Digital Libraries in Chemistry: Providing Access to Chemical Structure Information

Peter Willett

Department of Information Studies, University of Sheffield, Western Bank,
Sheffield S10 2TN, United Kingdom

Abstract

Chemical structures (either in 2D or in 3D) play a central role in the design of information systems to support research in the chemical sciences. This paper summarises the principal means of access, substructure searching and similarity searching, to databases of chemical structures, and shows how these access methods are used to support the discovery of novel bioactive molecules. The searching of textual and chemical databases is then compared, and the paper concludes with some current research areas.

Keywords

Chemical Databases, Chemoinformatics, Clustering, Digital Libraries, Docking, Drug Discovery, Information Retrieval, Molecular Diversity Analysis, Substructural Analysis, Textual Databases

1 Introduction

Computer databases of textual materials first became available some four decades ago, as a result of the computerisation of the operations of the large abstracting and indexing services and of the legal full-text services. Since then, a huge range of types of text have become available, most obviously *via* the Web, together with an increasing volume of multimedia material, so that digital libraries are now the concern not just of the information specialist but also of vast numbers of private and corporate users throughout academe, business and commerce. In particular, many of the sciences have developed techniques for the storage, retrieval and processing of the types of information that are specific to their disciplines. Perhaps the most widely known of these specialist digital libraries are the databases of biological sequences that now underlie research throughout the biological and medical sciences; indeed, a new sub-discipline has arisen, called *bioinformatics*, relating to the development and exploitation of these databases. In this paper, we discuss the analogous sub-discipline, called *chemoinformatics*, that relates to the specialised digital libraries of chemical structure information that are used to support research and development in the chemical sciences. Thus far, their application has been most notable in the discovery of novel drugs by the pharmaceutical industry and, albeit to a lesser extent, of novel pesticides and fungicides by the agrochemicals industry; however, the techniques that we describe are applicable to the processing of any type of chemical database.

Textual chemical information, such as the bibliographic details of a journal article describing the synthesis of a particular substance, or numerical chemical information, such as the melting point and the molecular weight for that substance, can be stored and retrieved using conventional database methodologies. Very different approaches, however, are required to process the two-dimensional (2D) or three-dimensional (3D) structures of chemical compounds: it is this type of information that is discussed here and that is the principal concern of chemoinformatics.

It is only within the last few years that chemoinformatics has come to be recognised as a distinct topic of study [1-3], but it is perhaps worth noting that many of its techniques are, in fact, of long standing. Thus, some of the basic methods that are used today for the representation and searching of chemical structures were developed at Chemical Abstracts Service some four decades ago, at the same time as this organisation was carrying out some of the earliest work on the computerisation of textual databases; since then, the Chemical Abstracts Service database has grown to include information on more than thirty million different molecules. The recent prominence of chemoinformatics is principally as a result of technological developments in chemistry and biology. Specifically, the methods of *combinatorial chemistry* and *high-throughput screening* allow the synthesis and biological testing, respectively, of huge arrays of molecules in parallel, thus producing a data explosion that has spurred the development of sophisticated informatics and data analytic methods.

In this paper, we discuss some of the principal aspects of chemoinformatics, focusing on the processing of databases of chemical structures. The next two sections describe the main searching methods, and the ways that these methods are used to assist in the discovery of new drugs; the paper then looks at some of the similarities between chemical and textual digital libraries, and concludes by summarising current areas of research.

2 Representation and Searching of Chemical Structure Information

2.1 Structure and Substructure Searching

The principal method of representation for a 2D chemical structure diagram is a labelled graph (called a *connection table*) in which the nodes and edges of a graph represent the atoms and bonds, respectively, of a molecule. A chemical database can hence be represented by a large number of such graphs, with searching historically being carried out using two types of graph isomorphism algorithms. *Structure searching* involves an exact-match search of a chemical database for a specific query structure: this is required, for example, to retrieve the biological assay results and the synthetic details associated with a particular molecule. Such a search is effected by means of a graph isomorphism search, in which the graph describing the query molecule is checked for isomorphism with the graphs of each of the database molecules. *Substructure searching* involves a partial-match search of a chemical database to find all those molecules that contain a user-defined query substructure, irrespective of the environment in which that substructure occurs; for example, a user interested in antibiotics might wish to search a database to find all molecules that contain the characteristic penicillin ring nucleus.

A substructure search is effected by checking the graph describing the query substructure for subgraph isomorphism with the graphs of each of the database molecules [4]. However, subgraph isomorphism is known to belong to the class of NP-complete computational problems, and thus substructure searching in databases of non-trivial size might be expected to be computationally infeasible. It is made possible by the use of an initial *screen search*, where a screen is a substructural feature, the presence of which is necessary, but not sufficient, for a molecule to contain the query substructure. These features are typically small, atom-, bond- or ring-centred fragment substructures that are algorithmically generated from a connection table when a molecule is added to the database that is to be searched. For example, the well-known augmented atom fragment consists of an atom, and those atoms that are bonded directly to it.

The fragments that have been chosen for use in a screening system are listed in a fragment coding dictionary, which will typically contain a few hundred or a few thousand carefully selected fragments. Each of the database structures is analysed to identify those screens from the coding dictionary that are present, and then represented for search by a fixed-length bit-string in which the non-zero bits correspond to the screens that are present. The query substructure is subjected to the same process and the screen search then involves checking the bit-strings representing each database structure for the presence of the screens that are encoded in the bit-string representing the query substructure.

Only a very small fraction of a database will normally contain all of the screens that have been assigned to a query substructure, and thus only these few molecules need to undergo the final, time-consuming graph-matching search. This checks to see whether there is an exact subgraph isomorphism between the graph representing the query substructure and the graphs representing each of the database structures that have passed the screen search. This simple, two-stage procedure (*i.e.*, screen searching and subgraph searching) has formed the basis for most operational 2D substructure searching systems to date.

Similar techniques are used for 3D substructure searching [4, 5], where there is a need to identify molecules that contain a query *pharmacophore*, i.e., a set of atoms having some specific geometric relationship to each other. Here, the nodes and edges of a chemical graph denote the atoms and the inter-atomic distances, and the fragments that are encoded in the bit-strings describe pairs or triplets of atoms and the associated inter-atomic distances. Only simple modifications to the 2D methods described previously are required to enable searchers for pharmacophores to be carried out. However, significant complexities needed to be overcome before these representations and searching methods were extended to encompass the fact that most molecules are *flexible*, i.e., they adopt not just a single, fixed 3D shape but can exist in some, many, or very may different shapes, depending on the temperature and the external chemical environment. This means that the separation between each pair of atoms is not necessarily fixed, but typically covers a range of possible distances. This increases the complexity of the matching operations that are required; in particular, the screening and subgraph isomorphism searches need an additional, *conformational* search, which takes account of the precise geometries and energies of the various shapes that each potential hit molecule can adopt [5].

2.2 Similarity Searching

Substructure searching, whether in 2D or in 3D, provides an invaluable tool for accessing databases of chemical structures. It does, however, have several limitations that are inherent in the retrieval criterion that is being used, which is that a database structure must contain the entire query substructure in precisely the form that has been specified by the user. Firstly, and most importantly, a substructure search requires that the user who is posing the query must already have acquired a well-defined view of what sorts of structures are expected to be retrieved from the database. This is clearly very difficult at the start of an investigation, when perhaps only one or two active structures have been identified and when it is not at all clear which particular feature(s) within them are responsible for the observed activity. Secondly, there is very little control over the size of the output that is produced by a particular query substructure. Accordingly, the specification of a common ring system, such as the benzodiazepine system that forms the nucleus of many tranquillisers, can result in the retrieval of many thousands of compounds from a chemical database. Finally, a substructure search results in a simple partition of the database into two discrete sub-sets (i.e., those structures that contain the query and those that do not) and there is no direct mechanism by which the retrieved molecules can be ranked in order of decreasing probability of activity.

These limitations are entirely analogous to those suffered by Boolean methods for text retrieval [6, 7]. In just the same way as Boolean retrieval has increasingly been complemented, or even supplanted, by best-match retrieval methods in text search engines, so substructure searching has now been augmented by chemical *similarity searching* [8, 9]. Similarity searching requires the specification of an entire *target* structure, rather than the partial structure that is required for substructure searching. The target molecule is characterised by a set of structural features, and this set is compared with the corresponding sets of features for each of the database structures. Each such comparison enables the calculation of a measure of similarity between the target structure and a database structure, and the database is then sorted into order of decreasing similarity with the target. The output from the search is a ranked list, where the structures that the system judges to be most similar to the target structure are located at the top of the list. Accordingly, if an appropriate measure of similarity has been used, the first database structures inspected will be those that have the greatest probability of being of interest to the user [8].

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. A normalised association coefficient, typically the Tanimoto coefficient, is used to give a numeric value to the similarity between the target structure and each database structure, with this value being in range of zero (no bits in common) to unity (all bits the same) [9].

While fragment-based measures such as the Tanimoto coefficient provide a simple (indeed simplistic) picture of the similarity relationships between pairs of molecules, they are both efficient (since they involve just the application of logical operations to pairs of bit-strings) and effective (since they have been shown to be capable of bringing together molecules that are judged by chemists to be structurally similar to each other) in operation. The latter characteristic is most surprising, given that the fragments that are used for the calculation of the similarities were originally designed to maximise the efficiency of substructure searching, not the effectiveness of similarity searching. Moreover, they describe only the 2D structures of molecules, and take only implicit account of the 3D structures, which are known to be of crucial importance in determining physical, chemical and biological properties. It should be noted here that there is much current interest in measures of 3D similarity but these have not, in general, been found to be as generally effective as the simpler 2D measures [9].

3 Use of Chemical Structure Information in Pharmaceutical Research

Many different scientific disciplines (such as synthetic organic chemistry, structural biology, pharmacology and toxicology) are needed to discover the new drugs that are the lifeblood of the pharmaceutical industry. The huge costs and extended timescales that characterise the industry mean that it is willing and able to make very substantial investments in any technology that can increase the speed with which drugs, *i.e.*, novel chemical molecules with beneficial biological properties, are brought to the market place (and similar comments apply to the pesticides and fungicides developed by the agrochemicals industry). The sophistication of current chemoinformatics systems is one manifestation of this investment.

Much modern drug research is based on high-throughput screening (HTS), which involves a battery of biological tests, called assays, that can be applied rapidly to very large numbers of molecules to see whether any of them exhibit any potentially useful level of biological activity. This HTS-based approach is typically applied to the molecules stored in a company's corporate database of molecules that have been synthesised by the company over the years. This database embodies much of a company's intellectual property and is thus an obvious source for the discovery of novel, patentable drugs. HTS typically throws up a few molecules with the desired activity and these, possibly augmented by existing, known drugs from competitor companies, can then be used for similarity searches of both corporate and public databases, such as those produced by Chemical Abstracts Service and by the Beilstein Institute.

Structurally similar molecules are of potential interest in the discovery of novel bioactive molecules because of the *Similar Property Principle* [10], which states that molecules that have similar structures will have similar properties. Hence, if the target structure has some interesting property, e.g., it lowers a person's cholesterol level or alleviates the symptoms of a migraine attack, then molecules that are structurally similar to it are more likely to exhibit that property than are molecules that have been selected from a database at random [11]. This has led to similarity searching being widely used for *virtual screening*, i.e., the ranking of databases in order of decreasing probability of activity. This is done so as to maximise the cost-effectiveness of biological testing, by focussing attention on just those few molecules that have the highest *a priori* probabilities of activity [12]

Similarity searching hence provides a simple and direct way of identifying further molecules for biological testing. As more and more actives are identified in this way, it becomes possible to delineate the precise substructural characteristics that are necessary for activity. Once these characteristics are known, it becomes possible to define a substructural query, either in 2D or in 3D, that can be used as the basis for a substructure search. This alternative, and more precise, form of virtual screening is normally carried out in an iterative manner, with molecules retrieved in the initial search being tested for activity, and the results (both positive and negative) of these biological tests being used to refine the query for the second and subsequent substructure searches.

Taken together, similarity searching and substructure searching can hence be expected to identify a pool of molecules that can form the starting point for a development project. This will involve other, more sophisticated types of virtual screening (such as the docking and substructural analysis methods described in Section 5), tests for biological activity that are much more rigorous than the simple assays used in HTS, and detailed and time-consuming tests for other characteristics such as stability, oral availability and toxicity.

It will hence be clear that searching of chemical databases plays a key role in modern approaches to drug discovery. In fact, there are several other types of database processing that are typically involved, including the docking, diversity analysis and substructural analysis approaches that are discussed in Section 5 below.

4 Relationships between Textual and Chemical Digital Libraries

Information retrieval (IR) provides the tools that are used to search digital libraries. IR has traditionally focussed on textual data, although it is now being extended to multimedia resources such as speech and image databases. However, the basic concepts of IR are applicable, in principle, to any type of data, and there are clear links between textual and chemical retrieval. Indeed, we have already noted one such relationship when discussing the reasons why similarity searching complements substructure searching, in much the same way as best-match searching complements Boolean searching in the textual context. In fact, it is often, though by no means invariably, the case that algorithms and data structures that can be applied to one type of database processing can also be applied to the other.

There are several reasons for this close link between textual and chemical retrieval. Firstly, there are clear similarities in the ways that the two types of database records are characterised. The documents in a text database are each typically indexed (manually or automatically) by some small number of keywords from the myriad of possible words. In just the same way, the molecules in a chemical database are each characterised by some small number of substructural features (the fragments that are encoded in a molecule's fragment bit-string) that are again chosen from some very large number of possible substructural features. Moreover, both types of attribute follow a well-marked Zipfian distribution, with the skewed distributions that characterise the frequencies of occurrence of characters, character substrings and words in text databases being mirrored by the comparable distributions for the frequencies of chemical substructural moieties. These shared characteristics mean that the two types of database are amenable to efficient processing using the same types of file structure. Finally, in just the same way as a document either is, or is not, relevant to some particular user query, so a molecule is active, or is not active, in some particular biological test. This means that analogous measures of retrieval performance (such as precision and recall, or variants thereof) can be used to assess search effectiveness in both chemical and textual retrieval systems.

That said there are differences, principally arising from the natures of the object representations that are used. The graph characterising a 2D or a 3D chemical structure bears a much closer relationship to its parent molecule than do the character-strings representing the words comprising a textual document. These chemical graphs can be regarded as direct manifestations of the underlying wave equations that describe a molecule, and it has thus proved possible to develop powerful simulation techniques that enable the prediction of many molecular properties from a knowledge of their 2D or 3D structure [2, 3]. Many of these molecular modelling tools have no direct textual equivalent, as the use of natural language raises a host of linguistic problems that do not arise in the chemical context. There are also other, more computational differences based on the fact that molecules are represented by graphs (or the substructural fragments that can be generated from them) and documents by linguistic texts (or the individual words or phrases that can be generated from them).

Even so, the similarities that do exist mean that there are many types of chemical database processing that have a direct textual analogue, and *vice versa*. The nature and the extent of these relationships have been discussed previously [13]: here, we present just two, similarity-based examples to illustrate the strong links that exist. First, we have mentioned previously the Similar Property Principle, which provides a rationale for the use of similarity-based approaches to virtual screening. This Principle can be regarded as the chemoinformatics equivalent of the Cluster Hypothesis in IR [14], which states that documents that are similar tend to be relevant to the same requests: simply replace "document" in the Cluster Hypothesis by "molecule" and "relevant to the same requests" by "exhibit the same biological properties" and one has the Similar Property Principle. Second, data fusion has been extensively used in IR to combine the rankings obtained from multiple best-match searches for the same query, for example searches using different weighting schemes or different types of index terms. These experiments have shown that retrieval effectiveness is generally enhanced, as compared to the use of a single retrieval mechanism, if the results of several different searches are combined, typically by applying some sort of averaging procedure to the rankings resulting from the different mechanisms. We have been able to show that comparable improvements are obtained in the chemical context by fusing the rankings resulting from the use of different measures of chemical similarity: this seems to provide a very simple way of enhancing the performance of current systems for similarity searching [15].

5 Current Research Topics

As noted above, there are many other types of database processing that are used in chemoinformatics, and we describe here three such areas that are currently the subject of much active research and development [2, 3]. These are: *substructural analysis*, a statistical technique that uses the incidence of substructures in active and in inactive molecules to prioritise compounds for testing; *docking*, checking to see whether the shape of a molecule is compatible with the 3D structure of the biological receptor site with which it needs to interact; and *molecular diversity analysis*, which ensures that the molecules that go forward for biological testing are as structurally disparate as possible.

Substructural analysis involves the calculation of weights that relate the presence of a molecular feature to the probability that that molecule is active in a particular biological test system. Specifically, given a database of compounds for which the biological activities are available (the so-called *training-set*), substructural analysis develops weights that can then be used to select new compounds, from amongst those in the so-called *test-set*, for biological testing.

The weights that are used are based on the numbers of active and inactive molecules that do possess, and that do not posses, particular features so that, for example, a feature that tends to occur only in active molecules will be given a greater weight than one that is more randomly distributed between the active and inactive members of the training-set. These features are typically the 2D or 3D fragment substructures that are encoded in the fragment bit-strings that are used for the first stage of substructure searching. Once the training-set has been analysed, the resulting weights can then be used to score the molecules in the test-set. A test-set molecule is analysed to determine the features present, and that molecule's score is obtained by summing the weights for those features; the test-set molecules are then ranked in order of decreasing sums-of-weights, so that chemical synthesis and biological assays can be focused on those compounds that occur at the top of the resulting ranking and that thus have the greatest probability of being active.

Substructural analysis was first described some three decades ago in a much-cited paper by Cramer *et al.* [16] but this approach to virtual screening is now enjoying a resurgence of interest as a result of the large amounts of structural and biological data that are becoming available from combinatorial chemistry and HTS. An example of the sorts of computational approach that are now being used in substructural analysis is provided in a recent paper by Wilton *et al.* [17].

It is perhaps worth noting here that substructural analysis is yet another chemical approach that has a direct textual analogue. Specifically, the weights that are calculated in substructural analysis mirror closely the *relevance weights* that are used in information retrieval to estimate the extent to which the presence of a specific index term in a document affects the probability that that document is relevant to a particular query [18].

Substructural analysis is a virtual screening method that requires information about the 2D (or 3D) structures of known active and known inactive molecules. Docking is a sophisticated method for virtual screening that additionally requires information about one of the biological pathways that is associated with the illness for which a therapy is required. Specifically, 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 the pathway. 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.

Shape matching is a computationally demanding task and one for which many algorithmic approaches have been suggested [19]. The original description of docking, by Kuntz *et al.* [20], considered the fitting of just a single molecule into a protein active site; 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). In fact, the fitting operation involves not just matching geometric characteristics, such as inter-atomic distances, but also chemical considerations such as the extent to which atoms of one type in the drug are compatible with the atoms that they are mapped to in the receptor site. This brings added complexity, in terms of both the mechanistic knowledge and the computational complexity that is required.

Current work focuses on flexible docking, where account is taken of the fact that molecules and proteins can adopt different shapes; thus, adopting the lock-and-key metaphor, rather than trying to fit a metallic key into metallic lock, one is actually trying to fit two non-rigid objects. Current systems for virtual screening enable the docking of databases of flexible molecules into a rigid receptor; the inclusion of both types of flexibility in an efficient and an effective manner is still probably some years away.

The final database application to be considered here is molecular diversity analysis [21]. Like virtual screening, this also seeks to maximise the cost-effectiveness of drug discovery, but takes as its starting point the need to maximise the *diversity* (a widely used term for structural heterogeneity or structural dissimilarity) of the molecules that are submitted for biological testing (rather than maximising the probability of activity, which is the main aim of virtual screening). Although HTS is very rapid, it is still costly and there is hence a need to minimise the numbers of molecules that are assayed. The Similar Property Principle means that structurally similar molecules are likely to give similar biological responses; thus, if one wishes to maximise the information that can be gained from a fixed number of molecules about the relationship between structure and activity, then one should try to ensure that the molecules submitted for HTS are as structurally diverse as possible.

The need for diversity may sound like a statement of the obvious, but the practical realisation of this has proved to be very difficult. The inherently subjective concept of diversity is normally quantified using similarity-based techniques that are a natural development of those discussed previously: thus, a diverse subset of the molecules in a database is selected by consideration of their inter-molecular structural similarities, typically as determined by use of fragment bit-strings and the Tanimoto coefficient. The problem is that there is an astronomical number of possible subsets that can be generated from a database of non-trivial size: it is hence infeasible to consider all of them so as to identify the most diverse subset that can then be submitted for HTS. There has thus been much interest in alternative approaches for selecting diverse sets of molecules that maximise the coverage of structural space, whilst minimising the numbers of molecules put forward for testing. Cluster analysis, or automatic classification, was the first such technique to be used for this purpose.

Cluster analysis is the process of subdividing a group of objects into groups, or clusters, of objects that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity. In the chemical context, the aim is to ensure that structurally similar molecules are clustered together, so that each cluster represents a well-marked part of the chemical space spanned by a database. Then a structurally diverse subset can be generated by selecting a single representative molecule from each of the clusters resulting from the application of an appropriate clustering method to that database [22]. The representative molecule for each cluster is either selected at random or selected as being the closest to the cluster centroid. These selected molecules are then tested in the bioassay of interest: if any of them prove to be active it is then appropriate to assay the other molecules in its cluster since the Similar Property Principle implies that these are also expected to be active.

Clustering is probably the simplest type of selection procedure, but several other algorithmic approaches are under active investigation, with the selection criteria increasingly being based not just on structural diversity but also on other characteristics (such as cost, pharmacokinetic properties, and ease of synthesis) that are necessary for a molecule to be considered as a potential drug.

6 Conclusions

Chemical structures might seem to present novel problems of representation and searching for the designers of digital library systems, were it not for the fact that techniques for these purposes have been under study for some four decades. There is now a wide range of both public and in-house chemical databases, together with associated software systems that provide not just conventional search-and-retrieve functions but also database-processing applications that play a key role in the design of new types of bioactive molecules.

References

1. Schofield, H., Wiggins, G. and Willett, P. Recent Developments in Chemoinformatics Education. Drug Discovery Today, Vol. 6, pp. 931-934, 2001.
2. Leach, A.R. and Gillet V.J. An Introduction to Chemoinformatics. Kluwer, 2003.
3. Gasteiger, J. and Engel, T. (Editors). Chemoinformatics. Wiley-VCH, 2003.
4. Barnard, J.M. Substructure Searching Methods: Old and New. Journal of Chemical Information and Computer Sciences, Vol. 33, pp. 532-538, 1993.
5. Martin, Y.C. and Willett, P. (Editors). Designing Bioactive Molecules: Three-Dimensional Techniques and Applications. American Chemical Society, 1998.
6. Salton, G. Automatic Text Processing. Addison-Wesley, 1989.
7. Sparck Jones, K. and Willett, P. (Editors). Readings in Information Retrieval. Morgan Kaufmann, 1997.
8. Carhart, R.E., Smith, D.H. and Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. Journal of Chemical Information and Computer Sciences, Vol. 25, pp. 64-73, 1985.
9. Willett, P., Barnard, J.M. and Downs, G.M. Chemical Similarity Searching. Journal of Chemical Information and Computer Sciences, Vol. 38, pp. 983-996, 1998.
10. Johnson, M.A. and Maggiora, G.M. (Editors). Concepts and Applications of Molecular Similarity. Wiley, 1990.
11. Martin, Y.C., Kofron, J.L. and Traphagen, L.M. Do Structurally Similar Molecules have Similar Biological Activities? Journal of Medicinal Chemistry, Vol. 45, pp. 4350-4358, 2002.
12. Bohm, H.-J. and Schneider, G. (Editors). Virtual Screening for Bioactive Molecules. Wiley-VCH, 2000.
13. Willett, P. Textual and Chemical Information Retrieval: Different Applications but Similar Algorithms. Information Research, 5(2), 1999, at URL http://InformationR.net/ir/5-2/infres52.html
14. van Rijsbergen, C.J. Information Retrieval. Butterworth, 1979.
15. Ginn, C.M.R., Willett, P. and Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. Perspectives in Drug Discovery and Design, Vol. 20, pp. 1-16, 2000.
16. Cramer, R.D., Redl, G. and Berkoff, C.E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. Journal of Medicinal Chemistry, Vol. 17, pp. 533-535, 1974.
17. Wilton, D., Willett, P., Mullier, G. and Lawson, K. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programmes. Journal of Chemical Information and Computer Sciences, Vol. 43, pp. 469-474, 2003.
18. Robertson, S.E. and Sparck Jones, K. Relevance Weighting of Search Terms. Journal of the American Society for Information Science, Vol. 27, pp. 129-146, 1976.
19. Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. Proteins, Vol. 47, pp. 409-443, 2002.
20. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E.. A Geometric Approach to Macromolecule-Ligand Interactions. Journal of Molecular Biology, Vol. 161, pp. 269-288, 1982.
21. Ghose, A.K. & Viswanadhan, V.N. (Editors). Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery. Marcel Dekker, 2001.
22. Willett, P., Winterman, V. & Bawden, D. 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, Vol. 26, pp. 109-118, 1986.