Published paper

Willett, P. (2011) *Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance*. Wiley Interdisciplinary Reviews - Data Mining and Knowledge Discovery, 1 (3). 241 - 251. ISSN 1942-4787

http://dx.doi.org/10.1002/widm.26

Similarity-Based Data Mining in Files of Two-Dimensional Chemical Structures using Fingerprint Measures of Molecular Resemblance

Abstract

This paper reviews the use of measures of inter-molecular similarity for processing databases of chemical structures, which play an important role in the discovery of new drugs by the pharmaceutical industry. The similarity measures considered here are based on the use of a fingerprint representation of molecular structure, where a fingerprint is a vector encoding the presence of fragment substructures in a molecule and where the similarity between pairs of such fingerprints is computed using an association coefficient such as the Tanimoto coefficient. The Similar Property Principle provides the basic rationale for the use of similarity methods in three important chemoinformatics applications: similarity searching, database clustering, and molecular diversity analysis. Similarity searching enables the identification of those molecules in a database that are most similar to a user-defined, biologically active query molecule, with data fusion providing an effective way of combining the results of multiple similarity searches. Cluster analysis, typically using the Jarvis-Patrick, Ward or divisive k-means clustering methods, enables the cost-effective selection of molecules for biological testing, for property prediction and for investigating database overlap. Molecular diversity analysis, typically using cluster-based, dissimilarity-based or optimisation-based approaches, enables the identification of structurally diverse sets of molecules, so as to ensure that the full chemical space spanned by a database is tested in the search for novel bioactive molecules.

Keywords: Association coefficient; Chemical database; Chemoinformatics; Cluster analysis; Cluster-based selection; Data fusion; Dissimilarity measure; Dissimilarity-based selection; Diversity; Divisive k-means clustering method; Drug discovery; Fingerprint; Fragment substructure; Jarvis-Patrick clustering method; Molecular diversity analysis; Pharmaceutical research; Similar Property Principle; Similarity measure; Similarity searching; Subset selection; Tanimoto coefficient; Ward's clustering method

INTRODUCTION

Research and development in the fine chemicals industry is driven by the need to discover novel molecules with useful physical, chemical or biological properties, e.g., lowering a person's cholesterol level in the pharmaceutical industry or having a pleasant aroma in the personal products industry. The structures, whether in two dimensions (2D) or three dimensions (3D), of chemical molecules hence form an extremely important component of a company's intellectual property and there has been interest for many years in computer techniques for processing databases of chemical structures [1, 2].

Chemical databases are used extensively in both the public and the private sectors. The longest-established public database system is the CAS Registry from Chemical Abstracts Service (at http://www.cas.org/), which contains all molecules that have been reported in the open chemical literature. Other important public databases include ChemSpider (at http://www.chemspider.com/), which brings together compound information from ca. 400 Web sources, and PubChem (at http://pubchem.ncbi.nlm.nih.gov/), which contains compound information contributed from governmental and academic sources with associated bioactivity data in many cases. Private, corporate databases play a key role in industrial research and development, and contain the records of all of the molecules that a specific organization has studied. Chemical databases can be very large: The Registry System now (late 2010) contains ca. 55 million molecules and is growing by ca. 12 thousand molecules a day; while the corporate database of a major pharmaceutical company may contain several million molecules, many of which have never been made public. The importance of such databases has driven the development of a specialist discipline, *chemoinformatics*, in just the same way as the increasing volumes of biological sequence data has driven the development of bioinformatics.

Chemoinformatics encompasses the interpretation of molecular spectra, the design of complex organic syntheses, the prediction of biological activity, and the analysis of the interactions between drugs and biological macromolecules such as proteins, *inter alia*, with these data being mined using statistical, graph theoretic, machine learning and evolutionary techniques [3, 4]. In this contribution, we review one such topic, the ways in which the concept of molecular similarity can be used to support drug discovery. Analogous techniques are used for the discovery of novel agrochemicals and other types of fine chemical, but the focus is on drug discovery since it is here that chemoinformatics has had its greatest impact; indeed, many of the techniques in current use have been developed by industrial, rather than academic, research groups.

The paper is structured as follows. The next section introduces the concept of molecular similarity and its quantification, and briefly describes how chemical structures can be encoded in machine-readable form. We then discuss the use of one type of representation (the chemical fingerprint, *vide infra*) for the three principal applications of molecular similarity, these being similarity searching, database clustering and molecular diversity analysis. Finally, we compare fingerprint-based approaches with more complex ways of processing molecular similarity data. Further details of the theory and practice of molecular similarity are available from the extensive literature that is available [5-9].

COMPUTING MOLECULAR SIMILARITIES

The Similar Property Principle

The identification of a novel molecule with a desired bioactivity, often referred to as *lead discovery*, is that stage of a drug programme where chemoinformatics makes its main contribution; it also contributes to the subsequent, *lead optimization* stage where the lead compound is systematically modified to obtain the best combination of activity, specificity, pharmacology etc. One of the ways in which chemoinformatics supports lead discovery is by drawing on what is commonly referred to as the Similar Property Principle. The Principle states that molecules that have similar structures will have similar properties; thus, if a molecule is known to exhibit the activity of interest, e.g., an existing drug for a disease of interest, it may be possible to identify potential bioactive substances by considering molecules that are structurally similar to the known compound. In 1990, Johnson and Maggiora edited the first book to deal with molecular similarity [10], and this is often cited as the source of the Principle; however, it had been discussed a decade before by Wilkins and Randic [11] and had almost certainly been known, albeit not as a formal principle, for many years prior to then. After all, medicinal chemistry has always made extensive use of analogy because if some relationship between the structures of molecules and their biological activities did not exist then drug discovery would be effectively a random process. The continuing success of the pharmaceutical industry over many years would suggest that this is not the case, and further evidence of the general validity of the Principle comes from the many experimental studies that have been carried out. The first detailed study of this type was by Willett and Winterman [12], who found that computed molecular similarities could be used to predict a range of physical, chemical and biological properties in a range of small datasets for which both structural and property information were available. There have been many subsequent examples of this approach to the evaluation of similarity procedures [13-17], and further supporting evidence for the general applicability of the Principle comes from studies in chemogenomics [18-21]. It must be emphasized that there are many exceptions to the Principle [22, 23], but it has been found to provide a very useful basis for the development of a range of similarity-based approaches for the processing of large chemical databases.

The degree of resemblance between two molecules is computed using a similarity measure, which has three components. First is the representation used to describe the two molecules that are to be compared, i.e., the manner in which the molecules are encoded for machine processing. Second, the weighting scheme that is used to prioritize (or de-prioritize) the contributions of different parts of the representation; related to weighting schemes are standardization schemes that are used to ensure that all parts of a representation contribute equally. Third, the similarity coefficient that computes the degree of resemblance between the molecules' representations. These three components are discussed further below.

Structure representations

Molecules are most commonly represented in the published literature by their names and/or by images of their 2D structure diagrams. Although familiar to the chemist, these are not suitable for detailed machine processing; instead, molecules are normally represented in chemical databases by *connection tables*, graphs in which the atoms and bonds of a molecule are denoted by the nodes and edges of a graph [4].

There are many different types of connection table, but they all provide an exact and explicit description of a molecule's topology that can be processed using the various types of isomorphism algorithm that permit the identification of areas of structural commonality in pairs of graphs [4]. In particular, a maximum common subgraph isomorphism provides a natural measure of molecular similarity since it identifies the largest overlap (in terms of atoms and bonds) when two chemical graphs are compared [24]. Isomorphism algorithms are effective in operation but are highly inefficient, requiring numbers of node-to-node comparisons that are factorial functions of the numbers of nodes in the graphs that are being compared. Whilst considerable effort has been devoted to maximising the efficiency of chemical graph matching [25] much use is made of simpler molecular representations that do not contain a complete description of molecular topology. Two types of simpler representation are of importance: *reduced graphs* and *fingerprints*. In a reduced graph, sets of individual atoms that are bonded together are merged into larger, reduced graph nodes, e.g., the six carbon atoms comprising a phenyl ring may be merged into a single node of type 'Ring' [26, 27]. Matching operations can then be carried out on these more compact molecular encodings, with substantial increases in efficiency; the new merged nodes are often designed to encode functionally important parts of molecules that are known to interact with proteins (and hence to exhibit some particular type of bioactivity) and searches using reduced graphs may hence also be more effective than when the full set of nodes is used.

In a fingerprint, a molecule is indexed by some number of chemical *fragments*. These fragments are typically small substructures that are generated automatically from a connection table so that a molecule might, e.g., be encoded using the fragments describing a phenyl group, a nitro group, and a carboxylic acid group (it must be emphasised that these example fragments are designed simply to illustrate the concept and a large body of work has gone into the design of atom-, bond- and ring-centred substructures for a whole range of chemoinformatics applications [28, 29]). The use of a set of fragments to characterise a molecule means that the contents of the molecule are indexed but not the precise way that these are linked together, whereas a connection table records the full topology of a molecule. In similar vein, a journal article might be indexed by a set of keywords and phrases, with the full text being required to understand the precise relationships between these textual elements. However, the fact that most text search engines employ this so-called 'bag of words' model very successful without recourse to sophisticated natural language processing suggests that an analogous simplified representation may be equally effective in the chemoinformatics context, as is clearly demonstrated in the remainder of this contribution.

The fragments generated for a molecule can be encoded in a bit-string, a binary vector in which bits are switched 'on' or 'off' depending whether particular fragments are present or absent in that molecule; there are several different ways in which fingerprints can be generated, and the reader is referred to the standard texts for details of fingerprinting procedures [3, 4]. There have been many comparisons of fingerprints, with the evidence to date suggesting that the most generally effective are based on circular substructures. These encode the immediate environment of each individual atom in a molecule, with the environment being defined as all of the atoms within some fixed number of bonds of the chosen, central atom. Such approaches have been known for many years, both for chemical substructure searching (*vide infra*) and for the analysis of spectroscopic data [30]. Early examples of circular substructures are described by Bremser [30], Attias [31] and Willett [32] *inter alia* with the Atom Environment fragments described by Bender *et al.* [33] a morerecent example. The most widely used circular substructures are those encoded in the Extended Connectivity Fingerprints and the Functional Connectivity Fingerprints included in the Pipeline Pilot software (from Accelrys Inc. at http://www.accelrys.com); these have been shown to be effective in a number of comparative studies of fingerprinting methods [34]. Fingerprints provide the basis for much of the processing that is carried out to support similarity applications in drug discovery, and hence form the focus of this review.

Weighting schemes and similarity coefficients

Having introduced fingerprints as the most common type of representation, we now turn to the other two components of a similarity measure: the weighting scheme and the similarity coefficient. There have been only a few studies of the use of weighting schemes for fingerprint-based similarity measures: the most detailed are two recent ones [35, 36], which show that encoding how frequently a fragment occurs within a molecule can give better results in some circumstances than encoding just its presence or absence (as is the case when conventional bit-strings are used). That said, the many successful applications of binary fingerprints over the years (as discussed below) suggest that they provide an appropriate representation in many cases.

Many different types of coefficient are available for computing the degree of resemblance between pairs of objects [8, 37], with the class of *association coefficients* having found most application in chemoinformatics. Association coefficients were originally developed to compare binary vectors and they are thus very well suited to the calculation of fingerprint-based similarities. An early study [12] suggested the use of the Tanimoto coefficient for molecular similarity studies, and it rapidly became the coefficient of choice from amongst the more than 20 association coefficients that have been published in the literature [38]. Indeed, unless stated otherwise, references in the chemoinformatics literature to molecular similarity will normally involve the use of fingerprint-based Tanimoto calculations. Given two molecules, *A* and *B*, having *a* and *b* bits switched 'on' in their fingerprints, and with *c* of these bits being in common (i.e., having *c* fragment substructures in common), then the Tanimoto similarity between *A* and *B* is given by

$$S_{A,B} = \frac{c}{a+b-c} \tag{1}$$

It will be seen that the coefficient takes values between zero and unity, these lower- and upper-bound values corresponding to having no bits in common and to having identical fingerprints, respectively. In the form shown in (1), the Tanimoto coefficient [39] is identical to the Jaccard coefficient [40]. However, the Tanimoto coefficient, and some other association coefficients, can be extended to encompass non-binary data, e.g., if a fingerprint encodes not just fragment incidences but the frequencies of occurrence. In this case, the coefficient is given by

$$S_{A,B} = \frac{\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sum_{j=1}^{j=n} \left(x_{jA}\right)^2 + \sum_{j=1}^{j=n} \left(x_{jB}\right)^2 - \sum_{j=1}^{j=n} x_{jA} x_{jB}} \tag{2}$$

where x_{jA} denotes the number of times that the *j*-th fragment occurs in *A* (and correspondingly for *B*) and where the summations are over the *n* elements of each fingerprint. It will be realized that this reduces to the simpler form if all the elements x_{jA} are zero or unity. The upper-bound for this formulation is again unity; the lower-bound is -1/3, or again zero if only positive values are allowed for the elements x_{jA} and x_{jB}.

Inspection of (2) will reveal that the numerator is a simple vector dot product, and many different similarity coefficients can be obtained merely by changing the precise form of the normalization that is applied to this product [37, 41, 42]. This being so, one may well ask why the Tanimoto coefficient is by far the most widely used coefficient in chemoinformatics applications of similarity. Arguably the most important reason is that its use was highlighted at a very early stage in the development of the subject. Specifically, studies in the mid-Eighties by Willett *et al.* showed that it gave results that were superior to the cosine coefficient for purposes of similarity searching (*vide infra*) [43] and that it was superior to the cosine and correlation coefficients and the average, Canberra and Euclidean distances for the prediction of molecular properties [12]. It was hence adopted for use by many subsequent workers in the field. The Tanimoto coefficient is not without its limitations, as noted by Flower [44] and by Fligner *et al.* [45]; however, extended comparative studies have shown that it performs at least as well as any of the other coefficients that have been suggested to date [38, 46].

Having described how molecular similarities can be calculated, the next three sections describe the main ways in which similarity concepts are currently used for processing chemical databases.

SIMILARITY SEARCHING

Early systems for querying chemical databases focussed on *substructure searching*, i.e., the retrieval of information about all molecules containing a specific query substructure [4]. An example of this would be a search for antibiotic substances, where a possible starting point would be to retrieve all of those molecules that contain the highly specific ring systems that characterise penicillin and cephalosporin antibiotics. Substructure searching is a powerful technique but one that requires that the searcher already knows the types of molecule that are of interest for the bioactivity of interest. This information is, of course, not generally available in the early stages of a discovery programme where less specific, more browsing-like searches need to be carried out: this can be effected by means of *similarity searching* [43, 47].

A similarity search assumes that a molecule is available, referred to variously as the *target structure* or the *reference structure*, that is of interest to the searcher, typically because it exhibits (or is expected to exhibit) the biological activity of interest, e.g., an existing drug molecule produced by a competitor company. When presented to a database, a similarity search ranks the database in order of decreasing similarity with the reference structure, so as to identify the nearest neighbours, i.e., those most similar to the reference structure. If the Similar Property Principle holds for the similarity measure that is being used, then the nearest neighbours are the molecules that have the greatest likelihood of exhibiting the reference structure's bioactivity and are hence prime candidates for biological testing if samples of these molecules are available to hand. If this is not the case, e.g., if the database being searched is a publicly available one rather than an in-house database, then these molecules will need to be synthesised or purchased; indeed, one of the main applications of similarity searching is for scanning suppliers' catalogues that contain millions of molecules that are available for purchase.

The first reports of similarity searching came from two pharmaceutical companies – Pfizer [43] and Lederle Laboratories [47] – and were at one in suggesting the use of fingerprint representations and an association coefficient as a simple but effective way of ordering a database in response to an input molecule of interest. The approach was rapidly and widely adopted and is now a standard feature in chemoinformatics software systems. The precise way in which it is implemented does, of course, vary from system to system, and there is still much discussion as to how to maximise the effectiveness of searching [7, 9, 48]. Two important areas of current research are *data fusion* and *scaffold hopping*, as discussed further below.

There have been many comparative studies of similarity measures over the years, but it has not proved possible to identify a single combination of fingerprint, weighting scheme and similarity coefficient that will give a consistently high level of performance across all the many different types of search that may be required by a medicinal chemist [49]. There has hence been much interest in data fusion, *viz* the idea of combining (or fusing) the rankings resulting from multiple similarity searches to give a new ranking that is expected to maximise the clustering of actives at the top of the fused ranking [50]. Two main approaches have been described, often referred to as *similarity fusion* and *group fusion*. In the more common similarity fusion approach, a single reference structure is searched against a database using multiple similarity measures, e.g., using several different types of fingerprint or of similarity coefficient. In group fusion, alternatively, multiple reference structures are searched against a database using a single similarity measure. The latter approach obviously requires the availability of multiple bioactive molecules, rather than just a single one as in conventional similarity searching, but has been found to be notably more effective in operation [6, 48, 51]. Current areas of research include, e.g., how different rankings should be combined [52] and whether it is possible to provide a theoretical model of the fusion process [53].

The nearest neighbours retrieved in a similarity search often have the same central ring system (or *scaffold*), as the reference structure, meaning that these neighbours may not be novel in so far as they may well be covered by an existing patent. There has hence been considerable discussion as to whether fingerprint-based similarity searching is appropriate for *scaffold hopping*, i.e., the identification of bioactive molecules with novel scaffolds [54], since one might reasonably expect that more sophisticated similarity measures involving 3D structural information would be necessary for this purpose. It is hence of interest to note that two of the leading groups in the area of molecular similarity, led by Sheridan and by Bajorath, have recently reported studies in which fingerprint-based similarity measures do exhibit a fair level of scaffold hopping ability [6, 55].

Similarity searching is the earliest, but still the most widely used, example of what is commonly referred to as *virtual screening*, i.e., the ranking of a database so that synthesis and biological testing activities can be focussed on those molecules that have the greatest probabilities of activity [56-60]. A simple association coefficient such as the Tanimoto coefficient clearly does not compute a probability *per se*, although there are more sophisticated types of similarity measure that do attempt this [61, 62]. None the less, the approach has been found to provide a database access mechanism that facilitates the identification of novel classes of compounds that would not be obtained from the more sophisticated types of virtual screening method that are available (*viz* pharmacophore analysis [63], machine learning [64] or ligand-protein docking [65]).

CLUSTER ANALYSIS

Cluster analysis involves grouping a set of objects (the molecules comprising a chemical database in the present context) into smaller groups, or clusters, in which the members of each cluster are similar to each other but dissimilar to the members of other clusters [66, 67]. Cluster analysis can hence be regarded as a natural extension of similarity searching: the latter identifies the nearest neighbours that are most similar to an input reference structure, while cluster analysis identifies groups of molecules that are highly similar (and many of which will in fact be nearest neighbours). The similarity measures that have been discussed previously are hence equally applicable to the clustering of chemical databases, with a clustering method processing the computed inter-molecular similarities to identify the groups that are present.

The principal application of cluster analysis in chemoinformatics has been to select molecules for biological testing. Despite substantial technological advances over the last few years, the testing of large numbers of molecules for bioactivity remains both time-consuming and expensive, and there is hence a need for methods that will ensure coverage of as wide a range of types of molecule as possible whilst minimising the costs of testing. If a database has been clustered then a cost-effective approach is to select one molecule from each of the clusters, with the selected molecule (which is often referred to as the *cluster representative*) in each case normally being that closest to the centre of the cluster. Only this selected subset of the database then undergoes biological testing. If a cluster representative proves to be bioactive then it will be appropriate to test the other molecules in that cluster; alternatively, the cluster can be removed from further consideration. Given an effective clustering method, this systematic, approach should ensure full coverage of all of the various structural types present in the database that is being studied [68]. Alternative approaches to the selection of database subsets are discussed further in the following section on Molecular Diversity Analysis.

Cluster-based selection of database subsets was first described over a quarter of a century ago [69]. It continues to be widely used, but is by no means the only application of cluster analysis. Substructure searching has been introduced above. It is one of the most important facilities in modern chemoinformatics systems but can result in very large hit-lists if it is not possible to specify a sufficiently detailed query substructure and/or if a big database, such as the CAS Registry, is being searched. In such cases, it can be helpful to cluster the molecules that have been retrieved, with the resulting cluster representatives then being used to obtain an overview of the range of structural types present in the hit-list [69]. Clustering can be used as a method of property prediction for structure-activity relationship (or SAR) studies [13], where SAR covers a range of methods for identifying statistical relationships between chemical structure and biological activity data [4]. An alternative, but related, application is to identify one, or some small number, of clusters that can then be analysed to determine the nature and the extent of any SAR present in the chosen sets of molecules [70]. Finally, clustering two or more databases together can serve to identify the extent of the overlap between sets of molecules from different sources [71]. For example, the degree of overlap between a corporate chemical database and one offered by a commercial supplier could be assessed by examining the contents of each cluster in turn: if a pharmaceutical company's molecules were notably under-represented in a particular cluster then it might be appropriate to purchase some of the vendor molecules to augment the corporate database.

Inspection of the pattern recognition, multivariate statistics and data mining literatures reveals a huge number of different clustering methods, with new approaches continuing to be described for a range of applications. Early studies of over 30 hierarchic and non-hierarchic methods [72] suggested that the best chemical classifications (in the sense of successfully grouping molecules with similar biological activities) were obtained using Ward's hierarchical-agglomerative method [73], with the non-hierarchical Jarvis-Patrick method [74] also performing well. Jarvis-Patrick is by far the more efficient of these two methods, and it was hence the method of choice for clustering large chemical datasets for many years. However, improvements in both hardware and software, coupled with further demonstrations of the greater effectiveness of Ward's method [13, 75] mean that this has now largely replaced the Jarvis-Patrick method for clustering databases containing up to ca. half-a-million structures. For larger files, the current standard is the divisive *k*-means method [76], with recent reported applications including SAR studies [77], comparing classifications based on substructural fragments and on ring scaffolds [78], and merging corporate databases [79].

Cluster analysis is hence extensively used in chemoinformatics, providing a simple, readily comprehendible way of grouping structurally related molecules. Its principal limitations are those of cluster analysis itself, such as the parameter-driven nature of many of the clustering methods that can be used, and the variant (and often non-unique) solutions that result from the use of different methods.

MOLECULAR DIVERSITY ANALYSIS

The use of clustering methods to identify subsets of databases (as described above) was the first approach to be used that comes under the general heading of *molecular diversity analysis*. This is the name given to techniques that maximise the degree of diversity (or dissimilarity or resemblance) in a set of molecules. As noted above when discussing cluster analysis, a cost-effective approach to biological testing involves a limited number of molecules that, taken together, describe the chemical space spanned by the complete set of molecules comprising a database (where this database could be an in-house corporate file, an external public or vendor database, or a set of molecules that could potentially be synthesised). The Similar Property Principle means that structurally similar molecules are likely to exhibit similar properties, and hence testing sets of similar molecules is unlikely to provide much more SAR information than would be obtained by testing just one or a few such molecules; instead, most information is likely to be obtained by testing sets of molecules that are as diverse, i.e., structurally dissimilar, as possible [80].

Many techniques for molecular diversity analysis have been described [81-83]; here we focus on those that make use of fingerprint-based calculations of inter-molecular dissimilarity (where the dissimilarity is normally the complement of the Tanimoto similarity). Similarity and dissimilarity are properties of a pair of molecules, whereas diversity is the property of a set of molecules (either an entire database or a subset thereof) and is computed by combining sets of pair-wise similarities or dissimilarities. For example, a common measure of diversity for a set of *n* molecules (and the measure considered in what follows) is the sum of the *n*(*n*-1)/2 pair-wise dissimilarities. Given this definition it is trivial in principle to identify the most diverse *n*-molecule subset of an *N*-molecule database (and hence the subset that should be submitted for biological testing) simply by computing the sum of the similarities for each possible *n*-molecule subset in turn. However, this is computationally infeasible, requiring consideration of up to

$$\frac{N!}{n!(N-n)!} \tag{3}$$

different subsets, and practical methods for subset selection hence make use of more approximate methods. There are three main approaches that use fingerprint-based similarities: cluster-based selection as described previously; and dissimilarity-based selection and optimisation-based selection as described below.

In dissimilarity-based selection the subset of selected molecules is initiated by choosing a molecule at random, then adding that molecule that is most dissimilar to the first molecule, then that molecule that is most dissimilar to the first two molecules, and so on until a subset of the desired size has been obtained [84]. An alternative, *sphere-exclusion* approach involves selecting an initial molecule and then excluding from further consideration all molecules that have a similarity greater than some threshold with the chosen molecule. In subsequent stages, that non-excluded molecule is chosen for inclusion in the subset that has the largest dissimilarity to those molecules that have already been selected, and further molecules excluded if they are nearest neighbours of the one that has been chosen [85] (other approaches have also been described [86]). These approaches involve the identification of the most dissimilar molecule at each stage, and different results can be obtained depending on how ‘most dissimilar’ is defined: the MaxMin approach is widely used, and involves selecting that molecule for inclusion that has the maximum dissimilarity to its nearest neighbour in the current subset of selected molecules [87].

The final approach involves use of a combinatorial optimisation procedure, with the optimisation being driven by the need to maximise the diversity of the chosen subset. The diversity is typically the sum of pair-wise dissimilarities for the molecules chosen for inclusion in the subset. Both genetic algorithms [88, 89] and simulated annealing [90, 91] have been used for this purpose, with Waldman *et al*. providing a detailed overview of the diversity criteria that can be employed [92]. The focus of the present review is the use of fingerprint methods; optimisation-based selection often uses additional information to ensure that the molecules chosen for inclusion in the subset are not just structurally diverse but also exhibit physicochemical properties typical of those for a drug [91, 93, 94]. A sophisticated example of this is provided by the work of Gillet, who uses Pareto optimization to obtain a family of equivalent solutions, each of which represents a different trade-off between the often conflicting requirements of the various objectives in the optimisation [95].

Molecular diversity analysis rapidly established itself as an effective tool for identifying structurally dissimilar sets of compounds. However, it came to be realised that the pursuit of diversity alone is not sufficient for the purposes of drug discovery: not only must the compounds selected for testing be structurally diverse, but they must also be drug-like, in the sense of exhibiting physical and chemical properties characteristic of known bioactive molecules. This can be achieved in part by using Pareto-based selection methods (as in the work of Gillet *et al*. mentioned above). Current drug-discovery programmes hence complement diversity methods with filters designed to ensure the drug-like nature of the compounds that are to be selected for testing [96, 97].

Conclusion

This review has discussed measures of inter-molecular structural similarity based on chemical fingerprints, and their use for three data mining applications in large chemical databases. Fingerprints provide a compact but effective description of the 2D substructures present in a molecule, but without an explicit description of the molecule’s topology. It is hence not unreasonable to suspect that the inclusion of such information would result in more effective measures of similarity, and one might reasonably ask why such a simple, indeed crude, approach to the quantification of similarity is still in widespread use more than a quarter-of-century after the first publications [43, 47]. The main reason is the huge numbers of similarity calculations required for the processing of large databases: there is hence a considerable premium associated with computational efficiency, and comparing two binary vectors is far faster than the operations required if more complex representations are adopted. We have noted previously that the graph-matching operations required for a full topology match are extremely time-consuming, and yet comparisons of fingerprints and graph matching for similarity searching [98] and for clustering [99] suggest that the former, more efficient approaches are of comparable or superior effectiveness. Thus, we do not observe the trade-off between efficiency and effectiveness that might have been expected.

Similar comments apply to the use of measures of 3D similarity. The shape of a molecule is often a key factor in determining whether a molecule will be bioactive, and many types of 3D representation have thus been reported in the similarity literature [7]. Examples include fingerprints that encode inter-atomic distance or angular information [100, 101], and descriptions of molecular shape [102, 103] and of the distribution of electrostatic charge around a molecule [104, 105]. However, these all need to take account of the fact that most molecules are flexible, i.e., they can adopt several or many different 3D shapes (called *conformations*) (where as a molecule has only a single 2D topology). These multiple conformations need to be considered if one wishes to provide a comprehensive description of molecular geometry, which has inevitable, and often large, computational requirements. Moreover, even if account is taken of conformational flexibility, there is again little evidence to suggest that the results are notably better than those obtained with 2D fingerprints [13, 49, 106].

At some point, it will surely prove possible to identify measures of topological and geometric similarity that are both efficient and effective in operation. Till then, 2D fingerprints provide an appropriate tool for data mining in the large databases that characterise pharmaceutical research.

References

1. Chen WL: **Chemoinformatics: past, present and future**. *Journal of Chemical Information and Modeling* 2006, **46**:2230-2255.
2. Willett P: **Chemoinformatics - a history**. *WIREs Computational Molecular Sciences* 2011, **1**:TBC.
3. Gasteiger J (ed.): **Handbook of Chemoinformatics**. Weinheim: Wiley-VCH; 2003.
4. Leach AR, Gillet VJ: **An Introduction to Chemoinformatics**, 2nd edition edn. Dordrecht: Kluwer; 2007.
5. Eckert H, Bajorath J: **Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches**. *Drug Discovery Today* 2007, **12**:225-233.
6. Sheridan RP: **Chemical similarity searches: when is complexity justified?** *Expert Opinion on Drug Discovery* 2007, **2**:423-430.
7. Willett P: **Similarity methods in chemoinformatics**. *Annual Review of Information Science and Technology* 2009, **43**:3-71.
8. Maggiora GM, Shanmugasundaram V: **Molecular similarity measures**. *Methods in Molecular Biology* 2010, **672**:39-100.
9. Bender A: **How similar are those molecules after all? Use two descriptors and you will have three different answers**. *Expert Opinion on Drug Discovery* 2010, **in press**.
10. Johnson MA, Maggiora GM (eds.): **Concepts and Applications of Molecular Similarity**. New York: John Wiley; 1990.
11. Wilkins CL, Randic M: **A graph theoretical approach to structure-property and structure-activity correlation**. *Theoretica Chimica Acta* 1980, **58**:45-68.
12. Willett P, Winterman V: **A comparison of some measures of inter-molecular structural similarity**. *Quantitative Structure-Activity Relationships* 1986, **5**:18-25.
13. Brown RD, Martin YC: **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* 1996, **36**:572-584.
14. Martin YC, Kofron JL, Traphagen LM: **Do structurally similar molecules have similar biological activities?** *Journal of Medicinal Chemistry* 2002, **45**:4350-4358.
15. He L, Jurs PC: **Assessing the reliability of a QSAR model's predictions**. *Journal of Molecular Graphics and Modelling* 2005, **23**:503-523.
16. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK: **Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR**. *Journal of Chemical Information and Computer Sciences* 2004, **44**:1912-1928.
17. Steffen A, Kogej T, Tyrchan C, Engkvist O: **Comparison of molecular fingerprint methods on the basis of biological profile data** *Journal of Chemical Information and Modeling* 2009, **49**:338-347.

18. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E: **Similarity metrics for ligands reflecting the similarity of the target proteins**. *Journal of Chemical Information and Computer Sciences* 2003, **43**:391-405.
19. Bostrom J, Hogner A, Schmitt S: **Do structurally similar ligands bind in a similar fashion?** *Journal of Medicinal Chemistry* 2006, **49**:6716-6725.
20. Cleves AE, Jain AN: **Robust ligand-based modeling of the biological targets of known drugs**. *Journal of Medicinal Chemistry* 2006, **49**:2921-2938.
21. Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK: **Quantifying the relationship among drug classes**. *Journal of Chemical Information and Modeling* 2008, **48**:755-765.
22. Kubinyi H: **Similarity and dissimilarity: a medicinal chemist's view**. *Perspectives in Drug Discovery and Design* 1998, **9-11**:225-232
23. Nikolova N, Jaworska J: **Approaches to measure chemical similarity - a review**. *Quantitative Structure-Activity Relationships and Combinatorial Science* 2003, **22**:1006-1026
24. Raymond JW, Willett P: **Maximum common subgraph isomorphism algorithms for the matching of chemical structures**. *Journal of Computer-Aided Molecular Design* 2002, **16**:521-533.
25. Rarey M, Ehrlich H-C: **Maximum common subgraph isomorphism algorithms and their applications in molecular science: A review**. *WIRES Computational Molecular Sciences* 2011, **1**:in press.
26. Gillet VJ, Downs GM, Ling A, Lynch MF, Venkataram P, Wood JV, Dethlefsen W: **Computer-storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical-structure retrieval**. *Journal of Chemical Information and Computer Sciences* 1987, **27**(3):126-137.
27. Rarey M, Dixon JS: **Feature trees: A new molecular similarity measure based on tree matching**. *Journal of Computer-Aided Molecular Design* 1998, **12**:471-490.
28. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures**. *Organic and Biomolecular Chemistry* 2004, **2**:3256-3266.
29. Sastry M, Lowrie JF, Dixon SL, Sherman W: **Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments**. *Journal of Chemical Information and Modeling* 2010, **50**:771-748.
30. Bremser W: **HOSE – a novel substructure code**. *Analytica Chimica Acta* 1978, **103**:355-365.
31. Attias R: **DARC substructure search system: a new approach to chemical information**. *Journal of Chemical Information and Computer Sciences* 1983, **23**:102-108.
32. Willett P: A s**creen set generation algorithm**. *Journal of Chemical Information and Computer Sciences* 1979, **19**:159-162.
33. Bender A, Mussa HY, Glen RC, Reiling S: **Molecular similarity searching using atom environments: information-based feature selection and a naive Bayesian classifier**. *Journal of Chemical Information and Computer Sciences* 2004, **44**:170-178.
34. Rogers D, Hahn M: **Extended-connectivity fingerprints**. *Journal of Chemical Information and Modeling* 2010, **50**:742-754.
35. Arif SM, Holliday JD, Willett P: **Analysis and use of fragment occurrence data in similarity-based virtual screening**. *Journal of Computer-Aided Molecular Design* 2009, **23**:655-668.
36. Arif SM, Holliday JD, P. W: **Inverse frequency weighting of fragments for similarity-based virtual screening**. *Journal of Chemical Information and Modeling* 2010, **50**:1340-1349.
37. Willett P, Barnard JM, Downs GM: **Chemical similarity searching**. *Journal of Chemical Information and Computer Sciences* 1998, **38**:983-996.
38. Holliday JD, Hu C-Y, Willett P: **Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings**. *Combinatorial Chemistry and High-Throughput Screening* 2002, **5**:155-166.

39. Rogers DJ, Tanimoto TT: **A computer program for classifying plants** *Science* 1960, **1960**:1115-1118.
40. Jaccard P: **Étude comparative de la distribution florale dans une portion des Alpes et des Jura**. *Bulletin de la Société Vaudoise des Sciences Naturelles* 1901, **37**:547-579.
41. Sokal RR, Sneath PH: **Principles of Numerical Taxonomy**. San Francisco: W.H. Freeman; 1963.
42. Gower JC: **Measures of similarity, dissimilarity and distance**. In: *Encyclopaedia of Statistical Sciences.* Edited by Kotz S, Johnson NL, Read CB. Chichester: John Wiley; 1982: 397-405.
43. Willett P, Winterman V, Bawden D: **Implementation of nearest-neighbour searching in an online chemical structure search system**. *Journal of Chemical Information and Computer Sciences* 1986, **26**:36-41.
44. Flower DR: **On the properties of bit string based measures of chemical similarity**. *Journal of Chemical Information and Computer Sciences* 1988, **38**:379-386.
45. Fligner MA, Verducci JS, Blower PE: **A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings**. *Technometrics* 2002, **44**:110-119.
46. Willett P: **Similarity-based virtual screening using 2D fingerprints**. *Drug Discovery Today* 2006, **11**:1046-1053.
47. Carhart RE, Smith DH, Venkataraghavan R: **Atom pairs as molecular-features in structure activity studies - definition and applications**. *Journal of Chemical Information and Computer Sciences* 1985, **25**:64-73.
48. Geppert H, Vogt M, Bajorath J: **Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation**. *Journal of Chemical Information and Modeling* 2010, **50**:205-216.
49. Sheridan RP, Kearsley SK: **Why do we need so many chemical similarity search methods?** *Drug Discovery Today* 2002, **7**:903-911.
50. Willett P: **Data fusion in ligand-based virtual screening**. *QSAR and Combinatorial Science* 2006, **25**:1143-1152.
51. Williams C: **Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance**. *Molecular Diversity* 2006, **10**:311-332.
52. Chen B, Mueller C, Willett P: **Combination rules for group fusion in similarity-based virtual screening**. *Molecular Informatics* 2010, **29**:533-541.
53. Whittle M, Gillet VJ, Willett P, Loesel J: **Analysis of data fusion methods in virtual screening: theoretical model**. *Journal of Chemical Information and Modeling* 2006, **46**:2193-2205.
54. Brown N, Jacoby E: **On scaffolds and hopping in medicinal chemistry**. *Mini-Reviews in Medicinal Chemistry* 2006, **6**:1217-1229.
55. Vogt M, Stumpfe D, Geppert H, Bajorath J: **Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening**. *Journal of Medicinal Chemistry* 2010, **53**:5707-5715.
56. Klebe G (ed.): **Virtual Screening: an Alternative or Complement to High Throughput Screening**. Dordrecht: Kluwer; 2000.
57. Böhm H-J, Schneider G (eds.): **Virtual Screening for Bioactive Molecules**. Weinheim: Wiley-VCH; 2000.
58. Stahura FL, Bajorath J: **Virtual screening methods that complement high-throughput screening**. *Combinatorial Chemistry and High-Throughput Screening* 2004, **7**:259-269.
59. Oprea TI, Matter H: **Integrating virtual screening in lead discovery**. *Current Opinion in Chemical Biology* 2004, **8**:349-358.
60. Alvarez J, Shoichet B (eds.): **Virtual Screening in Drug Discovery**. Boca Raton: CRC Press; 2005.
61. Abdo A, Salim N: **Similarity-based virtual screening with a Bayesian inference network**. *ChemMedChem* 2009, **4**:210-218.

62. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ: **Application of belief theory to similarity data fusion for use in analog searching and lead hopping**. *Journal of Chemical Information and Modeling* 2008, **48**:941-948.

63. Leach AR, Gillet VJ, Lewis RA, Taylor R: **3D pharmacophore methods in drug discovery**. *Journal of Medicinal Chemistry* 2010, **53**:539-558.

64. Goldman BB, Walters WP: **Machine learning in computational chemistry**. *Annual Reports in Computational Chemistry* 2006, **2**:127-140.

65. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S *et al*: **A critical assessment of docking programs and scoring functions**. *Journal of Medicinal Chemistry* 2006, **49**:5912-5931.

66. Sneath PHA, Sokal RR: **Numerical Taxonomy**. San Francisco: W. H. Freeman; 1973.

67. Everitt BS, Landau S, Leese M: **Cluster Analysis**, 4th edition edn. London: Edward Arnold; 2001.

68. Downs GM, Barnard JM: **Clustering methods and their uses in computational chemistry.** *Reviews in Computational Chemistry* 2002, **18**:1-40.

69. 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* 1986, **26**:109-118.

70. Nouwen J, Hansen B: **An investigation of clustering as a tool in quantitative structure-activity relationships (QSAR)**. *SAR and QSAR in Environmental Research* 1995, **4**:1-10.

71. Shemetulskis NE, Dunbar JB, Dunbar BW, Moreland DW, Humblet C: **Enhancing the diversity of a corporate database using chemical database clustering and analysis.** *Journal of Computer-Aided Molecular Design* 1995, **9**:407-416.

72. Willett P: **Similarity and Clustering in Chemical Information Systems**. Letchworth: Research Studies Press; 1987.

73. Ward JH: **Hierarchical grouping to optimize an objective function.** *Journal of the American Statistical Association* 1963, **58**:236-244.

74. Jarvis RA, Patrick EA: **Clustering using a similarity measure based on shared nearest neighbours**. *IEEE Transactions on Computers* 1973, **C-22**:1025-1034.

75. Downs GM, Willett P, Fisanick W: **Similarity searching and clustering of chemical-structure databases using molecular property data**. *Journal of Chemical Information and Computer Sciences* 1994, **34**:1094-1102.

76. Steinbach M, Karypis G, Kumar VA: **Comparison of Document Clustering Techniques**. In.: Department Computer Science & Engineering, University of Minnesota; 2000.

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

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

79. Engels MFM, Gibbs AC, Jaeger EP, Verbinnen D, Lobanov VS, Agrafiotis DK: **A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition**. *Journal of Chemical Information and Modeling* 2006, **46**:2651-2660.

80. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors**. *Journal of Medicinal Chemistry* 1996, **39**:3049-3059.

81. Dean PM, Lewis RA (eds.): **Molecular Diversity in Drug Design**. Amsterdam: Kluwer; 1999.

82. Lewis RA, Pickett SD, Clark DE: **Computer-aided molecular diversity analysis and combinatorial library design**. *Reviews in Computational Chemistry* 2000, **16**:1-51.

83. Maldonado AG, Doucet JP, Petitjean M, Fan B-T: **Molecular similarity and diversity in chemoinformatics: from theory to applications**. *Molecular Diversity* 2006, **10**:39-79.

84. Lajiness MS: **Dissimilarity-based compound selection techniques**. *Perspectives in Drug Discovery and Design* 1997, **7/8**:65-84.

85. Hudson BD, Hyde RM, Rahr E, Wood J: **Parameter based methods for compound selection from chemical databases**. *Quantitative Structure-Activity Relationships* 1996, **15**:285-289.

86. Pearlman RS, Smith KM: **Novel software tools for chemical diversity**. *Perspectives in Drug Discovery and Design* 1998, **9-11**:339-353.

87. Snarey M, Terrett NK, Willett P, Wilton DJ: **Comparison of algorithms for dissimilarity-based compound selection**. *Journal of Molecular Graphics and Modelling* 1997, **15**:372-385.

88. Brown RD, Martin YC: **Designing combinatorial library mixtures using a genetic algorithm**. *Journal of Medicinal Chemistry* 1997, **40**:2304-2313.

89. Sheridan RP, Kearsley SK: **Using a genetic algorithm to suggest combinatorial libraries**. *Journal of Chemical Information and Computer Sciences* 1995, **35**:310-320.

90. Hassan M, Bielawski JP, Hempel JC, Waldman M: **Optimization and visualization of molecular diversity of combinatorial libraries**. *Molecular Diversity* 1996, **2**:64-74.

91. Good AC, Lewis RA: **New methodology for profiling combinatorial libraries and screening sets: cleaning up the design with HARPick**. *Journal of Medicinal Chemistry* 1997, **40**:3926-3936.

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

93. Agrafiotis DK: **Multiobjective optimization of combinatorial libraries**. *Journal of Computer-Aided Molecular Design* 2002, **16**:335-356.

94. Brown RD, Hassan M, Waldman M: **Combinatorial library design for diversity, cost efficiency and druglike character**. *Journal of Molecular Graphics and Modelling* 2000, **18**:427-437.

95. Gillet VJ: **Designing combinatorial libraries optimized on multiple objectives**. *Methods in Molecular Biology* 2004, **275**:335-354.

96. Clark DE, Pickett SD: **Computational methods for the prediction of 'drug-likeness'**. *Drug Discovery Today* 2000, **5**:49-58.

97. Oprea TI: **Property distribution of drug-related chemical databases**. *Journal of Computer-Aided Molecular Design* 2000, **14**:251-264.

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

99. Raymond JW, Blankley CJ, Willett P: **Comparison of chemical clustering methods using graph-based and fingerprint-based similarity measures**. *Journal of Molecular Graphics and Modelling* 2003, **21**:421-433.

100. Fisanick W, Cross KP, Rusinko A: **Similarity searching on CAS Registry substances. 1. Global molecular property and generic atom triangle geometric searching.** *Journal of Chemical Information and Computer Sciences* 1992, **32**:664-674.

101. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF: **New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures**. *Journal of Medicinal Chemistry* 1999, **42**:3251-3264.

102. Ballester PJ, Richards WG: **Ultrafast shape recognition to search compound databases for similar molecular shapes**. *Journal of Computational Chemistry* 2007, **28**:1711-1723.

103. Hawkins PDC, Skillman AG, Nicholls A: **Comparison of shape-matching and docking as virtual screening tools**. *Journal of Medicinal Chemistry* 2007, **50**:74-82.

104. Mestres J, Rohrer DC, Maggiora GM: **MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches**. *Journal of Computational Chemistry* 1997, **18**:934-954.

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

106. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD: **Comparison of topological, shape, and docking methods in virtual screening**. *Journal of Chemical Information and Modeling* 2007, **47**:1504-1519.

Related Articles

Article ID	**Article title**
033	Drug design
133	Mining of chemical databases