Chemotypic Coverage: A New Basis for Constructing Screening Sublibraries

Mark Johnson,*[,†] Veer Shanmugasundaram,[‡] Gordon Bundy,[§] Darryl Chapman,[⊥] and Robert Kilkuskie[⊥]

Pannanugget Consulting, Kalamazoo, Michigan 49006, Pfizer Global Research & Development, Groton, Connecticut 06340, Independent Chemistry Consultant, Kalamazoo, Michigan 49006, and Michigan High Throughput Screening Center, Kalamazoo, Michigan 49006

Received July 25, 2008

Chemotypes are presented as a way of selecting a handful of nonoverlapping substructures for a molecular structure representing such things as ring systems and formal counterparts of functional groups. By taking positioning into account, chemotypes can be viewed as *a priori* postulates of the critical substructure of a compound should it become a lead. The construction and use of efficient single-coverage sublibraries involving cyclic systems, ring systems, and functional groups are presented and illustrated in a case study involving 12 competitive inhibitors of DHFR that emerged in screening the McMaster 50K training library. A 10K single-coverage sublibrary involving positioned functional groups uncovered 10 of the competitive inhibitors and all of the regions with interesting activities giving rise to an *a priori* enhancement ratio of 4.2. The reasoning underlying the construction of these sublibraries, the corresponding chemotypic logic of follow-up screening, and the consequential generation of multiscaffold SAR data are presented using the data in this case study.

INTRODUCTION

Deciding upon the exact set of compounds to screen can be viewed as a three-stage process. First, one specifies a relevant region of chemical space using notions related to chemical desirability, toxicity, and druglikeness[1] and, optionally, using more specific notions such as target-family directed masterkeys.[2] Then one forms a reference collection by filtering out the desired structures from a proprietary collection, creating relevant combinatorial libraries, or synthesizing or purchasing member compounds. At this point, the entire reference library might be screened, but as the size of the reference library increases, interest in screening a smaller sublibrary grows as well.

Formal methods of appropriately selecting a sublibrary from a reference library emerged with the advent of concepts of molecular similarity[3,4] and molecular diversity.[5,6] These methods partition the reference library into similarity neighborhoods or clusters and then select a specified number of structures from each neighborhood or cluster.[7] Follow-up screening of a hit can reflect any number of strategies. If one stays strictly within the similarity formalism, one would screen the remaining structures within the cluster containing the hit, those structures closer to the hit than to any other structure in the sublibrary, or those structures within a similarity neighborhood of the hit.

A medicinal chemist, on the other hand, may query the reference library reflecting a postulate as to the 2D structural features responsible for the activity of the hit. There is no natural counterpart to this 2D substructure query in the similarity formalism. The fragments of a fragment-based similarity measure are generally too small to represent reasonable substructure postulates. Fragment counts, distance-based keys, and pharmacophores are not 2D substructure postulates. The idea of a largest substructure present on all structures in a cluster[8] can be viewed as a 2D substructure postulate but represents more the synthesis of the notion of a cluster within the similarity formalism than the notion of a substructure within the substructure formalism.

Nilakantan et al.[9] partitions the reference collection into clusters whose members have the same cyclic system. As we shall see, this represents an entirely different formalism based on atom and bond properties and the notion of a largest connected fragment. In this case, it is easy to view the cyclic-system that defines the series as a 2D substructure postulate of the critical core of a structure should it become a lead with the side chains defining a minimally specified part of the surrounding environment. A number of other 2D-substructure related notions have been proposed in contexts that suggest their relevance as possible *a priori* categories of substructure postulates. Comprehensive characterizations[10−12] began to emerge following the early work by Bemis and Murcko on molecular frameworks, ring systems, linkers, and side chains.[13] More focused studies have been directed toward cyclic systems,[9] ring systems,[14] and functional groups.[15−17]

Associating a set of *a priori* 2D substructure postulates of the critical parts of a structure should it become a lead gives rise to the concept of a coverage sublibrary. One can set up a helpful metaphor by viewing a compound as a key chain in which the "keys" are the *a priori* 2D substructure postulates. A sublibrary will be called a single-coverage sublibrary if every "key" occurring on any compound in the

* Corresponding author phone: (269)370-6107; e-mail: mark@pannanugget.com.

[†] Pannanugget Consulting.

[‡] Pfizer.

[§] Independent Chemistry Consultant.

[⊥] Michigan High Throughput Screening Center.

10.1021/ci800250r CCC: $40.75
Published on Web 02/18/2009

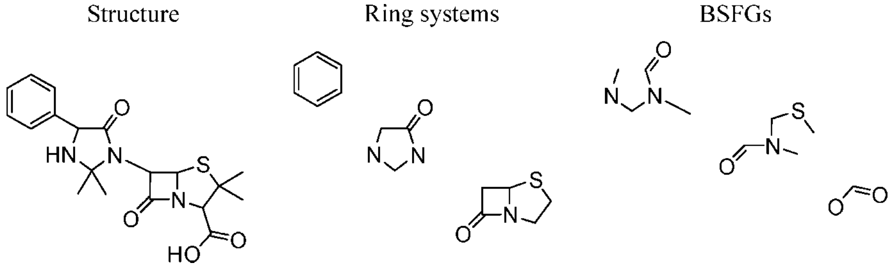


Figure 1. The ring systems and bond-separated functional groups of hetacillin.

reference library also occurs on at least one compound in that sublibrary. Thus, for example, if a particular positioning of a ring system exists on one or more compounds in the reference library, that same positioned ring system will exist on at least one compound in the single-coverage library. By this definition, the reference library is a single-coverage sublibrary of itself. However, an *efficient* single-coverage sublibrary rarely has more that one occurrence of the less common chemotypes found in the reference library.

Nilakantan et al.[9] determined that roughly 50−100 structures are needed to represent each cyclic system to be nearly certain of “finding at least one active compound in each true series for any given target”. The Michigan High-Throughput Screening Center (MHTSC) recently reported on the development of its screening library.[18] In the related reference library of ∼480K structures there were ∼100K different cyclic systems. Because the most common cyclic systems will be represented by hundreds of structures, the majority of cyclic systems in this reference library will be represented by only one or two structures. This leads to the question addressed in this study: What is the extent and nature of the information lost in screening a low-level coverage library rather than the reference library, in particular, an efficient single-coverage sublibrary.

The screening data[19] from the McMaster data-mining competition[20] provides an ideal case study for illustrating the logic in answering this question. It is large enough to illustrate a wide variety of issues that might be confronted, small enough so that it can be treated somewhat exhaustively, and by focusing on trimethoprim competitive DHFR inhibitors, one is reasonably assured that the activity regions of interest reflect a single mechanism of binding. We used only data from the training set as it reduced the scope of the calculations without sacrificing any points related to the rationale or losing any of the competitive inhibitors defining the active regions.

Four single-coverage sublibraries involving cyclic systems, ring systems, and functional groups and two “standard clustering” sublibraries for comparative purposes are examined to see which of the 12 competitive inhibitors are picked up either by the sublibrary itself or through follow-up screening. These sublibraries are roughly 20% to 30% the size of the reference library of 50K training structures. The first section presents the chemotypic categories used to construct the coverage sublibraries. The next section presents the basic protocols for constructing the sublibraries along with some minor algorithmic details for carrying out the protocols. Aspects of the operational logic and performance of the examined sublibraries are covered in the results section. By dividing the percent of all inhibitors that were uncovered by the percent of compounds screened, one obtains the *a priori* enhancement ratio. The last section discusses some of the large screening library design issues associated with *a priori* enhancement and with structural regions that might be totally ignored when using chemotypic sublibraries.

Defining Positioned Ring-Systems and Formally Defined Functional Groups As Possible Substructure Postulates. The particular chemotypic categories selected for this study reflect in part a preliminary study comparing eight different categories of potential substructure postulates with respect to their relevance to over 100 recognized drugs selected from 13 different therapeutic categories.[21] Ring systems are first defined here as the fragments (maximal connected substructures) that remain after deleting all nonring bonds (a bond whose breakage changes the number of fragments). These fragments are then augmented by any attached exocyclic double bonds. Bond-separated functional groups (BSFGs) are the fragments that remain after deleting all carbon−hydrogen bonds and all carbon−carbon single and aromatic bonds of a structure. Because of our use of hydrogen-reduced chemical-graph representations, the deletion of carbon−hydrogen bonds is superfluous. We see from Figure 1 that hetacillin has 3 ring systems and 3 BSFGs. Many of these property-based fragments would have counterparts using the Leadscope templates,[16] but the latter templates are likely to include a number of smaller functional groups which are not BSFGs such as the amine and sulfide groups in hetacillin.

The properties underlying the functional groups of Nilakantan et al.[12] are closely analogous to those used here although they, for example, distinguish aromatic hydroxyl and aliphatic hydroxyls which are not distinguished here until positioning considerations are taken into account. These BSFGs also differ from the functional groups of Xu and Johnson,[17] the latter of which cannot terminate in a single-bonded carbon.

The Merck Index[22] provides a convenient tool for empirically evaluating *a posteriori* the relevance of any 2D chemotypic postulate associated with a recognized drug. For example, the beta-lactam ring system in hetacillin occurs on 41 structures in the index, all of which are antibiotics. The corresponding lactam BSFG in hetacillin also occurs on these 41 structures but occurs on an additional 3 structures that are not antibiotics. Two of the latter 3 structures are depicted in Figure 3.

This result is surprising and merits further examination. Of the thousands of substructures of hetacillin, only six of these emerged as *a priori* chemotypic postulates of the critical core of hetacillin, and this was done *a priori*, i.e. without any consideration of what was known regarding antibiotic activity. Of these six, two were empirically

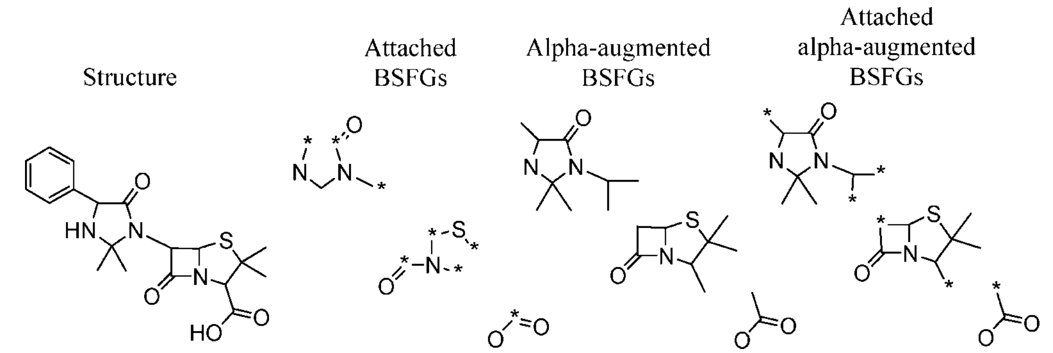
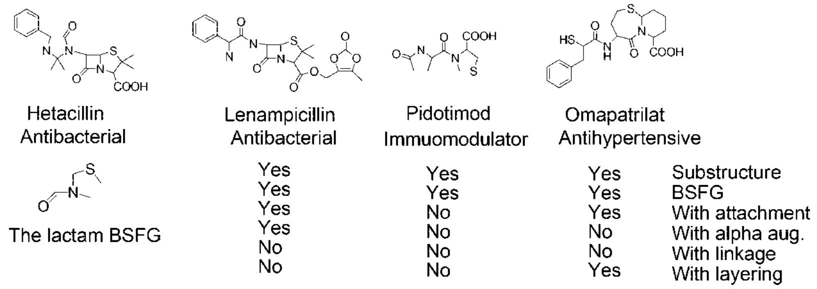


Figure 2. The BSFGs of hetacillin when attachment and alpha-augmentation are taken into account both separately and jointly.



The lactam BSFG	Lenampicillin	Pidotimod	Omapatrilat	
	Yes	Yes	Yes	Substructure
	Yes	Yes	Yes	BSFG
	Yes	No	Yes	With attachment
	Yes	No	No	With alpha aug.
	No	No	No	With linkage
	No	No	Yes	With layering

Figure 3. The varying occurrences of the lactam BSFG of hetacillin when various positioning constraints are taken into account.

substantiated in the following simple sense: First, the chemotype occurred on multiple structures. Second, a significant proportion of these were antibiotics. And third, the latter antibiotics differed substantially in those atoms and bonds that were not part of the postulated chemotypic substructure. This dramatic reduction in the search space for those substructure postulates most relevant to the chemotypic cores of active structures forms the fundamental rationale of using chemotypes as a basis for constructing screening sublibraries.

The fragments—ring systems and BSFGs in our case—that remain after deleting the atoms and bonds having some specified property will be referred to as chemotypic cores. In this lead-finding context, they function as rather crude *a priori* chemotypic postulates of the critical core. There are a number of ways by which they can be elaborated to function as much more specific postulates. One way is to simply add the atoms alpha to the core chemotype. Another is to distinguish those atoms that are bonded only to atoms of the chemotype from those that are also bonded to atoms outside of the chemotype. These two operations can be combined for additional specificity as illustrated in Figure 2 for the BSFGs of hetacillin.

Figure 3 illustrates how attachment and alpha-augmentation affect the occurrence of the lactam BSFG of hetacillin. The lactam BSFG of hetacillin has four attachment atoms as shown in Figure 2, but we see that the same occurrence structure in pidotimod has only three attachment atoms.

The occurrence of a chemotype can be additionally restricted by taking atom linkage and bond layering into account. Linkage is an attribute that qualitatively differentiates atoms by the structural types of bonds (side-chain, bridge, and ring) in which it participates. The three types can be combined eight ways if, as for a single-atom ion, the absence of any covalent bond is one of the possible combinations. As indicated in Figure 4, the carbon on the bridge to the lactam ring system has a ring and bridge linkage. Adding a methyl group at that carbon to form the methyl analog would change that carbon's linkage to side-chain&ring&bridge. The keto bonds in hetacillin are treated as terminal bonds and, consequently, as side-chain bonds in the linkage assignment (as well as in the layering assignment to follow) even though they are included in the ring system.

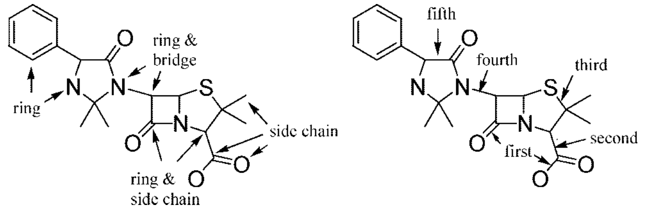


Figure 4. Atom linkage and bond layer. Hetacillin has no isolated, side-chain and bridge, and side-chain and bridge and ring linkage atoms.

Layering is an attribute that qualitatively differentiates bonds by the extent to which they are "topologically" buried within the structures. Terminal bonds are on the first layer; the remaining side-chain bonds are on the second layer; the bonds on a terminal ring system are on the third layer; the bonds on a bridge system attached to a terminal ring system are on the fourth layer; the bonds of the first inner ring system are on the fifth layer and so on and so forth.

We see that the terminal carbon bonded to the nitrogen in the BSFG depicted in the tabular part of Figure 3 has a ring and side-chain linkage in hetacillin but has a ring and bridge linkage in lenampicillin. The terminal carbon attached to the sulfur in that same BSFG has a side-chain and ring linkage in hetacillin but only a ring linkage in pidotimod and omapatrilat. If layering is taken into account, but not linkage, that same BSFG occurs on omapatrilat but not on either lenampicillin or pidotimod. The bond connecting the nitrogen and the keto carbon lies on ring layer 3 in hetacillin but on ring layer 5 in lenampicillin and on bridge layer 4 in pidotimod. Figure 3 nicely illustrates the fact that by taking different aspects of positioning into account, one can

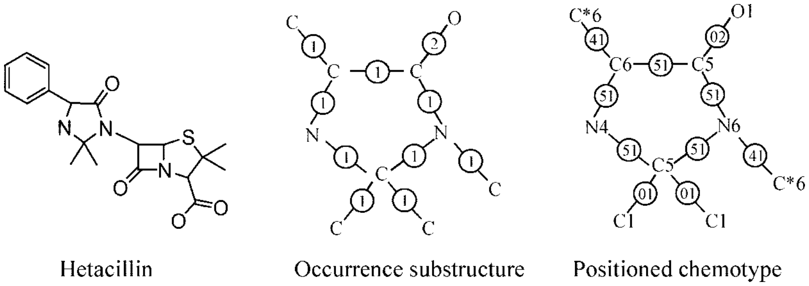


Figure 5. The positioned chemotype of the inner ring system of hetacillin together with its occurrence substructure. The labeled graph of the occurrence substructure has traditional atom types and circled bond types. Attachment and linkage are incorporated in the atom labels of the positioned chemotype, while the bond layer is incorporated in the circled bond labels.

dramatically alter the membership of the class of structures on which a particular core chemotype occurs.

The attachment information can be incorporated into the atom label with an appended asterisk. Atom-linkage assignments can be incorporated into the atom labels after mapping the eight possible atom-linkage possibilities to the numbers 0,1,...,7. Any one-to-one correspondence will suffice. Figure 5 illustrates the following correspondence: 0 - isolated; 1 - S; 2 - B; 3 - SB; 4 - R; 5 - SR; 6 - BR; 7 - SBR where S, B, and R stand for side-chain, bridge, and ring bond types, and combinations are represented by the concatenations of the corresponding letters. Thus an atom involved in only ring and bridge bonds would be assigned a linkage value of 6 appended to its atom label. Analogously, a more general bond type (when treated as a numeric value) can be formed that incorporates the layering information by multiplying the bond layer by 10 and adding the result to the bond type. Thus a level 3 single bond would be assigned a bond type of 31 and would correspond to a single bond on a terminal ring system. The result is a labeled graph[23] representation of a chemotype that takes the four aspects of positioning into account.

The labeled graphs for the positioned chemotype of the inner ring system for hetacillin and for its occurrence substructure are given in Figure 5.

The hydrogen-reduced graph representation was used to compute the chemotypes after converting alternating single and double bonds in resonance structures to “aromatic” bonds. Tautomeric distinctions were ignored in the sense that if there are multiple tautomers, only the tautomer as represented in the structure file was used. Tautomeric distinctions should not affect the conclusions in this case study as a result of the same representations being used to construct the various chemotypic sublibraries. Computations were carried out only on the largest component of a salt (or otherwise multicomponent structure) and only after halogens were assigned Hl as a generic atom type.

The following procedure was used to compute the cyclic systems after making the atom linkage and bond layer assignments:

1. Flag the side-chain bonds.

2. Remove the flags on those side-chain double bonds involving an atom incident to either a ring or bridge bond.

3. Delete the remaining flagged bonds.

The following procedure was used to compute the positioned ring systems after making the atom linkage and bond layer assignments:

1. Flag the side-chain and bridge bonds.

2. Remove the flags on those side-chain double bonds involving an atom incident to a ring bond.

3. Identify the components of the subgraph defined by the nonflagged bonds.

4. For each component proceed as follows:

a. Ignore the component if it has fewer than 5 atoms.

b. Remove all flags.

c. Flag those bonds involving an atom of the component.

d. Associate an asterisk with any atom involving both a flagged bond and a nonflagged bond.

e. Write out the subgraph consisting of the flagged bonds together with the associated asterisks and atom-linkage and bond-layer assignments.

The asterisked atoms following step 4d are bonded to an atom that is not part of the alpha-augmented ring system, i.e. are bonded to an atom beta to the ring system. The latter procedure was also used to obtain the positioned BSFGs except that step 2 was skipped, and in step 1 the carbon−carbon single bonds and the carbon−carbon aromatic bonds were flagged.

As a practical matter, each distinct chemotype was represented by a short chemotypic identifier (CID). Either a hash code[14] or an *a priori* assigned name[24] could be used. The resulting library is invariant to this choice so long as the assigned identifiers are identical if and only if the chemotypes are isomorphic. (Nonuniqueness, if rare, would give rise to slightly smaller single-coverage libraries. It would only affect the conclusions of this type of study if it resulted in a competitive inhibitor not being selected for a coverage library because a previous structure had a chemotype incorrectly assigned the same identifier. Such an event is highly unlikely and did not occur in this particular study.)

Because these four positioning constraints are defined independently of each other, one can choose to take any combination of these into account. Only two of the 16 possible combinations are considered in this study. All four positioning constraints were taken into account when constructing all but cyclic-system single-coverage sublibrary. The core ring system and BSFG chemotypes come into play in the discussion of follow-up screening and in the construction of the cyclic-system sublibrary.

METHODS

Four single-coverage sublibraries were constructed. Sublibrary 4, the smallest of these sublibraries, was based on the BSFG category. Sublibrary 3, the next smallest, was based on the ring-system category. Sublibrary 2 was based on the union of the two preceding chemotypic categories. Sublibrary 1, the largest, was based on the cyclic-system category.

The following generic protocol was used for generating the single-coverage sublibraries. The protocol starts with no structures in the list of accepted structures.

1. Get the next structure.

2. Does the structure contain a chemotype in the designated category not yet represented in the then-current list of accepted structures? If not, return to step 1.

3. Change the then-current list to include this structure.

4. Return to step 1 until all structures have been processed.

A structure was represented at step 2 in the protocol as a list of its CIDs. The CIDs of those chemotypes that were already represented at step 2 in the then-current list were stored and accessed at step 2 in the protocol as a binary tree.[25] This binary tree was updated at step 3.

This protocol is clearly sensitive to the ordering of the structures as given in the original file, and we shall see that this ordering is far from random. Issues regarding the optimization of these protocols are largely outside the scope of this study as this report focuses on those aspects of the chemotypic rationale that apply to all single-coverage protocols. However, we will show that each of the chemotypic single-coverage protocols is a worst-case scenario with regards to the structure selected to represent a particular chemotypic class.

The Meqi II software[26] was used to generate the chemotypes, to assign their CIDs, and to compute the coverage sublibraries. The authors are unaware of any cases in this study in which two chemotypes or structures having nonisomorphic labeled graph representations were assigned the same CID.

The two “standard cluster-based” sublibraries were constructed using SciTegic Pipeline Pilot version 6.0.2.0.[27] Tanimoto coefficients, computed on the MDL Public Key fingerprints, were used as the similarity measure. Clustering was based on a maximal dissimilarity partitioning algorithm. (Algorithmic details are given in the SciTegic Pipeline Pilot v6.0.2.0 User Manual). For comparative reasons, once the clusters are formed, we selected that member from each cluster that came earliest in the structure file. It will turn out that this latter choice did not represent the worst-case scenarios for the two clustering sublibraries. Thus, any comparisons of the chemotypic coverage sublibraries with these “standard cluster-based” sublibraries must take the latter choice into account as well as the many choices that must be made in constructing any cluster-based sublibrary.

If, as in our case, one’s attention is restricted to only those structures in the reference library, a screening effort that begins by screening a sublibrary can be viewed as consisting of two phases. In the first phase the sublibrary is constructed and screened. In the second phase, the hits available after this initial screening can be used to select additional structures from the reference library to be used for follow-up screening. The basis for constructing the sublibrary need bear no relationship to the basis for selecting structures for follow-up screening and can reflect different formalisms. For example, similarity searches involving the hits obtained using a single-coverage sublibrary involving functional groups could be used to select the additional structures from the reference library. However, we will use the same chemotypic formalism for both phases in the case of the single-coverage chemotypic sublibraries, as that is the formalism being explicated. Likewise, to simplify matters and to make things comparable, we will use the clustering formalism for both phases in the case of the cluster-derived sublibraries.

Selecting structures in the second phase is very simple under the clustering formalism. For each hit in the sublibrary, one selects those additional structures in the reference library that lie in the same cluster or class. Analogously, for each hit in the cyclic-system sublibrary, one selects those additional structures in the reference library that lie in that hit’s cyclic-system class.

Because most structures have more than one ring system and BSFG chemotype associated with them, the relevant follow-up screening procedure is somewhat more complicated. However, each such chemotype defines the class of those structures with that particular chemotype. Finding the structures in that class entails nothing more than a substring search of the CID for the relevant chemotype. The number of sublibrary hits in that class can be divided by the number of sublibrary structures in that class to form an activity ratio associated with its defining chemotype. By selecting that chemotype of a hit having the highest activity ratio, a single class of structures is associated with each hit. In case of ties, one could simply take the union of the relevant classes. Such ties are likely to be rare, and there were none in this case study. One now proceeds as in the clustering formalism, i.e. for each hit in the sublibrary, one selects those additional structures in the reference library that lie in the class of the selected chemotype.

RESULTS

Figure 6 shows the 12 inhibitors of DHFR that bound competitively[19] in the screening of the 50K library of the McMaster data-mining study.[20] The bold-faced numbers correspond to the structure numbers in the Zolli-Juran[19] study. The rank of the structure in the file of the training data is given directly below the structure number. With only 12 competitive inhibitors in a 50K file, it is clear from structures **1**−**5** and structures **6** and **7** that the structures are far from randomly sequenced in the file.

The competitive inhibitors present in each sublibrary are indicated with the number ‘1’ in Table 1. There we see that sublibraries 1−6 respectfully included 5, 7, 7, 4, 3, and 5 of the competitive inhibitors.

Table 1 also includes the size of each sublibrary. Selecting 10,000 or a fifth of the structures at random can be expected to pick up a fifth of the competitive inhibitors. Under this assumption of random selection, the likelihood of having the observed number or more of competitive inhibitors in a sublibrary is given in Table 2. There we see that the likelihood of obtaining 7 or more hits in the case of sublibrary 2 with ∼15,000 structures is 0.038 and obtaining that same number of hits in the case of sublibrary 3 with ∼10,000 structures is <0.01, i.e. both likelihoods are less than what might be expected under random selection. However, chance factors alone cannot be ruled out in the cases of sublibraries 1, 4, 5, and 6 in which only 5, 4, 3, and 4 inhibitors, respectively, were found based on the initial screening.

The competitive inhibitors that were not in the screening sublibraries but were found based on the follow-up screening

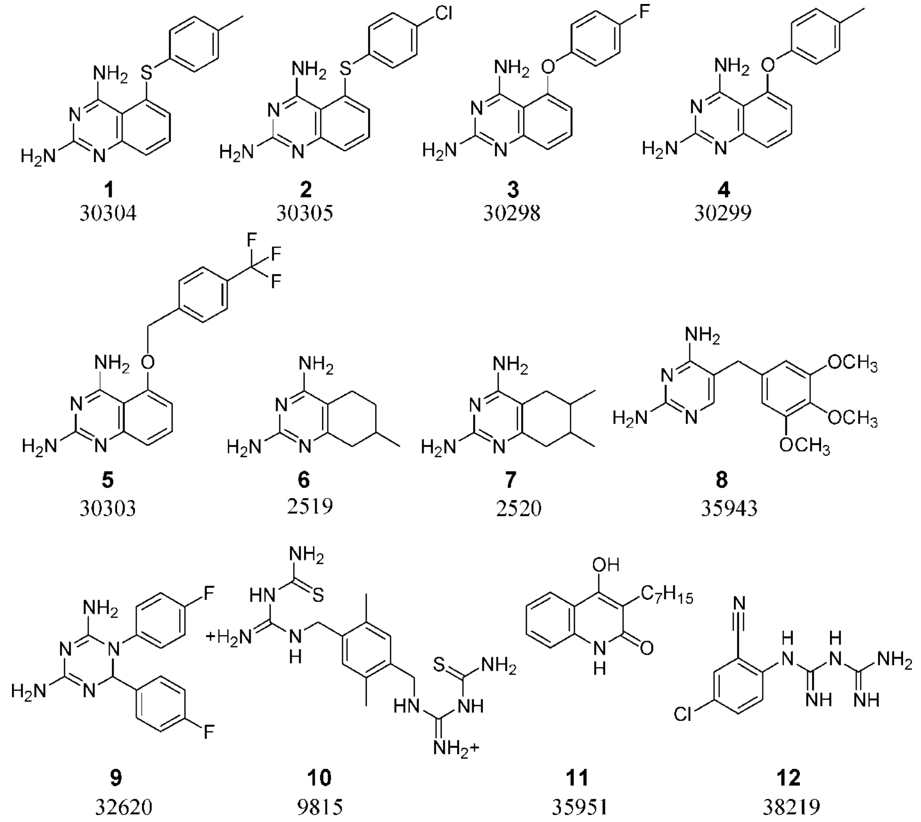


Figure 6. The 12 DHFR inhibitors that bind competitively in the McMaster study along with the ranks of their records in the structure file.

Table 1. Enhancement Ratios and Competitive Inhibitors Uncovered through the Use of Each Sublibrary[a]

				structure											
sublibrary	size	type	enhance-ment	1	2	3	4	5	6	7	8	9	10	11	12
1	16,600	CycSys	2.0	1	2	1	2	1	1	2	–	1	–	–	–
2	14,708	RngSys& BSFG	2.8	1	2	1	2	2	1	1	1	1	1	–	–
3	10,182	RngSys	3.7	1	2	1	2	2	1	1	1	1	1[b]	–	–
4	9,653	BSFG	4.3	2	2	1	2	2	1	2	3	1	1	–	–
5	10,000	cluster	2.5	1	2	1	–	2	1	2	–	–	–	–	–
6	15,000	cluster	1.9	1	1	1	2	2	1	2	–	–	–	–	–

[a] '1' marks those inhibitors found in the initial screening; '2' marks those inhibitors found using the follow-up screening protocol described in the Methods section; '3' marks the inhibitor found using a more sophisticated follow-up screening logic; '--' marks those inhibitors that were missed as a result of using the associated sublibrary. [b] Structure was included for the wrong reason.

Table 2. Probability of Getting at Least the Indicated Number of Hits for a Sublibrary of the Specified Size Had the Structures Been Selected at Random

	number of hits				
size of the sublibrary	≥4	≥5	≥6	≥7	≥8
15,000	0.51[a]	0.28	0.12	0.038	<0.01
10,000	0.21	0.073	0.019	<0.01	<0.001

[a] The probability was determined using the binomial distribution.[28] The likelihood of a successful trial was given by the ratio of the sizes of the sublibrary and the reference library.

protocol in the methods section are indicated in Table 1 by the letter '2'. For sublibrary one, 4 additional structures were screened, and 3 additional inhibitors were found yielding a follow-up hit ratio of 3 out of 4. For the remaining three chemotypic sublibraries, those values were 3 out of 5, 3 out of 4, and 5 out of 5, respectively. This significantly increased the number of competitive inhibitors uncovered after this follow-up screening to 8, 10, 10, and 9 for the four respective sublibraries. Based in Table 2, these numbers are far greater than what can be explained by chance factors alone and come at a negligible increase in the total number of compounds screened. For the two cluster libraries, those values were 3 out of 5 and 3 out of 7. This increased the number of competitive inhibitors uncovered to 6 and 7 with both values being statistically significant at the 0.05 level but not at the 0.01 level.

Figure 7 illustrates the occurrence substructures of the chemotypes that came into play for six different inhibitors. Aromatic bonds are denoted by dashed bonds so that the BSFGs of inhibitors **3** and **6** are distinguished pictorially. Note that because linkage is being taken into account, all three carbon−carbon bonds in the BSFG of inhibitor **6** are recognized as ring bonds. Although halogens are distinguished in the depicted structure, they are assigned 'Hl' as an atom type in the drawings of the chemotypes to reflect our ignoring such distinctions in the construction of the sublibraries and our analysis of the initial screening data.

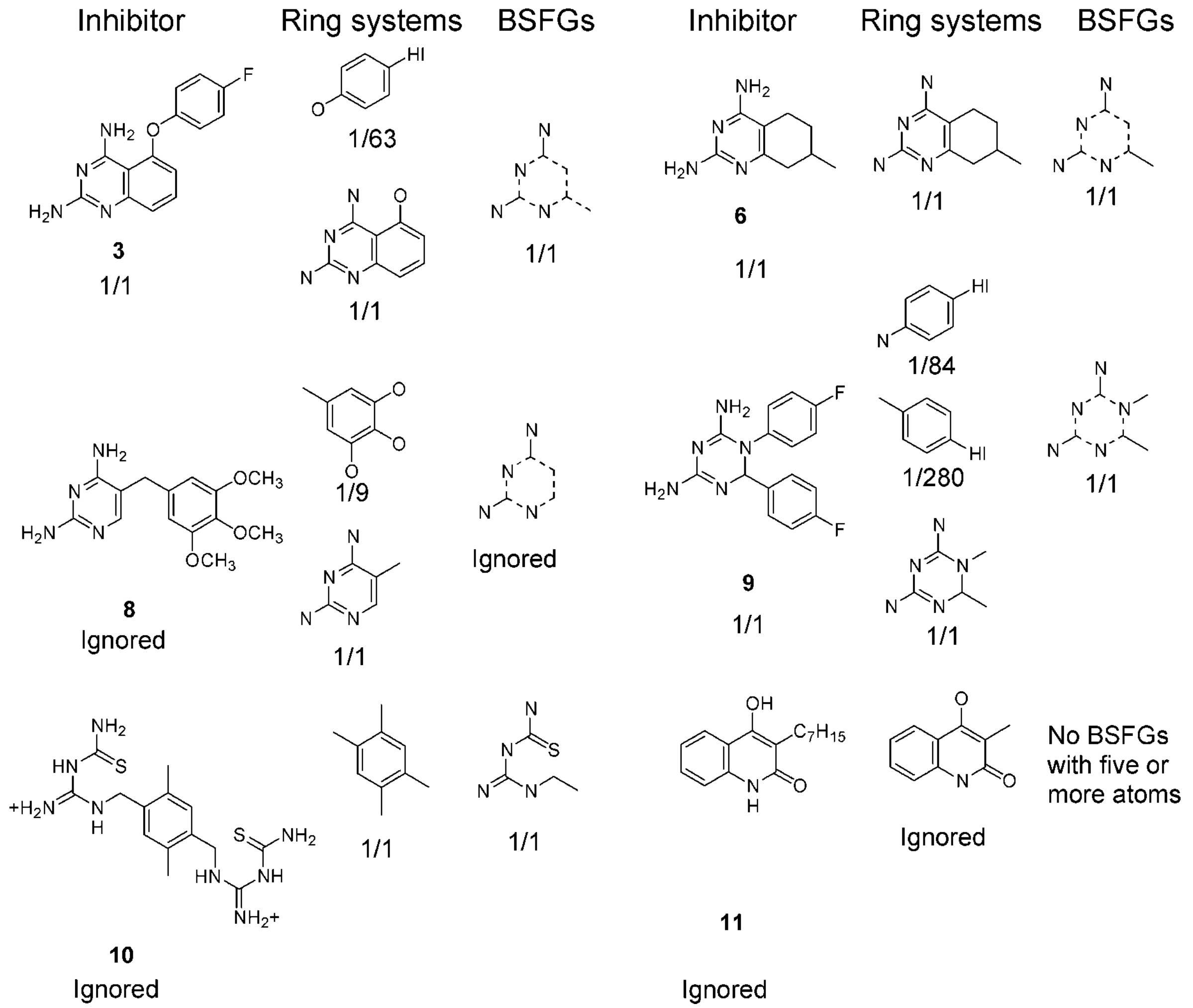


Figure 7. The structures of selected inhibitors and the occurrence substructures of their ring-system and BSFG fully positioned chemotypes. The occurrence substructures of the cyclic-systems are easily inferred from the structure. Below each occurrence substructure is the ratio of actives to inactives of those structures in the respective library having the relevant chemotype. The ratio is indicated as 'ignored' for those chemotypes occurring on an inhibitor that was not included in the relevant sublibrary. Inhibitor **11** had no BSFGs satisfying the five-or-more atom constraint.

Figure 7 also gives the activity ratios associated with the chemotypes of the six inhibitors. For example, inhibitor **3** occurred in all of the chemotypic sublibraries (as well as the two cluster sublibraries). It was the only compound with its particular cyclic system in the cyclic-system sublibrary and, consequently, was assigned an activity ratio of 1/1. This cyclic-system activity ratio is placed directly below the structure in Figure 7. Inhibitor **3** has two ring systems. Their corresponding fully positioned chemotypes had activity ratios of 1/63 for the benzene chemotype and 1/1 for the quinazoline chemotype resulting in the selection of the latter chemotype as the basis for follow-up screening. Inhibitor **3** had three BSFGs, but only one passed the five-or-more atom cutoff. Its fully positioned chemotype had an activity ratio of 1/1.

Two other examples merit explanation. Because inhibitor **8** was not present in the cyclic-system sublibrary, its cyclic system was ignored in the follow-up screening. This cyclic system was represented by 19 structures in the reference library of which only inhibitor **8** was active. Thus, the chances are small that this class would be represented by an active in an efficient single-coverage cyclic-system sublibrary had the structures been randomly permuted prior to implementing the sublibrary construction protocol. Similarly, inhibitor **8** is the only active out of four structures in the reference library that share its fully positioned BSFG. In this case there is a 1 in 4 chance that this class would be represented by an active in an efficient single-coverage BSFG sublibrary had the structures been randomly permuted prior to implementing the sublibrary construction protocol. Because inhibitor **8** was not the first of these four structures in the reference structure file, it was not a member of sublibrary 4. Consequently, this BSFG was ignored in the first follow-up screening tally. However, it was uncovered as indicated by the 3 in Table 1 using an enhancement to the follow-up screening protocol to be discussed shortly. Inhibitor **8** is the only structure in the reference library sharing its fully positioned pyrimidine ring system, and, consequently, inhibitor **8** would be present in every single-coverage ring-system sublibrary. This inhibitor's clusters in the 10K and 15K cluster sublibraries contained 10 and 3 compounds, respectively, and, consequently, there is a 1 in 10 and a 1 in 3 chance that these classes would be represented by an active had the structures been randomly permuted prior to implementing the sublibrary construction protocol. Because inhibitor **8** was not the first structure in these two classes as ordered in the reference structure file, it was not uncovered in either sublibrary 5 or sublibrary 6.

As just noted, the sizes of the chemotypic classes and clusters containing inhibitor **8** were 19, 1, 1, 4, 10, and 3 for sublibraries 1−6, respectfully. Contrary to what one might intuitively expect, inhibitor **8** was the *only* structure in its size-19 cyclic-system class that was also in its size-10 cluster in the 10K cluster sublibrary. Moreover, it was the *only* structure in its size-4 BSFG class that was also in its size-3 cluster in the 15K cluster sublibrary. This illustrates the dramatic differences in the shapes in chemical space of the structure classes that can arise when using different bases for designing sublibraries and protocols for follow-up screening.

Inhibitor **11** is the other example meriting attention. Neither of the two BSFGs of inhibitor **11** passed the five-or-more atom cutoff. Consequently, this inhibitor fell in a region of chemical space totally ignored by sublibrary 4, an issue to be discussed later.

The membership of a structural class or cluster based on a similarity measure is defined *a posteriori*, i.e. it is defined with respect to a particular set of compounds. In contrast, the membership of a chemotypic class is defined *a priori* in terms of particular atom and bond properties. Moreover, this class is represented by a visually recognizable and formally defined chemotype. Two advantages of having a chemotypic class representation will now be illustrated.

When two inhibitors share the same core chemotype but differ in the positioning of that core chemotype, attention is naturally directed to that core chemotype. Every distinct positioning of that core that is present in the reference library will be represented by at least one structure in a single-coverage sublibrary. We now note from Figure 7 that the most active BSFGs (in this case the only BSFGs) of inhibitors **3** and **6** are different positionings of the same core chemotype. This suggests that all of the structures in the reference library having that core chemotype should be screened so long as that core chemotype is shared by only a handful of other structures in sublibrary 4 that are not competitive inhibitors. There were only three such structures, namely **13**, **16**, and **17**. Thus, we see that 2 out of 5 structures in sublibrary 4 having this core chemotype were competitive inhibitors. The importance of that high activity ratio to follow-up screening is augmented by the diversity of the five structures. By enhancing our follow-up screening protocol to include all structures sharing this common core BSFG, trimethoprim, i.e. inhibitor **8**, is uncovered. This increases the number of competitive inhibitors found using sublibrary 4 to 10 with no appreciable increase in the number structures screened.

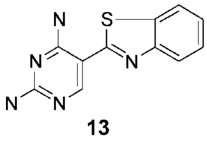
As a sidelight, picking up trimethoprim as a result of screening structures **3** and **6** is an interesting instance of scaffold hopping[29] within the chemotypic formalism, while screening structures **1**−**8** and **13**−**17** is an interesting instance of assembling an SAR data set for a lead finding effort.

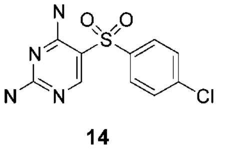
The same argument can be applied to inhibitors **1** and **3** vis-à-vis sublibrary 3 as the fully positioned ring systems associated with the highest activity ratios in these two inhibitors share the same quinazoline core. There were 32 structures in sublibrary 3 having a quinazoline ring system representing 31 different fully positioned quinazoline ring systems. Finding 2 actives out of 32 structures is not as encouraging as 2 out of 5 but would still suggest that other actives might be found with a very modest additional screening effort. After follow-up screening, the two subregions defined by the fully positioned quinazoline ring systems of inhibitors **1** and **3** were represented only by actives, i.e. inhibitors **1**−**2** and **3**−**5**, respectively. The other 29 subregions were represented by 60 structures in the reference library, none of which were competitive inhibitors. This enhancement to the follow-up screening protocol is not relevant when using sublibraries based on core chemotypes as in the case of cyclic systems.

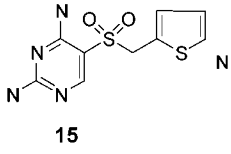
As another example of using chemotypes as an interpretable specification of the class of structures that they define, consider competitive inhibitor **10**. It was selected for sublibrary 3 because it was the only member in its fully positioned ring system class. This chemotype defines those structures that are 1,2,4,5-substituted benzenes with two methyl groups and such that the atoms alpha to the benzene ring are all carbons, all substituents are side chains and the two methyl groups are para to each other. This is not a meaningful class for which to select compounds for follow-up screening as its defining chemotype contains no hydrogen-bond donors or acceptors. It was selected for the ring-system sublibrary *not* because a 1,2,4,5-substituted benzene ring is critical to activity, but because it has a critical BSFG that happened to occur on the only compound in the reference library with that unlikely ring-system substitution pattern. Because this inhibitor occurred in the sublibrary for the wrong chemotypic reason, the number of competitive inhibitors found using sublibrary 3 was reduced to 9 when computing the *a priori* enhancement.

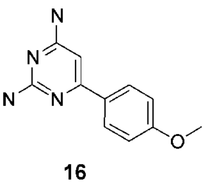
All six sublibraries failed to uncover competitive inhibitors **11** and **12** because neither inhibitor came first in the reference structure file in their respective classes as defined by the six sublibraries. The sizes of these classes for inhibitor **11** for sublibraries 1, 2, 3, 4, 5, and 6 were 27, 5, 5, ?, 3, and 2. A question mark is reported for the size of the class for sublibrary 4 as inhibitor **11** contained no BSFGs that satisfied the 5-atom cutoff, an important issue to be discussed later as 11,951 or roughly 25% of the structures in the reference library fell into this ignored class. The sizes of these classes for inhibitor **12** for sublibraries 1−6 were 3462, 10, 10, 23, 35, and 26, respectively. For sublibrary 2, the reported size is the smaller of the two sizes for the corresponding ring system and BSFG classes.

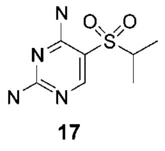
It will now be empirically argued that the latter two inhibitors do not come from chemotypically interesting DHFR activity regions and that, if they had, they would have most likely been represented by an active in sublibrary 3 in the case of inhibitor **11** and in sublibrary 4 in case of inhibitor


13
13912


14
24393


15
25773


16
14392


17
25174

12. Looking first at inhibitor **12**, neither its cyclic-system nor its alpha-augmented ring-system represent interesting activity regions as neither chemotype entails a proton-donator or acceptor. That its largest fully positioned BSFG does not represent an interesting activity region is reflected by the observation that it was the only active among 23 structures in the reference library that shared its BSFG and among the 61 structures in the reference library that shared its *core* BSFG. Moreover, those 61 structures entailed 5 different subregions defined by the different fully positioned versions of that core BSFG, each represented by one structure in sublibrary 4. Thus, had it been an interesting activity region with either a highly active subregion as defined by a fully positioned BSFG or two or more moderately active subregions, there is a good chance that one of those 5 representatives would have turned up active.

In the case of inhibitor **11**, sublibrary 4 implies little with regards to interest in the active region of inhibitor **11** other than it does not contain any structures with a core BSFG that passes our five-or-more atom restriction. It was the only active out of five structures in the reference library that shared its fully positioned ring system and only one active out of 57 structures that shared its ring system (as opposed to the 27 just reported that shared its cyclic system). Moreover, these 57 structures were dispersed among 12 subregions defined by different fully positioned versions of that ring system. Such statistics are not encouraging from a lead-development point of view. Had the region defined by this core ring system included a highly active subregion or two or more moderately active subregions, there is a strong chance at least one of the 12 representatives in sublibrary 3 of these subregions would have been represented by an active.

We now want to show that the critical aspects related to the performance of the four chemotypic sublibraries were independent of the ordering of the structures in the structure file. Certainly, the choice of a single representative from a chemotypic class would make no difference if (1) all of the structures in that class were inactive or (2) if all of the structures were active or (3) if there was only one structure in the class. Case 1 holds for all but the few classes containing at least one inhibitor. Inhibitors **1**−**7** are instances of classes in case 2 for all four sublibraries. Inhibitor **9** is an example of case 3 for all four sublibraries as is inhibitor **10** for sublibraries 2−4. Sublibrary 1 represented the worst-case scenario in failing to pick up inhibitor **10** as did all four sublibraries in failing to pick up inhibitors **11** and **12**. Finally, inhibitor **8** is an instance of case 3 for sublibraries 2 and 3 but is a worst-case scenario for sublibraries 1 and 4. Thus, regardless of how we might have permuted the structures prior to implementing the library construction protocols, the inhibitors reported as uncovered inhibitors in Table 1 would not have been missed. However, there would been a reasonable chance for each sublibrary to have uncovered one additional inhibitor. For example, there would be a 1 in 4 chance for sublibrary 1 to uncover inhibitor **8**, a 1 in 5 chance for sublibraries 2 and 3 to uncover inhibitor **11**, and a 1 in 23 chance for sublibraries 2 and 4 to uncover inhibitor **12**. The chance of selecting inhibitors **10** and **12** for sublibraries 1 and 3 is of no moment chemotypically as the implicated postulates of the critical core entail no proton donors or acceptors.

Similar worst-case scenarios for cluster sublibraries five and six would result in only 2 and 5 inhibitors being found upon follow-up screening, neither of which is statistically significant as seen from Table 2. With regards to sublibrary five, inhibitors **1** and **2** as well as inhibitors **6** and **7** came from size-3 clusters each of which included an inactive that would be the selected representative in a worst-case scenario. With regards to sublibrary six, inhibitors **1** and **2** both came from size-2 clusters each of which included an inactive. If, instead of picking the first structure in each cluster, we had used the PipelinePilot default of picking the structure closest to the center of each cluster, the resulting 10K sublibrary would have uncovered inhibitor **11** in addition to the six inhibitors reported as uncovered in Table 1. Picking up 7 inhibitors using a 10K sublibrary was statistically significant at the 0.01 level. Using this closest-to-the-center selection method in constructing the 15K cluster library did not change the number of uncovered inhibitors but did result in uncovering inhibitor **11** while failing to uncover inhibitor **1.**

Had these structures been screened with respect to a different target, we would want to uncover a different set of activity regions. Although we can only speculate as to how these sublibraries might perform as regarding a different set of activity regions, one can make informative statements regarding relative performance under appropriate assumptions regarding the nature of the activity regions. For example, suppose the activity regions correspond to the cyclic-system category. More specifically, suppose

1. That all actives have one of only a handful of cyclic systems;
2. That every compound having one of those cyclic systems is active.

Clearly, the single-coverage cyclic-system sublibrary would uncover every active in every one of the active regions picked up by the reference library, i.e. none would be missed by screening sublibrary 1. Each active region would be represented by one active when the sublibrary was screened. The remaining actives would be uncovered on follow-up screening.

But what percentage of the cyclic-system active regions and what percentage of the actives would be uncovered using the other five sublibraries? Here we only address the percentage of active regions that would be detected as a result of active structures being members of that sublibrary. For a group of targets with different active regions, but all satisfying our same two suppositions, that percentage would vary, but the average would approximate the percentage of cyclic systems represented in the various other sublibraries. These percentages are given in the cyclic-system column of Table 3.

It is not too surprising that sublibraries 3 through 5 which are roughly 60% of the size of the cyclic system sublibrary have slightly less than 40% of the cyclic systems represented, but it is somewhat surprising that sublibraries 2 and 6 which are roughly 90% of the size of the cyclic-system sublibrary represent only around 50% of the possible cyclic systems.

Staying with these two suppositions, consider what would happen should we have a hit using one of the other five sublibraries. Using sublibrary 4 as an example, one would operationally, but incorrectly, assume that a fully positioned BSFG was responsible for its activity and, consequently, those structures possessing that BSFG would be selected for

Table 3. Number of Structural Classes with Respect to Each Generational Basis Represented in Each Sublibrary Expressed As a Percentage of the Corresponding Number in the Bottom Row That Are Represented in the Reference Library

		generational basis					
sublibrary	type	cyclic system	RngSys & BSFG	ring system	BSFG	10K clustering	15K clustering
1	CycSys	100	63	65	62	67	60
2	RngSys & BSFG	50	100	100	100	68	58
3	RngSys	38	79	100	57	54	44
4	BSFG	36	78	58	100	47	40
5	10K cluster	36	50	54	46	100	65
6	15K cluster	51	63	66	60	98	100
total number		16600	19919	10656	9840	10000	15000

Table 4. Each Generational-Basis Column Gives the Efficiency of the Corresponding Sublibrary Relative to Each of the Other Sublibraries and to the Reference Library with Respect to Coverage of the Structural Classes That Basis Generated

		generational basis					
library	type	cyclic system	RngSys & BSFG	ring system	BSFG	10K clustering	15K clustering
ref	–	3.01	2.51	4.69	5.08	5.00	3.33
1	CycSys	1.00	1.79	2.52	2.77	2.49	1.84
2	RngSys & BSFG	1.76	1.00	1.44	1.52	2.18	1.69
3	RngSys	1.62	0.88	1.00	1.84	1.88	1.53
4	BSFG	1.61	0.84	1.63	1.00	2.05	1.62
5	10K cluster	1.66	1.37	1.83	2.27	1.00	1.02
6	15K cluster	1.77	1.62	2.23	2.59	1.53	1.00
ratio of max to min		1.77	2.13	2.52	2.77	2.49	1.84

follow-up screening. Without specifying exactly which cyclic systems represent the active regions, it is impossible to say what percentage of actives would show up in this follow-up screening. However, it is evident that the suggested BSFG generally would not occur on all of the structures sharing the cyclic system of the hit and generally would occur on a number of other structures that did not share the cyclic system of the hit. The same reasoning would hold for the other four sublibraries. Thus, using any one of sublibraries 2−6 when the active regions correspond to cyclic-system classes not only results in a failure to find 50% to 60% of the active regions but also results in a failure to find all of the actives sharing the cyclic systems of those hits that did turn up in the initial screening.

Substituting fully positioned BSFGs as the generational basis in these two suppositions yields an even more surprising result. Although sublibrary 1 has roughly 1.72 times more structures than sublibrary 4, based on the results in Table 3, the latter sublibrary represents 100/62 or roughly 1.63 times the number of fully positioned BSFG chemotypic classes. Thus, the number of such fully positioned BSFG classes represented by at least one structure in sublibrary 4 is 1.72 × 1.63 or roughly 2.8 times greater than the corresponding number for sublibrary 1 relative to its size. This and analogous relative efficiencies are given in Table 4 for all 36 sublibrary and generational-basis combinations.

To understand Table 4, note first that relative efficiency is always with respect to a particular generational basis and relative to the sublibrary created using that generational basis. Consequently, the efficiencies on the diagonal are all one, and the basic comparisons between sublibraries are made between values within a column. Intuitively, relative efficiency generally increases with the percent of structural classes represented by a sublibrary as indicated by the corresponding column in Table 3 and generally decreases with the relative size of the sublibrary. Because every structural class is, by definition, represented by at least one structure in the reference library regardless of the basis for generating the structural classes, its relative efficiency is simply the ratio of its size to that of the sublibrary with the relevant generational basis.

The size effect can largely be adjusted out by comparing sublibraries of comparable size. The efficiency of sublibrary 3 in the ring-system column is 1.63 times that of sublibrary 4 and 1.83 times that of sublibrary 5. The efficiency of sublibrary 4 in the BSFG column is 1.84 times that of sublibrary 3 and 2.27 times that of sublibrary 5. The efficiency of sublibrary 5 in the 10K-clustering column is 1.88 times that of sublibrary 3 and 2.05 times that of sublibrary 4. Similarly, the efficiency of sublibrary 1 in the cyclic-system column is 1.76 times that of sublibrary 2 and 1.77 times that of sublibrary 6. The efficiency of sublibrary 2 in the RngSys & BSFG column is 1.79 times that of sublibrary 1 and 1.62 times that of sublibrary 6. The efficiency of sublibrary 6 in the 15K clustering column is 1.84 times that of sublibrary 1 and 1.69 times that for sublibrary 2. Thus, among the three ∼10K-sized sublibraries, the relative efficiencies varied between 1.63 and 2.27, while among the three ∼15K-sized sublibraries, the relative efficiencies varied between 1.62 and 1.84.

Interestingly, in the 15K-cluster column, sublibrary 6 is seen to be only 1.02 times as efficient as sublibrary 5, while in the RngSys & BSFG column, the derived sublibrary, sublibrary 2 is seen to be less efficient than both sublibraries 3 and 4. The latter case reflects the fact that the percentage differences in coverage in the corresponding column of Table 3 is less than the percentage differences in the sizes of the corresponding sublibraries. Just why this is so and is not so in the former case (the 15K-cluster column) is somewhat puzzling.

These relative efficiencies reflect ideal situations largely governed by the second supposition that all structures in the formally defined structural classes are active. If we drop the second supposition but retain the first, the assumed active regions will correspond to structural classes as defined by the assumed generational basis that can contain both actives and inactives. If one of the active regions is represented by a structural class containing only one structure in the reference sublibrary, that structure will necessarily be included in the sublibrary created using that generational basis. If that structure is active, it will be found when screening either the reference library or the relevant sublibrary. If not, that active region will not be discovered in either case. The percentage of such singleton classes for the six generational bases and their corresponding sublibraries 1−6 are 61%, 49%, 47%, 51%, 23%, and 35%, respectively. Thus, based solely on supposition 1, the relevant chemotypic sublibrary will do just as well as the reference library in uncovering on average at least half of the active regions for those targets whose active regions correspond to the structural classes created by the corresponding generational basis.

From a more practical point of view, good performance of a sublibrary is a combination of success in turning up all of the competitive inhibitors in active regions meriting a lead optimization effort and screening a relatively small number of structures. If we restrict our attention to those competitive inhibitors discovered that were associated with a reasonable chemotypic postulate, then sublibraries 1−6 uncovered 8, 10, 9, 10, 6, and 7 inhibitors, respectively. This represents 63%, 83%, 75%, 83%, 50%, and 58% of the 12 competitive inhibitors and slightly higher percentages (80%, 100%, 90%, 100%, 60%, and 70%) if we exclude inhibitors **11** and **12** as a result of their being in largely inactive regions as defined by their cyclic systems, ring systems, and BSFGs. If we divide the first percentages by the sizes of the corresponding sublibraries relative to that of the reference library, i.e. by 33%, 30%, 20%, 20%, 20%, and 30%, we obtain empirically based *a priori* enrichment ratios of 1.9, 2.8, 3.8, 4.2, 2.5, and 1.9.

DISCUSSION AND SUMMARY

We have seen how vastly different can be the structural classes under the six generational bases considered in this study, an observation consistent with the findings of Cheng et al.[30] and the Rand analyses of Schuffenhauer et al.[31] Such differences are reflected in the relative efficiencies that differed by close to 3-fold vis-à-vis the bottom row of Table 4. This is reflected in *a priori* enhancement ratios in this case study that varied by more than 2-fold (4.2/1.9). Thus, strong inferences regarding performance magnitudes and orderings are simply not justified based upon a single case study. However, that the BSFG sublibrary performed well suggests there may be a class of targets for which such libraries will perform well. That one could enhance performance by a small change in the follow-up screening logic suggests other enhancements are likely to be forthcoming. That there were explicable differences in the performance of the four chemotypically based sublibraries regarding one target suggests that better chemotypic bases for generating coverage sublibraries applicable to broad screening will be found.

Although the single-coverage sublibrary based on the fully positioned BSFG chemotypic category performed well in this case, the fact that 11,951 or roughly 25% of the structures in the reference library had no core BSFGs satisfying the 5-atom restriction is intriguing. Should such sublibraries perform well across a broad range of targets or across a target family, ignoring such structures may be desired. If not, there are many tradeoffs by which it might be circumvented. By using the ring-system chemotypic category rather than the BSFG category, sublibrary 3 ignored only 311 or roughly 0.6% of the structures lacking a ring system under the same size restriction. Most, 279, of these were acyclic structures that were also ignored in sublibrary 1. This dramatic reduction in the chemotypically ignored region came at the cost of an *a priori* enhancement reduction of roughly 10%. By taking both BSFGs and ring systems into account, sublibrary 2 ignored only 108 structures or roughly 0.2% of the structures. These would be mainly acyclic structures with small BSFGs but came at a cost of a 50% increase in size and a corresponding 33% reduction in the *a priori* enrichment. Alternatively, one can select representatives using a dissimilarity based or any other relevant approach from this BSFG-ignored region. Each approach would come with its corresponding tradeoff in size and *a priori* enrichment.

The consideration of *a priori* enhancement ratios is important from two standpoints. The first is simply its utility as a measure of the reduction in the number of structures that must be screened per active found. In this particular case study, a 4.2-fold reduction was achieved in screening the BSFG sublibrary rather than the reference library. More importantly, it is a measure of the reduction in the number of structures per unrepresented active region that must be purchased so as to bring such regions within the representative scope of the reference library. For example, suppose one were to augment sublibrary 4 with 40K structures drawn from the same region of chemical space as was the reference library but did so using a single-coverage protocol analogous to that described in the Methods section. This would result in a screening library the same size as the McMaster training library but with 5 times the number of positioned BSFG classes being represented.

Exploring these suggestions necessitates the development of the chemotypic reasoning behind the construction and use of any chemotypic coverage sublibrary in broad-based screening. Moreover, the chemotypic logic of follow-up screening is relevant regardless of how the screening sublibrary is generated. The positioned and core chemotypes can be computed on any sublibrary. Once the *a priori* most logical chemotypic postulates have been determined for each active, they can be used to select the structures for follow-up screening in pretty much an automated fashion. The search for multiple active subregions subsumed in a larger region defined by a core chemotype (or chemotypes that take only one or two of the positioning-considerations into account) is more analysis intensive but could be effectively automated in cases where there are a large number of initial hits. Such types of reasoning should prove relevant across a broad range of targets and target families.

ACKNOWLEDGMENT

The authors would like to thank the members of the Michigan High-Throughput Screening Advisory Board, Jim Dehaven, John McCall, and Doug Morton, for their encouraging support as our efforts got underway in developing and implementing the idea of a chemotypic coverage sublibrary and the McMaster HTS Laboratory (http://hts.mcmaster.ca) for making available the screening data used in this study. We especially want to thank the reviewers for their time and significant contributions to the evolution of this research and its report.

REFERENCES AND NOTES

(1) Jacoby, E.; Schuffenhauer, A.; Popov, M.; Azzaoui, K.; Havill, B.; Schopfer, U.; Engeloch, C.; Stanek, J.; Acklin, P.; Rigollier, P.; Stoll, F.; Koch, G.; Meier, P.; Orain, D.; Giger, R.; Hinrichs, J.; Malagu, K.; Zimmermann, J.; Roth, H.-J. Key Aspects of the Novartis Compound Collection Enhancement Project for the Compilation of a Comprehensive Chemogenomics Drug Discovery Screening Collection. *Curr. Top. Med. Chem.* **2005**, *5*, 397–411.

(2) Müller, G. Medicinal Chemistry of Target Family-Directed Masterkeys. *Drug Discovery Today* **2003**, *8*, 681–691.

(3) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd.: Letchworth, England, 1987.

(4) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley: New York, NY, 1990.

(5) Dean, P. M. *Molecular Diversity in Drug Design*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1999.

(6) Martin, Y. Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J. Comb. Chem.* **2001**, *3*, 231–250.

(7) Harper, G.; Pickett, S. D.; Green, D. V. S. Design of a Compound Screening Collection for Use in High Throughput Screening. *Comb. Chem. High Throughput Screening* **2004**, *7*, 63–70.

(8) Gardner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007**, *47*, 354–366.

(9) Nilakantan, R.; Immermann, F.; Haraki, K. A Novel Approach to Combinatorial Design. *Comb. Chem. High Throughput Screening* **2002**, *5*, 105–110.

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

(11) Xu, J.; Stevenson, J. Drug-like Index: A New Approach to Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.

(12) Nilakantan, R.; Nunn, D. S.; Greenblatt, L.; Walker, G.; Harki, K.; Mobilio, D. A Family of Ring System-Based Structural Fragments for Use in Structure-Activity Studies: Database Mining and Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 1069–1077.

(13) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(14) Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. A Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic. *J. Med. Chem.* **1990**, *30*, 65–68.

(15) Cosgrove, D. A.; Willett, P. SLASH: A Program for Analyzing the Functional Groups in Molecules. *J. Mol. Graphics Modell.* **1998**, *16*, 19–32.

(16) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. Leadscope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.

(17) Xu, X.-J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.

(18) Johnson, M.; Bundy, D.; Chapman, D.; Kilkuskie, R. Michigan HTS Screening Center: Designing a Compound Library for Maximum Diversity. *SBS News [Soc. Biol. Screening]* **2007**, *29*, 3&6.

(19) Zolli-Juran, M.; Cechetto, J. C.; Hartlen, R.; Daigle, D. M.; Brown, E. D. High Throughput Screening Identifies Novel Inhibitors of *Escherichia coli* Dihydrofolate Reductase that are Competitive with Dihydrofolate. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2493–2496.

(20) Parker, C. McMaster University Data-Mining Competition: Computational Models on the Catwalk. *J. Biomol. Screening* **2005**, *10*, 647–648.

(21) Johnson, M. Specifying and Using Hierarchically-Organized Structural Categories. http://www.pannanugget.com/downloads.htm (accessed July 24, 2008).

(22) *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*; O'Neil, M. J., Heckelman, P. E., Koch, C. B., Roman, K. J., Eds.; Merck & Co, Inc.: Whitehouse Station, NJ, 2006.

(23) Behzad, M.; Chartrand, G.; Lesniak-Foster, L. *Graphs and Digraphs*; Wadsworth: Belmont, CA, 1979.

(24) Xu, Y.-j.; Johnson, M. A. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.

(25) Horowitz, E.; Sahni, S. *Fundamentals of Computer Algorithms*; Computer Science Press: Rockville, MD, 1978.

(26) *Meqi II, version 2.31*; Pannanugget Consulting: Kalamazoo, MI, 2008.

(27) *SciTegic Pipeline Pilot, version 6.0.2.0*; Accelrys, Inc.: San Diego, CA, 2007.

(28) Conover, W. J. *Practical Nonparametric Statistics*; John Wiley & Sons: New York, 1971.

(29) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.

(30) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. Four Association Coefficients for Relating Molecular Similarity Measures. *J. Chem. Inf. Model.* **1996**, *36*, 909–915.

(31) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **2007**, *47*, 325–336.

CI800250R