Published paper

Willett, P. (2011) *Similarity searching using 2D structural fingerprints*. Methods in Molecular Biology, 672. 133 - 158. ISSN 1064-3745
http://dx.doi.org/10.1007/978-1-60761-839-3_5

Published in *Methods in Molecular Biology*, **672** (2010) 133-158

Similarity Searching Using 2D Structural Fingerprints

Peter Willett

Abstract

This paper reviews the use of molecular fingerprints for chemical similarity searching. The fingerprints encode the presence of 2D substructural fragments in a molecule, and the similarity between a pair of molecules is a function of the number of fragments that they have in common. Although this provides a very simple way of estimating the degree of structural similarity between two molecules, it has been found to provide an effective and an efficient tool for searching large chemical databases. The review describes the historical development of similarity searching since it was first described in the mid-Eighties, reviews the many different coefficients, representations, and weightings that can be combined to form a similarity measure, describes quantitative measures of the effectiveness of similarity searching and concludes by looking at current developments based on the use of data fusion and machine learning techniques.

Key Words: Chemical databases; Chemoinformatics; Data fusion; Fingerprint; Fragment substructure; Machine learning; Similar Property Principle; Similarity coefficient; Similarity measure; Similarity searching; Weighting scheme.

1. Introduction

The Collins English Dictionary defines similar to be “showing resemblance in qualities, characteristics or appearance; alike but not identical” and the comparison of objects to determine their levels of similarity lies at the heart of many academic disciplines. Thus, archaeologists may study the relationships between pot shards from different historical sites; literary studies may involve comparing fragments of poetry from different works by – possibly – the same author; and modern systematics derives from the attempts of the medieval apothecaries to group medicinal plants. The definitions of similarity, and the purposes for which these definitions are employed, in these three applications are very different but they have in common the aim of synthesising new knowledge from a similarity-based analysis of that which already exists. Similarity concepts have long played an important role in chemistry (***1***); indeed one of the most striking examples is the work of Mendeleev that led to the establishment of the modern Periodic Table, by means of which he was able not only to classify the existing elements but also to predict the existence of elements that were then unknown.

In this chapter, we focus on one specific application of similarity in chemoinformatics: *similarity searching*, i.e., the ability to scan through a database of chemical molecules to find those that are most similar to some user-defined query molecule (***2-7***). In what follows, we shall normally refer to the query as the *reference structure*; an alternative name that is frequently used in the literature is the *target structure*, but we believe that the former name is to be preferred given the possibility of confusion with a biological target.

Similarity searching is one particular type of *virtual screening*. This is the use of a computational technique for selecting molecules for subsequent investigation, most obviously for testing for bioactivity in a lead-discovery programme. There are many different virtual screening methods available, but they all have the common aim of ranking a list of possible molecules so that those at the top of the ranking are expected to have the greatest probability of activity. Virtual screening methods differ in the amount of information that is available (***8-13***). Similarity searching has by far the smallest information requirement, needing just a single known bioactive molecule (as discussed further below), Examples of other approaches to virtual screening include: 2D or 3D substructure searching (which require the availability of a topological or geometric pharmacophore, respectively, these being derived from a small number of known bioactive molecules); machine learning methods (which require large numbers of both known active and known inactive molecules); and docking methods (which require the 3D structure of the biological target). The many methods that are now available have led to comparisons that seek to determine the relative effectiveness of different approaches to screening; the reader is referred to the literature for discussions of the strengths and weaknesses of similarity searching as compared to other screening approaches (see, e.g., (***7, 14-20***)).

This chapter seeks to present the basic principles of similarity searching, eschewing detailed discussion of individual approaches, and is structured as follows. Section 2 provides an introduction to similarity searching, and describes the *Similar Property Principle* that underlies the use of similarity as a tool for database searching. Section 3 discusses the three components – the *representation*, the *similarity coefficient* and the *weighting scheme* – that comprise a *similarity measure* for computing the degree of resemblance between two molecules; the focus of this chapter is one particular type of representation, the 2D *fingerprint*, and this representation is hence discussed in some detail in this section. Section 4 discusses the criteria that have been used to evaluate the retrieval effectiveness of different types of similarity searching procedure. Finally, Section 5 summarises recent work that involves the use of not just a single reference structure, as is used for conventional similarity searching, but multiple reference structures.

The coverage of this review is intentionally focused, considering only one representation of molecular structure (the *2D fingerprint*) and only one application of similarity (similarity searching). The reader is referred to the literature for more general discussions of chemoinformatics (***21-23***) and of other similarity-related topics, such as 3D similarity measures, cluster analysis, molecular diversity analysis, and reaction similarity (***3, 24-26***); for additional material specifically about similarity searching, it is worth noting that a characteristic of the field is that much of the work has been carried out by a limited number of research groups, most notably those directed by Bajorath (***27***) , Sheridan (***16***), and Willett (***28***).

2. The Similar Property Principle

The input to a similarity search is the reference structure for which related structures are required. In the drug-discovery context, the reference structure normally exhibits a potentially useful level of biological activity and might be, for example, a competitor's compound or a structurally novel hit from an initial high-throughput screening (HTS) experiment. Thus the reference structure is normally an entire molecule, rather than the partial structure that forms the basis for 2D or 3D substructure searching (that said, there has been some interest in similarity searches of molecules that are substructures or superstructures of the reference structures (***29-31***)). Each database structure is encoded using the same representation scheme as was used to encode the reference structure; the two representations are compared to ascertain the level of structural commonality using a similarity coefficient. In some cases, a weighting scheme is applied to one or both of the representations prior to the calculation of the similarity, with the aim of increasing the relative importance of particular features within the overall representation. The similarities are computed in this way for every molecule in the database that is being searched, and then the similarity values sorted into descending order. The molecules at the top of the resulting ranking, which are often referred to as the *nearest neighbours* as they are the closest in some sense to the reference structure, are then presented to the user as the output from the similarity search.

This approach to database access was first described by Carhart *et al*. (***32***) and by Willett *et al*. (***33***). Both of these studies found that effective measures of chemical similarity could be obtained by determining the numbers of 2D substructures common to a reference structure and a database structure, although the starting points for the two studies were rather different. Carhart *et al*., working at Lederle Laboratories, used the information about common fragments not just for similarity searching but also for substructural analysis (*vide infra*). The study by Willett *et al*. drew on earlier work by Adamson and Bush that reported probably the very first use of 2D fingerprints for the calculation of molecular similarity (specifically in the context of QSAR studies rather than for large-scale database applications) **(*34*)**. Willett *et al*. used the information about common fragments in a combined search system at Pfizer, where the computed similarities were used to rank the molecules retrieved in a substructure search; however, the authors soon realised that the initial substructure search was not necessary and that similarity searching on its own provided a novel way of accessing a chemical database.

Following these two initial studies, fragment-based similarity searching was adopted very rapidly in both commercial and in-house chemoinformatics systems. Its uptake was spurred by several factors: it provides a retrieval mechanism that is complementary to substructure searching; it uses the same basic data as existing substructure software, i.e., sets of 2D fingerprints; and it is both rapid and powerful in execution, encouraging interactive exploration of the range of structural types in a database **(*35*)**. These are all perfectly valid, but essentially pragmatic reasons for using similarity searching. There is, however, also a rational basis, which derives from what is known as the Similar Property Principle. The Principle states that molecules that have similar structures will have similar properties, and is normally ascribed to Johnson and Maggiora, whose 1990 book was the first to highlight the role of similarity in what we now refer to as chemoinformatics **(*25*)**. However, it had certainly been discussed prior to then, e.g., by Wilkins and Randic in 1980 **(*36*)**, and arguably underlies the whole area of drug discovery: if there was not some relationship between molecular structures (however these are represented in computational terms) and molecular properties then lead discovery and lead optimisation would be essentially random processes, which is certainly not the case. If the Principle holds then the molecules in a database that are most similar to a bioactive reference structure are (all other things being equal) those that are most likely to exhibit the reference structure's bioactivity. Ranking the database in order of decreasing similarity, where the similarity is defined using some quantitative measure of inter-molecular similarity, hence provides a rational way of prioritising compounds for biological testing and thus a firm basis for the development of similarity searching methods. It is appropriate to mention here the closely related concept of Neighbourhood Behaviour **(*37*)**, which involves relating absolute differences in bioactivity for pairs of molecules to the dissimilarities for those pairs of molecules. This concept has been used to categorise the effectiveness of molecular descriptors for molecular diversity applications **(*38-40*)**.

Given the importance of the Similar Property Principle, it is hardly surprising that there have been several attempts to demonstrate its applicability. Perhaps the first detailed study was that reported by Willett and Winterman, which showed that simple fingerprint-based similarities could be used to predict a range of physical, chemical and biological properties in small QSAR datasets (using a "leave-one-out" prediction approach that is discussed later in this review) (***41***). Having demonstrated that similarities in structure mirrored similarities in property, these authors then used differences in the strength of this relationship to compare different types of similarity measure. Specifically, they made the assumption that if the Principle holds for some particular dataset, then the extent of the relationship between structure and property that is obtained using some particular similarity measure provides a basis for evaluating the effectiveness of that measure, and hence for comparing the effectiveness of different types of similarity measure. Analogous results were obtained for their QSAR datasets when they were clustered using a range of hierarchic and non-hierarchic clustering methods (***42***). The latter work was extended to much larger datasets in two papers by Brown and Martin (***43, 44***). These studies were designed to compare the effectiveness of different clustering methods and different types of fingerprint for selecting structurally diverse database subsets, but their detailed experiments demonstrate clearly the general applicability of the Principle. A later paper by Martin *et al*. provided a direct evaluation of the Principle using structures that had been tested in over one-hundred assays at Abbott Laboratories (***45***). Whilst noting that there were cases where the Principle did not apply, the principal conclusion was that structurally similar compounds do indeed have similar bioactivities, with the latter increasing as the structural similarity is increased. These studies have been taken further in an interesting study by Steffen *et al*., who show that the Principle also applies when molecular bioactivities are considered across a range of assays, rather than just a single assay as in the other studies cited here (***46***).

Further demonstrations of the general validity of the Principle come from two near-contemporaneous studies of the applicability of QSAR models. Thus Sheridan *et al*. (***47***) and He and Jurs (***48***) showed that the more similar a molecule was to molecules in the training-set then the more likely it was that an accurate prediction could be made using the QSAR model that had been derived from that training-set. More recently, Bostrom *et al*. analysed sets of protein-ligand complexes from the Protein Data Bank to demonstrate that molecules that are structurally similar tend to bind to a biological target in the same way, i.e., in addition to eliciting the same biological response, similar molecules achieve this by means of the same mode of action (***49***). Finally, the Principle is attracting further support from work in chemogenomics, with recent studies demonstrating: that molecules with similar 2D fingerprints bind to structurally related biological targets (***50, 51***); that molecule-based similarities can suggest novel functional relationships between targets that exhibit little sequence similarity (***52, 53***); and that pairs of molecules acting on a common target are more likely to be similar than pairs of molecules that do not share a common target (***54***).

It should be noted that there are many exceptions to the Principle, a situation that Stahura and Bajorath refer to as the Similarity Paradox (***55***). This is especially the case if attention is focused on the relatively small numbers of structurally related molecules that are commonly encountered in QSAR studies (***5, 6, 56***), where it is not uncommon for very slight changes in structure to bring about large changes in activity (a phenomenon that has been referred to as an "activity cliff" (***57, 58***)). However, the Similar Property Principle does provide a highly appropriate basis for similarity searching, where similarities are typically computed for large, or very large, numbers of molecules spanning a huge range of structural classes.

3. Components of a Similarity Measure

Any database searching system must be both efficient (i.e., must involve the use of minimal computing resources, typically time and space) and effective (i.e., must retrieve appropriate items from the database that is being searched). Modern computer hardware and software enable highly efficient similarity searches to be carried out on even the largest chemical databases (at least when using the 2D fingerprint approaches that are considered in this chapter), and we hence focus on the factors that control effectiveness. This is determined by the nature of the measure that is used to compute the degree of resemblance between the reference structure and each of the database structures. A similarity measure has three components: the representation that is used to characterise the molecules that are being compared; the weighting scheme that is used to assign differing degrees of importance to the various components of these representations; and the similarity coefficient that is used to provide a quantitative measure of the degree of structural relatedness between a pair of (possibly weighted) structural representations.

3.1 Representations

Very many techniques are available for representing and encoding the structures of 2D chemical molecules (***23, 24, 59***) and many of these representations have been used for similarity searching (***16, 26, 60***). It is common to divide the many techniques into three broad classes of descriptor: whole molecule (sometimes called 1D) descriptors; descriptors that can be calculated from 2D representations of molecules; and descriptors that can be calculated from 3D representations.

Whole molecule descriptors are single numbers, each of which represents a different property of a molecule such as its molecular weight, the numbers of heteroatoms or rotatable bonds, or a computed physicochemical parameter such as logP. A single 1D descriptor is not usually discriminating enough to allow meaningful comparisons of molecules and a molecule is hence normally represented by several (or many) such descriptors (***61, 62***). 2D descriptors include topological indices and substructural descriptors. A topological index is a single number that typically characterises a structure according to its size and shape (***63, 64***). There are many such indices: the simplest characterise molecules according to their size, degree of branching and overall shape, while more complex indices take account of the properties of atoms as well as their connectivities. As with 1D descriptors, multiple different indices are normally combined for similarity searching (***65***). Substructure-based descriptors characterise a molecule by the substructural features that it contains, either by the molecule's 2D chemical graph, or by its fingerprint. Fingerprints are the focus of this chapter and are hence discussed in more detail below. They have been found to be at least as effective, if not more so, for virtual screening than chemical graphs (***66***) despite the fact that they provide a much less precise representation of a molecule's structure than does the underlying graph (which contains a full description of the molecule's topology). There is hence some interest in the use of simplified graph representations for virtual screening (***67-70***), and it is likely that work in this area will be developed further in the future. 3D descriptors are inherently more complex since they need to take account of the fact that many molecules are conformationally flexible (although some successful 3D similarity measures have assumed that a molecule can be represented by a single, low-energy conformation). Similarity measures have been reported that are based on inter-atomic distances (***71***), molecular surfaces (***72***), electrostatic fields (***73, 74***) and molecular shapes (***75, 76***) *inter alia*.

This chapter focuses on fingerprint-based similarity searching, and it is hence appropriate to discuss the various types of fingerprint that are available in more detail. Fingerprints enable effective similarity searching, but they were first developed for efficient substructure searching. This involves using a subgraph isomorphism algorithm to check for an exact mapping of the atoms and bonds in a query substructure onto the atoms and bonds of each database structure (***23, 24***). Graph matching algorithms are far too slow to enable interactive substructure searching of large files on their own, and it is hence necessary to use an initial *screening* search. This filters out of the great majority of the database structures that do not contain all of the substructural fragments present in the query substructure, with only those few molecules that do contain all of these fragments being passed on for the time-consuming graph-matching stage. The presence or absence of fragments in a query substructure or in a database structure is encoded in a binary vector that is normally referred to as a fingerprint.

There are two main ways of selecting the fragments that are encoded in a fingerprint (***23, 24, 77, 78***). In a *dictionary-based* approach, there is a pre-defined list of fragments, with normally one fragment allocated to each position in the bit-string. A molecule is checked for the presence of each of the fragments in the dictionary, and a bit set (or not set) when a fragment is present (or absent). The dictionary normally contains several different types of fragment. For example, an *augmented atom* contains a central atom together with its neighbouring atoms and bonds, and an *atom sequence* contains a specific number of connected atoms and their intervening bonds. The effectiveness of the dictionary is maximised if a statistical analysis is carried out of the sorts of molecules that are to be fingerprinted, so as to ensure that the most discriminating fragments are included (***79-81***). In a *molecule-based* approach, hashing algorithms are used to allocate multiple fragments to each bit-position. Here, a generic fragment type is specified, e.g., a chain of four connected non-hydrogen atoms, and a note made of all fragments of that type that occur in a given molecule. Each fragment is converted to a canonical form and then hashed using several (typically two or three) hashing algorithms to set bits in the fingerprint. The first widely used fingerprint of this sort was that developed by Daylight Chemical Information Systems Inc. (at http://www.daylight.com). This fingerprint encodes atom sequences up to a specified length (typically from 2 to 7 atoms), with each such sequence being hashed using multiple hashing procedures so that each bit is associated with multiple fragments and each fragment with multiple bit positions.

Both the dictionary-based and the molecule-based approaches are represented in the fingerprints encountered in operational chemoinformatics systems. For example, the fingerprints produced by Digital Chemistry (formerly Barnard Chemical Information, at http://www.digitalchemistry.co.uk,), by Sunset Molecular (at http://www.sunsetmolecular.com) and by Symyx Technologies (formerly MDL Information Systems at http://www.symyx.com) are dictionary-based, the Daylight fingerprints mentioned previously and the fingerprints produced by Accelrys (at http://www.accelrys.com) are molecule-based (using linear chains and circular substructures, respectively), and the Unity fingerprints produced by Tripos (at http://www.tripos.com) are based on both approaches.

Most of the fingerprints above were originally developed for efficient substructure searching, and it is perhaps surprising that they have also been found to provide a highly effective, alternative type of database access. There are also fingerprints that have been developed specifically for similarity searching (***14, 51, 82-87***). It is noteworthy that many of the newer types of fingerprint describe the atoms not by their elemental types but by their physicochemical characteristics, so as to enable the identification of database structures that have similar properties to the reference structure in a similarity search but that have different sets of atoms. This increases the chances of *scaffold-hopping*, i.e., the identification of novel classes of molecule with the requisite bioactivity (***88-91***). We should also note that the discussion here is restricted to fingerprints that encode structural fragments: other types of fingerprint used for similarity searching have involved other types of information such as property information (***46, 92, 93***) or affinities to panels of proteins (***94, 95***).

3.2 Weighting schemes

Most fingerprints are binary in nature, with each bit denoting the presence/absence of a substructural fragment in a molecule. However, the elements of a fingerprint can also contain non-binary information that assigns a weight, or degree of importance, to the corresponding features. Thus, a feature that had a large weight and that occurred in both the reference structure and a database structure would contribute more to the overall similarity of those molecules than would a common feature with a small weight. Weighting features in fingerprints lies at the heart of many approaches to substructural analysis and related machine-learning approaches where large amounts of training data are available (*vide infra*) **(*27, 96, 97*)**, but has been much less studied in the context of similarity searching, where the only information that is available is the reference structure and the database structures that are to be searched.

Willett and Winterman suggested that three types of weighting could be used for fingerprint-based similarity searching: weighting based on the number of times that a fragment occurred in an individual molecule; weighting based on the number of times that a fragment occurred in an entire database; and weighting based on the total number of fragments within a molecule **(*41*)**. Of these three types of weight, the last is accommodated in many of the common similarity coefficients (*vide infra*) since they include a factor describing the sizes (in terms of numbers of fragments) of the two molecules that are being compared, whilst studies of the second type of weight have been limited to date **(*98, 99*)**. However there have been several studies of the use of information about fragment occurrences in a single molecule **(*41, 43, 70, 84, 85, 100-102*)**. These studies have suggested that fingerprints encoding the occurrences of substructural fragments may be able to give better screening performance than conventional, binary fingerprints. However, the results have been far from consistent; and the performance differences often quite small; many of the previous studies were limited, either in terms of the numbers of molecules involved or in the extent to which the weighted and binary fingerprints differed; and there has been no attempt to explain the observed levels of performance. This situation has been addressed in a recent study by Arif *et al*. **(*103*)**, which has demonstrated conclusively the general superiority of occurrence-based weighting and also rationalised the different (and sometimes very different) levels of performance that were observed in experiments involving a range of weighting schemes, types of fingerprint and chemical databases. Their recommended scheme involves encoding both the reference structure and the database structures using the square root of a fragment's occurrence; the study was, however, limited to the use of the Tanimoto coefficient (*vide infra*) and it remains to be seen whether analogous results are obtained with other types of coefficient.

3.3 Similarity coefficients

The calculation of inter-object similarities by means of a similarity coefficient lies at the heart of cluster analysis, a multivariate data analysis technique that is used across the sciences and social sciences (***104***), and very many different similarity coefficients have thus been developed for this purpose (***105, 106***). Willett *et al*. provide an extended account of those that have been used for applications in chemoinformatics (***35***), focusing on the mathematical characteristics of the various coefficients that they discuss and, in particular, on the broad class of similarity coefficients known as *association coefficients*. These are all based on the number of fragments, i.e., bits in a fingerprint, common to the fingerprints describing a reference structure and a database structure, with this number normalised by some function based on the numbers of non-zero bits in the two fingerprints that are being compared. An example of an association coefficient is the Tanimoto coefficient. This was found to work well in Willett and Winterman's early similarity study of QSAR datasets (***41***) and was hence adopted as the coefficient of choice when the first operational searching systems were introduced a few years later. Subsequent work has demonstrated the appropriateness of this choice: the Tanimoto coefficient has been found to perform well in a wide range of applications, and not just similarity searching, and remains the yardstick against which alternative approaches are judged, despite the many years that have passed since Willett and Winterman's initial study in 1986. Like most association coefficients, the Tanimoto coefficient takes values between zero and unity when used with binary fingerprints: a value of zero corresponds to two fingerprints that have no bits in common, while a value of unity corresponds to two identical fingerprints (***35***).

Whilst widely used, the Tanimoto coefficient is known to give low similarity values in searches for small reference structures (where just a few bits are switched on in the reference structure's fingerprint) (***107-109***), and is also known to have an inherent bias towards specific similarity values (***110***). These observations spurred several comparative studies (summarised in (***28***)) that involved over 20 different fingerprint-based similarity coefficients. None of the coefficients was found to be consistently superior to the Tanimoto coefficient, and it was shown (both experimentally and theoretically) that most coefficients exhibit at least some degree of dependence on the sizes (i.e.., numbers of set bits) of the molecules that are compared in a similarity search. Later studies have focussed on the use of asymmetric coefficients, based on ideas first put forward by Tversky (***111***), for the calculation of inter-molecular structural similarities (***112, 113***). In a symmetric coefficient, the value of the coefficient is independent of whether a reference structure is mapped to a database structure or *vice versa*. This is not so with asymmetric coefficients and it has been suggested that this may be beneficial for database searching (***30, 114***), although the merits of such coefficients are still the subject of debate (***115, 116***).

The coefficients discussed thus far focus on the substructural fragments that are common to a reference structure and a database structure, i.e., those positions in the fingerprint where the bit is switched on. Information about the other bits, i.e., those that are switched off, may be included implicitly, typically via a contribution to the overall coefficient that reflects sizes of the two molecules that are being compared. Extended versions have been reported of the Tanimoto and Tversky coefficients where the overall value of the coefficient is the weighted sum of one coefficient based on the bits switched on and of one coefficient based on the bits switched off **(*109, 117*)**.

Association coefficients are specifically designed for use with binary data. If interval or ratio data is used, as would be the case if some form of fragment weighting scheme was to be employed in the generation of a fingerprint, other types of coefficient may then be appropriate. The Euclidean distance has been found to work well in many data analysis studies, both in chemoinformatics and more generally **(*35, 104*)**; however, Varin *et al.* **(*118*)** have recently suggested that a coefficient described by Gower and Legendre **(*119*)**, which reduces to the Tanimoto coefficient when applied to binary data, performs very well when weighted fingerprints are used for clustering and similarity searching.

4. Evaluation of Similarity Measures

It will be clear from the above that there are very many possible combinations of fingerprint, coefficient, and weighting scheme that could be used to build a similarity measure for similarity searching. It is hence reasonable to ask how one can assess the effectiveness of different measures and thus how one can identify the most appropriate for a particular searching application.

The aim of similarity searching, as of any virtual screening method, is to identify bioactive molecules and the evaluation of search effectiveness is hence normally carried out using datasets for which both structure and bioactivity data are available. There is, of course, a vast amount of such data available in corporate databases as a result of the massive biological screening programs carried out by industry, but intellectual property considerations mean that this rarely, or ever, becomes available for more general use. This is a severe limitation since the development of the science of similarity searching requires standard datasets that can be used for the evaluation and comparison of different methods as they become available. Instead, most reported studies of similarity measures make use of a limited number of public datasets for which both structural and activity data are available. Examples of such datasets that have become widely used include the *MDL Drug Data Report* database (available from Symyx Technologies at http://www.symyx.com), the *World Of Molecular Bioactivity* database (available from Sunset Molecular at http://sunsetmolecular.com/), the National Cancer Institute AIDS database (available from the National Library of Medicine

Developmental Therapeutics Programme at http://dtp.nci.nih.gov), and the *Directory of Useful Decoys* (DUD) database (available from http://www.dud.docking.org/).

Although standard datasets are widely used, it is important to recognise that they do have some limitations. First, they contain molecules that have been reported as exhibiting some particular bioactivity but may say nothing as to their activity or inactivity against other biological targets; instead, it is normally the case that the absence of activity information is taken to mean inactivity. Second, molecules that reach the published literature (and that are hence eligible for inclusion in such databases) may be only a small, carefully studied and high-quality subset of those that were actually synthesised and tested in a screening program. Third, the "me too" or "fast follower" nature of research in the pharmaceutical industry means that some structural classes are overly represented in a dataset. Finally, the numbers of molecules in these datasets are typically an order of magnitude less than in corporate databases, which may contain several million molecules. Notwithstanding these characteristics, the existence of these datasets does mean that there is a natural platform for evaluating new methods and for comparing them with existing methods.

The bioactivity data can be either *qualitative* (e.g., a molecule is categorised as either active or inactive) or *quantitative* (e.g., an IC50 value is available for a molecule), but the Similar Property Principle provides the basis for performance evaluation irrespective of the precise nature of the biological data. If the Principle does hold for a particular dataset, i.e., if structurally similar molecules have similar activities, then the nearest-neighbour molecules in a similarity search are expected to have the same activity as the bioactive reference structure. The effectiveness of a similarity measure can hence be quantified by determining the extent to which the similarities resulting from its use mirror similarities in the bioactivity of interest.

Several reviews are available of effectiveness measures that can be used when qualitative activity data are available **(*38, 120, 121*)**. Most if not all of the common measures can be regarded as a function of one or both of two underlying variables: the *recall* and the *precision*. Assume that a similarity search has been carried out, and a threshold applied to the resulting ranked list to retrieve some small subset, e.g., 1%, of the database. Then the recall is the fraction of the total number of active molecules retrieved in that subset; and the precision is the fraction of that subset that is active. A good search is one that maximises both recall and precision so that, in the ideal case, a user would be presented with all of the actives in the database without any additional inactives: needless to say, this ideal is very rarely achieved in practice.

Examples of measures that have been extensively used include the *enrichment factor*, i.e., the number of actives retrieved relative to the number that would have been retrieved if compounds had been picked from the database at random **(*122*)**, the numbers of actives that have been retrieved at some fixed position in the ranking **(*123*)**, and the Receiver Operating

Characteristic (or ROC curve) (***124, 125***). A ROC curve plots the percentage of true positives retrieved against the percentage of false positives retrieved at each position in the ranking (or at some series of fixed positions, e.g., the top 5%, the top 10%, the top 15% etc). ROC curves are widely used in machine learning and pattern recognition research but their use in virtual screening has been criticised (***126***) since no particular attention is paid to the top-ranked molecules, and it is these that would actually be selected for testing in an operational screening system. There is much current interest in the evaluation of virtual screening (based on similarity searching, docking or whatever) and it is likely to be some time before full agreement is reached as to the best approaches to evaluation (***127, 128***).

Similarity searching is normally used in the lead discovery stage of a drug discovery programme, when only qualitative biological data are available and when the evaluation criteria mentioned in the previous paragraph are appropriate. However, the Similar Property Principle can also be applied to the analysis of datasets with quantitative data, using a leave-one-out approach analogous to those used in QSAR studies (***121***). Assume that the activity value for the reference structure *R* is known and is denoted by *A*(*R*). A similarity search is carried out and some number of *R*'s nearest neighbours identified. The predicted activity value for *R*, *P*(*R*), is then taken to be the arithmetic mean of the known activity values for this set of nearest neighbours. The similarity search is repeated using different reference structures, and the correlation coefficient is then computed between the resulting sets of *A*(*R*) and *P*(*R*) values. A large correlation coefficient implies a good fit between the known and predicted bioactivities and hence strict adherence to the Similar Property Principle by the similarity search procedure that was used to generate the sets of nearest neighbours. This approach to performance evaluation was pioneered by Adamson and Bush (***34***); it formed the basis for Willett's extensive studies of similarity and clustering methods in the Eighties (***42***) and, more recently, was used in Brown and Martin's much-cited comparison of structural descriptors for compound selection (***43, 44***).

A focus on the number of active molecules retrieved by a similarity search is entirely reasonable, but the needs of lead discovery mean that it is also important to consider the structural diversity of those active molecules (***129***). Specifically, account needs to be taken of the scaffold-hopping abilities of the similarity search since, e.g., a search retrieving 25 active analogues that all have the same scaffold as the reference structure is likely to be of much less commercial importance than a search retrieving just five actives if each of these has a different scaffold. It is often suggested that fragment-based 2D similarity searching has only a limited scaffold-hopping capability, especially when compared with more complex (and often much more time-consuming) 3D screening methods. This suggestion is clearly plausible but there is a fair amount of evidence to suggest that 2D methods can exhibit non-trivial scaffold-hopping capabilities (***16***) MORE REFS FROM BOX

The evaluation criteria described above have been used in a very large, and constantly increasing, number of studies that discuss the effectiveness of similarity searching. Even a brief discussion of these many studies would require a totally disproportionate amount of space, and the reader is accordingly referred to the many excellent reviews that exist (***2, 5-7, 35, 60***).

5. Use of Multiple Reference Structures

As discussed thus far, similarity searching has involved matching a single bioactive reference structure against a database using a single similarity measure. Over the last few years, perhaps the principal development in the field of similarity searching has been the appearance of a range of methods that involve the use of additional information in generating a ranking of the database. It is possible to identify two broad classes of approach: the first class involves the use of *data fusion*, or *consensus*, methods; while the second class involves the use of *machine learning* methods to develop predictive models that can guide future searches given a body of training data. It is debateable where similarity searching stops and where machine learning starts, but the main difference is in the amounts of bioactivity data available and the way that data is used. One of the principal attractions of similarity searching as a tool for virtual screening is that it requires just a single known active molecule; whereas the application of machine learning to virtual screening requires a pool of molecules (this pool ideally including not just actives but also inactives) to enable the development of a predictive model. In this review we shall focus more on data fusion since work in this area is more tightly aligned to conventional similarity searching, but make some remarks about machine learning approaches at the end of the section.

The comparative studies referenced in Section 4 have typically sought to identify a single "best" similarity method; hardly surprisingly, it has not been possible to identify a single approach that is consistently superior to all others across a range of reference structures, biological targets and performance criteria (***7, 16***). The data fusion approach involves carrying out multiple similarity searches and then combining the resulting search outputs to give a single fused output that is presented to the searcher. For example, assume that three different types of 2D fingerprint are available. A search is carried out using the first fingerprint-type to describe the reference structure and each of the database structures, and the database ranked in decreasing order of the computed similarity. The procedure is repeated using each of the other two types of fingerprint in turn, and the three database rankings are then combined using a fusion rule, e.g., taking the mean rank for each database structure when averaged across the three rankings. Data fusion was first used for similarity searching in the mid-Nineties as discussed in an extensive review by Willett (***130***); analogous techniques are used in docking, where the approach is called *consensus scoring* (***131***).

Early studies of data fusion involved combining searches that were based on different types of structural representation. For example, Ginn *et al*. reported studies involving a wide range of types of representation (2D fingerprints, sets of physicochemical properties, Molecular Electrostatic Potential descriptors, and infra-red spectral descriptors) and of combination rules **(*132, 133*)**. This work, and analogous studies by the Sheridan group **(*122, 134*)**, suggested that fusion could give search outputs that were more robust, in the sense of offering a consistently high level of performance, than those obtainable from the use of a single type of similarity search. More recent work in this area has considered the combination of further types of representation, and the combination of searches that involve different similarity coefficients **(*135, 136*)**.

Thus far, we have considered data fusion to involve a single reference structure but multiple similarity measures, an approach that Whittle *et al*. refer to as *similarity fusion* **(*137*)**. The alternative, *group fusion* approach inverts the relationship between similarity measure and reference structure, so that the multiple searches that are input to the fusion procedure result from using multiple reference structures and a single similarity measure (e.g., the Tanimoto coefficient and 2D fingerprints). This idea seems to have been first reported by Xue *et al*. **(*138*)** and then by Schuffenhauer *et al*. **(*51*)** some time after the initial studies of similarity fusion; however, group fusion appears from the literature to have become much more widely used. Its popularity dates from a study by Hert *et al*. **(*123*)** who found that fusing the similarity rankings obtained from as few as ten reference structures enabled searches to be carried out that were comparable to even the very best from amongst many hundreds of conventional similarity searches using individual reference structures. Subsequent studies demonstrated the general validity of the approach, and it has now been widely adopted **(*139, 140*)**.

Hert *et al*. have also described a modification of conventional similarity searching that makes use of group fusion **(*141, 142*)**. A similarity search is carried out in the normal way using a single reference structure, and the nearest neighbours identified. The assumption is then made that they also are active, as is likely to be the case if the Similar Property Principle applies to the search. Each of these nearest neighbours is used in turn as a reference structure for a further similarity search, and the complete set of rankings (one from the original reference structure and one from each of the nearest neighbours) is then fused to give the final output ranking. This *turbo similarity searching* approach resulted in searches that were nearly always superior to conventional similarity searching (where just the initial reference structure is used) in its ability to identify active molecules, although performance appears to be crucially dependent on the effectiveness of the initial search based on the original reference structure **(*143*)**.

Most studies of fusion methods have found that they seem to work well in practice but have not provided any rationale for why this might be so (***130***). Two studies have addressed this question. An empirical study by Baber *et al*. (***144***) showed that active molecules are more tightly clustered than are inactive molecules (as would indeed be expected if the Similar Property Principle holds). Thus, when multiple scoring functions are used in similarity fusion, they are likely to repeatedly select many actives but not necessarily the same inactives, providing an enrichment of actives at the top of the final fused ranking. Whittle *et al*. provide a rigorous theoretical approach to the modelling of data fusion (***145, 146***). Their model suggests that the origin of performance enhancement for simple fusion rules can be traced to a combination of differences between the retrieved active and retrieved inactive similarity distributions and the geometrical difference between the regions of these multivariate distributions that the chosen fusion rule is able to access. Although their model gave predictions in accord with experimental data, it was concluded that improvements over conventional similarity searching would be obtained only if large amounts of training data are available; however, this is not normally the case in the early stages of drug-discovery programmes where similarity searching is most commonly used.

Group fusion requires multiple reference structures but the processing involves them being treated on an individual basis, with each one generating their own similarity ranking. It is arguable that this wastes available information since it takes no account of the relationships between the reference structures, as reflected in the bits that are, and that are not, set in their fingerprints. This is valuable information that can be correlated with the other information that we have available, i.e., that these reference structures are known to exhibit the activity that is being sought in the similarity search. Put simply, if a bit is set in many of the reference structures' fingerprints, then it seems likely that the corresponding 2D fragment is positively associated with the activity of interest, and this information can be used to enhance the effectiveness of a similarity search.

The relationship between fragment occurrences and bioactivity in large databases was first studied by Cramer *et al*. (***147***). Their *substructural analysis* approach (***148-151***) and the closely related *naïve Bayesian classifier* (***82, 142, 152-154***) are widely used examples of the application of machine learning methods to virtual screening (***97***). These applications require considerable amounts of training: this is normally HTS data that contains many examples of both active and inactive molecules. The use of such approaches for similarity searching typically uses training data based on the set of reference structures (for the actives) and on any large set of molecules from which the known actives have been removed (for the inactives). One example of this approach is the MOLPRINT system of Bender *et al*. (***82, 155***), who have used a naïve Bayesian classifier with atom-centred substructures chosen using a feature selection algorithm. However, the largest body of work in this area has been carried out by the

Bajorath group, who have used a Bayesian approach to derive functions that relate the probability of a molecule exhibiting bioactivity to the statistical distributions of the descriptor values for that molecule's descriptors (***156***). The procedure involves estimating the probability that a molecule will be active given a particular value of a descriptor, where the descriptor can be binary (as with a bit in a fingerprint) or non-binary (as with a molecular property). The probabilities of activity for different descriptors are assumed to be statistically independent, and it is hence possible to compute the overall probability of activity (or inactivity) for a molecule by taking the product of the individual descriptor probabilities. It should be noted that the independence assumption is generally incorrect (indeed, it is naïve, which is why the approach is called a naïve Bayesian classifier) but has been found to work well in practice. The overall approach is markedly more complex than with group fusion, where the reference structures are used for individual similarity searches; however, detailed comparisons suggest the greater search effectiveness of the Bayesian approach (***157***). An interesting application of this work is the ability to predict the probability that a similarity approach will be able to identify novel molecules that exhibit the reference structures' bioactivity when searching a particular database: if this probability is low then it may be worth considering an alternative type of structure representation for the search (***156***). Other recent studies by this group have included: ways of weighting the bits in fingerprints (***158***); the use of quantitative, rather than qualitative, bioactivities for the training data (***159***); and the use of a different machine learning tool, a support vector machine, for similarity searching (***160***).

We have thus considered two ways of using multiple reference structures: combining rankings based on each structure in turn (group fusion), and combining information about the bits that are and are not set in the structures' fingerprints. There is a much simpler approach, involving the combination of the multiple reference structures' fingerprints into a single, combined fingerprint (***51, 161***); however, this appears to be less effective than the other two approaches (***123, 162***). There is also a considerably more complex approach, which involves combining the actual chemical graphs of the reference structures (rather than fingerprints derived from those graphs) (***163***); however, this hardly comes within the scope of a review of fingerprint-based methods

6. Conclusions

Similarity searching of chemical databases using 2D structural fingerprints was first described almost a quarter of a century ago. Since that time, it has established itself as one of the most valuable ways of accessing a chemical database to identify novel bioactive molecules, providing a natural complement to the long-established systems for 2D substructure searching. It is now routinely used in the initial stages of virtual screening programmes, where very little structure-activity data may be available at the start of a research project, and has proved to be remarkably effective in this role, despite the inherent simplicity of the methods that are being used. There are very many different types of similarity measure that can be used to determine the similarity between a pair of molecules: at present, the Tanimoto coefficient and binary fingerprints are the method of choice, but it would be surprising if it did not prove possible to identify more effective ways of searching, e.g., using some type of fragment weighting scheme. Current research in similarity searching is looking at ways of exploiting the information that is available when multiple reference structures are available.

References

1. Rouvray, D. H. (1990) The evolution of the concept of molecular similarity, in *Concepts and Applications of Molecular Similarity* (Johnson, M. A., and Maggiora, G. M., Eds.), pp 15-42, John Wiley, Chichester.
2. Bender, A., and Glen, R. C. (2004) Molecular similarity: a key technique in molecular informatics, *Organic and Biomolecular Chemistry* ***2***, 3204-3218.
3. Dean, P. M., (Ed.) (1994) *Molecular Similarity in Drug Design*, Chapman and Hall, Glasgow.
4. Downs, G. M., and Willett, P. (1995) Similarity searching in databases of chemical structures, *Reviews in Computational Chemistry* ***7***, 1-66.
5. Maldonado, A. G., Doucet, J. P., Petitjean, M., and Fan, B.-T. (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications, *Molecular Diversity* ***10***, 39-79.
6. Nikolova, N., and Jaworska, J. (2003) Approaches to measure chemical similarity - a review, *Quantitative Structure-Activity Relationships and Combinatorial Science* ***22***, 1006-1026
7. Sheridan, R. P., and Kearsley, S. K. (2002) Why do we need so many chemical similarity search methods?, *Drug Discovery Today* ***7***, 903-911.
8. Alvarez, J., and Shoichet, B., (Eds.) (2005) *Virtual Screening in Drug Discovery*, CRC Press, Boca Raton.
9. Bajorath, J. (2002) Integration of virtual and high-throughput screening, *Nature Reviews Drug Discovery* ***1***, 882-894.
10. Böhm, H.-J., and Schneider, G., (Eds.) (2000) *Virtual Screening for Bioactive Molecules*, Wiley-VCH, Weinheim.
11. Klebe, G., (Ed.) (2000) *Virtual Screening: an Alternative or Complement to High Throughput Screening*, Kluwer, Dordrecht.
12. Lengauer, T., Lemmen, C., Rarey, M., and Zimmermann, M. (2004) Novel technologies for virtual screening, *Drug Discovery Today* ***9***, 27-34.
13. Oprea, T. I., and Matter, H. (2004) Integrating virtual screening in lead discovery, *Current Opinion in Chemical Biology* ***8***, 349-358.
14. Gedeck, P., Rhode, B., and Bartels, C. (2006) QSAR - how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets, *Journal of Chemical Information and Modeling* ***46***, 1924-1936.
15. McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.-F., and Cornell, W. D. (2007) Comparison of topological, shape, and docking methods in virtual screening, *Journal of Chemical Information and Modeling* ***47***, 1504-1519.
16. Sheridan, R. P. (2007) Chemical similarity searches: when is complexity justified?, *Expert Opinion on Drug Discovery* ***2***, 423-430.

17. Sheridan, R. P., McGaughey, G. B., and Cornell, W. D. (2008) Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results, *Journal of Computer-Aided Molecular Design* ***22***, 257-265.
18. Talevi, A., Gavernet, L., and Bruno-Blanch, L. E. (2009) Combined virtual screening strategies, *Current Computer-Aided Drug Design* ***5***, 23-37.
19. Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006) A critical assessment of docking programs and scoring functions, *Journal of Medicinal Chemistry* ***49***, 5912-5931.
20. Wilton, D., Willett, P., Lawson, K., and Mullier, G. (2003) Comparison of ranking methods for virtual screening in lead-discovery programs, *Journal of Chemical Information and Computer Sciences* ***43***, 469-474.
21. Bajorath, J., (Ed.) (2004) *Chemoinformatics. Concepts, Methods and Tools for Drug Discovery*, Humana Press, Totowa NJ.
22. Gasteiger, J., and Engel, T., (Eds.) (2003) *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim.
23. Leach, A. R., and Gillet, V. J. (2007) *An Introduction to Chemoinformatics*, 2nd edition ed., Kluwer, Dordrecht.
24. Gasteiger, J., (Ed.) (2003) *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim.
25. Johnson, M. A., and Maggiora, G. M., (Eds.) (1990) *Concepts and Applications of Molecular Similarity*, John Wiley, New York.
26. Willett, P. (2009) Similarity methods in chemoinformatics, *Annual Review of Information Science and Technology* ***43***, 3-71.
27. Eckert, H., and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches, *Drug Discovery Today* ***12***, 225-233.
28. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints, *Drug Discovery Today* ***11***, 1046-1053.
29. Hagadone, T. R. (1992) Molecular substructure similarity searching - efficient retrieval in two-dimensional structure databases, *Journal of Chemical Information and Computer Sciences* ***32***, 515-521.
30. Senger, S. (2009) Using Tversky similarity searches for core hopping: finding the needles in the haystack, *Journal of Chemical Information and Modeling* ***49***, 1514-1524.
31. Willett, P. (1985) An algorithm for chemical superstructure searching, *Journal of Chemical Information and Computer Sciences* ***25***, 114-116.
32. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom pairs as molecular-features in structure activity studies - definition and applications, *Journal of Chemical Information and Computer Sciences* ***25***, 64-73.
33. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of nearest-neighbour searching in an online chemical structure search system, *Journal of Chemical Information and Computer Sciences* ***26***, 36-41.
34. Adamson, G. W., and Bush, J. A. (1973) A method for the automatic classification of chemical structures, *Information Storage and Retrieval* ***9***, 561-568.
35. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching, *Journal of Chemical Information and Computer Sciences* ***38***, 983-996.
36. Wilkins, C. L., and Randic, M. (1980) A graph theoretical approach to structure-property and structure-activity correlation, *Theoretica Chimica Acta* ***58***, 45-68.
37. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. (1996) Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors, *Journal of Medicinal Chemistry* ***39***, 3049-3059.
38. Dixon, S. L., and Merz, K. M. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation, *Journal of Medicinal Chemistry* ***44***, 3795-3809.

39. Papadatos, G., Cooper, A. W. J., Kadirkamanathan, V., Macdonald, S. J. F., McLay, I. M., Pickett, S. D., Pritchard, J. M., Willett, P., and Gillet, V. J. (2009) Analysis of neighborhood behaviour in lead optimisation and array design, *Journal of Chemical Information and Modeling* ***49***, 195-208.

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

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

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

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

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

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

46. Steffen, A., Kogej, T., Tyrchan, C., and Engkvist, O. (2009) Comparison of molecular fingerprint methods on the basis of biological profile data *Journal of Chemical Information and Modeling* ***49***, 338-347.

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

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

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

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

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

52. Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I., and Shoichet, B. K. (2008) Quantifying the relationship among drug classes, *Journal of Chemical Information and Modeling* ***48***, 755-765.

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

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

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

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

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

58. Peltason, L., and Bajorath, J. (2007) SAR index: quantifying the nature of structure-activity relationships, *Journal of Medicinal Chemistry* ***50***, 5571-5578.

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

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

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

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

63. Kier, L. B., and Hall, H. L. (1986) *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York.

64. Lowell, H., Hall, H. L., and Kier, L. B. (2001) Issues in representation of molecular structure: The development of molecular connectivity, *Journal of Molecular Graphics and Modelling* ***20***, 4-18 check.

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

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

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

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

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

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

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

72. Mount, J., Ruppert, J., Welch, W., and Jain, A. N. (1999) Icepick: a flexible surface-based system for molecular diversity, *Journal of Medicinal Chemistry* ***42***, 60-66.

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

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

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

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

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

78. Brown, N. (2009) Chemoinformatics - an introduction for computer scientists, in *ACM Computing Surveys*.

79. Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M. (1973) Strategic considerations in the design of screening systems for

substructure searches of chemical structure files, *Journal of Chemical Documentation* ***13***, 153-157.

80. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002) Re-optimisation of MDL keys for use in drug discovery, *Journal of Chemical Information and Modeling* ***42***, 1273-1280.

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

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

83. Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M., and Davies, J. W. (2009) How similar are similarity searching methods? A principal components analysis of molecular descriptor space, *Journal of Chemical Information and Modeling* ***49***, 108-119.

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

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

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

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

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

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

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

91. Martin, Y. C., and Muchmore, S. (2009) Beyond QSAR: lead hopping to different structures, *QSAR & Combinatorial Science* ***28***, 797-801.

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

93. Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J. (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme, *Journal of Chemical Information and Computer Sciences* ***43***, 1151-1157.

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

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

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

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

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

99. Downs, G. M., Poirrette, A. R., Walsh, P., and Willett, P. (1993) Evaluation of similarity searching methods using activity and toxicity data, in *Chemical Structures 2. The International Language of Chemistry*. (Warr, W. A., Ed.), pp 409-421, Springer Verlag, Berlin.

100. Azencott, C.-A., Ksikes, A., Swamidass, S. J., Chen, J. H., Ralaivola, L., and Baldi, P. (2007) One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties, *Journal of Chemical Information and Modeling* ***47***, 965-974.

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

102. Olah, M., Bologa, C., and Oprea, T. I. (2004) An automated PLS search for biologically relevant QSAR descriptors, *Journal of Computer-Aided Molecular Design* ***18***, 437-449.

103. Arif, S. M., Holliday, J. D., and Willett, P. (2009) Analysis and use of fragment occurrence data in similarity-based virtual screening, *Journal of Computer-Aided Molecular Design* ***23***, 655-668.

104. Everitt, B. S., Landau, S., and Leese, M. (2001) *Cluster Analysis*, 4th edition ed., Edward Arnold, London.

105. Gower, J. C. (1982) Measures of similarity, dissimilarity and distance, in *Encyclopaedia of Statistical Sciences* (Kotz, S., Johnson, N. L., and Read, C. B., Eds.), pp 397-405, John Wiley, Chichester.

106. Hubálek, Z. (1982) Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation, *Biological Reviews of the Cambridge Philosophical Society* ***57***, 669-689.

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

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

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

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

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

112. Bradshaw, J. (1997) Introduction to Tversky similarity measure, in *MUG '97 - 11th Annual Daylight User Group Meeting* Laguna Beach CA.

113. Maggiora, G. M., Mestres, J., Hagadone, T. R., and Lajiness, M. S. (1997) Asymmetric similarity and molecular diversity, in *213th National Meeting of the American Chemical Society, April13-17, 1997*, San Francisco, CA.

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

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

116. Wang, Y., and Bajorath, J. (2008) Balancing the influence of molecular complexity on fingerprint similarity searching, *Journal of Chemical Information and Modeling* ***48***, 75-84.

117. Wang, Y., and Bajorath, J. (2009) Development of a compound-class directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching, *Journal of Chemical Information and Modeling* ***49***, 1369-1376.

118. Varin, T., Bureau, R., Mueller, C., and Willett, P. (2009) Clustering files of chemical structures using the Székely-Rizzo generalisation of Ward's method, *Journal of Molecular Graphics and Modelling* ***In press***.

119. Gower, J. C., and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification* ***5***, 5-48.

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

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

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

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

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

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

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

127. Jain, A. N., and Nicholls, A. (2008) Recommendations for evaluation of computational methods, *Journal of Computer-Aided Molecular Design* ***22***, 133-139.

128. Nicholls, A. (2008) What do we know and when do we know it?, *Journal of Computer-Aided Molecular Design* ***22***, 239-255.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

143. Gardiner, E. J., Gillet, V. J., Haranczyk, M., Hert, J., Holliday, J. D., Malim, N., Patel, Y., and Willett, P. (2009) Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance, *Statistical Analysis and Data Mining* ***2***, 103-114.

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

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

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

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

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

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

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

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

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

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

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

155. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors: evaluation of performance, *Journal of Chemical Information and Computer Sciences* ***44***, 1708-1718.

156. Vogt, M., Nisius, B., and Bajorath, J. (2009) Predicting the similarity search performance of fingerprints and their combination with molecular property descriptors using probabilistic and information theoretic modeling, *Statistical Analysis and Data Mining* ***2***, 123-134.

157. Vogt, M., and Bajorath, J. (2008) Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints, *Chemical and Biological Drug Design* ***71***, 8-14.

158. Wang, Y., and Bajorath, J. (2008) Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics, *Journal of Chemical Information and Modeling* ***48***, 1754-1759.

159. Vogt, I., and Bajorath, J. (2007) Analysis of a high-throughput screening data set using potency-scaled molecular similarity algorithms, *Journal of Chemical Information and Modeling* ***47***, 367-375.

160. Geppert, H., Horvath, T., Gartner, T., Wrobel, S., and Bajorath, J. (2008) Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds, *Journal of Chemical Information and Modeling* ***48***, 742-746.

161. Shemetulskis, N. E., Weininger, D., Blankey, C. J., Yang, J. J., and Humblet, C. (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets, *Journal of Chemical Information and Computer Sciences* ***36***, 862-871.

162. Tovar, A., Eckert, H., and Bajorath, J. (2007) Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity, *ChemMedChem* ***2***, 208-217.

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