2D) and 3D representations of molecule structure. The techniques that are used are of widespread applicability since molecules - either traditional small molecules or, increasingly, biological macromolecules - play a key role across the physical and life sciences, and in many modern industries (agrochemicals, flavours and fragrances, food science, materials, petrochemistry, and – most importantly in the context of this review - pharmaceuticals).

The importance of the chemical structure has meant that it has formed the focus of four previous ARIST reviews, these appearing approximately once a decade; the processing of non-structural chemical information is described by Bottle and Rowland (1993) and by Maizel (1998). The first ARIST review (Tate, 1967) appeared shortly after the creation of the Chemical Abstracts Service (CAS) Registry System, the principal source for information about the chemical molecules reported in the world's chemical and related literatures (Leiter, Morgan, & Stobaugh, 1965; Shively, 2007; Weisgerber, 1997). The second appeared eleven years later (Rush, 1978), by which time the basic techniques for the representation and substructure searching (*vide infra*) of databases of 2D structures were well established, as were the first operational systems, both public and corporate (Ash & Hyde, 1975). A further eleven years passed before the appearance of the third review (Lipscomb, Lynch, & Willett, 1989), the intervening period having seen the extension of 2D techniques to encompass similarity and Markush searching (*vide infra*) and the first reports of systems for 3D substructure searching. The latter had become well established by the time that the most recent review was reported by Paris (1997), this review also including early work on the analysis of molecular diversity (*vide infra*). As the focus of this review is the processing of 2D and 3D chemical molecules and as techniques for this are by now well established, the reader may be querying the need for a further review of the subject. That there is such a need is perhaps best demonstrated by the fact that the title of this review contains a word – chemoinformatics – that did not even exist until 1998. And yet within just a few years chemoinformatics has come to be recognised as playing a key role in chemical research, and has spawned the publication of three textbooks (Bunin, Bajorath, Siesel, & Morales, 2007; Gasteiger & Engel, 2003; Leach & Gillet, 2003) and the introduction of several specialist academic programmes (Schofield, Wiggins, & Willett, 2001; Wild & Wiggins, 2006).

THE EMERGENCE OF CHEMOINFORMATICS

The recent recognition of chemoinformatics has come about principally as a result of changes in two of the basic technologies that underlie drug discovery, which is currently the most important application domain for chemoinformatics. The pharmaceutical industry has traditionally discovered new drugs by a time-consuming and costly process that involved medicinal chemists synthesising novel molecules that were then tested by biologists to see whether the molecules exhibited beneficial therapeutic properties (Lombardino & Lowe, 2004). This synthesise-and-test procedure is in two parts: lead-discovery involves identifying a molecule, called the *lead*, with the desired biological activity; and lead-optimisation involves identifying that member of the lead's chemical class, called the *candidate*, that has the best combination of activity, side-effects, synthetic feasibility, and ease-of-delivery. The candidate molecule is then passed on for testing in patients during three phases of clinical testing, each more rigorous and more time-consuming than the previous one (Ekins, 2006; Rang, 2006). Only once these tests have been carried out successfully can a pharmaceutical company obtain a license from a regulatory authority, such as the Food and Drug Administration in the USA or the Medicines and Healthcare Products Regulatory Agency in the UK, to market the molecule as a commercial product. More than a decade can pass between the start of a new research programme and the launch of a new drug, meaning that drug discovery is very expensive. The costs of drug discovery have been estimated in a much-cited (and much discussed) paper (DiMasi, Hansen, & Grabowski, 2003) that quoted a figure of $802 million (based on 2000 prices) as the mean research and development cost when averaged over 68 randomly selected new drugs from ten pharmaceutical firms.

Computers have been used for many years in all aspects of pharmaceutical research and development to increase the cost-effectiveness of drug discovery (Boyd & Marsh, 2006). Technological developments in the early Nineties suggested that it might be possible to reduce significantly the time (and hence the cost) required for the lead-discovery and lead-optimisation stages. Specifically, *combinatorial chemistry* is the name given to a body of techniques that allow large numbers of molecules to be synthesised in parallel. Assume, for example, that a chemist wants to react an acid with an amine to create an amide, and that multiple amides need to be evaluated. The conventional approach would be to react one acid with one amine to produce one amide, and then to repeat the process with sequences of different acids and amines. Combinatorial chemistry allows hundreds of different acids to be reacted with hundreds of different amines to create tens of thousands of amides at the same time, in a manner analogous to the way that a massively parallel processor allows the same set of computer operations to be applied simultaneously to large numbers of data elements (Terrett, 1998). Such large sets of compounds, called *chemical libraries*, are normally created using sophisticated robotics systems that control the dispensing of chemicals and the carrying out of the reaction. Robotics also lies at the heart of the subsequent *high-throughput screening* (HTS) systems that are used to test the biological activities of the resulting reaction products (Hertzberg & Pope, 2000; Huser, 2006). A parallel mode of operation is again used, initially with samples of 96-compounds being tested at a time, then 384 compounds and now 1536 compounds as the standard. The combination of these two technologies means that there has been an explosion in the amounts of chemical and biological data that need to be analysed and, hopefully, rationalised in the search for potential leads and then candidates. Computer techniques had already been used for many years, but the introduction of combinatorial synthesis and HTS led to a significant interest in the use of chemoinformatics techniques in lead-discovery and lead-optimisation, in much the same way as the availability of massive amounts of sequence information spurred the development of bioinformatics to support research in molecular biology.

Indeed, the new technologies spurred a new name – chemoinformatics – which seems to have first been used in 1998: "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 optimization" (Brown, 1998). This definition clearly ties the discipline to the pharmaceutical industry: whilst that has been its most important application area thus far, chemoinformatics techniques are equally applicable to many types of specialty-chemical (as noted above). A more general definition is hence that of Paris: "Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information" (Warr, 1999). The inclusion of the bracketed "o" in Paris' definition is significant since there has been much discussion as to whether the word should be chemoinformatics – presumably by analogy with bioinformatics – or the etymologically-informed cheminformatics. The usage of the two terms, and of others, has been discussed in some detail by Willett (2007); we shall use chemoinformatics in this review.

It is important to emphasize that whilst chemoinformatics is a new word it is not a new discipline: indeed, as noted above, there have already been four previous appearances in ARIST, and an early review of the subject took as its title "Chemoinformatics - a new name for an old problem" (Hann & Green, 1999). The novelty arises from the difference in the scale of operation and implementation, and in the much greater integration than previously of two rather different applications of computers to drug discovery. One application was the creation and searching of databases of chemical molecules, initially in 2D and then later in 3D. The previous ARIST reviews have focused on these approaches, which are often referred to as *chemical structure handling* and which involve the use of sophisticated database methods to process files that can contain millions of molecules. The other application was the prediction of bioactivity in previously untested molecules. Techniques here include the use of highly detailed *computational chemistry* approaches to compute the structures and properties of molecules (Clark, 2004; Cramer, 2004), *molecular modelling*, which uses a range of simpler, and computationally much less-demanding, approaches to the same end (Hinchcliffe, 2003; Leach, 2001), and *quantitative structure-activity relationship* (QSAR) studies, which seek to establish statistical relationships between chemical structure and biological activity (Kubinyi, 1997a, 1997b). These predictive techniques were able to handle only very small files, a few tens of molecules (or a few hundreds at most), when compared with those routinely handled by the various database approaches. This limitation arose not just from the computational requirements of some of the techniques but also from the fact that large amounts of experimental, and hence costly, bioactivity training-data may be required, whereas only limited amounts may be available. The rise of chemoinformatics has been spurred by the integration of these two distinct approaches so as to develop methods of activity prediction that can be used at the database level.

The integration of the database and modelling approaches is reflected in the title of the core journal in the field, the *Journal of Chemical Information and Modeling*. This started life in 1961 as the *Journal of Chemical Documentation*, at a time when the principal focus of the subject was the (printed) chemical literature. It changed its name to the *Journal of Chemical Information and Computer Sciences* in 1975 to reflect the by-then central role of computer methods in the processing of chemical information, and adopted its current title as recently as 2005 to reflect the substantial changes that have taken place in the discipline over the past few years; the effects of these changes in bibliometric terms have been discussed elsewhere (Onodera, 2001; Willett, 2007). Apart from the *Journal of Chemical Information and Modeling*, the other most important journals for chemoinformatics material are the *Journal of Computer-Aided Molecular Design*, the *Journal of Molecular Graphics and Modelling*, and *QSAR and Combinatorial Science*, although an increasing number of journals across the chemical and life sciences now include chemoinformatics papers (Willett, 2007).

SCOPE OF THE REVIEW

Having introduced the subject of chemoinformatics, we now describe the format of this review, which is rather different in scope from many that appear in ARIST in that this one is designedly not comprehensive in scope. There are two reasons for this. First, “chemoinformatics is a vast discipline, standing on the interface of chemistry, biology, and computer science” (Agrafiotis, Bandyopadhyay, Wegner, & van Vlijmen, 2007) and a comprehensive discussion is not feasible given the confines of the typical ARIST review, with the standard text containing four volumes and over 2000 pages (Gasteiger, 2003). Second, many of the topics that are of concern to researchers in chemoinformatics have a strong medicinal chemistry focus and thus require considerable knowledge of chemistry and biology to understand the details of the work that is being carried out. For example, one of the most important topics in the current literature (Willett, 2007) is the computation of the energies of interaction that are involved when a potential drug molecule, or *ligand*, binds to a biological target such as the binding site of an enzyme (Raha & Merz, 2005). The accurate computation of binding energies plays a central role in the scoring functions of many protein-ligand docking algorithms (*vide infra*), but an understanding of the relevant papers requires the reader to have some knowledge of medicinal chemistry in general, and of molecular interactions in particular. The prediction of a molecule’s pharmacokinetic and toxic properties (as discussed very briefly near the end of this review) is another focus of interest that requires considerable specialist knowledge, and there are other such topics in chemoinfomatics that are likely to be equally far removed from the interests and knowledge of ARIST readers, However, as we discuss in the following two paragraphs, there are some aspects of chemoinformatics that are markedly similar to the topic of information retrieval (IR), one of the most important components of library and information science. The review here will hence focus on that aspect of chemoinformatics that is most closely related to IR, *viz* the methods that are used to select molecules from databases of chemical structures, in much the same way as research in IR emphasises methods for selecting records from document databases. *Screening* is the name given to the selection of molecules for bioactivity testing, and the use of computer methods for the selection of molecules is hence generally referred to as *virtual screening*. In fact, we shall restrict the review still further, focusing on those methods for virtual screening that are based on the concept of molecular similarity and that are closely related to those used in IR.

IR systems have traditionally considered the processing of textual documents, but are increasingly also being applied to the processing of multimedia documents (Downie, 2003; Rasmussen, 1997; Smeaton, 2004). Multimedia is normally considered to encompass speech, image and video retrieval but the basic algorithms and data structures that are used in IR are more generally applicable: specifically, it is my view that many are applicable to the processing of chemical structure information. The basis for this view is the very extensive studies of textual and chemical retrieval that have been carried out at the University of Sheffield over many years (Lynch & Willett, 1987), where we have found that studies using textual data have often yielded results that, with relatively little modification, can be applied to chemical data, and *vice versa*.

The relationships that we have identified between IR and chemoinformatics have been explored at some length (Willett, 2000), and so will not be repeated here; however, it is worth briefly summarising the similarities that exist in the ways that chemical and textual database records are characterised. First, the documents in a text database are each typically indexed by some small number of keywords, in just the same way as the 2D or 3D molecular representations in a chemical database are each characterised by some small number of substructural features chosen from a much larger number of potential attributes, i.e., the bits that are set in a fingerprint to denote the presence of a fragment substructure (*vide infra*). Second, both types of attribute follow a well-marked Zipfian distribution, with the skewed distributions that characterise the frequencies of occurrence of characters, character substrings and words in text databases being mirrored by the comparable distributions for the frequencies of chemical moieties. Thus, the overwhelming majority of all of the many millions of molecules that have ever been made contain the element carbon but even the tenth most frequent element, iodine, occurs only about one thousandth as frequently, with the great majority of the elements having vanishingly small frequencies of occurrence; similar distributions are observed for other types of chemical substructure (Lynch, 1977). These shared characteristics mean that the two types of database are amenable to efficient processing using the same type of file structure. For example, one of the first studies of what would now be referred to as text signature searching (Barton, Creasey, Lynch, & Snell, 1974) arose from previous studies of chemical bit-string processing (Adamson, Cowell, Lynch, McLure, Town, & Yapp, 1973); and our early work on chemical similarity searching used an inverted-file search algorithm that had originally been devised for computing inter-document similarities (Willett, 1981). Third, and importantly in the context of this review, in just the same way as a document either is, or is not, relevant to some particular user query, so a molecule is active, or is not active, in some particular biological test. This means that performance measures that have been developed for evaluating the effectiveness of IR systems (in terms of the numbers of relevant and non-relevant documents retrieved) can also be used for evaluating the performance of systems for virtual screening (in terms of the numbers of active and inactive molecules retrieved) (Edgar, Holliday, & Willett, 2000). In addition, just as document test collections, such as Cranfield in the early days or TREC (Text Retrieval Conference) now, play an important role in the development and validation of IR systems, so novel virtual screening methods are routinely tested on publicly available databases of chemical molecules and associated biological activity data. Typical files that are used are: the MDL Drug Data Report (MDDR) database from Symyx Technologies Inc., a set of ca. 130K molecules that have been reported in the literature as having undergone biological testing; ca. 40K molecules that have been tested for HIV-1 activity in the National Cancer Institute’s anti-AIDS programme; and the increasing volumes of data available in the PubChem system, which provides access to the structural and bioassay data that is being generated by the Molecular libraries Roadmap Initiative of the National Institutes of Health. Other such public datasets are becoming available to complement the huge volumes of structural and activity data held within corporate chemoinformatics systems.

The review is structured as follows. The next section provides an historical overview of the development of the subject up to the late Nineties when Paris’ review appeared in ARIST (more extended historical accounts have been reported by Chen (2006) and Willett (2008)). This section is intended to introduce chemoinformatics’ basic technologies to the reader who has no familiarity with the subject, and thus to provide a basis for the main review. This describes the techniques that are used for similarity-based virtual screening in chemical databases, emphasising the computer techniques that are used and making less reference to those approaches that require significant knowledge of chemistry and/or biology for detailed comprehension. We next describe the use of data fusion methods for combining the results of similarity searches. There follow less detailed discussions of two similarity-related topics (chemical clustering, and molecular diversity analysis), and then of other approaches to virtual screening (including machine learning, ligand-protein docking, and drug-likeness and ADMET prediction studies).

HISTORICAL BACKGROUND

The material in this historical overview has been rather arbitrarily divided into two parts: methods for processing databases of 2D chemical structures; and methods for 3D databases and for the prediction of biological activity using QSAR methods. This division has been chosen given the two principal sources on which modern chemoinformatics has drawn, i.e., the archival functions of chemical databases and the predictive functions of molecular modelling and QSAR (*vide supra*). These and other topics are reviewed in more detail in two textbooks (Gasteiger & Engel, 2003; Leach & Gillet, 2003) and in very considerable detail in the multi-volume standard work, the *Handbook of Chemoinformatics* (Gasteiger, 2003).

Processing Databases of 2D Structures

Searching databases of chemical structures lies at the heart of chemoinformatics, and has been studied for many years. A searching method requires some computer-readable representation that encodes a molecule's structure, and many types of representation have been suggested. Two have been of particular significance in the development of the subject: *linear notations* and *connection tables*. Both were studied in the early Sixties but it was the linear notation that first became widely established, initially the Dyson/IUPAC notation and then the Wiswesser Line Notation (or WLN), which formed the basis for both in-house and public chemical information systems during the Sixties and early Seventies (Ash & Hyde, 1975). More recently the Simplified Molecular Input Line Entry Specification (or SMILES) notation has become an important component of in-house systems, where it is used as a simple input mechanism, for compound registration (i.e., adding a new molecule to a database) and structure search (as discussed below), and as an interface between a range of chemoinformatics programs (Weininger, 1988). There is also much current interest in the International Chemical Identifier (or InChI), which has been developed as an open-source, non-proprietary notation for encoding chemical compounds (Coles, Day, Murray-Rust, Rzepa, & Zha, 2005). An InChI is generated algorithmically from a connection table and is being increasingly used by publishers and database suppliers as a standard compound identifier; the development of this standard is overseen by the International Union of Pure and Applied Chemistry (2007). A connection table provides an explicit encoding of a molecule's topology (i.e., the way that its components are connected together) and can hence be considered as a *graph*, in which the atoms and bonds of a molecule are represented by the nodes and edges of a graph (Diestel, 2000; Wilson, 1996). The use of such a graph representation permits the application of graph-match algorithms for a range of searching and structure matching applications in chemoinformatics (Gasteiger & Engel, 2003; Leach & Gillet, 2003). Connection tables attracted attention when they were chosen to form the basis for the first version of the CAS Registry System (Leiter, Morgan, & Stobaugh, 1965) and they are now the most important type of structure representation; many different formats have been described in the literature, but it is normally possible to convert from one to another without too much trouble (Barnard, 1990; Dalby, Nourse, Hounshell, Gushurst, Grier, Leland, et al., 1992).

Once chemical compounds could be stored in machine-readable form, it was natural to think of ways in which the resulting databases could be searched. The simplest type of search mechanism is *structure searching*, i.e., checking for the presence or absence of a specific molecule in a database. This is very simple to implement if a canonical (i.e., unique and unambiguous) character-string representation of a molecule is available, such as a WLN; structure searches of connection-table files were more problematic till the development by Morgan (1965) at CAS of a simple canonicalisation procedure that, with enhancements (Freeland, Funk, O'Korn, & Wilson, 1979; Wipke & Dyott, 1974), continues to be widely used, although the new InChI notation (*vide supra*) has developed an alternative canonicalisation procedure (Warr, 2003).

Probably the most important retrieval facility in chemical databases is *substructure searching*, which involves checking for the presence of a partial structure in a complete molecule (Barnard, 1993). The first algorithm for substructure searching was described half-a-century ago (Ray & Kirsch, 1957) and involved a backtracking subgraph-isomorphism procedure that exhaustively searched a connection table for the presence of the query pattern. This algorithm is effective, in that it enables searches to be carried out with complete recall, but it is also highly inefficient, and several years were to pass before the introduction of *set-reduction* (Sussenguth, 1965) provided a rapid and simple technique for the identification of subgraph isomorphism. Set-reduction, and developments such as the fast algorithm due to Ullmann (1976), enabled significant increases in search speeds but were still totally unable to make subgraph matching feasible on databases of non-trivial size in a reasonable amount of time. Large-scale substructure searching only became a viable retrieval option with the introduction of fragment-based *screening* methods. These involve the application of a filter that allows only those molecules, a very small minority, that contain all of the query's substructural fragments to pass on to the time-consuming subgraph-isomorphism search. This idea is analogous to the use of keyword-based IR to filter a database prior to the application of sophisticated natural language processing techniques such as automatic summarisation.

The presence of fragments (either in a query substructure or in a database structure) is encoded in a binary vector (also called a *bit-string* or a *fingerprint*). Two main approaches have been developed for selecting the fragments that are used for screening: either each bit-position in the binary vector corresponds to one, or a small number, of fragments selected from a dictionary of possible screens (see, e.g., Adamson, Cowell, Lynch, McLure, Town, & Yapp (1973)); or hashing algorithms are used to allocate multiple fragments to each bit-position (see, e.g., (Feldman & Hodes, 1975)). These screening approaches were rapidly developed for first batch (Graf, Kaindl, Kniess, Schmidt, & Warszawski, 1979) and then online (Attias, 1983; Dittmar, Farmer, Fisanick, Haines, & Mockus, 1983) systems for 2D substructure searching in the CAS Registry System, and they also formed the basis for the new generation of in-house chemoinformatics systems that became available in the Eighties (Ash, Warr, & Willett, 1991).

Following the establishment of systems for 2D structure and substructure searching, the next development was the introduction of *similarity searching*. This involves the user submitting an entire molecule (referred to as the *reference* structure or the *target* structure), typically one that has previously exhibited activity in a biological screening experiment. The search calculates a measure of similarity between the reference structure and each of the molecules, and then ranks the database in order of decreasing similarity. The rationale for such a retrieval mechanism is the *Similar Property Principle* (Johnson & Maggiora, 1990), which states that molecules that are structurally similar are likely to have similar properties (and which is discussed in detail later in the review). Thus, if a bioactive target structure is searched for, then the top-ranked, nearest-neighbour molecules are also likely to possess that activity: these molecules are hence prime candidates for biological testing, as compared to other molecules that occur further down the ranking. Similarity searching will be effective if, and only if, an appropriate similarity measure is employed, and there has hence been much discussion as to what sorts of measure should be employed (Bender & Glen, 2004; Dean, 1994; Nikolova & Jaworska, 2003; Sheridan & Kearsley, 2002; Willett, Barnard, & Downs, 1998). Of the measures that have been suggested, by far the most common are those based on the numbers of substructural fragments common to a pair of molecules, this number being normalised in the range zero (no fragments in common) to unity (all fragments in common) by means of an association coefficient (usually the Tanimoto coefficient (Willett, Barnard, & Downs, 1998)). The quantification of structural resemblance by comparing sets of fragment substructures was first used for predicting chemical and biological properties in small datasets (Adamson & Bush, 1973), but was later applied in two near-contemporaneous papers (Carhart, Smith, & Venkataraghavan, 1985; Willett, Winterman, & Bawden, 1986a) that appeared over a decade later and that clearly demonstrated the power and the simplicity of fragment-based approaches for chemical database applications. Similarity searching was thus rapidly adopted (not least because the requisite fragment occurrence data was already available from its use for 2D substructure searching), and continues to be one of the principal building-blocks of any system for chemoinformatics, as discussed below in the main review.

The methods discussed thus far have been designed to search databases of individual, specific molecules. However, account also needs to be taken of the *generic* (or Markush) structures that often play a key role in defining the claims made in chemical patents, since a single generic structure provides a simple way of encoding large numbers, or even an infinite number, of distinct individual molecules (Barnard, 1984). Patents represent a vital information resource for the chemical industry, and the early fragment-based systems for the retrieval of chemical patents were developed by industrial organisations, e.g., Derwent Information Ltd. (Nübling & Steidle, 1970) and International Documentation in Chemistry (Rössler & Kolb, 1970). Such systems are limited in that they cannot provide a detailed graph representation of a Markush structure, and it was not till the end of the Eighties that operational systems were developed that allowed effective substructure searches of patent databases to be carried out. These systems (Berks, 2001; Fisanick, 1990; Shenton, Norton, & Fearns, 1988) were based in large part on over a decade of academic research (as reviewed by Lynch & Holliday (1996)) that developed techniques for the graph-based screening and searching of a fully explicit 2D representation of each of the molecules covered by a patent.

The techniques described above have focused on the representation and searching of chemical molecules, but chemistry is based on the inter-conversion of molecules by means of reactions. It was hence recognised at an early stage in the development of chemoinformatics that there was a need to represent and to search reaction information. The basic problem is much more complex than with molecules, since in a reaction one must consider not just a single molecule but linked ensembles of molecules, i.e., the sets of reactants and of products in the reaction; moreover, one must consider both the constituent molecules and the *reaction sites*, i.e., the parts of the reacting molecules where the reaction has taken place. Much of the early work on reaction information processing focused on ways of representing the reaction sites, since the substructural transformation engendered by a reaction is a vital component of many of the queries put to a reactions database. Vleduts was the first person to realise that a reaction site could be identified automatically by a detailed comparison of the connection tables of the reactants and products, this comparison identifying the parts of the reacting molecules where structural changes had occurred (Vleduts, 1963). The idea is a simple one but almost two decades were to pass before a matching procedure was described that could identify reaction sites effectively and efficiently using a maximum common subgraph (MCS) isomorphism procedure (Willett, 1980); this graph-based approach has since been much enhanced (Chen, Nourse, Christie, Leland, & Grier, 2002). Once the reaction sites had been detected, it was possible to carry out substructure searches for reacting molecules and/or reaction sites and the early Eighties saw the first operational reaction-database systems (Willett, 1986), with later work adding facilities for reaction similarity searching (Moock et al., 1988) and reaction classification (InfoChem, 2007).

Information is a necessary precursor of knowledge, and the Seventies onwards saw much interest in *expert systems*, knowledge-based information systems that could go beyond simple retrieval to provide expert advice analogous to that provided by a human expert in a specific domain. Much early expert-systems work involved processing chemical knowledge, specifically knowledge about chemical syntheses and knowledge about the elucidation of structures from spectral information. The first of these applications, *computer-aided synthesis design* (or CASD), was first suggested as a possible area of research in Vleduts' automatic reaction-indexing paper (Vleduts, 1963). This suggested that if one could store information about common reactions and about the conditions under which those reactions could be used successfully, then an automated inference engine could suggest a series of reactions that, taken in sequence, would result in the synthesis of a desired molecule in acceptable yield. This idea was embodied in OCSS, the first operational CASD program (Corey & Wipke, 1969), with subsequent developments (Corey, Wipke, Cramer, & Howe, 1972; Wipke, Ouchi, & Krishnan, 1978) occasioning much interest during the Seventies and Eighties, despite the very large amounts of high-quality chemical knowledge that needed to be encoded if an acceptable level of performance were to be achieved (Loftus, 1991; Wipke & Howe, 1977). A less labour-intensive approach to CASD was pioneered by Gasteiger and his collaborators (Blair, Gasteiger, Gillespie, Gillespie, & Ugi, 1974; Gasteiger & Jochum, 1978), who took very simple starting-point molecules and then created complex synthetic pathways using approximate physicochemical calculations to estimate the feasibility of the suggested molecules. The reader is referred to a recent review for further details of CASD (Ott, 2004).

In fact, the work on CASD was preceded by that on *computer-aided structure elucidation* (CASE), which was the first major scientific application of expert systems technology and which involves the use of (principally) spectroscopic information for determining the identity of an unknown compound (Gray, 1986). Starting in 1965, the DENDRAL project developed a range of methods for structure generation and elucidation based on the analysis of mass spectral information (Lindsay, Buchanan, Feigenbaum, & Lederberg, 1980). These methods have formed the basis for much subsequent research, using not just mass spectra but also infra-red and nuclear magnetic resonance (NMR) data (Bremser, Klier, & Meyer, 1975; Gray, 1986; Munk, 1998; Sasaki, Abe, Ouki, Sakamoto, & Ochiai, 1968; Shelley, Hays, Munk, & Roman, 1978). The basic approach in CASE is to use the spectral information to suggest substructural fragments that may be present in the unknown molecule. These fragments are then connected together in all possible ways, and the spectra simulated for each of the resulting molecules: these predicted spectra can then be compared with the available experimental data to confirm or to deny the presence of larger substructural moieties. This plan-generate-test procedure is continued until a molecule has been suggested that is fully compatible with the experimental data. The reader is referred to a recent review for further details of CASE (Pretsch, Tóth, Munk, & Badertscher, 2003)

Finally in this section, we mention molecular diversity analysis, which was just starting to come to the fore at the time of Paris's ARIST review. Molecular diversity analysis takes as its starting point the need to maximise the diversity (i.e., the structural heterogeneity or the structural dissimilarity) of the molecules that are submitted for biological testing. Although HTS is very rapid, it is still costly if implemented on the very large scales that characterise pharmaceutical research programmes, and there is hence a strong economic imperative to minimise the numbers of molecules that are assayed. An implication of the Similar Property Principle is that structurally similar molecules are likely to give similar results when biological testing is carried out; thus, the information that can be gained from a set of molecules about the relationship between structure and activity will be maximised if those that are submitted for HTS are as structurally diverse as possible (Martin, Willett, Lajiness, Johnson, Maggiora, Martin, et al., 2001; Patterson, Cramer, Ferguson, Clark, & Weinberger, 1996).

The need for diversity may sound like a statement of the obvious, but the practical realisation of this has proved to be very difficult. Early approaches involved selecting a diverse subset of a database by consideration of their inter-molecular structural similarities, typically as determined by use of fragment bit-string similarity measures (Bawden, 1993; Lajiness, 1990; Martin, Blaney, Siani, Spellmeyer, Wong, & Moos, 1995). The problem is the astronomical number of possible subsets that can be generated from a database of non-trivial size: it is infeasible to consider all of them so as to identify the most diverse subset that can then be submitted for HTS. There has thus been much interest in alternative approaches for selecting diverse sets of molecules that maximise the coverage of structural space, whilst minimising the numbers of molecules put forward for testing. Cluster analysis, or automatic classification, was the first such technique to be used for this purpose (Brown & Martin, 1996; Shemetulskis, Dunbar, Dunbar, Moreland, & Humblet, 1995; Willett, Winterman, & Bawden, 1986b), but several other algorithmic approaches are now being used (*vide infra*).

Processing 3D Information and the Prediction of Biological Activity

Thus far, we have focused on techniques that are suitable for handling databases of 2D chemical structures. However, the activities and properties of molecules are crucially dependent on their 3D characteristics; there is thus a need to extend 2D techniques to encompass not just the topologies but also the geometries of molecules, this leading naturally to a discussion of the tools that are available for predicting biological activity.

Methods for 2D substructure searching based on bit-string screening and subgraph isomorphism matching were well established by the mid-Seventies. Whilst providing a valuable means of accessing a chemical database, either corporate or public, such techniques took no account of the geometric characteristics of the individual molecules. Gund was the first person to suggest that graph-based methods could also be applied to the retrieval of 3D chemical structures, with the nodes and edges of a graph being used to represent the atoms and inter-atomic distances, respectively, in a 3D molecule (Gund, 1977; Gund, Wipke, & Langridge, 1974). The resulting inter-atomic distance matrix could then be inspected for the presence of a query *pharmacophore*, or *pharmacophoric pattern*, i.e., the arrangement of structural features in 3D space necessary for a molecule to bind at an active site (Güner, 2000). The importance of Gund's work was widely recognised but no further significant developments occurred for almost a decade. There were two reasons for this apparently surprising lack of interest: first, there was no obvious way to obtain 3D atomic coordinate data for the structures in corporate chemical databases, the principal source of the novel bioactive compounds that drive the pharmaceutical industry; second, while Gund had demonstrated the feasibility of 3D substructure searching, his search algorithm could only be used with very small numbers of molecules and was not obviously scalable to the database context. The development of operational systems for 3D substructure searching in the late-Eighties and early Nineties was due to the emergence of near-contemporaneous solutions to both of these long-standing problems.

The obvious source of 3D structural data is the X-ray crystal structures in the Cambridge Structural Database (CSD) produced by the Cambridge Crystallographic Data Centre (Allen, 2002). However, even now there are only ca. 400K molecules in the Database for which an experimental structure determination has been published, and many of these structures are of crystallographic, rather than of pharmaceutical, interest. The creation of large searchable databases hence required the development of software that could generate computational, rather than experimental, structures. The late-Eighties saw the introduction of several programs that met these criteria and that thus enabled companies to convert their corporate databases to 3D form. The two most important programs are CONCORD (Pearlman, 1987) and CORINA (Hiller & Gasteiger, 1987), which permit the rapid computation of a reasonably accurate low-energy structure for a large fraction of the molecules in a typical corporate database (Green, 1998).

In a series of papers starting in 1986 (Jakes & Willett, 1986), Willett and collaborators showed that the bit-string screening and graph-matching techniques that had been developed for efficient 2D substructure searching could be modified to enable rapid searching of 3D databases for pharmacophoric patterns, with operational systems soon being reported (Jakes, Watts, Willett, Bawden, & Fisher, 1987; Sheridan et al., 1989). These early systems were, however, limited in that they took no account of the flexibility that characterises many molecules: specifically, a molecule may exist in some, or many, different geometric forms (called *conformations*) that have different shapes and energies, with those of lowest energy being the most stable. Accordingly, a pharmacophore search of a database of rigid structures, in which each molecule is represented by just a single, low-energy conformation, is likely to miss large numbers of matching molecules that can adopt a conformation containing the query pattern but that are represented in the database by a low-energy conformation that does not contain this pattern. Two main approaches to flexible 3D searching have been described in the literature (Warr & Willett, 1997). ChemDBS3D was developed by Chemical Design Limited and was the first system to allow flexible 3D searching; this was achieved by summarising the conformational space (i.e., the full range of conformations that a molecule can adopt) of a molecule by selecting a number of its low-energy conformations when it is added to a database, with a rigid-searching algorithm being applied to each of the conformations describing a molecule (Murrall & Davies, 1990). Alternatively, work at Sheffield and subsequently by Tripos Inc. and MDL Information Systems Inc. developed representational and searching techniques that allow exploration of the entire conformational space of a molecule when a pharmacophoric query is submitted (Clark, Willett, & Kenny, 1992; Hurst, 1994; Moock, Henry, Ozkaback, & Alamgir, 1994).

Research and development of systems for 3D substructure searching was driven in part by the knowledge that the 3D structure of a molecule plays a key role in determining its biological activity. 3D substructure searching was hence seen as one way of linking two previously separate areas of research: 2D database searching and the prediction of biological activity by means of molecular modelling and QSAR techniques. By the late Eighties, the latter had already reached an advanced stage of development when applied to the detailed analysis of datasets containing small numbers of structurally related molecules (or *analogues*) such as are commonly considered in medicinal chemistry lead-optimization programmes. Database searching, conversely, has as its aim the provision of simple retrieval mechanisms for even the largest files of compounds, covering the full range of structural types. This focus had resulted in an emphasis on computational efficiency and highly scalable algorithms, in marked contrast to the sophisticated processing methods that had been developed to rationalise the relationships between chemical structure and biological activity.

Perhaps the most important single contribution to the development of QSAR was work by Hansch in the early Sixties that introduced the use of multivariate statistical methods, specifically multiple regression, to correlate molecular physicochemical properties (most importantly the octanol-water partition coefficient but also other properties such as molar refractivity) with quantitative data describing the biological activity of molecules (Hansch & Fujita, 1964; Hansch, Maloney, Fujita, & Muir, 1962). The importance of this approach is evidenced by it rapidly becoming known as *Hansch analysis* and being used, with continuing development (Martin, 1978), to the present day (Hansch & Leo, 1995; Hansch, Hoekman, Leo, Weininger, & Selassie, 2002). A characteristic, some would say a limitation, of Hansch analysis is that it is based on the correlation of properties (which may be difficult to measure or to compute), rather than structures (which are readily available in 2D or 3D form) with activity. Shortly after the appearance of Hansch's initial paper, Free and Wilson (1964), working at Smith Kline & French, developed an alternative approach to QSAR that again used multiple regression but that correlated activity data with structural variables indicating the presence or absence of substituents on a common core ring-system.

Application of the Free-Wilson approach requires assuming that a given substituent at a given position on a fixed, central ring-system makes an additive and constant contribution to the overall activity of a molecule that contains it, irrespective of the other substituents that are present; moreover, like Hansch analysis, it is restricted to the analysis of small sets of analogues (specifically sets that involve an invariant central ring system). These limitations were tackled by the introduction of *substructural analysis*, an approach to the prediction of biological activity that can be applied to large, structurally diverse datasets (such as the compounds in a corporate database that have been tested in a primary biological screen) (Cramer, Redl, & Berkoff, 1974). Substructural analysis is applicable when the activity data is binary in nature (i.e., active or inactive), and involves calculating weights that describe the extent to which the presence of a particular fragment in a molecule is associated with the likelihood that that molecule will be active (or inactive). The probability that a previously untested molecule is active can then be computed by combining the weights for its constituent fragments, this step again making the assumption that a fragment's contribution to activity is independent of the precise substructural environment in which it occurs. Although not realised at the time, substructural analysis was the first chemical application of machine learning (*vide infra*).

The SAR methods described above have been based on the properties of molecules or their 2D structures, without direct consideration of the 3D geometries of the molecules that were being analysed. With improved computer technology, 3D graphics started to be used for visualisation purposes in the early Eighties (Gund, Andose, Rhodes, & Smith, 1980; Langridge, Ferrin, Kuntz, & Connolly, 1981) and then for the detailed modelling of molecules. An important early study (Marshall, Barry, Bosshard, Dammkoehler, & Dunn, 1979) described the first efficient algorithms for searching the conformational spaces of molecules to identify the potential shapes that they could adopt. Not only did this work facilitate the computation of a range of types of geometry-related molecular properties but it also inaugurated the concept of *pharmacophore mapping*. Assume that several molecules have all been found to exhibit some particular biological activity, and that information is available as to the atoms (or other molecular features) that comprise the pharmacophoric pattern that is involved in binding to the biological target. Then the *active analogue* approach (Marshall, Barry, Bosshard, Dammkoehler, & Dunn, 1979) provided the first systematic way to identify those conformations that would enable the points comprising the pharmacophore in each molecule to adopt a common arrangement. An IR analogy might be the analysis of search outputs to identify words and phrases occurring in known relevant documents.

The active-analogue approach provided the first automated procedure for the identification of pharmacophoric patterns, which in turn provided the spur for the work on 3D database searching described previously in this section. Crandell and Smith (1983) and then Brint and Willett (1987) described the use of MCS algorithms to identify automatically the sets of points comprising the pharmacophore, thus removing the need for the user to specify this information prior to the conformational-analysis stage of pharmacophore mapping. This graph matching work led in due course to the development by Martin *et al*. (1993) of DISCO, the first of many operational systems for pharmacophore mapping (Güner, 2000). Once efficient and effective tools had become available for structure generation and for conformational searching, it was not long before workers in QSAR began to consider how these techniques could be used to assist in the prediction of biological activity. Several 3D-QSAR approaches have been described in the literature (Doweyko, 2004; Greco, Novellino, & Martin, 1998; Kubinyi, Folkers, & Martin, 1998). An example of these is Comparative Molecular Field Analysis (CoMFA) (Cramer, Patterson, & Bunce, 1988), which correlates bioactivity with descriptors derived from 3D grid representations of molecular steric or electrostatic fields (i.e., the distribution of shape and of charge around a molecule in 3D space) and which is now very widely used in lead-optimisation studies (Cramer, DePriest, Patterson, & Hecht, 1993; Cramer & Wendt, 2007).

Pharmacophore mapping, CoMFA and related tools permit the analysis of datasets where 3D structural information is available for a set of potential drug molecules (known as *ligands*), but more sophisticated methods are available if the 3D structure of the biological target has been determined experimentally (using the methods of X-ray crystallography, protein NMR or homology modelling). *Structure-based design* methods seek to identify compounds that might bind to a biological target of known geometry, with two approaches being of importance: *docking*, where one scans a database to find molecules that are complementary to a binding site; and *de novo* design, where one assembles molecules from scratch that are complementary to a binding site. Fitting a molecule into a protein is analogous to fitting a key into a lock, an idea that was first described in the DOCK program (Kuntz, Blaney, Oatley, Langridge, & Ferrin, 1982). This docked a 3D structure from the Cambridge Structural Database into a *receptor* (i.e., a protein that binds to a specific ligand), focusing initially on shape complementarity, as represented by a sphere-based description of the geometries of the receptor and of the potential ligands. Early extensions to the program included the additional use of chemical complementarity and the sequential docking of multiple molecules, thus providing an extremely sophisticated approach to database scanning; more recent developments in docking are discussed later in the review. Docking is very effective in suggesting molecules that are complementary to the receptor (i.e., potential ligands) but is restricted to molecules contained in the database that is being searched. *De novo* structure generation goes one stage further in suggesting novel substances, by building molecules into the active site of a protein so as to ensure a high level of complementarity (Gillet & Johnson, 1998). The construction of a new molecule involves joining together atoms or small substructural fragments to form increasingly larger species until one is achieved that provides a satisfactory degree of fit to the active site. The combinatorial explosion of possible substances is alleviated by the application of strict constraints based both on the nature of the active site and on the stability and the synthetic feasibility of the grown ligands. The process is exceedingly complex, and computationally demanding, and progress has thus been quite slow since the first such programs were reported in the late Eighties and early Nineties (Böhm, 1992; Danziger & Dean, 1989; Gillet, Johnson, Mata, & Sike, 1990; Moon & Howe, 1991; Nishibata & Itai, 1991); however, the ability to suggest novel, previously unrecognised structural classes suggests that such techniques will become of increasing importance in lead-discovery programmes in the future (Schneider & Fechner, 2005).

SIMILARITY-BASED VIRTUAL SCREENING

Introduction

Virtual screening is the generic name given to the computational methods that are used for selecting molecules from a chemical database, and there are three main ways in which this can be done: selecting molecules for acquisition from amongst the published catalogues of commercial chemical suppliers; selecting molecules from an existing database, which in drug research normally means a company’s corporate database of structures that it has synthesised over the years; or selecting molecules from a virtual database, i.e., a set of molecules that has been created by a computer program of some sort but that could be synthesised if required. The principal selection criterion in virtual screening is to maximise the probability that the molecules selected by computer processing will prove to be active when subjected to biological testing, thus maximising the cost-effectiveness of the testing process. The ranking of a database in order of decreasing probability of activity is a task that is clearly analogous to the many models of IR that seek to order the documents in a database in order of decreasing probability of relevance (Spärck Jones & Willett, 1997; van Rijsbergen, 1979).

The virtual screening approaches that can be used in any particular circumstances depend principally upon the amounts and types of data that are available. If just a single active molecule is available, such as a competitor's compound or a natural product, then similarity searching can be used, in which a database is ranked in decreasing order of similarity to the known active structure. If several structurally related actives have been identified then pharmacophore mapping can be carried out to ascertain common patterns of features that may be responsible for the observed activity, with a 3D substructure search of the database then being carried out to identify further molecules that contain the pharmacophore, or a less-precise 2D substructure search if there are few geometric commonalities. If it is not possible to identify a common pharmacophore, as often occurs with heterogeneous sets of actives (e.g., the initial hits from an HTS programme), and if significant numbers of both active and inactive molecules are available, then these can be used as training data for a machine learning system. Finally, if the 3D structure of the biological target is known then a docking study can be carried out to identify those database molecules that are complementary to the binding site. Docking is an example of *structure-based virtual screening*; the other approaches are examples of *ligand-based virtual screening* (since information is available about the ligands for some biological target, rather than about the target itself). Both types of virtual screening are likely to be used in a research project (Prathipati, Dixit, & Saxena, 2007) and their importance for drug discovery has occasioned many excellent reviews, to which the reader is referred for further details of the topics discussed here (Alvarez & Shoichet, 2005; Bajorath, 2002; H.-J. Böhm & Schneider, 2000; Klebe, 2000; Lengauer, Lemmen, Rarey, & Zimmermann, 2004; Lyne, 2002; Maldonado, Doucet, Petitjean, & Fan, 2006; Oprea, 2002; Oprea & Matter, 2004; Walters, Stahl, & Murcko, 1998).

Of the virtual-screening approaches that are available, similarity searching is the most closely related to procedures in IR, specifically to the classical *ad hoc* retrieval problem where a database is ranked in decreasing order of similarity to the query. In (textual) IR, the query is a set of words or phrases describing the user's information need; in chemoinformatics, the query is a known bioactive molecule (i.e., the reference structure). Pharmacophore searching and machine learning can both be thought of as chemoinformatics equivalent of query expansion by relevance feedback. Pharmacophore searching uses features extracted from sets of bioactive molecules (as against sets of relevant documents); and machine learning uses weighted features extracted from sets of active and inactive molecules (as against relevant and non-relevant documents).

This section focuses on similarity searching, describing in some detail both the methods themselves and the techniques available to compare the effectiveness of different methods. Subsequent sections will discuss the use of *data fusion* to combine multiple screening methods, an approach that seeks to overcome the limitations of individual methods, and of two other techniques in which measures of molecule similarity play a central role, *viz* the clustering of chemical databases and molecular diversity analysis.

The Similar Property Principle

Similarity searching was first described (Carhart et al., 1985; Willett et al., 1986a) in the mid-Eighties as a complement to existing systems for 2D substructure searching. Although highly effective, substructure searching systems have several obvious limitations: there is little control over the volume of output that is obtained in response to a query; the output is not ranked in any way (other than by date of accession to the database); and query formulation can be complex for the non-expert. Similarity searching suffers from none of these problems, and was hence rapidly adopted as a standard method for database-access, in much the same way as ranked-output searching was rapidly adopted in operational IR systems as a way of alleviating the manifest problems of Boolean text searching (Davis & McKim, 1999; Salton, 1989; Spärck Jones & Willett, 1997). 2D similarity searching was thus well established by 1997 when Paris' ARIST review appeared, and he hence described it only briefly, referring the reader to a then-recent, detailed review (Downs & Willett, 1995). However it is now, some ten years later, the subject of significant renewed interest: indeed, it is probably the most widely used technique for virtual screening. There are several reasons for its popularity, including its conceptual simplicity, ease of use, availability (not just in-house but also since 2006 for searching the CAS Registry System), and proven effectiveness (Bender & Glen, 2004; Sheridan, 2007; Sheridan & Kearsley, 2002; Willett et al., 1998).

The rationale for similarity-based virtual screening is the Similar Property Principle, which has been introduced previously and which also underlies much research into molecular diversity (as discussed later in this review when discussing the concept of neighbourhood behaviour). It must be emphasised that the Principle is, in fact, just an assumption: but if the assumption is valid for some dataset then the nearest neighbours of a bioactive reference structure are also likely to be active, with database molecules that appear lower down the similarity ranking being de-emphasised when considering which molecules should be submitted for biological testing. There are many exceptions to the Principle (Kubinyi, 1998; Maldonado, Doucet, Petitjean, & Fan, 2006; Nikolova & Jaworska, 2003; Stahura & Bajorath, 2002), but these do not invalidate its use in drug research. Indeed, if there were not at least some relationship between chemical similarity and biological similarity then it would be very difficult to develop rational approaches for drug discovery that took account of the structures of the molecules. Willett and Winterman (1986) carried out the first detailed investigation of the extent of the correlation between chemical (i.e., structural) similarity and biological activity, using an approach based on simulated property prediction that has formed the basis for many subsequent studies. Important examples include work carried out by the research groups under Bajorath (Eckert & Bajorath, 2006; Eckert & Bajorath, 2007; Godden, Stahura, & Bajorath, 2004), Martin (Brown & Martin, 1996, 1997b; Martin, Kofron, & Traphagen, 2002), Maggiora (Cheng, Maggiora, Lajiness, & Johnson, 1996; Shanmugasundaram, Maggiora, & Lajiness, 2005) and Sheridan (Kearsley et al., 1996; Sheridan, 2000, 2007; Sheridan & Miller, 1998) *inter alia*. It is worthy of note that many of these studies have been carried out by groups based in the pharmaceutical industry, rather than in academe as would be the case for most scientific disciplines.

We note here a further relationship between IR and chemoinformatics, since the Similar Property Principle has a direct IR analogue in the Cluster Hypothesis, which states that similar documents tend to be relevant to the same requests (van Rijsbergen, 1979; Willett, 1988). The Hypothesis is a direct equivalent of the Principle: as can be demonstrated by replacing "documents" with "molecules" and "be relevant to the same requests" by "exhibit the same biological properties". Indeed, it was our studies of hierarchic document clustering (El-Hamdouchi & Willett, 1989; Griffiths, Robinson, & Willett, 1984) that inspired our extensive studies of chemical clustering (Willett, 1987), an approach that is now extensively used in compound-selection programmes (*vide infra*).

Three recent papers have provided further evidence for the general applicability of the Similar Property Principle. The first two of these studies (He & Jurs, 2005; Sheridan, Feuston, Maiorov, & Kearsley, 2004) showed that the predictive power of a QSAR model is dependent on the extent to which the test-set molecules, for which predictions are required, are similar to the training-set molecules, on which the QSAR model has been based. Thus, strong structural similarities between the test-set and training-set molecules enable the making of accurate bioactivity predictions, as would be the case if the Principle is, indeed, correct (although it has been suggested that there are some situations where such QSAR behaviour might not be observed (Maggiora, 2006)). The third study (Bostrom, Hogner, & Schmitt, 2006) used data from the Protein Data Bank (the principal repository of information regarding the 3D structures of proteins (Berman, Battistuz, Bhat, Blum, Bourne, Burkhardt, et al., 2002)) to show that structurally similar molecules bind in the same way, i.e., they not only exhibit the same bioactivity but they also do this by the same mode of action. The Similar Property Principle also underlies recent work in the emerging field of *chemogenomics*, which seeks to establish the relationships that exist between small molecules and biological macromolecules that might be drug targets (Jacoby, 2006; Kubinyi & Muller, 2004). For example, two recent studies have shown that molecules with similar 2D fingerprints bind to protein targets that are in the same structural class (Paolini, Shapland, van Hoorn, Mason, & Hopkins, 2006; Schuffenhauer, Floersheim, Acklin, & Jacoby, 2003), another has demonstrated that it is possible to use molecule-based similarities to suggest novel functional relationships between protein targets that have little sequence similarity to each other (Keiser et al., 2007), and a fourth that pairs of drugs that act at the same biological target are more similar to each other than pairs of drugs that do not have a common target (Cleves & Jain, 2006).

Similarity Coefficients And Weighting Schemes

The extent to which the Principle holds for a particular set of compounds will be crucially dependent on the effectiveness of the measure that is used to quantify the degree of similarity between a pair of molecules, and much research into molecular similarity (and also molecular diversity analysis) has focussed on the key components of a similarity measure that control the effectiveness of searching. There are three such components: the representation that is used to characterise the molecules that are being compared; the weighting scheme that is used to assign differing degrees of importance to the various components of these representations; and the similarity coefficient that is used to provide a quantitative measure of the degree of structural relatedness between a pair of structural representations. Thus far, there have been only a few reports in the literature on the extent to which the weighting scheme affects the utility of a similarity measure. The principal finding has been that fingerprints that note the frequency of occurrence of a fragment are more effective in operation than those that note merely its presence or absence (Chen & Reynolds, 2002; Ewing, Baber, & Feher, 2006; Fetchner, Paetz, & Schneider, 2005), as exemplified in the commercial HQSAR fingerprints produced by Tripos Inc. (Tong, Lowis, Perkins, Chen, Welsh, Goddette, Heritage, & Sheehan, 1998) and the Pipeline Pilot fingerprints produced by SciTegic Inc. (Hassan, Brown, Varma-O'Brien, & Rogers, 2006). However, the general lack of interest is rather surprising given the prominence of weighting in the IR context, as exemplified, e.g., by Spärck Jones' extended series of studies of the weighting of index terms in documents and queries (Robertson & Spärck Jones, 1976; Spärck Jones, 1972; Spärck Jones, Walker, & Robertson, 2000). Instead, most interest in chemoinformatics has focussed on the ways that the molecules are represented and on the coefficients that are used to quantify the similarity of two such representations.

Early work on similarity coefficients (Willett & Winterman, 1986) demonstrated that a simple association coefficient - the Tanimoto coefficient – provided an effective way of comparing 2D fingerprints and this rapidly established itself as the standard coefficient for similarity-based virtual screening, despite the very many alternative coefficients that might be used for this purpose (Willett, Barnard, & Downs, 1998). However, Flower noted that the Tanimoto typically yields low similarity values when the reference molecule in a similarity search has just a few bits set in its fingerprint (Flower, 1988). This marked size-dependency was confirmed in later studies (Dixon & Koehler, 1999; Fligner, Verducci, & Blower, 2002), and it was also demonstrated that the coefficient has an inherent bias towards certain similarity values (Godden, Xue, & Bajorath, 2000). These problems spurred studies of the characteristics of over 20 fingerprint-based similarity coefficients (Holliday, Hu, & Willett, 2002; Salim, Holliday, & Willett, 2003). The work was arguably unsuccessful, in that it did not prove possible to identify a coefficient (or a combination of coefficients) with a consistently better level of performance than the Tanimoto; however, the studies did show that most coefficients, and not just the Tanimoto, had inherent biases dependent on the sizes of the molecules that were being retrieved. These empirical findings, and similar ones that had been noted previously in research on molecular diversity analysis, were later rationalised by a mathematical analysis that related a coefficient's degree of bias to the relative sizes (i.e., numbers of bits) of a reference molecule and the database-molecules with which it was being compared (Holliday, Salim, Whittle, & Willett, 2003). The majority of the coefficients studied by Holliday *et al*. were symmetric, in the sense that they gave the same result whether the reference structure was compared to a database structure, or *vice versa*. There have, however, been some calls for the use of asymmetric coefficients, based on ideas first put forward by Tversky (1977), for the calculation of inter-molecular structural similarities (Bradshaw, 1997; Maggiora, Mestres, Hagadone, & Lajiness, 1997). A recent study has suggested that the use of such a coefficient can be beneficial for database searching (Chen & Brown, 2006), although this conclusion has been criticised on the grounds that the observed effects are a result of the size biases noted above (Wang, Eckert, & Bajorath, 2007).

Structural Representations For 2D Similarity Searching

A huge number of types of descriptor has been suggested for characterising the structures of chemical molecules (Gasteiger, 2003; Glen & Adams, 2006; Todeschini & Consonni, 2002), and many of these have been used in studies of molecular similarity (e.g., (Bender & Glen, 2004; Brown & Martin, 1998; Willett, Barnard, & Downs, 1998), *inter alia*). It is common to divide them into three main classes: whole molecule (sometimes called 1D) descriptors; descriptors that can be calculated from 2D representations of molecules; and descriptors that can be calculated from 3D representations (which are discussed in the next section).

Whole molecule descriptors are single numbers, each of which represents a different property of a molecule such as its molecular weight, the numbers of heteroatoms or rotatable bonds, or a computed physicochemical parameter such as logP (the logarithm of the octanol/water partition coefficient as used in Hansch analysis (*vide supra*)). A single 1D descriptor is not usually discriminating enough to allow meaningful comparisons of molecules and a molecule is hence normally represented by several (or many) such descriptors, possibly after standardisation so that they are all measured on the same scale. For example, a set of 104 such properties has been used by Godden *et al.* for similarity-based virtual screening (Godden, Furr, Xue, Stahura, & Bajorath, 2004) and molecular diversity analysis (Godden, Xue, Kitchen, Stahura, Schermerhorn, & Bajorath, 2002). It is also convenient to mention here affinity fingerprints, which are vectors containing a molecule's binding affinities or docking scores obtained with a reference panel of proteins: examples of the use of such fingerprints are reported by Briem and Lessel (2000), Kauver et al. (1995) and Dixon and Villar (1998).

2D descriptors include topological indices and substructural descriptors. A topological index is a single number that typically characterises a structure according to its size and shape (Kier & Hall, 1986). Many different topological indices have been devised and, as for whole molecule properties, multiple different indices are normally used in combination (Estrada & Uriarte, 2001). Substructure-based descriptors characterise a molecule by the substructural features that it contains, either by the molecule's connection table (i.e., the underlying 2D chemical graph), or by its fragment bit-string: they are the most important type of descriptor for similarity studies and are thus described in some detail.

Early comparative studies demonstrated both the effectiveness and the efficiency of 2D fingerprints for similarity searching (Downs & Willett, 1995) and these continue to be very widely used (as reviewed by Hert, Willett, Wilton, Acklin, Azzaoui, Jacoby, et al. (2004b). In the past, the fingerprints were normally those used for substructure searching but there has been recent interest in alternative types that are intended specifically for similarity searching. The fragments in substructure-searching systems are typically small patterns of atoms and bonds, with the atoms being denoted by their elemental types, whereas the newer types of fingerprint describe atoms in terms of their associated properties. The aim of this change is to facilitate the retrieval of molecules that have similar properties to the reference structure in a similarity search but that have different sets of atoms, thus permitting the identification of new classes of molecules with the requisite bioactivity. The ability to discover such novel structural types is commonly referred to as *scaffold-hopping*, and is of considerable commercial importance since it provides a way of circumventing an existing patent (Böhm, Flohr, & Stahl, 2004; Brown & Jacoby, 2006; Schneider, Schneider, & Renner, 2006).

An early example of the use of atom-types is the work of Kearsley et al. (1996) at Merck, who characterised atoms as belonging to one of seven classes: cations; anions; hydrogen bond donors; hydrogen bond acceptors; atoms that are both donors and acceptors; hydrophobic atoms (i.e., atoms that prefer to interact with non-aqueous solvents rather than with water); and all others. A related, four-part classification is used to characterise atoms in the Similog fingerprints developed by Schuffenhauer *et al.* (2003) at Novartis, where a fragment consists of a triplet of atoms, together with the numbers of bonds separating each pair of them, and the frequency of occurrence of that triplet in the molecule. A similar triplet description has been described by Ewing et al. (2006), and a further four-part classification scheme is used in the CATS descriptor, which is based on counts of pairs of atoms separated by up to ten bonds (Schneider, Neidhart, Giller, & Schmid, 1999). The SciTegic fingerprints encode circular substructures surrounding each of the atoms in a molecule, with the atoms being described either by their elemental type or by means of a six-part classification scheme. The former type of SciTegic fingerprint was found to be the most effective in an extended comparison of a large number of 2D fingerprints for similarity searching, including not just all those described here but also those in other commercial and in-house chemoinformatics systems (Hert, Willett, Wilton, Acklin, Azzaoui, Jacoby, et al., 2004b); however, a comparably extended study found a much less consistent pattern of behaviour in a comparison of 2D fingerprints for QSAR applications (Gedeck, Rhode, & Bartels, 2006).

Arguably the most obvious way to compute the similarity between two 2D molecules is to compare their underlying chemical graphs, i.e., their connection tables, using an MCS isomorphism algorithm. However, these are extremely time-consuming, and thus not generally appropriate for use in a database context (where the MCS must be computed between the reference structure and each of the molecules in the database that is being searched) unless an approximate procedure is used (Hagadone, 1992; Sheridan & Miller, 1998). A recent, exact MCS algorithm has been described that is able to perform thousands of similarity comparisons a second (Raymond, Gardiner, & Willett, 2002a, 2002b). The method uses a number of sophisticated screening steps to eliminate compounds from the MCS calculation, together with a rapid graph-matching step that pre-processes the graphs so that the atoms become bonds and *vice versa*, this maximising the speed with which subgraph mis-matches can be identified. The algorithm has become the method of choice for computing similarities between 2D chemical graphs, and its availability enabled a comparison to be carried out for the first time of the effectiveness of fingerprint-based and graph-based virtual screening (Raymond & Willett, 2002). Surprisingly, it was found that the simpler, and computationally far more efficient, fingerprint representations were in no way inferior to the full chemical graphs in terms of the numbers of active molecules retrieved, although there were differences in the identities of the actives retrieved by the two approaches.

Rather than using the full chemical graph, there has been interest in using a condensed form that subsumes some of the individual atoms into higher-level structural aggregates: examples of this approach are the ***reduced graph*** (Gillet, Willett, & Bradshaw, 2003) and the *feature tree* (Rarey & Dixon, 1998). The level of aggregation can be chosen to highlight functional features in a molecule, such as ring systems or hydrophobic regions, in a manner analogous to the atom-typing described previously. A condensed graph provides a level of representation that is less complex than a full chemical graph, and that hence allows much more rapid processing. Indeed, just as a full graph can be searched at the graph level or at the level of fingerprints derived from it, so can a reduced graph or feature tree, and there have been several evaluations of the effectiveness of both types of representation for virtual screening (Barker, Cosgrove, Gardiner, Gillet, Kitts, & Willett, 2006; Barker, Gardiner, Gillet, Kitts, & Morris, 2003; Harper, Bravi, Pickett, Hussain, & Green, 2004; Rarey & Stahl, 2001; Stiefl, Watson, Baumann, & Zaliani, 2006; Stiefl & Zaliani, 2006; Takahashi, Sukekawa, & Sasaki, 1992). Bohl *et al*. (2006) describe a reduced-graph representation of the 3D structure of complex ring systems that has been found effective for scaffold-hopping searches.

Conventional similarity searching involves using a single reference structure as the basis for a search, but it may be the case that several actives are available. If they have structural commonalities it may be possible to derive a 2D or 3D pharmacophore, but this is often not possible, e.g., a set of diverse hits from an HTS experiment or molecules from the journal or patent literatures. There has hence been interest in ways of combining the information contained in the structures of the actives. Hessler *et al*. (2005) have suggested superimposing the feature-trees of individual molecules to form a multiple feature tree (or MTree) that is then matched against each of the molecules in the database. This overlaying of multiple actives is similar to an approach that has been described for the analysis of HTS data and the generation of 2D pharmacophores (Brown, Willett, Wilton, & Lewis, 2003). However, most work on using multiple actives has involved fingerprint representations, and Hert *et al*. (2004a) have reported a detailed comparison of ways in which the fingerprints of a set of known actives could be exploited. This study found that the most effective of the tested approaches involved carrying out a conventional similarity search using each of the set of actives in turn as the target, and then merging the individual rankings to give a single, combined ranking: this is an example of *data fusion*, which is described in some detail later in the review.

Structural Representations For 3D Similarity Searching

Similarity searching, at both the operational and the research level, continues to be dominated by the use of 2D representations; that said, the 3D characteristics of a molecule are a key determinant of biological activity and there have hence been many descriptions of measures that could be used for 3D searching. The methods can be organised on the basis of: whether they consider just single, low-energy conformations or take account of the conformational flexibility of most molecules; and whether they require that the molecules that are being compared are aligned with each other prior to the similarity calculation. Alignment procedures, in which molecules are superimposed so as maximise the degree of structural fit, are described in an excellent review by Lemmen and Lengauer (2000).

Substructural descriptors have been discussed previously when considering 2D representations for similarity searching. This approach can be extended to encompass the use of geometric information such as pairs of atoms and the associated inter-atomic distances, with early such studies being reviewed by Willett *et al.* (1998). The approach can also encompass conformationally flexible molecules by superimposing the fingerprints for multiple low-energy conformers of a molecule, and this has formed the basis for fingerprints based on fragments describing three-point or four-point pharmacophores. Here, atoms are classified by their potential pharmacophore characteristics (as discussed previously for 2D fingerprints) and then the distances computed between sets of three or sets of four atoms to yield the fragment that is encoded in a fingerprint (Mason, Morize, Menard, Cheney, Hulme, & Labaudiniere, 1999; Pickett, Mason, & McLay, 1996). These pharmacophore descriptors have been shown to be effective search tools (Good, Cho, & Mason, 2004; Zhang & Muegge, 2006) and are now available in commercial chemoinformatics software systems. The 3D graphs themselves can be compared using an MCS procedure, but this is even more time-consuming than in the 2D case, especially when the molecules are allowed to be flexible (Raymond & Willett, 2003).

Studies of 3D QSAR have demonstrated the importance of molecular fields and this has been reflected in the use of field-based descriptors for 3D similarity searching. The basic approach involves computing the electrostatic, steric or hydrophobic field at each point of a 3D grid surrounding a molecule and then aligning the grids describing two molecules so as to maximise the fit of the computed field values. Similarities based on the alignment of grids were first described by Carbo *et al.* (1980) but the approach aroused little interest because of the computational costs required to align the two molecules and then to compute the field-based similarity. Good *et al.* (1992) described an approach based on the use of Gaussian approximations to the exact form of the distribution of charge around a molecule (the molecular electrostatic potential, or MEP) that permitted a substantial increase in the speed of the similarity calculation. Gaussian methods are widely used in computational chemistry to enhance the efficiency of quantum mechanical calculations, and their use by Good et al. encouraged the use of MCS algorithms and genetic algorithms that could carry out the alignment stage sufficiently rapidly to allow the implementation of MEP-based 3D similarity searching (Thorner, Wild, Willett, & Wright, 1996; Thorner, Willett, Wright, & Taylor, 1997; Wild & Willett, 1996). A subsequent study showed that such approaches yielded noticeably different sets of molecules than did conventional 2D fingerprints searches (Schuffenhauer, Gillet, & Willett, 2000) hence making it suitable for scaffold-hopping applications (Bohl et al., 2002). The effectiveness of MEP-based similarity searching was further demonstrated in studies by Mestres and co-workers (Mestres & Knegtel, 2000; Mestres, Rohrer, & Maggiora, 1997; Mestres, Rohrer, & Maggiora, 2000), with CRESSET being an example of a commercial software package that uses electrostatics-based similarity searching for drug discovery (Cheeseright, Mackey, Rose, & Vinter, 2006; Low, Buck, Cooke, Cushnir, Kalindjian, Kotecha, et al., 2005).

Gaussian methods can also be used to characterise the steric characteristics of molecules (Good & Richards, 1993; Grant, Gallardo, & Pickup, 1996), with ROCS being a commercial package that uses this approach to similarity searching (Hawkins, Skillman, & Nicholls, 2007; Rush, Grant, Mosyak, & Nicholls, 2005). A recent paper described a fingerprint representation that uses the Gaussian approach to shape similarity together with a standard dictionary of molecular shapes (in a manner reminiscent of the affinity fingerprints mentioned previously). Each bit in the fingerprint describing a molecule is set if that molecule has a Gaussian shape similarity with the corresponding standard shape that exceeds some threshold value (Haigh, Pickup, Grant, & Nicholls, 2005) (an analogous idea, but using 2D similarities to a set of standard molecules, forms the basis of the SIBAR (for Similarity-Based SAR) system (Klein, Kaiser, Kopp, Chiba, & Ecker, 2002)). Ballester and Richards report a simpler shape fingerprint that summarises the main components of the distribution of distances in a molecule (Ballester & Richards, 2007a, 2007b). All of these approaches enable the detection of shape resemblance in both an efficient and an effective manner.

There are many further types of 3D structural descriptor that can be generated using the methods of computational chemistry, but these are currently far too time-consuming to be used routinely in a database context.

Evaluation Of Similarity Measures

The combination of the many similarity coefficients, structure representations and search algorithms that are available means that there are now a very large number of different methods that could be used for similarity-based virtual screening: there is hence a need to measure the effectiveness of the various procedures. The most important evaluation criterion in the context of virtual screening is the ability to retrieve molecules from a database that prove, on inspection, to share the same bioactivity as the reference structure: only if this is the case will a particular similarity measure be useful for virtual screening (in much the same way as an effective IR procedure is one that is able to retrieve relevant documents). There have been several reviews of performance criteria (Dixon & Merz, 2001; Edgar, Holliday, & Willett, 2000; Willett, 2004), these including: the enrichment factor, i.e., the number of actives retrieved relative to the number that would have been retrieved if compounds had been picked from the database at random (Kearsley, Sallamack, Fluder, Andose, Mosley, & Sheridan, 1996); the numbers of actives that have been retrieved at some fixed position in the ranking, e.g., the top 1% of the ranked database (Hert, Willett, Wilton, Acklin, Azzaoui, Jacoby, et al., 2004a); and (increasingly) the receiver operating characteristic curve (Cuissart, Touffet, Crémilleux, Bureau, & Rault, 2002; Triballeau, Acher, Brabet, Pin, & Bertrand, 2005) although the use of this last criterion for virtual screening experiments has been criticised recently (Truchon & Bayly, 2007).

The numbers of active molecules retrieved, in essence a recall parameter, is clearly of great importance, but Good *et al*. have emphasised the need for structural novelty in comparing screening methods in the industrial context (Good, Hermsmeier, & Hindle, 2004). Thus, a search method retrieving 50 molecules that prove to be active but that all belong to the same structural class as the reference structure is probably less useful than a method that retrieves five molecules that prove to be active but that belong to two different structural classes, i.e., the scaffold-hopping capability that has been discussed above. Good *et al*. hence argue that comparisons of screening methods should focus less on the numbers of active molecules and more on the numbers of active classes, although Hert *et al*. (2004b) found that the two approaches provided very similar rankings of methods when evaluating virtual screening experiments using 2D fingerprints.

It is not possible in a review of this sort to discuss the details of the very many comparisons of similarity methods that have been carried out; instead the reader is referred to existing reviews of molecular similarity (Bender & Glen, 2004; Glen & Adams, 2006; Maldonado, Doucet, Petitjean, & Fan, 2006; Nikolova & Jaworska, 2003; Sheridan & Kearsley, 2002; Willett, Barnard, & Downs, 1998). Reference will, however, be made to the work of Brown and Martin (1996, 1997b). These much-cited studies involved an extensive comparison of a range of structural descriptors, both 2D and 3D, for the clustering of chemical databases and for the analysis of ligand binding. The study is notable not only for the breadth of the experiments but also for the conclusion that simple 2D fingerprints were superior to more sophisticated 3D descriptors. This would appear to be counter-intuitive, given the importance of steric effects in determining biological activity, but subsequent experiments have shown that 2D representations are indeed generally superior to 3D representations for similarity-based virtual screening (and also diversity analysis). There are several reasons why this might be so. For example, conformational effects may best be left unconsidered at the simple similarity level, although they are, of course, of paramount importance in more sophisticated types of processing such as ligand docking (in much the same way as simple "bag of words" approaches are often found to be more effective for IR than sophisticated approaches based on natural language processing). Alternatively, it may be that the experimental environment, where averaging takes place over large numbers of reference structures, favours approaches that perform at a reasonable level in all (or most) circumstances, which is certainly the case with 2D fingerprints. Finally, it may simply be that the appropriate way of encoding 3D information remains to be identified. Whatever the reason, there is a continuing debate as to the relative effectiveness of 2D and 3D measures (see, e.g., (Cramer, Jilek, Guessregen, Clark, Wendt, & Clark, 2004; Cruciani, Pastor, & Mannhold, 2002; Gedeck, Rhode, & Bartels, 2006; Makara, 2001; Matter, 1997; Schuffenhauer, Gillet, & Willett, 2000; Sheridan & Kearsley, 2002) *inter alia*).

Finally here, there has been some interest in assessing the statistical significance, or otherwise, of computed inter-molecular similarity values. The basic approach has been to use randomisation procedures of various sorts to generate null hypotheses against which observed distributions of similarity values can be compared. The first such study was by Bradshaw and Sayle, assessing the significance of fingerprint-based similarities using an approach based on those that have been developed to assess the significance of protein sequence homology (Bradshaw & Sayle, 1997). Other fingerprint-based studies have been reported (Holliday, Salim, & Willett, 2005; Keiser, Roth, Armbruster, Ernsberger, Irwin, & Shoichet, 2007), as has the significance of MCS-based similarities (Sheridan & Miller, 1998).

COMBINATION OF SIMILARITY SEARCHES USING DATA FUSION

Given the range of similarity methods that are now available, there have been many comparative studies that seek to identify the "best" virtual-screening method; however, it seems inherently unlikely that any single method could be expected to perform equally well under all circumstances (Sheridan, 2007; Sheridan & Kearsley, 2002). A more realistic approach, instead, is to carry out multiple searches and to combine the results using the methods of *data fusion* (Hall, 1992; Klein, 1999). This refers to a range of techniques for the detection, capture, pre-processing and combination of multiple sources of digital information, typically the signals output by some type (or types) of sensor. Data fusion methods were first developed for military applications but are now used in a wide range of contexts: here, we shall consider the combination of different rankings of a database that result from the use of different searching methods, an approach that is well established in the IR literature (Belkin, Kantor, Fox, & Shaw, 1995; Croft, 2000).

The basic idea is a simple one. Assume that it is possible to carry out multiple similarity searches, e.g., searches using three different types of 2D fingerprint. A search is carried out using the first fingerprint-type to describe the reference structure and each of the database structures, and the database ranked in decreasing order of the computed similarity. The procedure is repeated using the other two types of fingerprint, and the three database rankings are then combined using a fusion rule. A typical rule involves summing the ranks obtained by each specific molecule in each of the searches; the final output is then obtained by ranking the database molecules in decreasing order of these computed sums. Data fusion was first used for similarity searching in the mid-Nineties (Ginn, Turner, Willett, Ferguson, & Heritage, 1997; Kearsley, Sallamack, Fluder, Andose, Mosley, & Sheridan, 1996; Sheridan, Miller, Underwood, & Kearsley, 1996) with applications in docking, where the approach is called *consensus scoring*, appearing shortly afterwards (Charifsen, Corkery, Murcko, & Walters, 1999; Clark, Strizhev, Leonard, Blake, & Matthew, 2002; Stahl & Rarey, 2001). There has now been extensive work in these two areas, as detailed in two recent reviews (Feher, 2006; Willett, 2006).

Early studies of data fusion involved combining searches that were based on different types of structural representation. For example, two papers by the group at Merck involved searching using fingerprints based on pairs of atoms (described either by elemental type or by physicochemical properties) and the associated inter-atomic distances (either through-bond or through-space). Searches were carried out using pairs of types of fingerprint and then the final score for a database molecule was the sum of its scores in the two individual searches, or its higher rank position in the two searches (Kearsley, Sallamack, Fluder, Andose, Mosley, & Sheridan, 1996; (Ginn, Turner, Willett, Ferguson, & Heritage, 1997; Kearsley, Sallamack, Fluder, Andose, Mosley, & Sheridan, 1996; Sheridan, Miller, Underwood, & Kearsley, 1996). A wide range of types of representation - including 2D fingerprints, vectors of physicochemical properties (such as molecular weight, number of rings, and molecular volume), MEP descriptors, and infra-red spectral descriptors – and of combination rules were studied by Ginn *et al*. (1997, 2000). Both the Merck and Sheffield studies suggested that combined searches yielded at least as many bioactive molecules as could the best of the individual searches that were being combined: since the latter often varied unpredictably from one search to another search, it was concluded that the use of a fusion rule would generally provide a more consistent level of search performance than would a single similarity measure. There is continuing interest in combining different representations for similarity searching (Kogej, Engkvist, Blomberg, & Muresan, 2006; Zhang & Muegge, 2006), and there has also been interest in combining different similarity coefficients (Holliday, Hu, & Willett, 2002; Salim, Holliday, & Willett, 2003).

There are many different fusion rules, including the use of voting procedures, both weighted and unweighted arithmetic functions, and statistical procedures that require the availability of training data (Feher, 2006). A comparison of six fusion rules for similarity searching showed that the best virtual-screening performance came from the use of a logistic regression procedure; this requires training data and it was hence suggested that it might best be used in the lead-optimisation of a discovery project when such data would have become available (Baber, Shirley, Gao, & Feher, 2006). Training data is also required for the conditional probability fusion rule (Raymond, Jalaie, & Bradley, 2004), which seeks to estimate the probability that a database molecule is active given the value of its similarity score to a reference structure. The overall score for the database molecule is then the product of the probabilities computed for each of the individual scoring functions; a similar approach has been described for ranking documents in IR (Manmatha, Rath, & F. Feng, 2001). Several other comparisons of fusion rules have been reported (Oda, Tsuchida, Takakura, Yamaotsu, & Hirono, 2006; Yang, Chen, Shen, Kristal, & Hsu, 2005; Zhang & Muegge, 2006).

The discussion of data fusion thus far has assumed that a single reference structure forms the basis for a series of searches, each using a different representation or coefficient. The use of a single molecule but multiple similarity measures is referred to by Whittle *et al*. (2004) as *similarity fusion*. An alternative approach, *group fusion*, involves a single similarity measure (e.g., the Tanimoto coefficient and 2D fingerprints) but multiple reference structures. Drawing on earlier work (Schuffenhauer, Floersheim, Acklin, &Jacoby, 2003; Xue, Stahura, Godden, & Bajorath, 2001), Hert *et al*. (2004a) studied the search-effectiveness of group fusion, comparing it with conventional similarity searching and with other ways of combining the results of multiple similarity searches. Using one particular fusion rule – specifically taking the maximum similarity coefficient for a database molecule with each of the multiple reference structures – they found that picking as few as ten active reference structures and combining them using group fusion enabled searches to be carried out that were comparable to even the very best from amongst many hundreds of conventional similarity searches using individual reference structures. Further studies (Hert, Willett, Wilton, Acklin, Azzaoui, Jacoby, et al., 2006; Whittle, Gillet, Willett, Alex, & Loesel, 2004) demonstrated that the benefits of group fusion are greatest when the sought actives are structurally diverse; conventional similarity searching or similarity fusion, conversely, are most effective when the actives are strongly clustered in structural space. The group fusion approach has now been taken up by other workers, who have confirmed its effectiveness as a general tool for similarity searching (Williams, 2006; Zhang & Muegge, 2006).

Hert *et al*. (2005, 2006) have also described a modification of conventional similarity searching that makes use of group fusion. Given a bioactive reference structure, the top-ranked structures resulting from a similarity search are expected to have a high probability of activity as a consequence of the Similar Property Principle. The assumption was then made that such molecules are indeed active (rather than just probably active) and that they can hence be used as the reference structures for further similarity searches; the rankings from the set of searches are then combined using group fusion. Extensive testing demonstrated that this activity-assumption resulted in searches that were nearly always superior to conventional similarity searching (where just the initial reference structure is used) in its ability to identify active molecules, with some of the increases in performance being quite marked.

Most studies of combination methods have found that they result in an improvement in screening performance. In some cases, it has been found that the combined search is better than even the best of the individual searches that are being fused; more generally, the combined search is found to be comparable to the best individual search or better than the average function. Importantly, the best individual search is quite variable, whereas an effective fusion rule is robust to changes in reference structure, database and biological activity, ensuring a consistent level of search performance. There is hence considerable experimental support for the use of data fusion or consensus scoring but it is only recently that there has been interest in the theoretical basis of the various combination methods that have been studied, whereas this has been the subject of debate in the IR context for some years (Beitzel, Jensen, Chowdhury, Grossman, Goharian, & Frieder, 2004; Hsu & Taksa, 2005; Lee, 1997; Ng & Kantor, 2000). The first investigation in the chemoinformatics area was a simulation study (Wang & Wang, 2001). This simulation demonstrated that performance initially increased rapidly with an increase in the number of scoring functions, but then levelled off after three or four functions had been included in the combination, and that the consensus performance was superior to any individual scoring function. The latter result was explained by the simple statistical fact that the mean of repeated samplings will tend to be closer to the true value than any individual sampling: in other words, multiple rankings will better reproduce the ideal database ranking than will any individual ranking. However, the study has been criticised (Baber, Shirley, Gao, & Feher, 2006; Verdonk, Berdini, Hartshorn, Mooij, Murray, Taylor, et al., 2004) as it assumes that the scoring functions that are being combined are of comparable effectiveness and are independent; neither of which is likely to be the case in practice. An empirical study of data fusion (Baber, Shirley, Gao, & Feher, 2006) showed that active molecules are more tightly clustered than are inactive molecules (as would be expected if the Similar Property Principle holds.) Thus when multiple scoring functions are used they are likely to repeatedly select many actives but not necessarily the same inactives, a finding that is paralleled by work in IR (Lee, 1997).

A rigorous theoretical approach to the modelling of data fusion has been reported recently (Whittle, Gillet, Willett, & Loesel, 2006a, 2006b). The theoretical model shows that the origin of performance enhancement for simple fusion rules can be traced to a combination of differences between the retrieved active (i.e., true positives) and retrieved inactive (i.e., false positives) similarity distributions and the geometrical difference between the regions of these multivariate distributions that the chosen fusion rule is able to access. There are many factors involved, e.g., the fusion of just two lists of similarity values (e.g., similarity searches using the Tanimoto coefficient and the cosine coefficient) depends on no less than eight distinct distributions. These complex interactions mean that it is extremely difficult to predict with any degree of accuracy the exact level of search performance that will result from the use of data fusion. The analysis was able to demonstrate that improvements over conventional similarity searching should be obtainable on a routine basis if large amounts of training data are available; however, this is not normally the case in the early stages of drug-discovery programmes where similarity searching is most commonly used.

OTHER APPLICATIONS OF MOLECULAR SIMILARITY

Measures of inter-molecular structural similarity lie at heart of similarity-based virtual screening, as they also do at the heart of two related topics: the clustering of chemical structures and methods for molecular diversity analysis.

Clustering Databases Of Chemical Molecules

Similarity searching involves comparing a single molecule, the reference structure, with each of the molecules in a database; clustering involves (in many clustering methods) comparing every molecule in a database with every other molecule, and thus many of the comments above regarding the types of measure that are used and their effectiveness are equally applicable to the calculation of the similarities involved in clustering applications. Clustering is the process of subdividing a group of objects (chemical molecules in the present context) into groups, or clusters, of objects that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity (Everitt, Landau, & Leese, 2001; Sneath & Sokal, 1973). It is thus possible to obtain an overview of the range of structural types present within a dataset by selecting one (or some small number) of the molecules from each of the clusters resulting from the application of an appropriate clustering method to that dataset. The representative molecule for each cluster is either selected at random or selected as being the closest to the centre of that cluster, and can be used to maximise the efficiency of random screening in lead-discovery programmes. Thus, if a representative molecule proves active when tested in the bioassay of interest then it is appropriate to assay the other compounds in its cluster since these may also exhibit the activity of interest; alternatively, if the representative molecule proves to be inactive then attention should be transferred to another cluster (Downs & Willett, 1994).

Very many different clustering methods have been described in the literature. Early studies of over 30 hierarchic and non-hierarchic methods (Willett, 1987) showed that the best results were obtained from Ward's hierarchical-agglomerative method (Ward, 1963), with the non-hierarchical nearest-neighbour method of Jarvis and Patrick (1973) performing almost as well. The latter is far faster in operation and was hence the method of choice for many years not only for selecting molecules for random screening but also for clustering the outputs of substructure searches that retrieve very large numbers of molecules, thus providing the searcher with an overview of the structural classes that contain the substructure of interest. However, the method does have limitations (see, *e.g.*, (Doman, Cibulskis, Cibulskis, McCray, & Spangler, 1996)) and improvements in hardware and software, such as the efficient reciprocal nearest neighbours algorithm (Murtagh, 1985), and further comparative studies (Brown & Martin, 1996) have led to Ward's method becoming accepted as the standard method for files containing up to ca. half-a-million structures. Ward's is an example of an agglomerative hierarchical clustering method, in which the most similar molecules or clusters of molecules are progressively merged until all of the molecules are in a single cluster. The sequence or mergers, or agglomerations, results in a hierarchy, called a *dendrogram*, and a level in this hierarchy must then be chosen to obtain a set of clusters. A comparison of the many methods available for this purpose (Wild & Blankley, 2000) showed that the Kelley index (Kelley, Gardner, & Sutcliffe, 1996) seemed to yield the most generally useful clusters of molecules.

Developments in clustering since the last ARIST review (Paris, 1997) have focused on the evaluation of further types of clustering method. For example, the Ward and Jarvis-Patrick methods were compared with two clustering methods that had been developed for the processing of gene expression data and one that had been developed for the fuzzy clustering of chemical graphs (Raymond, Blankley, & Willett, 2003). The study found that the two gene-expression methods (Ben-Dor, Shamir, & Yakhini, 1999; Yin & Chen, 1994) provided robust and effective approaches to the clustering of a range of datasets, and suggested that these methods were worthy of further consideration. The experiments involved both fingerprint-based and graph-based measures of 2D similarity, with no obvious advantage being identified from the use of the latter, more time-consuming measure. A study of the fuzzy k-means method for clustering chemical structures (Holliday, Rodgers, Willett, Chen, Mahfouf, Lawson, et al., 2004) showed that this gave generally superior results to the conventional, non-fuzzy k–means clustering method, a non-hierarchic relocation procedure (Everitt, Landau, & Leese, 2001). However, the most important development has been the identification of the *divisive k-means* clustering method (Boecker, Derksen, Schmidt, Teckentrup, & Schneider, 2005; Steinbach, Karypis, & Kumar, 2000) as an effective way of clustering chemical datasets that are too large for processing by Ward's method. The method, which has also been used in bioinformatics by Sultan et al. (2002), is an hierarchical version of the k-means method, and has been shown to be well suited to applications such as comparing the effectiveness of classifications based on substructural fragments and on ring scaffolds (Schuffenhauer, Brown, Ertl, Jenkins, Selzer, & Hamon, 2007) and the merging of corporate databases (Engels, Gibbs, Jaeger, Verbinnen, Lobanov, & Agrafiotis, 2006).

Downs and Barnard (2002) provide an authoritative overview of developments in chemical clustering.

Molecular Diversity Analysis

Much of the recent work in clustering has been carried out to support work in molecular diversity analysis. This has been introduced previously and takes as its starting point the need to maximise the diversity of the molecules that are submitted for biological testing (rather than maximising the probability of activity, which is the main aim of virtual screening) (Dean & Lewis, 1999; Ghose & Viswanadhan, 2001; Gorse, 2006; Lewis, Pickett, & Clark, 2000). Although HTS is very rapid, it is still costly and there is hence a need to minimise the numbers of molecules that are assayed. The Similar Property Principle means that structurally similar molecules are likely to give similar biological responses; thus, if one wishes to maximise the information that can be gained from a fixed number of molecules about the relationship between structure and activity, then one should try to ensure that the molecules submitted for HTS are as structurally diverse as possible. This has been formalised in the concept of *Neighbourhood Behaviour*, which is, in essence, a quantitative reformulation of the Similar Property Principle that focuses on differences in biological activity and in molecular similarity. Specifically, the approach considers absolute differences in biological activity (as measured on some quantitative scale, rather than the active/inactive nature of much bioactivity data) for all pairs of molecules in a dataset, and plots these differences against the dissimilarity values for these pairs. If the molecular description that is being used exhibits a good Neighbourhood Behaviour then there will be few points in the resulting plot that correspond to a large difference in property being associated with a small dissimilarity value; however, the remainder of the plot is expected to be heavily populated. Analysis of such plots hence provides a simple way of validating the performance of a molecular descriptor for similarity and diversity applications. Initial studies (Dixon & Merz, 2001; Patterson, Cramer, Ferguson, Clark, & Weinberger, 1996) used QSAR datasets, containing small numbers of structurally related molecules, for the comparison of descriptors, but the utility of the approach on a large-scale has been demonstrated recently by Perekhodtsev (2007). This study involved combining multiple literature sources to produce four large sets of structurally heterogeneous molecules, and showed that excellent Neighbourhood Behaviour was exhibited by simple 2D fingerprints.

Molecular diversity methods established themselves in the mid-Nineties: initial studies were discussed in the previous ARIST review (Paris, 1997), and an historical overview of much of this early work has been reported by Martin *et al*. (2001). Since then, four main approaches to the selection of compounds have been identified. The first, and still probably the most widely used, is cluster analysis, with a database subset being obtained by selecting one, or a small number, of the compounds from each of the clusters identified by the clustering method (*vide supra*). Cluster-based selection methods require the explicit computation of inter-molecular similarity; *partition-based* (or *cell-based*) methods compute the similarities implicitly by defining a low-dimensional grid in which molecules can be located. The grid is defined by a small number of physicochemical properties such as molecular weight or calculated logP, and a molecule is allocated to that cell in the grid that corresponds to its particular values for the chosen properties. A diverse subset is then obtained by selecting one or more molecules from each cell. Mason and Pickett (1997) review early work on this approach, whilst Bayley and Willett (1999) describe statistical criteria for the design of the grid. Partition-based methods are widely used, normally by means of the BCUT software (Pearlman & Smith, 1998, 1999).

Cluster-based and partition-based selection involve finding the most similar molecules before identifying a diverse subset: *dissimilarity-based selection* methods derive from early work by Bawden (1993) and Lajiness (1990) and tackle the problem directly. In essence, the methods select a molecule for inclusion in the chosen subset, and then repeatedly add to the subset that molecule from the database that is most dissimilar to those already in the selected subset, with the dissimilarity calculations continuing until the subset is of the required size. A comparison of algorithms for this purpose (Snarey, Terrett, Willett, & Wilton, 1997) highlighted the general effectiveness of the MaxMin algorithm, in which that database molecule is selected at each stage that has the maximum dissimilarity to its nearest neighbour in the subset of already-chosen molecules. This comparison used 2D fingerprints, and these are probably the most popular structure representation for selection purposes. Sets of a few physicochemical properties are also widely employed (as for partition-based selection), with the use of a limited numbers of descriptors enabling the use of a particularly rapid implementation of the MaxMin algorithm (Agrafiotis & Lobanov, 1999). *Sphere-exclusion* algorithms are closely related to dissimilarity-based algorithms, but use a dissimilarity threshold to eliminate molecules from consideration that are close to previously selected molecules. The threshold defines the radius of an exclusion sphere and each time that a molecule is selected for inclusion in the chosen subset, then all database molecules within the threshold distance are excluded from further consideration (Hudson, Hyde, Rahr, &

Wood, 1996; Pearlman & Smith, 1998); a similar idea underlies the single-pass clustering methods that have been described for chemical applications (Butina, 1999; Taylor, 1995). The OptiSim algorithm involves concepts from both dissimilarity-based and sphere-exclusion selection, with a user-defined parameter specifying the extent to which the two types of selection criterion are reflected in the final set of molecules that is selected (Clark, 1997; Clark & Langton, 1998). Systematic selection methods such as those described above are now widely used, although there have been suggestions that the resulting compound selections are not obviously superior to those obtained by random selection (Schuffenhauer, Brown, Selzer, Ertl, & Jacoby, 2006).

An alternative approach to compound selection formulates the identification of the most diverse subset as a problem in combinatorial optimisation. Whilst methods based on D-optimal designs have been described (Martin, Blaney, Siani, Spellmeyer, Wong, & Moos, 1995), most work has involved the use of genetic algorithms or simulated annealing. The approach requires the availability of a *diversity index*, some quantitative measure of the degree of structural diversity in a set of molecules: an optimal (or, more likely, near-optimal) subset can then be obtained by exploring the space of all possible subsets to find that with the largest value of the chosen index. There have been reports of methods for the design of structurally diverse combinatorial libraries using both genetic algorithms (Brown & Martin, 1997a; Gillet, Willett, & Bradshaw, 1997; Sheridan & Kearsley, 1995) and simulated annealing (Agrafiotis, 1997; Good & Lewis, 1997; Hassan, Bielawski, Hempel, & Waldman, 1996) to effect the optimisation. Many diversity indices have been described: for example, Gillet *et al.* (1997) quantified the diversity of a set of molecules by the mean pair-wise complement of the Tanimoto coefficient when averaged over all the pairs of 2D fingerprints for molecules in the subset whilst Good and Lewis (1997) quantified the diversity by the number of distinct three-point 3D pharmacophores identified in the molecules in the subset. Waldman *et al.* (2000) provide a detailed review of diversity indices and of their use for the selection of molecules.

The scoring function in the program by Good and Lewis (1997) involved not just the maximisation of the structural diversity but also an attempt to ensure an approximately even distribution of a set of molecules across three properties that provide a crude, but rapidly computable, measure of molecular shape. The idea of including a range of factors in a scoring function has since been studied in detail. Initial studies used a simple weighted sum approach (Agrafiotis, 2002; Brown, Hassan, & Waldman, 2000; Gillet, Willett, Bradshaw, & Green, 1999), but this requires specifying the weights for the scoring function’s individual components, which is particularly difficult if these components are in competition (e.g., maximising diversity is often in conflict with ease of synthesis). This problem has been addressed by means of a multi-objective genetic algorithm (MOGA) in which the various components are optimised independently and a family of equivalent solutions is obtained by means of Pareto optimization: each solution then represents a different trade-off between the often conflicting requirements of the various objectives (Gillet, 2004; Gillet, Khatib, Willett, Fleming, & Green, 2002).

OTHER VIRTUAL SCREENING METHODS

We have noted previously that similarity searching is but one of a range of methods that are available for virtual screening, and this section briefly describes developments in some of these other methods. There have historically been four main approaches - similarity searching, 3D pharmacophore searching, machine learning and docking – but these have now been augmented by two further, and closely related, approaches: drug-likeness and ADMET (for absorption, distribution, metabolism, excretion and toxicity, *vide infra*) prediction. Of these, similarity searching has been discussed in great detail above, as it forms the principal focus of the review; while pharmacophore mapping followed by 3D substructure searching had already become well-established by the time of the previous ARIST review (Paris, 1997), although there is continuing interest, as exemplified in studies by Dixon *et al.* (2006), Steindl *et al.* (2006) and Kurogi and Guner (2001), *inter alia*.

Machine Learning

Machine learning involves the assignment of objects of unknown classification (these objects comprising the *test-set*) to one of two or more existing classes, the classification rule having been derived from analysis of a set of objects of known classification (these objects comprising the *training-set*). For example, in the IR context, this might involve the categorisation of a newswire to one or more types of news story, or the assignment of a document to the class of relevant (or non-relevant) for some particular query. When machine learning is used in chemoinformatics, the task is to categorise a molecule as either active or inactive. In practice, some of the methods available are used for ranking, rather then categorisation, with a scoring scheme being used to rank molecules in decreasing probability of activity, rather than setting a threshold value above which the molecule will be judged active.

As noted previously, the first such method to be used in chemoinformatics was substructural analysis (Cramer, Redl, & Berkoff, 1974). After relative neglect for many years, there has been renewed interest in this approach (Capelli, Feriani, Tedesco, & Pozzan, 2006; Cosgrove & Willett, 1998; Medina-Franco, Petit, & Maggiora, 2006; Schreyer, Parker, & Maggiora, 2004), not least because it has been realised that the weighting schemes used in substructural analysis are very close to those used in naive Bayesian classifiers, one of the most widely used tools for machine learning (Hert, Willett, Wilton, Acklin, Azzaoui, Jacoby, et al., 2006). Indeed, some of the weighting schemes that are used in substructural analysis are closely related to the weights developed by Robertson-and Spärck Jones (1976) for use in IR relevance-feedback systems. Examples of the use of naive Bayesian classifiers (Bender, Mussa, Glen, & Reiling, 2004; Xia, Maliski, Gallant, & Rogers, 2004) are becoming increasingly common, spurred in part by the availability of this approach in the popular SciTegic software system (Hassan, Brown, Varma-O'Brien, & Rogers, 2006; Rogers, Brown, & Hahn, 2005). Although substructural analysis was first described over three decades ago, neural networks were the first machine-learning approach to be widely applied in chemoinformatics, especially for QSAR applications from the early Nineties onwards. There is a continuing literature, but we refer the reader to the book by Zupan and Gasteiger (1999) for a detailed account of work in what is by now a well-established area.

Decision trees have been extensively used in chemoinformatics, the aim being to construct a tree that can predict correctly the activity or otherwise of test-set molecules by seeing whether they are assigned to a node containing mainly active, or mainly inactive members of the training-set. An early example of the use of decision trees was reported by A-Razzak and Glen (1992), who used trees based on physicochemical properties to derive simple rules for the rationalisation of QSAR datasets; simple chemical and physicochemical properties have also been used for categorising molecules as drug-like or non-drug-like (Ajay, Walters, & Murcko, 1998; Wagener & van Geerestein, 2000). At the heart of any decision-tree procedure is the splitting criterion that is used to decide how the tree should be sub-divided at each level in the hierarchy. Given the importance of fragment substructures in chemoinformatics, it is not surprising that chemical fragments have been used for tree-construction, most notably in the work of Rusinko and colleagues, who have described the use of their SCAM program (Statistical Classification of Activities of Molecules) for the analysis of HTS and QSAR data (Chen, Rusinko, & Young, 1998; Hawkins, Young, & Rusinko, 1997; Rusinko, Farmen, Lambert, Brown, & Young, 1999). Over-fitting (Hawkins, 2004) can be a problem with decision trees, and there has hence been interest in the use of random forests (Breiman, 2001) for chemoinformatics applications, where a robust categorisation of a test-set molecule is obtained by matching it against multiple decision trees (e.g., (Cannon, Bender, Palmer, & Mitchell, 2006; Svetnik, Liaw, Tong, Culberson, Sheridan, & Feuston, 2003; Tong, Hong, Fang, Xie, & Perkins, 2003)).

Kernel methods figure prominently in the machine-learning literature (Christianini & Shawe-Taylor, 2000) and are now starting to be used in chemoinformatics. The methods have traditionally used small numbers of non-binary descriptors to characterise objects, but Harper *et al*. (2001) have developed the use of a kernel method, called binary kernel discrimination (or BKD), that scores test-set molecules by a function based on the ratio of the sums of its weighted Hamming distances from the active and from the inactive molecules in the training-set. Detailed studies of BKD have shown its effectiveness for the analysis of both agrochemical and pharmaceutical screening data (Chen, Harrison, Pasupa, Wilton, Willett, Wood, et al., 2006; Jorissen & Gilson, 2005; Willett, Wilton, Hartzoulakis, Tang, Ford, & Madge, 2007; Wilton, Harrison, Willett, Delaney, Lawson, & Mullier, 2006). However, most attention has focussed on the use of the support vector machine (SVM), which seeks to identify a hyperplane that maximally separates the two classes present in the training-set. The first chemoinformatics application appeared in 2001 (Burbridge, Trotter, Buxton, & Holden, 2001) and since then there has been a very large number of publications looking at not only virtual screening (Jorissen & Gilson, 2005; Saeh, Lyne, Takasaki, & Cosgrove, 2005; Warmuth, Liao, Ratsch, Mathieson, Putta, & Lemmen, 2003) but also at other important applications such as separation of drugs from non-drugs (Byvatov, Fechner, Sadowski, & Schneider, 2003; Muller, Ratsch, Sonnenburg, Mika, Grimm, & Heinrich, 2005), and the prediction of aqueous solubility (Lind & Maltseva, 2003) and of mutagenicity (Mahe, Ueda, Akutsu, Perret, & Vert, 2005), *inter alia*.

Further applications of machine learning in chemoinformatics are reviewed by Goldman and Walters (2006).

Docking

The first docking program was reported in the early Eighties (*vide supra*), with a study of non-peptidic inhibitors for HIV protease (DesJarlais, Seibel, Kuntz, Furth, Alvarez, de Montellano, et al., 1990) being one of the first uses of the approach for virtual screening. However, it is only within the last decade that docking has come to play an important, albeit still computationally demanding, role in drug discovery programmes. Its recent popularity has come about for two main reasons: the increasing availability of 3D structural information for many proteins of pharmaceutical interest (arising from developments in the technology of X-ray crystallography and protein NMR (Congreve, Murray, & Blundell, 2005)); and the incorporation of ligand flexibility in docking algorithms, so that screening is not restricted to files of rigid molecules (the further incorporation of protein flexibility is still at an early stage (Erickson, Jalaie, Robertson, Lewis, & Vieth, 2004)).

There are now about 30 different docking programs available (Leach, Shoichet, & Peishoff, 2006), with important current examples including DOCK (Ewing, Makino, Skillman, & Kuntz, 2001), FlexX (Rarey, Kramer, Lengauer, & Klebe, 1996), GLIDE (Friesner, Banks, Murphy, Halgren, Klicic, Mainz, et al., 2004) and GOLD (Jones, Willett, Glen, Leach, & Taylor, 1997). These programs are designed to carry out three principal tasks. The first is predicting the binding mode of a known active molecule, i.e., identifying the conformation that is adopted by a molecule when it binds to the active site of a protein. This information can provide a rationalisation for the observed activity and can also be of benefit in identifying related molecules that could make comparable, or superior, interactions with the protein. The second task is virtual screening, the subject of this review, something that is now being carried out on a production-line basis in most major pharmaceutical companies, typically using large Linux clusters to provide the necessary computational resources. As each molecule is docked into the protein, a score is computed that measures the degree of complementarity of that molecule to that binding-site. The scores can then be used to produce a rank ordering of the molecules in the database that is being searched, and hence to prioritise molecules for further testing. These scores could ideally be used for the third task, the accurate prediction of bioactivity, but this is generally beyond the capabilities of current docking programs.

The availability of many docking programs has led to comparative studies, in which two or more programs are applied to the same data and a comparison is made of the numbers of active molecules retrieved. In a very comprehensive example, Warren *et al*. (2006) evaluated the use of ten different docking programs and 37 different scoring functions with eight different protein structures. The evaluation considered all of the three tasks outlined above, and involved workers from five different GlaxoSmithKline sites in three different countries. Such studies clearly provide useful information about the operational characteristics of the various programs, but often involve proprietary datasets, which makes it difficult to generalise the results. Thus, just as with TREC and similar IR experiments, there is a need for standard datasets (Huang, Shoichet, & Irwin, 2006), evaluation measures (Cornell, 2006) and experimental protocols (Cole, Murray, Nissink, Taylor, & Taylor, 2005).

The reader is referred to a special issue of the *Journal of Medicinal Chemistry* (Leach, Shoichet, & Peishoff, 2006) and a recent book (Alvarez & Shoichet, 2005) for further details of docking-based approaches to virtual screening.

Drug-Likeness and ADMET Prediction

An important consideration in virtual screening is the need to ensure the *drug-likeness* or *drugability* of the molecules that are being considered for biological testing. Two approaches have proved to be of value.

The first approach seeks to determine whether a new molecule more closely resembles the molecules in a file of known drugs or the molecules in a file of (assumed) non-drugs. This is done using the methods of machine learning, with a predictive function being developed from the available training data that can subsequently be applied to the processing of previously untested molecules. Three near-contemporary papers reported on the use of decision trees (Ajay, Walters, & Murcko, 1998), neural networks (Sadowski & Kubinyi, 1998) and substructural analysis (Gillet, Willett, & Bradshaw, 1998) to discriminate between drugs and non-drugs, and there have been many subsequent such papers (e.g., (Byvatov, Fechner, Sadowski, & Schneider, 2003; Frimurer, Bywater, Naerum, Lauritsen, & Brunak, 2000; Wagener & van Geerestein, 2000; Zernov, Balakin, Ivaschenko, Savchuk, & Pletnev, 2003)). The second approach is based on analyses of files of known drugs. This work has shown that drugs have physical and physicochemical properties that are significantly different from those of non-drug molecules. This difference was first highlighted in a much-cited paper by Lipinski *et al.* (1997). Here, the authors analysed structures that had undergone preliminary clinical trials and noted that many that were orally active satisfied simple physicochemical constraints that could be summarised by what has come to be called the “Rule of Five”. For example, most of the molecules that were studied had molecular weights that were less than 500 and had no more than five hydrogen-bond donor or ten hydrogen-bond acceptor features. Molecules that violated several of these “rules” were hence less likely to be active when tested and could thus be removed from further consideration in a screening programme. This simple idea has spawned many enhancements in terms of the physicochemical properties that are considered, the precise numeric values for these properties, the differences in properties between marketed drugs and lead molecules, and the use of rule-of-five constraints in diversity-selection procedures (Brown, Hassan, & Waldman, 2000; Gillet, 2004).

There is now a very large literature on drug-likeness (e.g., (Egan, Walters, & Murcko, 2002; Hann, Leach, & Harper, 2001; Oprea, 2000; Oprea, Davis, Teague, & Leeson, 2001; Proudfoot, 2002; Teague, Davis, Leeson, & Oprea, 1999; Veber, Johnson, Cheng, Smith, Ward, & Kopple, 2002)), with new reports appearing frequently (e.g., (Monge, Arrault, Marot, & Morin-Allory, 2006; Sirois, Hatzakis, Wei, Du, & Chou, 2005; Verheij, 2006; Vieth & Sutherland, 2006)). Lipinski (2005) provides an overview of recent developments, with Leeson and Springthorpe (2007) discussing the use of drug-likeness in practical medicinal chemistry.

Drugability studies can be regarded as a stepping stone to a much more difficult problem: the prediction of ADMET properties, where ADMET denotes absorption, distribution, metabolism, excretion and toxicity. Research into QSAR has provided a body of tools that can be used to make reasonably accurate predictions, in many cases, of the biological activity of previously untested molecules. Now a molecule cannot possibly be a drug if it does not exhibit the desired activity, initially *in vitro* and then *in vivo*: however, many other criteria must be satisfied if a molecule is to become first a lead, then a candidate and finally a marketed drug. In particular, it must exhibit satisfactory pharmacokinteic properties and must not be toxic. Unsatisfactory ADMET characteristics are one of the principal causes of failure in drug-discovery programmes and hence drug-likeness studies provide a way of probing molecules' ADMET profiles indirectly. Methods for ADMET prediction tools try to model this behaviour explicitly using techniques such as multiple regression or support vector machines (*vide supra*). The four most important pharmacokinetic properties are absorption, distribution, metabolism and excretion. These correspond, respectively, to how rapidly the molecule gets into the bloodstream, is distributed around the body, is converted to other, possibly non-active substances, and is removed from the body. Toxicity relates to the ways in which a drug interacts not just with its chosen biological target but also with other substances in the body, such interactions being the cause of many of the side-effects that patients may experience when taking drugs.

The development of computer tools for the prediction of ADMET is one of the most important current areas in chemoinformatics, but will not be discussed further here since the methods are intimately bound up with details of human biology and thus rather far from the principal focus of the review. Instead, the reader is referred to some of the many excellent discussions that are available (Clark, 2005; Clark & Grootenhuis, 2002; Hyland, Jones, & van de Waterbeemd, 2006; van de Waterbeemd, 2005; van de Waterbeemd & Gifford, 2003; Yu & Adedoyin, 2003).

CONCLUSIONS

Chemoinformatics has evolved steadily since its beginnings in the late Sixties and early Seventies, and now plays a key role in the lead-discovery and lead-optimisation stages of pharmaceutical (and other types of specialty chemical) research programmes. In this review, we have focused on the use of similarity-based methods for virtual screening, a technique that is now a routine complement to experimental screening using HTS.

The Similar Property Principle, the chemical equivalent of the Cluster Hypothesis in IR, provides a firm basis for the use of similarity methods, and underlies not just similarity searching but also related applications such as the clustering of chemical databases and the analysis of molecular diversity. Thus far, the most successful similarity methods have been based on 2D representations of molecular structure, typically 2D fragment bit-strings, despite the much greater amounts of information that can be exploited if a 3D structural representation is available. One of the key tasks facing researchers in chemoinformatics over the next few years is how to exploit this additional source of information, the potential value of which has already been demonstrated by the importance of applications such as pharmacophore searching, flexible docking and 3D-QSAR. Other developments that we may expect over the next decade include: more effective ways of combining the various virtual-screening methods that are now available; chemoinformatics applications of new machine-learning algorithms as they are discovered by the data-mining community; more robust scoring functions for ligand-protein docking; and, the most challenging task, the development of methods for ADMET prediction that are comparable in performance to current QSAR methods for the prediction of bioactivity.

REFERENCES

A-Razzak, M., & Glen, R. C. (1992). Applications of rule-induction in the derivation of quantitative structure-activity relationships. *Journal of Computer-Aided Molecular Design, 6*, 349-383.

Adamson, G. W., & Bush, J. A. (1973). A method for the automatic classification of chemical structures. *Information Storage and Retrieval, 9*, 561-568.

Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., & Yapp, A. M. (1973). Strategic considerations in the design of screening systems for substructure searches of chemical structure files. *Journal of Chemical Documentation, 13*, 153-157.

Agrafiotis, D. K. (1997). Stochastic algorithms for maximising molecular diversity. *Journal of Chemical Information and Computer Sciences, 37*, 841-851.

Agrafiotis, D. K. (2002). Multiobjective optimization of combinatorial libraries. *Journal of Computer-Aided Molecular Design, 16*, 335-356.

Agrafiotis, D. K., Bandyopadhyay, D., Wegner, J. K., & van Vlijmen, H. (2007). Recent advances in chemoinformatics. *Journal of Chemical Information and Modeling, 47*, 1279-1293.

Agrafiotis, D. K., & Lobanov, V. S. (1999). An efficient implementation of distance-based diversity measures based on k-d trees. *Journal of Chemical Information and Computer Sciences, 39*, 51-58.

Ajay, Walters, W. P., & Murcko, M. A. (1998). Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *Journal of Medicinal Chemistry, 41*, 3314-3324.

Allen, F. H. (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallographica, B58*, 380-388.

Alvarez, J., & Shoichet, B. (Eds.). (2005). *Virtual screening in drug discovery*. Boca Raton: CRC Press.

Ash, J. E., & Hyde, E. (Eds.). (1975). *Chemical information systems*. Chichester: Ellis Horwood.

Ash, J. E., Warr, W. A., & Willett, P. (Eds.). (1991). *Chemical structure systems*. Chichester: Ellis Horwood.

Attias, R. (1983). DARC substructure search system: A new approach to chemical information. *Journal of Chemical Information and Computer Sciences, 23*, 102-108.

Baber, J. C., Shirley, W. A., Gao, Y., & Feher, M. (2006). The use of consensus scoring in ligand-based virtual screening. *Journal of Chemical Information and Modeling, 46*, 277-288.

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery, 1*, 882-894.

Ballester, P. J., & Richards, W. G. (2007a). Ultrafast shape recognition for similarity search in molecular databases *Proceedings of the Royal Society A, 463*, 1307-1321.

Ballester, P. J., & Richards, W. G. (2007b). Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry 28*, 1711-1723.

Barker, E. J., Cosgrove, D. A., Gardiner, E. J., Gillet, V. J., Kitts, P., & Willett, P. (2006). Scaffold-hopping using clique detection applied to reduced graphs. *Journal of Chemical Information and Modeling, 46*, 503-511.

Barker, E. J., Gardiner, E. J., Gillet, V. J., Kitts, P., & Morris, J. (2003). Further development of reduced graphs for identifying bioactive compounds. *Journal of Chemical Information and Computer Sciences, 43*, 346-356.

Barnard, J. M. (1990). Draft specification for revised version of the standard molecular data (SMD) format. *Journal of Chemical Information and Computer Sciences, 30*, 81-96.

Barnard, J. M. (1993). Substructure searching methods - old and new. *Journal of Chemical Information and Computer Sciences, 33*, 532-538.

Barnard, J. M. (Ed.). (1984). *Computer handling of generic chemical structures*. Aldershot: Gower.

Barton, I. J., Creasey, S. E., Lynch, M. F., & Snell, M. J. (1974). An information-theoretic approach to text-searching in direct-access systems *Communications of the Association for Computing Machinery, 17*, 345-350.

Bawden, D. (1993). Molecular dissimilarity in chemical information systems. In W. A. Warr (Ed.), *Chemical structures 2* (pp. 383-388). Heidelberg: Springer-Verlag.

Bayley, M. J., & Willett, P. (1999). Binning schemes for partition-based compound selection. *Journal of Molecular Graphics and Modelling, 17*, 10-18.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., Goharian, N., & Frieder, O. (2004). Fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology, 55*, 859-868.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. B. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management, 31*, 431-448.

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology, 6*, 281-297.

Bender, A., & Glen, R. C. (2004). Molecular similarity: a key technique in molecular informatics. *Organic and Biomolecular Chemistry, 2*, 3204-3218.

Bender, A., Mussa, H. Y., Glen, R. C., & Reiling, S. (2004). Molecular similarity searching using atom environments: information-based feature selection and a naive Bayesian classifier. *Journal of Chemical Information and Computer Sciences, 44*, 170-178.

Berks, A. H. (2001). Current state of the art of Markush topological search systems. *World Patent Information 23*, 5-13.

Berman, H. M., Battistuz, T., Bhat, T. N., Blum, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The Protein Data Bank. *Acta Crystallographica, D58*, 899-907.

Blair, J., Gasteiger, J., Gillespie, C., Gillespie, P. D., & Ugi, I. (1974). Representation of the constitutional and stereochemical features of chemical systems in the computer-assisted design of syntheses. *Tetrahedron, 30*, 1845-1859.

Boecker, A., Derksen, S., Schmidt, E., Teckentrup, A., & Schneider, G. (2005). A hierarchical clustering approach for large compound libraries. *Journal of Chemical Information and Modeling, 45*, 807-815.

Bohl, M., Dunbar, J., Gifford, E. M., Heritage, T., Wild, D. J., Willett, P., et al. (2002). Scaffold searching: automated identification of similar ring systems for the design of combinatorial libraries. *Quantitative Structure-Activity Relationships, 21*, 590-597.

Bohl, M., Wendt, B., Heritage, T. H., Richmond, N., & Willett, P. (2006). Unsupervised 3D ring template searching as an ideas generator for scaffold hopping by LAMDA, RigFit and Field-Based Similarity Search (FBSS) methods. *Journal of Chemical Information and Modeling, 46*, 1882-1890.

Böhm, H.-J. (1992). The computer-program LUDI - a new method for the *de novo* design of enzyme-inhibitors. *Journal of Computer-Aided Molecular Design, 6*, 61–78.

Böhm, H.-J., Flohr, A., & Stahl, M. (2004). Scaffold hopping. *Drug Discovery Today: Technologies, 1*, 217-224.

Böhm, H.-J., & Schneider, G. (Eds.). (2000). *Virtual screening for bioactive molecules.* Weinheim: Wiley-VCH.

Bostrom, J., Hogner, A., & Schmitt, S. (2006). Do structurally similar ligands bind in a similar fashion? *Journal of Medicinal Chemistry, 49*, 6716-6725.

Bottle, R. T., & Rowland, J. F. B. (Eds.). (1993). *Information sources in chemistry* (4th ed.). London: Bowker-Saur.

Boyd, D., & Marsh, M. M. (2006). History of computers in pharmaceutical research and development: a narrative. In S. Ekins (Ed.), *Computer applications in pharmaceutical research and development* (pp. 3-50). Hoboken NJ: John Wiley.

Bradshaw, J. (1997). Introduction to Tversky similarity measure. Retrieved 22nd July 2007, from http://www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html

Bradshaw, J., & Sayle, R. A. (1997). Some thoughts on significant similarity and sufficient diversity. Retrieved 22nd July 2007, from http://www.daylight.com/meetings/emug97/BradshawSignificant_Similarity.html).

Breiman, L. (2001). Random forests. *Machine Learning, 36*, 5-32.

Bremser, W., Klier, M., & Meyer, E. F. (1975). Mutual assignment of subspectra and substructures structure elucidation by 13C-NMR. *Organic Magnetic Resonance 7*, 97-105.

Briem, H., & Lessel, U. F. (2000). *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes. *Perspectives in Drug Discovery and Design, 20*, 231-244.

Brint, A. T., & Willett, P. (1987). Algorithms for the identification of three-dimensional maximal common substructures. *Journal of Chemical Information and Computer Sciences, 27*, 152-158.

Brown, F. K. (1998). Chemoinformatics: what is it and how does it impact drug discovery? *Annual Reports in Medicinal Chemistry, 33*, 375-384.

Brown, N., & Jacoby, E. (2006). On scaffolds and hopping in medicinal chemistry. *Mini-Reviews in Medicinal Chemistry 6*, 1217-1229.

Brown, N., Willett, P., Wilton, D. J., & Lewis, R. A. (2003). Generation and display of activity-weighted chemical hyperstructures. *Journal of Chemical Information and Computer Sciences, 43*, 288-297.

Brown, R. D., Hassan, M., & Waldman, M. (2000). Combinatorial library design for diversity, cost efficiency and druglike character. *Journal of Molecular Graphics and Modelling, 18*, 427-437.

Brown, R. D., & Martin, Y. C. (1996). Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences, 36*, 572-584.

Brown, R. D., & Martin, Y. C. (1997a). Designing combinatorial library mixtures using a genetic algorithm. *Journal of Medicinal Chemistry, 40*, 2304-2313.

Brown, R. D., & Martin, Y. C. (1997b). The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of Chemical Information and Computer Sciences, 37*, 1-9.

Brown, R. D., & Martin, Y. C. (1998). An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR and QSAR in Environmental Research, 8*, 23-39.

Bunin, B. A., Bajorath, J., Siesel, B., & Morales, G. (2007). *Chemoinformatics: Theory, practice, & products*. Dordrecht: Springer.

Burbridge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry, 26*, 5-14.

Butina, D. (1999). Unsupervised data base clustering based on Daylight's fingerprint, Tanimoto similarity: a fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences, 39*, 747-750.

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences, 43*, 1882-1889.

Cannon, E. O., Bender, A., Palmer, D. S., & Mitchell, J. B. O. (2006). Chemoinformatics-based classification of prohibited substances employed for doping in sport. *Journal of Chemical Information and Modeling, 46*, 2369-2380.

Capelli, A. M., Feriani, A., Tedesco, G., & Pozzan, A. (2006). Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands. *Journal of Chemical Information and Modeling, 46*, 659-664.

Carbó, R., Leyda, L., & Arnau, M. (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *International Journal of Quantum Chemistry, 17*, 1185-1189.

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular-features in structure activity studies - definition and applications. *Journal of Chemical Information and Computer Sciences, 25*, 64-73.

Charifsen, P. S., Corkery, J. J., Murcko, M. A., & Walters, W. P. (1999). Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry, 42*, 5100-5109.

Cheeseright, T., Mackey, M., Rose, S., & Vinter, A. (2006). Molecular field extrema as descriptors of biological activity: definition and validation. *Journal of Chemical Information and Modeling, 46*, 6650-6676.

Chen, B., Harrison, R. F., Pasupa, K., Wilton, D. J., Willett, P., Wood, D. J., et al. (2006). Virtual screening using binary kernel discrimination: effect of noisy training data and the optimisation of performance. *Journal of Chemical Information and Modeling, 46*, 478-486.

Chen, L., Nourse, J. G., Christie, B. D., Leland, B. A., & Grier, D. L. (2002). Over 20 years of reaction access systems from MDL: a novel reaction substructure search algorithm. *Journal of Chemical Information and Computer Sciences, 42*, 1296-1310.

Chen, W. L. (2006). Chemoinformatics: past, present and future. *Journal of Chemical Information and Modeling, 46*, 2230-2255.

Chen, X., & Brown, F. K. (2006). Asymmetry of chemical similarity. *ChemMedChem, 2,* 180-182.

Chen, X., & Reynolds, C. H. (2002). Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences, 42*, 1407-1414.

Chen, X., Rusinko, A., & Young, S. S. (1998). Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors. *Journal of Chemical Information and Computer Sciences, 38*, 1054-1062.

Cheng, C., Maggiora, G., Lajiness, M., & Johnson, M. A. (1996). Four association coefficients for relating molecular similarity measures. *Journal of Chemical Information and Computer Sciences, 36*, 909-915.

Christianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge Universtity Press.

Clark, D. E. (2005). Computational prediction of ADMET properties: recent developments and future changes. *Annual Reports in Computational Chemistry, 1*, 133-151.

Clark, D. E., & Grootenhuis, P. D. J. (2002). Progress in computational methods for the prediction of ADMET properties. *Current Opinion in Drug Discovery & Development, 5*, 382-390.

Clark, D. E., Willett, P., & Kenny, P. W. (1992). Pharmacophoric pattern matching in files of 3D chemical structures: use of smoothed bounded-distance matrices for the representation and searching of conformationally-flexible molecules. *Journal of Molecular Graphics, 10*, 194-204.

Clark, R. D. (1997). OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. *Journal of Chemical Information and Computer Sciences 37*, 1181-1188.

Clark, R. D., & Langton, W. J. (1998). Balancing representativeness against diversity using optimisable k-dissimilarity and hierarchical clustering. *Journal of Chemical Information and Computer Sciences, 38*, 1079-1086.

Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F., & Matthew, J. B. (2002). Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling, 20*, 281-295.

Clark, T. (2004). *A handbook of computational chemistry: A practical guide to chemical structure and energy calculations* (2nd edition). Chichester: John Wiley.

Cleves, A. E., & Jain, A. N. (2006). Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medicinal Chemistry, 49*, 2921-2938.

Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., & Taylor, R. (2005). Comparing protein-ligand docking programs is difficult. *Proteins, 60*, 325-332.

Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S., & Zha, Y. (2005). Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic and Biomolecular Chemistry, 3*, 1832-1834.

Congreve, M., Murray, C. W., & Blundell, T. L. (2005). Structural biology and drug discovery. *Drug Discovery Today, 10*, 895-907.

Corey, E. J., & Wipke, W. T. (1969). Computer-assisted design of complex organic syntheses. *Science, 166*, 178-193.

Corey, E. J., Wipke, W. T., Cramer, R. D., & Howe, W. J. (1972). Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *Journal of the American Chemical Society, 94*, 421-431.

Cornell, W. D. (2006). Recent evaluations of high-throughput docking methods for pharmaceutical lead finding - consensus and caveats. *Annual Reports in Computational Chemistry, 2*, 297-323.

Cosgrove, D. A., & Willett, P. (1998). SLASH: A program for analysing the functional groups in molecules. *Journal of Molecular Graphics and Modelling, 16*, 19-32.

Cramer, C. J. (2004). *Essentials of computational chemistry, theories and models* (2nd ed.). Hoboken NJ: Wiley.

Cramer, R. D., DePriest, S. A., Patterson, D. E., & Hecht, P. (1993). The developing practice of Comparative Molecular Field Analysis. In H. Kubinyi (Ed.), *3D QSAR in drug design. Theory, methods and applications* (pp. 443-485). Leiden: ESCOM.

Cramer, R. D., Jilek, R. J., Guessregen, S., Clark, S. J., Wendt, B., & Clark, R. D. (2004). "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *Journal of Medicinal Chemistry, 47*, 6777-6791.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative Molecular-Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society, 110*, 5959-5967.

Cramer, R. D., Redl, G., & Berkoff, C. E. (1974). Substructural analysis. A novel approach to the problem of drug design. *Journal of Medicinal Chemistry, 17*, 533-535.

Cramer, R. D., & Wendt, B. (2007). Pushing the boundaries of 3D-QSAR. *Journal of Computer-Aided Molecular Design, 21*, 23-32.

Crandell, C. W., & Smith, D. H. (1983). Computer-assisted examination of compounds for common three-dimensional substructures. *Journal of Chemical Information and Computer Sciences, 23*, 186-197.

Croft, W. B. (2000). Combining approaches to information retrieval. In W. B. Croft (Ed.), *Advances in information retrieval* (pp. 1-36). Dordrecht: Kluwer.

Cruciani, G., Pastor, M., & Mannhold, R. (2002). Suitability of molecular descriptors for database mining. A comparative analysis. *Journal of Medicinal Chemistry, 45*, 2685-2694.

Cuissart, B., Touffet, F., Crémilleux, B., Bureau, R., & Rault, S. (2002). The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *Journal of Chemical Information and Computer Sciences, 42*, 1043-1052.

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A., et al. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, 244-255.

Danziger, D. J., & Dean, P. M. (1989). Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proceedings of the Royal Society B, 236*, 101-113.

Davis, C. H. & McKim, G. W. (1999). Systematic weighting and ranking: Cutting the Gordian knot. *Journal of the American Society for Information Science, 50*, 626-628.

Dean, P. M. (Ed.). (1994). *Molecular similarity in drug design*. Glasgow: Chapman and Hall.

Dean, P. M., & Lewis, R. A. (Eds.). (1999). *Molecular diversity in drug design*. Amsterdam: Kluwer.

DesJarlais, R. L., Seibel, G. L., Kuntz, I. D., Furth, P. S., Alvarez, J. C., de Montellano, P. R. O., et al. (1990). Structure-based design of nonpeptidic inhibitors specific for the human immunodeficiency virus 1 protease. *Proceedings of the National Academy of Sciences (USA) 87*, 6644-6648

Diestel, R. (2000). *Graph theory*. New York: Springer-Verlag.

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics, 22*, 151-185.

Dittmar, P. G., Farmer, N. A., Fisanick, W., Haines, R. C., & Mockus, J. (1983). The CAS Online search system. I. General system design and selection, generation and use of search screens. *Journal of Chemical Information and Computer Sciences, 23*, 93-102.

Dixon, S. L., & Koehler, R. T. (1999). The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *Journal of Medicinal Chemistry, 42*, 2887-2900.

Dixon, S. L., & Merz, K. M. (2001). One-dimensional molecular representations and similarity calculations: methodology and validation. *Journal of Medicinal Chemistry, 44*, 3795-3809.

Dixon, S. L., Smondyrev, A. M., Knoll, E. H., Rao, S. N., Shaw, D.E., & Friesner, R. A. (2006). PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening. 1. Methodology and preliminary results. *Journal of Computer-Aided Molecular Design*, 20, 647-671.

Dixon, S. L., & Villar, H. O. (1998). Bioactive diversity and screening library selection via affinity fingerprinting. *Journal of Chemical Information and Computer Sciences, 38*, 1192-1203.

Doman, T. N., Cibulskis, J. M., Cibulskis, M. J., McCray, P. D., & Spangler, D. P. (1996). Algorithm5: a technique for fuzzy similarity clustering of chemical inventories. *Journal of Chemical Information and Computer Sciences, 36*, 1195-1204.

Doweyko, A. M. (2004). 3D-QSAR illusions *Journal of Computer-Aided Molecular Design, 18*, 587-596.

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Scence and Technology, 37*, 295-340.

Downs, G. M., & Barnard, J. M. (2002). Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry 18*, 1-40.

Downs, G. M., & Willett, P. (1994). Clustering of chemical-structure databases for compound selection. In H. van de Waterbeemd (Ed.), *Advanced computer-assisted techniques in drug discovery* (pp. 111-130). New York: VCH.

Downs, G. M., & Willett, P. (1995). Similarity searching in databases of chemical structures. *Reviews in Computational Chemistry, 7*, 1-66.

Eckert, H., & Bajorath, J. (2006). Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds *Journal of Medicinal Chemistry, 49*, 2284-2293.

Eckert, H., & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches. *Drug Discovery Today, 12*, 225-233.

Edgar, S. J., Holliday, J. D., & Willett, P. (2000). Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *Journal of Molecular Graphics and Modelling, 18*, 343-357.

Egan, W. J., Walters, W. P., & Murcko, M. A. (2002). Guiding molecules towards drug-likeness. *Current Opinion in Drug Discovery & Development, 5*, 540-549.

Ekins, S. (Ed.). (2006). *Computer applications in pharmaceutical research and development*. Hoboken NJ: Wiley-Interscience.

El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *Computer Journal, 32*, 220-227.

Engels, M. F. M., Gibbs, A. C., Jaeger, E. P., Verbinnen, D., Lobanov, V. S., & Agrafiotis, D. K. (2006). A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *Journal of Chemical Information and Modeling, 46*, 2651-2660.

Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., & Vieth, M. (2004). Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry, 47*, 45-55.

Estrada, E., & Uriarte, E. (2001). Recent advances on the use of topological indices in drug discovery research. *Current Medicinal Chemistry, 8*, 1573-1588.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th edition). London: Edward Arnold.

Ewing, T. J. A., Baber, J. C., & Feher, F. (2006). Novel 2D fingerprints for ligand-based virtual screening. *Journal of Chemical Information and Modeling, 46*, 2423-2431.

Ewing, T. J. A., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design, 15*, 411-428.

Feher, M. (2006). Consensus scoring for protein-ligand interactions. *Drug Discovery Today, 11*, 421-428.

Feldman, A., & Hodes, L. (1975). An efficient design for chemical structure searching. I. The screens. *Journal of Chemical Information and Computer Sciences, 15*, 147-152.

Fetchner, U., Paetz, J., & Schneider, G. (2005). Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR and Combinatorial Science, 24*, 961-967.

Fisanick, W. (1990). The Chemical Abstracts Service generic chemical (Markush) storage and retrieval capability, part 1. Basic concepts. *Journal of Chemical Information and Computer Sciences, 30*, 145-155.

Fligner, M. A., Verducci, J. S., & Blower, P. E. (2002). A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics, 44*, 110-119.

Flower, D. R. (1988). On the properties of bit string based measures of chemical similarity. *Journal of Chemical Information and Computer Sciences, 38*, 379-386.

Free, S. M., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry, 7*, 395-399.

Freeland, R., Funk, S., O'Korn, L., & Wilson, G. (1979). The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *Journal of Chemical Information and Computer Sciences, 19*, 94-98.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). GLIDE: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry, 47*, 1739-1749.

Frimurer, T. M., Bywater, R., Naerum, L., Lauritsen, L. N., & Brunak, S. (2000). Improving the odds in discriminating "drug-like" from non "drug-like" compounds. *Journal of Chemical Information and Computer Sciences, 40*, 1315-1324.

Gasteiger, J. (Ed.). (2003). *Handbook of chemoinformatics*. Weinheim: Wiley-VCH.

Gasteiger, J., & Engel, T. (Eds.). (2003). *Chemoinformatics: A textbook*. Weinheim: Wiley-VCH.

Gasteiger, J., & Jochum, C. (1978). EROS – a computer program for generating sequences of reactions. *Topics in Current Chemistry, 74*, 93-126.

Gedeck, P., Rhode, B., & Bartels, C. (2006). QSAR - how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling, 46*, 1924-1936.

Ghose, A. K., & Viswanadhan, V. N. (Eds.). (2001). *Combinatorial library design and evaluation: principles, software tools and applications in drug discovery*. New York: Marcel Dekker.

Gillet, V. J. (2004). Designing combinatorial libraries optimized on multiple objectives. *Methods in Molecular Biology, 275*, 335-354.

Gillet, V. J., & Johnson, A. P. (1998). Structure generation for de novo design. In Y. C. Martin & P. Willett (Eds.), *Designing bioactive molecules: Three-dimensional techniques and applications* (pp. 149-174). Washington DC: American Chemical Society.

Gillet, V. J., Johnson, A. P., Mata, P., & Sike, S. (1990). Automated structure design in 3D. *Tetrahedron Computer Methodology, 3*, 681-696.

Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., & Green, D. V. S. (2002). Combinatorial library design using a multiobjective genetic algorithm. *Journal of Chemical Information and Computer Sciences, 42*, 375-385.

Gillet, V. J., Willett, P., & Bradshaw, J. (1997). The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *Journal of Chemical Information and Computer Sciences, 37*, 731-740.

Gillet, V. J., Willett, P., & Bradshaw, J. (1998). Identification of biological activity profiles using substructural analysis and genetic algorithms. *Journal of Chemical Information and Computer Sciences, 38*, 165-179.

Gillet, V. J., Willett, P., & Bradshaw, J. (2003). Similarity searching using reduced graphs. *Journal of Chemical Information and Computer Sciences, 43*, 338-345.

Gillet, V. J., Willett, P., Bradshaw, J., & Green, D. V. S. (1999). Selecting combinatorial libraries to optimize diversity and physical properties. *Journal of Chemical Information and Computer Sciences, 39*, 169-177.

Ginn, C. M. R., Turner, D. B., Willett, P., Ferguson, A. M., & Heritage, T. W. (1997). Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *Journal of Chemical Information and Computer Sciences, 37*, 23-37.

Ginn, C. M. R., Willett, P., & Bradshaw, J. (2000). Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design, 20*, 1-16.

Glen, R. C., & Adams, S. E. (2006). Similarity metrics and descriptor spaces - which combinations to choose? *QSAR and Combinatorial Science, 25*, 1133-1142.

Godden, J. W., Furr, J. R., Xue, L., Stahura, F. L., & Bajorath, J. (2004). Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *Journal of Chemical Information and Computer Sciences, 44*, 21-29.

Godden, J. W., Stahura, F. L., & Bajorath, J. (2004). POT-DMC: a virtual screening method for the identification of potent hits. *Journal of Medicinal Chemistry, 47*, 5608-5611.

Godden, J. W., Xue, L., & Bajorath, J. (2000). Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences, 40*, 163-166.

Godden, J. W., Xue, L., Kitchen, D. B., Stahura, F. L., Schermerhorn, E. J., & Bajorath, J. (2002). Median partitioning: A novel method for the selection of representative subsets from large compound pools. *Journal of Chemical Information and Computer Sciences, 42*, 885-893.

Goldman, B. B., & Walters, W. P. (2006). Machine learning in computational chemistry. *Annual Reports in Computational Chemistry, 2*, 127-140.

Good, A. C., Cho, S. J., & Mason, J. S. (2004). Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. *Journal of Computer-Aided Molecular Design, 18*, 523-527.

Good, A. C., Hermsmeier, M. A., & Hindle, S. A. (2004). Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments *Journal of Computer-Aided Molecular Design, 18*, 529-536.

Good, A. C., Hodgkin, E. E., & Richards, W. G. (1992). Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *Journal of Chemical Information and Computer Sciences, 32*, 188-191.

Good, A. C., & Lewis, R. A. (1997). New methodology for profiling combinatorial libraries and screening sets: cleaning up the design with HARPick. *Journal of Medicinal Chemistry 40,* 3926-3936.

Good, A. C., & Richards, W. G. (1993). Rapid evaluation of shape similarity using Gaussian functions. *Journal of Chemical Information and Computer Sciences, 33,* 112-116.

Gorse, A.-D. (2006). Diversity in medicinal chemistry space. *Current Topics in Medicinal Chemistry 6,* 3-18.

Graf, W., Kaindl, H. K., Kniess, H., Schmidt, B., & Warszawski, R. (1979). Substructure retrieval by means of the BASIC Fragment Search Dictionary based on the Chemical Abstracts

Service Chemical Registry III System. *Journal of Chemical Information and Computer Sciences, 19*, 51-55.

Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry, 17*, 1653-1666.

Gray, N. A. B. (1986). *Computer assisted structure elucidation*. New York: John Wiley.

Greco, G., Novellino, E., & Martin, Y. C. (1998). 3D-QSAR methods. In Y. C. Martin & P. Willett (Eds.), *Designing bioactive molecules: three-dimensional techniques and applications*. Washington DC: American Chemical Society.

Green, D. V. S. (1998). Automated three-dimensional structure generation. In Y. C. Martin & P. Willett (Eds.), *Designing bioactive molecules: three-dimensional techniques and applications* (pp. 47-71). Washington DC: American Chemical Society.

Griffiths, A., Robinson, L. A., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation 40*, 175-205.

Gund, P. (1977). Three-dimensional pharmacophoric pattern searching. *Progress in Molecular and Subcellular Biology, 5*, 117-143.

Gund, P., Andose, J. D., Rhodes, J. B., & Smith, G. M. (1980). Three-dimensional molecular modelling and drug design. *Science, 208*, 1425-1431.

Gund, P., Wipke, W. T., & Langridge, R. (1974). *Computer searching of a molecular structure file for pharmacophoric patterns.* Paper presented at the International Conference on Computers in Chemical Research and Education Ljubljana, July 12-17 1973.

Güner, O. (Ed.). (2000). *Pharmacophore perception, development and use in drug design*. La Jolla CA: International University Line.

Hagadone, T. R. (1992). Molecular substructure similarity searching - efficient retrieval in two-dimensional structure databases. *Journal of Chemical Information and Computer Sciences, 32*, 515-521.

Haigh, J. A., Pickup, B. T., Grant, J. A., & Nicholls, A. (2005). Small molecule shape-fingerprints. *Journal of Chemical Information and Modeling, 45*, 673-684.

Hall, D. L. (1992). *Mathematical techniques in multisensor data fusion*. Northwood, MA: Artech House.

Hann, M., & Green, R. (1999). Chemoinformatics - a new name for an old problem? *Current Opinion in Chemical Biology, 3*, 379-383.

Hann, M., Leach, A. R., & Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of Chemical Information and Computer Sciences, 41*, 856-864.

Hansch, C., & Fujita, T. (1964). Rho sigma pi analysis: a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society, 86*, 1616-1626.

Hansch, C., Hoekman, D., Leo, A., Weininger, D., & Selassie, C. D. (2002). Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. Chemical Reviews, 102, 783 -812.

Hansch, C., & Leo, A. (1995). *Exploring QSAR. Fundamentals and applications in chemistry and biology*. Washington DC: American Chemical Society.

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature, 194*, 178-180.

Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. S., & Leach, A. R. (2001). Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences, 41*, 1295-1300.

Harper, G., Bravi, G. S., Pickett, S. D., Hussain, J., & Green, D. V. S. (2004). The reduced graph descriptor in virtual screening and data- driven clustering of high-throughput screening data. *Journal of Chemical Information and Computer Sciences, 44*, 2145-2156.

Hassan, M., Bielawski, J. P., Hempel, J. C., & Waldman, M. (1996). Optimization and visualization of molecular diversity of combinatorial libraries. *Molecular Diversity, 2*, 64-74.

Hassan, M., Brown, R. D., Varma-O'Brien, S., & Rogers, D. (2006). Cheminformatics analysis and learning in a data pipelining environment *Molecular Diversity, 10*, 283-299.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Compuer Sciences, 44*, 1-12.

Hawkins, D. M., Young, S. S., & Rusinko, A. (1997). Analysis of a large structure-activity data set using recursive partitioning. *Quantitative Structure-Activity Relationships, 16*, 296-302.

Hawkins, P. D. C., Skillman, A. G., & Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry, 50*, 74-82.

He, L., & Jurs, P. C. (2005). Assessing the reliability of a QSAR model's predictions. *Journal of Molecular Graphics and Modelling, 23*, 503-523.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004a). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences, 44*, 1177-1185.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004b). Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic and Biomolecular Chemistry, 2*, 3256-3266.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2005). Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. *Journal of Medicinal Chemistry, 48*, 7049-7054.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2006). New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching. *Journal of Chemical Information and Computer Sciences, 46*, 462-470.

Hertzberg, R. P., & Pope, A. J. (2000). High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology, 4*, 445-451.

Hessler, G., Zimmermann, M., Matter, H., Evers, A., Naumann, T., Lengauer, T., et al. (2005). Multiple-ligand-based virtual screening: methods and applications of the MTREE approach. *Journal of Medicinal Chemistry, 48*, 6575-6584.

Hiller, C., & Gasteiger, J. (1987). Ein automatisierter molekülbaukasten. In J. Gasteiger (Ed.), *Software-entwicklung in der Chemie 1* (pp. 53-66). Berlin: Springer Verlag.

Hinchcliffe, A. (2003). *Molecular modelling for beginners*. Chichester: Wiley.

Holliday, J. D., Hu, C.-Y., & Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High-Throughput Screening, 5*, 155-166.

Holliday, J. D., Rodgers, S. L., Willett, P., Chen, M.-Y., Mahfouf, M., Lawson, K., et al. (2004). Clustering files of chemical structures using the fuzzy k-means clustering method. *Journal of Chemical Information and Computer Sciences, 44*, 894-902.

Holliday, J. D., Salim, N., Whittle, M., & Willett, P. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences, 43*, 819-828.

Holliday, J. D., Salim, N., & Willett, P. (2005). On the magnitudes of coefficient values in the calculation of chemical similarity and dissimilarity. *American Chemical Society Symposium Series, 894*, 77-95.

Hsu, D. F., & Taksa, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval, 8*, 449-480.

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry, 49*, 6789-6801.

Hudson, B. D., Hyde, R. M., Rahr, E., & Wood, J. (1996). Parameter based methods for compound selection from chemical databases. *Quantitative Structure-Activity Relationships, 15*, 285-289.

Hurst, T. (1994). Flexible 3D searching - the directed tweak technique. *Journal of Chemical Information and Computer Sciences, 34*, 190-196.

Huser, J. (Ed.). (2006). *High-throughput screening in drug discovery*. Weinheim: Wiley-VCH.

Hyland, R., Jones, B., & van de Waterbeemd, H. (2006). Utility of human/human-derived reagents in drug discovery and development: an industrial perspective. *Environmental Toxicology and Pharmacology, 21*, 179-183.

InfoChem GmbH (2007). *CLASSIFY. The InfoChem Reaction Classification Program*. Retrieved 21st December 2007, from http://www.infochem.de/content/downloads/classify.pdf

International Union of Pure and Applied Chemistry (2007). *The IUPAC International Chemical Identifier (InChI^{TM}).* Retrieved 21st December 2007, from http://www.iupac.org/inchi/.

Jacoby, E. (Ed.). (2006). *Chemogenomics. Knowledge-based approaches to drug discovery*. London: Imperial College Press.

Jakes, S. E., Watts, N., Willett, P., Bawden, D., & Fisher, J. D. (1987). Pharmacophoric pattern-matching in files of 3D chemical structures - evaluation of search performance. *Journal of Molecular Graphics, 5*(1), 41-48.

Jakes, S. E., & Willett, P. (1986). Pharmacophoric pattern-matching in files of 3D chemical structures - selection of interatomic distance screens. *Journal of Molecular Graphics, 4*, 12-20.

Jarvis, R. A., & Patrick, E. A. (1973). Clustering using a similarity measure based on shared nearest neighbours. *IEEE Transactions on Computers, C-22*, 1025-1034.

Johnson, M. A., & Maggiora, G. M. (Eds.). (1990). *Concepts and applications of molecular similarity*. New York: John Wiley.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology, 267*, 727-748.

Jorissen, R. N., & Gilson, M. K. (2005). Virtual screening of molecular databases using a support vector machine. *Journal of Chemical Information and Computer Sciences, 45*, 549-561.

Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, A., Bukar, R., et al. (1995). Predicting ligand binding to proteins by affinity fingerprinting. *Chemistry & Biology, 2*, 107-118.

Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., & Sheridan, R. P. (1996). Chemical similarity using physicochemical property descriptors. *Journal of Chemical Information and Computer Sciences, 36*, 118-127.

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology, 25*, 197-206.

Kelley, L. A., Gardner, S. P., & Sutcliffe, M. J. (1996). An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. *Protein Engineering, 9*, 1063-1065.

Kier, L. B., & Hall, H. L. (1986). *Molecular connectivity in structure-activity analysis*. New York: Wiley.

Klebe, G. (Ed.). (2000). *Virtual screening: An alternative or complement to high throughput screening*. Dordrecht: Kluwer.

Klein, C. T., Kaiser, D., Kopp, S., Chiba, P., & Ecker, G. F. (2002). Similarity based SAR (SIBAR) as tool for early ADME profiling. *Journal of Computer-Aided Molecular Design, 16*, 785-793.

Klein, L. A. (1999). *Sensor and data fusion concepts and applications* (2nd ed.). Bellingham: SPIE Optical Engineering Press.

Kogej, T., Engkvist, O., Blomberg, N., & Muresan, S. (2006). Multifingerprint based similarity searches for targeted class compound selection. *Journal of Chemical Information and Modeling, 46*, 1201-1213.

Kubinyi, H. (1997a). QSAR and 3D QSAR in drug design. Part 1. *Drug Discovery Today, 2*, 457-467

Kubinyi, H. (1997b). QSAR and 3D QSAR in drug design. Part 2. *Drug Discovery Today, 2*, 538-546.

Kubinyi, H. (1998). Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design, 9-11*, 225-232

Kubinyi, H., Folkers, G., & Martin, Y. C. (Eds.). (1998). *3D QSAR in drug design*. Leiden: Kluwer/ESCOM.

Kubinyi, H., & Muller, G. (Eds.). (2004). *Chemogenomics in drug design*. Weinheim: Wiley VCH.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology, 161*, 269-288.

Kurogi, Y., & Guner, O. F. (2001). Pharmacophore modeling and three-dimensional database searching for drug design using CATALYST. *Current Medicinal Chemistry, 8,* 1035-1055.

Lajiness, M. S. (1990). Molecular similarity-based methods for selecting compounds for screening. In D. H. Rouvray (Ed.), *Computational chemical graph theory* (pp. 299-316). New York: Nova Science Publishers.

Langridge, R., Ferrin, T. E., Kuntz, I. D., & Connolly, M. L. (1981). Real-time color graphics in studies of molecular interactions. *Science, 211*, 661-666.

Leach, A. R. (2001). *Molecular modelling: Principles and applications* (2nd ed.). Harlow: Pearson Education.

Leach, A. R., & Gillet, V. J. (2003). *An introduction to chemoinformatics*. Dordrecht: Kluwer.

Leach, A. R., Shoichet, B. K., & Peishoff, C. E. (2006). Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *Journal of Medicinal Chemistry, 49*, 5851-5855.

Lee, J. H. (1997). Analyses of multiple evidence combination *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, 20*, 267-276.

Leeson, P. D. & Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, *6*, 881-890.

Leiter, D. P., Morgan, H. L., & Stobaugh, R. E. (1965). Installation and operation of a registry for chemical compounds. *Journal of Chemical Documentation, 5,* 238-242.

Lemmen, C., & Lengauer, T. (2000). Computational methods for the structural alignment of molecules. *Journal of Computer-Aided Molecular Design, 14*, 215-232.

Lengauer, T., Lemmen, C., Rarey, M., & Zimmermann, M. (2004). Novel technologies for virtual screening. *Drug Discovery Today, 9*, 27-34.

Lesk, A. M. (2005). *An introduction to bioinformatics* (2nd ed.). Oxford: Oxford University Press.

Lewis, R. A., Pickett, S. D., & Clark, D. E. (2000). Computer-aided molecular diversity analysis and combinatorial library design. *Reviews in Computational Chemistry 16*, 1-51.

Lind, P., & Maltseva, T. (2003). Support vector machines for the estimation of aqueous solubility. *Journal of Chemical Information and Computer Sciences, 43*, 1855-1859.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). *Applications of artificial intelligence for organic chemistry: The DENDRAL project*. New York: McGraw-Hill.

Lipinski, C. A. (2005). Filtering in drug discovery. *Annual Reports in Computational Chemistry, 1*, 155-168.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 23*, 3-25.

Lipscomb, K. J., Lynch, M. F., & Willett, P. (1989). Chemical structure processing. *Annual Review of Information Science and Technology, 24*, 189-238.

Loftus, F. (1991). Computer-aided synthesis design. In J. E. Ash, W. A. Warr & P. Willett (Eds.), *Chemical structure systems* (pp. 222-262). Chichester: Ellis Horwood.

Lombardino, J. G., & Lowe, J. A. (2004). The role of the medicinal chemist in drug discovery – then and now. *Nature Reviews Drug Discovery, 3*, 853-862.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science.* Chichester: Wiley.

Low, C. M. R., Buck, I. M., Cooke, T., Cushnir, J. R., Kalindjian, S. B., Kotecha, A., et al. (2005). Scaffold hopping with molecular field points: Identification of a cholecystokinin-2 (CCK2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based CCK2 antagonists. *Journal of Medicinal Chemistry, 48*, 6790-6802.

Lynch, M. F. (1977). Variety generation - a re-interpretation of Shannon's mathematical theory of communication and its implications for information science. *Journal of the American Society for Information Science, 28*, 19-25.

Lynch, M. F., & Holliday, J. D. (1996). The Sheffield generic structures project - a retrospective review. *Journal of Chemical Information and Computer Sciences, 36*, 930-936.

Lynch, M. F., & Willett, P. (1987). Information retrieval research in the Department of Information Studies, University of Sheffield: 1965-1985. *Journal of Information Science, 13*, 221-234.

Lyne, P. D. (2002). Structure-based virtual screening: an overview. *Drug Discovery Today, 7,* 1047-1055.

Maggiora, G. M. (2006). On outliers and activity cliffs - why QSAR often disappoints. *Journal of Chemical Information and Modeling, 46,* 1535.

Maggiora, G. M., Mestres, J., Hagadone, T. R., & Lajiness, M. S. (1997). Asymmetric similarity and molecular diversity, Paper presented at the *213th National Meeting of the American Chemical Society, April 13-17, 1997*. San Francisco, CA.

Mahe, P., Ueda, N., Akutsu, T., Perret, J. L., & Vert, J. P. (2005). Graph kernels for molecular structure-activity relationship analysis with support vector machines. *Journal of Chemical Information and Modeling, 45*, 939-951.

Maizel, R. E. (1998). *How to find chemical information* (3rd ed.). New York: Wiley.

Makara, G. M. (2001). Measuring molecular similarity and diversity: total pharmacophore diversity. *Journal of Medicinal Chemistry, 44,* 3563-3571.

Maldonado, A. G., Doucet, J. P., Petitjean, M., & Fan, B.-T. (2006). Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular Diversity, 10,* 39-79.

Manmatha, R., Rath, T., & F. Feng. (2001). Modelling score distributions for combining the outputs of search engines. *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval,* 267-275.

Marshall, R. R., Barry, C. D., Bosshard, H. E., Dammkoehler, R. A., & Dunn, D. A. (1979). The conformational parameter in drug design: The active analogue approach in computer-assisted drug design. In E. C. Olson & R. E. Christoffersen (Eds.), *Computer-assisted drug design* (pp. 205-226). Washington DC: American Chemical Society.

Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., & Moos, W. H. (1995). Measuring diversity - experimental-design of combinatorial libraries for drug discovery. *Journal of Medicinal Chemistry, 38,* 1431-1436.

Martin, Y. C. (1978). *Quantitative drug design. A critical introduction*. New York: Marcel Dekker.

Martin, Y. C., Bures, M. G., Danaher, E. A., Delazzer, J., Lico, I., & Pavlik, P. A. (1993). A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *Journal of Computer-Aided Molecular Design, 7,* 83-102.

Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activities? *Journal of Medicinal Chemistry, 45,* 4350-4358.

Martin, Y. C., Willett, P., Lajiness, M., Johnson, M., Maggiora, G. M., Martin, E., et al. (2001). Diverse viewpoints on computational aspects of molecular diversity. *Journal of Combinatorial Chemistry, 3,* 231-250.

Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., & Labaudiniere, R. F. (1999). New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry, 42*, 3251-3264.

Mason, J. S., & Pickett, S. D. (1997). Partition-based selection. *Perspectives in Drug Discovery and Design, 7-8*, 85-114.

Matter, H. (1997). Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry, 40,* 1219-1229.

Medina-Franco, J. L., Petit, J., & Maggiora, G. M. (2006). Hierarchical strategy for identifying active chemotype classes in compound databases. *Chemical Biology & Drug Design, 67*, 395-408.

Mestres, J., & Knegtel, R. M. A. (2000). Similarity versus docking in 3D virtual screening. *Perspectives in Drug Discovery and Design 20,* 191-207.

Mestres, J., Rohrer, D. C., & Maggiora, G. M. (1997). MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches. *Journal of Computational Chemistry, 18,* 934-954.

Mestres, J., Rohrer, D. C., & Maggiora, G. M. (2000). A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. 2. The relationship between alignment solutions obtained from conformationally rigid and flexible matching. *Journal of Computer-Aided Molecular Design, 14*, 39-51.

Monge, A., Arrault, A., Marot, C., & Morin-Allory, L. (2006). Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Molecular Diversity, 10*, 389-403.

Moock, T. E., Grier, D. L., Hounshell, W. D., Grethe, G., Cronin, K., Nourse, J. G., et al. (1988). Similarity searching in the organic reaction domain. *Tetrahedron Computer Methodology, 1*, 117-128.

Moock, T. E., Henry, D. R., Ozkaback, A. G., & Alamgir, M. (1994). Conformational searching in ISIS/3D databases. *Journal of Chemical Information and Computer Sciences, 34*, 184-189.

Moon, J. M., & Howe, W. J. (1991). Computer design of bioactive molecules: A method for receptor-based *de novo* ligand design. *Proteins, 11*, 314-328.

Morgan, H. (1965). The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. *Journal of Chemical Documentation, 5*, 107-113.

Muller, K. R., Ratsch, G., Sonnenburg, S., Mika, S., Grimm, M., & Heinrich, N. (2005). Classifiying 'drug-likeness' with kernel-based learning methods. *Journal of Chemical Information and Computer Sciences, 45*, 249-253.

Munk, M. E. (1998). Computer-based structure determination: then and now. *Journal of Chemical Information and Computer Sciences, 38*, 997-1009.

Murrall, N. W., & Davies, E. K. (1990). Conformational freedom in 3-D databases. 1. Techniques. *Journal of Chemical Information and Computer Sciences, 30*, 312-316.

Murtagh, F. (1985). *Multidimensional clustering algorithms.* Vienna: Physica Verlag.

Ng, K. B., & Kantor, P. B. (2000). Predicting the effectiveness of naïve data fusion on the basis of system characteristics. *Journal of the American Society for Information Science, 51*, 1177-1189.

Nikolova, N., & Jaworska, J. (2003). Approaches to measure chemical similarity - a review. *Quantitative Structure-Activity Relationships and Combinatorial Science, 22*, 1006-1026

Nishibata, Y., & Itai, A. (1991). Automatic creation of drug candidate structures based on receptor structure. *Tetrahedron, 47*, 8985-8990.

Nübling, W., & Steidle, W. (1970). Documentation Ring of the chemical/pharmaceutical industry - aims and methods. *Angewandte Chemie International Edition in English, 9*, 596-598.

Oda, A., Tsuchida, K., Takakura, T., Yamaotsu, N., & Hirono, S. (2006). Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *Journal of Chemical Information and Modeling, 46*, 380-391.

Onodera, N. (2001). A bibliometric study on chemical information and computer sciences focusing on literature of JCICS. *Journal of Chemical Information and Computer Sciences 41*, 878-888.

Oprea, T. I. (2000). Property distribution of drug-related chemical databases. *Journal of Computer-Aided Molecular Design, 14*, 251-264.

Oprea, T. I. (2002). Virtual screening in lead discovery: a viewpoint. *Molecules, 7*, 51-62.

Oprea, T. I., Davis, A. M., Teague, S. J., & Leeson, P. D. (2001). Is there a difference between leads and drugs? A historical perspective. *Journal of Chemical Information and Computer Sciences, 41*, 1308-1315.

Oprea, T. I., & Matter, H. (2004). Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology, 8*, 349-358.

Orengo, C. A., Thornton, J. M., & Jones, D. Y. (Eds.). (2002). *Bioinformatics*. Abingdon: Bios Scientific Publishers Ltd.

Ott, M. A. (2004). Cheminformatics and organic chemistry. Computer-assisted synthetic analysis. In J. H. Noordik (Ed.), *Cheminformatics developments: History, reviews and current research* (pp. 83-109). Amsterdam: IOS Press.

Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., & Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature Biotechnology, 24*, 805-815.

Paris, C. G. (1997). Chemical structure handling by computer. *Annual Review of Information Science and Technology, 32*, 271-337.

Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., & Weinberger, L. E. (1996). Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry, 39*, 3049-3059.

Pearlman, R. S. (1987). Rapid generation of high quality approximate 3D molecular structures. *Chemical Design Automation News, 2*, 1-7.

Pearlman, R. S., & Smith, K. M. (1998). Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design, 9-11*, 339-353.

Pearlman, R. S., & Smith, K. M. (1999). Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences, 39*, 28-35.

Perekhodtsev, G. D. (2007). Neighbourhood behavior: Validation of two-dimensional molecular similarity as a predictor of similar biological activities and docking scores. *QSAR and Combinatorial Science, 26*, 346-351.

Pickett, S. D., Mason, J. S., & McLay, I. M. (1996). Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *Journal of Chemical Information and Computer Sciences, 36*, 1214-1223.

Prathipati, P., Dixit, A., & Saxena, A. K. (2007). Computer-aided drug design: integration of structure-based and ligand-based approaches in drug design. *Current Computer-Aided Drug Design, 3*, 133-148.

Pretsch, E., Tóth, G., Munk, M. E., & Badertscher, M. (2002). *Computer-aided structure elucidation. Spectra interpretation and structure generation.* Weinheim: Wiley-VCH.

Proudfoot, J. R. (2002). Drugs, leads and drug-likeness: an analysis of some recently launched drugs. *Bioorganic Medicinal Chemistry Letters, 12*, 1647-1650.

Raha, K., & Merz, K. M. (2005). Calculating binding free energy in protein-ligand interactions. *Annual Reports in Computational Chemistry, 1*, 113-130.

Rang, H. P. (Ed.). (2006). *Drug discovery and development. Technology in transition*. Edinburgh: Churchill Livingstone.

Rarey, M., & Dixon, J. S. (1998). Feature trees: a new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design, 12*, 471-490.

Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology, 261*, 470-489.

Rarey, M., & Stahl, M. (2001). Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design, 15*, 497-520.

Rasmussen, E. M. (1997). Indexing images. *Annual Review of Information Science and Technology, 32*, 169-196.

Ray, L. C., & Kirsch, R. A. (1957). Finding chemical records by digital computers. *Science, 126,* 814-819.

Raymond, J. W., Blankley, C. J., & Willett, P. (2003). Comparison of chemical clustering methods using graph-based and fingerprint-based similarity measures. *Journal of Molecular Graphics and Modelling, 21*, 421-433.

Raymond, J. W., Gardiner, E. J., & Willett, P. (2002a). Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *Journal of Chemical Information and Computer Sciences, 42*, 305-316.

Raymond, J. W., Gardiner, E. J., & Willett, P. (2002b). RASCAL: calculation of graph similarity using maximum common edge subgraphs. *Computer Journal, 45*, 631-644.

Raymond, J. W., Jalaie, M., & Bradley, P. P. (2004). Conditional probability: a new fusion method for merging disparate virtual screening results. *Journal of Chemical Information and Computer Sciences, 44*, 601-609.

Raymond, J. W., & Willett, P. (2002). Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design, 16*, 59-71.

Raymond, J. W., & Willett, P. (2003). Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *Journal of Chemical Information and Computer Sciences, 43*, 908-916.

Robertson, S. E., & Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129-146.

Rogers, D., Brown, R. D., & Hahn, M. (2005). Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *Journal of Biomolecular Screening, 10*, 682-686.

Rössler, S., & Kolb, A. (1970). The GREMAS system, an integral part of the IDC system for chemical documentation. *Journal of Chemical Documentation, 10*, 128-134.

Rush, J. E. (1978). Handling chemical structure information. *Annual Review of Information Science and Technology, 13*, 209-262.

Rush, T. S., Grant, J. A., Mosyak, L., & Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry, 48*, 1489-1495.

Rusinko, A., Farmen, M. W., Lambert, C. G., Brown, P. L., & Young, S. S. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. *Journal of Chemical Information and Computer Sciences, 39*, 1017-1026.

Sadowski, J., & Kubinyi, H. (1998). A scoring scheme for discriminating between drugs and nondrugs. *Journal of Medicinal Chemistry, 41*, 3325-3329.

Saeh, J. C., Lyne, P. D., Takasaki, B. K., & Cosgrove, D. A. (2005). Lead hopping using SVM and 3D pharmacophore fingerprints. *Journal of Chemical Information and Modeling, 45*, 1122-1133.

Salim, N., Holliday, J. D., & Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *Journal of Chemical Information and Computer Sciences, 43*, 435-442.

Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley.

Sasaki, S. I., Abe, H., Ouki, T., Sakamoto, M., & Ochiai, S. (1968). Automated structure elucidation of several kinds of aliphatic and alicyclic compounds. *Analytical Chemistry, 40*, 2220-2223.

Schneider, G., & Fechner, U. (2005). Computer-based *de novo* design of drug-like molecules. *Nature Reviews Drug Discovery, 4*, 649-663.

Schneider, G., Neidhart, W., Giller, T., & Schmid, G. (1999). "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angewandte Chemie-International Edition, 38*, 2894-2896.

Schneider, G., Schneider, P., & Renner, S. (2006). Scaffold-hopping: how far can you jump? *QSAR and Combinatorial Science, 25*, 1162-1171.

Schofield, H., Wiggins, G., & Willett, P. (2001). Recent developments in chemoinformatics education. *Drug Discovery Today, 6*, 931-934.

Schreyer, S. K., Parker, C. N., & Maggiora, G. M. (2004). Data shaving: a focused screening approach. *Journal of Chemical Information and Computer Sciences, 44*, 470-479.

Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P., & Jacoby, E. (2006). Relationships between molecular complexity, biological activity, and structural diversity *Journal of Chemical Information and Modeling, 46*, 525-535.

Schuffenhauer, A., Floersheim, P., Acklin, P., & Jacoby, E. (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences, 43*, 391-405.

Schuffenhauer, A., Gillet, V. J., & Willett, P. (2000). Similarity searching in files of three-dimensional chemical structures: Analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *Journal of Chemical Information and Computer Sciences, 40*, 295-307.

Schuffenhauer, A., Brown, N. A., Ertl, P., Jenkins, J. L., Selzer, P., & Hamon, J. (2007). Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *Journal of Chemical Information and Modeling, 47*, 325-336.

Shanmugasundaram, V., Maggiora, G. M., & Lajiness, M. S. (2005). Hit-directed nearest-neighbor searching. *Journal of Medicinal Chemistry, 48*, 240-248.

Shelley, C. A., Hays, T. R., Munk, M. E., & Roman, H. V. (1978). An approach to automated partial structure expansion. *Analytica Chimica Acta, 103*, 121-132.

Shemetulskis, N. E., Dunbar, J. B., Dunbar, B. W., Moreland, D. W., & Humblet, C. (1995). Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design, 9*, 407-416.

Shenton, K., Norton, P., & Fearns, E. A. (1988). Generic searching of patent information. In W. A. Warr (Ed.), *Chemical structures: the international language of chemistry* (pp. 169-178). Berlin: Springer Verlag.

Sheridan, R. P. (2000). The centroid approximation for mixtures: Calculating similarity and deriving structure-activity relationships. *Journal of Chemical Information and Computer Sciences, 40*, 1456-1469.

Sheridan, R. P. (2007). Chemical similarity searches: when is complexity justified? *Expert Opinion on Drug Discovery, 2*, 423-430.

Sheridan, R. P., Feuston, B. P., Maiorov, V. N., & Kearsley, S. K. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences, 44*, 1912-1928.

Sheridan, R. P., & Kearsley, S. K. (1995). Using a genetic algorithm to suggest combinatorial libraries. *Journal of Chemical Information and Computer Sciences, 35*, 310-320.

Sheridan, R. P., & Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today, 7*, 903-911.

Sheridan, R. P., & Miller, M. D. (1998). A method for visualizing recurrent topological substructures in sets of active molecules. *Journal of Chemical Information and Computer Sciences, 38*, 915-924.

Sheridan, R. P., Miller, M. D., Underwood, D. J., & Kearsley, S. K. (1996). Chemical similarity using geometric atom pair descriptors. *Journal of Chemical Information and Computer Sciences, 36*, 128-136.

Sheridan, R. P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K. S., & Venkataraghavan, R. (1989). 3Dsearch: a system for three-dimensional substructure searching. *Journal of Chemical Information and Computer Sciences, 29*, 255-260.

Shively, E. (2007). CAS surveys its first 100 years. *Chemical and Engineering News, 84*(24), 41-53.

Sirois, S., Hatzakis, G., Wei, D., Du, Q., & Chou, K.-C. (2005). Assessment of chemical libraries for their drugability. *Computational Biology and Chemistry, 29*, 55-67.

Smeaton, A. F. (2004). Indexing, browsing, and searching of digital video. *Annual Review of Information Science and Technology, 38*, 371-407.

Snarey, M., Terrett, N. K., Willett, P., & Wilton, D. J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling, 15*, 372-385.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco: W. H. Freeman.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*, 11-21.

Spärck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of retrieval: development and comparative experiments. *Information Processing and Management,, 36*, 779-840.

Spärck Jones, K., & Willett, P. (Eds.). (1997). *Readings in information retrieval*. San Francisco: CA: Morgan Kaufmann.

Stahl, M., & Rarey, M. (2001). Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry, 44*, 1035-1042.

Stahura, F. L., & Bajorath, J. (2002). Bio- and chemo-informatics beyond data management: crucial challenges and future opportunities. *Drug Discovery Today, 7*, S41-S47.

Steinbach, M., Karypis, G., & Kumar, V. A. (2000). *Comparison of document clustering techniques*. Technical Report 00-034: Department of Computer Science & Engineering, University of Minnesota.

Steindl, T. M., Schuster, D., Wolber, G., Laggner, C., & Langer, T. (2006). High-throughput structure-based pharmacophore modelling as a basis for successful parallel virtual screening. *Journal of Computer-Aided Molecular Design*, *20*, 703-715.

Stiefl, N., Watson, I. A., Baumann, K., & Zaliani, A. (2006). ERG: 2D pharmacophore descriptions for scaffold hopping. *Journal of Chemical Information and Modeling, 46*, 208-220.

Stiefl, N., & Zaliani, A. (2006). A knowledge-based weighting approach to ligand-based virtual screening. *Journal of Chemical Information and Modeling, 46*, 587-596.

Sultan, M., Wigle, D. A., Cumbaa, C. A., Maziarz, M., Glasgow, J., Tsao, M. S., et al. (2002). Binary tree-structured vector quantization approach to clustering and visualizing microarray data. *Bioinformatics, 18*, 111-119.

Sussenguth, E. H. (1965). A graph-theoretic algorithm for matching chemical structures. *Journal of Chemical Documentation, 5*, 36-43.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences, 43*, 1947-1958.

Takahashi, Y., Sukekawa, M., & Sasaki, S. (1992). Automatic identification of molecular similarity using reduced-graph representation of chemical-structure. *Journal of Chemical Information and Computer Sciences, 32*, 639-643.

Tate, F. A. (1967). Handling chemical compounds in information systems. *Annual Review of Information Science and Technology, 2*, 285-309.

Taylor, R. (1995). Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *Journal of Chemical Information and Computer Sciences, 35*, 59-67.

Teague, S. J., Davis, A. M., Leeson, P. D., & Oprea, T. (1999). The design of leadlike combinatorial libraries. *Angewandte Chemie International Edition in English, 38*, 3743-3748.

Terrett, N. K. (1998). *Combinatorial chemistry*. Oxford: Oxford University Press.

Thorner, D. A., Wild, D. J., Willett, P., & Wright, P. M. (1996). Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials. *Journal of Chemical Information and Computer Sciences, 36*, 900-908.

Thorner, D. A., Willett, P., Wright, P. M., & Taylor, R. (1997). Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs. *Journal of Computer-Aided Molecular Design, 11*, 163-174.

Todeschini, R., & Consonni, V. (2002). *Handbook of molecular descriptors*. Weinheim: Wiley-VCH.

Tong, W., Hong, H., Fang, H., Xie, Q., & Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences, 43*, 525-531.

Tong, W., Lowis, D. R., Perkins, R., Chen, Y., Welsh, W. J., Goddette, D. W., Heritage, T. W., & Sheehan, D. M. (1998). Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *Journal of Chemical Information and Computer Sciences*, *38*, 669-677.

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., & Bertrand, H.-O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor type 4. *Journal of Medicinal Chemistry, 48*, 2534-2547.

Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling, 47*, 488-508.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*, 327-352.

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *Journal of the ACM, 16*, 31-42.

van de Waterbeemd, H. (2005). From *in vivo* to *in vitro*/*in silico* ADME: progress and challenges. *Expert Opinion on Drug Metabolism & Toxicology, 1*, 1-4.

van de Waterbeemd, H., & Gifford, E. (2003). ADMET *in silico* modelling: towards prediction paradise? *Nature Reviews Drug Discovery, 2*(3), 192-204.

van Rijsbergen, C. J. (1979). *Information retrieval*. London Butterworth.

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry, 45*, 2615-2623.

Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T. M., Murray, C. W., Taylor, R. D., et al. (2004). Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences, 44*, 793-806.

Verheij, H. J. (2006). Leadlikeness and structural diversity of synthetic screening libraries. *Molecular Diversity, 10*, 377-388.

Vieth, M., & Sutherland, J. J. (2006). Dependence of molecular properties on proteomic family for marketed oral drugs. *Journal of Medicinal Chemistry, 49*, 3451-3453.

Vleduts, G. E. (1963). Concerning one system of classification and codification of organic reactions. *Information Storage and Retrieval, 1*, 117-146.

Wagener, M., & van Geerestein, V. J. (2000). Potential drugs and nondrugs: prediction and identification of important structural features. *Journal of Chemical Information and Computer Sciences, 40*, 280-292.

Waldman, M., Li, H., & Hassan, M. (2000). Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *Journal of Molecular Graphics and Modelling, 18*, 412-426.

Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening - an overview. *Drug Discovery Today, 3*, 160-178.

Wang, R., & Wang, S. (2001). How does consensus scoring work for virtual library screening? An idealized computer experiment. *Journal of Chemical Information and Computer Sciences, 41*, 1422-1426.

Wang, Y., Eckert, H., & Bajorath, J. (2007). Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem, 2*, 1037-1042.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236-244.

Warmuth, M. K., Liao, J., Ratsch, G., Mathieson, M., Putta, S., & Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences, 43*, 667-673.

Warr, W. A. (1999). Balancing the needs of the recruiters and the aims of the educators. Paper presented at the 218th American Chemical Society National Meeting, New Orleans, August 22-26, 1999. Retrieved 22nd July 2007, from http://www.warr.com/warrzone2000.html.

Warr, W. A. (2003). *IUPAC Project Meeting 12-14 November 2003: Extensible Markup Language (XML) Data Dictionaries and Chemical Identifier*. Retrieved 22nd July 2007, from http://www.warr.com/inchi.pdf

Warr, W. A., & Willett, P. (1997). The principles and practice of 3D database searching. In Y. C. Martin & P. Willett (Eds.), *Designing bioactive molecules: three-dimensional techniques and applications* (pp. 73-95). Washington: American Chemical Society.

Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry, 49*, 5912-5931.

Weininger, D. (1988). SMILES, a chemical language and information-system.1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*, 31-36.

Weisgerber, D. W. (1997). Chemical Abstracts Service Chemical Registry System: history, scope and impacts. *Journal of the American Society for Information Science, 48*, 349-360.

Whittle, M., Gillet, V. J., Willett, P., Alex, A., & Loesel, J. (2004). Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *Journal of Chemical Information and Computer Sciences, 44*, 1840-1848.

Whittle, M., Gillet, V. J., Willett, P., & Loesel, J. (2006a). Analysis of data fusion methods in virtual screening: theoretical model. *Journal of Chemical Information and Modeling, 46*, 2193-2205.

Whittle, M., Gillet, V. J., Willett, P., & Loesel, J. (2006b). Analysis of data fusion methods in virtual screening: similarity and group fusion. *Journal of Chemical Information and Modeling, 46*, 2206-2219.

Wild, D. J., & Blankley, C. J. (2000). Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *Journal of Chemical Information and Computer Sciences, 40*, 155-162.

Wild, D. J., & Wiggins, G. D. (2006). Challenges for chemoinformatics education in drug discovery. *Drug Discovery Today, 11*, 436-439.

Wild, D. J., & Willett, P. (1996). Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. *Journal of Chemical Information and Computer Sciences, 36*, 159-167.

Willett, P. (1980). The evaluation of an automatically indexed, machine-readable chemical reactions file. *Journal of Chemical Information and Computer Sciences, 20*, 93-96.

Willett, P. (1981). A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing and Management, 17*, 53-60.

Willett, P. (Ed.). (1986). *Modern approaches to chemical reaction searching*. Aldershot: Gower.

Willett, P. (1987). *Similarity and clustering in chemical information systems*. Letchworth: Research Studies Press.

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management, 24*, 577-597.

Willett, P. (2000). Textual and chemical information retrieval: Different applications but similar algorithms, *Information Research*, *5*. Retrieved 22nd July 2007, from http://InformationR.net/ir/5-2/infres52.html).

Willett, P. (2004). The evaluation of molecular similarity and molecular diversity methods using biological activity data. *Methods in Molecular Biology, 275*, 51-63.

Willett, P. (2006). Data fusion in ligand-based virtual screening. *QSAR and Combinatorial Science, 25*, 1143-1152.

Willett, P. (2007). A bibliometric analysis of chemoinformatics. *Aslib Proceedings, in the press*.

Willett, P. (2008). From chemical documentation to chemoinformatics: fifty years of chemical information science. *Journal of Information Science, in the press*.

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences, 38*, 983-996.

Willett, P., Wilton, D. J., Hartzoulakis, B., Tang, R., Ford, J., & Madge, D. (2007). Prediction of ion channel activity using binary kernel discrimination. *Journal of Chemical Information and Modeling, 47,* in the press.

Willett, P., & Winterman, V. (1986). A comparison of some measures of inter-molecular structural similarity. *Quantitative Structure-Activity Relationships, 5*, 18-25.

Willett, P., Winterman, V., & Bawden, D. (1986a). Implementation of nearest-neighbour searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences, 26*, 36-41.

Willett, P., Winterman, V., & Bawden, D. (1986b). Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *Journal of Chemical Information and Computer Sciences, 26*, 109-118.

Williams, C. (2006). Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Molecular Diversity, 10*, 311-332.

Wilson, R. (1996). *Introduction to graph theory* (4th edition). Harlow: Longman.

Wilton, D. J., Harrison, R. F., Willett, P., Delaney, J., Lawson, K., & Mullier, G. (2006). Virtual screening using binary kernel discrimination: analysis of pesticide data. *Journal of Chemical Information and Modeling, 46*, 471-477.

Wipke, W. T., & Dyott, T. M. (1974). Stereochemically unique naming algorithm. *Journal of the American Chemical Society, 96*, 4825-4834.

Wipke, W. T., & Howe, W. J. (Eds.). (1977). *Computer assisted organic synthesis*. Washington: American Chemical Society.

Wipke, W. T., Ouchi, G. I., & Krishnan, S. (1978). Simulation and evaluation of chemical synthesis - SECS: an application of artificial intelligence techniques. *Artificial Intelligence, 11*, 173-193.

Worboys, M. F. (1995). *GIS: a computer perspective*. London: Taylor and Francis.

Xia, X. Y., Maliski, E. G., Gallant, P., & Rogers, D. (2004). Classification of kinase inhibitors using a Bayesian model. *Journal of Medicinal Chemistry, 47*, 4463-4470.

Xue, L., Stahura, F. L., Godden, J. W., & Bajorath, J. (2001). Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *Journal of Chemical Information and Computer Sciences, 41*, 746-753.

Yang, J.-M., Chen, Y.-F., Shen, T.-W., Kristal, B. S., & Hsu, D. F. (2005). Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling, 45*, 1134-1146.

Yin, P. Y., & Chen, L. H. (1994). A new non-iterative approach for clustering. *Pattern Recognition Letters, 15*, 125-133.

Yu, H., & Adedoyin, A. (2003). ADME-Tox in drug discovery: integration of experimental and computational techniques. *Drug Discovery Today, 8*, 852-861.

Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., & Pletnev, I. V. (2003). Drug discovery using support vector machines. Case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of Chemical Information and Computer Sciences, 43*, 2048-2056.

Zhang, Q., & Muegge, I. (2006). Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring *Journal of Medicinal Chemistry, 49*, 1536-1548.

Zupan, J., & Gasteiger, J. (1999). *Neural networks in chemistry and drug design*. New York: Wiley-VCH.