Identification and Selection of "Privileged Fragments" Suitable for Primary Screening

Eleonora Gianti

Computational Sciences Group, Department of Chemistry (Congenia s.r.l.), Genextra S.p.A., Milan MI 20100, Italy

Luca Sartori*

Computational Sciences Group, Department of Chemistry (DAC s.r.l.), Genextra S.p.A., Milan MI 20100, Italy

Received June 30, 2008

The use of small molecule libraries for fragment-based primary screening (FBS) is a well-known approach to identify protein binders in the low affinity range. However, the search, analysis, and selection of suitable screening fragments can be a lengthy process, because of the large number of compounds that must be analyzed for different levels of ring/substituents identification and submitted to selection/exclusion criteria based on their physicochemical properties. The purpose of the present work is to propose a strategy to identify substructures from databases of known drugs, which can be used as templates for the generation of libraries of "privileged fragments" that are able to provide high-quality hits. The entire process has been developed integrating Pipeline Pilot (Accelrys Inc., San Diego, CA; http://www.accelrys.com) native components and user-defined molecular files containing ISIS-like substructure query features (Symyx, San Ramon, CA; http://www.symyx.com). The method is effortless, easy to put in place, and fast enough to be iteratively applied to different sources of druglike compounds.

INTRODUCTION

Fragment-based approaches currently are well-established tools for lead identification.[1−10] They are typically based on biophysical screening methods, in particular, protein crystallography and nuclear magnetic resonance (NMR) techniques.

X-ray methods, which offer the main advantage to go directly into the structural information, have been successfully applied in the identification and optimization of lead compounds in several drug discovery projects.[11,12] However, the use of this technique is sometimes limited by difficulties in obtaining crystals that can be due, for instance, to particular conformations assumed by the active site of the protein in crystals, or to mechanisms of diffusion and molecular stability.[13−15] In addition, amphiphilic membrane proteins are, by their specific nature, difficult to crystallize.

NMR-based screening represents an alternative and, to a large extent, complementary approach to the identification of weak binders from ligand mixtures.[16−18] From a technical point of view, this technique provides a powerful set of tools[19−22] to study the interaction of proteins with small molecules, such as TINS,[23] STD,[24] water-LOGSY,[25] and 3-FABS.[26] The latter has multiple advantages, because the substrate molecule is tagged with one or more F atoms, usually CF3. In this case, fragment mixtures can be analyzed without the need to check overlapping signals of the members, because what is observed is the high-intensity signal of the fluorinated substrate. Moreover, these types of assays can also be used as early indicators of the druggability of a given target.[27−31]

Because of their ability to detect weak intermolecular interactions under quasi-natural conditions, and the applicability to targets that cannot be crystallized, the use of NMR techniques has significantly increased over the past few years.

A common requirement of all the fragment-based paradigms is the need for a screening collection of small molecules.

Many fragment-based screening (FBS) libraries have been reported in the literature, with sizes ranging from ∼1500 molecules up to ∼15 000 molecules. The size of the library is usually dictated by different factors, such as the cost of compounds acquisition, the assay conditions, the throughput of the screening technique, and the acceptable number of compounds in the mixtures. Furthermore, many pharmacological targets may require specific moieties (e.g., ATPases, metallo-proteases, kinases, etc.) and it is not always useful or economically wise to screen very large libraries, knowing that, sometimes, focused ones can afford better results in a much shorter time.

Whatever the size collection, its chemical diversity, a suitable profile of physicochemical properties and the availability of the selected fragments are key factors to be considered during compound selection.[32−38]

In particular, to build well-performing screening libraries, compounds typically must be soluble (up to 1 mM in dimethyl sulfoxide (DMSO) and 1 mM in water) and chemically stable, with low molecular weight ($MW < 300$) and appropriate polar surface area ($PSA < 60$ Å).

Quite often, the search, analysis, and selection of suitable small molecules turns out to be a lengthy process, because a large number of compounds must be both submitted to

* Author to whom correspondence should be addressed. E-mail: luca.sartori@ifom-ieo-campus.it.

10.1021/ci800219h CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/08/2008

different levels of ring/substituents identification and checked for their compliance to the criteria set on chemical and physical properties.[39] Moreover, the continuously increasing number of compounds available through many vendors and catalogues, usually interspersed with compounds with undesired properties or chemical features, makes the acquisition of effective fragment screening libraries quite a problematic task.[40−42]

Various approaches to set up FBS libraries have been proposed; in most cases, these are based on the identification of recurring fragments from known drugs. For example, Murcko et al. developed a computational method to analyze the properties of known drugs. They stated that major frameworks from known drugs can simply be represented by two or more ring systems linked by a chain and concluded that the diversity of common shapes is very limited in the set of the known drugs.[22,43−46]

Lewell and colleagues[47] developed the so-called "retrosynthetic combinatorial analysis procedure" (RECAP), which is based on dissecting drug molecules according to a set of predefined bond cleavage rules. This molecular fragmentation method leads to the identification of a set of building blocks that can be used for designing lead-like libraries. Thanks to its reaction oriented fragmentation rules, RECAP brings directly to chemical accessibility and can be considered the reference for the "make" approach.

More recently, Fesik and co-workers proposed to enrich screening libraries with small molecules that have been experimentally shown to frequently bind to proteins.[38] In the present work, "privileged fragments" are suggested as compounds that satisfy both physicochemical properties (i.e., the "rule of 3"[48]) and molecular shape requirements (e.g., to have topologies similar to those of common drugs), making them particularly suitable for small-molecule screening.

Within this approach, a novel computational strategy was developed to identify and isolate molecular moieties from known drugs, which are potentially important for their biological activities.[49,50] To do this, databases of known drugs were stratified according to different levels of scaffold decoration (in particular, rings and ring assemblies) and all the generated substructures were extracted. The subsequent application of this fragmentation procedure to virtual screening databases allowed the enrichment of the generated leadlike fragment set with more "novel" and "original" moieties.

Ultimately, a collection of privileged fragments was produced; structures were extracted from commercially available databases using similarity and substructure searches, and then they were filtered by physicochemical properties and undesired chemical moieties. The entire process has been developed integrating Accelrys Pipeline Pilot native components and molecular files that have Symyx ISIS-like substructure query features. The method is fast enough to be iteratively applied to different sources of druglike compounds, as well as to various catalogues of suppliers, to enhance the chemical diversity of generated screening libraries and easily identify molecules with the best chance to meet desirable druglike requirements.

MATERIALS AND METHODS

Druglike Databases. The generation of the set of privileged fragments, originating from "druglike" topologies, has been performed searching different type of sources, as reported in Figure 1.

Commercial Databases. Symyx Drug Data Report (MDDR). Originally produced by MDL and Prous Science and now distributed by Symyx, MDDR is a database covering the patent literature, journals, meetings, and congresses. It contains ~180 000 biologically relevant compounds and well-defined derivatives (v. 2007), integrated with therapeutic action and biological activity data.[51]

Symyx Comprehensive Medicinal Chemistry (CMC). Derived from the Drug Compendium in Pergamon's Comprehensive Medicinal Chemistry (CMC), the Symyx Comprehensive Medicinal Chemistry database provides three-dimensional (3D) models and important biochemical properties, including drug class and log P and pK_a values for ~9000 pharmaceutical compounds (v. 2006) used or studied as medicinal agents in humans.[52]

Freely Accessible Collection. The ZINC Database from Shoichet Laboratory, Department of Pharmaceutical Chemistry, University of California, San Francisco (UCSF) was used.

ZINC is a free database of commercially available compounds appropriate for virtual screening. It contains over 4.6 million compounds in ready-to-dock, 3D formats, grouped in several sets of molecules.[53,54] The ZINC "druglike" subset (version 2006-05-02) was chosen, which contained ~2 367 000 compounds, in agreement with the Lipinski[55] counts.

The subset of ZINC that was used can be considered to be quite different from the other databases of known drugs, because it contains a large number of compounds that have been designed using combinatorial chemistry approaches and are suitable for exploratory activities such as docking or virtual screening. The use of the ZINC druglike set expanded the number and diversity of the generated fragments to cover regions of the druglike chemical space otherwise unexplored and suggested interesting evaluations and comparisons in terms of number and quality of fragments produced from different database sources (data not shown).

Molecular Filter. An automatic Pipeline Pilot protocol, called, for simplicity, "Molecular Filter", has been developed and dedicated to apply different types of filters at distinct steps of the process implementation.

Molecular Filter is a collection of various types of filters that conceptually can be divided into three main areas:

(1) The "Organic Filter", which checks molecules for the presence of inorganic atoms. Organic atoms are defined as H, C, N, O, P, S, F, Cl, Br, and I. Molecules that contain only these atom types are identified as organic; any molecule that contains any other atom type (including R, and query atoms such as A and Q) is defined as inorganic and discarded.

(2) "Rule of three (Ro3) Filters",[48] which are a set of filters based on physicochemical properties previously calculated by means of independent protocols (see the Physicochemical Properties section). Properties thresholds were set according to Ro3, as follows: $A \log P \leq 3$; PSA ≤ 60 Å, and $\log S_w > -3$. Molecular weight limits were custom set, depending on the implementation step.

(3) The "Undesirable Moieties Filter", which discards a set of molecules showing undesirable substructures, either

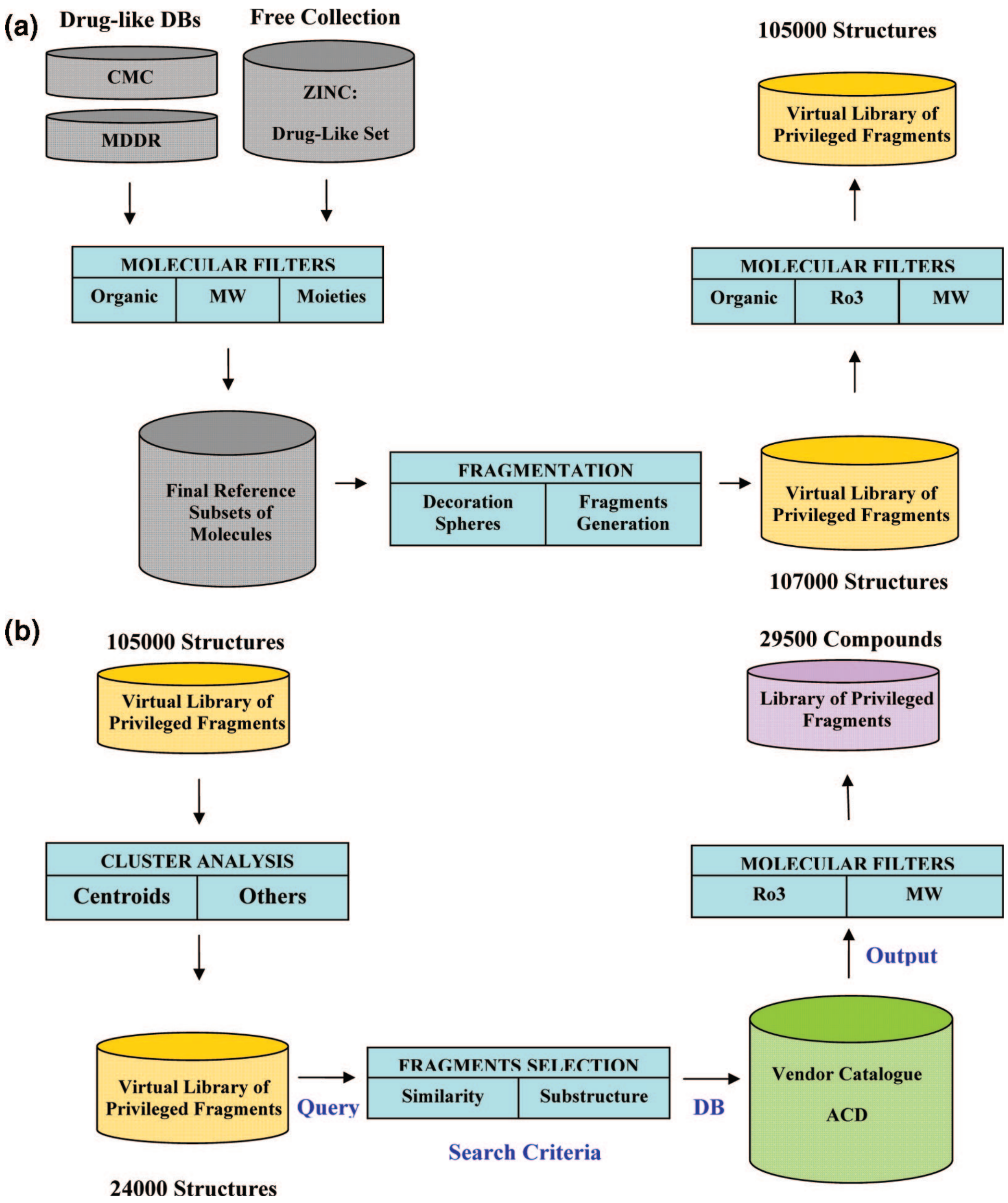

Figure 1. Workflow of the implementation: (a) a set of privileged fragments are generated upon fragmentation of the databases of known drugs or druglike compound collections; (b) virtual fragments are used as query input toward commercial catalogues to collect libraries suitable for FB primary screening.

because of expected high toxicity, or because they represent a Michael acceptor, which is likely to produce artifactual inhibition, because of covalent modifications of the target.

In addition, only structures that contain at least one ring bond were considered. This type of structural filter was implicitly implemented by the application of the structure fragmentation algorithm reported here, that dismisses all of the sidechains generated upon the fragmentation of the original molecules.

Molecular Filter has been used three times to discard different types of molecules. Initially, it has been applied to the starting compound collections (CMC, MDDR, and ZINC) to discard molecules with inorganic atoms and to remove undesirable moieties. As for the Ro3 filters, all the physicochemical properties components were inactivated, except for molecular weight, which was set to MW ≤ 800.

Molecular Filter then has been applied to the collection of virtual fragments obtained from the fragmentation of druglike databases. In that case, the adjunctive feature to remove structures that contain only C, H, or halogens has been added to the basic Organic Filter. Ro3 rules were applied and MW limits were set to MW ≤ 300, and a lower range limit of MW ≥ 80 was also set, to eliminate solvents and molecules that were too small or trivial, such as the over-represented benzene ring.

The last implementation of the Molecular Filter protocol was applied to the screening library of real fragments collected from vendor catalogues. At that stage, fragments that satisfy the physicochemical properties according to Ro3 criteria were retained to compose the final collection; the only exception to these rules was represented by the molecular weight (MW), which was set to a maximum of 400 to enhance diversity of the screening molecules.

Physicochemical Properties. Pipeline Pilot components were used to calculate all the required physicochemical properties. In particular:

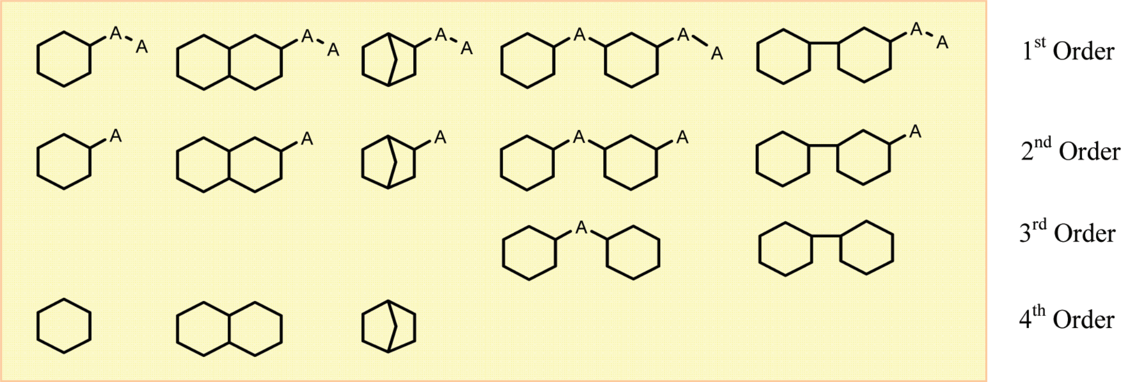

Figure 2. Decoration spheres: according to the definition of fragmentation rules, different orders (or levels) of scaffolds decorations are defined to generate a collection of privileged fragments. In this way, structural features of known drugs are transferred to the set of generated fragments. Hexagons represent a generic ring of any size and bond order; for simplicity, query feature definitions are not reported.

(1) "*Molecular Solubility*, which is the method implemented in Pipeline Pilot Component to estimate the solubility is the multiple linear regression model published by Tekto et al.[56]

(2) "*A log P*", where the corresponding Pipeline Pilot Component uses the method as described by the published work of Ghose et al.[57]

(3) "*Log D*", for which the Pipeline Pilot Component is based on the published work by Csizmadia et al.[58]

(4) "*Molecular Surface Area*" and "*Polar Surface Area*" are calculated using a method based on the published work by Ertl et al.[59]

(5) "*Molecular Weight*" and the presence or absence of desired/undesired atoms were calculated using the corresponding Pipeline Pilot components.[60]

Clustering. The Pipeline Pilot[60] "Cluster Molecules" component has been used to generate clusters from the virtual set of privileged fragments obtained from the fragmentation of the druglike data sets. The clustering method used was the maximal dissimilarity partitioning algorithm (Max-Min method of Waldman et al.[61]) and the so-called 2D functional class fingerprints (FCFPs) were selected as cluster descriptors, detailed at the second iteration order of neighborhoods (four bond distances around the starting point atom), corresponding to a maximum diameter of 4 (FCFP_4).

Different from substructural descriptors (such as SYMYX-MDL keys or Daylight[62] path-based fingerprints), FCFPs are structure-based fingerprints, where each feature represents a structural motif. They are fast to calculate and offer the main advantage of considering, for each molecule, a much larger list of features rather than other types of fingerprints (such as those based on binary bit arrays). Finally, information about tertiary and quaternary centers or stereo chemical data also can be included.

Commercial Catalogues of Reagents. The final step of any library generation process is represented by the purchase of commercially available compounds (see Figure 1a, b). Therefore, with our set of privileged fragments extracted from databases of known drugs, a commercial source has been used to search for available compounds. A specific one has been chosen, but the procedure can be applied to any vendor catalogue.

Symyx Available Chemical Directory (ACD). Here, ACD (v. 2006) has been used, which contains more than 490 000 structures.[63]

PROCESS IMPLEMENTATION

The central purpose of the present work was to select a library of small molecules, the so-called "privileged fragments", with two main characteristics: physicochemical properties falling within (or as close as possible) the ranges defined by the Ro3 rules (see Molecular Filter section) and topological shapes with a high degree of similarity to those of known biologically active compounds. In addition, to obtain novel and original starting points for lead optimization, privileged fragments were also extracted from virtual screening databases, which could potentially provide structural features not yet included in known drugs.

These fragments have been used as query input toward commercial catalogues (ACD) to build libraries of small molecules likely to afford hits from their fragment-based screening (as reported in the general schema of Figure 1).

Structure Fragmentation Algorithm (I): Definition of "Decoration Spheres". Initially, a set of substructure definitions, called "decoration spheres", was created and used to describe and classify each fragment generated from the reference set of druglike molecules. According to these rules and by means of the application of the algorithm described here, fragments with a decreasing level of complexity (from the first decoration sphere to the fourth decoration sphere), in terms of side-chain decorations around a central core, were produced; the generated molecules were finally grouped according to their differentially detailed sidechains, as belonging to distinct "decoration spheres".

The first-order sphere of substituent comprises either any sidechain of two generic atoms (A, excluded hydrogen) one bond away from each ring assembly or any sidechain of one atom when the original side-chain length is greater than two atoms. The process removes all the side chains and breaks the connections between the various ring assemblies present in any compound, with the exception of bonds directly connecting two ring systems and bonds connecting two rings and one atom in between (see the fourth and fifth columns in Figure 2). Similarly, the second-order sphere includes any generic atom (A, excluding hydrogen) limited to one bond away from each ring assembly. The third-order sphere removes the sidechains completely, with the exception of the protruding double bonds that contain heteroatoms (C=O, C=S), but keeps the bonds that connect two ring assemblies and the single generic atoms between two or more ring assemblies.

Ultimately, in the fourth level of decoration, only rings, ring assemblies, and bridged ring assemblies are considered. Figure 2 shows a description and visualization of generic shapes of fragment types.

Structure Fragmentation Algorithm (II): Fragments Generation. The fragmentation process of chemical structures, based on the previously described definition of "decoration spheres", has been implemented using a computational protocol that integrates both Accelrys Pipeline Pilot native components and molecular files that contain substructure queries with the appropriate features required for structure dissection. The procedure is based on a series of steps, each finalized to the generation of a different type of fragment with decreasing levels of decoration, as described in Figure 2. For each step, the generated dataset was used as input for the following phase of the procedure. The outputs are also saved as independent results, primarily for process debugging, but also to obtain details on each step.

The process started with the application of Molecular Filter protocol (see the section, Molecular Filter, for details) to the CMC, MDDR, and ZINC databases (see the section, Materials and Methods, for details). When filters were applied, final reference subsets were subjected to the fragmentation procedure as reported (see Figure 1):

CMC: ~9000 molecules

MDDR: ~170 000 molecules

ZINC druglike set: ~2 300 000 molecules

The input molecules passed through a series of Pipeline Pilot components, in particular "*Substructure Maps*", "*Delete Bonds and Isolate Rings*", which were used to generate scaffolds of the first order of complexity via the mapping and removal of substructure 1 reported in Figure 3. This section of the protocol allowed the identification of either any chain of two generic atoms away from ring assembly or one atom in the case of chain length longer than two atoms. Then, by mapping and removing substructure 2 in Figure 3, any atom up to one bond away from pure ring assemblies was eliminated, thus producing scaffolds as defined by the second-order sphere. The process continued, using substructure queries 3 and 4 (refer to Figure 3) to isolate the stream fragments with any type of connected ring from the main process. The two queries mapping any chain on ring (substructures 5 and 6 reported in Figure 3) then were used to eliminate sidechains and to keep only bridge atoms and directly connected ring systems. Finally, the "out of the box" Pipeline Pilot[60] component "*Generate Fragments*" was used to produce, after the removal of alpha atoms, the fourth decoration sphere of substituents. In particular, the following types of molecular fragments were generated:

(1) Ring assemblies: contiguous ring structures, including fused rings and bridge systems;

(2) Bridge assemblies: rings sharing two or more bonds; and

(3) Rings: individual rings.

Chemical structures coming from the four steps were finally linked together to compile a unique list of fragments, using a merger based on canonical smiles.

At the end of this step, a virtual collection of ~107 000 privileged fragments was available. After the application of the Molecular Filter (see related paragraph for details), a final set of 105 000 molecules was submitted for cluster analysis. (See Figure 4 for Ro3 profiles of each sphere.)

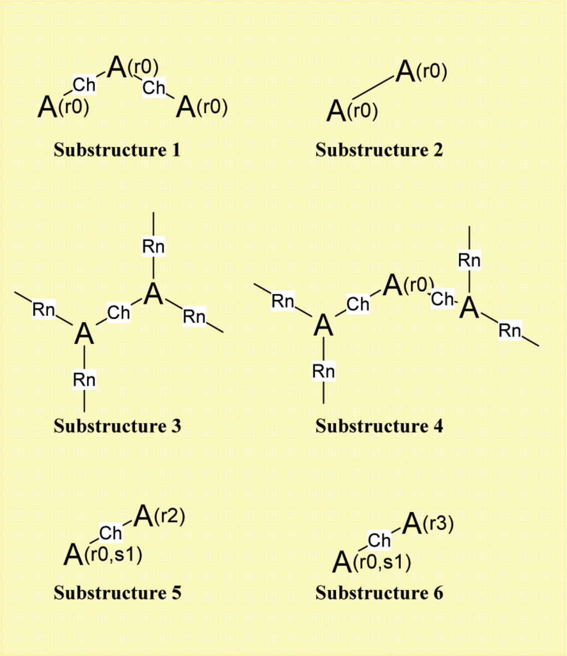

Figure 3. Fragments Generation. Substructures used to generate the four complexity orders of fragments reported in Figure 2. Substructure 1 maps either any chain of two atoms (A) away from a ring assembly or one atom when chain length is greater than two atoms; it is used to generate privileged fragments of the first decoration order. Substructure 2 is used to recognize any atom (A) up to one bond away from ring assembly and to generate, after the deletion of mapped bonds of the input molecules, fragments of the second decoration order. The substructure queries 3 and 4 were used to separate the third substituent residues from the main process stream; then, substructures 5 and 6 were used to eliminate sidechains and keep only bridge atoms (A) and directly connected ring assemblies.

Cluster Analysis and Privileged Fragments Selection. Cluster analysis was used to group privileged fragments showing similar topological and physicochemical properties and to select a subset of representative virtual molecules. They have been used as a query set to generate the library to be purchased by extraction of suitable molecules from vendor catalogues.

The virtual library of ~105 000 privileged fragments generated ~9500 clusters, with an average distribution of 20 molecules for cluster. Examples of clusters of privileged fragments are reported in Figure 5.

Cluster analysis offered a convenient and unavoidable tool to extract real privileged fragments from vendor collections.

The selection of representative privileged fragments started with the set of centroids from all clusters, together with a selection of nearest neighbors comprised of those within 0.1 maximum distances to the closest one. The Pipeline Pilot property "*Distance to Closest*", with values ranging from 0 to 1, gives an indication of the molecular diversity within each cluster. To enhance the sampling of molecules with higher chemical diversity, fragments with a distance to centroids of >0.57 were also kept. Finally, clusters that, upon visual inspection, proved to be particularly interesting were entirely included, to generate a final set of ~24 000 privileged fragments.

Privileged Fragments Library Generation. To select compounds to be purchased, the entire set of 24 000 fragments that were obtained from the cluster analysis was

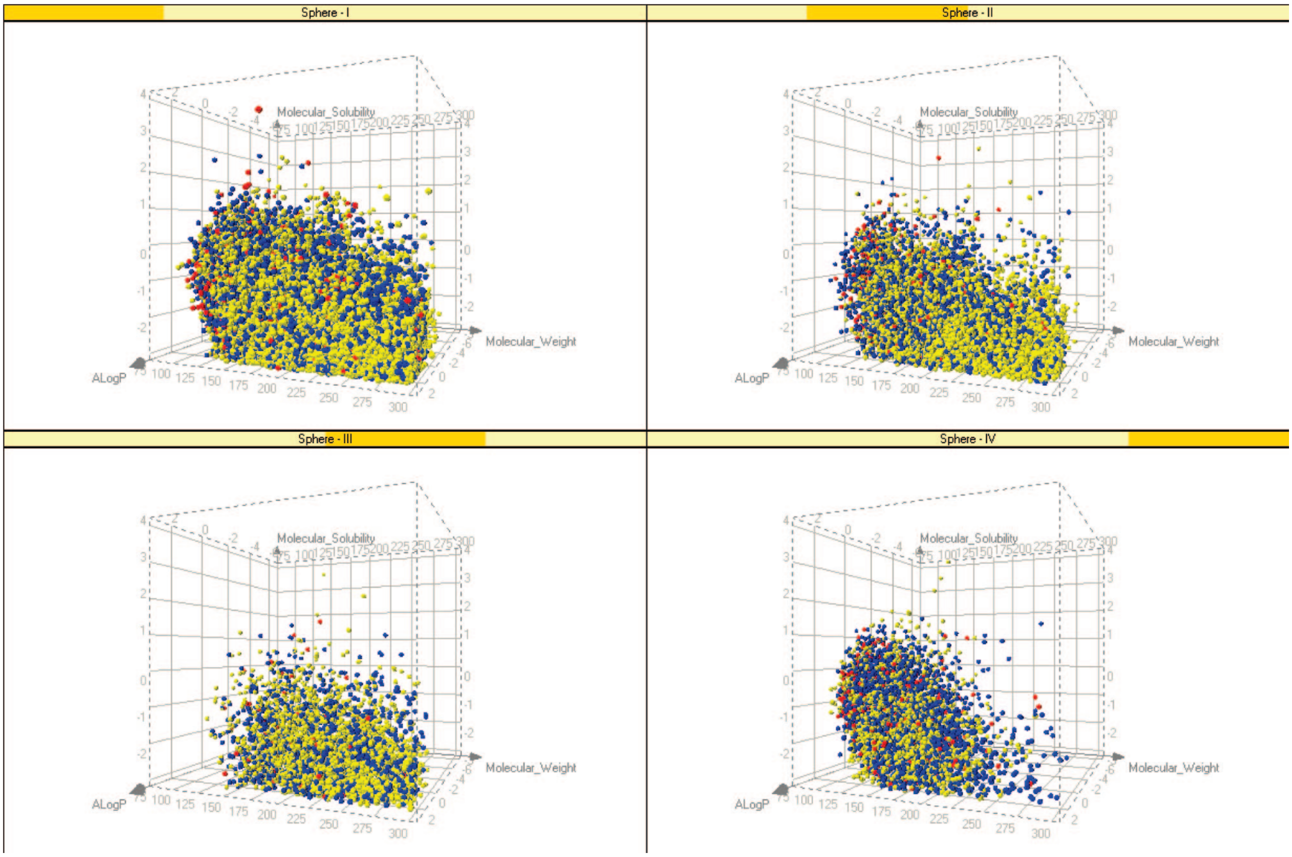

Figure 4. Distribution of physicochemical properties (molecular weight, solubility, and *A* log *P*) of privileged fragments obtained upon fragmentation of the CMC, MDDR, and ZINC databases. Single panels display fragments distribution according to different decoration orders (from Sphere I to IV). Privileged fragments are colored by the database of origin; red = CMC, blue = MDDR, and yellow = ZINC.

used. Substructure searches and Tanimoto similarity criteria were used to extract real molecules from vendor catalogues (in our case, ACD, but many other collections can be added).

Because substructure searches and the Tanimoto similarity can retrieve molecules with properties exceeding fragment-like criteria, a final set of filters (see the section, Molecular Filter) has been applied. Moreover, particular consideration has been maintained for the lower limit of MW, which was set to 80 to exclude non-specific and over-represented substructures, such as the benzene ring, from the purchase list of compounds.

At the end of this procedure, a library of 29 500 privileged fragments had been obtained.

RESULTS AND DISCUSSION

To obtain a reasonable set of privileged fragments, the procedure has been applied to compounds that either are known drugs or are molecules produced at different stages of drug discovery and development. Adding ZINC (a druglike subset) to the CMC and MDDR databases, it was possible to collect a large initial set of more than 2 300 000 molecules. The four spheres of scaffolds that were extracted can be regarded as different levels of decorations around a central core, with the intent of automatically generating a large number of possible atom connectivities that are present in druglike molecules.

One of the main advantages of this fragmentation algorithm is that it can preserve particular moieties, such as directly connected rings or rings that are connected with one atom (see sphere III in Figure 2), still removing the other sidechains.

Furthermore, because the definitions of substructure and fragmentation implementation are based on a modular concept, if modifications are required to address particular moieties (e.g., exclusion of undesired substructures), this can be implemented without changes to the entire process.

The sets of privileged fragments generated from all the iterative steps for each database were submitted to molecular properties calculation and filtering according to "rule-of-three" (Ro3) filters,[48] producing 3900 structures from the CMC database, 30 000 from the MDDR database, and 92 000 from the ZINC database.

To avoid redundancies, fragments coming from the aforementioned lists were merged based on canonical smiles, thus producing a final set of ~105 000 unique structures (i.e., a set of molecules where each of them is contained once and only once, despite the original step that generated it).

Fragments showing similar topological and physicochemical properties were then grouped by cluster analysis with the Pipeline Pilot "*Cluster Molecules*" component, which allows easier navigation within chemical classes and faster evaluation of suitable properties for X-ray and NMR screening and leads to an immediate selection of 24 000 representative cluster members. (Details regarding the selection criteria are reported in the proper paragraph and examples in Figure 5.)

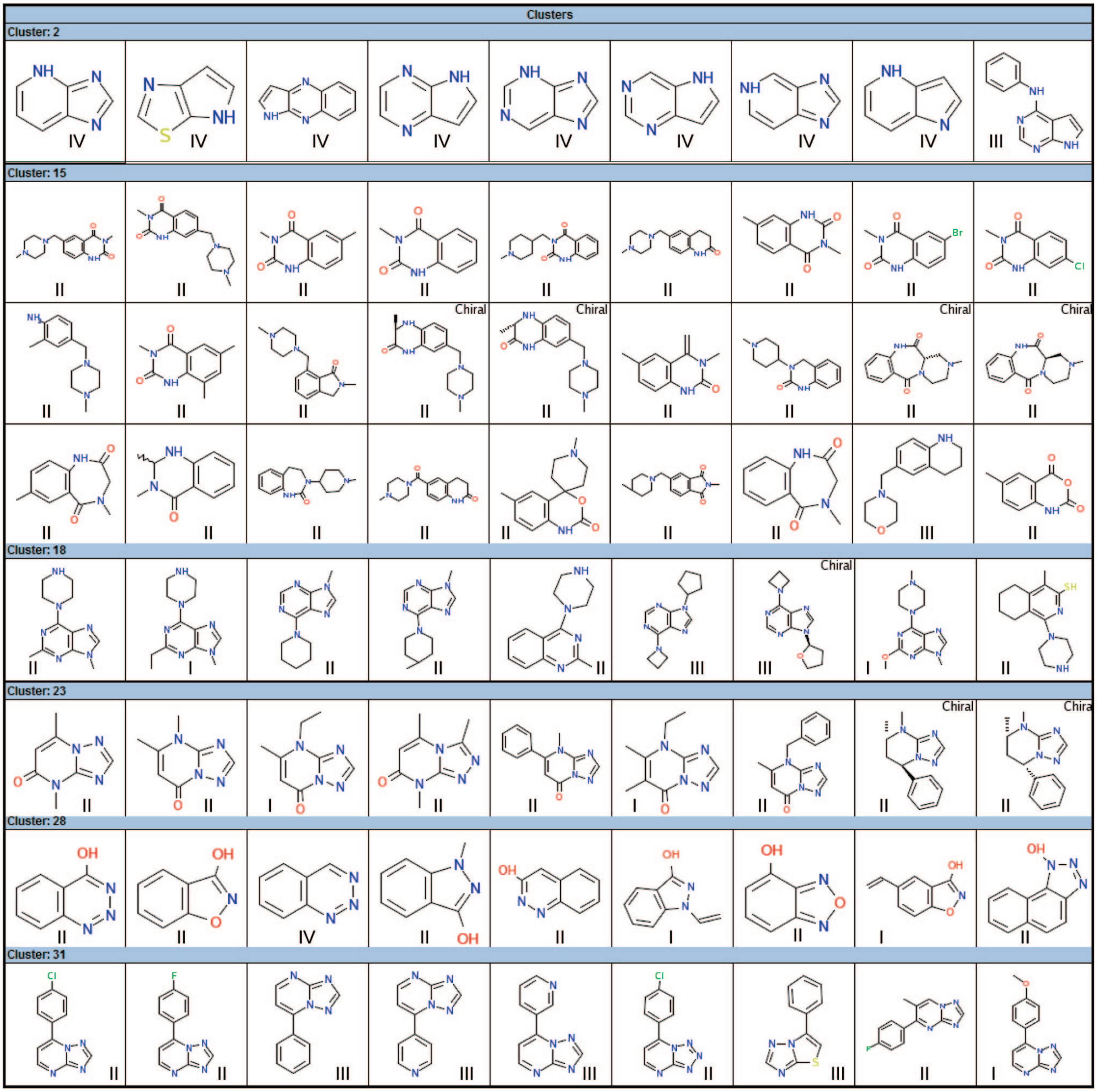

Figure 5. Representative clusters of privileged fragments, obtained from the fragmentation of reference databases, are reported. Clusters were generated by means of maximal dissimilarity partitioning algorithm, using FCFP_4 descriptors. For any privileged fragment, the decoration order is also reported (from sphere I to sphere IV). Cluster analysis has been used to select a set of privileged fragments from the virtual collection obtained from the fragmentation step. Selected fragments were used to collect a real library, by searching vendor catalogues for Tanimoto similarities and substructure searches.

To search for commercially available compounds (see Figure 1), these cluster-selected molecules were used as query templates toward the ACD database, according to distinct and complementary search criteria:

(1) Similarity, based on SciTegic FCFP_4 fingerprints (*0.8, Tanimoto, Pipeline Pilot*);

(2) Substructure, ISIS/Host rules (*ISIS SSS*[64]).

Compounds that did not match Ro3 requisites or contain undesired chemical moieties were discarded, and the finally obtained collection of molecules contained ~29 500 unique structures (2000 compounds matching the Tanimoto Similarity and 28 000 kept by Substructure Searches).

The same extraction step can be applied to any other vendor catalogue, either starting from the entire set of privileged fragments (105 000) or using different criteria to generate diverse subsets such as random selection or those based on cluster approach.

To evaluate the ability of this fragmentation procedure to generate screening collections with a high rate of druglike molecules, and to compare random selection versus cluster analysis, two series of representative datasets have been generated and submitted to supplementary analysis.

In particular, by random selection of 500, 1000, 5000, and 15 000 structures from the virtual library of 105 000 fragments and by applying the same criteria in terms of searching techniques and filters (see above), four fragment collections (A Series) have been extracted from ACD.

Table 1. MDDR Coverage Enrichment Curve (% of Extracted Drugs), Using Sets of 500, 1000, 5000, and 15 000 Privileged Fragments Obtained by Random Selection: A Series, Symyx ACD

A series	extracted from MDDR	% MDDR
500	14000	8
1000	45000	25
5000	110000	61
15000	140000	77

Table 2. MDDR Coverage Enrichment Curve (% of Extracted Drugs), Using Sets of 500, 1000, 5000, and 15 000 Privileged Fragments, Selected by Cluster Analysis: B Series, Symyx ACD

B series	extracted from MDDR	% MDDR
500	50000	28
1000	74000	41
5000	135000	75
15000	156000	86
29500	163000	90

For comparison, equivalent series of collections were generated by random extraction of 500, 1000, 5000, and 15 000 structures from the final library of 29 500 privileged fragments obtained from ACD, using cluster-selected fragments as the query set (B Series).

Structures coming from these two groups were used to extract drugs from the MDDR database, which was selected as being representative of drug databases, in accordance with the same search criteria previously reported:

(1) Similarity, based on SciTegic FCFP_4 fingerprints (*0.8, Tanimoto, Pipeline Pilot*);

(2) Substructure, ISIS/Host rules (*ISIS SSS* [64]).

Again, lists of unique compounds were generated by merging, based on canonical smiles, structures obtained from the similarity and substructure searching criteria.

Results obtained from the extraction procedure, using random-generated or cluster-selected fragment collections, are reported in Tables 1 and 2 respectively.

It is not surprising that it was possible to cover a good percentage of MDDR drugs using fragments that were generated upon the fragmentation of a set of starting molecules that comprises this database. Indeed, this consideration adopts a new concept when considering that MDDR represents a small fraction (~7.3%) of the starting set of molecules submitted to the fragmentation procedure (see Figure 1a). Moreover, the enrichment curves were calculated for a library of molecules extracted from a vendor catalogue, such as ACD.

As a matter of fact, starting from a small set of 500 privileged fragments, ~14 000 drugs (8% of the entire MDDR database) were extracted, covering a large amount of MDDR activity classes. When the query dataset contains 1000 molecules, the extracted list of drugs totals 45 000 compounds and the percentage of MDDR coverage becomes 25%. However, to make the difference, it was necessary to start from a collection of at least 5000 fragments; this query set retrieved a list of ~110 000 drugs, covering 61% of the MDDR database and obtaining almost the same profile in terms of activity classes as the entire MDDR. Ultimately, the last set of 15 000 fragments derived ~140 000 drugs (corresponding to 77% of the MDDR database) with the activity class profile entirely covered (A Series; see Figure 6).

Interestingly, when evaluating results coming from datasets of fragments randomly extracted from the library generated by cluster analysis, the enrichment curve and the distribution activity classes are even better than those previously described.

Indeed, starting from the first query set of just 500 fragments, the number of molecules retrieved from the MDDR database is >50 000 (28% of coverage). Using the set of 1000 fragments, the percentage of coverage amounts to 41% over the MDDR database, with more than 73 000 extracted molecules. Once again, the minimum number of privileged fragments able to generate the same activity profile as the entire MDDR database was 5000, from which 135 000 molecules (~75% of the MDDR database) were derived. However, it was using the library of 15 000 fragments, which a very high number of drugs (~156 000) were kept from the MDDR database (86%). Finally, the maximum coverage of MDDR drug space was obtained using the entire library of real privileged fragments (136 000 retrieved drugs, 90% MDDR coverage) (B Series; see Figure 7).

In conclusion, as gathered from the comparisons performed among several fragment collections described here, the best compromise between the number of molecules to be screened and the highest degree of similarity with known drugs is represented by a library of at least 15 000 privileged fragments selected by cluster analysis. This query set covered the quasi-total (~86%) of a collection of biological active molecules, as the MDDR database is.

Despite the good coverage obtained using just 15 000 fragments, as in the acquisition step a significant number of molecules could turn out to be unavailable (out of stock, low purity,...), a vendor catalogue (ACD) was searched for the entire dataset of ~24 000 fragments, thus generating a final selection of 29 500 compounds, which made us reasonably confident to be able to set up a suitable collection of privileged fragments for FB screenings.

CONCLUSIONS

In the present work, a novel procedure to detect privileged fragments, which are defined as compounds that satisfy the physicochemical property criteria ("rule of three" (Ro3) filters [47]) and molecular shape requirements (similarity to known drugs based on two-dimensional (2D) topological descriptors) is described.

The theoretical basis of this procedure originates from the concept, proposed by Bemis and Murcko, of the so-called "pharmacological promiscuity":[44]

> "Drugs which possess common topological shapes of known drugs are quite different in polarity, conformation, hydrogen-bonding potential, and other properties; they bind to different classes of receptor and they serve different pharmacological needs. And yet, they all have the same topological shape".

According to this assumption, the developed structure dissection procedure allowed the identification of privileged fragments that represented special substructures of increasing orders of complexity, in terms of decorations around the main

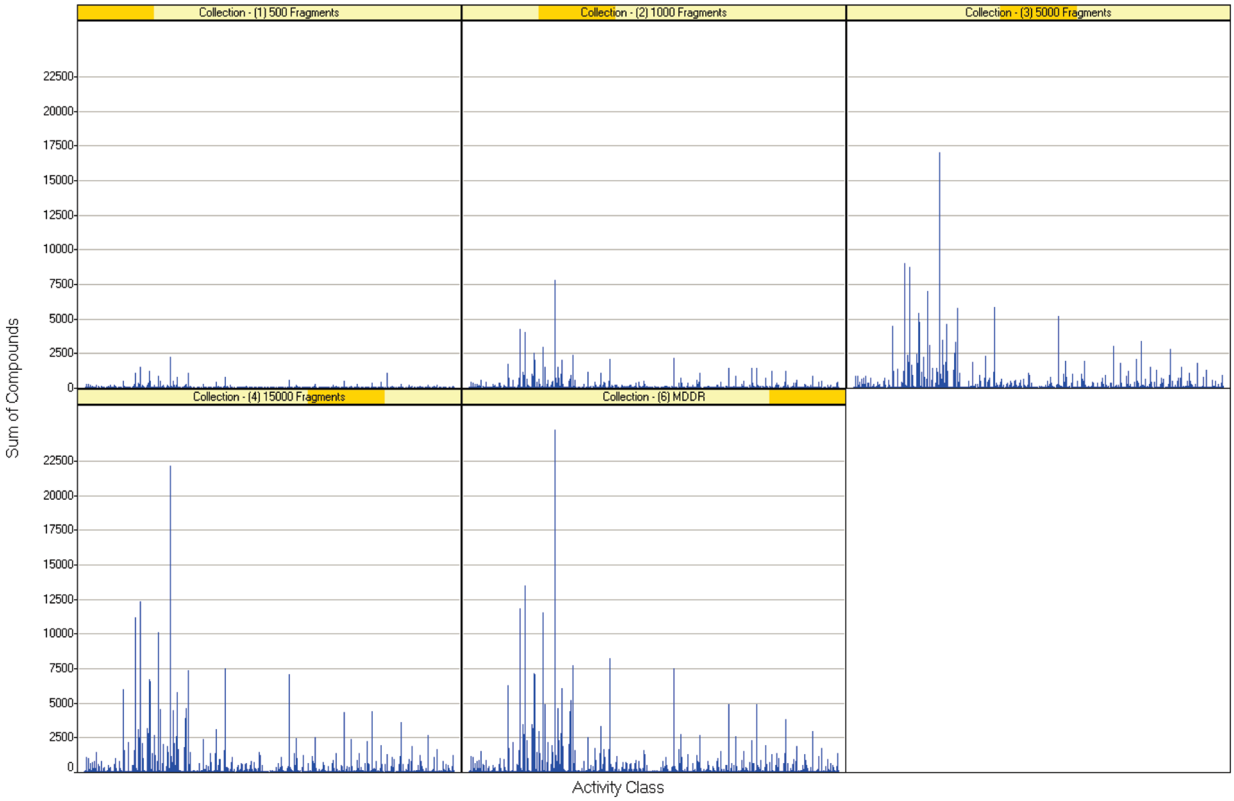

Figure 6. Bar-chart profile shows activity class distributions over the entire MDDR drugs (Collection (6)). Similar visualizations are reported for MDDR subsets (Collections (1)−(4)) obtained by similarity or substructure searches, starting from 500, 1000, 5000, or 15 000 random selected from privileged fragments library (A Series, Symyx ACD).

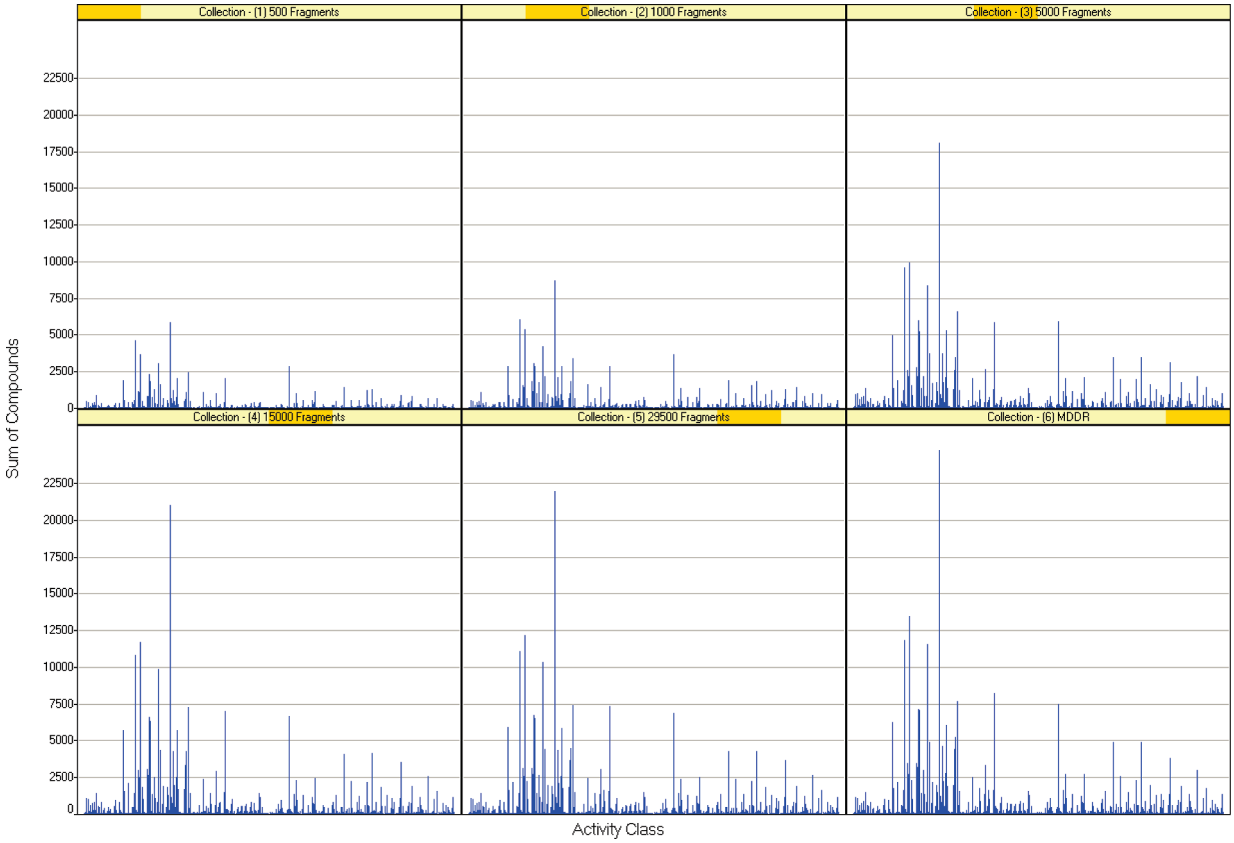

Figure 7. Bar-chart profile showing the activity class distributions over the entire MDDR drugs (collection (6)). Similar visualizations are reported for MDDR subsets (collection (1)−(5)) obtained by similarity or substructure searches, starting from 500, 1000, 5000, 15 000, or the entire collection (29 500) of privileged fragments selected by cluster analysis (B Series, Symyx ACD).

core present into the original compounds. The application of this fragmentation procedure to the druglike subset of the ZINC database can be used to expand the search to "novel" and "original" moieties, and put the basis for the enrichment, in terms of chemical diversity, of the initial set of privileged fragments. This operation preserved the topological moieties of drugs, potentially important for their biological activities, to be further used as structural queries to search for similar shapes in the chemical space of vendor's commercial catalogues.

In summary, a novel molecular dissection technique for designing fragment libraries through a simple computational procedure is suggested; the implementation is fast enough to be iteratively applied to different sources of druglike compounds, as well as many other types of collections.

The results obtained from the validation procedure demonstrated that this approach is a valuable strategy for building high-quality screening libraries, which are well-suited to provide valuable hits by fragment-based (FB) screening.

ACKNOWLEDGMENT

We would like to thank Dr. Mario Varasi, Head of Chemistry Department at Genextra, for the support and the very useful discussions about fragment concepts and molecular diversity. We would like to thank Dr. Daniele Fancelli for critical analysis of the approach and help in reviewing the manuscript.

Supporting Information Available: Protocol containing fragmentation procedures and corresponding molfiles are provided. The protocol is written in XML programming language and must be run using Pipeline Pilot version 6.1. Molfiles are produced using Symyx ISIS/Draw, and can be read with any ASCII file reader. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Leach, A. R.; Hann, M. M.; Burrows, J. N.; Griffen, E. J. Fragment screening: an introduction. *Mol. BioSyst.* **2006**, *2* (9), 430–46.

(2) Zartler, E. R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.* **2005**, *9* (4), 366–70.

(3) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.* **2004**, *47* (14), 3463–3482.

(4) Verdonk, M. L.; Hartshorn, M. J. Structure-guided fragment screening for lead discovery. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (4), 404–10.

(5) McGrath, M. E.; Sprengeler, P. A.; Hirschbein, B.; Somoza, J. R.; Lehoux, I.; Janc, J. W.; Gjerstad, E.; Graupe, M.; Estiarte, A.; Venkataramani, C.; Liu, Y.; Yee, R.; Ho, J. D.; Green, M. J.; Lee, C. S.; Liu, L.; Tai, V.; Spencer, J.; Sperandio, D.; Katz, B. A. Structure-guided design of peptide-based tryptase inhibitors. *Biochemistry* **2006**, *45* (19), 5964–73.

(6) Gill, A. L.; Frederickson, M.; Cleasby, A.; Woodhead, S. J.; Carr, M. G.; Woodhead, A. J.; Walker, M. T.; Congreve, M. S.; Devine, L. A.; Tisi, D.; O'Reilly, M.; Seavers, L. C.; Davis, D. J.; Curry, J.; Anthony, R.; Padova, A.; Murray, C. W.; Carr, R. A.; Jhoti, H. Identification of novel p38alpha MAP kinase inhibitors using fragment-based lead generation. *J. Med. Chem.* **2005**, *48* (2), 414–26.

(7) Gill, A. New lead generation strategies for protein kinase inhibitors - fragment based screening approaches. *Mini-Rev. Med. Chem.* **2004**, *4* (3), 301–311.

(8) Liu, G.; Xin, Z.; Pei, Z.; Hajduk, P. J.; Abad-Zapatero, C.; Hutchins, C. W.; Zhao, H.; Lubben, T. H.; Ballaron, S. J.; Haasch, D. L.; Kaszubska, W.; Rondinone, C. M.; Trevillyan, J. M.; Jirousek, M. R. Fragment screening and assembly: a highly efficient approach to a selective and cell active protein tyrosine phosphatase 1B inhibitor. *J. Med. Chem.* **2003**, *46* (20), 4232–5.

(9) Kolb, P.; Caflisch, A. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.* **2006**, *49* (25), 7384–92.

(10) Babaoglu, K.; Shoichet, B. K. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2006**, *2* (12), 720–3.

(11) Sharff, A.; Jhoti, H. High-throughput crystallography to enhance drug discovery. *Curr. Opin. Chem. Biol.* **2003**, *7* (3), 340–345.

(12) Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.* **2005**, *48* (2), 403–413.

(13) Congreve, M.; Aharony, D.; Albert, J.; Callaghan, O.; Campbell, J.; Carr, R. A.; Chessari, G.; Cowan, S.; Edwards, P. D.; Frederickson, M.; McMenamin, R.; Murray, C. W.; Patel, S.; Wallis, N. Application of fragment screening by X-ray crystallography to the discovery of aminopyridines as inhibitors of beta-secretase. *J. Med. Chem.* **2007**, *50* (6), 1124–32.

(14) Gill, A.; Cleasby, A.; Jhoti, H. The discovery of novel protein kinase inhibitors by using fragment-based high-throughput x-ray crystallography. *ChemBioChem* **2005**, *6* (3), 506–12.

(15) Lesuisse, D.; Lange, G.; Deprez, P.; Benard, D.; Schoot, B.; Delettre, G.; Marquette, J. P.; Broto, P.; Jean-Baptiste, V.; Bichet, P.; Sarubbi, E.; Mandine, E. SAR and X-ray. A new approach combining fragment-based screening and rational drug design: application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem.* **2002**, *45* (12), 2379–87.

(16) Hajduk, P. J.; Betz, S. F.; Mack, J.; Ruan, X.; Towne, D. L.; Lerner, C. G.; Beutel, B. A.; Fesik, S. W. A Strategy for High-Throughput Assay Development Using Leads Derived from Nuclear Magnetic Resonance-Based Screening. *J. Biomol. Screen.* **2002**, *7* (5), 429–432.

(17) Dalvit, C.; Pevarello, P.; Tatò, M.; Veronesi, M.; Vulpetti, A.; Sundstrom, M. Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* **2000**, *18* (1), 65–68.

(18) Liu, G.; Xin, Z.; Liang, H.; Abad-Zapatero, C.; Hajduk, P. J.; Janowick, D. A.; Szczepankiewicz, B. G.; Pei, Z.; Hutchins, C. W.; Ballaron, S. J.; Stashko, M. A.; Lubben, T. H.; Berg, C. E.; Rondinone, C. M.; Trevillyan, J. M.; Jirousek, M. R. Selective protein tyrosine phosphatase 1B inhibitors: Targeting the second phosphotyrosine binding site with non-carboxylic acid-containing ligands. *J. Med. Chem.* **2003**, *46*, 3437–3440.

(19) Dalvit C.; Ardini E.; Flocco M. M.; Fogliatto G.; Mongelli N.; Veronesi M. Fluorine NMR spectroscopy for biochemical screening of inhibitors of an enzyme by using a CF3-labeled substrate 2005, PCT Int. Appl.: WO 2005005978 A2 20050120.

(20) Hajduk, P. J.; Mack, J. C.; Olejniczak, E. T.; Park, C.; Dandliker, P. J.; Beutel, B. A. SOS-NMR: A saturation transfer NMR-based method for determining the structures of protein-ligand complexes. *J. Am. Chem. Soc.* **2004**, *126*, 2390–2398.

(21) Dalvit C.; Stockman B. J.; Flocco M. M.; Veronesi M. Use of fluorine-19 NMR for high throughput screening 2004, PCT Int. Appl.: WO 2004051214 A2 20040617.

(22) Fejzo, J.; Lepre, C. A; Peng, J. W.; Bemis, G. W.; Ajay Murcko, M. A.; Moore, J. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* **1999**, *6* (10), 755–769.

(23) Vanwetswinkel, S.; Heetebrij, R. J.; van Duynhoven, J.; Hollander, J. G.; Filippov, D. V.; Hajduk, P. J.; Siegal, G. TINS. Target Immobilized NMR Screening: An Efficient and Sensitive Method for Ligand Discovery. *Chem. Biol.* **2005**, *12* (2), 207–216.

(24) Mayer, M.; Meyer, B. Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectra. *Angew. Chem., Int. Ed.* **1999**, *35*, 1784–1788.

(25) Dalvit, C.; Fogliatto, G. P.; Stewart, A.; Veronesi, M.; Stockman, B. WaterLOGSY as a method for primary NMR screening: Practical aspects and range of applicability. *J. Biomol. NMR* **2001**, *21*, 349–359.

(26) Dalvit, C.; Ardini, E.; Flocco, M.; Fogliatto, G. P.; Mongelli, N.; Veronesi, M. A general NMR method for rapid, efficient, and reliable biochemical screening. *J. Am. Chem. Soc.* **2003**, *125* (47), 14620–14625.

(27) Tsao, D. H.; Sutherland, A. G.; Jennings, L. D.; Li, Y.; Rush, T. S., 3rd.; Alvarez, J. C.; Ding, W.; Dushin, E. G.; Dushin, R. G.; Haney, S. A.; Kenny, C. H.; Malakian, A. K.; Nilakantan, R.; Mosyak, L. Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorg. Med. Chem.* **2006**, *14* (23), 7953–7961.

(28) Huth, J. R.; Sun, C. Utility of NMR in lead optimization: fragment-based approaches. *Comb. Chem. High Throughput Screening.* **2002**, *5* (8), 631–643.

(29) Hajduk, P. J.; Shuker, S. B.; Nettesheim, D. G.; Craig, R.; Augeri, D. J.; Betebenner, D.; Albert, D. H.; Guo, Y.; Meadows, R. P.; Xu, L.; Michaelides, M.; Davidsen, S. K.; Fesik, S. W. NMR-based

modification of matrix metalloproteinase inhibitors with improved bioavailability. *J. Med. Chem.* **2002**, *45* (26), 5628–5639.

(30) Wyss, D. F.; Arasappan, A.; Senior, M. M.; Wang, Y. S.; Beyer, B. M.; Njoroge, F. G.; McCoy, M. A. Non-Peptidic Small-Molecule Inhibitors of the Single-Chain Hepatitis C Virus NS3 Protease/NS4A Cofactor Complex Discovered by Structure-Based NMR Screening. *J. Med. Chem.* **2004**, *47* (10), 2486–2498.

(31) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48* (7), 2518–2525.

(32) Baurin, N.; Aboul-Ela, F.; Barril, X.; Davis, B.; Drysdale, M.; Dymock, B.; Finch, H.; Fromont, C.; Richardson, C.; Simmonite, H.; Hubbard, R. E. Design and Characterization of Libraries of Molecular Fragments for Use in NMR Screening against Protein Targets. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2157–2166.

(33) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29* (1), 55–67.

(34) Jacoby, E.; Davies, J.; Blommers, M. J. J. Design of small molecule libraries for NMR screening and other applications in drug discovery. *Curr. Top. Med. Chem.* **2003**, *3* (1), 11–23.

(35) Huth, J. R.; Sun, C.; Sauer, D. R.; Hajduk, P. J. Utilization of NMR-derived fragment leads in drug design. *Methods Enzymol.* **2005**, *394*, 549–571.

(36) Villar, H. O.; Yan, J.; Hansen, M. R. Using NMR for ligand discovery and optimization. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 387–391.

(37) Vajda, S.; Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discovery Dev.* **2006**, *9* (3), 354–62.

(38) Hajduk, P. J.; Gomtsyan, A.; Didomenico, S.; Cowart, M.; Bayburt, E. K.; Solomon, L.; Severin, J.; Smith, R.; Walter, K.; Holxman, T. F.; Stewart, A.; McGaraughty, S.; Jarvis, M. F.; Kowaluk, E. A.; Fesik, S. W. Design of Adenosine Kinase Inhibitors from the NMR-Based Screening of Fragments. *J. Med. Chem.* **2000**, *43* (25), 4781–4786.

(39) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325–336.

(40) Gasteiger, J.; Marsili, M.; Hutchings, M. G.; Saller, H.; Loew, P.; Roese, P.; Rafeiner, K. Models for the representation of knowledge about chemical reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 467–476.

(41) Merlot, C.; Domine, D. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594–602.

(42) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivashchenko, A. A. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.* **2001**, *42*, 249–258.

(43) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(44) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.

(45) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42* (25), 5095–5099.

(46) Moore J.; Bemis G. W.; Lepre C. A.; Fejzo J.; Peng J. W.; Wilson K. P.; Murcko M. A. Methods for identifying drug cores for drug discovery 1998, PCT Int. Appl. WO 9857155 A1 19981217.

(47) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications to combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(48) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A. 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, (8), 876–887.

(49) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 255–271.

(50) Clark, D. E.; Pickett, S. D. Computational methods for prediction of 'druglikeness'. *Drug Discovery Today* **2000**, *5*, 49–58.

(51) Symyx Drug Data Report (MDDR). SYMYX, San Ramon, CA, http://www.mdl.com (accessed Sep 18, 2008).

(52) Symyx Comprehensive Medicinal Chemistry (CMC). SYMYX, San Ramon, CA, http://www.mdl.com (accessed Sep 18. 2008).

(53) ZINC. Shoichet Laboratory, Department of Pharmaceutical Chemistry, University of California, San Francisco (UCSF), http://blaster.docking.org/zinc/ (accessed September 18, 2008).

(54) Irwin, J. J.; Shoichet, J. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–82.

(55) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. Review. *J. Pharmacol. Toxicol. Methods* **2000**, *44* (1), 235–49.

(56) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.

(57) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragment Methods: An Analysis of AlogP and CLogP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.

(58) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865–871.

(59) Ertl, P.; Rohde, B.; Selzer, P. 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.* **2000**, *43*, 3714–3717.

(60) Pipeline Pilot, Accelrys, Inc., San Diego CA, USA (http://www.accelrys.com) (accessed September 18, 2008).

(61) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity and Combinatorial Libraries. *Mol. Diversity* **1996**, *2*, 64–74.

(62) Daylight Chemical Information Systems, Inc., Aliso Viejo, CA (http://www.daylight.com)(accessed September 18, 2008).

(63) SYMYX, San Ramon, CA, http://www.mdl.com (accessed September 18, 2008).

CI800219H