Article:
Willett, P. (2008) From chemical documentation to chemoinformatics: fifty years of chemical information science. Journal of Information Science, 34 (4). pp. 477-499. ISSN: 1741-6485

doi: 10.1177/0165551507084631

JIS

From chemical documentation to chemoinformatics: fifty years of chemical information science

Peter Willett[1]

University of Sheffield

Abstract

This paper summarises the historical development of the discipline that is now called 'chemoinformatics'. It shows how this has evolved, principally as a result of technological developments in chemistry and biology during the past decade, from long-established techniques for the modelling and searching of chemical molecules. A total of 30 papers, the earliest dating back to 1957, are briefly summarised to highlight some of the key publications and to show the development of the discipline.

Keywords: Chemical documentation; Chemical structures; Chemoinformatics; Drug discovery; History; Informatics; Molecules; Pharmaceutical research

1. Introduction

Chemistry is, and has been for many years, one of the most information-rich academic disciplines. The very first journal devoted to chemistry was *Chemisches Journal*, which was published 1778-1784 and then, under the name of *Chemische Annalen*, till 1803 [1]. The growth in the chemical literature during the 19th century led to a recognition of the need for comprehensive abstracting and indexing services for the chemical sciences. The principal such service is Chemical Abstracts Service (CAS), which was established in 1907 and which acts as the central repository for the world's published chemical (and, increasingly, life-sciences) information. The size of this repository is impressive: at the end of its first year of operations, the CAS database contained ca. 12K abstracts; by the end of 2006, this had grown to ca. 25M abstracts with ca. 1M being added each year. Most chemical publications will refer to one or more chemical substances. The structures of these substances form a vitally important part of the chemical literature, and one that distinguishes chemistry from many other disciplines. The CAS Registry System was started in 1965 to provide access to substance information, initially registering just small organic and inorganic molecules but now also registering biological sequences [2]. At the end of 1965 there

[1] *Correspondence to: Prof. Peter Willett, Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK; p.willett@sheffield.ac.uk.*

were ca. 222K substances in the System; by the end of 2006 this had grown to ca. 89M substances, of which ca. one-third were small molecules and the remainder biological sequences, with ca. 1.5M being added each year. There are also many additional molecular structures in public databases such as the Beilstein Database [3], and corporate files, in particular those of the major pharmaceutical, agrochemical and biotechnology companies.

The presence of chemical structures requires very different computational techniques from those used for processing conventional textual information. These specialised techniques - now referred to by the name of *chemoinformatics* as discussed further below - have developed steadily over the fifty years that have passed since the founding of the Institute of Information Scientists in 1958. This paper provides an historical overview of the development of these techniques by highlighting some of the key papers that have been published over the years. The focus is on the representation and searching of small molecules; the reader is referred elsewhere for the processing of textual chemical information (e.g., [4-6]) and of biological sequence information (e.g., [7,8]). Readers interested in the history of chemoinformatics are referred to Williams and Bowden's *Chronology of Chemical Information Science* [1], Metanomski's history of the Division of Chemical Information (formerly the Chemical Literature Group and then the Division of Chemical Literature) of the American Chemical Society [9], and Chen's recent historical review [10].

2. Historical development of the field

The importance of chemical information was recognised in 1961 by the establishment of what has since become the core journal for the field, the *Journal of Chemical Documentation* (as it was then named) published by the American Chemical Society. The focus on the documentation of the literature is evidenced by the journal's title, and an inspection of early tables-of-content demonstrates the importance of the published literature and of manual, rather than computerised, information processing. Very soon, however, papers began to appear in the journal that focused on the computer handling of structural information so that, for example, the first issue of Volume 2 contained articles describing the use of fragmentation-code and linear-notation systems based on punched card systems.

The Sixties and Seventies were a time of intensive research, with techniques being introduced that are, with appropriate development, still playing an important role in present-day systems. Examples include: the introduction of efficient algorithms for (sub)structure searching of databases of chemical molecules and for the indexing of databases of chemical reactions; the application of expert-systems technology to chemical problems; and the use of statistical correlation methods for the prediction of molecular properties (called QSAR for quantitative structure-activity relationships). The early Seventies saw the publication of the first two books devoted to the computer handling of chemical structure information. The first of these was that by Lynch *et al*., providing not just a snapshot of the current state-of-the-art but also summarising much of his early research at the University of Sheffield into methods for searching molecules and reactions [11]. The second was the proceedings of a NATO Advanced Study Institute held in Noordwijkerhout in Holland and attended by all the key researchers of the time [12], with the published proceedings including contributions from what have proved to be the three longest-lived academic research groups in the field: the DARC group in Paris under Dubois; the group in Sheffield initially under Adamson and Lynch and more recently under Gillet and Willett; and the group under Gasteiger, initially in Munich and more recently at Erlangen-Nuremberg. Then, in 1975, there was published what became the standard text for the next decade, Ash and Hyde's *Chemical Information Systems* [13], with further books from the same lead-author and publisher appearing in 1985 [14] and then in 1991 [15]. 1975 also saw the first change of name of the core journal, with its new title - the *Journal of Chemical Information and Computer Sciences* - emphasising the centrality of computerised techniques in the handling of chemical information.

The Eighties and early Nineties were – in part at least - developmental in nature, with much of the work building on techniques that had been introduced in the two previous decades. For example, similarity and generic searching methods were developed to complement structure and substructure searching, and the much enhanced computer technology of the time enabled the widespread implementation of operational systems, both public and in-house, for searching chemical databases. Perhaps the major enhancement was the move from two-dimensional

(2D, i.e., the conventional chemical structure diagram) to three-dimensional (3D, i.e., full atomic coordinate information) representations of molecular structure. This move was spurred by the appearance of structure-generation programs that permitted the conversion of 2D structure databases to 3D form, the latter necessitating the extension of existing systems for 2D searching and structure-property correlation to encompass the increased dimensionality of the structure representation. Developments in this period are exemplified by the third issue of Volume 25 of the *Journal of Chemical Information and Computer Sciences*, which celebrated the silver anniversary of the founding of the journal, and the sixth issue of Volume 3 of *Tetrahedron Computer Methodology*, which contained the papers from the first major symposium on 3D structure handling to be held by the American Chemical Society. Noordwijkerhout has been mentioned already as the location of the 1973 NATO Advanced Study Institute: 1987 saw it being the venue for the first of what has proved to be a three-yearly International Conference on Chemical Structures [16]. This rapidly established itself as the principal conference in the field, with the next in the series to be held in June 2008. The other major international conference dedicated to chemoinformatics has been held in Sheffield, co-sponsored by the Chemical Structure Association Trust and the Molecular Graphics and Modelling Society, every three years since 1998, in the year preceding the Noordwijkerhout meeting.

The commercial importance of many chemicals has meant that industry – in particular the pharmaceutical industry - has for long played a vitally important role in the development of chemical structure handling. The pharmaceutical industry is based on the synthesis of novel molecules that exhibit useful biological activities. Both the synthesis of molecules and the subsequent testing of these molecules for bioactivity underwent dramatic changes in the Nineties: taken together, the developments in these two areas resulted in significant increases in the volumes of data associated with pharmaceutical research programmes. Specifically, combinatorial chemistry provided the ability to synthesise not just one, but hundreds or even thousands, of structurally related compounds at a time; and high-throughput screening provided the ability to test very small samples of these large numbers of molecules at a time. There was thus an explosion in the volumes of structural and biological data that needed to be assimilated and rationalised, and this resulted in the emergence of what has come to be called chemoinformatics to deal with these requirements [17]. To quote Brown (who first used the term):

> *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* [18].

This definition ties chemoinformatics very firmly to pharmaceutical research; whilst many of the techniques have their roots in that industry, they are of much broader applicability, as noted by Paris and quoted by Warr:

> *Chem(o)informatics is a generic term that encompasses the design, creation, organisation, storage, management, retrieval, analysis, dissemination, visualisation and use of chemical information [19].*

Note the use of "chem(o)informatics" above since there has been some discussion as to whether the term should be "cheminformatics" or "chemoinformatics" [17]: we shall use the latter form here. Finally, a particularly succinct definition is given by Gasteiger

> *Chemoinformatics is the application of informatics methods to solve chemical problems* [20].

Chemoinformatics is not really new: instead, it is the integration of two previously separate aspects of chemical structure-handling. The group of researchers considered thus far had developed techniques to store, search and process the molecules in databases of chemical structures, so as to identify useful (in some sense) sets of compounds; a second, almost totally distinct group of researchers had for many years developed techniques to model and to correlate the structures of those molecules with biological properties, so as to enable the prediction of bioactivity in previously untested molecules [21-23]. At the risk of simplification, the former techniques were designed to handle the large numbers of molecules (hundreds of thousands or even millions) that would exist in a database; the latter techniques were designed to handle the few tens (or the few hundreds at most) of molecules for which the appropriate biological training data was available. Chemoinformatics is thus, at heart, a very specialised type of data mining, involving the analysis of chemical and biological information to support the discovery of new bioactive molecules [18,20].

The recent emergence of chemoinformatics has been marked by the publication of several new books, and by a further change in the title of the core journal, which has been called the *Journal of Chemical Information and Modeling* since 2005. The new title makes clear the linkage between chemical information (the archival and repository functions of chemical information systems) and the modelling and prediction of biological activity (the QSAR functions of molecular modelling systems) noted above. The emergence is also evidenced by the appearance of specialist educational programmes in chemoinformatics. Skolnik suggested that four elements are needed to characterise a discipline: a body of active researchers; a forum for the interaction of these researchers; a journal for the presentation of leading-edge papers in the discipline; and roots in the educational structure [24]. The first three of these had been present for many years, but it was not till the start of this century that the fourth was achieved with the appearance of several masters-level programmes in chemoinformatics in higher-education institutions in Europe and in the USA: the need for these is discussed by Schofield *et al*. [25] while Wild and Wiggins provide a detailed review of current provision [26].

3. Selection of key papers

This section reviews some of the most important papers in the development of chemoinformatic techniques. In each case, we briefly summarise the key paper, mention a few significant subsequent publications and give a recent review article (if available) to summarise the current status. More detailed accounts of all the topics considered here are provided in the textbooks by Leach and Gillet [27] and by Gasteiger and Engels [28], and in the extremely comprehensive *Handbook of Chemoinformatics* edited by Gasteiger [29]. As with any selection of key papers, the choice is inevitably biased by my own research interests and those of my colleagues, especially as I have spent my working life in an institution that has hosted for some four decades one of the most active research groups in the field [30,31]: as Alexander Pope noted

To observations which ourselves we make, we grow more partial for th'observer's sake.

Table 1 lists the papers that have been chosen for discussion, many of which will be very familiar to workers in the field. Others are less well known. In some cases, this is because they are by now very old, such as the seminal – and I use the word advisedly – contributions by Ray and Kirsch [32] and by Vleduts [33] to the searching of molecules and of reactions, respectively. In other cases, the original work was over-shadowed by subsequent publications: for example, the 1981 paper by Lynch *et al*. [34] was merely the first in a sequence of over twenty publications on the representation and searching of generic structures; and the 1983 graph-matching paper by Crandell and Smith [35] resulted a decade later in the first successful commercial system for pharmacophore mapping [36] (*vide infra*). In still other cases, the importance of the work was simply not fully recognised at the time, e.g., the paper by Adamson and Bush [37] on comparing fragment bit-strings to compute molecular similarity preceded the first descriptions of similarity searching systems by over a decade [38,39]; indeed, it was this 1973 paper that was one of the principal drivers for work in Sheffield on fingerprint-based searching and clustering that commenced in the Eighties [40] and that continues to the present day [41]. Even with such less well-recognised papers included, the thirty selected papers in Table 1 had attracted a total of 7363 citations in the *Web of Science* by May 2007, with four of those in the drug discovery area [42-45] each receiving over 500 citations in the literature.

The focus here is on the computational techniques underlying operational systems, but there are at least two further ways in which we could chart the development of chemoinformatics. The first approach would be by reference to the major operational systems (which are, of course, based on the techniques considered here). Examples of such milestones include the appearance of: the batch [67] and then online [68,69] implementations of the CAS Registry System; the NIH-EPA Structure and Nomenclature Search System (the first fully interactive structure and data retrieval system) [70]; ORAC [71] and REACCS [72] (the first in-house systems for storing and searching chemical reactions); and the Cambridge Structural Database [73] and the Protein Data Bank [74] (the two principal sources of experimental 3D coordinate data for organic molecules). The second approach would be by reference to the activities of those individuals responsible for the most important research findings. There is, however, a ready source of such information, this being the winners of the Herman Skolnik Award of the

1. Ray and Kirsch (1957) Substructure searching of connection tables [32]
2. Hansch *et al*. (1962) Correlation of bioactivity with physicochemical properties [46]
3. Vleduts (1963) Indexing of reactions and generation of synthetic pathways [33]
4. Free and Wilson (1964) Correlation of bioactivity with substituent contributions [43]
5. Morgan (1965) Canonicalisation of connection tables [47]
6. Sussenguth (1965) Set reduction technique for substructure searching [48]
7. Hyde *et al*. (1967) Conversion of WLNs to connection tables [49]
8. Corey and Wipke (1969) Interactive computer-aided synthesis design [50]
9. Crowe *et al*. (1970) Selection of dictionary-based screens [51]
10. Topliss and Costello (1972) Sample-feature ratios in QSAR [52]
11. Adamson and Bush (1973) Calculation of inter-molecular structural similarity [37]
12. Cramer *et al*. (1974) Correlation of bioactivity with substructural fragments [53]
13. Blair *et al*. (1974) Non-interactive computer-aided synthesis design [54]
14. Feldman and Hodes (1975) Selection of superimposed-coding screens [55]
15. Ullmann (1976) Efficient algorithm for substructure searching [56]
16. Gund (1977) Possibility of 3D substructure searching [57]
17. Lynch and Willett (1978) Indexing of chemical reactions [58]
18. Marshall *et al*. (1979) Active analogue approach for pharmacophore mapping [59]
19. Lynch *et al*. (1981) Representation and searching of Markush structures [34]
20. Kuntz *et al*. (1982) Protein-ligand docking [44]
21. Crandell and Smith (1983) Graph matching approach for pharmacophore mapping [35]
22. Jakes and Willett (1986) Selection of distance screens for 3D substructure searching [60]
23. Cramer *et al*. (1988) CoMFA method for 3D QSAR [42]
24. Danziger and Dean (1989) *De novo* molecular design [61]
25. Gasteiger *et al*. (1990) Generation of 3D atomic coordinates [62]
26. Johnson and Maggiora (1990) Similar property principle [63]
27. Martin *et al*. (1995) Computer selection of diverse molecules [64]
28. Brown and Martin (1996) Comparison of methods for compound selection [65]
29. Patterson *et al*. (1996) Neighbourhood behaviour [66]
30. Lipinski *et al*. (1997) Physicochemical properties of drug-like molecules [45]

Table 1*. Key papers in the development of chemoinformatics*

Division of Chemical Information of the American Chemical Society. The Division established this Award in 1976 to recognize outstanding contributions to, and achievements in the theory and practice of, chemical information science. It is named in honour of the first recipient, Herman Skolnik (the editor of the *Journal of Chemical Documentation* and then the *Journal of Chemical Information and Computer Sciences* from 1961 to 1982), and the awardees (listed at http://acscinf.org/dbx/awards/skolnik.asp, together with links to supporting information) comprise many of the pioneers in the field.

We have identified five broad categories of technique to structure the discussion below, these being techniques for: searching databases of 2D molecules; searching databases of patents, reactions and 3D molecules; quantitative structure-activity relationships and molecular modelling; knowledge-based systems; and diversity analysis and drug-likeness. It must be emphasised that these categories are rather arbitrary; for example, one could have separate sections for processing information about molecules and reactions, with the latter encompassing material about reaction databases and synthesis planning that are currently discussed in two separate sections (Sections 5 and 7, respectively). The reader should also note that the categories chosen here overlap to some considerable extent, e.g., pharmacophore mapping is discussed in Section 6, despite one of its main applications being for 3D database searching as discussed in Section 5.

4. Searching databases of 2D molecules

Database search is one of the principal facilities in any information system, and much of the early research in chemoinformatics targeted the development of efficient access mechanisms for databases of 2D structures. The key paper here, and the earliest to be discussed in this review, is that by Ray and Kirsch [32]. These authors were the first to describe an automated procedure for *substructure searching*, i.e., the identification of those molecules in a file that contain a user-defined query substructure. The paper described the use of a *connection table*, a labelled graph representation in which the nodes and edges of a graph encode the atoms and bonds of a 2D chemical structure diagram, to describe each of a file of 200 steroid molecules, and then the use of a subgraph-isomorphism algorithm based on an exhaustive, depth-first tree-search to analyse each molecule's connection table for the presence of a query substructure.

The connection table format was chosen for the CAS Registry System, which started operations in 1965 following several years of intensive research and development [75]. An important criterion in the development of the System was the need to provide a unique machine-readable identifier for each distinct molecule. The creation of a unique, or *canonical*, graph requires that the nodes of the graph are numbered, and for a graph containing N nodes there are up to $N!$ possible sets of numberings. Drawing on ideas first presented by Gluck [76], Morgan [47] described an algorithm to impose a unique ordering on the nodes in a graph, and hence to generate a canonical connection table that could provide a unique molecular identifier for computer processing (in much the same way as systematic nomenclature uniquely describes a molecule in a printed subject index). With subsequent enhancements [77,78], the Morgan algorithm continues to lie at the heart of the CAS Registry System (and also of many other chemoinformatics systems) right up to the present day.

The issue of *Journal of Chemical Documentation* that contained Morgan's paper on graph canonicalisation also contained the paper by Sussenguth describing his *set reduction* algorithm. Ray and Kirsch's substructure search algorithm was certainly effective but was also extremely inefficient, requiring a huge amount of backtracking. Sussenguth [48] realised that much of the backtracking could be eliminated, and hence the number of atom-to-atom comparisons minimised, by partitioning a graph into sets of nodes that possessed common characteristics, e.g., nodes of type nitrogen linked to not more than two other nodes. Nodes from the graph describing the query substructure then need to be considered for matching only against those nodes from a database structure that possess the same characteristics. It is interesting to note that this work was carried out in collaboration with Gerard Salton, as part of his pioneering work on statistical methods for retrieval that laid the foundation for present-day information retrieval systems [79]. The idea of linking query nodes to database nodes is fundamental to all substructure searching algorithms, including the refinement procedure that lies at the heart of the subgraph isomorphism algorithm due to Ullmann [56]. Although not designed specifically for the processing of chemical graphs, subsequent studies showed that it was particularly well suited to these sorts of graphs [80,81], and the Ullmann algorithm now forms the basis for most current substructure searching systems, both 2D and 3D.

However, even the use of set reduction was not sufficient to enable subgraph searching of chemical databases with acceptable response times, and it was only with the introduction of fragment-based screening methods that substructure searching became feasible on a large scale. The idea of screening is a simple one: to filter out that great fraction of the molecules in the search file that do not contain all of the substructural fragments that are

contained in the query substructure (in much the same way as keywords are used to filter searches of text databases). This idea was first suggested by Ray and Kirsch, who experimented with a simple molecular-formula screen but who realised that more sophisticated approaches might be required for large-scale operations [32]. This is now normally done using a fragment *bit-string*, in which the presence or absence of small substructural fragments in a molecule is encoded in a binary vector. Two main approaches have been developed for selecting the fragments that are used for screening: *dictionary-based* approaches in which there is a pre-defined list of fragments with normally one fragment allocated to each position in the bit-string; and *fingerprint-based* approaches, where hashing algorithms are used to allocate multiple fragments to each bit-position.

Effective dictionary-based screening requires that the fragments encoded in the bit-string have been selected so as to maximise the degree of filtering. In particular, the best use will be made of the available bits if the selected fragments are statistically independent of each other and if they occur with intermediate frequencies of occurrence [82]: if the fragments occur very frequently (in the database that is to be searched) then their presence in a query will eliminate only a small fraction of the database; if the fragments occur very infrequently in the database that is to be searched then they are most unlikely to be specified in a query. The use of frequency criteria was first studied by Lynch *et al*. in a series of papers, commencing with a study by Crowe *et al*. [51] of the frequencies of occurrence of bond-centred fragments. Subsequent papers in this series considered atom- and ring-centred fragments, culminating in a prototype system [83] that strongly influenced the subsequent design of dictionary-based screening systems such as that used for the CAS Registry System [67,68]. The screening methods developed by Feldman and Hodes at the National Institutes of Health do not make use of a dictionary of carefully selected fragments; instead, fragments are grown in an algorithmic fashion, one atom at a time, until they meet a frequency criterion, with a superimposed coding procedure being used to allocate multiple bits to each fragment and multiple fragments to each bit position [55]. The fragments here are hence closely tuned to the specific database that is to be screened, and the methods have strongly influenced the design of subsequent fingerprint-based screening systems.

Substructure searching requires the availability of a substructural query, this in turn requiring detailed insights into the structural requirements for biological activity. A common alternative situation is when the only information available is the existence of a known active molecule, such as a literature compound or a competitor's product. Use can then be made of the *Similar Property Principle*, which states that molecules that have similar structures will have similar properties. It is not clear where this was first stated explicitly; many people (including the present author) have cited the 1990 book by Johnson and Maggiora [63] as the source but the Principle had certainly been articulated prior to the book's publication in 1990. For example, writing in 1980, Wilkins and Randic [84] noted that it is

> *generally accepted that molecules of similar structural form may be expected to show similar biological or pharmacophoric patterns,*

and such considerations clearly underlie the 1973 paper by Adamson and Bush that is discussed further below. It is, however, appropriate to include the Johnson and Maggiora book [63] as one of the key contributions since it was the first publication to highlight the role played by similarity in a whole range of chemoinformatics applications, including database searching, property prediction, and computer-aided synthesis design *inter alia*. While there are many minor exceptions to the Principle, there is now a considerable body of experimental evidence for its general correctness [40,85-87]: the Principle provides a rational basis for a wide range of applications in chemoinformatics, including not just similarity searching and molecular diversity analysis (both of which are discussed later in this paper) but also database clustering and property prediction *inter alia*.

In a similarity search, a known active molecule, often referred to as a *reference* or *target* structure, is compared with all of the molecules in a database; molecules that are structurally similar to the reference structure are more likely to be active than are molecules that have been selected at random from a database. A similarity search hence provides a simple way of identifying further compounds for testing, and is thus one example of the more general concept of *virtual screening*, i.e., the use of computational, rather than biological, methods to identify bioactive molecules [88-90]. At the heart of any similarity searching system is the measure that is used to quantify the degree of structural resemblance between the target structure and each of the structures in the database that is to be searched. There are many such measures but by far the most common are those obtained by comparing the fragment bit-strings that are used for 2D substructure searching, so that two molecules are judged

as being similar if they have a large number of bits, and hence substructural fragments, in common. This approach was first described by Adamson and Bush for property prediction purposes [37]; over a decade was to pass, however, before bit-string similarities started to be used on a large scale for database searching, as exemplified by the first operational similarity searching systems at Lederle Laboratories [38] and Pfizer UK [39]. Similarity searching has now established itself as one of the most important tools for virtual screening, with widespread and continuing interest in the development of new similarity measures and search algorithms [91-93].

Thus far, we have considered only connection table records of 2D molecules. There is, however, an alternative type of representation that was very popular in the early days, at least in part since it had significantly lower processing costs at a time when computers were orders-of-magnitude slower than is the case today. This representation was the *line notation*, which encodes the topology of a molecule in an implicit form in an alphanumeric string, rather than explicitly as in a connection table. The Wiswesser Line Notation (WLN) system enjoyed widespread use for both in-house and public chemical information systems during the Sixties and early Seventies [13]. In 1967, Hyde *et al*. showed that it was possible to convert between WLN and a connection table [49], thus opening the way to providing full substructure searching capabilities on WLN-based systems, something that had previously been difficult to achieve with a high degree of effectiveness [94]. Although we have chosen normally to ignore systems papers, the paper by Hyde *et al*. is included here not just because of the relationship between linear notations and connection tables, but also because it (and two subsequent papers by this group [95,96]) described CROSSBOW, probably the first fully integrated in-house chemoinformatics system that allowed for compound registration, substructure searching and, importantly, structure display. This work set a standard that has driven the development of in-house systems ever since. WLN is now of historical interest only, but two other line notations – the SMILES [97] and IUPAC International Chemical Identifier (InChI) (details available at http://www.iupac.org/inchi/) notations – are widely used today as convenient input and storage representations.

5. Searching databases of reactions, patents and 3D molecules

Chemistry is as much about reactions as it is about molecules, but the development of databases of chemical reactions lagged behind the development of databases of chemical molecules for many years. The principal problem was the need to characterise not just the sets of reactant and product molecules, but also the *reaction sites* (or *reaction centres*), i.e., those parts of the reacting molecules where the substructural transformation takes place and which are the focus of many reaction queries. The key role of the reaction site in developing reaction database systems was first highlighted by Vleduts in an important 1963 paper [33]. This paper made three contributions: it suggested that the sites could be detected by comparing the connection tables of the reactant and product molecules to identify the structural commonalities and differences; it described a simple classification scheme based on bonds broken or formed that could be used to organise and to search a database of chemical reactions; and it considered the use of computers to suggest synthetic pathways (as discussed further in Section 7 below). Starting in 1967 [98], Lynch and collaborators studied a range of comparison methods for the mapping of reactant and product atoms, but some ten years passed before an effective and efficient procedure for reaction-site detection was identified, based on a maximum common subgraph (MCS) isomorphism algorithm [58]. Lynch and Willett's original procedure used an approximate MCS algorithm that was based on an adaptation of the Morgan algorithm, but the operational systems that soon emerged (such as CASREACT [99] and REACCS [100]) used exact graph-matching procedures for reaction-site detection that, with appropriate development, are used to the present day [101].

Specialised techniques are also required for the handling of the structural information that occurs in chemical patents. In many cases, a patent will describe specific chemical molecules, but it may also describe *generic*, or *Markush*, structures, which encode many, or even an infinite number of, different specific molecules in a single representation. The individual specific molecules normally result from variations in the nature, position and frequency of substituents on a central ring system, or scaffold, with further complexities arising from, e.g., a variable substituent itself being capable of variation. These complexities drove a long-term research programme

by Lynch's group to develop the connection-table, screening and atom-by-atom components of conventional substructure searching systems so that they could be used to represent and to search Markush structures. The basic strategy is outlined by Lynch *et al.* [34], this being the first of over 20 papers that are summarised and reviewed by Lynch and Holliday [102]. These studies resulted in a body of algorithms and data structures that provided much of the theoretical and practical basis for the current sophisticated systems for structure-based access to generic chemical structures [103]. Many of the techniques that were developed on this project have found further application in the representation and searching of *combinatorial libraries*, large (and sometimes extremely large) sets of structurally related molecules that can be generated using the techniques of combinatorial synthesis [104] as discussed further in Section 8.

Thus far, we have considered only 2D molecular representations. However, the 3D structure of a molecule is of crucial importance in determining its properties, and it is thus hardly surprising that interest turned to the provision of facilities for 3D searching, often referred to as *pharmacophore* searching where a pharmacophore is the geometric pattern of features in a drug that interacts with a biological receptor [105]. Pharmacophore searching was first described by Gund. His 1977 paper (there is an earlier, difficult-to-obtain report in the proceedings of a 1973 conference [106]) showed that the graph matching techniques that were by then well established for 2D substructure searching could be applied to 3D substructure searching using graphs in which the edges denoted inter-atomic distances, rather than bonds as in a conventional connection table [57]. However, there was little interest in this remarkable achievement for over a decade, because two problems had to be addressed before searching systems could be developed.

The first problem was the lack of data. The principal source of experimental atomic coordinate data for small molecules is the Cambridge Structural Database (CSD) produced by the Cambridge Crystallographic Data Centre [73,107], which started in 1964 and which now contains the 3D structures for ca. 400K molecules. Although the database is a vital resource for drug research, it contains only a very small fraction of the molecules that might be of interest to a pharmaceutical company. Accurate 3D structures for many molecules can be obtained using computational techniques such as quantum mechanics, molecular dynamics and molecular modelling, but these are too time-consuming for large numbers of molecules. There was hence much interest in 1987 when two programs for *structure generation* were reported. These could rapidly convert a 2D structure into a reasonably accurate 3D structure, thus opening up the possibility of converting chemical databases to 3D form. These two programs were CONCORD, developed by Pearlman and co-workers [108], and CORINA, developed by Gasteiger and co-workers [109]: despite many subsequent programs for structure generation [110], these two remain the principal sources of computed 3D structures to the present-day. The two listed papers are both hard-to-get – one is in an informal newsletter and the other in a German conference proceedings – so the paper chosen for inclusion here is that by Gasteiger *et al*. in the 1990 issue of *Tetrahedron Computer Methodology* mentioned previously [62].

The second problem was that while Gund had demonstrated that 3D substructure searching was possible, the matching algorithm was far too slow for large-scale processing. We have noted that operational systems for 2D substructure searching only became feasible with algorithmic developments such as the use of set-reduction and of bit-string screening. In a series of papers, Willett and co-workers developed the basic algorithmic techniques necessary for efficient 3D substructure searching. In the first of these, they reported the development of a screening system in which the bit-strings encoded the distances between pairs of heavy atoms in a molecule as a distance-range, these ranges being chosen using frequency-based methods analogous to those developed previously for selecting screens for 2D substructure searching [60]. Subsequent papers compared subgraph isomorphism algorithms that could be used for 3D substructure searching, and reported the first operational 3D searching system developed in collaboration with Pfizer UK; they then extended their techniques to take account of the fact that most molecules are not rigid but can, instead, exhibit some degree of flexibility owing to the existence of one or more rotatable bonds in a molecule. The work of the Sheffield group is summarised by Willett [111] and 3D substructure searching, both rigid and flexible, is now a standard facility for in-house chemoinformatics systems [105,112].

There is a further type of 3D database search that is now one of the key components of systems for virtual screening: *protein-ligand docking* or, more simply, docking. Docking assumes that a 3D structure has been obtained, typically by X-ray crystallography, of the biological receptor, such as the active site of an enzyme, that

is involved in a biological pathway of interest. The "lock-and-key" theory of drug action assumes that a drug fits into a biological receptor in much the same way as a key fits a lock; thus, if the shape of the lock is known, one can identify potential drugs by scanning a 3D database to find those molecules that have shapes that are complementary to the shape of the receptor. The original description of docking, by Kuntz *et al*. [44], considered the fitting of just a single molecule into a protein active site, with the molecule and the binding site being described by sets of spheres that were checked for a steric match. However, it was soon realised that if this fitting operation was repeated for all of the molecules in a database then docking could provide a highly sophisticated approach to virtual screening, with a database being ranked in order of decreasing goodness of fit with the active site (and hence in decreasing likelihood of activity). Developments in the basic technique involved matching not just geometric but also chemical characteristics of a molecule and a protein; however, the recent surge in interest in docking has come about as the result of systems that take account of the inherent flexibility of many small molecules, with docking systems such as GOLD [113] and FlexX [114] being used on a very large scale in industrial lead-discovery programmes. Rester [115] and Leach *et al*. [116] summarise the current state-of-the-art, the latter in the preface to a special issue of *Journal of Medicinal Chemistry* that focuses on studies of protein-ligand docking.

6. Quantitative structure-activity relationships and molecular modelling

Chemoinformatics is principally concerned with the lead-discovery and lead-optimisation stages of drug discovery: finding one or more exemplars of a class of compounds that exhibits the bioactivity of interest; and then identifying those members of that class that possess the best combination of potency, synthetic feasibility, pharmacokinetic properties (e.g., solubility and metabolic stability) and minimal side-effects. Early studies of computational methods in drug discovery focussed on the second of these two stages, whilst modern work also contributes to the lead discovery stage, most obviously by means of virtual screening as described previously.

The classical approach to quantitative structure-activity relationships (QSAR) is *Hansch analysis*. In a series of papers in the early Sixties, Hansch and his co-workers showed that it was possible to use multiple linear regression (MLR) to derive statistically significant correlations between the biological activities of sets of structural analogues and experimental or computed physicochemical parameters that describe the molecules' steric, electrostatic and hydrophobic properties; once the correlation has been obtained, the resulting equation can be used to predict the bioactivity of previously untested molecules. The first such paper was published in 1962 [46], with this being followed by several others that, taken together, provided an approach to lead optimisation that has played a crucial role in the development of QSAR and that continues to be used to the present day [117,118]. Many different physicochemical parameters have been used in Hansch analysis, but by the far the most common is the octanol-water partition coefficient, which has spurred the development of many different programs for the calculation of this important descriptor [119]. Just two years after Hansch's seminal paper, Free and Wilson published a further technique for lead optimisation that was again based on MLR but that used structural, rather than physicochemical, variables in the analysis [43]. The basic idea is that the presence of a specific substituent at a specific position on a ring scaffold makes a constant and additive contribution to the overall activity of those molecules that contain it. These contributions are obtained using MLR, and then used to suggest new analogues for synthesis and testing.

The use of MLR to predict biological activities was rapidly adopted as a key tool for lead optimisation. There is, however, a problem – common to many statistical and machine-learning methods – that is related to the *sample-feature ratio*, i.e., the ratio of the number of variables to the number of observations. Specifically, it is possible to derive seemingly strong correlations even if no meaningful correlation exists in practice when the value of this ratio is less than some threshold value, typically 5-10 being quoted in the literature. The importance of these statistical considerations was first demonstrated by Topliss and Costello [52] (see also [120]), who showed that seemingly good QSAR correlations could be obtained using random variables if sufficient of them were included in the predictive equation; indeed the crucial factor is the number of variables considered for inclusion (a number that is often greater than the number included in the final equation) [121,122]. Such statistical

considerations need to be taken into account whenever a new predictive tool becomes available, as evidenced by studies of the applications of neural networks to the prediction of biological activity [123]. Sample-feature ratios continue to be of importance given the very large numbers of descriptors that can now be generated for a molecule [124,125], although techniques such as cross-validation and data scrambling can help to confirm the significance of potential structure-activity correlations [126,127]. Other problems associated with the use of MLR for QSAR are discussed by Wold and Dunn [128].

A limitation of Free-Wilson analysis is the large number of analogues that need to be synthesised and tested if multiple substituent positions are allowed on the central scaffold, and this has restricted the use of the approach as originally described. However, the applicability of the method is significantly enhanced if the location-specific criterion is relaxed, and the biological activity is expressed merely in terms of the presence of a substituent (or, more generally, substructural fragment) rather than its location. Thus Adamson and co-workers correlated several biological and physical properties with the occurrences of fragments generated from connection tables or WLN using MLR (see, e.g., [129,130]), an approach that foreshadows the commercial HQSAR package [131]. However, the most important development of Free-Wilson analysis is probably *substructural analysis*, as first described by Cramer *et al*. in 1974 [53]. This used qualitative, i.e., active/inactive, biological data, and also allowed the analysis of large, structurally diverse datasets, thus enabling the analysis of the screening data that forms one of the principal components of lead-discovery programmes. Substructural analysis involves calculating a weight for each fragment (often denoted by a particular bit-location in a fingerprint) that is used to characterise the molecules in the training data, this weight being a function of the numbers of active and inactive molecules that contain this fragment [132]. A score is then obtained for a molecule of unknown activity by summing the weights for its constituent fragments. The resulting score represents the molecule's probability of activity, and untested molecules can hence be prioritised for screening in order of decreasing probability of activity; the anti-cancer screening programme that was carried out during the Eighties by the National Cancer Institute [133] is an important example of such an approach. Substructural analysis is important not just in its own right but also as the first example of machine learning being used on a large scale in chemoinformatics since, although not realised at the time [134], substructural analysis is an example of a naive Bayesian classifier, a machine-learning technique that is now widely used for the analysis of biological screening data [135].

The QSAR methods discussed thus far take no explicit account of the 3D structures of molecules, despite the fact that molecular size and shape is a key factor in determining the interactions between a potential drug molecule and its biological receptor. The methods of computational chemistry – quantum mechanics, molecular dynamics and molecular modelling – provide effective tools for analysing the conformations that molecules can adopt in 3D space but, as noted previously, these can be very demanding of computational resources. Marshall *et al.* were the first to describe a conformational searching procedure that was sufficiently rapid in operation to investigate the shapes of sets of molecules such as those that might be encountered in a QSAR analysis [59]. With programs such as this, and then structure-generation procedures such as CONCORD and CORINA, it was not long before QSAR methods started to appear that sought to correlate bioactivity with the 3D structures of molecules. Two papers were published in 1988 describing the Hypothetical Active Site Lattice (HASL) approach of Doweyko [136] and the Comparative Molecular Field Analysis (CoMFA) approach of Cramer *et al*. [42]. The latter has been far more widely used, not least because it rapidly became available as a successful commercial product. A molecule in a CoMFA analysis is placed at the centre of a regular 3D grid, and the steric and electrostatic interaction energies between the molecule and a standard probe then computed at each point in the grid. The resulting sets of interaction energies for each molecule are then correlated with those molecules' bioactivities using not MLR but an alternative multivariate technique, Partial Least Squares (PLS). PLS describes the variations in the bioactivity by means of latent variables that are linear combinations of the original variables, i.e., the grid-point interaction energies. The use of latent variables, rather than the original variables, makes analysis of the resulting correlation equations rather more difficult than in the case of MLR, but this is compensated for by the fact that PLS can handle datasets with very large numbers of variables, i.e., with sample-feature ratios that would preclude the use of MLR. This it does by means of a multiple-sampling technique known as cross-validation that ensures the statistical significance of the resulting predictive equations. There are several factors that need to be taken into account when carrying out a CoMFA analysis [137] but these have not prevented it becoming the method of choice for present-day QSAR [138].

Many of the techniques that are used in molecular modelling are too time-consuming for use in chemoinformatics applications, although this is starting to change [139]. There is, however, one such technique that has proved of considerable value, and that is the application of methods for conformational searching, i.e., a detailed exploration of the conformations that a molecule can adopt in 3D space, to the identification of *pharmacophoric patterns*, where a pharmacophoric pattern, or pharmacophore, is the geometric arrangements of features that is responsible for some particular type of bioactivity. A pharmacophoric pattern can be used both to rationalise the activities of molecules and to act as the query for a 3D substructure search to find new molecules that contain the pattern and that are hence also possible actives. The active analogue approach of Marshall *et al.* [59] was the first automated procedure for pharmacophore mapping. Given a set of molecules with a common biological activity, the low-energy conformational space of each of the molecules is explored to find a conformation (or conformations) that allows the chosen features (typically hydrogen-bond donors or acceptors, or the centres of aromatic rings) to appear in the same geometric arrangement in all of the molecules. The effectiveness of the approach was demonstrated by the identification of the pharmacophore common to a set of diverse angiotensin-converting enzyme (ACE) inhibitors [140], and its efficiency was later increased by means of an improved conformational searching algorithm [141].

The active analogue approach is widely used but does require the specification of the matching features prior to the conformational search, implying some knowledge of the protein-ligand interactions that are involved in the observed bioactivity. This limitation was first overcome in a study by Crandell and Smith [35], who described the use of an MCS algorithm to find 3D patterns common to sets of molecules. The work was carried out as an aid to structure elucidation (see below) but is also clearly applicable to the problem of pharmacophore mapping. The Crandell-Smith algorithm involves a breadth-first search and becomes very slow if multiple molecules are required. However, a detailed study of a range of algorithms for 3D MCS detection [142] demonstrated the general efficiency of the clique-detection algorithm of Bron and Kerbosch [143] for this application. An operational implementation of the Bron-Kerbosch algorithm by Martin *et al.* [36] resulted in DISCO, the first widely used program for pharmacophore mapping and a direct influence on the many such programs that are now available [105,144].

7. Knowledge-based systems

It may be argued that the term "knowledge-based" is rather non-specific since any computer system must have at least some knowledge of the types of data that are to be processed and the results that are required. However, it is a term that has come to be applied to a class of systems that encode human expertise – either explicit or implicit – in machine-readable form to facilitate the solution of some problem, normally one that cannot be tackled efficiently by a conventional, deterministic computer program. Such systems, often referred to as "expert systems", "intelligent knowledge-based systems" or "fifth generation computer systems" were much to the fore during the Eighties and early Nineties; they are rather less prominent now, with many of the basic techniques that were developed then having been assimilated into more conventional types of computer system. Interestingly, much of the early work on expert systems was carried out in the field of structural chemistry, with three applications being of particular importance: computer-aided structure elucidation (CASE); computer-aided synthesis design (CASD); and *de novo* design.

Structure elucidation is the task of identifying an unknown molecule given knowledge of its properties, which can be of any type although spectral properties have been the principal focus of work in CASE. There are two ways in which a computer can be used to assist the analyst when faced with an unknown molecule. The first, and simpler approach, is to carry out a database search, matching the spectrum of the unknown molecule with those available in an existing database; a complete or partial match can then suggest the identity or the principal substructural components of the unknown molecule [145,146]. The second, expert-systems approach derives from some of the very earliest work in the area of expert systems. This was the DENDRAL project at Stanford University for the analysis of mass spectra [147], which derived from work by Lederberg [148]. Given the mass spectrum and the molecular formula of the unknown molecule, the program would exhaustively generate all

possible molecules satisfying these constraints. The spectrum of each generated molecule would be computed, and then compared with the source data, this process identifying further constraints that could be included in subsequent iterations of the generation cycle. The process continues until the unknown has been identified or until it is not possible to identify any further constraints. Although massively influential in the development of expert systems in general, DENDRAL was never as successful in practice as some of its proponents claimed [146]. However, the techniques that were developed for generating structures have proved to be of widespread applicability, and systems based on a range of types of spectral data (including not just mass spectra but also nuclear magnetic resonance and infra-red spectroscopy) are widely used [149], as are systems for searching databases of spectra [150-152].

Another application of expert systems to structural chemistry is the area of computer-aided synthesis design (CASD), as reviewed recently by Ott [153]. CASD was first suggested as a possible area of research in Vleduts' 1963 paper on automatic reaction indexing that has been mentioned in Section 5 above [33]. Given stored information about the most common reactions and the conditions under which they could be applied successfully, Vleduts suggested that a computer could be used to generate a sequence of reactions that would result in the generation of a user-specified synthetic target in acceptable yield. Descriptions of synthesis in the chemical literature move from the starting materials to the final products, even though a synthetic chemist will normally design the synthesis by starting with the final product and then working backwards until known starting materials are reached. This *retrosynthetic* approach was the basis for the first published description of a CASD system: Corey and Wipke's OCSS (for Organic Chemical Simulation of Syntheses) program [50]. The retrosynthetic approach attracted much attention in the Seventies and Eighties, with programs such as LHASA [154] and SECS [155] undergoing extensive development, much of it in collaboration with industry who saw such programs as a complement to the work of their synthetic chemists [156]. However, it came to be realised that very large amounts of synthetic knowledge needed to be captured from the chemists and then encoded in machine-readable form before the programs could be expected to perform at a reasonable level of effectiveness [157]. Ugi and Gillespie [158] were the first to advocate an alternative approach in which the computer would take a set of starting materials and then generate synthetic pathways by the making and breaking of bonds. The program CICLOPS operated in a fully automated mode that was not constrained by the existing chemical knowledge, and that could thus generate all syntheses that were mathematically feasible [54], whereas retrosynthetic programs involve considerable interaction with the chemist running the program. Gasteiger and colleagues suggested that the effectiveness of CICLOPS and similar programs could be enhanced by the inclusion of chemical knowledge, in the shape of computed physical properties such as heats of formation and reaction enthalpies, that could be used to assess the feasibility of the suggested molecules [159]. The resulting EROS program has undergone extensive development over the years [160] and there is an associated program, called WODCA, for reaction prediction that uses both forward and backward planning [161]. An important component of any CASD program is a module to predict the synthetic feasibility of the molecules under consideration, and such procedures are now being used more generally in drug discovery programmes [162].

The final example of knowledge-based systems to be described here are the *de novo* design programs, which produce novel molecules that possess specific properties, typically the ability to fit within the binding site of a biological target such as an enzyme. They can hence be regarded as complementary to the docking programs discussed in Section 5: docking identifies known molecules that are able to fit into the binding site, where as *de novo* design generates unknown molecules with this ability. The approach was first described by Danziger and Dean [61], whose HSITE program identified those regions in a receptor that could form strong hydrogen-bonding interactions with a ligand, thus specifying geometric constraints that must be satisfied by potential ligands. A new molecule is then designed by placing appropriate molecular fragments – typically individual atoms or substructures chosen from a dictionary - into the binding site at locations that satisfy these constraints, and then by linking these fragments together to form a connected entity. Other programs soon followed; examples that continue to be widely used include LUDI [163], which includes one of the most widely used functions for estimating the energy of binding for a suggested molecule, and SPROUT [164], which includes both hydrogen-bonding and hydrophobic interactions in its scoring function. Most work on *de novo* design has focussed on molecules that will fit a binding site, but any type of constraint can be used, for example ranges of values for chemical and physical properties [165]. Schneider and Fechner review the current state-of-the-art, and include several examples of the use of *de novo* programs in the design of bioactive molecules, while emphasising that the

principal role of such programs is to suggest novel structural types for further consideration rather than to provide fully-fledged lead molecules [166].

8. Diversity analysis and library design

The last area of research to be discussed here is also the most recent, being driven by the developments in combinatorial chemistry and high-throughput screening that took place in the early Nineties. These technological improvements meant that it was now possible to synthesise and to test vastly more molecules than had previously been the case. However, real-world experiments were still expensive and there was hence much interest in computational methods to ensure that those molecules that were put forward for testing could provide the maximum amount of information to support lead discovery in a cost-effective manner. This requirement led initially to work on maximising the structural *diversity* (i.e., the level of dissimilarity) of the molecules that are submitted for biological testing, whilst minimising the numbers of molecules that are tested. The Similar Property Principle discussed in Section 4 implies that structurally similar molecules are likely to exhibit the same bioactivity, and hence the synthesis of large numbers of structurally related molecules is unlikely to result in a commensurate amount of useful structure-activity data. Patterson *et al.* [66] postulated the related, but distinct, concept of *Neighbourhood Behaviour*, which states that structurally dissimilar molecules may give different biological responses. The maximum amount of structure-activity information that can be extracted from testing some fixed number of molecules (as determined by the testing capacity that is available) will thus be obtained by selecting a set of molecules that are as diverse as possible.

Structural diversity had long been recognised as an important factor in the selection of compounds for testing [167], but it was the vast libraries of compounds that became available as a consequence of combinatorial chemistry that focused interest on computer techniques for diversity analysis. Two problems were of initial interest: selecting a set of molecules from those already available, either from a corporate database or from commercial suppliers; or selecting a set of molecules that could be obtained from an appropriately designed combinatorial synthesis. The basic problem of selecting the most-diverse subset of a set of available (or possible) molecules is a very simple one; however, it is also one that is computationally infeasible, as there is an astronomical number of subsets that can be chosen from a dataset of non-trivial size. A wide range of techniques was hence suggested for selecting diverse sets of molecules, whilst not being able to guarantee the identification of the optimally diverse subset. Examples of these techniques are described in detail in two books [168,169] and a recent review [170], with Martin *et al.* providing an interesting overview of the early history of molecular diversity analysis [171]. We exemplify this work by two of the earliest, and most heavily cited, studies, those by Martin *et al.* [64] and by Brown and Martin [65].

Combinatorial synthesis operates by reacting together sets of reactant molecules in parallel to yield a set of products called a combinatorial library, e.g., sets of acids and sets of amines to yield a combinatorial library of amides. Martin *et al.* [64] described the use of similarity measures based on fingerprints and on computed properties to select diverse sets of reactants, with the expectation that this would yield a diverse set of products in the resulting combinatorial library. This approach was rapidly taken up, and there is now an extensive literature on the use of reactant-based selection to ensure diverse combinatorial libraries. Later work focussed on the diversity of the final library rather than of the input reactants, and demonstrated that such product-based approaches could yield more diverse libraries, albeit at the cost of increased computational complexity [172]. The paper by Brown and Martin [65] compared different types of clustering method and structural descriptor in terms of their ability to predict a range of types of property, and hence of their suitability for compound selection. The study is notable not only in terms of the very detailed comparisons that were carried out (and also in a second, related paper [173]) but also because it concluded that simple, 2D descriptors were at least as effective as more sophisticated 3D descriptors for database-scale operations such as compound selection. The latter, surprising result has been confirmed in several subsequent studies of compound selection, although it is probably the case that an appropriate level of 3D representation has yet to be identified, rather than that 3D representations are inherently

less suitable for database-scale operations. It is also the case that much 3D information is implicit in the 2D structure of a molecule, especially if there are few rotatable bonds.

Initial studies of methods for diversity analysis focussed on the identification of sets of molecules that were structurally diverse, but it was soon realised that other factors also needed to be taken into account when selecting molecules for biological testing. For example, Gillet and co-workers have described the use of multiobjective optimisation to ensure the design of libraries that are not only structurally diverse but that contain molecules whose physicochemical properties resemble those of known drugs [174]. There is widespread interest in such *drug-likeness* (or *drugability*) studies, driven in large part by the "Rule of Five" first suggested by Lipinski *et al.* [45]. Methods of combinatorial synthesis had been rapidly adopted by the pharmaceutical industry; however, whilst these had resulted in very large numbers of molecules, they had not resulted in significantly larger numbers of bioactive molecules than could be obtained using conventional synthetic methods. Lipinski *et al.* analysed over two-thousand molecules that had entered phase II clinical trials (i.e., the phase in drug discovery where a potential drug is given to a small number of patients for initial studies of efficacy and side-effects) and observed that many of these obeyed simple physicochemical constraints that were simple multiples of five, e.g., the molecular weight should be less than 500 and the molecule should contain not more than five hydrogen-bond donor features. Molecules not satisfying these physicochemical constraints were likely to exhibit poor absorption or permeation, thus providing an obvious filtering mechanism for the selection of molecules and for the design of new combinatorial libraries. Subsequent studies have involved more detailed analyses of the physicochemical requirements for activity [175], the differences in properties between leads (i.e., molecules that are considered appropriate for detailed study in a drug-discovery programme) and drugs (i.e., molecules that get to the stage of being administered to patients) [176,177], the use of machine-learning tools to differentiate between databases of drugs and (assumed) non-drugs [178-180], and the development of analogous techniques for the design of agrochemicals [181].

9. Conclusions

In this paper, we have sought to highlight some of the major contributions to the historical development of chemoinformatics, although considerations of length inevitably mean that many other important papers have had to be omitted or merely mentioned in passing. However, it is hoped that this personal selection – however biased – is sufficient to make clear the intellectual debts that we owe to the early pioneers, many of whose techniques are still in widespread use many years after they were first published.

Chemoinformatics is, of course, continuing to develop, with three areas of particular importance for the next few years. The first, already mentioned, area is the use of machine learning methods for virtual screening. Machine learning and data mining is the subject of intense research in computer science, with the resulting methodologies starting to be applied in very many application areas, including chemoinformatics. Thus techniques such as decision trees, kernel discrimination, and support vector machines have been rapidly adopted for chemoinformatics applications and this trend will undoubtedly continue as new techniques become available [182]. The second area is ADMET prediction (standing for absorption, distribution, metabolism, excretion and toxicity). Work in QSAR over many years has resulted in reasonably effective methods for the prediction of biological activity; the aim now is to extend these methods to enable the prediction of these more complex types of pharmacokinetic and biological properties. The work mentioned previously on drug-likeness can be regarded as a first step in this direction, with measures of drug-likeness representing an implicit codification of the pharmacokinetic properties required for a molecule to be not just bioactive but also potentially a drug; ADMET prediction studies try to model these properties explicitly [183]. The third area arises from the observation that QSAR uses computationally simple, but surprisingly effective, techniques to model the requirements for bioactivity. As noted previously, the emergence of chemoinformatics has been driven in large part by the scaling-up of these techniques, which had traditionally been aimed at just a few tens of molecules, to the very large datasets that characterise modern pharmaceutical research. Given the successes that have been achieved thus far, it seems not unreasonable to expect that further increases in effectiveness could be achieved by application of the

sophisticated techniques of computational chemistry. These have traditionally been aimed at the detailed analysis of small numbers of molecules, but improvements in computer hardware and software mean that the methods are starting to be applied on a significantly larger scale than heretofore, as exemplified by the work of Beck *et al*. [139], and this trend can only increase further. In brief, we can expect the next fifty years to be at least as productive as the fifty years that have passed since the founding of the Institute of Information Scientists.

Acknowledgements. I thank Val Gillet and Wendy Warr for their comments on this paper.

References

[1] R.V. Williams and M.E. Bowden, Chronology of Chemical Information Science (at http://www.libsci.sc.edu/bob/chemnet/chchron.htm).

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

[3] S.R. Heller (ed.), *The Beilstein Online Database - Implementation, Content, and Retrieval* (American Chemical Society, Washington DC, 1990).

[4] R.T. Bottle and J.F.B. Rowland (eds.), *Information Sources in Chemistry* 4th ed. (Bowker-Saur, London, 1993).

[5] R.E. Maizel, *How to Find Chemical Information*, 3rd ed. (Wiley, New York, 1998).

[6] W.A. Warr and C. Suhr, *Chemical Information Management* (VCH, Weinheim, 1992).

[7] C.A. Orengo, J.M. Thornton and D.Y. Jones (eds.), *Bioinformatics* (Bios Scientific Publishers Ltd, Abingdon, 2002).

[8] A.M. Lesk, *An Introduction to Bioinformatics*, 2nd ed. (Oxford University Press, Oxford, 2005).

[9] W.V. Metanomski, *50 Years of Chemical Information in the American Chemical Society 1943-1993* (American Chemical Society, Washington DC, 1993).

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

[11] M.F. Lynch, J.M. Harrison, W.G. Town and J.E. Ash, *Computer Handling of Chemical Structure Information* (Macdonald London, 1971).

[12] W.T. Wipke, S. Heller, R. Feldmann and E. Hyde (eds.), *Computer Representation and Manipulation of Chemical Information* (John Wiley, New York, 1974).

[13] J.E. Ash and E. Hyde (eds.), *Chemical Information Systems* (Ellis Horwood, Chichester, 1975).

[14] J.E. Ash, P.A. Chubb, S.E. Ward, S.M. Welford and P. Willett (eds.), *Communication, Storage and Retrieval of Chemical Information* (Ellis Horwood, Chichester, 1985).

[15] J.E. Ash, W.A. Warr and P. Willett (eds.), *Chemical Structure Systems* (Ellis Horwood, Chichester, 1991).

[16] W.E. Warr (ed.), *Chemical Structures. The International Language of Chemistry* (Springer-Verlag, Berlin, 1988).

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

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

[19] W.A. Warr, Paper presented at the 218th American Chemical Society National Meeting, New Orleans, August 22-26, 1999, 1999.

[20] J. Gasteiger, The central role of chemoinformatics, *Chemometrics and Intelligent Laboratory Systems* 82 (2006) 200-209.

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

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

[23] H. Kubinyi, From narcosis to hyperspace: the history of QSAR, *Quantitative Structure-Activity Relationships* 21 (2002) 348-356.

[24] H. Skolnik, The journal for chemical information and computer scientists: a 25-year perspective, *Journal of Chemical Information and Computer Sciences* 25 (1985) 137-140.

[25] H. Schofield, G. Wiggins and P. Willett, Recent developments in chemoinformatics education, *Drug Discovery Today* 6 (18) (2001) 931-934.

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

[27] A.R. Leach and V.J. Gillet, *An Introduction to Chemoinformatics* (Kluwer, Dordrecht, 2003).

[28] J. Gasteiger and T. Engel (eds.), *Chemoinformatics: A Textbook* (Wiley-VCH, Weinheim, 2003).

[29] J. Gasteiger (ed.), *Handbook of Chemoinformatics* (Wiley-VCH, Weinheim, 2003).

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

[31] N. Bishop, V.J. Gillet, J.D. Holliday and P. Willett, Chemoinformatics research at the University of Sheffield: a history and citation analysis, *Journal of Information Science* 29 (2003) 249-267.

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

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

[34] M.F. Lynch, J.M. Barnard and S.M. Welford, Computer storage and retrieval of generic chemical structures in patents, Part 1. Introduction and general strategy, *Journal of Chemical Information and Computer Sciences* 21 (1981) 148-150.

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

[36] Y.C. Martin, M.G. Bures, E.A. Danaher, J. Delazzer, I. Lico and P.A. Pavlik, A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists, *Journal of Computer-Aided Molecular Design* 7 (1) (1993) 83-102.

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

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

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

[40] P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth, 1987).

[41] P. Willett, Similarity-based virtual screening using 2D fingerprints, *Drug Discovery Today* 11 (2006) 1046-1053.

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

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

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

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

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

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

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

[49] E. Hyde, F.W. Matthews, L.H. Thomson and W.J. Wiswesser, Conversion of Wiswesser notation to a connectivity matrix for organic compounds, *Journal of Chemical Documentation* 7 (1967) 200-204.

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

[51] J.E. Crowe, M.F. Lynch and W.G. Town, Analysis of structural characteristics of chemical compounds in a large computer-based file. I. Non-cyclic fragments, *Journal of the Chemical Society (C)* (1970) 990-996.

[52] J.G. Topliss and R.J. Costello, Chance correlations in structure-activity studies using multiple regression analysis, *Journal of Medicinal Chemistry* 15 (1972) 1066-1068.

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

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

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

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

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

[58] M.F. Lynch and P. Willett, The automatic detection of chemical reaction sites, *Journal of Chemical Information and Computer Sciences* 18 (1978) 154-159.

[59] R.R. Marshall, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler and D.A. Dunn, 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* Vol. 112, (American Chemical Society, Washington DC, 1979).

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

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

[62] J. Gasteiger, C. Rudolph and J. Sadowski, Automatic generation of 3D atomic coordinates for organic molecules, *Tetrahedron Computer Methodology* 3 (1990) 537-547.

[63] M.A. Johnson and G.M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity* (John Wiley, New York, 1990).

[64] E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong and W.H. Moos, Measuring diversity - experimental-design of combinatorial libraries for drug discovery, *Journal of Medicinal Chemistry* 38 (9) (1995) 1431-1436.

[65] R.D. Brown and Y.C. Martin, 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 (1996) 572-584.

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

[67] W. Graf, H.K. Kaindl, H. Kniess, B. Schmidt and R. Warszawski, 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 (1979) 51-55.

[68] P.G. Dittmar, N.A. Farmer, W. Fisanick, R.C. Haines and J. Mockus, 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 (1983) 93-102.

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

[70] S.R. Heller, G.W.A. Milne and R.J. Feldmann, A computer-based chemical information system, *Science* 195 (1977) 253-259.

[71] A.P. Johnson, Computer aids to synthesis planning, *Chemistry in Britain* 21 (1) (1985) 59-67.

[72] W.T. Wipke, Exploring Reactions with REACCS, 188th National Meeting of the American Chemical Society, American Chemical Society, Philadelphia PA, 1984.

[73] F.H. Allen, S. Bellard, M.D. Brice, B.A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B.G. Hummelink-Peters, O. Kennard, W.D.S. Motherwell, J.R. Rogers and D.G. Watson, The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information, *Acta Crystallographica B* 35 (1979) 2331-2339.

[74] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: a computer-based archival file for macromolecular structures, *Journal of Molecular Biology* 112 (1977) 535-542.

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

[76] D.J. Gluck, A chemical structure storage and search system developed at DuPont, *Journal of Chemical Documentation* 5 (1965) 43-51.

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

[78] R. Freeland, S. Funk, L. O'Korn and G. Wilson, The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula, *Journal of Chemical Information and Computer Sciences* 19 (1979) 94-98.

[79] G. Salton and E.H. Sussenguth, Some flexible information retrieval systems using structure matching procedures, *AFIPS Conference Proceedings, Spring Joint Computer Conference* 25 (1964) 587-597.

[80] A.T. Brint and P. Willett, Pharmacophoric pattern matching in files of 3-D chemical structures: comparison of geometric searching algorithms, *Journal of Molecular Graphics* 5 (1987) 49-56.

[81] G.M. Downs, M.F. Lynch, P. Willett, G.A. Manson and G.A. Wilson, Transputer implementations of chemical substructure searching algorithms, *Tetrahedron Computer Methodology* 1 (1988) 207-217.

[82] L. Hodes, Selection of descriptors according to discrimination and redundancy - application to chemical-structure searching, *Journal of Chemical Information and Computer Sciences* 16 (1976) 88-93.

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

[84] C.L. Wilkins and M. Randic, A graph theoretical approach to structure-property and structure-activity correlation, *Theoretica Chimica Acta* 58 (1980) 45-68.

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

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

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

[88] H.-J. Böhm and G. Schneider (eds.), *Virtual Screening for Bioactive Molecules.* (Wiley-VCH, Weinheim, 2000).

[89] G. Klebe (ed.), *Virtual Screening: an Alternative or Complement to High Throughput Screening* (Kluwer, Dordrecht, 2000).

[90] J. Alvarez and B. Shoichet (eds.), *Virtual Screening in Drug Discovery* (CRC Press, Boca Raton, 2005).

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

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

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

[94] J.E. Crowe, P. Leggate, B.N. Rossiter and J.F.B. Rowland, The searching of Wiswesser line notations by means of a character-matching serial search, *Journal of Chemical Documentation* 13 (1973) 85-92.

[95] L.H. Thomson, E. Hyde and F.W. Matthews, Organic search and display using a connectivity matrix derived from Wiswesser notation, *Journal of Chemical Documentation* 7 (1967) 204-209.

[96] E. Hyde and L.H. Thomson, Structure display, *Journal of Chemical Documentation* 8 (1968) 138-146.

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

[98] J.E. Armitage and M.F. Lynch, Automatic detection of structural similarities among chemical compounds, *Journal of the Chemical Society (C)* (1967) 521-528.

[99] P.E. Blower and R.C. Dana, Creation of a chemical reaction database from the primary literature, in: P. Willett (ed.), *Modern Approaches to Chemical Reaction Searching*, (Gower, Aldershot, 1986).

[100] T.E. Moock, J.G. Nourse, D. Grier and W.D. Hounshell, The implementation of atom-atom mapping and related features in the Reaction Access System (REACCS), in: W.A. Warr (ed.), *Chemical Structures. The International Language of Chemistry*, (Springer-Verlag, Berlin, 1988).

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

[102] M.F. Lynch and J.D. Holliday, The Sheffield Generic Structures Project - a retrospective review, *Journal of Chemical Information and Computer Sciences* 36 (1996) 930-936.

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

[104] G.M. Downs and J.M. Barnard, Techniques for generating descriptive fingerprints in combinatorial libraries, *Journal of Chemical Information and Computer Sciences* 37 (1997) 59-61.

[105] O. Güner (ed.), *Pharmacophore Perception, Development and Use in Drug Design* (International University Line, La Jolla CA, 2000).

[106] P. Gund, W.T. Wipke and R. Langridge, Computer searching of a molecular structure file for pharmacophoric patterns, *International Conference on Computers in Chemical Research and Education* Vol. 3 (Elsevier, Amsterdam, City, Year).

[107] F.H. Allen, The Cambridge Structural Database: a quarter of a million crystal structures and rising, *Acta Crystallographica Section B-Structural Science* 58 (2002) 380-388.

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

[109] C. Hiller and J. Gasteiger, Ein automatisierter molekülbaukasten, in: J. Gasteiger (ed.), *Software-Entwicklung in der Chemie 1*, (Springer Verlag, Berlin, 1987).

[110] D.V.S. Green, Automated three-dimensional structure generation, in: Y.C. Martin, P. Willett (eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, (American Chemical Society, Washington DC, 1998).

[111] P. Willett, Searching for pharmacophoric patterns in databases of three-dimensional chemical structures, *Journal of Molecular Recognition* 8 (1995) 290-303.

[112] Y.C. Martin and P. Willett (eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications* (American Chemical Society, Washington, 1998).

[113] G. Jones, P. Willett and R.C. Glen, A genetic algorithm for flexible molecular overlay and pharmacophore elucidation, *Journal of Computer-Aided Molecular Design* 9 (1995) 532-549.

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

[115] U. Rester, Dock around the clock - current status of small molecule docking and scoring, *QSAR & Combinatorial Science* 25 (2006) 605-615.

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

[117] Y.C. Martin, *Quantitative Drug Design. A Critical Introduction* (Marcel Dekker, New York, 1978).

[118] C. Hansch and A. Leo, *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology* (American Chemical Society, Washington DC, 1995).

[119] R. Mannhold and H. van de Waterbeemd, Substructure and whole molecule approaches for calculating log P, *Journal of Computer-Aided Molecular Design* 15 (2001) 337-354.

[120] J.G. Topliss and R.P. Edwards, Chance factors in studies of quantitative structure-activity relationships, *Journal of Medicinal Chemistry* 22 (1979) 1238-1244.

[121] D.J. Livingstone and D.W. Salt, Judging the significance of multiple linear regression models, *Journal of Medicinal Chemistry* 48 (2005) 661-663.

[122] D.W. Salt, S. Ajmani, R. Crichton and D.J. Livingstone, An improved approximation to the estimation of the critical F values in best subset regression, *Journal of Chemical Information and Modeling* 47 (2007) 143-149.

[123] D.T. Manallack, D.D. Ellis and D.J. Livingstone, Analysis of linear and nonlinear QSAR data using neural networks, *Journal of Medicinal Chemistry* 37 (1994) 3758-3767.

[124] D.J. Livingstone, The characterisation of chemical structures using molecular properties, *Journal of Chemical Information and Computer Sciences* 40 (2000) 195-209.

[125] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2002).

[126] A. Golbraikh and A. Tropsha, Beware of q²!, *Journal of Molecular Graphics and Modelling* 20 (2002) 269-276.

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

[128] S. Wold and W.J. Dunn, Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability, *Journal of Chemical Information and Computer Sciences* 23 (1983) 6-13.

[129] G.W. Adamson and J.A. Bush, A method for relating the structure and properties of chemical compounds, *Nature* 248 (1974) 406-407

[130] G.W. Adamson and D. Bawden, A method of structure-activity correlation using Wiswesser Line Notation, *Journal of Chemical Information and Computer Sciences* 15 (1975) 215-220.

[131] W. Tong, D.R. Lowis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage and D.M. Sheehan, 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 (1998) 669-677.

[132] A. Ormerod, P. Willett and D. Bawden, Comparison of fragment weighting schemes for substructural analysis, *Quantitative Structure-Activity Relationships* 8 (1989) 115-129.

[133] L. Hodes, G.F. Hazard, R.I. Geran and S. Richman, A statistical-heuristic method for automated selection of drugs for screening, *Journal of Medicinal Chemistry* 20 (1977) 469-475.

[134] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, 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 (2006) 462-470.

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

[136] A.M. Doweyko, The hypothetical active site lattice: an approach to modelling active sites from data on inhibitor molecules, *Journal of Medicinal Chemistry* 31 (1988) 1396-1406.

[137] H. Kubinyi, G. Folkers and Y.C. Martin (eds.), *3D QSAR in Drug Design* (Kluwer/ESCOM, Leiden, 1998).

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

[139] B. Beck, A. Horn, J.E. Carpenter and T. Clark, Enhanced 3D-databases: a fully electrostatic database of AM1-optimized structures *Journal of Chemical Information and Computer Sciences* 38 (1998) 1214-1217.

[140] D. Mayer, C.B. Naylor, I. Motoc and G.R. Marshall, A unique geometry of the active site of angiotensin-converting enzyme consistent with structure-activity studies, *Journal of Computer-Aided Molecular Design* 1 (1987) 3-16.

[141] R.A. Dammkoehler, S.F. Karasek, E.F.B. Shands and G.R. Marshall, Constrained search of conformational hyperspace, *Journal of Computer-Aided Molecular Design* 3 (1989) 3-21.

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

[143] C. Bron and J. Kerbosch, Algorithm 457. Finding all cliques of an undirected graph, *Communications of the ACM* 16 (1973) 575-577.

[144] Y.C. Martin, Pharmacophore mapping, in: Y.C. Martin, P. Willett (eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, (American Chemical Society, Washington, 1998).

[145] N.A.B. Gray, *Computer Assisted Structure Elucidation* (John Wiley, New York, 1986).

[146] N.A.B. Gray, Computer-aided structure elucidation, in: J.E. Ash, W.A. Warr, P. Willett (eds.), *Chemical Structure Systems*, (Ellis Horwood, Chichester, 1991).

[147] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: the DENDRAL Project* (McGraw-Hill, New York, 1980).

[148] J. Lederberg, Topological mapping of organic molecules, *Proceedings of the National Academy of Sciences, USA* 53 (1965) 134-139.

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

[150] W. Bremser, HOSE – a novel substructure code, *Analytica Chimica Acta* 103 (1978) 355-365.

[151] B.A. Jezl and D.L. Dalrymple, Computer Program for the Retrieval and Assignment of Chemical Environments and Shifts to Facilitate Interpretation of Carbon-13 Nuclear Magnetic Resonance Spectra, *Analytical Chemistry* 47 (1975) 203-207.

[152] W.A. Warr, Spectral databases, *Chemometrics and Intelligent Laboratory Systems* 10 (1991) 279-292.

[153] M.A. Ott, Cheminformatics and organic chemistry. Computer-assisted synthetic analysis, in: J.H. Noordik (ed.), *Cheminformatics Developments: History, Reviews and Current Research*, (IOS Press, Amsterdam, 2004).

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

[155] W.T. Wipke, G.I. Ouchi and S. Krishnan, Simulation and Evaluation of Chemical Synthesis - SECS: an application of artificial intelligence techniques, *Artificial Intelligence* 11 (1978) 173-193.

[156] W.T. Wipke and W.J. Howe (eds.), *Computer Assisted Organic Synthesis* (American Chemical Society, Washington, 1977).

[157] F. Loftus, Computer-aided synthesis design, in: J.E. Ash, W.A. Warr, P. Willett (eds.), *Chemical Structure Systems*, (Ellis Horwood, Chichester, 1991).

[158] I. Ugi and P.D. Gillespie, Matter preserving synthetic pathways and semi-empirical computer assisted planning of syntheses, *Angewandte Chemie International Edition* 10 (1971) 915-919.

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

[160] W.D. Ihlenfeldt and J. Gasteiger, Computer-assisted planning of organic syntheses: the second generation of programs, *Angewandte Chemie International* 34 (1995) 2613-2633.

[161] J. Gasteiger, W.D. Ihlenfeldt, P. Rose and R. Wanke, Computer-assisted prediction and synthesis design, *Analytica Chimica Acta* 235 (1990) 65-75.

[162] J.C. Baber and M. Feher, Predicting synthetic accessibility: application in drug discovery and development, *Mini Reviews in Medicinal Chemistry* 4 (2004) 681-692.

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

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

[165] R.C. Glen and A.W.R. Payne, A genetic algorithm for the automated generation of molecules with constraints, *Journal of Computer-Aided Molecular Design* 9 (1995) 181-202.

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

[167] P. Willett, V. Winterman and D. Bawden, 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 (1986) 109-118.

[168] P.M. Dean and R.A. Lewis (eds.), *Molecular Diversity in Drug Design* (Kluwer, Amsterdam, 1999).

[169] A.K. Ghose and V.N. Viswanadhan (eds.), *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery* (Marcel Dekker, New York, 2001).

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

[171] Y.C. Martin, P. Willett, M. Lajiness, M. Johnson, G.M. Maggiora, E. Martin, M.G. Bures, J. Gasteiger, R.D. Cramer, R.S. Pearlman and J.S. Mason, Diverse viewpoints on computational aspects of molecular diversity, *Journal of Combinatorial Chemistry* 3 (2001) 231-250.

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

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

[174] V.J. Gillet, W. Khatib, P. Willett, P.J. Fleming and D.V.S. Green, Combinatorial library design using a multiobjective genetic algorithm, *Journal of Chemical Information and Computer Sciences* 42 (2) (2002) 375-385.

[175] T.I. Oprea, Property distribution of drug-related chemical databases, *Journal of Computer-Aided Molecular Design* 14 (3) (2000) 251-264.

[176] M.M. Hann, A.R. Leach and G. Harper, Molecular complexity and its impact on the probability of finding leads for drug discovery, *Journal of Chemical Information and Computer Sciences* 41 (3) (2001) 856-864.

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

[178] Ajay, W.P. Walters and M.A. Murcko, Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?, *Journal of Medicinal Chemistry* 41 (18) (1998) 3314-3324.

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

[180] V.J. Gillet, P. Willett and J. Bradshaw, Identification of biological activity profiles using substructural analysis and genetic algorithms, *Journal of Chemical Information and Computer Sciences* 38 (2) (1998) 165-179.

[181] E.D. Clarke and J.S. Delaney, Physical and molecular properties of agrochemicals: an analysis of screen inputs, hits, leads, and products, *Chimia* 57 (2003) 731-734.

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

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