ELSEVIER

Knowledge-based chemoinformatic approaches to drug discovery

Arup K. Ghose, Torsten Herbertz, Joseph M. Salvino and John P. Mallamo

Department of Chemistry, Cephalon, 145 Brandywine Parkway, West Chester, PA 19380, USA

The modern drug discovery process is steadily becoming more information driven. Structural, physicochemical and ADME–Tox property profiles of reference (successful) ligands, along with structural information of their target proteins, have been extremely useful for early-stage drug discovery. Recently, databases of known biologically active ligands (knowledgebases) have become more focused toward different protein-target classes. The number of new chemoinformatics tools used to analyze structures and properties of successful molecules has also increased enormously. Scientists in this area are exploring new physicochemical properties and appropriate drug sets to understand druglike properties. In this review, the various uses of the ligand knowledgebases in the drug discovery process have been critically reviewed.

Drug discovery is a complex and risky process. Despite increased spending in research and development, the number of drugs launched has declined in recent years [1]. Only 11% of drugs entering clinical development reach the market place, most are withdrawn from further studies mainly for reasons associated with efficacy, toxicity, drug metabolism and pharmacokinetics. Can we enhance the success rate from systematic knowledge use? Chemoinformatics approaches to drug discovery (Figure 1) are based on the premise that knowledge of successful and failed ligand structures and properties are valuable for hit enrichment, appropriate lead identification and lead optimization. It is assumed that staying within the physicochemical property ranges where most successful compounds reside and satisfying their structural features or similarities will statistically enhance the chances of success in hit finding, identifying the best lead from multiple hits and generating the best candidate during lead optimization. In this review, we will focus first on the therapeutic protein-target specific ligand knowledgebases that are available, then on the physicochemical properties of the approved drugs or partially successful compounds. In the final section we will focus on the structural aspects of the ligands.

Understanding druglike molecules is the primary objective of every knowledge-based drug discovery approach. There have been numerous attempts to predict drug-likeness [2–4]. Among these studies, Lipinski's Rule of Five (ROF) emerged as a guide to provide medicinal chemists with a simple means of predicting potential problems regarding solubility and permeability, factors that greatly influence drug absorption [4]. Thus, Lipinski's rules have been adopted in the broadest sense to define druglike properties. Over time, various groups [5,6] began to realize that a large portion of commercially available compounds included in databases such as the Available Chemical Database (ACD) (http://www.mdl.com/products/knowledge/index.jsp) satisfied Lipinski's rules. Databases of approved drugs and compounds that have progressed into clinical trials are readily accessible. For example, Corwin Hansch dedicated a volume of *Comprehensive Medicinal Chemistry* [7] to associating chemical structures with their physicochemical properties. Molecular Design Limited (MDL) (http://www.mdl.com/products/knowledge/index.jsp) concurrently started to distribute this information as the Comprehensive Medicinal Chemistry (CMC) database. Consequently, investigators [8] have studied physicochemical and structural profiles of various classes of drugs. Others have focused solely on the smaller number of marketed oral drugs that have successfully overcome hurdles in the drug-development process. Thus, deciding what the most appropriate knowledgebase is, or what the compound set and the physicochemical and substructural properties as guidelines in developing proprietary compound libraries for hit enrichment are in lead optimization and candidate selection, is important.

Corresponding author: Ghose, A.K. (aghose@cephalon.com)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.10.012

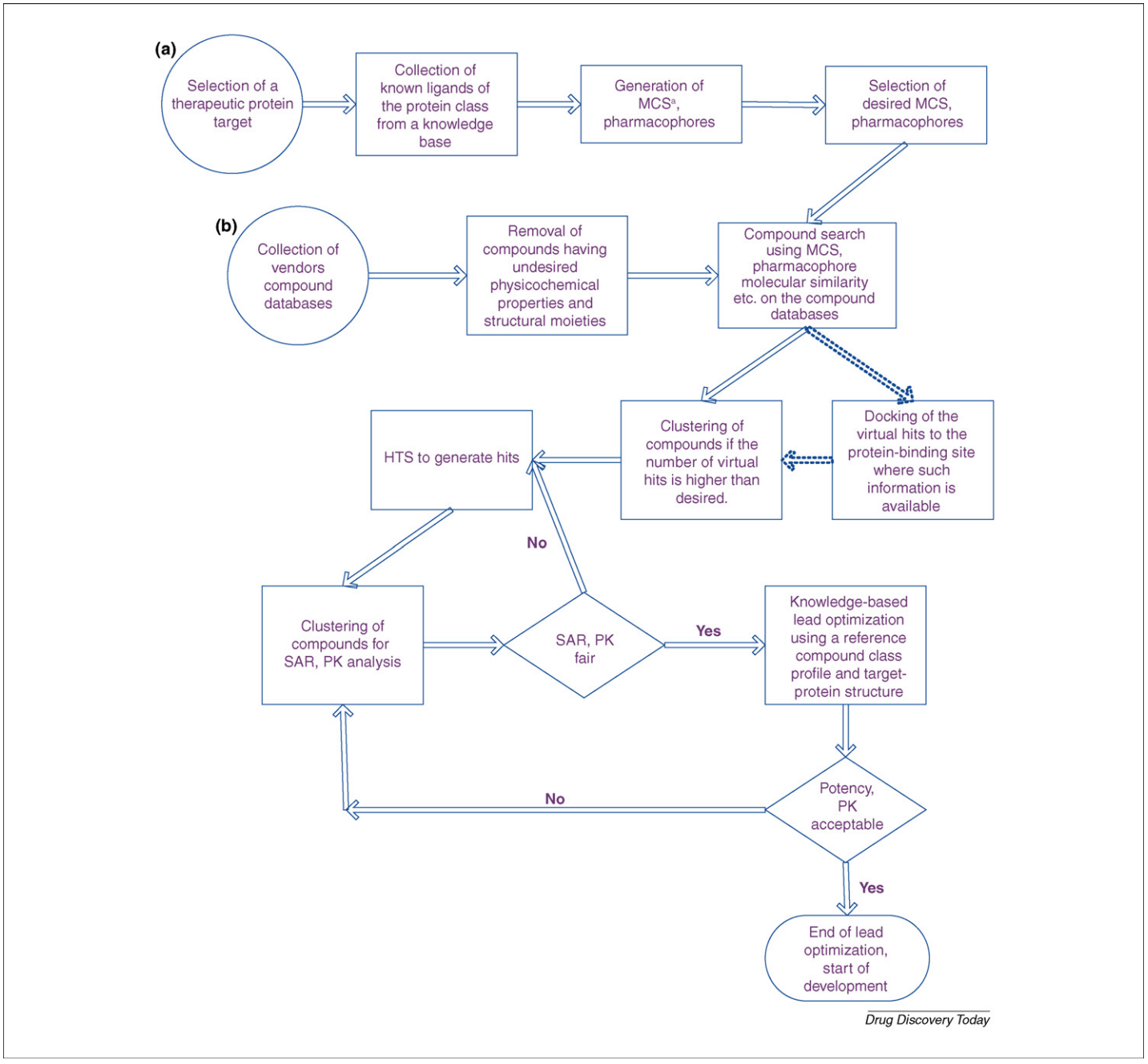

FIGURE 1

A chemoinformatics-driven drug discovery approach. Dotted arrows represent optional steps. This approach starts with two parallel steps: **(a)** selection of a therapeutic target protein and associated known ligands of the target-protein class; and **(b)** creation of a virtual database of commercially available compounds. The databases of known successful compounds can be used to learn important common substructures and physicochemical proerty ranges for this class of ligands. The database of commercially available compounds should be used for virtual screening using the common substructures and these physicochemical property ranges. Virtual screening hits can be acquired and used for HTS. The 'real' hits can be used further for data mining in the virtual database. The process continues until a viable 'lead' is found. A rigorous compound synthesis will only start in the lead optimization process. [a]MCS, Maximal common substructure.

Selection of the knowledgebase

Databases emphasizing various subclasses of biologically active compounds have rapidly expanded and increased the use of repositories such as the World Drug Index (WDI) (http://www.scientific.thomson.com/products/wdi/), MDL Drug Data Report (MDDR) (http://www.mdl.com/products/knowledge/drug-data_report/index.jsp) and CMC (http://www.mdl.com/products/knowledge/medicinal_chem/). The majority of these databases focus on classification of G-protein-coupled receptor (GPCR) ligands (http://www.jubilantbiosys.com/products.htm; http://www.inpharmatica.co.uk/GPCR/Index.htm; http://www.gvkbio.com/informatics/dbprod.htm; http://www.aureus-pharma.com/Pages/Seminars/seminar_abstracts.php). This is consistent with the fact that GPCRs contribute substantially to the number of protein targets in human therapeutics, where protein structural information is at a minimum. Pharmaceutical industry interest

in kinase targets (http://www.jubilantbiosys.com/products.htm; http://www.gvkbio.com/informatics/dbprod.htm; http://www.eidogen-sertanty.com/products.html), toxicity and metabolism (http://www.mdl.com/products/pdfs/metab-tox_broc.pdf; http://www.lhasalimited.org/index.php) also generates considerable efforts in databases. The selection of a knowledgebase depends on the goal of the study; for example, one such specialized database might prove to be an excellent source for a thorough compilation of a range of biologically active molecules for a specific target. By contrast, the WDI and MDDR contain an excellent summary of biologically active molecules from various therapeutic areas.

Physicochemical property profile

Initial interest in specialized databases focused on the physicochemical properties of compounds in late-stage clinical development or approved drugs [2,4,8]. More-recent efforts [9–11] have used all approved drugs, or oral drugs only, in the physicochemical property analysis; because analysis of drugs from various phases of clinical development revealed that the mean molecular weight of compounds in earlier development phases was much higher than of those in later stages of development and marketed drugs [10]. In addition, it was found that compounds that were more lipophilic tend to be discontinued before Phase III. A decreased number of H-bond acceptors and more-constrained molecules (i.e. molecules with few rotatable bonds) is observed in late-stage compounds. Thus, using the smaller set of Phase III candidates or launched oral drugs arguably provides a more meaningful set of molecules on which to base profiling. However, anticancer and antifungal compounds can also be unusual and should not be used for the basis of profiling molecular properties for orally available drugs [8] because these compounds often react chemically with biological molecules. Recently, several scientists [11] raised the issue of variation in physicochemical profile over time, and analyzed available drug databases in a time-dependent manner. Computationally derived values and experimentally derived measurements have been demonstrated to be useful. Computational properties might be the only option for profiling early-stage compounds, which are made or purchased in small amounts, and designed conceptual compounds, which are yet to be made. As a program progresses, it is crucial that calculated property trends are confirmed with experimentally obtained values because fewer compounds are made in large enough quantities for these measurements.

Early analysis of drug databases

Lipinski derived a simple set of rules by profiling 2245 molecules from the Derwent Drug Index (http://www.scientific.thomson.com/products/wdi/). According to the ROF, poor drug absorption and permeation are likely to be observed when any two of the following conditions are met: (i) there are greater than five H-bond donors (ii) molecular weight is greater than 500; (iii) the calculated log P (clogP) is greater than five; and (iv) the molecule has more than ten H-bond acceptors. Four rules, and all of them are divisible by five – leading to the name Rule of Five. These rules catalogue factors that strongly influence drug absorption [11]. The ROF was derived from compounds reaching at least Phase II clinical trials through to approved drugs. Many of these drug candidates did not reach the market. Coincidently, it is possible to purchase a few million compounds that comply with the ROF for HTS studies [12], yet the drug approval rate is declining. Thus, profiling the smaller number of marketed oral drugs can provide a more reasonable measure of the druglike physicochemical properties required. These compounds might have already passed drug metabolism and pharmacokinetics (DMPK) studies, formulation studies, toxicity studies and various clinical obstacles, and can be synthesized on a commercial scale.

Compounds of different therapeutic classes

It is important to realize that physicochemical properties for drugs with molecular targets in the central nervous system (CNS) versus the periphery, or drugs for topical applications, will vary greatly. Therefore, compound properties should be profiled based on a therapeutic objective. For example, CNS drugs differ greatly from the other classes of drugs with regards to lipophilicity [8,13,14]. This is because CNS drugs typically need to cross the blood–brain barrier to reach their molecular target. Furthermore, in the CNS drug class, antipsychotic drugs are considerably more hydrophobic (mean log P = 4.1) than antidepressants (mean log P = 3.1) or hypnotic drugs (mean log P = 2.2). CNS drugs tend to be more lipophilic than anticancer, cardiovascular or anti-infective drugs.

Profiling inhibitors of different protein classes

Inhibitors of a specific class of proteins (e.g. kinases) can be considered as the complementary image of the active site and, because proteins of the same class can have a similar binding pocket, we might expect similarity in the properties of these inhibitors. Although there are few published works [15,16] that describe the simple physicochemical property profiles of the ligands of various protein classes, many chemical vendors are profiling compounds based upon protein-class ligand properties to bundle their compounds to customers. Morphy [15] studied the property profiles of inhibitors of several protein classes such as GPCRs, kinases, nuclear receptors and integrin receptors, among others.

New properties for drug profiling

Although clogP and molecular weight are historically the two most studied parameters in the analysis of drug properties, some scientists [17] have explored the relationship between molecular-surface properties, specifically polar surface area (PSA), and oral absoprtion. Twenty structurally and physicochemically diverse model drugs – ranging from 0.3% to 100% measured oral absorption – showed an excellent sigmoidal relationship between the absorbed fraction (FA) and the dynamic PSA after oral administration in humans. Drugs that are sufficiently absorbed (% of FA >90%) had a PSA $\leq$60 Å^2, whereas drugs that are <10% absorbed had a PSA $\geq$140 Å^2. Kelder *et al.* [18] calculated PSA for a more extensive set of compounds, to study similar relationships. The authors used 776 CNS active compounds and 1590 non-CNS drugs that had reached, at least, Phase II in the clinic. They found a clear difference in the distribution of the PSA between CNS and non-CNS drugs. They concluded that orally active molecules that are transported passively by a transcellular route should not exceed a calculated PSA of ~120 Å^2. For efficient penetration of the blood–brain barrier the PSA should be <60–70 Å^2. After several groups claimed that PSA accurately predicted cell permeability and blood–brain barrier penetration, fast calculations of PSA were developed

using fragment- or atom-based methods [19]. Ertl *et al.* [19] claimed that fragment-based methods gave topological PSA (TPSA) values that are identical to 3D structure-based PSA values. Egan *et al.* [20] built a statistical pattern-recognition model of passive intestinal absorption. They used PSA and AlogP98 [21] in their model, and computed the 95% and 99% confidence ellipses for well-absorbed compounds using 199 compounds that are described in the literature as >90% absorbed or having oral bioavailability >90%. They also used a set of 35 poorly absorbed compounds. Most commercial molecular-modeling software companies provide modules to predict various pharmacokinetic properties along with the physicochemical properties. The computed values of complex pharmacokinetic properties should be used with caution because these models were often developed from small sets of compounds – compared with the vast number of compounds that are used in early drug discovery.

Time dependency of druglike property profiles

Veith *et al.* [22] did an interesting study on the relationships of drug properties as a function of their product-launch date. If the average value of various key physicochemical properties shift as a function of the launch date, scientists might need to create a dynamic definition of a druglike molecule by incorporating this time progression. This analysis demonstrated that the mean property values for oral drugs did not vary substantially with respect to launch time between 1982 and 2002. In another study, Leeson and Davis [11] calculated the mean and median of several physicochemical properties of drugs during three time frames: pre-1983, 1983–1992 and 1993–2002 (Table 1). They reached a slightly different conclusion. According to their study, mean values of lipophilicity, PSA and H-bond-donor counts remained the same, whereas other mean values (e.g. of H-bond-acceptor counts) have changed. The authors suggest that the invariant properties are the most important oral druglike physicochemical properties.

Property upper limit or the property range

By discussing the physicochemical properties of drugs, most scientists either provide an upper limit (e.g. the ROF) or the mean and/or median (Table 1). However, it is clear that the property distributions are closely Gaussian (bell-shaped) in nature rather than uniform across the range [8,20]. The studies by Ghose *et al.* [8] and Egan *et al.* [20] show that it is appropriate to select compounds in the most populated regions of property space (i.e. below the maximum of the population curve) as those occupied by marketed drugs or any other reference set. The percentage of drugs that fall within the desired range can be subjective. It might be appropriate to satisfy physicochemical property ranges of 80% or 95% of the approved drugs: (i) compounds falling within the 80% range can be considered good; (ii) compounds with properties outside this range but within the 95% range can be considered moderate; (iii) anything beyond these ranges might need attention. The 95% range is considerably larger and can accommodate less-successful candidates. It is useful to set different reference points for different therapeutic applications, such as CNS drugs for CNS lead optimization, versus non-CNS drugs for non-CNS lead optimization.

TABLE 1

Mean physicochemical properties of oral drugs at different times[a], median values given in parentheses

	Pre-1983	1983–1992	1993–2002
	***n* = 864**	***n* = 175**	***n* = 154**
MW[b]	331 (310)	374. (359)	382 (357)
ClogP	2.27 (2.31)	2.39 (2.36)	2.61 (2.38)
%PSA	21.1 (18.5)	20.9 (19.0)	21.2 (19.97)
OH^+NH	1.81 (1)	1.75 (1)	1.80 (1.5)
O + N	5.14 (4)	6.33 (6)	6.32 (6)
HBA	2.95 (2)	3.66 (3)	3.82 (4)
RotB	4.97 (4)	6.29 (6)	6.58 (6)
Ring	2.56 (3)	2.77 (3)	3.02 (3)

[a] Adapted, with permission, from Ref. [11]. Here, *n* is the number of oral drugs.
[b] MW, molecular weight; clogP, computed octanol–water partition coefficient (calculated using the Hansch and Leo approach); %PSA, PSA ÷ (total surface area); HBA, H-bond acceptors; RotB, number of rotatable bonds; Ring, number of rings in the molecule.

Knowledge-based complex mathematical models

Immediately after Lipinski's ROF [4] became popular, several groups [5,6,21] applied ROF to the available chemical databases such as ACD (http://www.mdl.com/products/knowledge/index.jsp) with the hope of buying ready-made druglike molecules. To their surprise, ~80% of compounds in the ACD satisfied Lipinski's ROF. In other words, the ROF cannot differentiate between drugs and chemicals. The authors, either using neural net or other statistical methods, developed more-complex expressions for druglike molecules or inhibitor classes. These complex models were more successful in differentiating between drugs and chemicals. The core of all such methods involves identifying two sets of compounds belonging to the positive class and the negative class followed by applying statistical methods or artificial intelligence [23].

Pharmacophore, molecular similarity and maximal common substructure

Pharmacophore, molecular similarity and maximal common substructure (MCS) are three inter-related approaches that are used to invoke the structural knowledge of known ligands in drug research. The term pharmacophore was introduced almost a century ago by Ehrlich (see Ref. [24]) and has been defined as the essential substructural features, and their geometric arrangement, in a molecule known to have a defined biological recognition event or activity. The ultimate goal of a pharmacophore definition [25] is to explain which minimum set and arrangement [26,27] of functional groups or structural features can account for biological activity. This understanding can be used to identify different classes of active compounds conforming to these pharmacophores, and to enhance activities by optimizing 'alignment' with them. The molecular similarity approach [28,29] takes the whole structures of the active compounds and divides it into smaller substructures (fingerprints), usually containing 1–6 atoms extending bonds outward from a central atom. It determines the similarity of two molecules by the number of common fingerprints and the total number of fingerprints in the two molecules. By contrast, MCS (http://www.simulationplus.com/classpharmer/classpharmer.html; http://www.scitegic.com/products_services/pipeline_pilot.htm) is a single contiguously connected common subgraph of a molecular structure present in a specified fraction of all molecules.

Derivation and uses of a pharmacophore

Medicinal chemists develop conceptual pharmacophore hypotheses based on SAR trends and cumulative experience. The derivation of substructure trends is largely manual in nature, aided by database systems and built-in SAR utilities. Either an extensive SAR or the activities of a few compounds can lead experienced medicinal chemists to recognize certain substructures that are necessary for binding to a receptor or diverse receptors in the same gene family. Such scaffolds or substructures were also termed 'privileged structures' [30], a concept that has been further developed in the design of combinatorial libraries [31]. When a ligand has more than one pharmacophore it is necessary to have a holder or linker substructure. The structural requirement of the holder or linker can be best understood from 3D pharmacophore geometry determination [26]. There are two types of uses for pharmacophores: (i) to find novel molecules with similar biological activity; (ii) to improve the biological activity and pharmacokinetic properties of a compound of interest. Both objectives can be achieved by replacing a pharmacophore with a similar pharmacophore or by modifying the nonpharmacophoric part. The use of pharmacophore modeling has increased enormously in recent years because of the availability of databases with millions of commercially available compounds. Protein-target class-based pharmacophore identification [32,33] (and its uses) is a routine procedure in modern drug discovery.

Molecular similarity search

Molecular similarity and diversity are two inversely related parameters widely used in database searching for acquiring compounds and generating libraries [34,35]. One obvious advantage of similarity searching over a pharmacophore-based search is that it does not require a set of structurally related compounds of similar biological activity to derive a commonality. Here, even one active molecule can be used to search a database for related compounds. Some scientists [36], however, raised the question whether compounds with a high similarity value really posses similar biological activity. It might be true that a small structural change could make a fairly potent compound an inactive one, but making similar compounds is the basis of medicinal chemistry for drug discovery research. Thus, similarity searching became a standard tool in drug discovery and most molecular modeling and database companies provide such tools. Several groups [37–39] explored the use of atomic physicochemical properties to evaluate molecular similarities. Such approaches are not yet commercially available but they are especially useful for identifying chemically different but biologically similar compounds.

The concept of a maximal common substructure and its derivation

The concept of an MCS is not only the most recent among the three molecular similarity approaches but is also available in several chemoinformatics packages (http://www.simulationplus.com/classpharmer/classpharmer.html; http://www.scitegic.com/products_services/pipeline_pilot.htm) for analyzing a set of active compounds and for identifying the derived MCS in molecular databases. Algorithms for MCS combine clique-detection and the clustering family of approaches, rather than a systematic fragmentation path taken by the molecular similarity approach. For a more in-depth method description, the works of Xu [40] and Raymond and Willett [41] can be consulted. The basic concept of MCS is illustrated in Figure 2. MCS seeks to arrive at a classification of structures without differentiating the extent of biological activity. All active compounds are treated equally and similar inactive compounds are not taken into account. A given class contains molecules that contain a common subgraph in which all vertices (i.e. atoms) and edges (i.e. bonds) are contiguously connected. This distinguishes it from 2D or 3D pharmacophore descriptions of molecules in which features are not necessarily directly connected.

Key aspects of MCS classification include desired class size and affiliation of a single structure to multiple classes. It is desirable to identify large common substructures and to have a sufficient number of molecules in a particular class. Recent examples of MCS usage, either alone or in conjunction with fingerprint clustering, show the value of the approach [42,43]. It has also been used to find the most common chemical replacements (i.e. bioisosteres) in druglike compounds [44].

Available software for maximal common substructure

MCS is a core component of the SimulationsPlus ClassPharmer program (http://www.simulationsplus.com/classpharmer/classpharmer.html); this implementation requires minimal user input to achieve meaningful classifications. Class size and the size of the substructure are dynamically linked and automatically balanced in response to dataset composition and user selection. Only qualitative input is required for the number of classes that a given compound can belong to. In addition, classifications can be limited to closed, complete ring structures and exact atom matches versus the more permissive fuzzy rings and atoms.

MCS can also be found as a component in Pipeline Pilot from SciTegic (http://www.scitegic.com/products_services/pipeline_pilot.htm); however, unlike ClassPharmer, Pipeline Pilot enables more user input, concomitant with larger variability in results. For example, it is necessary to specify a fraction of the reference set that must contain a common substructure to qualify it as an MCS. An 'optimal' classification might require fine-tuning various parameters in an iterative approach.

Application of maximal common substructure in virtual screening

A probable workflow scenario to assemble a knowledge-based screening collection using MCS analysis includes four steps:

(i) Curate a set of known active ligands from a molecular knowledgebase (http://www.mdl.com/products/knowledge/index.jsp).

(ii) Extract appropriate MCS subgraph information with a program of choice.

(iii) Examine MCS output to assess value of subgraphs for ligand design or database searching. Molecular fragments that are small (e.g. benzene) and ubiquitous do not provide much useful information because they are found in many ligands for all target classes. However, a medium size fragment (e.g. diphenyl methane for GPCRs or 2,6-diamino pyridines for kinases) can afford an adequate query or a starting point for chemistry. Larger fragments can also create problems. They might not be available in a commercial compound database. This step is somewhat subjective, as well as target- and

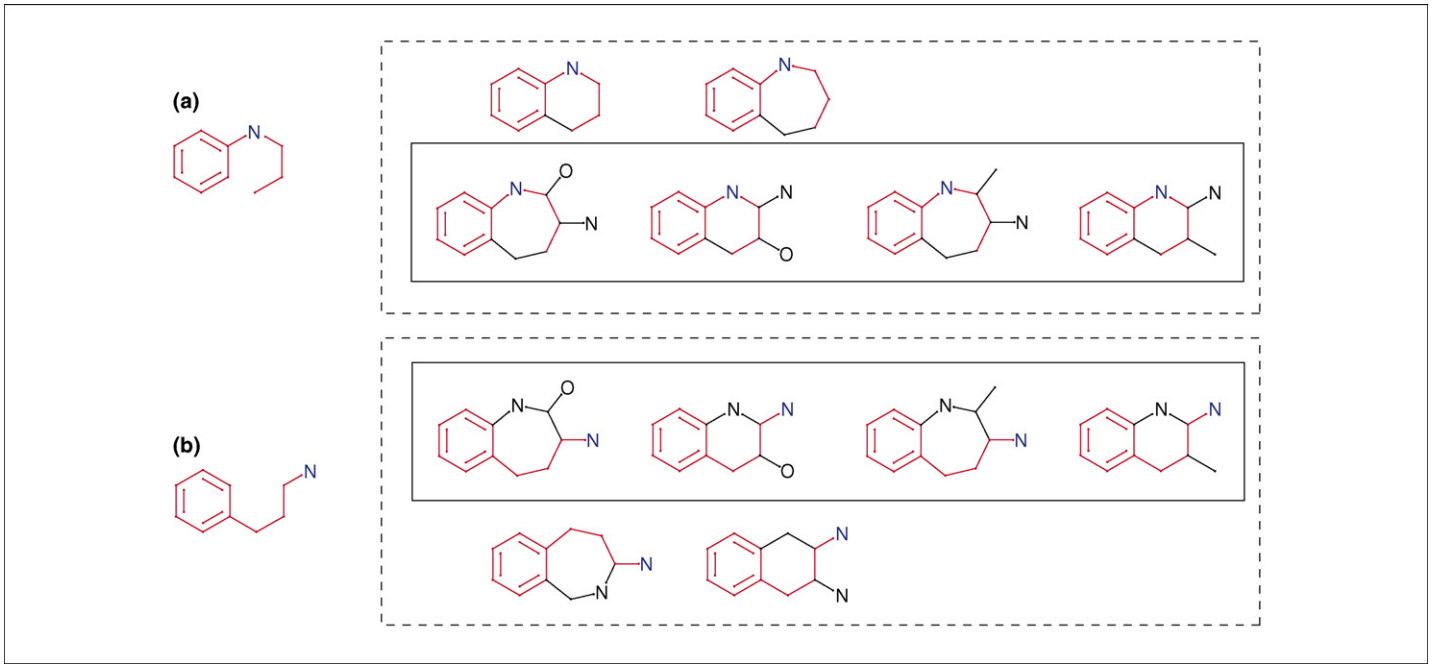

FIGURE 2

Maximal common substructures (MCSs) for a series of molecules. Two classes are highlighted, with their common subgraphs **(a)** and **(b)** and members of a class are shown inside dotted boxes; atoms and bonds not part of the MCS for that class are black, structures common to both classes are shown in solid boxes. The derived MCS depends on several constraints that these programs set in their analysis. One major component of the constraints is the fraction or number of compounds that should satisfy the substructure. The bigger the fraction the smaller the substructure becomes. MCSs that are too small are not useful for virtual screening because too many compounds will map to them. Conversely, MCSs that are too large might not be useful because few commercially available compounds will contain such a substructure.

project-specific, and can benefit from multiple iterations. A simple substructure search in readily available vendor databases should provide an answer as to whether too many, very few or an interesting number of compounds are returned.

(iv) The MCS-based approach usually concludes with a vendor database search for leadlike or druglike compounds, or the design of focused, target-biased compound libraries.

Maximal common substructure example on kinase inhibitors

The kinase ligands in the MDDR database were classified with a Pipeline Pilot protocol built around the MCS component, as well as in ClassPharmer (http://www.simulationplus.com/classpharmer/classpharmer.html). The ultimate aim of this analysis was to identify molecules for purchase that contain structural features associated with hinge-region-binding ability (i.e. having complementary H-bond-acceptor or -donor functionality). Both programs generated several subgraphs, containing known kinase-hinge-binder motifs, a few are shown in Figure 3a. In addition, both programs generated subgraphs that have no substructural features to be effective hinge binders (Figure 3b). Many of the latter examples result from the fact that kinase inhibitors, in addition to a hinge-binding core, also contain solubilizing groups and hydrophobic specificity and affinity functions. Although naphthyl, biphenyl and diphenyl methyl groups, among others, are ubiquitous in ligands for many target classes, particularly GPCRs, they alone are not relevant subgraphs for kinase-hinge-region binders. Likewise, MCSs containing basic amines and alkyl ethers probably derive from the solubilizing portion of the kinase inhibitors and not the hinge-binding core.

Undesired functional group screening

A compound with a reactive functional group can denature a protein or react with the reagents of the biological assay and can give false hits [45,46]. Prolonged storage of such compounds can also be a problem because samples can deteriorate over time. Many chemists would not include such compounds in their screening library. However, uniformity of opinion in the selection of a set of functional groups is not so clear or consistent [47]. The reactivity of a compound cannot always be predicted from the existence of a functional group in the molecule [48]. Also, promiscuity can arise from other properties [49]. A reasonable approach to define an undesired functional group could be knowledge-based, performing a substructure search on the approved drugs. If there are many nontoxic drugs approved with that functionality, it might be a good idea to remove the substructure from the 'undesired' list. Considering that reactivity often changes because of neighboring group effects, criteria are often a compromise between properties of a lead that cannot be 'fixed' versus losing a potential compound that might be 'repaired' by medicinal chemistry techniques. There is no unique solution to this issue.

Concluding remarks

To improve productivity, knowledgebase-guided decisions must be incorporated into the discovery and development process. The selection of appropriate knowledgebases is crucial to making an informed decision for selecting a lead or promoting a lead compound to the development phase. There is a major advantage in using *in silico* physicochemical property-prediction criteria

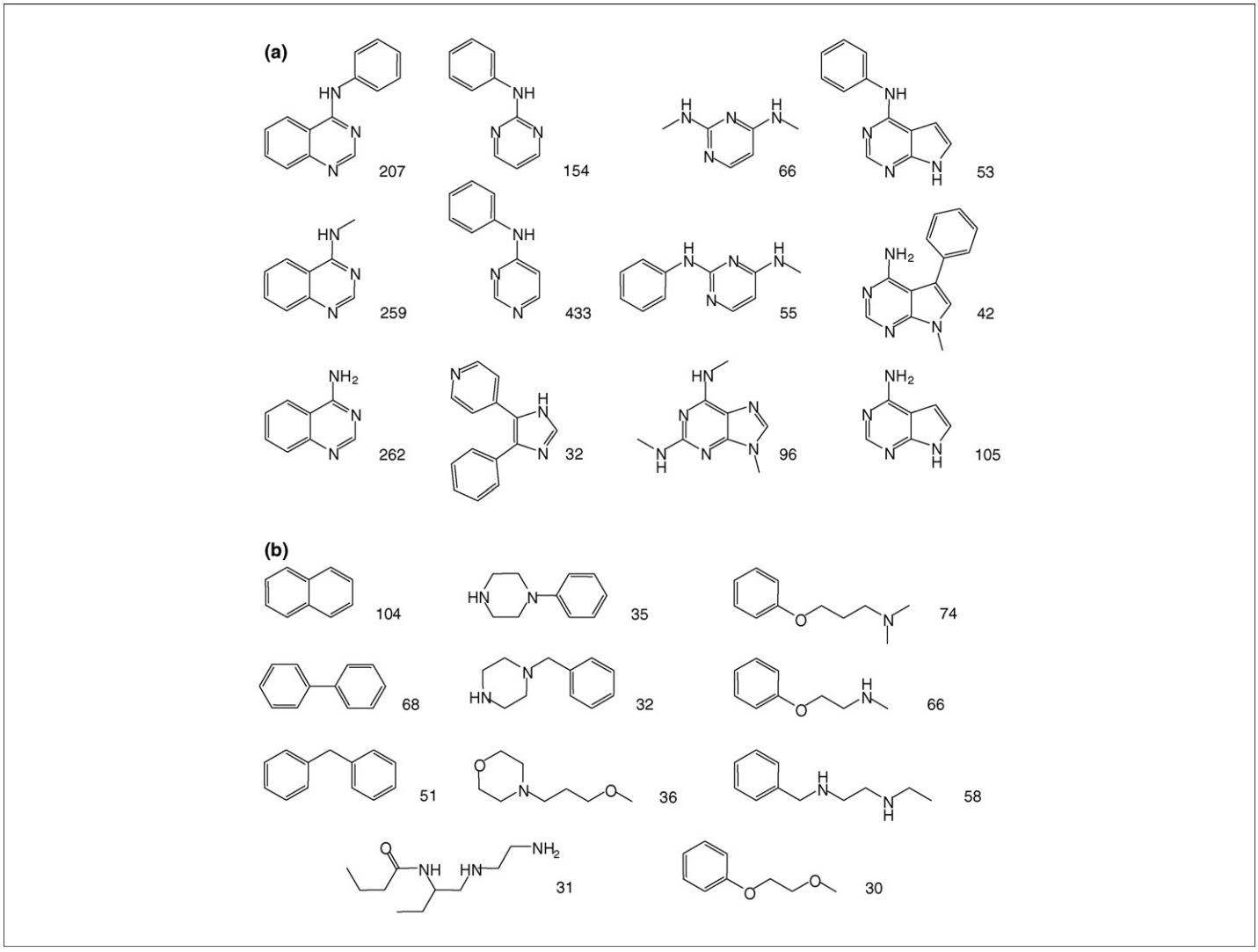

FIGURE 3

Maximum common substructures (MCSs) derived from known kinase ligands. (a) Examples of molecular graphs judged to be relevant MCSs derived within Pipeline Pilot for a series of kinase ligands; numbers indicate how many molecules of the set contain that core. **(b)** Molecular graphs judged not to be relevant MCS for hinge-binding affinity in kinase ligands. Most often, the second set of substructures are needed to improve the pharmacokinetic properties. These substructures alone cannot provide kinase inhibitory activity.

because this does not require synthesis of the molecules of interest. A model developed with *in silico* properties can be readily used in the design of future molecules. The quality of computational assessment of physicochemical properties is superior to ADME–Tox property assessment. It is a common experience that ADME–Tox properties are difficult to predict from a global model and the predicted properties should be used with caution [50]. Pharmacophore modeling, molecular similarity and MCS should be used extensively to assemble compound libraries for HTS. The majority of compounds in lead optimization should fall within the appropriate property range, derived from a set of successful reference compounds. However, small percentages of compounds should be made that have properties beyond this range so that hypotheses can be recalibrated and knowledge will grow with time. However, drug discovery is a complex but approximate science and chance factors can have a key role. By incorporating knowledge-based approaches researchers hope to bias the drug discovery process towards success and lessen the impact of detrimental chance factors.

References

1 Pharmaceutical Research and Manufacturers of America, Pharmaceutical Industry Profile 2005 (Washington, DC, March)

2 Bemis, G.W. and Murcko, M.A. (1996) Properties of known drugs. I. Molecular frameworks. *J. Med. Chem.* 37, 2887–2893

3 McGregor, M.J. and Pallai, P.V. (1997) Clustering large databases of compounds using MDL ''Keys'' as structural descriptors. *J. Chem. Inf. Comput. Sci.* 37, 443–448

4 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

5 Ajay, *et al.* (1998) Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* 41, 3314–3324

6 Sadowski, J. and Kubinyi, H. (1998) Atom-type based neural net model of drug-likeness. *J. Med. Chem.* 41, 3325–3329

7 Hansch, C. *et al.* eds (1990) *Comprehensive Medicinal Chemistry*, Pergamon Press, Oxford
8 Ghose, A.K. *et al.* (1999) A knowledge-based approach in designing combinatorial and medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drugs databases. *J. Comb. Chem.* 1, 55–68
9 Lajiness, M.S. *et al.* (2004) Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Devel.* 7, 470–477
10 Wenlock, M.C. *et al.* (2003) A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* 46, 1250–1256
11 Leeson, P.D. and Davis, A.M. (2004) Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* 47, 6338–6348
12 Baurin, N. *et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totaling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* 44, 643–651
13 Ghose, A.K. *et al.* (1998) Prediction of hydrophobic properties of small organic molecules using fragmental methods: an analysis of AlogP and ClogP methods. *J. Phys. Chem.* 102, 3762–3772
14 Lipinski, C.A. (2005) Filtering in drug discovery. In *Annual reports in computational chemistry* (Vol. 1) (Spellmeyer. *D. C., Ed.)*, pp. 155–168, Elsevier
15 Morphy, R. (2006) The influence of target family and functional activity on the physicochemical properties of preclinical compounds. *J. Med. Chem.* 49, 2969–2978
16 Pirard, B. and Pickett, S.D. (2000) Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.* 40, 1431–1440
17 Palm, K. *et al.* (1997) Polar molecular surface properties predict the intestinal absorption of drugs in human. *Pharm. Res.* 14, 568–571
18 Kelder, J. *et al.* (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* 16, 1514–1519
19 Ertl, P. *et al.* (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* 43, 3714–3717
20 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877
21 Ghose, A.K. *et al.* (2000) Quantitative structure and physicochemical property-based scoring scheme to evaluate drug-likeness of small organic compounds. Book of abstracts, 219th ACS national meeting, San Francisco, CA, March 26–30, CINF-002. American Chemical Society, Washington, D.C.
22 Vieth, M. *et al.* (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* 47, 224–232
23 Manallack, D.T. *et al.* (2002) Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* 42, 1256–1262
24 Gund, P. (2000) Evolution of pharmacophore concept in pharmaceutical research. In *Pharmacophore perception, development and the use in drug design* (Guner, O.F., ed.), pp. 3–12, IUL-press, La Jolla
25 Ghose, A.K. and Wendoloski, J.J. (1998) Pharmacophore modelling: methods, experimental verification and applications. *Perspect. Drug Discov. And Des.* 9, 253–271
26 Ghose, A.K. *et al.* (1995) Determination of pharmacophoric geometry for collagenase inhibitors using a novel computational method and its verification using molecular dynamics, NMR and X-ray crystallography. *J. Am. Chem. Soc.* 117, 4671–4682
27 Takeuchi, Y. *et al.* (1998) Derivation of three-dimensional pharmacophores model of substance P antagonists bound to the neurokinin-1 receptor. *J. Med. Chem.* 41, 3609–3623
28 Johnson, A.M. and Maggiora, G.M., eds) (1990) *Concepts and applications of molecular similarity*, Wiley, New York
29 Willett, P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996
30 Patchett, A.A. and Nargund, R.P. (2000) Privileged structures – an update. In *Annual reports in medicinal chemistry (Vol. 35)* (Doherty, A.M., ed.), pp. 289–298, Elsevier
31 Tounge, B.A. and Reynolds, C.H. (2004) Defining privileged reagents using subsimilarity comparison. *J. Chem. Inf. Comput. Sci.* 44, 1810–1815
32 Klabunde, T. and Hessler, G. (2002) Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* 3, 928–944
33 Prien, O. (2005) Target-family-oriented focused libraries for kinases – conceptual design aspects and commercial availability. *ChemBioChem* 6, 500–505
34 Chen, X. and Reynolds, C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.* 42, 1407–1414
35 Bender, A. *et al.* (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* 44, 1708–1718
36 Martin, Y.C. *et al.* (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358
37 Ghose, A.K. *et al.* (1989) Structural mimicry of adenosine by the antitumor agents 4-methoxy- and 4-amino-8-(β-D-ribofuranosylamino)pyrimido[5,4-d]pyrimidine as viewed by a molecular modeling method. *Proc. Natl. Acad. Sci. U. S. A.* 86, 8242–8246
38 Itai, A. *et al.* (1988) A receptor model for tumor promoters: rational superposition of teleocidins and phorbol esters. *Proc. Natl. Acad. Sci. U. S. A.* 85, 3688–3692
39 Wildman, S.A. and Crippen, G.M. (2001) Evaluation of ligand overlap by atomic parameters. *J. Chem. Inf. Comput. Sci.* 41, 446–450
40 Xu, J. (1996) GMA: a generic match algorithm for structural homomorphism, isomorphism and maximal common substructure match and its application. *J. Chem. Inf. Comput. Sci.* 36, 25–34
41 Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Design* 16, 521–533
42 Stahl, M. *et al.* (2005) A robust clustering method for chemical structures. *J. Med. Chem.* 48, 4358–4366
43 Schnur, D.M. *et al.* (2006) Are target-family privileged substructures truly privileged? *J. Med. Chem.* 49, 2000–2009
44 Sheridan, R.P. (2002) The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* 42, 103–108
45 Rishton, G.M. (1997) Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today* 2, 384–386
46 Pearce, B.C. *et al.* (2006) An empirical process for the design of high-throughput-screening deck filters. *J. Chem. Inf. Model* 46, 1060–1068
47 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing set of compounds. *J. Med. Chem.* 47, 4891–4896
48 Cusack, K.P. *et al.* (2004) A ^{13}C NMR approach to categorizing potential limitations of α,β-unsaturated carbonyl systems in drug-like molecules. *Bioorg. Med. Chem. Lett.* 14, 5503–5507
49 Feng, B.Y. and Shoichet, B.K. (2006) A detergent-based assay for the detection of promiscuous inhibitors from virtual and high-throughput screening. *Nature Protocols* 1, 550–553
50 Lombardo, F. *et al.* (2003) *In silico* ADME prediction: data, models, facts and myths. *Mini Reviews Med. Chem.* 3, 861–875