Chemoinformatics techniques for data mining in files of two-dimensional and three-dimensional chemical molecules.

Peter Willett

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom.
Email: p.willett@sheffield.ac.uk

Abstract. Chemoinformatics is the name given to a body of computer techniques that are used to process information pertaining to the two-dimensional (2D) and three-dimensional (3D) structures of chemical molecules. This paper introduces some of these techniques, starting with those that are used to represent and search for biologically active molecules in the pharmaceutical and agrochemical industries. These industries have created extensive databases of both 2D and 3D structures and a variety of data mining tools are routinely used to support the discovery of novel pharmaceuticals and agrochemicals. Two types of tool are considered here: molecular diversity analysis methods, which ensure that a research programme will consider as wide a range of different types of structure as possible in the search for biological activity; and virtual screening methods, which can rank a database so that synthesis and biological testing can be restricted to those with high *a priori* probabilities of activity.

Keywords: chemical databases; chemoinformatics; molecular diversity analysis; virtual screening

1 Introduction

Many different scientific disciplines (such as synthetic organic chemistry, structural biology, pharmacology and toxicology) are needed to discover the new drugs that are the lifeblood of the pharmaceutical industry. The huge costs and extended timescales that characterise the industry mean that it is willing and able to make very substantial investments in any technology that can increase the speed with which drugs, *i.e.*, novel chemical molecules with beneficial biological properties, are brought to the market place (and similar comments apply to the herbicides, insecticides and fungicides developed by the agrochemicals industry). One such technology is what is increasingly referred to as *chemoinformatics*. This term was introduced by Brown, who stated that "The use of information technology and management has become a crucial part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization" [1]. This is clearly a very broad definition, covering as it does all aspects of information technology and information management: here we focus on one of the most important areas of chemoinformatics, the techniques that are used to process information pertaining to the two-dimensional (2D) and three-dimensional (3D) structures of chemical molecules.

It is only within the last few years that chemoinformatics has come to be recognised as a distinct topic of study [2-5], this prominence arising principally as a result of technological developments in chemistry and biology. Specifically, the methods of *combinatorial chemistry* and *high-throughput screening* allow the synthesis and biological testing, respectively, of huge arrays of molecules in parallel. Taken together, these developments have resulted in a data explosion that has spurred the development of sophisticated informatics and data analytic methods. This paper provides an introduction to some of the techniques used in modern chemoinformatics systems, focusing on tools that are available for data mining in files of 2D and 3D chemical structures.

2 Representation and searching of chemical structures

The principal method of representation for a 2D chemical structure diagram is a labelled graph (called a *connection table*) in which the nodes and edges of a graph represent the atoms and bonds, respectively, of a molecule. A chemical database can hence be represented by a large number of such graphs, with searching historically being carried out using two types of graph isomorphism algorithms. *Structure searching* involves an exact-match search of a chemical database for a specific query structure: this is required, for example, to retrieve the biological assay results and the synthetic details associated with a particular molecule. Such a search involves a graph isomorphism search, in which the graph describing the query molecule is checked for isomorphism (or structural equivalence) with the graphs of each of the database molecules. *Substructure searching* involves a partial-match search of a chemical database to find all those molecules that contain a user-defined query substructure, irrespective of the environment in which that substructure occurs; for example, a user interested in antibiotics might wish to search a database to find all molecules that contain the characteristic penicillin ring nucleus. A typical search output is illustrated in Figure 1.

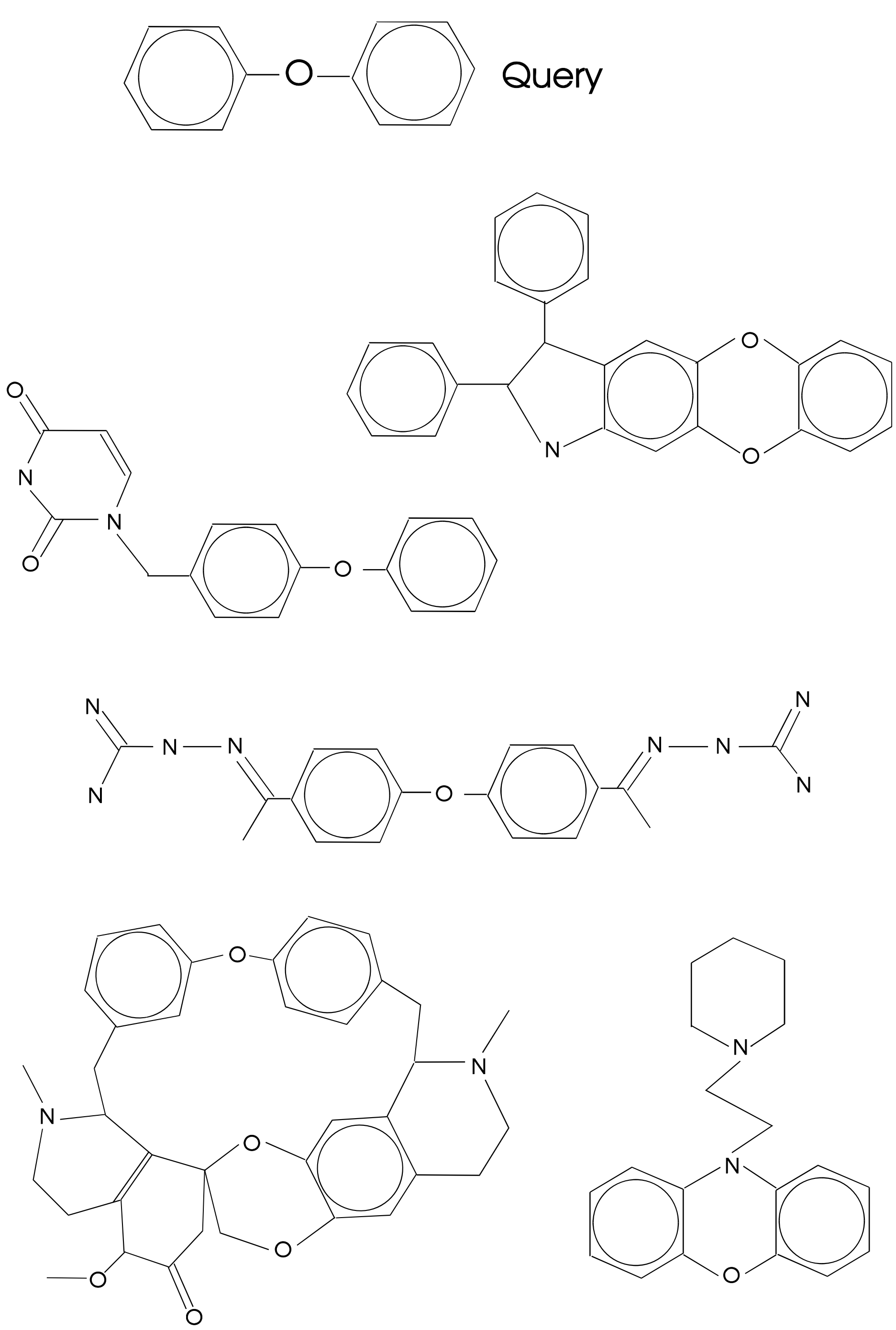

Figure 1. Example of a 2D substructure search. The search is for the diphenyl ether query substructure at the top of the figure, below which are shown five of the hits resulting from a search of the National Cancer Institute database of molecules that have been tested in the US government anti-cancer programme (see URL http://dtp.nci.nih.gov/). This database is also used for the searches described in Figures 2 and 3.

A substructure search involves checking the graph describing the query substructure for subgraph isomorphism (or structural inclusion) with the graphs of each of the database molecules [6]. However, subgraph isomorphism is known to belong to the class of NP-complete computational problems, and thus substructure searching in databases of non-trivial size might be expected to be computationally infeasible. It is made possible by the use of an initial *screen search*, where a screen is a substructural feature, the presence of which is necessary, but not sufficient, for a molecule to contain the query substructure. These features are typically small, atom-, bond- or ring-centred fragment substructures that are algorithmically generated from a connection table when a molecule is added to the database that is to be searched. One common approach to screening involves listing the fragments that have been chosen for use as screens in a fragment coding dictionary, which will typically contain a few hundred or a few thousand carefully selected fragments. Each of the database structures is analysed to identify those screens from the coding dictionary that are present, and the structure is then represented for search by a fixed-length bit-string in which the non-zero bits correspond to the screens that are present. The query substructure is subjected to the same process and the screen search then involves checking the bit-strings representing each database structure for the presence of the screens that are encoded in the bit-string representing the query substructure. Only a very small fraction of a database will normally contain all of the screens that have been assigned to a query substructure, and thus only these few molecules need to undergo the final, time-consuming graph-matching search. This checks to see whether there is an exact subgraph isomorphism between the graph representing the query substructure and the graphs representing each of the database structures that have passed the screen search. This simple, two-stage procedure (*i.e.*, screen searching and subgraph searching) has formed the basis for most operational 2D substructure searching systems to date.

Similar techniques are used for 3D substructure searching [7], where there is a need to identify molecules that contain a query *pharmacophore*. A pharmacophore, or *pharmacophoric pattern*, is a set of atoms having some specific geometric relationship to each other (as illustrated by the anti-leukemic pharmacophore [8] shown in Figure 2). Here, the nodes and edges of a chemical graph denote the atoms and the inter-atomic distances, and the fragments that are encoded in the bit-strings describe pairs or triplets of atoms and the associated inter-atomic distances. Only simple modifications to the 2D methods described previously are required to enable searches for pharmacophores to be carried out, such as that shown in Figure 2. However, significant complexities needed to be overcome before these representations and searching methods were extended to encompass the fact that most molecules are *flexible,* i.e., they adopt not just a single, fixed 3D shape but can exist in some, many, or very many different shapes, depending on the temperature and the external chemical environment. This means that the separation between each pair of atoms is not necessarily fixed, but typically covers a range of possible distances. This increases the complexity of the matching operations that are required; in particular, the screening and subgraph isomorphism searches need an additional, *conformational* search, which takes account of the precise geometries and energies of the various shapes that each potential hit molecule can adopt [9].

Substructure searching, whether in 2D or in 3D, provides an invaluable tool for accessing databases of chemical structures when the searcher already knows the sorts of structures that are expected to be retrieved from the database. This is clearly very difficult at the start of an investigation, when perhaps only one or two active structures have been identified and when it is not at all clear which particular feature(s) within them are responsible for the observed activity. *Similarity searching* has been developed to address this problem, and as a complement to substructure searching [10]. Similarity searching requires the specification of an entire *target* structure (or *reference* structure), rather than the partial structure that is

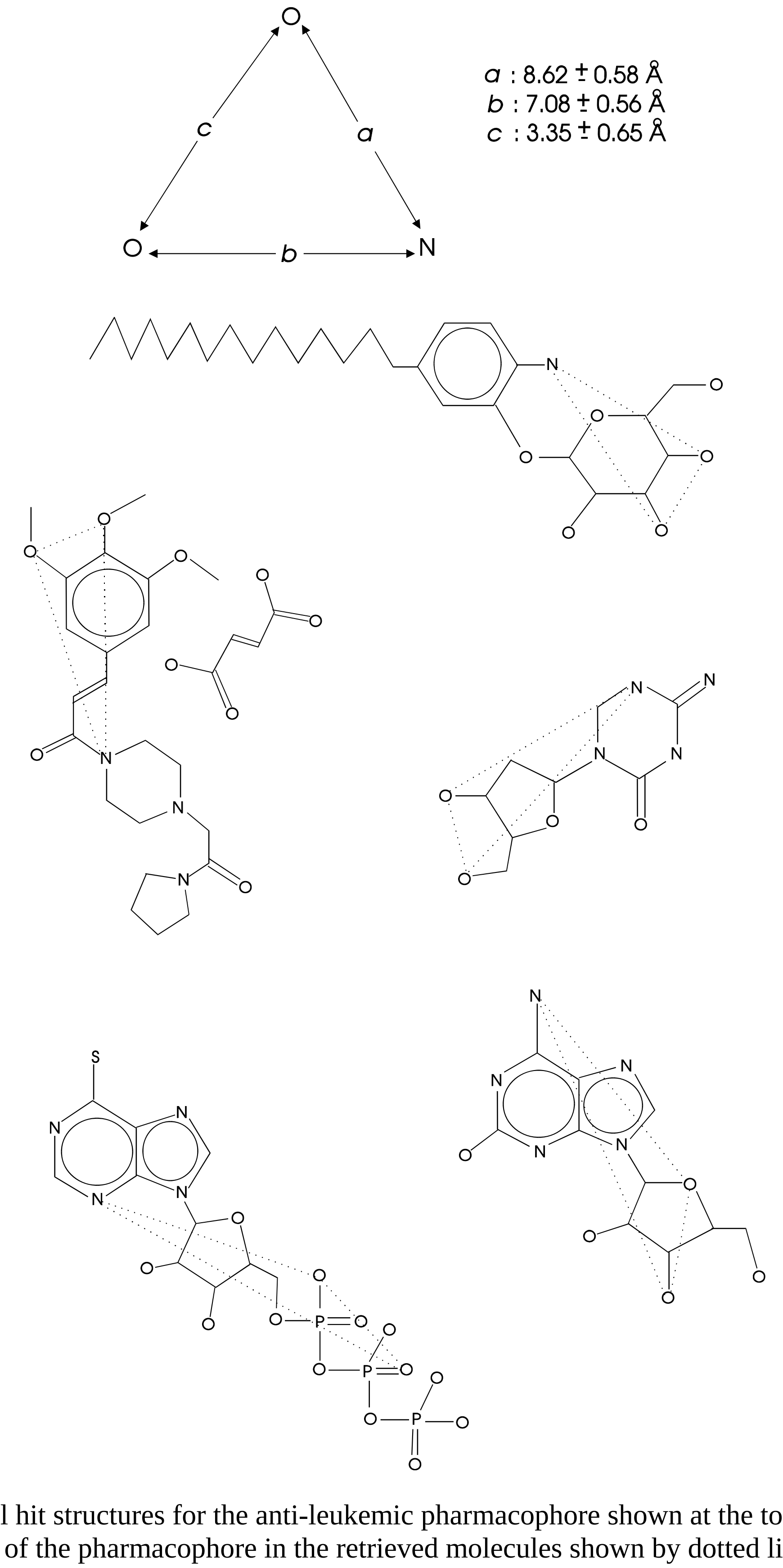

Figure 2. Typical hit structures for the anti-leukemic pharmacophore shown at the top of the page, with the presence of the pharmacophore in the retrieved molecules shown by dotted lines.

required for substructure searching. The target molecule is characterised by a set of structural features, and this set is compared with the corresponding sets of features for each of the database structures. Each such comparison enables the calculation of a measure of similarity between the target structure and a database structure, and the database is then sorted into order of decreasing similarity with the target. The output from the search is a ranked list, where the structures that the system judges to be most similar to the target structure are located at the top of the list, and are hence the first to be presented to the searcher.

An effective similarity searching system requires an appropriate way of quantifying the degree of structural resemblance between the target structure and each of the structures in the database that is to be searched [10-13]. There are many such *similarity measures* but by far the most common are those obtained by comparing the fragment bit-strings that are used for 2D substructure searching, so that two molecules are judged as being similar if they have a large number of bits, and hence substructural fragments, in common. A normalised association coefficient, typically the Tanimoto coefficient, is used to give a numeric value to the similarity between the target structure and each database structure. If these structures have *A* and *B* bits set in their fragment bit-strings, with *C* of these in common, then the Tanimoto coefficient is defined to be

$$\frac{C}{A+B-C}.$$

The value of the Tanimoto Coefficient for bit-string similarities lies in the range of zero (no bits in common) to unity (all bits the same); a more complex version of the Coefficient is available for handling non-binary data [10]. An example of a 2D similarity search based on the Tanimoto Coefficient is shown in Figure 3. While fragment-based measures such as the Tanimoto coefficient provide a simple (indeed simplistic) picture of the similarity relationships between pairs of molecules, they are both efficient (since they involve just the application of logical operations to pairs of bit-strings) and effective (since they have been shown to be capable of bringing together molecules that are judged by chemists to be structurally similar to each other) in operation. The latter characteristic is most surprising, given that the fragments that are used for the calculation of the similarities were originally designed to maximise the efficiency of substructure searching, not the effectiveness of similarity searching. Moreover, they describe only the 2D structures of molecules, and take only implicit account of the 3D structures, which are known to be of crucial importance in determining physical, chemical and biological properties. It should be noted here that there is much current interest in measures of 3D similarity based on fingerprints that encode the geometric arrangement of atom triplets or atom quartets [14]. thus far, however, such approaches have not, been found to be as generally effective as the simpler 2D measures for database applications [15, 16]. Methods for the representation and searching of molecular surfaces and molecular fields are also under active investigation [12, 17-20].

It will be seen from Figure 3 that there is a close family relationship between the target structure and its nearest neighbours. This is of potential value in the search for novel bioactive molecules because of the *Similar Property Principle* [21], which states that molecules that have similar structures will have similar properties. Hence, if the target structure has some interesting property, e.g., it lowers a person's cholesterol level or alleviates the symptoms of a migraine attack, then molecules that are structurally similar to it are more likely to exhibit that property than are molecules that have been selected from a database at random. The Principle is clearly only an approximation that does not hold in all cases [22], but it does provide a rational basis for similarity-based access to chemical databases.

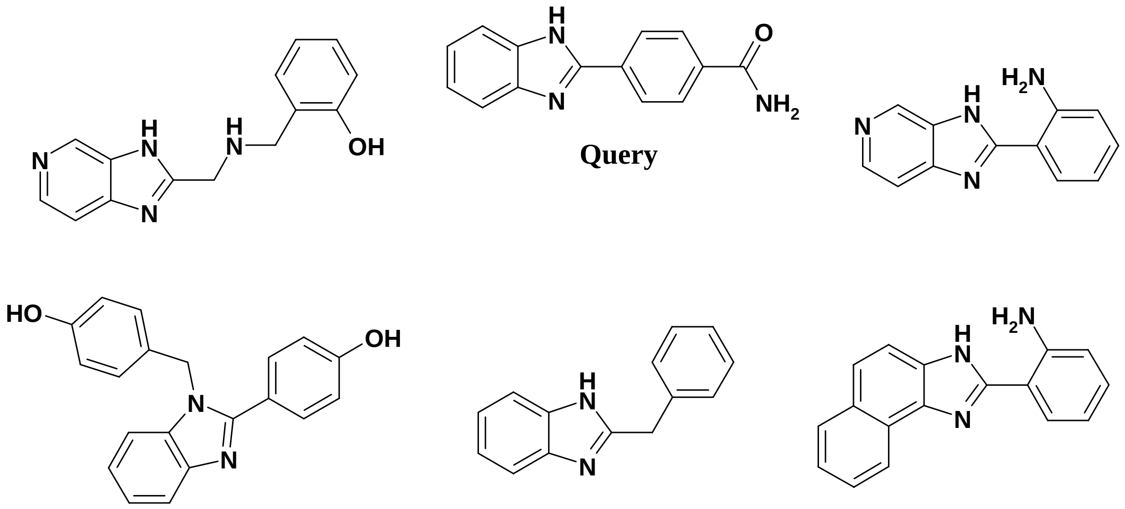

Figure 3. Example of a 2D similarity search, showing a query molecule and five of its nearest neighbours. The similarity measure for the search is based on fragment bit-strings and the Tanimoto coefficient.

Having described the basic searching methods available to access a database of 2D or 3D molecules to find those that have particular structural characteristics, we now discuss some of the other data mining techniques that can be applied to such databases. Specifically, we discuss two techniques: *molecular diversity analysis* methods [23, 24], which ensure that a research programme will consider as wide a range of different types of structure as possible in the search for biological activity; and *virtual screening* methods [25-28], which can rank a database so that synthesis and biological testing can be restricted to those with high *a priori* probabilities of activity.

3 Molecular diversity analysis methods

Molecular diversity analysis is the name given to a body of techniques that seek to enhance the cost-effectiveness of drug discovery by maximising the diversity of the molecules that are submitted for biological testing (rather than maximising the probability of activity, which is the main aim of the virtual screening techniques discussed in Section 4). We have noted above that structurally similar molecules are likely to give similar biological responses; thus, to maximise the structure-activity information that can be gained from a fixed number of molecules, one should try to ensure that the molecules submitted for testing should be as structurally diverse as possible. This requirement may sound like a statement of the obvious, but the practical realisation of this has proved to be very difficult.

The inherently subjective concept of diversity is normally quantified using similarity-based techniques that are a natural development of those discussed previously: thus, a diverse subset of the molecules in a database is selected by consideration of their inter-molecular structural similarities, typically as determined by use of fragment bit-strings and the Tanimoto coefficient. There is a trivial algorithm available to identify the most diverse *n*-compound subset of an *N*-compound database (where, typically, *n*<<*N*). This algorithm involves generating each of the

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

possible subsets and calculating their diversities using a *diversity index* (some function of the inter-molecular similarities in the chosen subset): the optimal subset is then that group of compounds that has the greatest value of the diversity index. The problem is that the factorials in the expression above mean that there is an astronomical number of possible subsets that can be generated from a database of non-trivial size: it is hence infeasible to consider all of them so as to identify the most diverse subset. There has thus been much interest in alternative approaches for selecting diverse sets of molecules that maximise the coverage of structural space, whilst minimising the numbers of molecules put forward for testing. Here, we will exemplify this work by consideration of two of the approaches that have been used: clustering and dissimilarity-based compound selection (DBCS).

Cluster analysis, or *clustering*, is the process of subdividing a group of objects (chemical molecules in the present context) into groups, or *clusters*, of objects that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity [29, 30]. It is thus possible to obtain an overview of the range of structural types present within a dataset by selecting one, or some small number, of the molecules from each of the clusters resulting from the application of an appropriate clustering method to that dataset. The representative molecule (or molecules) for each cluster is either selected at random or selected as being the closest to the centre of that cluster, and this representation is then put forward for biological testing. Very many different clustering methods have been described in the literature, and a considerable amount of effort has gone into comparing the effectiveness of the various methods for clustering chemical structures (see, *e.g.*, [31]). Clustering methods can produce *overlapping* clusters, in which each object may be in more than one cluster, or *non-overlapping* clusters, in which each object occurs in only one cluster. Of these, the latter are far more widely used, and are of one of two types: *hierarchical* methods and *non-hierarchical* methods. An hierarchical clustering method produces a classification in which small clusters of very similar molecules are nested within larger and larger clusters of less closely-related molecules. The classification is normally generated by means of an *agglomerative* procedure: this generates a classification in a bottom-up manner, by a series of agglomerations (or fusions) in which small clusters, initially containing individual molecules, are fused together to form progressively larger clusters.

There are many hierarchic agglomerative methods, all of which can be implemented by means of the basic algorithm shown in Figure 4, where a *point* is either a single molecule or a cluster of molecules. This procedure is known as the *stored matrix* algorithm since it involves random access to the inter-molecular similarity matrix throughout the entire cluster-generation process. Individual hierarchical agglomerative methods differ in the ways in which the most similar pair of points is defined and in which the merged pair is represented as a single point. Although simple in concept, the algorithm is demanding of both computer time and computer storage and more efficient algorithms are available for specific methods. Thus, the well-known Ward's method can be implemented by what is known as the *reciprocal nearest neighbour* (RNN) algorithm. In this, a path is traced through the similarity space until a pair of points is reached that are more similar to each other than they are to any other points, *i.e.*, they are RNNs. These RNN points are fused to form a single new point, and the search continues until the last unfused point is reached. The basic RNN algorithm is thus as shown in Figure 5, where NN(X) denotes the nearest neighbour for the point X, and the final hierarchy is then created from the list of RNN fusions that has taken place.

1. Calculate the inter-molecular similarity matrix.
2. Find the most similar pair of points in the matrix and fuse them into a cluster to form a new single point.
3. Calculate the similarity between the new point and all remaining points.
4. Repeat Steps 2 and 3 until only a single point remains, *i.e.*, until all of the molecules have been fused into one cluster.

Figure 4. Stored matrix algorithm for hierarchic agglomerative clustering methods.

1. Mark all molecules, *I*, as unfused.
2. Starting at an unfused *I*, trace a path of unfused nearest neighbours (NN) until a pair of RNNs is encountered, *i.e.*, trace a path of the form J := NN(I), K := NN(J), L := NN(K)..... until a pair is reached for which Q = NN(P) and P = NN(Q).
3. Add the RNNs *P* and *Q* to the list of RNNs along with the distance between them, mark *Q* as fused and replace the centroid of *P* with the combined centroid of *P* and *Q*.
4. Continue the NN-chain from the point in the path prior to *P*, or choose another unfused starting point if *P* was a starting point.
5. Repeat Steps 2-4 until only one unfused point remains.

Figure 5. Reciprocal nearest neighbours algorithm for hierarchic agglomerative clustering methods.

```
1.  Identify the top-$K$ nearest neighbours for each of the $N$ molecules in the dataset.
2.  Create an $N$-element array, Label, that contains a cluster label for each of the $N$
        molecules in the dataset.  Initialise Label by setting each element to its array
    position, thus assigning each molecule to its own initial cluster;
3.  For each pair of molecules, $I$ and $J$ ($I < J$)
        If they have at least $K_{min}$ of their top-$K$ nearest neighbours in common
        and each is in the top-$K$ nearest-neighbour list of the other
        then replace all occurrences of the Label entry for $J$ with the Label entry for $I$.
4.  The members of each cluster then all have the same entry in the final Label.
```

Figure 6. Algorithm for the Jarvis-Patrick clustering method.

Once the cluster hierarchy has been produced, some means is required to identify a set of clusters from which molecules can be selected. This is normally achieved by applying a threshold similarity to the hierarchy and identifying the clusters present in the resulting partition (*i.e.*, a set of non-overlapping groups having no hierarchical relationships between them) of the dataset.

A non-hierarchical method, conversely, generates a partition of a dataset directly. There is a combinatorial number of possible partitions, making a systematic evaluation of them totally infeasible, and many different heuristics have thus been described to allow the identification of good, but possibly sub-optimal, partitions [29-31]. An example is the *Jarvis-Patrick nearest-*

neighbour method, which is much less demanding of computational resources than the hierarchical methods and which has been extensively used for clustering chemical databases.

The Jarvis-Patrick method, which is detailed in Figure 6, involves the use of a list of the top K nearest neighbours for each molecule in a dataset, *i.e.*, the K molecules that are most similar to it. Once these lists have been produced for each molecule in the dataset that is to be processed, two molecules are clustered together if they are nearest neighbours of each other and if they additionally have some minimal number of nearest neighbours, K_{min}, in common. The user has to specify the value of K_{min}, and it is generally necessary to experiment with a range of K_{min} values until roughly the required number of clusters is obtained. Many variants of this basic approach have been described in the literature.

Dissimilarity-based methods seek to identify a subset comprising the n most diverse molecules in a dataset containing N molecules (where, typically, $n << N$). However, as noted above, the astronomical number of such subsets means that heuristic, and sub-optimal, approaches need to be considered. Thus far, two major classes of algorithm have been described: *maximum-dissimilarity* algorithms and *sphere-exclusion* algorithms [32].

The basic maximum-dissimilarity algorithm for selecting a size-*n Subset* from a size-*N Dataset* is shown in Figure 7. This algorithm permits many variants depending upon the precise implementation of Steps 1 and 3. Possible mechanisms for the choice of the initial compound in Step 1 include: choosing a compound at random; choosing that compound that is most dissimilar to the other compounds in *Dataset*; or choosing that compound that is nearest to the centre (in some sense) of *Dataset*, *inter alia*. Step 3 in the figure requires a quantitative definition of the dissimilarity between a single compound in *Dataset* and the group of compounds that comprise *Subset*, so that the most dissimilar molecule can be identified in each iteration of the algorithm. There are several ways in which "most dissimilar" can be defined, with each definition resulting in a different version of the algorithm and hence in the selection of a different subset (much as different hierarchic agglomerative clustering methods result from the use of different similarity criteria in the stored-matrix algorithm of Figure 4).

The alternative sphere-exclusion approach involves the specification of a threshold dissimilarity *t*, which can be thought of as the radius of a hypersphere in multi-dimensional chemistry space. A compound is selected, either at random or using some rational basis, for inclusion in *Subset* and the algorithm then excludes from further consideration all those other compounds within the sphere centred on that selected compound, as shown in Figure 8. Many variants are again possible, depending upon the manner in which Stage 2 is implemented. Thus, one can choose that molecule that is most dissimilar to the existing *Subset*, in which case different results will be obtained (as with the maximum dissimilarity algorithms) depending upon the dissimilarity definition that is adopted.

Cluster-based and dissimilarity-based algorithms of the sort discussed here are now widely used to select structurally heterogeneous sets of compounds for input to biological screening programmes. Increasingly, the compounds are selected not just on the basis of chemical diversity but also on the basis of other characteristics (such as cost, pharmacokinetic properties, and ease of synthesis) that are necessary if a molecule is to be considered seriously as a potential drug.

1. Initialise *Subset* by transferring a compound from *Dataset*.
2. Calculate the dissimilarity between each remaining compound in *Dataset* and the compounds in *Subset*.
3. Transfer to *Subset* that compound from *Dataset* that is most dissimilar to *Subset*.
4. Return to Step 2 if there are less than *n* compounds in *Subset*.

Figure 7. Dissimilarity-based compound selection using a maximum-dissimilarity algorithm.

1. Define a threshold dissimilarity, *t*.
2. Transfer a compound, *J*, from *Dataset* to *Subset*.
3. Remove from *Dataset* all compounds that have a dissimilarity with *J* of less than *t*.
4. Return to Step 2 if there are compounds remaining in *Dataset*.

Figure 8. Dissimilarity-based compound selection using a sphere-exclusion algorithm.

1. Execute a similarity search of a chemical database for some particular target structure using two, or more, different measures of inter-molecular similarity.
2. Note the rank position, r_i, or the similarity score, s_i, of each individual database structure in the ranking resulting from use of the *i*-th similarity measure.
3. Combine the various rankings using a fusion rule to give a new combined score for each database structure
4. Rank the resulting combined scores, and then use this ranking to calculate a quantitative measure of the effectiveness of the search for the chosen target structure.

Figure 9. Combination of similarity rankings using data fusion.

1. For each bit-position *j* identify the training-set actives and training-set inactives that have the *j*-th position set and not set; use this information to calculate the weight for bit-position *j* using the chosen weighting scheme.
2. For each molecule *I* sum the weights for all of those bit-positions for which the bit is set.
3. Rank the molecules in the database in decreasing order of the sums-of-weights computed in Step 2.

Figure 10. Ranking compounds by means of substructural analysis

4 Virtual screening methods

The principal aim of discovery research programmes is the identification of a *lead*, a compound that has the desired biological activity (e.g., lowers a person's blood pressure, or reduces the size of a tumour), that has appropriate pharmacokinetic characteristics (e.g., it is soluble and does not metabolise too rapidly) and that does not have any obvious side-effects. Over the years, pharmaceutical companies have built up corporate databases containing hundreds of thousands (or millions) of drug-like compounds, and many millions of similar compounds are now available from commercial suppliers. These repositories provide the obvious starting place in the search for new leads: given the vast numbers of compounds that need to be considered there is much interest in the use of techniques that can rapidly focus-in on that relatively small fraction that has a high *a priori* probability of activity. The identification of such candidate compounds is normally referred to as virtual screening.

In principle, any technique that can rank a database in order of decreasing probability of activity can be used for virtual screening; in practice, the methods available are determined largely by the amounts of structural and biological information that are available. At the heart of most virtual screening methods is the Similar Property Principle that has been mentioned previously: if some molecule or molecules are known to exhibit the biological activity of interest then a sensible virtual screening strategy is to find other molecules that are structurally similar to the known active(s). The most obvious approach is hence to use similarity searching of a database, using as the target structure any one molecule that is already known to be active; this could either be a hit from initial biological screening or a compound from the published literature, e.g., one specified in a competitor's patent. As exemplified in Figure 3, the nearest-neighbours retrieved by the search will contain many substructural features in common with the target structure, and are hence obvious candidates for biological testing [10, 15, 33]. We are currently studying the use of *data fusion* to increase the performance of similarity-based virtual screening [34]. Here, we combine rankings resulting from several different measures of structural similarity to give a single, combined ranking as the output of the search, as shown in Figure 9. For example, one could carry out similarity searches using different types of 2D fingerprint, or different types of similarity coefficient; we have found that such combined searches often result in a level of search effectiveness that is better than that resulting from a conventional similarity search using just a single similarity measure.

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

Once the (in)activities of a fair number of molecules have been established, it becomes possible to use techniques from the area of computer science known as *machine learning*. These techniques assume the availability of a *training-set*, i.e., sets of both known active and known inactive molecules that can be used to develop a tool that can be applied to molecules of unknown activity, the *test-set*, and predict their (in)activities with a fair degree of confidence. The best-established approach is called *substructural analysis* [35], which is based on the assumption that a given substructural feature makes some fixed contribution to the overall activity of a molecule, irrespective of the other substructures that are present in that molecule. This is likely to be a very drastic assumption but one that, if accepted, enables the calculation of weights that relate the presence of a molecular feature to the probability that a molecule containing it is biologically active. A (very simple) example of such a weight might be as follows: assume that the *j*-th fragment in a 2D fragment bit-string occurs in A_j active and I_j inactive molecules; then a plausible weight would be

$$\frac{A_j}{A_j + I_j}.$$

There are many such schemes that have been described in the literature, differing in the precise ways that they use the fragment-occurrence data for the training-set molecules [36]. Whichever weighting scheme is used, the weights are calculated for each of the fragments present in a set of molecules. These fragments are often those encoded in the bit-strings that are used for 2D substructure and/or similarity searching, in which case the weights are obtained by analysis of the fingerprints for the training-set molecules. The resulting weights are then used to select new compounds for biological testing: specifically, the sum is calculated of the weights for the fragment-substructures that are present in a molecule, and the compounds are sorted into decreasing order of the sums-of-weights (see Figure 10). The top-ranked molecules in the resulting sorted list are those that have the greatest likelihood of activity (if it is assumed that the structural characteristics of the test-set are not too different from those of the training-set). Substructural analysis provides a simple but surprisingly effective way of rationalising large volumes of structural and activity information so as to produce meaningful rankings of the as yet untested compounds in a database. Many new methods for machine learning are now becoming available, and some of these methods are starting to be applied to the virtual screening problem, e.g., recent work on support vector machines and kernel discrimination methods [37, 38].

Substructural analysis requires information about the 2D (or 3D) structures of known active and known inactive molecules. The final virtual screening approach to be discussed here, docking, additionally requires information about one of the biological pathways that is associated with the illness for which a therapy is required. Specifically, docking assumes that a 3D structure has been obtained, typically by X-ray crystallography, of the biological receptor that is involved in the pathway, such as the active site of an enzyme. The "lock-and-key" theory of drug action assumes that a drug fits into a biological receptor in much the same way as a key fits a lock; thus, if the shape of the lock is known, one can identify potential drugs by scanning a 3D database to find those molecules that have shapes that are complementary to the shape of the receptor.

Shape matching is a computationally demanding task for which many algorithmic approaches have been suggested [39]. The original description of docking, by Kuntz *et al.* [40], considered the fitting of just a single molecule into a protein active site; however, it was soon realised that if this fitting operation was repeated for all of the molecules in a database then docking could provide a highly sophisticated approach to virtual screening. In fact, two types of computational procedure are required for docking: a search algorithm that can explore the space of possible protein-ligand geometries; and a scoring function that is used to evaluate the likelihood of each possible geometry, so as to identify the most probable geometries, and hence (hopefully) the true binding mode. The same scoring function can also be used to rank geometries from different potential ligands, so that a database can be ranked in order of decreasing goodness of fit with the active site, and hence in decreasing likelihood of activity.

Modern docking systems involve not just matching the geometric characteristics, such as inter-atomic distances, of the database molecules and the target protein, but also chemical considerations such as the extent to which atoms of one type in the drug are compatible with the atoms that they are mapped to in the receptor site. This brings added complexity, in terms of both mechanistic knowledge and the computational power that is required. The computational requirements are increased still further when, as is increasingly the case, account is taken of the fact that molecules and proteins can adopt different shapes; thus, adopting the lock-and-key metaphor, rather than trying to fit a metallic key into metallic lock, one is actually trying to fit together two non-rigid objects. Current systems for virtual screening enable the docking of databases of flexible molecules into a rigid receptor; efficient and effective processing of both types of flexibility is still probably some years away.

It will be realised that the various data mining tools that are used for virtual screening vary considerably in their sophistication and in the associated information and computational requirements. It is thus common for the approaches to be used in sequence: similarity searching is used initially to identify a few actives; these actives are then analysed to generate a pharmacophore model for 3D searching; once a fair amount of testing has been carried out, it is possible to build a training-set for machine learning; and then docking can be used once a 3D structure is available for the biological target. It is also common to use an initial filtering step to eliminate from further consideration any molecules whose physicochemical properties are such as to render them unlikely to be able to act as a drug [41, 42]. Examples of such *drug-likeness* or *drugability* filters include substructure searches to eliminate molecules containing reactive or toxic substructures and analyses of the values of simple properties (such as molecular weight, the octanol-water partition coefficient and the numbers of rotatable bonds, hydrogen-bond donors and acceptors) in known drug molecules.

5. Conclusions

Public and private databases contain the machine-readable structure representations (normally in 2D but increasingly also in 3D) of many millions of chemical molecules.

Chemoinformatics provides a range of tools that can be used for data mining in these files, so as to assist directly in the discovery of novel bioactive molecules. With the increasing costs of drug discovery, it is likely that more use will be made of such tools, with the availability of more powerful software and hardware enabling more accurate predictions of activity, and thus enhancing the cost-effectiveness of research.

References

1. Brown, F. K. Chemoinformatics: what it is and how does it impact drug discovery? *Ann. Report. Med. Chem.* **1998**, *33*, 375.
2. Leach, A. R.; Gillet V. J. *An Introduction to Chemoinformatics*; Kluwer: Amsterdam, **2003**.
3. Gasteiger, J.; Engel, T. Eds. *Chemoinformatics*; Wiley-VCH: Weinheim, **2003**.
4. Gasteiger, J. Ed. *Handbook of Chemoinformatics. From Data to Knowledge*. Wiley-VCH: Weinheim, **2003**.
5. Bajorath, J. Ed. *Chemoinformatics. Concepts, Methods and Tools for Drug Discovery*. *Methods in Molecular Biology, vol. 275*. Humana Press Inc.: Totowa, NJ, **2004**.
6. Barnard, J.M. Substructure searching methods: old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532.
7. Willett, P. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J. Mol. Recognit.* **1995**, *8*, 290.
8. Zee-Cheng, K. Y.; Cheng, C. C. Common receptor-complement feature among some anti-leukemic compounds. *J. Pharm. Sci.* **1970**, *59*, 1630.
9. Martin, Y. C.; Willett, P. (Editors). *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; American Chemical Society: Washington DC, **1998**.
10. Willett, P.; Barnard, J. M.; Downs, G.M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
11. Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903.

12. Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - a review. *QSAR Comb. Sci*. **2003**, *22*, 1006.
13. Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem*. **2004**, *2*, 3204.
14. Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudinere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications. *J. Med. Chem*. **1999**, *42*, 3251.
15. Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572.
16. Matter H.; Potter T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci*. **1999**; *39*, 1211.
17. Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 80.
18. Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000,** *14*, 573.
19. Zauhar, R. J.; Moyna, G.; Tian, L. F.; Li, Z. J.; Welsh, W. J. Shape signatures: A new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003,** *46*, 5674.
20. Rush, T. S., Grant, J. A., Mosyak, L., Nicholls, A. A shape-based 3-D scaffold hopping method and its applications to a bacterial protein-protein interaction. *J. Med. Chem*. **2005**, *48*, 1489.
21. Johnson, M. A.; Maggiora, G.M. Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, **1990**.
22. Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspect. Drug Discov. Design* **1998**, *9-11*, 225.
23. Dean, P. M.; Lewis, R. A. Eds. *Molecular Diversity in Drug Design*; Kluwer: Amsterdam, **1999**.
24. Ghose, A. K.; Viswanadhan, V. N. Eds. *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery*; Marcel Dekker: New York, **2001**.
25. Bohm, H.-J.; Schneider, G. Eds. *Virtual Screening for Bioactive Molecules*. Wiley-VCH: Weinheim, **2000**.
26. Klebe, G. Ed. *Virtual Screening: An Alternative or Complement to High Throughput Screening*; Kluwer: Dordrecht, **2000**.
27. Bajorath, J. Integration of virtual and high throughput screening. *Nature Rev. Drug Discov*. **2002**, *1*, 882.
28. Oprea, T.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol*. **2004**, *8*, 349.
29. Everitt, B. S.; Landau, S.; Leese, M. *Cluster Analysis*; 4th ed.; Arnold: London, **2001**.
30. Arabie, P.; Hubert, L. J.; De Soete, G. Eds. *Clustering and Classification*; World Scientific: Singapore, **1996**.
31. Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, **1987**.
32. Snarey, M.; Terret, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **1997**, *15*, 372.
33. Martin, Y. C.; Kofron, J. L.; Traphagen, L.M. Do structurally similar molecules have similar biological activities? *J. Med. Chem*. **2002**, *45*, 4350.
34. Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Design* **2000**, *20*, 1.
35. Cramer, R. D.; Redl, G.; Berkoff, C.E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem*. **1974**, *17*, 533.
36. Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Activ. Relat.* **1989**, *8*, 115.
37. Wilton, D. J.; Willett, P.; Mullier, G.; Lawson, K. Comparison of ranking methods for virtual screening in lead-discovery programmes. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 469.
38. Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a Support Vector Machine**.** *J. Chem. Inf. Model*. **2005**; *45*, 549.
39. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409.
40. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol*. **1982**, *161*, 269.
41. Lipinski, C. A.; Lombardo, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev*. **1997**, *23*, 3.
42. Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aid. Mol. Design* **2000**, *14*, 251.