Journal of Computer-Aided Molecular Design, **16:** 311–323, 2002.
KLUWER/ESCOM
© 2002 *Kluwer Academic Publishers. Printed in the Netherlands.*

Filtering databases and chemical libraries

Paul S. Charifson and W. Patrick Walters
Vertex Pharmaceuticals, 130 Waverly St, Cambridge, MA 02139, USA

Introduction

In the current climate of high-throughput chemistry and screening, there are many compounds that can be synthesized and screened. Recent experience (and common sense) suggests, however, that not all compounds which can be synthesized or which are present in compound collections are worthy of screening on any given target. 'Hit-rates' are typically less than 1% [1] for HTS performed on pharmaceutical screening libraries. Although many strategies exist for attempting to improve 'hit-rates' (i.e. enrichment) including diversity approaches and focused library design, the simplest way of accomplishing this goal is to remove compounds which have a low probability of providing useful information at the 'lead-generation' stage of a drug discovery project. This reductionist approach utilizes one of the most fundamental scientific tenets: the process of elimination.

It would be fair to state that such filtration approaches, although useful, possesses significant limitations and are best used in conjunction with other techniques. For example, it is a common practice to remove all potential compounds possessing 'undesirable' functional groups and then perform a diversity analysis on the remaining compounds. Another example would be to take all compounds consistent with a defined chemistry with a ClogP less than some desired value and dock them into a protein binding site. One might then graphically evaluate only those compounds with 'favorable' contact scores and select a final set of compounds for synthesis and/or screening. In each of these cases, there is a degree of subjectivity employed in defining what is 'favorable' or 'undesirable'. These 'threshold' values are usually derived from experiences within a given organization or the collective experience across the pharmaceutical industry.

Another key issue related to *how* such filters are employed is *when* to use such filters. Typically, the types of filters discussed in this chapter are employed either upfront or at the backend of a design or selection cycle. The advantage to using these filters at the beginning of the process is that there are reduced numbers of compounds to consider at sequential steps of selection. Therefore, if more computationally demanding methods are employed after the initial filtration, the entire process is rendered more efficient. The disadvantage with this approach is that if the filters are too strict, compounds may be removed upfront which may possess other desirable features and which could be 'corrected' later on in a medicinal chemistry optimization cycle. If the type of filter being used is more speculative in nature (e.g. a binary classification scheme for oral absorption) then, it may be more useful to employ it after other techniques (e.g. docking into a protein binding site) have been performed. This may be preferable so that more intuition (or other information) can be used in the final stages of selection.

The use of any type of filter-based approach as a 'black-box' will inevitably lead to compounds being missed (false negatives). For example the argument is often made that most of these approaches would discard many natural-products or semi-synthetic analogs of natural products that are clearly useful medicinal agents. However, it must be clearly understood from the outset whether the goal is to identify a 'druggable' compound or a 'drug'. It is often argued that medicinal chemists can take a 'lead' compound and engineer 'drug-likeness' in later. Alternatively, others believe that one should start with a 'lead' which possesses global molecular properties in the 'ball-park' of existing drugs (e.g. lack of known toxic or metabolically labile functionality, adequate solubility and permeability characteristics, etc.). Thus, efficient use of simple filters to aid in the ultimate selection of compounds to synthesize and/or screen depends largely on the environment in which these filters are to be used.

Methods based on molecular connectivity

As was mentioned in the introduction, collections of screening compounds often contain a large number of molecules which are undesirable and should not be screened. One may wish to reject compounds for a variety of reasons including:

- calculated physical properties (molecular weight, logP, etc.),
- the presence of functional groups known to be reactive or toxic,
- the absence of functional groups known to impart biological activity.

In this section, we will discuss methods for rapidly removing compounds which have little chance of providing suitable leads in a drug discovery program. The techniques discussed here do not require the presence of 3D coordinates and are typically referred to as '2D filters'. This is actually a bit of a misnomer. These filters rely purely on molecular connectivity and can be just as easily applied to 1D notation such as SMILES [2, 3] strings. The first thing that should be pointed out is these filters can be calculated very rapidly, it is often possible to process hundreds of compounds per second on a typical workstation. This makes connectivity based filters an execellent precursor to more computationally expensive techniques such as docking which require 1 to 5 minutes per molecule. The second important point is that these filters can dramatically reduce the number of molecules which must be considered by 3D methods. As we will show below, these simple filters can often remove 50 to 75 percent of the compounds in commercial screening collections.

A. Property Filters

Many researchers over the years have attempted to show that drug-like molecules tend to have certain properties. For example, logP, molecular weight, and the number of hydrogen bonding groups have been correlated with oral bioavailability. A common approach to filtering is to limit selections to those molecules whose calculated properties fall within a specified range. Perhaps the most well known example of this approach is the 'Rule of 5', which was published in 1997 by Lipinski and coworkers [4]. The authors carried out an analysis of 2,245 compounds from the World Drug Index. Only those compounds with a USAN (United States Adopted Name) or INN (International Nonproprietary Name) and an entry in the 'indications and usage field' of the database were included in the analysis. The assumption is that compounds meeting these criteria have entered human clinical trials and therefore must possess many of the desirable characteristics of drugs. It was found that in a high percentage of compounds, the following rules were true: hydrogen bond donors ≤ 5, hydrogen bond acceptors ≤ 10, molecular weight ≤ 500, and logP ≤ 5. The majority of the violations came from natural products including antibiotics, antifungals, vitamins and cardiac glycosides. The authors suggest that, despite their violations of the 'rule of 5', such compounds are orally bioavailable because they possess groups which act as substrates for transporters.

The work carried out by Lipinski has been further validated by a similar analyses performed on other drug databases. Ghose and co-workers [5] examined the computed physical properties of 6,304 compounds taken from the Comprehensive Medicinal Chemistry Database. They established qualifying ranges which cover more than 80% of the compounds in the set. Ranges were established for AlogP [6], molar refractivity, molecular weight, and number of atoms. Oprea recently published a similar analysis of a number of other databases (MDDR, CMC, Current Patents Fast-alert, New Chemical Entities and ACD)[7].

Figure 1 shows the distributions of molecular weight and ClogP for six collections of commercially available screening compounds. The collections vary in size from 50,000 to 150,000. As a reference, we also show distributions for a subset of 5,115 compounds from the Comprehensive Medicinal Chemistry Database (CMC). The CMC contains compounds which have been used or studied as medicinal agents in humans. The CMC has been widely used as reference set in studies which seek to identify drug-like compounds. It is apparent from both figures that molecules in the commercial screening databases tend be much larger and more lipophilic than drug molecules. One could argue that the objective of a screening exercise is to identify a lead and not a drug, and that sub-optimal physical properties could be optimized in the course of a medicinal chemistry effort. However, a number of recent studies [8–10] have shown that the end result of a medicinal chemistry program tends to be a compound which is larger and more lipophilic than the original lead. These studies highlight the need for filtering programs which can identify lead molecules with desirable physical properties.

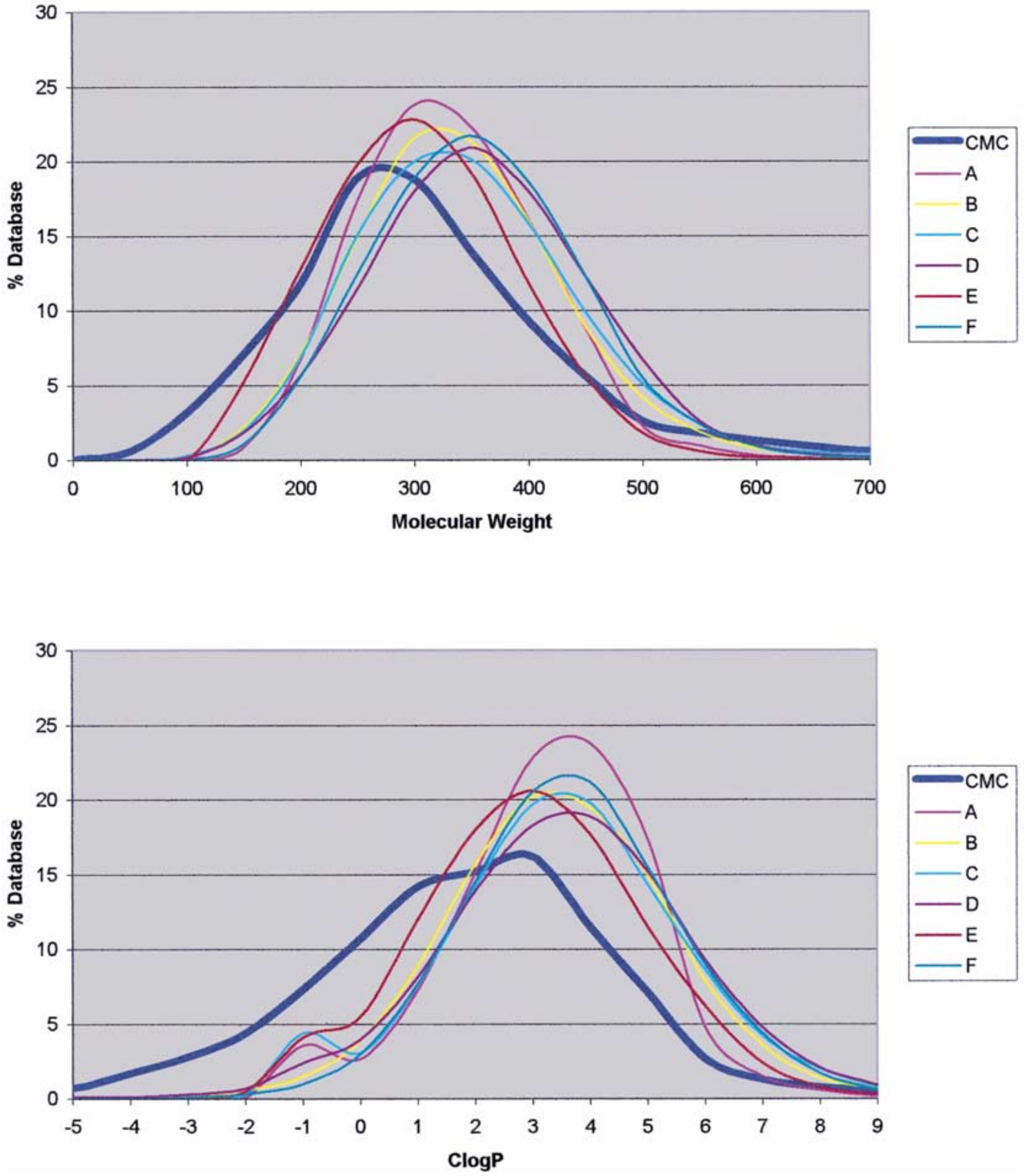

Figure 1. Distributions of molecular weight and ClogP for a drug database and six commercial screening collections.

B. Functional Group Filters

Another highly effective method of filtering large collections of compounds is to eliminate compounds containing moieties known to be problematic. A number of groups have published sets of rules for identifying compounds which may be toxic, reactive, or may interfere with biological assays [11–13]. One example of such a filtering approach is the REOS program developed at Vertex Pharmaceuticals. The program's primary function is to analyze a database of screening compounds and 'filter out' molecules which may be problematic (swill). REOS is a hybrid method which combines a set of functional group filters with some simple counting schemes similar to those in the rule of 5. Initial filtering is based on a set of 7 property

Table 1. The default property filters used by the REOS program

Property	Min	Max
molecular weight	200	500
LogP	−5	5
HB donors	0	5
HB acceptors	0	10
formal charge	−2	2
rotatable bonds	0	8
# heavy atoms	15	50

filters. The default values for these filters are shown in Table 1.

One may initially question the validity of setting a minimum value for the number of heavy atoms and the molecular weight. As mentioned earlier, the final product of a lead optimization effort is typically larger and more hydrophobic than the initial lead. Thus, it would be better to start with a small lead. However, while there are cases where very low molecular weight compounds have been found to be highly potent [14], these tend to be rare. In order to improve the odds of finding an inhibitor, REOS typically employs the molecular weight and heavy atom minima specified in Table 1.

In addition to the property filters, REOS allows the user to remove compounds using a set of more than 200 functional group filters. Rather than providing a simple 'accept/reject' facility, REOS allows the user to specify a maximum allowed quantity for each functional group. For instance, if one is selecting screening compounds, it is usually desirable to eliminate aldehydes due to their reactivity with biological nucleophiles. In this case, the maximum number of allowed aldehydes would be set to 0. However, if one is using aldehydes in the synthesis of a combinatorial library, it may be desirable to select reagents containing only one aldehyde. This can be easily accomplished by setting the maximum allowed value to 1. Examples of the functional group filters employed by REOS are listed in Figure 2.

In REOS, functional group filters are specified using the SMARTS [15] pattern matching language developed at Daylight Chemical Information Systems. SMARTS is an extended version of the SMILES [2, 3] notation developed specifically for substructure searching. For instance, an acid halide can be specified by the pattern C(=O)[F,Cl,Br,I]. In this example, the parentheses are used to indicate a branch while the square brackets are used to specify a set of options. With a simple modification, this pattern could be used to specify sulfonyl halides as well [C,S](=O)[F,Cl,Br,I]. More complex rules can also be defined for abstract atom types such as hydrogen bond donor or lipophilic atom. Variations of the SMARTS language are also available in Unity suite of programs from Tripos [16], the Molecular Operating Environment (MOE) from The Chemical Computing Group [17], and the open source OELib Toolkit from OpenEye Scientific Software [8].

Table 2 shows the results of a REOS analysis of seven databases. The first database consists of 5,115 compounds from the CMC database. The other six databases consist of commercially available screening compounds. The screening databases contain between 50,000 and 150,000 compounds. The first column (Passed Property Filters) shows the percentage of each database which passed the property filters. Parameters for this analysis were set to the default values listed in Table 1. Some readers may find it surprising that approximately 20% of the compounds in the CMC database do not meet the Lipinski-like criteria. Failures can primarily be attributed to compounds with molecular weight > 500 (613 compounds) and logP > 5 (683 compounds). A large percentage the high molecular weight compounds were either antibacterials or antineoplastics. This reflects the history of drug discovery efforts in these areas, which have traditionally been based on screening large collections of natural products. Many of the compounds with high logP values were from classes considered to be CNS active (antiparkinsonian, antipsychotic, antidepressant). This is consistent with the fact the CNS compounds tend to be more lipophilic than other biologically active molecules [18].

The second column in Table 2 (Passed FG Filters) shows the number of compounds from each database that passed the functional group filters. The rejection rate for the CMC is considerably smaller than that of the screening databases. However, it is surprising that approximately 25% of the CMC compounds were eliminated by the functional group filters. The largest number of rejections (185) was due to the presence of nitro groups. Nitro groups tend to activate aromatic rings [19] and may increase a molecule's tendency to generate false positives under assay conditions. In addition, nitro compounds tend to be colored and may interfere with assays which employ a spectrophotometric readout. Another 125 compounds were

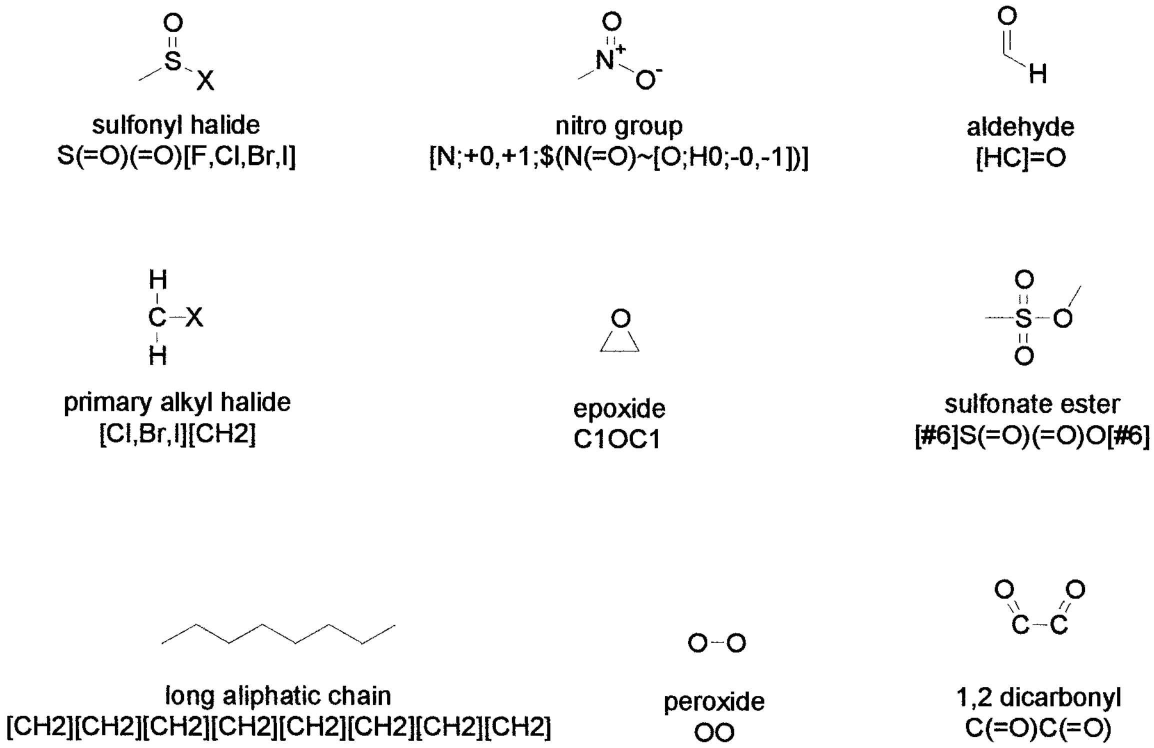

Figure 2. Some of the functional group filters used by the REOS program and the corresponding SMARTS representations.

rejected because they contained atom types other than C,O,N,S,P,F,Cl,Br,I,Na,K,Mg,Ca, and Li. The majority of these rejections were due to antineoplastics containing Pt and As, antacids containing Al or Si, or vitamins containing Fe and Co.

The third column in Table 2 (Passed All Filters) shows the intersection of the two lists described above, i.e., the compounds which passed both the Rule of 5 filters and the functional group filters. It is quite striking that in every case more than half the molecules are eliminated.

Techniques such REOS can also be extremely valuable when applied to the design of combinatorial libraries. When analyzing a compbinatorial library, the REOS analysis is typically carried out in three steps.

1. Filter the reagents. In this step reactive and toxic reagents are removed. Reagents which will clearly create products that violate the molecular weight limits are also removed.
2. Check the reagents for compatibility with the chemistry. For example, when synthesizing amides one can simplify the chemistry by removing acids containing basic amines and amines containing acidic functionality.
3. Filter the products. This allows a final consideration of properties such a logP which cannot be calculated from reagents in a straightforward fashion. In addition, this step is necessary for functional group filters which specify a maximum count. For instance, let us suppose we have rule which only permits molecules with fewer than 3 bromine atoms. Two reagents having 2 bromine atoms each would pass the filter, but their product having 4 bromine atoms would fail.

The number of compounds eliminated will vary with the composition of the library. In the case of large multi-component screening libraries we often see a ten-fold reduction in the size of the library. When dealing with smaller focused libraries the number of compounds removed can be quite small.

Tools such as REOS can provide a significant reduction in the number of compounds that should be screened. Careful filtering can also decrease the amount of time required for followup screens by reducing the number of false positives. While functional group filtering is highly effective, one must avoid the temptation to become overzealous. It is important to periodically examine the results of a filtering analy-

Table 2. A REOS analysis of a drug database and six commercial collections of screening compounds.

	% Passed Property Filters	% Passed FG Filters	% Passed All Filters
CMC	81	76	63
vendor A	74	57	43
vendor B	68	48	35
vendor C	66	50	35
vendor D	63	52	34
vendor E	77	65	51
vendor F	66	62	35

Table 3. Functional groups used in the scoring scheme developed by Andrews et al.

	Score
Charged Groups	
Carboxylate	8.2
Phosphate	10
N+	11.5
Polar Groups	
N	1.2
OH	2.5
CO	3.4
O or S ethers	1.1
Halogens	1.3
Nonpolar Groups	
C (sp2)	0.7
C (sp3)	0.8

Table 4. Functional groups used in the scoring scheme developed by Muegge et al.

amine*
amide
alcohol
ketone
sulfone
sulfonamide
carboxylic acid*
carbamate
guanidine*
amidine*
urea
ester

sis to insure that 'interesting' molecules are not being inadvertently eliminated.

C. Identifying 'Interesting' Molecules

Another productive filtering approach is to eliminate compounds which do not posses 'interesting' functionality. While the exact definition of interesting functionality is subjective, medicinal chemists typically agree that the functional groups listed in Tables 3 and 4 tend to impart biological activity. A relatively simple means of filtering is to score molecules based on the presence of these functional groups and eliminate molecules with low scores. One of the earliest examples of this approach is the work of Andrews and coworkers [20]. In this paper, the authors used a set of 200 drug molecules to derive a set of 'intrinsic binding energies' for the 10 functional groups shown in Table 3. The inherent binding affinity of a small molecules was then estimated by summing the intrinsic binding energies and subtracting an entropic factor. While this method was not originally designed as a filtering tool, it has been widely used for this purpose [12].

A more recent example of the 'functional group' approach to filtering is the work of Muegge and coworkers [21]. In this work, each molecule is assigned a score based on the presence of structural fragments typically found in drugs. The fragments used in this study are listed in Table 4. A molecule is given one point for each non-overlapping pharmacophoric element. Molecules with a score between 2 and 7 are classified as drugs, otherwise they are classified as non-drugs. The authors note that many drugs affecting the central nervous system are relatively small and only contain a single pharmacophoric group. A second filter is defined in order to avoid classifying these molecules as non-durgs. Compounds containing a single pharmacophoric group can only be classified as drugs if they contain one of the groups marked with an asterisk in Table 4.

D. Efficiency Considerations

While the methods described above are computationally efficient, calculating properties for a large combinatorial library can still be time consuming. For instance, calculating the 'rule of 5' parameters for a set of 10,000 molecules takes about 2 minutes on a typical workstation. Consequently, the same calculation for a library of a billion molecules would require more than two days. In the case of combinatorial libraries, the time requirements can be greatly reduced. It is often possible to estimate the properties of the combinatorial products based on the properties of the building blocks used in the reaction. By carrying out the calculation in this fashion, the time requirements scale as sum of the number of building blocks rather than the product. This procedure is straightforward for simple properties such as molecular weight or number of hydrogen bonding groups. However, calculating other properties such as logP can be more complicated due to the fact the fragments used in the calculation may span more than one combinatorial building block.

Barnard and coworkers describe an approach to rapid property calculation in which they create a Markush representation of a combinatorial library from a reaction scheme and a set of precursors [22]. This representation is then used to rapidly calculate SMILES, molecular fingerprints, and 'rule of five' properties for a set of combinatorial products. Using this procedure, the authors are able to calculate properties at a rate of approximately 10,000 compounds per second.

An alternative approach to rapid calculation of molecular properties was published by Shi and coworkers [23]. This paper presents two complimentary approaches. The 'direct reactants' method can be employed when the product properties do not depend on fragments which span multiple combinatorial building blocks. This method was applied to properties such as molecular weight, number of hydrogen bonding groups, number of ionizable groups and SLOGP. In order to deal with the interaction of neighboring fragments, the authors developed a second method which utilizes a set of 'basis products'. These basis products are formed by combining each reactant with its simplest possible reaction partner. For instance, a set of basis products for an amide reaction could be formed by combining amines with acetic acid and combining acids with methyl amine. The basis products method was used to calculate solvent accessible volume and surface area as well as CLOGP. The correlations (R^2) between the properties calculated on whole molecules and those calculated using the two methods described here ranged between 0.90 and 0.99. The authors also describe an efficient tree searching algorithm which allows the rapid selection of combinatorial products meeting a desired property profile.

E. Other Methods

The majority of the methods discussed above were developed by translating the collected knowledge of scientists involved in drug discovery into a computer program. While this approach has considerable merit, it is dependant on the availability of local experts and often introduces a considerable amount of human bias into the predictions. An alternate approach taken by some researchers is to present a computer program with a set of labeled examples of drugs and non-drugs and allow the program to 'learn' to distinguish the two classes. The development of computer programs which can learn classification rules forms the basis of a branch of Computer Science known as Machine Learning. Machine learning programs operate by examining a set of training examples, each of which is assigned to belong to a particular class. A learning program then derives a rule or set of rules that assign new examples to these classes [24, 25].

In addition to being able to correctly classify molecules as drugs or non-drugs, it would be advantageous for a machine learning program to express its rules in a form that could be easily understood by humans. Both Ajay [26] and Wagener [27] have published papers which describe the use of inductive machine learning programs to derive a 'decision tree' for distinguishing drugs from non-drugs. These decision trees can be used to construct a set of human readable rules which classify molecules with a 70–80% accuracy. In the work published by Ajay, the training and test sets consisted of 3500 compounds each from the CMC (drugs) and the ACD (non-drugs). The machine learning program C4.5 [28] was used with a of set of seven one-dimensional descriptors to produce a decision tree. The decision tree was able to correctly classify approximately 80% of the CMC compounds and approximately 70% of the ACD compounds.

Neural networks [26, 29] are another automated classification method which has received a great deal of recent attention. This technique uses a highly connected network (modeled after a biological nervous system) to create an output classification based on a set of input values. The neural network is initially presented with a set of labeled examples and a set attributes describing these examples. The program then 'learns' the relationship between these variables and the desired output by creating nonlinear relationships between the attributes. A neural network designed to identify drug-like compounds is typically presented with a set of molecules labeled as drug (1) or non-drug (0) and a set of descriptors for each molecule. The program then learns relationships between the descriptors and uses these relationships to assign a score between 0 and 1 to each molecule. Those molecules with scores closer to 1 are classified as drugs while those with scores closer to 0 are classified as non-drugs.

One advantage to this approach is that it allows the user to assign a confidence to each prediction. A molecule with a score 0.95 can be assumed to have a higher probability of being a drug than molecule with a score of 0.55. This approach also allows one to adjust the number of false positives or false negatives produced by the neural network. Raising the threshold above which a compound is considered a drug causes the number of compounds predicted as drugs to decrease but also reduces the number of false positives. Conversely, decreasing the threshold causes more compounds to be classified as drugs but also increases the number of false positives.

Three-dimensional filtering methods

In this section we consider methods in which compounds are represented as three-dimensional structures and are docked into protein binding sites. The primary goal of these methods is to reduce the number of molecules considered for experimental evaluation. This is clearly of fundamental importance when one is considering large numbers of possible candidates that might be present in the design of large chemical libraries or even searching large corporate databases. The first part of this section describes approaches that either attempt to limit the ultimate number of compounds considered based upon either energetics or maintenance of established key interactions. Additionally, this section addresses methods that reduce the ultimate time needed for efficient orientational searching within protein binding sites. The second part of this section primarily describes using pharmacophores as prescreens for docking.

Structure-based approaches

A. Scoring

When one attempts to calculate the strength of interaction between two molecules and use this criterion for selection purposes they are, in fact, implementing a simple filtering process. These 'scoring functions' are typically used to rank large sets of compounds as part of a virtual screening process against a macromolecular target with an experimentally determined three-dimensional structure [30]. In practice, this is usually accomplished by utilization of a 'cutoff' or 'threshold' value of some empirical [31], knowledge-based [32], or energetic function [33] which crudely approximates the thermodynamics of binding. One example of this was given by Shoichet, et al. [34] in an attempt to discover novel inhibitors of thymidylate synthase (TS). Commercially available compounds were docked into TS and orientations of each molecule were filtered for steric fit. All docked orientations that did not possess any bad contacts were then evaluated for their non-bonded interaction energy. The best scoring orientation of each molecule was determined and a ligand solvation correction was applied for the final ranking. Of the approximately 1000 ligands docked in such a manner, the top scoring compounds were tested for their ability to inhibit TS. Several compounds were identified from this process with an $IC_{50} < 10\ \mu M$.

B. Interaction monitoring

In addition to intermolecular energies or scores as filters, the next most fundamental approach entails insuring that only compounds that make specific types of interactions can pass on to the next stage of evaluation. Examples of this would include: maintenance of a key hydrogen bond or hydrophobic interaction between the macromolecule and ligand, ligand-metal interaction, volume overlap with a known ligand, and subsite occupancy. In our own work, we have found significant improvement in confirmed hit rates and identification of new scaffold classes by employment of such filters. In some of our docking studies into inosine monophosphate dehydrogenase, we found that considering only ligands capable of making a key hydrogen bond improved our overall hit-rates from approximately 2% to 5%. Additionally, in our jnk-3 kinase virtual screening efforts, we only considered compounds both capable of making a conserved backbone hydrogen bond as well as possessing a volume overlap of $> 70\%$ with the original crystallographic ligand. This approach resulted in a hit rate of $\sim 7\%$ (i.e. 7 out of 106 compounds tested had an $IC_{50} < 50\ \mu M$) and represented five novel scaffold classes. The above types of filtering can be carried out either as a post-processing step or can be integrated as a part of the program. For example, in DOCK [33, 35–37] version 3.5, the concept of critical clusters was introduced. This feature enables the program to force matching to include members from a particular group (or groups) of site points in every match. This type of simple filter is not only useful to focus docking around a key residue or residues in an active site, but also dramatically decreases the number of possible ligand-receptor matches and results in increased program speed.

C. Pose reduction in protein-ligand docking

Some of the approaches described above could be used either to select viable orientations of a docked molecule or may be used to choose among different molecules. Stahl and Bohm [38] have addressed the former issue by applying several useful filter functions to remove all docked poses with certain unfavorable properties. A pose is defined as the orientation of a given ligand conformation within the protein binding site. For each pose generated for a protein-ligand complex, they have calculated the following four properties: the fraction of the ligand volume buried inside the binding pocket, the size of lipophilic cavities (voids) along the protein-ligand interface, the solvent-accessible surface area of nonpolar parts of the ligand, and the number of close contacts between nonhydrogen-bonded polar atoms of the ligand and the protein (repulsive polar interactions). These four terms are used to filter out the majority of the calculated solutions in a post-processing implementation. The remaining solutions can then be rescored with any other metric one deems appropriate. The above terms are sometimes referred to as 'penalty functions' and when taken together reflect the fact that high affinity protein-ligand complexes typically possess a high degree of steric complementarity. It should not be expected that any of the four filters mentioned above would perform well as a standalone scoring function or possess any ability to rank different molecules. However, it would be straightforward to implement any combination of these filters as part of a pre-scoring approach to reduce the number of poses being considered by a more time-consuming function (e.g. an energy-based evaluation).

The computation of void volumes is another useful method for pose reduciton. Figure 3 graphically depicts the void volume or the volume of 'unfilled' lipophilic space between the protein and ligand surfaces for Rosiglitazone bound to PPARγ[39]. In this figure, the thickness of void region is color coded with blue representing the area of greatest thickness. One might use this type of graphical representation as part of a lead optimization cycle by looking for blue areas of the void surface where synthetically accessible substituents might be placed. This simply helps focus attention on areas of the protein-ligand interface where increased substitution may lead to improved binding affinity. We have also found void volumes to be useful as a nongraphical pre-scoring filter to reduce the number of poses one needs to consider for actual scoring. We analyzed sixty-nine high-resolution crystallographic systems predominantly comprised of kinases and proteases in which corresponding ligands were docked in to their respective proteins and 200 unique poses were saved for each system. For each of these systems, we then calculated the void volumes for all poses and saved the fifty poses with the smallest void volumes for each system. We found that in 70% of the cases a docked pose was within 1.0 Å RMSD of the crystallographic pose while in 80% of the cases a docked pose was within 1.5 Å RMSD of the crystallographic pose. This means that we can rapidly discard 75% of the poses up front prior to scoring. If the scoring method we were using involved an energy evaluation or minimization, this reduction in the number of poses considered would result in a significant time savings. In such instances where an energy-based method is intended to select the final pose one could also use rapid empirical functions upfront to reduce the number of poses ultimately considered.

In the same sixty-nine study systems referred to above, we found that several empirical methods are adequate to triage the number of considered poses down to ten or so poses which have a reasonable likelihood of possessing a pose close to the crystallographic pose. Figure 4 shows the result after further cluster-

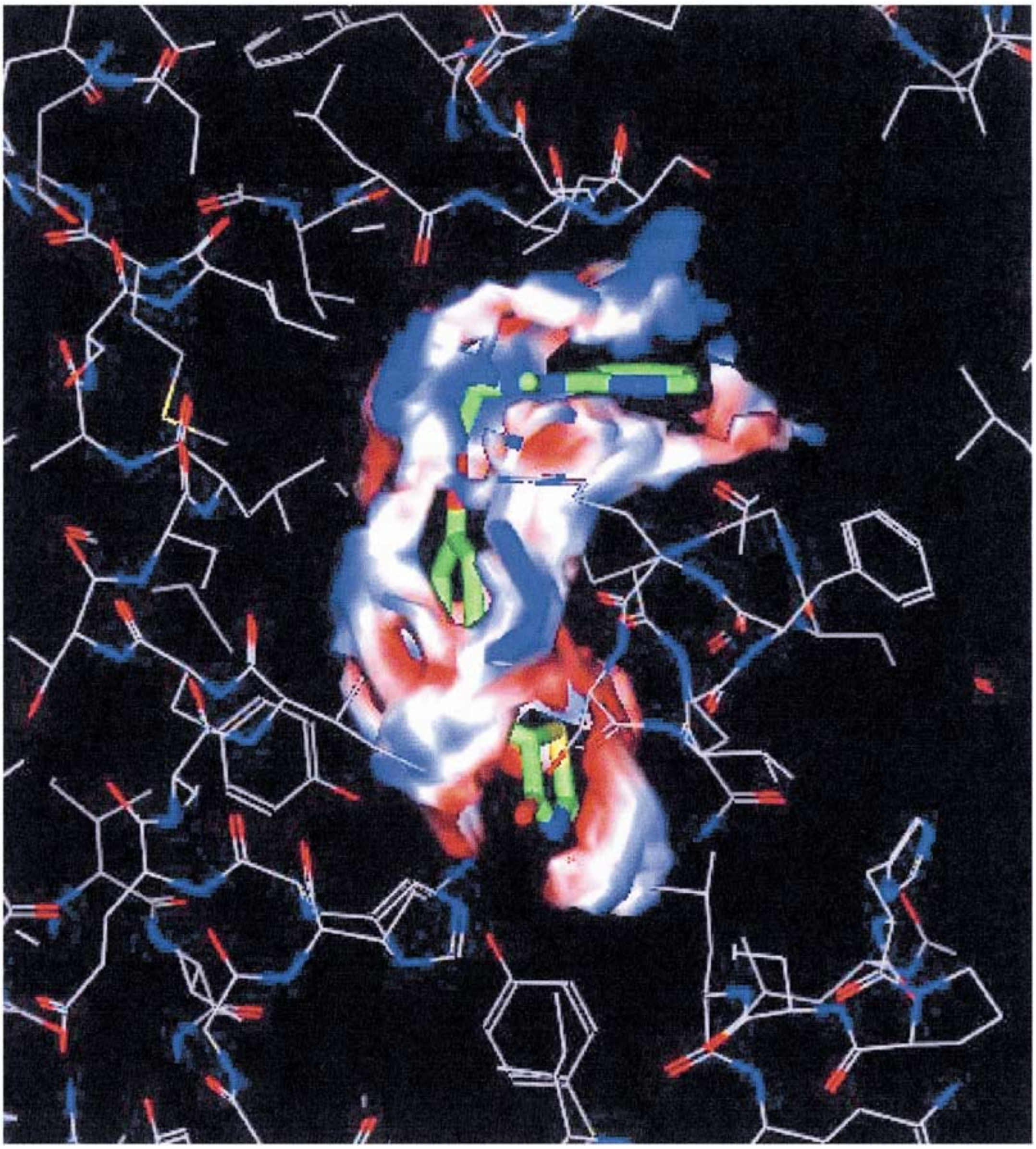

Figure 3. Void volume for the protein-ligand interface of Rosiglitazone bound to PPARγ. the thickness of the void volume is color coded such that the thickest portions of the surface are blue and the thinnest portions are red, with white being of intermediate thickness. The void volume was calculated and displayed by the program, VIDA [50].

ing of the set of 200 poses against the scores for each empirical function. This suggests for all of the empirical functions shown, with the exception of PMF [32], that the top ten scoring poses contains a pose within 1.5 Å of the crystallographic pose. This corresponds to roughly a 95% reduction in the total number of poses that would have to be considered for a more expensive energy evaluation. Figure 4 also suggests that for the piecewise linear potential (PLP) [40] that the top 5 scoring poses find a pose within 1.5 Å of the crystallographic pose 84% of the time. Thus a function such as PLP might provide a useful approach for triaging large numbers of poses down to smaller numbers of poses to be considered for more rigorous evaluation.

One area that is poorly addressed in the area of pose reduction and scoring, in general, is that of penalties for mismatches. In the simplest form, a simple bump check attempts to remove poses that clash

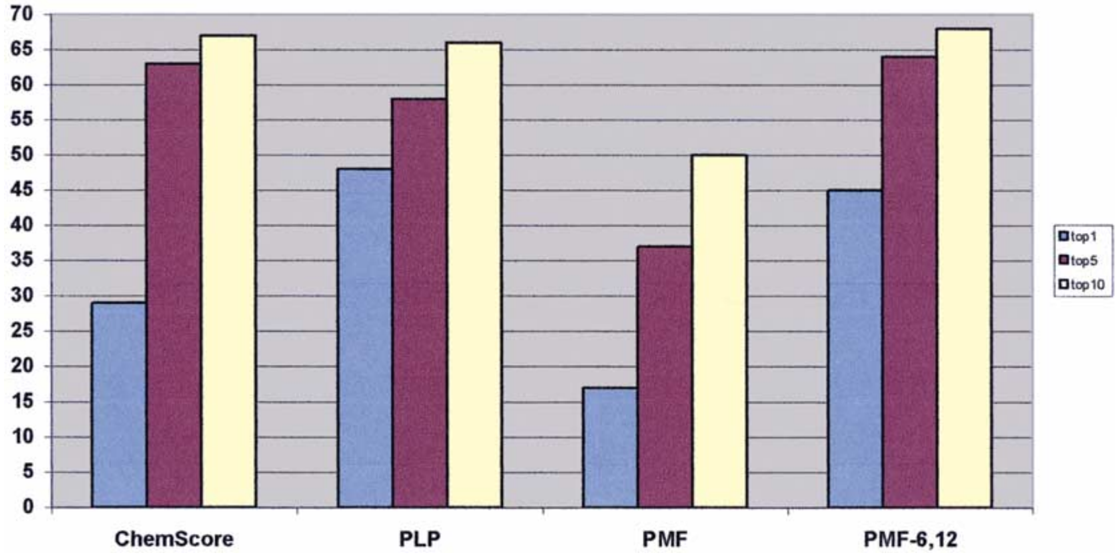

Figure 4. Performance of several scoring functions to correctly identify the crystallographic pose. The corresponding crystallographic ligands were removed from sixty-nine high resolution crystal complexes and docked back in to their respective proteins using the program FRED [51]. The top 200 poses (using ChemScore) were kept for each system and then rescored with the functions shown and then clustered against each scoring function using a clustering threshold of 1.5 Å. The X-axis thus denotes the portion of the sixty-nine complexes for which the crystallographic pose was correctly identified within the top 1, 5 or 10 scoring poses for each function. PLP is the piecewise linear potential [40], PMF is the Potential of Mean Force [32], and PMF-6, 12 is Lennard-Jones 6-12 potential fitted to the PMF data [52].

with a respective binding site. Typically, these types of simple estimations of intermolecular repulsion are either under or overestimated. Although some formal VdW potentials can do a better job of estimating repulsions, they are much slower when considering thousands of poses for large numbers of molecules. As stated above, Stahl and Bohm attempted to account for repulsive polar interactions in their filter function work [38]. Unfortunately, there are few other examples of mismatch penalties in this field. In earlier work, Bohm [41] also included a similar type of penalty for repulsive polar interactions in his empirical scoring function. More recently, one of he available scoring functions used by the docking program, Glide [42] (i.e. GlideScore) that is based on ChemScore [43] utilizes buried polar terms to penalize for electrostatic mismatches. Clearly, there is much opportunity in this area for further enhancement.

Ligand-based approaches

In this section, we describe filtering approaches in which some description of ligand features deemed necessary for binding are used as a prescreen for further model development. This primarily involves determination of some key pharmacophoric elements, searching a 3D database for compounds that possess these elements and then docking these molecules into a protein binding site. It is well established to use similarity searching based on known actives and then dock those molecules into proteins [34]. It also common to dock known actives into a binding site and then derive a pharmacophore [44]. When considering large numbers of compounds, however, such as in chemical libraries, it might make sense to use any knowledge upfront as a prescreen for docking. If sufficient knowledge exists to determine a three-dimensional pharmacophore, this information can contribute to a significant increase in efficiency. In this context, the pharmacophore can be extracted from ligand features of known active compounds or from the binding site requirements, if a ligand-bound structure is available [45–48]. It is inherently faster to perform a preliminary 3D-database search based on a pharmacophore then to dock the same molecule into a protein active site. In some cases, a pharmacophore key derived from a ligand can be compared with a pharmacophore key of its target-binding site without ever docking the ligand into the site [49]. This approach was applied to combinatorial library design to identify inhibitors of thrombin and Factor Xa [48]. In this design, reagent combinations were chosen in an attempt to maximize the total number of pharmacophoric matches between multiple pharmacophores derived from the structure of the binding sites and the virtual four-component library. This approach insures that the maximal number of binding modes and subsites within the binding site are searched by the final reagent selection.

Conclusions

Ultimately, the best filter of all is that of human intervention. Our pattern recognition abilities allow for differing amounts of experience and intuition to be employed. This is quite difficult to capture in the context of a computer program. Nonetheless, when considering large numbers of compounds, the types of approaches outlined in this chapter have proven useful to reduce the number of ultimate compounds considered. Thus, the final selection can be made from a smaller set with increased probability of having the desired activity and that has a maximal degree of overlap with the properties associated with known drugs.

References

1. Thorpe, D.S., Chan, A.W.E., Binnie, A., Chen, L.C., Robinson, A., Spoonamore, J., Rodwell, D., Wade, S., Wilson, S., Ackerman-Berrier, M., Yeoman, H., Walle, S., Wu, Q., Wertman, K.F., Biochem. Biophys. Res. Commun., 266 (1999) 62–65.
2. Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31–36.
3. Weininger, D., Weininger, A. and Weininger, J.L., J. Chem. Inf. Comput. Sci., 29 (1989) 97–101.
4. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug Delivery Rev., 23 (1997) 3–25.
5. Ghose, A.K., Viswanadhan, V.N. and Wendelowski, J.J., J. Comb. Chem., 1 (1999) 55–67.
6. Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J., J. Phys. Chem. A., 102 (1998) 3762–3772.
7. Oprea, T.I., J. Comput. Aided Mol. Des., 14 (2000) 251–264.
8. Lipinski, C.A., J Pharmacol. Toxicol. Methods, 44 (2000) 235–249.
9. Oprea, T.I., Davis, A.M., Teague, S.J. and Leeson, P.D., J. Chem. Inf. Comput. Sci., (2001)
10. Hann, M.M., Leach, A.R. and Harper, G., J. Chem. Inf. Comput. Sci., 41 (2001) 856–864.
11. Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L. and Ramsden, N., J. Chem. Inf. Comput. Sci., 39 (1999) 897–902.
12. Leach, A.R., Bradshaw, J., Green, D.V.S., Hann, M.M. and Delany, III, J.J., J. Chem. Inf. Comput. Sci., (1999) 1161–1172.
13. Rishton, G.M., Drug Discovery Today, 2 (1997) 382–385.
14. Kuntz, I.D., Chen, K., Sharp, K.A. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 96 (1999) 9997–10002.
15. James, C.A., Weininger, D. and Delany, J., Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
16. Unity, Tripos, St. Louis, MO, http://www.tripos.com
17. Molecular Operating Environment, Chemical Computing Group, Montreal, Quebec, Canada, http://www.chemcomp.com
18. Ajay, Bemis, G.W. and Murcko, M.A., J. Med. Chem., (1999) 4942–4951.
19. March, J., Advanced Organic Chemistry : Reactions, Mechanisms, and Structure, John Wiley & Sons, 1992
20. Andrews, P.R., Craik, D.J. and Martin, J.L., J. Med. Chem., 27 (1984) 1648–1657.
21. Muegge, I., Heald, S.L. and Brittelli, D., J. Med. Chem., 44 (2001) 1841–1846.
22. Barnard, J.M., Downs, G.A., von Scholley-Pfab, A. and Brown, R.D., J. Mol. Graph. Model., 18 (2000) 452–463.
23. Shi, S., Peng, Z., Kostrowicki, J., Paderes, G. and Kuki, A., J. Mol. Graph. Model., 18 (2000) 478–496.
24. Weiss, S.M. and Kulikowski, C.A., Computer Systems That Learn, Morgan Kaufmann, 1991.
25. Mitchell, T.M., Machine Learning, McGraw Hill, 1997.
26. Ajay, Walters, W.P. and Murcko, M.A., J. Med. Chem., 41 (1998) 3314–3324.
27. Wagener, M. and van Geerestein, V.J., J. Med. Chem., 40 (2000) 280–292.
28. Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
29. Sadowski, J. and Kubinyi, H., J. Med. Chem., 41 (1998) 3325–3329.
30. Stahl, M. and Rarey, M., J. Med. Chem., 44 (2001) 1035–1042.
31. Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., J Med Chem, 42 (1999) 5100–5109.
32. Muegge, I. and Martin, Y.C., J. Med. Chem., 42 (1999) 791–804.
33. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comp. Chem., 13 (1992) 505–524.
34. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445–1450.
35. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Molec. Biol., 161 (1982) 269–288.
36. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722–729.
37. Shoichet, B.K., Bodian, D.L. and Kuntz, I.D., J. Comp. Chem., 13 (1992) 380–397.
38. Stahl, M. and Bohm, H.J., J. Mol. Graph. Model., 16 (1998) 121–132.
39. PDB refcode 2PRG
40. Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B. and Freer, S.T., Chem. Bio., 2 (1995) 317–324.
41. Böhm, H.-J., J. Comput.- Aided Mol. Design, 12 (1998) 309–323.
42. First Discovery Technical Notes Manual, version 1.8, Schrodinger, Inc., New York, NY,
43. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.PP., J. Comput.-Aided Mol. Design, 11 (1997) 425–445.
44. Volter, K.E., Embrey, K.J., Pierens, G.K. and Quinn, R.J., Eur J Pharm Sci, 12 (2001) 181–194.
45. Clark, D.E., Westhead, D.R., Sykes, R.A. and Murray, C.W., J. Comput.-Aided Mol Des, 10 (1996) 397–416.
46. Moro, S., Li, A.H. and Jacobson, K.A., J. Chem. Inf. Comput. Sci., 38 (1998) 1239–1248.
47. Murray, C.M. and Cato, S.J., J. Chem. Inf. Comput. Sci., 39 (1999) 46–50.

48. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C. and Labaudiniere, R.F., J. Med. Chem., 42 (1999) 3251–3264.
49. Mason, J.S. and Cheney, D.L., Pac. Symp. Biocomput., (2000) 576–587.
50. VIDA, OpenEye Scientific Software, Santa Fe, NM, http://www.eyesopen.com
51. FRED, OpenEye Scientific Software, Sante Fe, NM, http://www.eyesopen.com
52. Pearlman, D.A. and Walters, W.P., manuscript in preparation.