Journal of Computer-Aided Molecular Design, 11 (1997) 447–452.
KLUWER/ESCOM

J-CAMD 408

Database diversity assessment: New ideas, concepts, and tools

Ramaswamy Nilakantan*, Norman Bauman** and Kevin S. Haraki

Wyeth-Ayerst Research, North Middletown Road, Pearl River, NY 10965, U.S.A.

Received 2 January 1997
Accepted 23 April 1997

Keywords: Similarity; Comparison; Database; Ring; Ring-cluster; Combinatorial

Summary

We present some new ideas for characterizing and comparing large chemical databases. The comparison of the contents of large databases is not trivial since it implies pairwise comparison of hundreds of thousands of compounds. We have developed methods for categorizing compounds into groups or series based on their ring-system content, using precalculated structure-based hashcodes. Two large databases can then be compared by simply comparing their hashcode tables. Furthermore, the number of distinct ring-system combinations can be used as an indicator of database diversity. We also present an independent technique for diversity assessment called the 'saturation diversity' approach. This method is based on picking as many mutually dissimilar compounds as possible from a database or a subset thereof. We show that both methods yield similar results. Since the two methods measure very different properties, this probably says more about the properties of the databases studied than about the methods.

Introduction

There is a continuous need in the pharmaceutical industry to generate novel chemicals. In pharmaceutical companies, novel chemical compounds are obtained using traditional synthesis, compound acquisition, and, most recently, by combinatorial chemistry. The latter two are ways of adding large numbers of compounds to a company's collection in a very short time. In order not to waste resources, it is necessary to assess the quality of the compounds being purchased or generated by combinatorial chemistry beforehand. One aspect of particular importance is the chemical and structural diversity of the new compounds and their relationship to those already present in the company's collection. In this paper, we present methods for assessing the diversity of large databases and to compare two or more databases in order to select compounds from one to enrich the other.

Excellent methods to search for compounds in two- and three-dimensional (2D and 3D) databases are available. Reviews of current 2D and 3D methods can be found in Refs. 1 and 2. In our own laboratory, we have developed methods for 2D similarity searching [3,4] as well as for 3D pharmacophore searching [5,6]. However, the area of studying database characteristics has only recently received attention [7–10].

Shemetulskis et al. [7] describe a method to compare two large databases and select a subset from one to augment the structural diversity of the other. For example, a pharmaceutical company might want to augment its corporate database (let us call it 'A') by buying compounds from another database 'B'. The method is as follows. First each database is clustered separately using the Jarvis–Patrick (J–P) method [11]. Then a sub-database is constructed by retaining one compound from each cluster representing the cluster centroid. Several such sub-databases can be produced by varying the J–P clustering parameters. These reduced sub-databases are used for further analysis. Mixed subsets are then prepared by combining compounds from the two databases A and B and are clustered using the J–P method. The number of clusters of pure or nearly pure B composition are an indication of the novelty of database B with respect to A. If there are none, then the database B does not have the potential to enrich A significantly. On the other hand, if there are several clusters of pure B compounds or nearly pure B compounds, then these could add structural variety to A. Thus, these pure or nearly pure clusters are

*To whom correspondence should be addressed.
**Present address: 18 Ridgetop Drive, Tomkins Cove, NY 10986, U.S.A.

selected and recommended for purchase to augment the corporate database.

This method, while quite straightforward, is computationally very expensive to implement. A clustering method such as the J–P method requires calculations of many nearest-neighbor lists. To calculate these lists, one needs to compare each compound in the database with every other, a task which can rapidly get out of hand for very large databases of millions of compounds. When comparing one database with several other databases as potential sources of acquiring new compounds, this method will prove extremely time consuming and cumbersome. Another paper from the same group [8] describes an algorithm 'Stigmata', which uses molecular fragment fingerprints to determine structural commonalties within sets of compounds. This approach, however, is not appropriate for large compound sets and is not intended to be a database comparison tool.

Boyd et al. [9] describe 'HookSpace', a program to assess the diversity of large databases. This program is based on analyzing the spatial relationships between functional groups in each molecule and representing the results graphically. It therefore works only on 3D databases, where it provides a way to study the diversity of functional group orientations. It does not, however, provide a way to compare two databases and pick out specific compounds from one to enrich the other.

Martin et al. [10] describe a method (closely related to Ref. 8) to compare databases by using fragment fingerprints. It involves calculating a 2048-bit fingerprint for each molecule in a database, and then logically 'OR'ing the fingerprints of all the molecules to get a fingerprint for the whole database. The fingerprints of two different databases can then be compared to get an idea of how different they are. This method can give an idea of how different two databases are, without actually identifying the compounds from one to enrich the other.

Cummins et al. [12] have published a method for comparing databases by characterizing the compounds by using properties such as topological indices, reducing the dimensionality of the descriptor space by factor analysis, and viewing the compounds graphically. Their results indicate that compounds in all databases are generally tightly clustered in some areas of the descriptor space, probably indicating that this level of description is not able to resolve fine differences among molecules. Their method does not provide a neat way to select novel compounds from one database to enrich another.

Another interesting piece of work by Pickett et al. [13] uses a partitioning of databases on the basis of the pharmacophores they contain. A large set of pharmacophores is used, and each database is characterized by the set of pharmacophores it contains. This method of course relies on 3D structures.

The above discussion shows that there is a need for methods to characterize compounds rapidly based on their structures and to use such a characterization to compare databases and select novel compounds from one database to enhance another. We present some ideas and methods to do this. We also present two methods for quantifying database diversity.

Methods and Results

In order to compare two databases, we first have to categorize the compounds within each of them. In order to keep the computation feasible, we have to avoid pairwise comparisons. With this in mind, we set out to devise simpler methods that can characterize the compounds in a database according to some carefully preselected structural criteria, i.e., to classify them into intuitively reasonable chemical series. In an earlier paper [14], we described a ring-system-based structural query system that could be used as a powerful adjunct to conventional searching tools. We presented methods for extracting, storing, organizing, and retrieving ring-systems in large databases. (Recently, Bemis and Murcko [15] have used similar ideas to characterize compounds in a database into ring-system, linker, and side chains.) Continuing the theme of our earlier paper, we realized that the *ring-cluster*, or the combination of ring-systems in a compound, is an extremely useful feature that can be exploited for rapid database characterization and comparison.

It is well known that most large organic chemical databases contain at most only about 10–25% of acyclic compounds [14]. Most marketed drugs possess one or two ring-systems. For example, a set of 618 marketed drugs from the World Drug Index (WDI) [16] that we experimented with had only 30 acyclic compounds. Therefore, if we are willing to put all the acyclic compounds into one structural class, we can classify the remaining compounds by their *content of ring-systems*, or, in our terminology, *ring-clusters*. Compounds of the same ring-cluster may have their various ring-systems interconnected in different ways, and the ring-systems themselves may be decorated with functional groups in different ways. Our studies indicate that in most pharmaceutical databases, compounds of the same ring-cluster often belong to the same synthetic series. Conversely, given a synthetic series, most members generally fall into one ring-cluster group. (Of course, if different ring-systems are made as part of a series, they would fall into different ring-cluster groups.) Actually, some of these ideas grew out of a project for retrospective identification of synthetic series in our corporate database. It should be realized, however, that no method can correctly identify all the synthetic series; indeed, the term 'synthetic series' has no formal structural definition.

Definitions

Ring-system A set of atoms connected together into one or more 'cycles' is a ring-system. In graph-theoretical

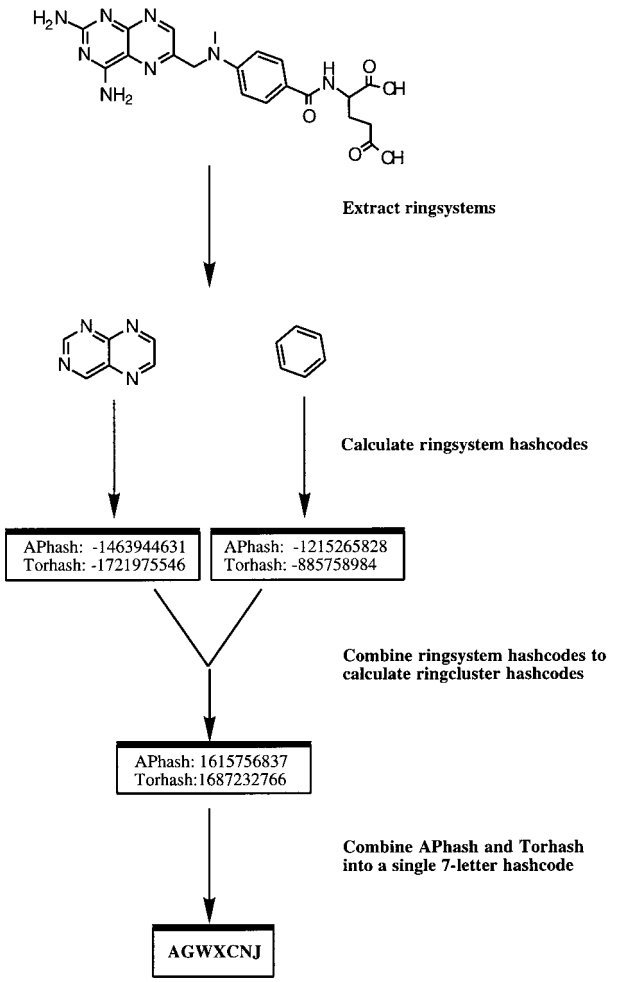

Fig. 1. Derivation of hashcodes for methotrexate: a pair of hashcodes is calculated for each ring-system, one based on the atom-pair descriptor and the other on the topological-torsion descriptor. To calculate ring-cluster hashcodes, the hashcodes of all the different ring-systems are added together, thus obtaining a pair of hashcodes (atom-pair and topological-torsion) for the ring-cluster. This pair of hashcodes is then turned into a single seven-letter code by exclusive 'OR'ing them and converting the resulting unsigned 32-bit integer into a base-26 number, with the digits represented by the letters of the alphabet.

terms, each connected component remaining after deletion of all the acyclic edges of the molecular graph is called a ring-system. Ring-systems are perceived by using a spanning-tree representation of the connection table as discussed in Ref. 14.

Ring-cluster The set of ring-systems in a molecule is collectively called a ring-cluster. Note that the ring-systems in a cluster need not be distinct; thus, compounds with one pyridine ring and one benzene ring fall in a different cluster from compounds with one pyridine ring and two benzene rings.

Use of hashcodes to encode ring-systems

In our earlier paper [14], we discussed the use of a pair of hashcodes based on the *atom-pair* and *topological-torsion* descriptors [3,4]. To recapitulate, each ring-system is resolved into its atom-pair (or topological-torsion) descriptor set. Each descriptor is encoded into a single 32-bit integer. Hashcodes are constructed for atom-pair descriptors by calculating the sum of the products of each descriptor right-shifted by 2 bits with the same descriptor right-shifted by 13 bits, ignoring overflows. A similar calculation is made for the topological-torsion descriptors, using the product of the unshifted descriptor with the same descriptor right-shifted by 16 bits. The descriptor hashcodes are summed, ignoring overflows, to create ring-system hashcodes. These hashcodes depend only on the covalent structure of the ring-system and do not distinguish among stereoisomers and geometric isomers. The two hashcodes constructed for each ring-system, taken together, were found to encode each ring-system uniquely [14], within isomerism, although in principle there could be accidental collisions. This method is very fast because it is independent of the numbering of the atoms, thus avoiding canonicalization, and because it uses only integer arithmetic.

Hashcodes to encode ring-clusters

In order to encode ring-clusters, we simply add together the hashcodes of all the constituent ring-systems, again ignoring overflows. Thus, we derive a pair of 32-bit hashcodes (atom-pair and topological-torsion) for each ring-cluster. For further compression, the two numeric hashcodes are exclusive 'OR'ed, and the result (taken as an unsigned 32-bit value) is converted to a base-26 number. The base-26 digits 0–25 are expressed as A–Z to form a seven-letter hashcode. Thus, each ring-cluster in the database (and each compound) is characterized by a single seven-letter hashcode. Although some resolution is lost when the two numeric hashcodes are combined, collisions are still very rare, well under 0.1%. This makes the comparison of large sets of compounds very easy and convenient. Figure 1 illustrates the procedure used to calculate hashcodes taking methotrexate as an example.

Use of ring-cluster hashcodes to characterize databases

The identification of ring-systems and ring-clusters is fast, taking about 3 h on a VAX 4000-700 for a database of about a million compounds. Once hashcodes are computed for each compound's ring-cluster, the compound ID numbers and hashcodes are loaded into ORACLE [17] tables for further analysis. Using these tables, it is possible to group compounds into chemical series as indicated earlier. Compounds having the same set of ring-systems (same ring-cluster) are often part of the same synthetic series. Thus, with a single database query, it is possible to group compounds into series. We can also study the distribution of compounds having various ring-clusters.

Use of ring-cluster hashcodes to compare databases

Perhaps more interestingly, it is possible to use these hashcodes to compare two or more databases. One could

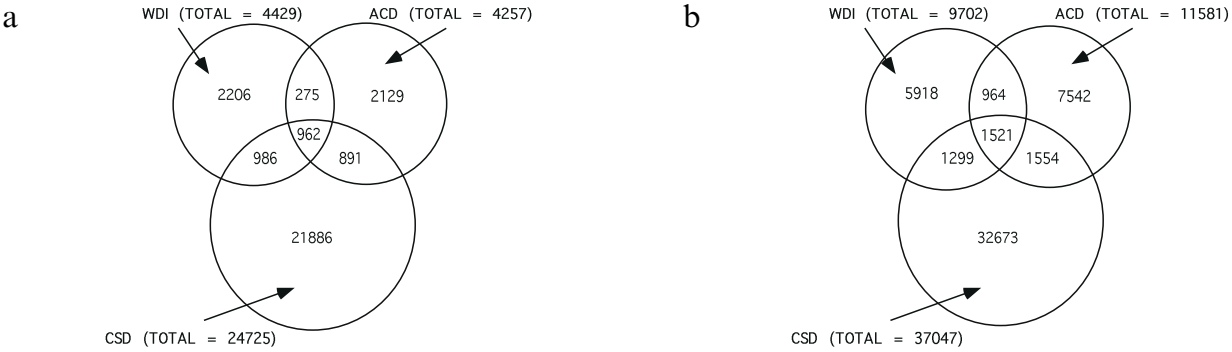

Fig. 2. Venn diagrams showing the (a) ring-system and (b) ring-cluster data for WDI, ACD, and CSD databases.

find those series that are not present in one's own corporate database, but are found in a commercially available database. Thus, ring-cluster hashcodes can be an extremely fast means of rationally identifying compounds to purchase, taking only as long as it takes to execute a few ORACLE [17] queries. (Of course, for each new database, the ring-system and ring-cluster tables would have to be constructed once. This, as mentioned earlier, takes about 3 h on a VAX 4000-700 for a database of a million compounds. But for the smaller databases offered for sale, the calculation is usually a matter of minutes.) This technique is perhaps the only one available for identification of common synthetic series in two or more databases.

In combinatorial chemistry synthesis planning, it is often desirable to maximize the diversity of the compounds synthesized. Using the computer, it is possible to construct several alternative virtual libraries and compare them using the ring-cluster method in order to help decide which ones to actually make.

Our method is somewhat related to that of Martin et al. [10] in that both use a set of descriptors to characterize databases. It is also related to that of Pickett et al. [13] in that both methods partition databases on the basis of a set of structural features. The Pickett method uses 3D structures and a set of pharmacophores to characterize and compare databases. We expect these techniques will complement each other since they look at different aspects of chemical structure. Our method is unique in the use of ring-system combinations to (i) characterize databases and (ii) select compounds from one database to enrich another.

Calculation of ring-clusters for four public-domain databases

In order to illustrate the technique, we applied it to four well-known public-domain databases, viz., World Drug Index (WDI) [16], Available Chemicals Directory (ACD) [18], the public-domain portion of the NCI database (NCI3D) [19], and the MACCS [20] version of the Cambridge Structural Database (CSD) [21]. The results are shown in the first four columns of Table 1. Note that the last two columns of Table 1 refer to database diversity measures and will be discussed later in the paper. The ring-system and ring-cluster characteristics of these four databases are shown separately in the table, illustrating the utility of the method in comparing large collections of compounds for variety and structural diversity. It can be seen that the CSD has a much higher diversity of ring-clusters. This is to be expected since the CSD is a collection of compounds from various laboratories all over the world whose crystal structures are available. Large synthetic series are not usually found in the CSD. On the other hand, the ACD, NCI3D, and WDI (candidate drug molecules) contain many structural series. It can be seen that ACD and NCI3D have essentially the same richness of ring-clusters. Table 1 also shows that while there are a slightly greater number of distinct ring-systems in WDI than in ACD, the latter contains a larger number of distinct ring-clusters. In other words, the ACD contains compounds with more combinations of ring-systems than the WDI. The ACD and NCI3D have approximately the same normalized number of ring-clusters.

Figures 2a and b show a comparison of three of the four (WDI, ACD, and CSD) databases using Venn diagrams. Comparisons involving four or more databases are equally easy to do, but difficult to illustrate. Figure 2a shows the ring-system data and Fig. 2b shows ring-cluster data. From these diagrams, we can see the number of ring-systems and ring-clusters unique to each of the three databases. Similarly, the number of ring-systems found in

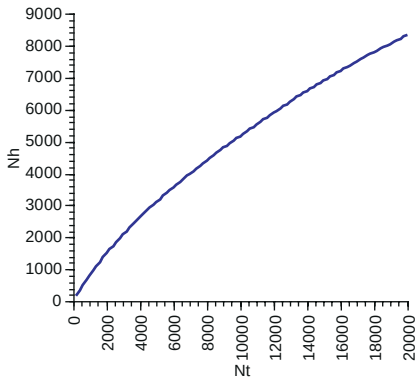

Fig. 3. Plot of the number of hits versus the number of tries in running the DIVPIK program. This program picks a subset of compounds from a database such that no two of them have a similarity above a user-specified threshold.

TABLE 1
NUMBER OF DISTINCT RING-SYSTEMS, RING-CLUSTERS AND DIVERSITY INDEXES FOR WDI, ACD, CSD, AND NCI3D DATABASES

Name of the database	Number of compounds in the database	Number of distinct ring-systems	Number of distinct ring-clusters	Number of distinct ring-clusters (normalized)	Diversity index D
WDI	52 695	4 429	9 702	2055	0.26
ACD	170 509	4 257	11 581	2214	0.35
NCI3D	126 554	5 905	11 611	2275	0.35
CSD	99 354	24 725	37 047	7414	0.46

two of the databases but not in the third can be obtained directly from the diagram. Pairwise comparisons of databases can be easily derived from the diagrams. For example, the number of ring-systems found in WDI but not in CSD can be obtained from the diagram as 2206 + 275 = 2481. Finally, the figure shows the number of ring-systems and ring-clusters present in all three databases. This type of data enables one to get a qualitative flavor of different databases and also helps in compound acquisition programs. For example, it is clear that the CSD database has the richest collection of ring-systems and ring-clusters. We can also see that while there is significant overlap between the databases, each database has a considerable number of unique ring-systems and ring-clusters.

Estimating database diversity

Ring-cluster-based approach The ring-cluster method can also be used to derive an indicator of the diversity or structural richness of a database. The number of distinct ring-clusters is a numerical indicator of database diversity. In order to normalize this number by database size, we divide the raw number of ring-clusters by the logarithm of the number of compounds in the database. (This is an admittedly intuitive normalization, justified only by the well-known theorem that the incidence of random record-breaking is asymptotic to the natural logarithm of the number of tries [22].) This normalized ring-cluster count can be calculated for various databases and can be used to get an idea of their relative structural richness. Table 1 (column 5) shows the numbers for WDI, ACD, NCI3D, and CSD databases.

The saturation diversity approach The ring-cluster-based approach to estimating database diversity has obvious disadvantages. As mentioned before, it deals only with ring-compounds, and is most useful with multi-ring compounds. It is not applicable to acyclic compounds. Moreover, it does not recognize the ways in which ring-systems are interconnected. There is therefore a need for an independent method at least to validate our results. We have developed a method to do this which we discuss below. In an earlier publication from our laboratory [3], SIMPRB, a program which picks out compounds similar in a 2D sense to a given probe compound, was discussed. A sister program DIVPIK, which is not discussed in that paper, was also developed in our laboratory at that time. It is a variation of SIMPRB, modified to pick a set of mutually dissimilar compounds. DIVPIK picks successive compounds at random from the database (or subset thereof), comparing each newly picked compound with those already selected to check that its similarity to any one of them is no more than a user-specified amount. (Other diversity-based compound-picking programs have since been published, including a very efficient one by Holliday et al. [23].) As the DIVPIK program runs, it is possible to collect data on the number of tries it makes and the number of successes it has in picking compounds. It is thus possible to plot a graph of the number of tries versus the number of successes. Such a graph is shown in Fig. 3. (It should be noted that the algorithm is order-dependent. However, since we are dealing with large numbers of trials on large databases, the resulting statistics do not differ much from trial to trial.) As we start picking compounds from the database, the number of successful picks depends on the number of compounds remaining in the database. Therefore, initially we have a large number of successful picks, but as the number of remaining compounds begins to diminish, the number of successful picks also diminishes, finally reducing to zero. In order to derive a diversity estimate for a database, we proceed as follows. First we choose a suitably large random subset of N compounds from the database (we have used N = 30 000 for each of the databases we studied). Then we run the DIVPIK program using a similarity cutoff of 0.6 on this set until no more compounds can be picked. The maximum number picked, H_{max}, as a proportion of the total number of compounds gives us an indication of the diversity of the database. Thus, we have

$$D = H_{max}/N$$

where D is the *diversity index*.

We carried out these calculations on a series of four databases: WDI, ACD, NCI3D, and CSD. For the CSD database, the MACCS version supplied by the Cambridge Crystallographic Data Centre was used. For each database, the calculations were carried out on two randomly chosen 30 000-compound subsets and the results (which were within 1% of each other) were averaged. The diversity indexes for the four databases are shown in Table 1, column 6. Comparing the normalized ring-cluster counts (one diversity measure) with the corresponding D values

(another diversity measure), we see that they follow the same trend. The relative diversities of the four databases as measured by the ring-cluster count are in the order WDI < ACD ≈ NCI3D < CSD. When measured using the saturation diversity approach, we obtain essentially the same order: WDI < ACD = NCI3D < CSD. It should be pointed out that the saturation diversity method is slow when compared to the ring-system method, but serves as an independent check on the validity of the approach. Of course, the saturation diversity approach cannot be used for selective compound acquisition.

2D versus 3D

The method we have described is entirely 2D, using only connection tables. There are 3D approaches available such as the method described in Ref. 9. These methods seek common geometric features among compounds in a database and use these to group them into clusters. The use of 3D features in comparing databases has several drawbacks. One problem is that of conformational flexibility: since compounds are flexible, any one conformation or even a finite number of low-energy conformations does not really cover all the structural possibilities. Also, in unpublished work, we found that purely geometric features such as distances are not generally differentially distributed in large databases, i.e., different databases have the same profile. In particular, we found that a set of 644 marketed drugs showed the same profile as the large databases. In small sets of compounds however, geometric characteristics may be useful in segregating compounds into sets sharing common features or pharmacophores.

Comparison with clustering approaches

As described in the Introduction section, databases can also be compared using standard clustering approaches such as in Ref. 7. However, such approaches have several drawbacks. First, all clustering methods are sensitive to the choice of clustering parameters. Thus, while clustering may be suitable as an aid to browsing, its use in comparing databases is limited. Also, when a database grows in size, the clustering will have to be done all over again, and might disturb old cluster memberships. Finally, clustering is computationally expensive, requiring pairwise comparison of all the compounds under study.

Conclusions

We have presented a method for the rapid structural characterization of chemical databases using ring-clusters. The method makes it possible to compare databases to get a qualitative idea of their content and to help select compounds for acquisition. Furthermore, normalized ring-cluster counts can be used as a measure of database diversity. We have also presented an independent method to estimate database diversity, which uses results from our program to pick a diverse set of structures from a database. We show that the results from the two methods for diversity measurement are consistent and validate each other.

Acknowledgements

We would like to thank Drs. Dominick Mobilio and Gary Walker for reading the manuscript carefully and making useful suggestions.

References

1 Barnard, J.M., J. Chem. Inf. Comput. Sci., 33 (1993) 532.
2 Downs, G.M. and Willett, P., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. 7, VCH, New York, NY, U.S.A., 1996, pp. 1–66.
3 Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
4 Nilakantan, R., Bauman, N., Dixon, J.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 27 (1987) 82.
5 Sheridan, R.P., Rusinko III, A., Nilakantan, R. and Venkataraghavan, R., Proc. Natl. Acad. Sci. USA, 86 (1989) 8165.
6 Sheridan, R.P., Nilakantan, R., Rusinko III, A., Bauman, N., Haraki, K.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 29 (1989) 255.
7 Shemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., Moreland, D.W. and Humblet, C., J. Comput.-Aided Mol. Design, 9 (1995) 407.
8 Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J. and Humblet, C., J. Chem. Inf. Comput. Sci., 36 (1996) 862.
9 Boyd, S.M., Beverley, M., Norskov, L. and Hubbard, R.E., J. Comput.-Aided Mol. Design, 9 (1995) 417.
10 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
11 Jarvis, R.A. and Patrick, E.A., IEEE Trans. Comput., C22 (1973) 1025.
12 Cummins, D.J., Andrews, C.W., Bentley, J.A. and Cory, M., J. Chem. Inf. Comput. Sci., 36 (1996) 750.
13 Pickett, S.D., Mason, J.S. and McLay, I.M., J. Chem. Inf. Comput. Sci., 36 (1996) 1214.
14 Nilakantan, R., Bauman, N., Haraki, K.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 30 (1990) 65.
15 Bemis, G.W. and Murcko, M.A., J. Med. Chem., 39 (1996) 2887.
16 World Drug Index (WDI), developed and published by Derwent Publications, London, U.K.
17 ORACLE, a database management system distributed by Oracle Corporation.
18 Available Chemicals Directory (ACD), a database of commercially available compounds distributed by MDL Information Systems, San Leandro, CA, U.S.A.
19 NCI3D, the public-domain portion of the National Cancer Institute's database distributed by MDL Information Systems, San Leandro, CA, U.S.A.
20 MACCS, an acronym for Molecular Access System, a chemical database management system supplied by MDL Information Systems, San Leandro, CA, U.S.A.
21 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
22 Durrett, R., In Probability, Theory and Examples, Wadsworth, Belmont, CA, U.S.A., 1991, pp. 45–46.
23 Holliday, J.H., Ranade, S.H. and Willett, P., Quant. Struct.–Act. Relatsh., 14 (1995) 501.