

# Chemical substructures in drug discovery

Cédric Merlot, Daniel Domine, Christophe Cleva and Dennis J. Church

The widespread use of HTS and combinatorial chemistry techniques has led to the generation of large amounts of pharmacological data, which, in turn, has catalyzed the development of computational methods designed to reduce the time and cost in identifying molecules suitable for pharmaceutical development. This review focuses on the use of substructure-based *in silico* techniques for lead discovery, an effective and increasingly popular approach for augmenting the chance of selecting drug-like compounds for preclinical and clinical development.

Cédric Merlot

Daniel Domine

Christophe Cleva

Dennis J. Church\*

Serono Pharmaceutical

Research Institute

14, ch. des Aulx

1228-Plan-les-Ouates

Geneva, Switzerland

\*e-mail:

dennis.church@serono.com

▼ The small molecule discovery process has been greatly influenced by recent advances in HTS and combinatorial chemistry, which has led to an unprecedented wealth of experimental data. Unfortunately, the widespread adoption of these techniques has also been accompanied by several disappointments, such as an insignificant rise in the number of compounds entering clinical development despite increased R&D expenditures worldwide [1].

Many reasons have been forwarded to explain the limitations of current discovery strategies, such as prohibitive costs, logistical requirements, the quality of the data that are being produced, the sheer size of chemical space and the appropriateness of molecules that are being progressed through development pipelines [2–4]. Unforeseen toxicity accounts for a 50% attrition rate of compounds in the clinical candidate stage [4], suggesting that many inadequate molecules are being pursued in hits-to-leads and lead optimization. In conjunction with the number of new targets that are expected from genomics and bioinformatics initiatives [5,6], the current cost and relative inefficacy of discovery research makes for a clear need to use more productive methods for the identification of clinical candidates.

As a result, recently there has been much interest in rationalizing the discovery process

by using *in silico* techniques, which aim to identify the molecular properties that are at the basis of a compound's biological effect(s). These techniques are used to select structures that are most likely to succeed in preclinical development. Although computational techniques for studying structure–activity relationships (SAR) have been available for many years now, newer substructure-based analytical methods are being used to conduct systematic analyses of hundreds of thousands of data points at a time and are able to supply researchers with predictive models for evaluating the most probable outcomes of screening, profiling and even toxicological experiments before they are even run [7–13]. This article looks briefly at these methods, together with examples of their use in HTS data analysis, compound set design, virtual screening, ADMET property prediction and selectivity profiling.

## Chemical substructures

Traditional SAR methods focus on the study of descriptors that are related to the chemical structures of compounds undergoing analysis. A veritable plethora of chemical descriptors have been devised over the years [14], and, intuitively, one would think that those reflecting the three-dimensional (3D) properties of a molecule would be the most effective because the binding of a small molecule to a drug target is a 3D-dependent event. Unfortunately, this is rarely the case because most 3D-based methods rely on computational simplifications that reduce the efficacy of the analytical process [3,15,16], thereby making them less effective than 2D-based techniques, such as those that employ chemical substructures [17,18]. Indeed, most 2D-based descriptors present the advantages of being self-explanatory, easy to calculate and simple to interpret [19], thus making them among the most

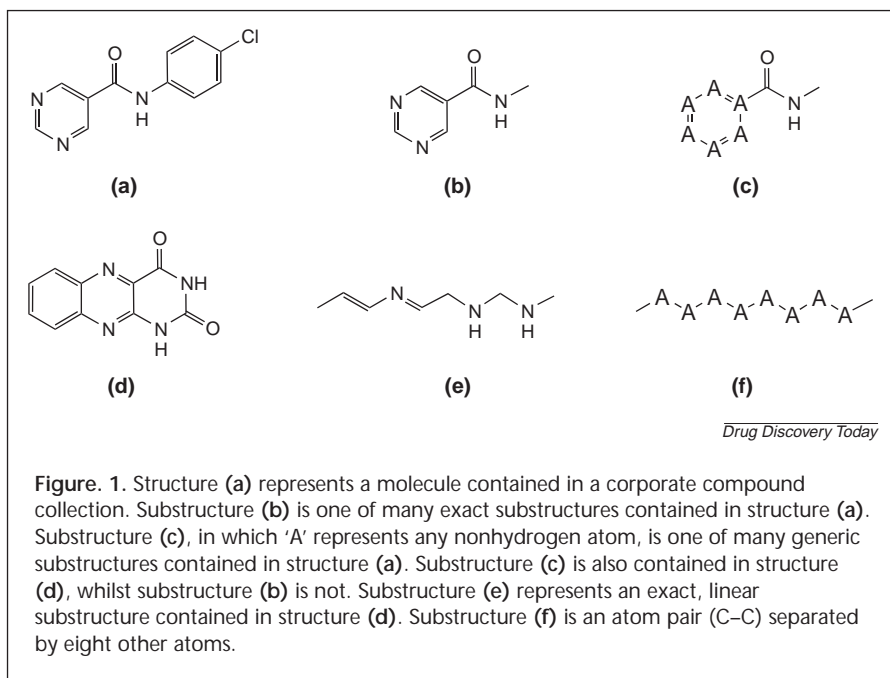
widely used descriptors in pharmaceutical research today [3].

But what exactly is a chemical substructure? Mathematically speaking, a substructure is a subgraph of the graph that is associated with a chemical structure, and is correspondingly labeled (or colored) in a manner that reflects the nature of the atoms and bonds comprised in the original (parental) molecule. Such descriptors were first used in chemical retrieval systems and were later used for data analysis and compound classification [20,21]. Molecular access system (MACCS) keys are the most commonly encountered example [22], but remain of limited analytical use because it is difficult to describe the infinity of chemical space with any finite list of small chemical fragments [21,23].

Accordingly, several groups have turned towards using larger lists of larger fragments for data analysis. These have the advantage of being more readily interpretable by medicinal chemists because one does not need to decipher endless lists of smaller substructures and they are easier to understand than many other 2D descriptors, which are often abstract, theoretical constructs [3,24–26]. Typically, the fragments used in this approach represent exact, generic, linear and/or branched substructures of 6–40 atoms (Fig. 1). An exact fragment is defined as a set of atoms and their connections, whereas a generic fragment is defined as a 'fuzzy' substructure in which atoms and/or bonds are denoted by symbols representing lists of possible atoms, bonds or even structural properties such as 'H-bond acceptor' or the like. Atom pairs [27] represent a special case of the use of generic linear fragments, in which the atoms flanked by the extremities are denoted with pseudo-atoms representing any type of atom. As a molecular graph contains fewer atom pairs than exact or generic fragments, these descriptors are easy to generate and can be supplemented with other substructures to retain information on the local environment (i.e. topological torsions) [28].

### Analytical methods

Numerous analytical techniques, such as multiple linear regression, partial least squares, clustering and neural networks, can, at least in theory, be used to process discovery data. However, the wealth of information produced by most discovery units is such that many methods are too cumbersome to be applied to the analysis of corporate

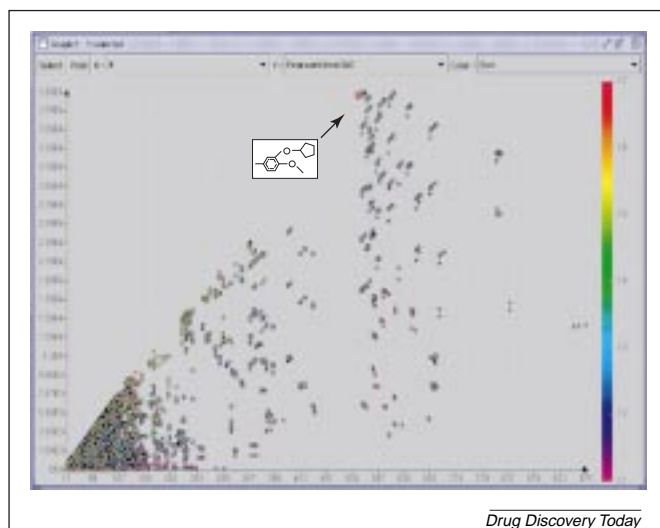


*Drug Discovery Today*

databases. As a result, several substructure-based techniques have been devised to process large amounts of information [2,3].

One of the simplest ways to perform a substructure analysis is to use a predefined list or 'dictionary' of fragments to check for the presence or absence of each substructure in a set of active molecules, and to then use a statistical test to identify fragments that are associated with the outcome of interest. This approach has been implemented in the commercially available software package LeadScope (<http://www.leadscope.com>) [29], which uses a 27,000 fragment dictionary and has been successfully applied to problems ranging from compound selection in HTS to toxicity prediction [30]. However, it is clear that even a dictionary of 100,000 carefully selected substructures cannot be used to identify all the possible pharmacologically active fragments, so several potentially interesting substructures are missed [3,31,32].

This problem might be circumvented by an approach made available by Bioreason (<http://www.bioreason.com>) [12], which groups compounds into clusters using MACCS keys and then scores each cluster on the basis of the average activity of the compounds contained in it. Substructures associated with activity are identified as the maximal common substructures (MCS) found in high-scoring clusters, which are then used as starting points for constructing larger, pharmacologically active fragments using a multidomain classification algorithm that comprises an annealing process. The end result is a method that allows one to detect fragments that are not explicitly enumerated in



**Figure 2.** A list of 1450 phosphodiesterase IV (PDE IV) inhibitors were exported from MDL's Drug Data Report [22], fragmented and the substructures were scored using a chi-square related test statistic. Each point on the graph represents a unique substructure (see example in inset). The x-axis shows the number of inhibitors containing a given fragment, whilst the y-axis is inversely related to the probability of chance occurrence of each fragment in the set of active structures. Color coding is used to indicate the fragment size. The substructures exhibiting the highest scores are the ones that are most unlikely to be present on the basis of chance alone, so it is assumed that they represent important determinants for PDE IV inhibition.

the fragment dictionary, thus providing the analyst with a distinct advantage for dealing with large sets of structurally diverse molecules.

It is also possible to replace predefined fragment dictionaries with alternative statistical techniques, such as recursive partitioning (RP) [7]. As suggested by its name, RP is a reiterative process whereby a set of structures is partitioned according to the presence or absence of a discriminating fragment contained in an exhaustive list of atom pairs and topological torsions, the fragment of choice is the one that provides the largest possible spread of mean activity between the subset of molecules that contain it and those that do not. Partitioned subsets are then repartitioned according to the next best descriptor (and so on and so forth), producing a hierarchically organized tree in which the nodes represent fragments that are most likely to be at the basis of a given outcome. Although RP can be a powerful technique for identifying combinations of fragments associated with biological activity [7,8], the method does have a few drawbacks. For example, molecules can only fall within a single given branch of each tree, so if a structure contains more than one pharmacologically active fragment (as many do) RP may not be successful in identifying all of them [32]. Moreover, RP can be sensitive to the

false-negative and false-positive data contained in unbalanced HTS datasets [12,32], whereas atom pairs contain little structural information *per se*. As a result, it is often necessary to reassemble atom pairs and topological torsion descriptors into easier-to-interpret structural moieties [33].

The MultiCASE (<http://www.multicase.com>) method provides an alternative to conducting substructure analyses and then having to reassemble the fragments into more accessible results [13]. It works by enumerating all linear fragments of up to ten atoms contained in a list of test molecules, and then assigning a statistical score value to each fragment according to its association with the outcome of interest. Although MultiCASE has been validated extensively for therapeutic and/or toxicological outcomes, it can encounter difficulties when dealing with datasets of more than a few thousand molecules.

Finally, discrete substructural analysis (DSA) is another method for identifying substructures associated with pharmacological outcomes that does not rely on the use of a predefined fragment dictionary [10]. It is based on a process whereby the structures undergoing analysis are exhaustively fragmented, and each substructure is then scored according to its probability of chance occurrence in the dataset at hand. The end result is an easy-to-interpret graphical output in which the fragments associated with a given outcome display high score values and can be further ranked on the basis of their size or incidence in the subset of active structures (Fig. 2). Furthermore, the fragments used can be exact, generic, linear and/or branched, and are typically very large, often more than 25 atoms [10,11,34], thereby providing chemists with clear examples of what substructure(s) a molecule needs to contain to exhibit a given pharmacological property, and this without having to reconstruct fragments or interpret descriptors after analysis. In this context, DSA is capable of rapidly identifying local and/or global maxima in datasets in excess of 100,000 chemical structures, which is impractical with many other methods described to date.

## Applications

### HTS

The most obvious application of substructure analysis is processing screening results. Although substructure-based information can be used in hits-to-leads chemistry to increase compound potency, bypass patents and/or to predict ADMET properties [3,10], the same information can also be used to direct sequential screening campaigns in an attempt to accelerate the lead discovery process. Sequential screening is an iterative technique that employs *in silico* methods to process screening results and then uses the information to select compounds for subsequent rounds of

testing. As a result, new compounds are selected from areas of chemical space occupied by chemotypes that have a high chance of displaying a desired activity, whereas molecules that do not contain the required determinants are removed from the screening process [2,35]. Substructure analysis is a very efficient way to identify those preferred chemical subspaces and routinely leads to 20-fold increases in screening hit rates [10,35,36].

Another use of substructure analysis in HTS is to detect false-positive results and 'frequent hitters' [10,12,31]. Many HTS datasets contain molecules that appear to be active, yet are structurally dissimilar from all other hits. This may be because the compounds are truly active and that no other molecules of the same chemotype were detected. Alternatively, the compounds may represent false-positive data points. This can be caused by many factors, such as pipetting errors, intrinsic fluorescence and the propensity of some compounds to form aggregates and the like [37]. Although it is difficult to determine the actual cause of a false-positive result, such compounds can be identified by checking whether their substructures are associated with activity by using several substructure-based analytical techniques [7,9,10,12,13].

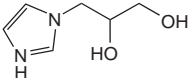
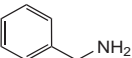
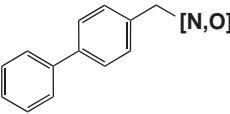
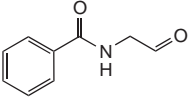
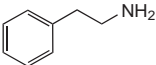
Nevertheless, several issues remain pertaining to the use of substructure analysis in screening operations, such as the failure of some methods to account for the notion of bioisosterism [3]. Many compounds that differ from a structural point of view are in fact equivalent in the sense that they bind to targets in a similar manner. However, methods that rely solely on the use of exact substructures have difficulty in detecting these equivalencies, a hurdle that can be removed by using generic substructures instead [10,11,36,38].

Finally, substructure analysis is an efficient tool for assembling initial, first generation compound collections for HTS, and represents an interesting complement to diversity-related approaches [39,40]. For example, the investigation of kinases has produced more than 300 pharmacologically active scaffolds that cover 80% of the kinase space, which can be used to build compound collections [41]. Again, the use of generic fragments is an important step in this process because it allows novel compounds to be identified [10].

#### *Compendium of chemical determinants*

Another interesting use of substructure analysis is in the construction of a database or 'compendium' of pharmacologically

**Table 1. Database of pharmacologically active chemical determinants<sup>a</sup>**

Substructure	Outcomes				
	Adenosine A <sub>1</sub> agonism	Angiotensin AT <sub>2</sub> antagonism	Cathepsin L inhibition	Adrenergic α <sub>2</sub> antagonism	Dopamine D <sub>1</sub> agonism
	93	0	3.2	2.3	0
	0.1	0	0.1	71	2.1
	0.7	109	0	0	0
	0	0.1	45	0	0
	6.1	0	0	11	112

<sup>a</sup>The table above shows a small portion of the Serono Pharmacological Compendium, which contains exact and generic fragments ('substructures', left hand column) annotated for their associations with various pharmacological effects ('outcomes', top row). In this case, the associations were determined using a modification of Fischer's Exact Test as per the DSA method [10]. The complete database contains thousands of substructures annotated for over 500 endpoints, including pharmacological, cell-based, animal model-based, toxicity, ADME and therapeutic outcomes. Such tables are simple to use. Molecules containing any of the substructures shown in the left hand column are most likely to exhibit effects that correspond to the outcome(s) for which they exhibit the highest score value(s) (see highlighted scores). For example, the substructure in the third row is strongly associated with angiotensin AT<sub>2</sub> receptor antagonism (score of 109 versus < 1 for other outcomes).

active chemical determinants (Table 1) [10]. Such a compendium can be viewed as representing a spreadsheet comprised of many thousands of rows and columns, in which each row corresponds to an exact or generic fragment that is annotated for its association with various biological end points. Such end points can be molecular, cell-based, animal model-based, toxicological, metabolic and/or therapeutic and can be created from corporate, literature and/or patent data.

Needless to say, such databases are extremely useful for designing compound collections because the researcher only needs to enter a set of outcomes to obtain a subset of substructures that code for the desired properties. Alternatively, a compendium can be used in the reverse to predict the most probable biological effects of novel compounds, a property that is useful in hits-to-leads and lead optimization chemistry. A database containing thousands of fragments associated with over 500 biological outcomes has been constructed for this purpose at Serono [10,11,34,36] and it is probable that similar repositories will be commercially available from other sources in the future.

### Virtual screening

Virtual screening is currently the most widely used method for compound selection [17,42] and can take many forms, such as filtering molecules for computed physicochemical properties, 2D-based substructure searching and/or 3D pharmacophore searching. 2D-based virtual screening consists of searching for structures that contain one or more given fragment(s) in a chemical database, a process that takes only seconds in repositories containing millions of molecules. In practice, searches often start by using generic fragments and/or lists of exact fragments in a manner that identifies as many candidate molecules as possible (typically thousands of structures), after which the subset is reduced by additional substructure searching, diversity analysis, similarity analysis and/or property-based filtering [3]. Interestingly, one would think that it would be difficult to retrieve novel chemical series on the basis of 2D searching, but this is not the case. Indeed, the use of generic fragments allows the chemist to identify active compounds with unexpected chemical structures, which allows one to bypass prior art [10,11]. Table 2 shows the results obtained for target-focused compound sets that were de-

signed using DSA at Serono. In each case, the compound sets were assembled using literature data; however, they delivered novel pharmacologically active chemotypes that, in the majority of cases, were of higher chemical tractability, potency and drug-likeness than those derived from a random set that was tested for control purposes.

Nevertheless, one drawback of substructure-based virtual screening is its propensity to deliver only a handful of candidates for testing, particularly when multiple nested searches are conducted in commercial compound collections, which does not allow the chemical space around a family of chemical determinants to be explored sufficiently [43]. Moreover, many commercially available structures exhibit poor ADMET properties, so the usefulness of virtual screening without using additional property filters remains questionable. One way to address this problem is to use substructure analysis to design and screen virtual libraries of billions of novel targeted, drug-like compounds.

**Table 2. Example of hit rate improvement using substructure analysis<sup>a</sup>**

Target	Class	Random set	Corresponding DSA Set	DSA-induced enhancement
1	7TM Receptor	0.10%	8.77%	87.7
2	7TM Receptor	0.10%	9.49%	94.9
3	7TM Receptor	0.50%	6.25%	12.5
4	7TM Receptor	0.78%	0.58%	0.7
5	Steroid Receptor	0.34%	5.58%	16.4
6	Kinase	0.15%	4.04%	26.9
7	Kinase	0.23%	4.21%	18.3
8	Kinase	0.24%	0.54%	2.3
9	Kinase	0.64%	3.10%	4.8
10	Ion Channel	0.58%	4.77%	8.2
11	Ion Channel	0.22%	0.87%	4.0
12	Protease	0.12%	10.02%	83.5
13	Phosphatase	0.47%	3.71%	7.9
14	Polymerase	0.52%	3.12%	6.0
Average enhancement:				26.7 Fold

<sup>a</sup>Focused compound sets of ~2000 molecules were generated using literature and patent data for 14 different targets representing seven target classes as previously described [10]. A set of 1280 randomly selected molecules was assembled for control purposes. The hit rates obtained with the target-focused sets (see 'DSA set') were almost always higher than those obtained by random screening, providing on average a 26.7-fold enhancement. Interestingly, many of the compounds identified in the focused sets represented novel pharmacologically active structures (in the sense of application and/or composition of matter, see ref. 10), whereas a significant proportion of those identified in the random set represented variants of known ligands, kinase inhibitors and the like. However, it should be noted that substructure analysis is not infallible, as operator errors and/or incorrect starting hypotheses can lead to failure in improving hit rates (see fourth and eighth entries).

### *Virtual combinatorial libraries*

Virtual combinatorial libraries (VCLs) are databases consisting of many hundreds of combinatorial reaction schemes that, at least in theory, can lead to the generation of billions of novel compounds. In this way, VCLs capture the chemist's knowledge in the form of combinatorial libraries [44], only a few of which will be selected during the virtual screening process for synthesis and testing [42,45]. Provided that the libraries are designed with protocols that comprise a few simple steps, the delay between subsequent rounds of HTS can be reduced so that the overall discovery process is considerably shortened.

Of course, the main problem in constructing a VCL lies in populating it with adequate structures. It is attractive to try to cover the majority of chemical space with non-targeted libraries in an attempt to introduce chemical novelty [43,46], but in practice this is unrealistic [19,42]. Indeed, whatever the number of reaction schemes contained in a VCL, these will never lead to products that represent more than a small part of chemical space. In agreement with this, it has been estimated that up to  $10^{100}$  structures can be synthesized using current methods [42,47], that there are potentially  $10^{60}$  small molecules of less than 30 non-hydrogen atoms [48] and that up to  $10^{40}$  pharmacologically relevant compounds are likely to exist [2,45]. Even if hardware continues to follow Moore's law [49], it is obvious that even a VCL containing trillions of compounds fails to represent such a large number of structures in any meaningful way.

Consequently, most VCL design work is focused on reducing the number and size of libraries to fit individual needs [42]. A current trend is to assemble collections that are directed towards families of targets as opposed to focusing on a single receptor or enzyme [50]. Scaffold selection is typically based on the chemist's insight, but can also involve selecting novel chemotypes whose structures bear determinants that are strongly associated with activity on a given family of targets [10,11]. This second approach is identical to the use of substructure analysis and/or a compendium in the design of first generation compound sets for sequential screening, a method that has been used to generate a VCL of  $4 \times 10^9$  novel, target-focused compounds at Serono.

Alternatively, the size of a VCL can be restricted by using more stringent inclusion criteria. For example, substructures can be generated from basic chemical principles and used to narrow the size of collections by assuring chemical accessibility. However, some groups insist that their virtual reaction products be novel before inclusion to increase the chance of claiming intellectual property in the event of activity. This is easy to do because one only needs to

conceive of novel central scaffolds and/or of novel reagents to ensure novelty for every structure that contains them. It is also important to focus on properties related to drug-likeness [42,51–53], which can be used as filters during the design phase [52–55]. This limits the chance of working on leads with poor ADMET properties [6] and can take on several forms, such as limiting the molecular weight or number of rotatable bonds in virtual reaction products [54], optimizing compounds for lipophilicity [42] and/or avoiding non-drug-like functional groups. However, such filters should be used with caution because non-drug-like compounds still provide precious structural information that can be used to optimize chemical series [42,55]. Moreover, undesirable moieties can sometimes be removed to turn non-drug-like hits into exciting leads [42].

### *Toxicity prediction*

Failures in the R&D process account for a substantial part of the drug development cost because many compounds are discarded after substantial investment owing to unforeseen toxicological effects. This has led to an accrued interest in applying substructure analysis to the problem of toxicological prediction [6], which is a process aimed at identifying the structural features that confer toxicological properties to molecules, the said structural features being termed toxicophores [56,57].

Most often toxicophores are used as filters to avoid potentially toxic compounds early on in the library design stage [6,42] and/or to flag compounds that are going to be difficult to progress [10]. Two approaches to toxicophore identification have been described. The first involves rule-based systems that require expert input [30], whereas the second relies on the identification of chemical determinants that are associated with toxicity as per every standard substructure analysis. Analyses are conducted on parent compounds and/or their metabolites. In the former case, the resulting toxicophores can be toxic *per se* or correspond to substructures in a parent molecule that confer toxicity once metabolized. A classic example is the aniline substructure (i.e. an aromatic amine), which is considered by many to represent a toxicophore even though the fragment requires metabolic activation to reveal its carcinogenic and/or mutagenic behavior [57].

In view of the above, many of the methods used for analyzing HTS data can be applied directly to the problem of toxicological prediction [9,10,13], as further illustrated by the adoption by the FDA of the MultiCASE technique for assessing drug toxicity [58–60]. However, it should be recognized that selecting compounds for HTS differs from trying to eliminate toxic ones. In the former case, one is interested in identifying active molecules, so it suffices to

assemble a set of structures that comprise fragments that are strongly associated with activity. In the latter case, however, one is interested in assessing the potential toxicity of individual compounds. Although known toxicophoric fragments can be used to eliminate undesirable molecules in a preliminary filtering step [57,61], more accurate predictions require the in-depth analysis of each structure [57–62].

#### ADME prediction

ADME prediction is another area that has begun to attract considerable attention [6,62]. In view of the difficulty in evaluating both *in vitro* and *in vivo* properties, ADME prediction is a particularly challenging area in which most *in silico* methods have been directed at increasing compound throughput or diminishing the need for animal sacrifices [63]. Accordingly, there are very few reports describing the use of substructure analysis in the assessment of ADME properties, although this will probably change in the future.

The majority of drugs are metabolized by cytochrome P450 and/or by extrahepatic enzymes [51,56]. 3D pharmacophore models have been used to describe this process [6] and it is foreseen that docking algorithms will also soon be used to predict metabolism [64]. An alternative approach involves screening structures for the presence or absence of fragments known to be associated with cytochrome P450 inhibition (or induction), such as those contained in a fragment database [10].

Substructure-based analytical methods are also starting to be used for predicting compound permeability and plasma protein binding [51]. In the first case, a method based on the use of partial least squares has been described and appears to be a useful complement to 3D-based methods, which have difficulty in dealing with the problem of passive diffusion. Along the same lines, the MultiCASE technique has been used to identify biophores (i.e. chemical fragments) that are associated with plasma protein binding, which is a useful way to supplement standard lipophilicity calculations [17].

Finally, there are no reports describing the use of substructure analysis in the prediction of blood-brain barrier permeability, but, again, this may change in the future. Other methods are typically used for this purpose [51], as well as for the prediction of solubility [65–67]. Interestingly, when fragment-based approaches are used for assessing solubility, substructures are regarded more as contributing groups than as representing a single pharmacophoric moiety that is at the basis of a given outcome [68]. In this context, it should be noted that the prediction of ADME properties remains a challenge using any computational

technique, particularly in view of the paucity of reliable data [51].

#### Selectivity

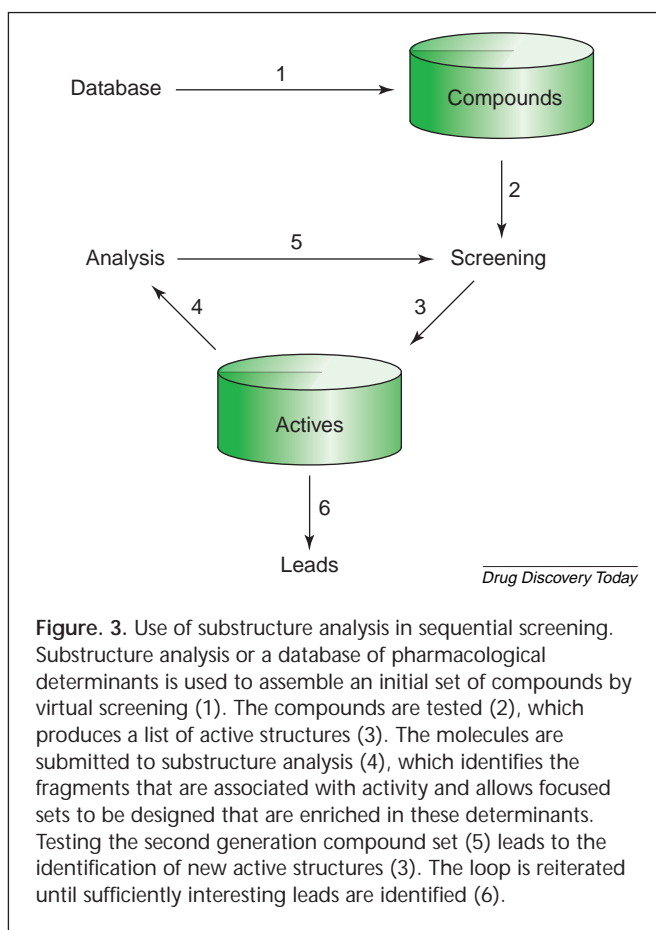
A last application of substructure analysis is in the prediction of compound selectivity. In this regard, databases of chemical determinants are invaluable tools for predicting the secondary pharmacological actions of compounds [10,11]. This can be done either by using a compendium, such as a dictionary, to check for the presence or absence of substructures that are associated with various actions as previously described, or simply by checking for the presence of a specific fragment in a compound of interest [10,11].

#### Conclusion

Substructure analysis is rapidly changing the way hit and lead candidates are being identified for pharmaceutical development. Although substructure-based computational techniques have been available for nearly three decades [69], the more recently devised methods address many of the previous limitations relating to the use of small sets of fragment keys, the need to handle and interpret difficult-to-understand indices and the impossibility of analyzing more than a few hundred structures at a time [7–13]. In addition, several easy-to-use substructure-based analytical tools are now commercially available, thereby allowing discovery groups to circumvent the need to randomly test hundreds of thousands of molecules to identify novel, drug-like chemotypes.

Overall, the heuristic potency of substructure-based predictive methods is now well established for HTS data analysis, compound set design, virtual screening, *in silico* selectivity profiling and toxicological prediction. One of the many interesting future applications of substructure analysis is in high throughput sequential screening (Fig. 3). Although the need to cherry pick and plate individual compounds can make this a very time consuming process [35], implementing workflows in which privileged substructures are extracted from datasets and then used to select compound plates that are enriched in closely related variants of the same said substructures is faster and is a method that is being used successfully in several laboratories [8,11,39].

Another exciting application of substructure analysis is in the compilation of databases of privileged substructures, which can provide a useful framework for decision making. Not only can such repositories be used for virtual screening, as described above, but they can also be used to design compound sets for identifying orphan receptor-ligand pairs to elucidate targets in cell-based chemical genomics experiments, to build semiautonomous discovery



workstations and/or to evaluate the hits and lead series. In this last regard, knowing that a lead candidate bears one or more chemical determinants that are strongly associated with, for example, the inhibition of a cardiac ion channel, the induction of hepatic enzymes and/or a nephrotoxic effect is a clear advantage for selecting molecules for further progression.

Finally, ADME property prediction is another field in which substructure analysis is likely to represent a useful addition to currently available techniques. Much remains to be explored in this area, and, although both solubility and passive transport modeling appear to be better addressed using other methods, preliminary reports indicate that substructure analysis may prove useful in predicting both central nervous system activity (indicative of blood-brain barrier permeability) and metabolism [70,71].

To conclude, modern substructure-based analytical methods are useful additions to the discovery scientist's armory of computational techniques. They are affordable, readily accessible to the non-specialist and provide discovery units with the capacity to predict the outcome of screening experiments before they are even run, which remains a remarkable advantage in an industry in which

random HTS and step-by-step medicinal chemistry remain so often the norm.

## References

- 1 Bajorath, J. (2001) Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* 6, 989-995
- 2 Engels, M.F.M. and Ventkatarangan, P. (2001) Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* 4, 275-283
- 3 Merlot, C. *et al.* (2002) Fragment analysis in small molecule discovery. *Curr. Opin. Drug Discov. Dev.* 5, 391-399
- 4 Boguslavsky, J. (2001) Minimizing risk in hits to leads. *Drug Discov. Devel.* 4, 26-30
- 5 Joseph-McCarthy, D. (2002) An overview of *in silico* design and screening: toward efficient drug discovery. *Curr. Drug Discov. March*, 20-23
- 6 Manly, C.J. *et al.* (2001) The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today* 6, 1101-1110
- 7 Rusinko, A. *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017-1026
- 8 Jones-Hertzog, D.K. (1999) Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J. Pharmacol. Toxicol.* 42, 207-215
- 9 Roberts, G. *et al.* (2000) LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* 40, 1302-1314
- 10 Church, D.J. and Colinge, J. (2000) Method of operating a computer system to perform a discrete substructural analysis. *Eur. Pat. Appl. EP 00/309114 PCT Int. Appl. WO 02, 2002*
- 11 Cleva, C. *et al.* (2002) *Privileged substructure searching for focused set design. Presentation at the 224<sup>th</sup> ACS meeting*, August 18-22, Boston
- 12 Nicolaou, C. *et al.* (2000) Method and system for artificial intelligence directed lead discovery through multi-domain clustering. *PCT Int. Appl. WO 00/049539*.
- 13 Klopman, G. and Tu, M. (1999) Diversity analysis of 14156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure-activity relational expert system MCASE. *J. Med. Chem.* 42, 992-998
- 14 Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* 40, 195-209
- 15 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7, 903-911
- 16 Root, D.E. (2002) Global analysis of large-scale chemical and biological experiments. *Curr. Opin. Drug Discov. Dev.* 5, 355-360
- 17 Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* 41, 233-245
- 18 Bajorath, J. (2002) Virtual screening in drug discovery: methods, expectations and reality. *Curr. Drug Discov. March*, 24-28
- 19 Langer, T. and Hoffmann, R.D. (2001) Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.* 7, 509-527
- 20 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. *J. Chem. Inf. Comput. Sci.* 36, 572-584
- 21 Rhodes, N. and Willett, P. (2000) Bit-string methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* 40, 210-214
- 22 MDL Information Systems Inc San Leandro, CA, USA. <http://www.mdll.com>
- 23 MacGuish, J. *et al.* (2001) Ties in proximity and clustering compounds. *J. Chem. Inf. Comput. Sci.* 41, 134-146
- 24 Cosgrove, D.A. and Willett, P. (1998) SLASH: A program for analyzing the functional groups in molecules. *J. Mol. Graph. Model.* 16, 19-32
- 25 Randic, M. *et al.* (2001) On structural interpretation of several distance related topological indices. *J. Chem. Inf. Comput. Sci.* 41, 593-601

- 26 Randic, M. and Zupan, J. (2001) On interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* 41, 550–560
- 27 Carhart, R.E. *et al.* (1985) Atom pairs as molecular features in structure-activity studies: definitions and applications. *J. Chem. Inf. Comput. Sci.* 25, 64–73
- 28 Nilakantan, R. (1987) Topological torsions: a new molecular descriptor for SAR application comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* 27, 82–85
- 29 Roberts, G. *et al.* (2000) LeadScope: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* 40, 1302–1314
- 30 Johnson, D.E. (2001) Chem-tox informatics: data mining using medicinal chemistry building block approach. *Curr. Opin. Drug Discov. Dev.* 4, 92–101
- 31 Roche, O. (2002) Development of virtual screening method for identification of 'frequent hitters' in compound libraries. *J. Med. Chem.* 45, 137–142
- 32 Nicolaou, C.A. *et al.* (2002) Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* 42, 1069–1079
- 33 Blower, P. *et al.* (2002) On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* 42, 393–404
- 34 Sauer, W. *et al.* (2001) *Cost-effective discovery using predictive substructural analysis. Presentation at the Drug Discovery Technology meeting*, Boston
- 35 Valler, M.J. and Green, D. (2000) Diversity screening versus focused screening in drug discovery. *Drug Discov. Today* 5, 286–293
- 36 Domine, D. *et al.* (2001) *High-throughput lead discovery using predictive substructural analysis. 221<sup>st</sup> ACS Meeting*, San Diego, CA, USA COMP-099
- 37 McGovern, S.L. *et al.* (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* 45, 1712–1722
- 38 Cho, S.J. *et al.* (2000) Binary formal inference-based recursive modeling using multiple atom and physicochemical property class pair and torsion descriptors as decision criteria. *J. Chem. Inf. Comput. Sci.* 40, 668–680
- 39 Young, S.S. *et al.* (2002) Initial compound selection for sequential screening. *Curr. Opin. Drug Discov. Dev.* 5, 422–427
- 40 Abt, M. *et al.* (2001) A sequential approach for identifying lead compounds in large chemical databases. *Stat. Sci.* 16, 154–168
- 41 Kubinyi, H. (2002) High throughput in drug discovery. *Drug Discov. Today* 7, 707–709
- 42 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 169–178
- 43 Nilakantan, R. *et al.* (2002) A novel approach to combinatorial library design. *Comb. Chem. High Throughput Screen.* 5, 105–110
- 44 Hecht, P. (2002) High-throughput screening: beating the odds with informatics-driven chemistry. *Curr. Drug Discov.* January, 21–24
- 45 Ritchie, T.J. (2001) Chemoinformatics: manipulating chemical information to facilitate decision-making in drug-discovery. *Drug Discov. Today* 6, 813–814
- 46 Tropsha, A. and Zheng, W. (2002) Rational principles of compound selection for combinatorial library design. *Comb. Chem. High Throughput Screen.* 5, 111–123
- 47 Beroza, P. *et al.* (2002) Chemoproteomics as basis for post-genomic drug discovery. *Drug Discov. Today* 7, 807–814
- 48 Bohacek, R.S. *et al.* (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16, 3–50
- 49 Augen, J. (2002) The evolving role of information technology in the drug discovery process. *Drug Discov. Today* 7, 315–323
- 50 Lahana, R. (2002) Cheminformatics – decision making in drug discovery. *Drug Discov. Today* 7, 898–900
- 51 Clark, D.E. *et al.* (2002) Progress in computational methods for the prediction of ADMET properties. *Curr. Opin. Drug Discov. Dev.* 5, 382–390
- 52 Teague, S.J. *et al.* (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* 38, 3743–3748
- 53 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315
- 54 van Dongen, M. (2002) Structure-based screening and design in drug discovery. *Drug Discov. Today* 7, 471–478
- 55 Viswanadhan, V.N. (2002) Knowledge-based approaches in the design and selection of compound libraries for drug-discovery. *Curr. Opin. Drug Discov. Dev.* 5, 400–406
- 56 Williams, D.P. and Naisbitt, D.J. (2002) Toxicophores: groups and metabolic routes associated with increased safety risks. *Curr. Opin. Drug Discov. Dev.* 5, 104–115
- 57 Barratt, M.D. and Rodford, R.A. (2001) The computational prediction of toxicity. *Curr. Opin. Chem. Biol.* 5, 383–388
- 58 Schwetz, BA *et al.* (1999) Science at the FDA: improving the scientific basis of regulation through collaboration with 'stakeholders' [http www.fda.gov/oc/oha/fdascience.htm](http://www.fda.gov/oc/oha/fdascience.htm)
- 59 On the World Wide Web URL: [http www.fda.gov/ohrms/dockets/dailys/02/May02/051002/99N-2079\\_emc-000002-01.pdf](http://www.fda.gov/ohrms/dockets/dailys/02/May02/051002/99N-2079_emc-000002-01.pdf)
- 60 Benz, D. *Presentation at the 22<sup>nd</sup> Annual Meeting of the American College of Toxicology*, Washington, DC. &-Nov-01
- 61 Durham, S.K. and Pearl, G.M. (2001) Computational methods to predict drug safety liabilities. *Curr. Opin. Drug Discov. Dev.* 4, 110–115
- 62 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055
- 63 van de Waterbeemd, H. (2002) High-throughput and *in silico* techniques in drug metabolism and pharmacokinetics. *Curr. Opin. Drug Discov. Dev.* 5, 33–43
- 64 Waszkowycz, B. (2002) Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discov. Dev.* 5, 407–413
- 65 Engkvist, O. and Wrede, P. (2002) High-throughput, *in silico* prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* 42, 1247–1249
- 66 Raevsky, O.A. *et al.* (2002) SLIPPER-2001 – Software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. *J. Chem. Inf. Comput. Sci.* 42, 540–549
- 67 Advanced Chemistry Development. On the World Wide Web URL: [http://www.acdlabs.com/products/phys\\_chem\\_lab/aqsol](http://www.acdlabs.com/products/phys_chem_lab/aqsol)
- 68 Klopman, G. and Zhu, H. (2001) Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* 41, 439–445
- 69 Cramer, R.D. *et al.* (1974) Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* 17, 533–535
- 70 Engkvist, O. *et al.* (2003) Prediction of CNS activity of compound libraries using substructure analysis. *J. Chem. Inf. Comput. Sci.* 43, 155–160
- 71 Lee, P.W. (2001) *Metabolism expert system: Management of metabolism information and knowledge. Presentation at the 222<sup>th</sup> ACS meeting*, August 26–30, Chicago.

## Conference reports

*Drug Discovery Today* Publications is pleased to publish the highlights from international conferences.

Conference participants who wish to cover a particular meeting should contact:

Dr Christopher Watson, *Drug Discovery Today*, 84 Theobald's Road, London, UK WC1X 8RR

e-mail: DDT@drugdiscoverytoday.com