Data Shaving: A Focused Screening Approach

Suzanne K. Schreyer,*,† Christian N. Parker,‡ and Gerald M. Maggiora§

Chemical Computing Group, 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec H3A 2R7, Lead Discovery Center, Room 2605, Novartis Institutes for BioMedical Research Inc., 100 Technology Square, Cambridge Massachusetts 02139, and Department of Pharmacology & Toxicology, College of Pharmacy, University of Arizona, 1703 East Mabel, P.O. Box 210207, Tucson, Arizona 85721

Received July 24, 2003

The number of compounds available for evaluation as part of the drug discovery process continues to increase. These compounds may exist physically or be stored electronically allowing screening by either actual or virtual means. This growing number of compounds has generated an increasing need for effective strategies to direct screening efforts. Initial efforts toward this goal led to the development of methods to select diverse sets of compounds for screening, methods to cluster actives into related groups of compounds, and tools to select compounds similar to actives of interest for further screening. In this work we extend these earlier efforts to exploit information about inactive compounds to help make rational decisions about which sets of compounds to include as part of a continuing screening campaign, or as part of a focused follow-up effort. This method uses the information from inactive compounds to "shave" off or deprioritize compounds similar to inactives from further consideration. This methodology can be used in two ways: first, to provide a rational means of deciding when sufficient compounds containing certain structural features have been tested and second as a tool to enhance similarity searching around known actives. Similarity searching is improved by deprioritizing compounds predicted to be inactive, due to the presence of structural features associated with inactivity.

INTRODUCTION

High-throughput screening (HTS) has become a central part of finding new leads in the drug discovery process, and a number of successful examples has been reported.[1] However, as compound collections continue to grow, the cost of conducting a high-throughput-screening campaign increases (even with the current advances in assay miniaturization to reduce assay volumes, which reduces reagent and compound consumption).[2] In terms of scale, it has been reported that approximately 3.5 billion compounds have been virtually screened,[3] while over 5 million unique compounds are reported to be commercially available.[4] Thus, a "screen-them-all" approach is neither feasible nor economically practical, even for the largest of pharmaceutical companies.

This has led to numerous methods to try and improve the efficiency of the screening process. The objective of many of these methods is to classify compounds using structural descriptors to help organize and simplify data analysis. Methods to describe molecular structure have become widely accepted as the bases for the design of diverse, representative screening libraries.[5−9] Molecular descriptors can then be used to organize the results of screening by means of clustering techniques,[10] by the presence of predefined structural features,[11] or by identifying features that group compounds due to the presence of some minimum common substructure.[12]

Concurrently, there have been efforts to focus screening using a variety of tools. Where structural information about the target is known, efforts are being made to conduct virtual screening.[13] Where no information about the target structure is known, preliminary screening data can be used to generate models to predict the activity of untested compounds, thus improving the efficiency of subsequent screening. These methods include: design of experiment strategies,[14] binary quantitative structure−activity relationships (QSARs),[15] recursive partitioning,[16] and extensions of this method,[17] pharmacophore models,[18,19] and artificial intelligence techniques.[20] The goal of all of these methods is to generate models that identify structural features of compounds associated with active molecules. However, for data derived using HTS, the vast majority of compounds tested are inactive, so models describing features associated with activity may be based on as little as 0.01−10% of the screening data. While efforts have been made to describe the effect of biased data sets upon modeling, these have still shown that at least 20% of the compounds should be active in order to develop robust activity predictions.[21,22] However, methods are being introduced that try to overcome this constraint.[23] Further compounding this problem are three additional issues. First is the fact that sets of compounds chosen for screening are often chosen to be maximally diverse. As a result there are usually very few examples of structurally related active compounds around which to identify the molecular features important for activity. Second, screening data contain a high degree of uncertainty due to inherent noise and single point activity determination. The final problem confounding these attempts to model HTS data is the difficulty of knowing which set of molecular descriptors and clustering methods will be most suited to the screen under study.[24,25]

* Corresponding author phone: (514) 393-1055; fax: (514) 874-9538; e-mail: sschreyer@chemcomp.com.
† Chemical Computing Group.
‡ Novartis Institutes for Biomedical Research, Inc.
§ University of Arizona.

10.1021/ci030025s CCC: $27.50
Published on Web 01/23/2004

Molecular descriptors generally fall into either fragment-based or whole-molecule-based measures. Fragment-based descriptors use the presence of common substructures that may be predefined (e.g., Leadscope[26] or Molecular Equivalence Numbers[27] (meqnum)). The advantage of having precomputed descriptions of the compounds is that this facilitates computational analysis of very large data sets. Another possible benefit of such descriptors is that the predefined fragments can be chosen to represent accepted pharmacophoric features. Fragment-based descriptors can also be calculated from the data set under study to identify appropriate atom pair descriptors or maximum common substructures (Bioreason). These have the advantage of possibly being more relevant to the data set under study but are more computational intensive to generate. Additionally, fragment-based descriptors can be presented as categorical designations that may be arranged hierarchically (as is the case for LeadScope and meqnum) or can be presented as a binary fingerprint matrix of zeros and ones indicating the presence or absence of specific substructural fragments. Because the number of bits in the fingerprints is usually quite large, such representations of compounds may be simplified by reducing the number of descriptors through hashing.[28] Whole-molecule descriptors, which describe features such as molecular weight, calculated lipophilicity, or the number of rotatable bonds, are another type of descriptor. Yet, no matter how the descriptors are generated, they are used to assess the resulting chemistry space in order to find relationships between compound descriptors and the observed activities.

In this paper, we introduce a technique that recognizes these caveats and in fact tries to exploit them. First, inactive data are modeled to identify molecular features that are associated with inactive compounds. Such a model can then be used to deprioritize compounds predicted to be inactive from further testing. Second, using the data from the active compounds, substructural features associated with activity can be identified. In effect this process suggests structural hypotheses or potential pharmacophores to explain the observed activity. Third, by adding similarity searching around the known active molecules (using a different set of molecular descriptors and clustering methodology), one can identify additional actives and subsequently improve the identification of features required for activity. This similarity searching phase can also be improved by applying the rules found to identify inactive compounds to further filter these similar compounds to help improve the screening efficiency.

We call this method "data shaving" by analogy to the "gene-shaving" method developed by Hastie[29] et al. In the gene-shaving analysis of gene chip microarray data, a principal-component analysis is first performed and any genes showing little variance, hence containing little information, are then removed from the data set. Thus, only the results from genes that potentially contain more information are retained. Our method by analogy tries to retain information-rich structures and to disregard information from compounds giving either poor information or those that have already been adequately tested. This then allows screening efforts to be focused into new untested areas of chemistry space.

The result of using this method as part of a high-throughput screening campaign is that "stopping" rules can be generated as sufficient data are obtained about inactive molecules containing particular molecular features, which can form the basis for the decision not to screen additional compounds. This has the effect of directing the screening away from the sorts of compounds previously tested and toward novel structures, thus improving the screening efficiency.

MATERIALS AND METHODS

Data Sets. A series of three proprietary dissimilarity libraries and the NCI AIDS data set were used in the analyses. For the Pharmacia data sets, three libraries were designed for maximum diversity with minimal overlap of compounds between the three. These libraries were constructed at different times using three different perceptions of molecular diversity: The PB library (79 095 compounds) was generated using a maximum dissimilarity analysis,[30] the PC library (67 513 compounds) was generated using a fuzzy clustering method called Algorithm5,[31] and the PD library (74 772 compounds) was selected from commercial sources using the Diverse Solutions Software package. These compound libraries were then used in a whole cell assay with resulting activities of 9.6% for PB (7574 actives), 11.7% for PC (7502 actives), and 9.9% for PD (7415 actives). For analysis using data shaving, the entire collection of compounds was used.

For the NCI AIDS data set, three subsets were generated (SS1, SS2, SS3) from the 42 687 compounds. These almost equally sized subsets were generated using a random selection program (written in Matlab) into 10 429 compounds for SS1, 15 985 compounds for SS2, and 9812 compounds for SS3. Inactive-based rules were generated using data set SS1 and applied to SS2 and SS3. This will be described in greater detail in subsequent sections of the paper.

Meqnum Methodology. Xu and Johnson[32] have developed the concept of molecular equivalency indices (MEQI) to classify molecules according to their structural and topological features. Molecules are generally represented as chemical graphs, which provide the two-dimensional depiction of molecular structure familiar to chemists. In addition, hydrogen atoms are removed so that all chemical graphs used in this work are "hydrogen-suppressed". In this work MEQIs are generated by parsing chemical structures into their simple component parts, each component being labeled by a specific molecular equivalency number or meqnum. Meqnums are *codes* that are uniquely assigned to molecular structures or substructures. Each MEQI is composed of a list of meqnums and thus is also called a composite meqnum. The meqnums correspond to either the functional groups or to other features such as the ring systems of a particular molecule. Technically, each composite meqnum can be parsed into a particular individual substructural feature, designated as meqnum in this work, which refers to a particular generic functional group shared by many compounds. The meaning should be clear from the context. Meqnums can also contain other types of information such as the number of non-hydrogen atoms in a molecule, the number of atoms in the cyclic system of a molecule, and/or the number of rings in a molecule (vide infra). For example, the numbers of non-hydrogen atoms of benzene, naphthalene, and biphenyl are respectively 6, 10, and 12, which are identical to their cyclic system sizes, while their corresponding numbers of rings are 1, 2, and 2.

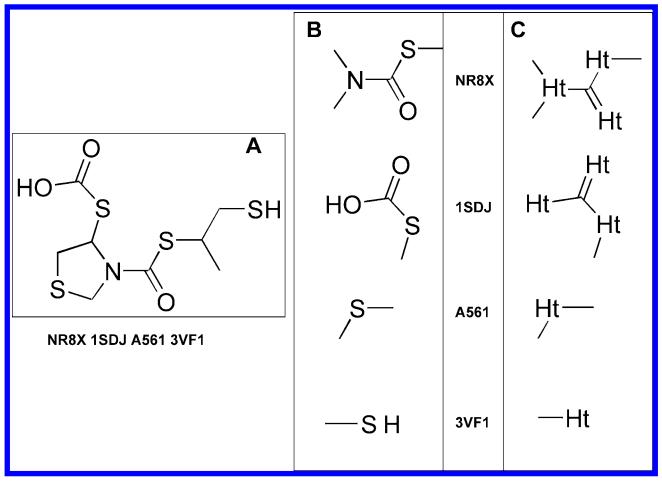
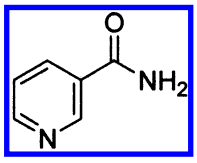


Figure 1. Illustration of the decomposition of a molecule into its α-augmented functional groups. (A) The composite meqnum associated with (B) the α-augmented functional groups, and (C) the coarse-grained functional group representation in which all heteroatoms are designated by the symbol "ht" and are thus treated as equivalent, and all halogens are designated by the symbol "hl".

If ring system meqnums are used, this method is very similar to the ring-based structural query system described by Nilakantan et al.[33] In contrast to most studies employing meqnums, where the focus is on various types of ring structures or cyclic systems, here we will focus on functional group meqnums that result from "decomposing" molecules into sets of nonoverlapping fragments. The algorithm used to do this is based upon partitioning the atoms in a molecule into sets of "spacers" and "nonspacers." The set of nonspacer atoms contains all non-carbon atoms as well as carbon atoms connected by bonds with multiple bond character (e.g., partial double, double, or aromatic) to a heteroatom. The remaining carbons constitute the set of spacer atoms. The fragments that remain after deleting all spacer atoms are called maximal functional groups, abbreviated here as functional groups. As an example, nicotinamide, has two functional groups, one



related to the nitrogen atom in its pyridine ring (C∼N∼C) and one related to the amide functionality (O═C−N). The functional group obtained by replacing the pyridine ring by a saturated piperdine ring is simply the single nitrogen atom.

Because the functional groups are nonoverlapping, atoms are not shared between them. Functional groups can be augmented by adding all spacer carbon atoms α to any of the nonspacer atoms. Such functional groups are called α-augmented functional groups. Because spacer carbons may be associated with more than one functional group, α-augmented functional groups may overlap with one another, but only through their "α-augmenting" carbon atoms. Figure 1 illustrates the decomposition of a molecule into its α-augmented functional groups using the method of Johnson and Xu.[34] The meqnum given below the first panel in the figure is a *composite* of the meqnums associated with the α-augmented functional groups depicted in the second panel. In a composite meqnum the individual meqnums are ordered by size, that is, the number of non-hydrogen atoms they contain. If a particular meqnum occurs multiple times within a molecule, it will also appear multiple times in the composite meqnum associated with that molecule. If two different meqnums of the same size are present, they are ordered lexicographically. A large set of molecules will contain many possible individual meqnums, although each molecule will contain relatively few.

Compound	Activity	GW84	Z5HA	TWH6	1VD2	Cycsys size	Ringsys cnt
Cmpd 1	I	1	0	0	0	9	1
Cmpd 2	A	0	1	1	1	19	2
Cmpd 3	I	0	0	0	1	27	4

Figure 2. Flow chart displaying the (A) matrix depicting the bit vector of 1's and 0's, where a "1" corresponds to the occurrence of a specific primitive meqnum and a "0" corresponds to its absence. Additional columns in the binary table are associated with the number of atoms in the cyclic system, the number of rings, and whether the compound is active (A) or inactive (I).

For a library of compounds, the composite meqnums are calculated, using Meqi.[35] In the present work, we will use a coarse-grained functional group representation in which all heteroatoms are designated by the symbol "ht" and are thus treated as equivalent, as depicted in the third panel of Figure 1. Likewise all halogens are taken to be equivalent and are designated by the symbol "hl." The coarse-grained fragments depicted in the figure provide a lower resolution representation of the structural information in molecules. However, as will be shown in the subsequent analysis, it nevertheless provides sufficient resolution to be of use in the analysis of low-resolution screening data. This representation leads to what Xu and Johnson call aryl−hetero functional-group meqnums (AHFG) and will be used throughout the present analysis.

Structural information encoded in the composite meqnum associated with each molecule is then transcribed into a binary table in which each column corresponds to an instance of a specific meqnum and each row corresponds to a given molecule. Thus, each molecule can be described by a bit vector of 1's and 0's, where a "1" corresponds to the occurrence of a specific meqnum and a "0" corresponds to its absence. For simplicity, specific accounting of multiple occurrences is not considered in this work. Additional columns in the binary table are associated with the number of atoms in the cyclic system, the number of rings, and whether the compound is active (A) or inactive (I), which is depicted in Figure 2.

The table is then transformed into the new table, called a *meqnum fingerprint matrix*, depicted in Figure 3, which shows the number of active and inactive molecules in the data set for which a given AHFG meqnum occurs at least once. This information is then combined to give the percent activity for each meqnum. The meqnum fingerprint matrix plays a key role in the analysis presented here.

Data Shaving. (a) Rationale. Usually the majority of diverse compounds screened against any particular biological activity are inactive. This should be intuitively obvious since if every biological system were to respond to every chemical in its environment, selectivity and specificity for biological ligands would be impossible. The concept of molecular similarity predicts that similar compounds will have similar biological activity.[36] So, by identifying compound classes that are predominantly inactive, one can make a decision to

Meqnum	Actives	Inactives	Total	%I	%A
1VD2	1373	6032	7405	81.46	18.54
FD7C	415	5583	5998	93.08	6.92
C08B	766	4736	5502	86.08	13.92
Z5HA	433	4177	4610	90.61	9.39
JMEQ	1264	2138	3402	62.85	37.15
U2AN	102	3212	3314	96.92	3.08
R815	172	2892	3064	94.39	5.61
TWH6	261	2215	2476	89.46	10.54
VZPK	77	1249	1326	94.19	5.81
4AN0	100	1011	1111	91.00	9.00
SY42	110	864	974	88.71	11.29
2PVV	63	905	968	93.49	6.51
7GH4	280	624	904	69.03	30.97
3E3P	78	817	895	91.28	8.72
C4R2	129	711	840	84.64	15.36
Z3QA	60	754	814	92.63	7.37
UTN7	20	618	638	96.87	3.13
U7SZ	271	336	607	55.35	44.65
XWQC	46	509	555	91.71	8.29
K6WT	76	458	534	85.77	14.23
3HGZ	43	442	485	91.13	8.87
3UG0	60	307	367	83.65	16.35
SVYH	20	332	352	94.32	5.68
L3M5	14	316	330	95.76	4.24
T2L8	30	229	259	88.42	11.58
6R84	12	209	221	94.57	5.43

Figure 3. Table 1 result from the transformation of the binary matrix in Figure 2. This shows the number of active and inactive molecules in the data set in which a given meqnum occurs at least once. This information is then combined to give the percent activity for each meqnum.

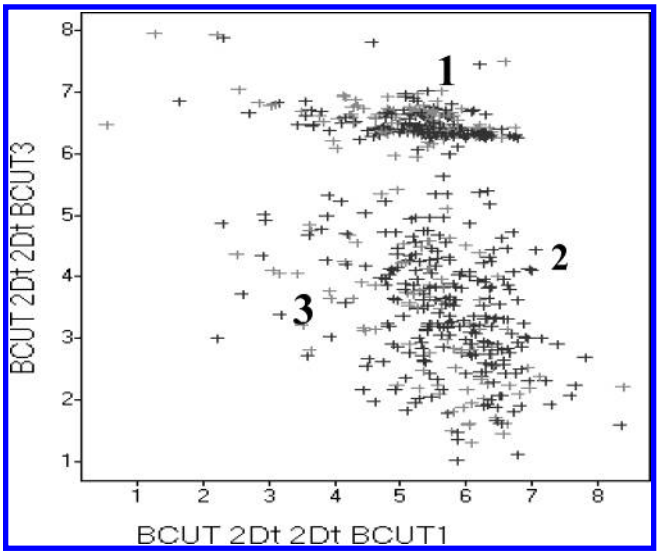


Figure 4. Representative BCUT space of a diverse library used to illustrate the three regions. In region 1, this area has been screened adequately given the density of the compounds; hence a stopping rule can be formulated. Region 2 contains mostly inactive compounds, with little advantage in continued searching of this region for actives, as the cost in resources may outweigh any advantage gained. Region 3 contains a mix of active and inactive compounds that can be mined for information content.

stop screening such types of compounds. For example, if one were to screen a library of compounds for inhibition of a matrix metalloproteinase and 100 sterol compounds were found to be inactive, it would make sense to deprioritize screening of additional sterols. A similar argument could be made for active compounds, if the goal of screening is to identify as many different classes of active compounds as possible. So when screening for matrix metalloproteases, once hydroxamates had been identified as a class of active compounds, additional hydroxamates would be deprioritized so as to facilitate discovery of new classes of inhibitors. However, since the majority of compounds tested are inactive, it is usually more difficult to have tested sufficient active compounds to be able to make such a rational decision, especially if libraries of compounds chosen for their dissimilarity are being screened. In essence this describes the strategy outlined here in which structural features of compounds that are associated with inactivity are identified and additional compounds with these features are deprioritized from further screening.

(b) Procedure. As described previously, once compounds are classified according to their meqnums, the relationship between meqnums and activity level is determined using a heuristic approach based on frequency distributions. This is a preliminary analysis for several reasons. Namely, primary screening data are often based on a single result at a fixed concentration, with considerable variability and noise associated with the inhibition values. As a result the standard deviation associated with the inhibition values tends to be highest around the active/inactive interface. Thus, the interface between actives and inactives tends to be "fuzzy" and dependent upon the way in which the activity cutoff is determined.

Further, the most common meqnums often correspond to structural features that may not provide clear differences between active and inactive compounds. For example, the most common meqnum (in the data sets presented here) is usually C08B, which codes for the presence of a halogen; halogens are usually found equally in active and inactive compounds, so this meqnum does not provide useful information. Thus, careful examination of the type of meqnum, in the most frequent distributions, is still required to determine if a structural feature is really of interest.

As mentioned earlier it has been estimated that data sets with at least 20% of the compounds being active are needed in order to generate robust models relating activity to structure.[37] Our analysis of high-throughput-screening data, which contain a much smaller percentage of active compounds, supports this observation. Because of the small number of dissimilar actives identified from each screening library, it is possible to miss structural features that are present in the active compounds, depending on the data set used.

Method. The actual mechanics of data shaving are relatively simple, utilizing a series of heuristic rules. By using the preponderance of information on structural features associated with inactive compounds, it would be rational to use this information to identify (a) regions of chemistry space that have been sufficiently screened, and are not likely to yield further new information; (b) regions of chemistry space that are predominately inactive, so further screening in this area is relatively inefficient; and (c) regions of chemistry space that are of interest because they contain a high proportion of actives. In this sense, data shaving serves to prioritize the search for biological activity in chemistry space. An example of this approach is shown schematically using BCUT[38] descriptors in Figure 4, where the three regions are identified. Region 1 has been screened adequately given the density of the compounds; hence, a stopping rule can be formulated. Region 2 contains mostly inactive compounds Thus, little advantage is gained in continued searching of this region for actives, as the cost in resources may outweigh

any advantage gained. Region 3 contains a mix of active and inactive compounds that can be mined for further information. By focusing on region 3 and by identifying and removing the structural features associated with inactivity from region 2, the structural features linked to the active compounds may be easier to spot. This strategy is similar to similarity searching using activity-weighted fragment descriptors[39] but instead deprioritizes substructural fragments associated with inactivity as greater statistical confidence can be made in assigning inactivity. This results in a more efficient search, as well as enhancing the overall activity of the region, as prioritization will result in a focused search in the relevant areas of chemistry space.

The reliance on expert opinion and manipulation are a necessary result of the nature of screening data. However, in future papers we will show how these approaches can be rationalized using machine learning approaches. A more detailed description of the procedure is as follows:

(A) A histogram showing the active and inactive compound ratios associated with each of the composite meqnums is generated using one data set. Each composite meqnum codes for a number of specific substructural features. These composite meqnums are further divided into their individual meqnums by a Perl script written by Chad Storer at Pharmacia.

(B) After being parsed into meqnums, a descriptor matrix is generated where each compound is represented by a 0 or 1 for the absence or presence of the structural feature, respectively. The matrix also includes information on the cyclic system size, the ring count, and the activity (Figure 2).

(C) Only meqnums that are represented by 100 or more compounds are considered for further analysis. This identifies classes of less than 100 compounds that have not been sufficiently sampled to make a rational decision whether to continue screening or to stop testing such compounds due to either their inactivity or their confirmed activity.

(D) Meqnums that contain predominantly inactive compounds are then identified. This is usually achieved by removal of meqnums that do not contain significant numbers of actives. In the example given in Figure 3, meqnums that do not contain between 20 and 30% actives out of all compounds with that meqnum were not considered further.

(E) The descriptions of ring size and ring count are also used to identify sets of compounds that are mostly inactive.

(F) This procedure has identified a series of meqnums that represent predominantly inactive compounds. These meqnums can then be used as rules ("stopping rules") and applied to subsequent data sets to deprioritize such compounds from further screening.

(G) Removal of compounds in meqnums that are predominantly inactive or which have not been sufficiently sampled results in a data set enriched with active compounds. This also results in a focused chemistry space which makes it easier to identify those structural features relevant to activity.

(H) As a final step it is possible to use these "stopping rules" to help search for additional active compounds using a technique we term "similarity focusing". Similarity focusing is conducted using one or more active compounds as reference compounds for similarity searching based upon ISIS or Cousin fingerprint[40] and the Tanimoto similarity index. In this way compounds, similar in structure to the active "seed" compounds, that have not yet been tested are identified. Because these compounds were identified using a different set of molecular descriptors, many will contain meqnums that have been associated with inactivity. These compounds are then deprioritized from testing, helping to increase the screening efficiency.

Although the preceding discussion has been applied to Pharmacia data sets, the same procedure was performed on the NCI AIDS data subsets, with the added caveat that in terms of identifying active compounds both the active (CA) and the moderately active (CM) compounds were classified as active.

RESULTS AND DISCUSSION

Pharmacia Data Sets. Based on the procedure outlined above, data shaving was first applied to the PD data set of 74 772 compounds. This data set was used to generate the appropriate stopping rules that were applied to the other two data sets (PC, PB). The composite meqnums associated with each compound were parsed into a meqnum fingerprint matrix (vide supra). Frequency distributions of each meqnum were generated for active and inactive compounds in order to identify the most prevalent meqnum associated with activity. From this table the percent inactive (and percent activity) associated with the meqnum was also generated. To extract the significant meqnums principally associated with inactivity, certain filtering rules were applied. First, only meqnums associated with a sufficient number of active or inactive compounds are retained for further analysis. It was determined that there should be a total of at least 100 compounds associated with a given meqnum for the meqnum to be retained as statistically relevant, which is consistent with the findings of Nilakantan et al.[41] This reduced the initial 648 meqnums generated from the PD data set to 54. Meqnums associated equally with both active and inactive compounds are removed, as these give little information since they usually describe small and less significant substructural features. An example of this is the removal from further consideration of C08B, which represents a halogen, and 1VD2, which represents a single heteroatom. Both substructural features are present equally in active and inactive compounds and so do not provide any significant information on their relationship to activity. Then, using the frequency distribution, the percentage inactivity associated with each meqnum was ranked. Due to the prevalence of the inactives in this data set, only meqnums containing greater than 90% inactivity were considered. This reduces the number of meqnums to 39 from the previous 54 meqnums. The choice of meqnums significantly associated with *inactivity* in the PD data set is then taken to be the set {U2AN, Z5HA, R815, and FD7C}.

The final filtering was done by examining the histogram of the ring count (number of ring systems present), and the cyclic system count (number of atoms involved in a cyclic system). For example, both benzene and naphthalene will return a ring count of 1 but can be differentiated by their cyclic system count of 6 and 10 atoms, respectively. This will serve to remove any obviously undesirable compounds such as very large structures, compounds with no rings (not

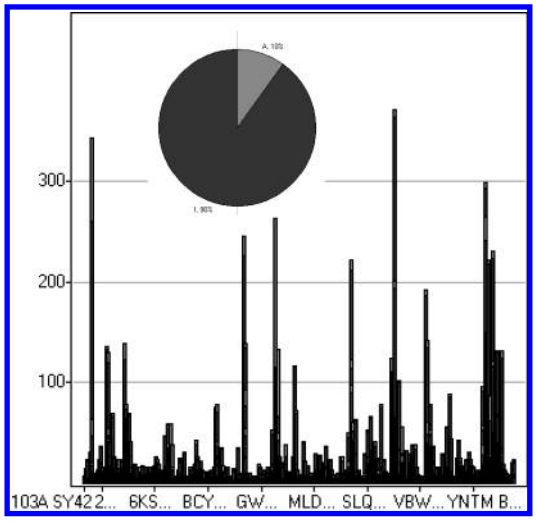


Figure 5. Histogram showing the resultant distribution of compounds in the PD data set used to generate the frequency distribution and identify features common to inactives. Compounds are depicted in their respective composite meqnum bins. Initial activity for this data set is 10%.

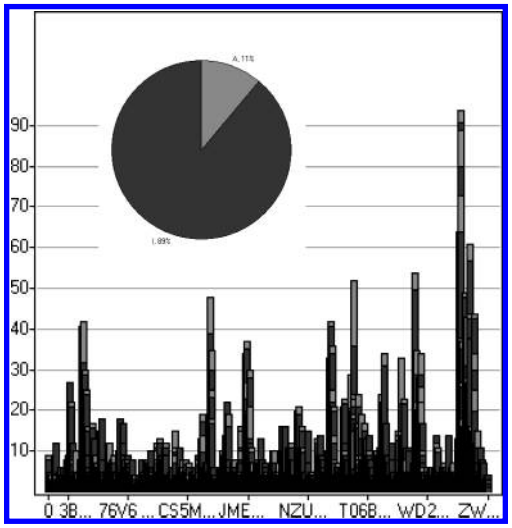


Figure 6. Histogram showing the resultant distribution of compounds (67 513) in the PC data set according to their respective composite meqnums. Initial activity for this data set is 11%.

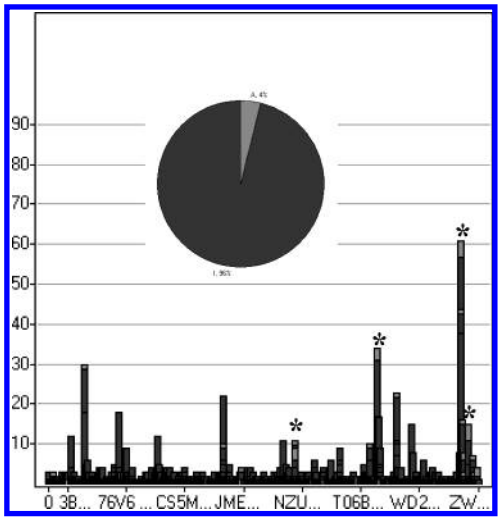


Figure 7. Compounds containing the meqnum {U2AN}. These are predominantly inactive in the PC data set with the overall activity of 4%. Those composite meqnums with activity rates greater than 20% (labeled with asterisks) and are retained.

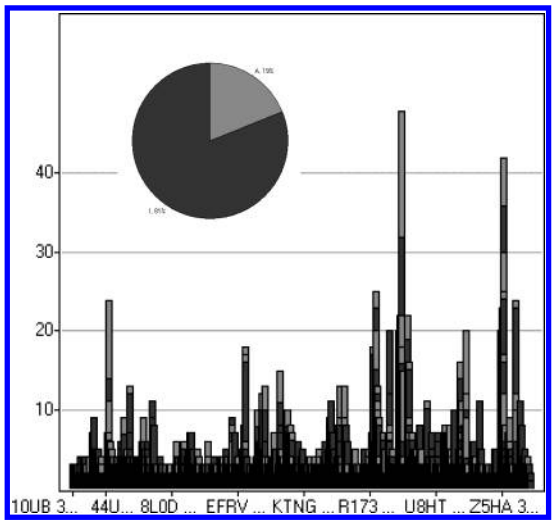


Figure 8. Histogram showing the resultant distribution of compounds (67 513) in the PC data set according to their respective composite meqnums, after data shaving. Activity for this data set is now 19%.

generally considered druglike), and systems with large fused rings, etc. From the PD data set any compounds that contain no rings or greater than four rings were also filtered out. This filtering does not change the significant meqnums, but it does reduce the number of compounds in the data sets. The meqnums that contain the most compounds are given in Figure 3, and it can be clearly seen from the frequency distribution that certain AHFG meqnums are predominately associated with the inactive compounds, namely, the set {U2AN, Z5HA, R815, and FD7C}.

To evaluate the effect of applying these data shaving rules to other data sets, an example of the systematic application of data shaving is shown in Figures 5−10. In this example the rules generated from the above inactive compounds are applied to two other compound data sets (PC and PB) chosen to be dissimilar. Since these compound libraries cover the same space, the structural features associated with one of the data sets are also present in the other data sets. Thus, substructural features encoded by their meqnums and associated with inactive compounds identified from the PD data set (Figure 5) are also prevalent among the inactives in the other two data sets.

The PC data set (Figure 6) contains 67 513 compounds of which 7502 are active (11%). However, by applying the stopping rules derived from the PD data set (i.e. not testing compounds with either no rings or greater than four rings or compounds containing the following set of meqnums {U2AN, Z5HA, R815, FD7C}) to select compounds from the PC dataset for testing would have resulted in a hit rate of 19% actives, as shown in Figure 8.

In addition to the potential enhanced activity, many of the meqnums identified in this sequential screening strategy allow new stopping rules to be identified. For example, in subsequent follow-up screening, it is assumed that compounds containing meqnums associated with inactive structural features will not be screened further, as little further information is generated from them. Instead, the focus of follow-up screening will shift to both untested and undersampled regions of chemistry space. Focused similarity searches around the active compounds initially identified

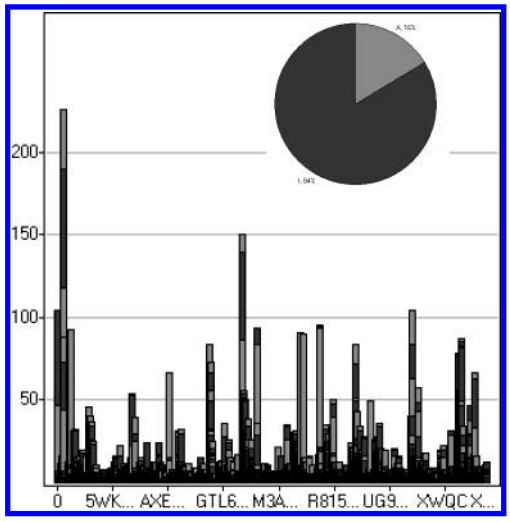


Figure 9. Histogram shows the composite meqnums for compounds generated from a search in the Pharmacia database for compounds containing at least 67% similarity (Cousin keys and Tanimoto similarity) to the active compounds in the PD data set. From this similarity-focused search 16% activity is achieved.

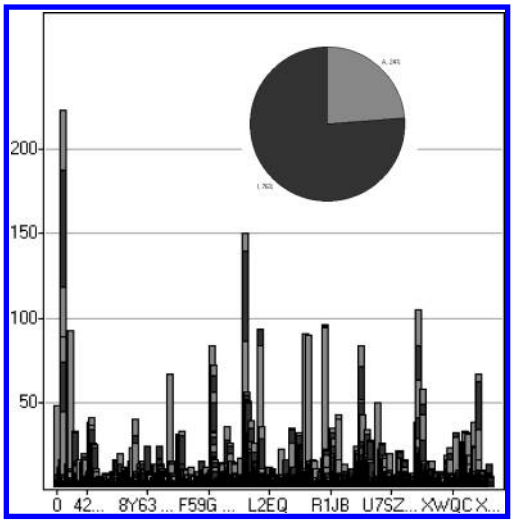


Figure 10. Same data set from Figure 11, after data shaving to remove inactive meqnum found from the initial PD data set, as described in the paper. Activity is increased to 24%.

complement this shift and lead to enhanced screening efficiency. This can, in conjunction with the deprioritization of compounds containing meqnums associated with inactivity, lead to dramatic increases in screening efficiency. Taking the actives from the PD data set, and using a similarity search about these actives (67% similarity cutoff), resulted in the data set shown in Figure 9, which has an activity of 16%. Applying the same rules generated previously, the new focused space now contains 24% active compounds (Figure 10). In this case, compounds not previously tested in the assay can be readily identified and sent for follow-up screening. This has the added advantage that the suggested compounds are extracted from a *focused* chemistry space, which has either not been tested enough or contains a significant portion of structural features associated with activity, resulting in an improved approach to screening.

Alternatively, this strategy can be applied retrospectively to help focus attention onto regions of chemistry space where structural features of possible importance can be more readily identified. Applying these stopping rules to the PC data set so that excess or redundant information is trimmed away results in a prioritization of the data. Figure 7 shows an example of this with the PC data set. The compounds containing the meqnum U2AN are predominantly inactive, as the overall activity for this subset is 4%. However, further examination of the compounds containing this meqnum reveals that some combinations of features with this feature have activity rates greater than 20%. The compounds containing these combinations of structural features are labeled with asterisks in Figure 7. Thus, this retrospective analysis has helped identify combinations of structural features that may represent classes of compounds for further analysis because they may represent false positives or new classes of interest. This example highlights the need for expert evaluation of the screening data during this process to extract as much information as possible.

The rules generated from the PD data set were also applied to the PB data set, which has 79 095 compounds of which 7574 are active (9.6%). After data shaving was carried out, the resulting activity was 16.9% without follow-up by similarity-focused searching and subsequent data shaving of this space.

Active-Based Search. As mentioned, the rules generated for data shaving are based on the inactive compounds. A similar shaving procedure was done using rules based on the active compounds in order to show the difference in information that results when this much smaller group of compounds is used. Using the PD data set and generating a similar frequency distribution as before, but considering only actives, rules which identify meqnums associated with actives were found. This was more difficult as heuristic examination revealed an overlap between active and inactive compounds, as relative amounts of the inactives were much greater than the active compounds. However, after carefully choosing meqnums thought to associate with actives to a greater percent than with inactives {**TWH6**, **K6WT**, 1VD2, JMEQ, **3HGZ**} (where the bold meqnums indicate those found only in that data set), these rules were applied to the subsequent data sets PC and PB. No significant increase was found for the PC data set (from 11 to 11.8%), while a slight increase was seen in the activity of the PB data set (from 9.6 to 12.6%). These results clearly show the advantage of basing structural rules on the larger number of inactive compounds, as these cover the chemistry space of all the data sets and, due to their greater numbers, provide a more robust model.

Repeating this analysis on the PC and PB data sets identified the following structural features as associated with activity: (1VD2 2PVV SY42 FD7C **C4R2** JMEQ **R815**) for PC and (2PVV SY42 FD7C **LM46** 1VD2 JMEQ **N5E**) for PB, where the bold meqnums again indicate those found only in that data set. These differences again highlight how the low hit rate confounds the identification of general rules for identifying active compounds compared to the success found identifying consistent rules to identify inactive compounds.

NCI Data Set. The data shaving procedure was also applied to the NCI AIDS data set of 42 687 compounds. In this case, the categorical target was the active (CA + CM) or inactive (CI) compounds, while the activity is given by IC50 values (converted to log(1/ IC50)). The variables are the MEQI determined for each subset, where each subset was defined previously. The rules generated for the inactive

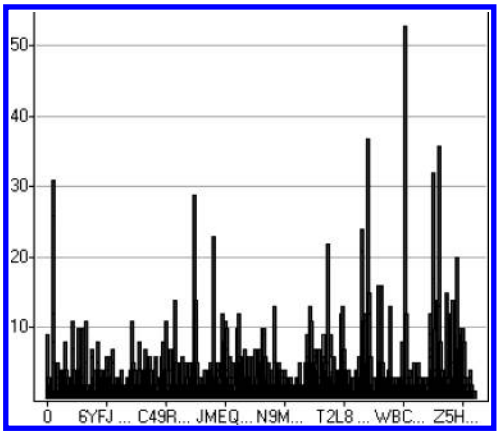


Figure 11. Examination of the composite meqnums that result from SS1. These reveal predominately inactive compounds, with few actives (2.9%).

compounds were based on SS2, and applied to SS1 and SS3—although the same meqnums are found to be important in all subsets, so the subsets used in rule generation are interchangeable.

It should be noted that the NCI AIDS data set is technically not screening data in several key ways. For instance, the NCI data uses IC50 values rather than single-point percentage inhibition values, as is generally the case for screening data. Additionally, the compounds present in the NCI data set were not selected using a dissimilarity-based procedure, with the result that the NCI data set has a number of identical compounds that were submitted several times for testing or as part of a congenic series of compounds. In addition the data set contains many compounds with either no rings or very large rings as this compound set contains natural products that are often excluded from pharmaceutical company screening collections. The result is that this set of compounds while often used to illustrate data mining applications is very different from the data sets often experienced in a drug discovery setting.

Note that the results from the Pharmacia data set were not identified as a specific section so it is not done here either. Subsets from the 42 687 compound data (NCI AIDS) were generated from a randomized split into three subsets (SS1, SS2, SS3). Focusing on the rules generated by SS2 (the largest subset), the most significant meqnums for the inactives were found to be {U2AN, 1VD2, 4AN0, FD7C, R815, Z5HA}. This is a greater number of meqnums for a much smaller data set than that obtained from the Pharmacia data sets, and part of the explanation for this lays in the nature of the NCI AIDS data set. Examination of the meqnums that result from SS1 (Figure 11) reveals an overwhelming amount of inactive compounds, with few actives seen. In fact, most of the actives exist as singletons—or single compounds in a meqnum bin. The most prevalent meqnum bins contain approximately 97% inactives; hence, the information from the inactives tends to represent the only source of significant information about this data set. Thus, the application of data shaving should be of value to remove redundant information, and prioritize the significant structural features. In fact, once data shaving is performed, and excessive inactive information is removed, the structural features important to active compounds are easier to spot, as shown schematically in Figure 12. This figure results after data shaving of similarity-focused SS1, with the untested compounds not shown in the figure for clarity. As shown, certain structural groups now are more clearly identified.

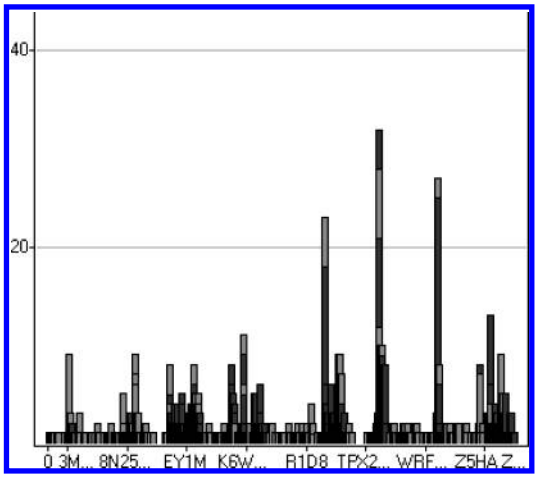


Figure 12. Results after data shaving. Excessive inactive information is removed, and active structural features are easier to spot. This figure results after data shaving of similarity-focused SS1, with the untested compounds not shown in the figure for clarity. As shown, certain structural groups now are more clearly identified.

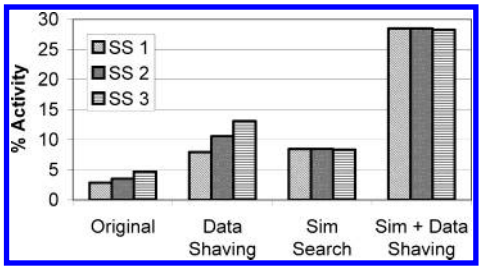


Figure 13. Using the rules generated by SS2, final results for the NCI data sets are shown. A 2-fold increase in activity is observed from the original data set to the first application of data shaving. Similarity searches with the actives in the original data set, SS2, result in approximately the same, or smaller, increase in activity than observed when data shaving is applied.

Since the process of data shaving has been discussed in depth for the Pharmacia data sets, only the final results for the NCI data sets are shown (Figure 13). A 2-fold increase in activity is observed from the original data set to the first application of data shaving. Similarity searches with the actives in the original data set, SS2, results in approximately the same, or smaller, increase in activity than that observed when data shaving is applied. However, application of data shaving to the compounds identified by similarity searches around actives results in a much greater increase in the percentage of actives found. As noted previously, the main advantage of data shaving is its ability to focus the search for compounds for further screening on areas of chemistry space that offer enhanced information content or to shift the search from areas of chemistry space that have already been adequately tested and offer the potential for little further significant information. Thus, data shaving offers preliminary stopping rules combined with a prioritization on structural features that can offer information to enhance activity.

CONCLUSIONS

While many previous analyses of screening data have focused on active compounds, it is generally the case that

active compounds form only a small minority of the compounds tested in any given screen. The majority of information garnered from a screen is inactive data, often overwhelmingly so. Due to the large variability and noise present in screening data, focusing on the small number of active compounds actually further diminishes the confidence that should be placed in the information. We have utilized the greater amount of information available from inactive compounds to develop a data analysis strategy that generates stopping rules that indicate when areas of chemistry space have been adequately tested. This strategy has the effect of focusing screening on those areas of chemistry space that are either undersampled or contain more significant structural information, for enhanced follow-up screening.

By incorporating data shaving of the inactives as part of an overall screening strategy, we have observed dramatic improvements in screening efficiency, the development of effective stopping rules, and the improvement of hit follow-up, even allowing preliminary SAR hypothesis generation. While the strategy described here is based on accepted methods of data visualization, such as the generation of substructural features from compounds using a graph-theoretic approach, and frequency distributions of active and inactive compounds according to their structural features, this method still employs a predominantly heuristic approach. For example, examination of the nature of important substructural features, generation of the inactive-based rules, and the data shaving process require human judgment.

This heuristic approach forces the future direction of this method into more stochastic methods. An information-theoretic approach will be discussed in a follow-up paper, which will introduce the generation of the inactive-based rules based on combinations of recursive partitioning and information theoretic procedures. Other ways in which this strategy can be applied include its application using other types of molecular descriptors (such as BCUTS or MOE descriptors) to identify areas of the chemistry space associated with inactivity rather than activity against specific targets. Thus, this is a more rigorous data-analysis approach for defining the rules for data shaving than the intuitive ad hoc approach described here. However, the present approach was used in this preliminary paper in order to introduce the concept of data shaving and to lead the reader through the steps and motivations for this approach.

In effect this paper also serves to highlight several weaknesses of the current approach, which will be addressed by the more stochastic approach. The weaknesses include the possibility of missing relevant active compounds in highly inactive regions of space and the lack of formal and definitive rules as to the number of compounds that need to be screened to define a stopping rule. In fact, the rules are based upon the data sets being used and so may need to be reevaluated as the size of the data sets change. The cutoff value for the number of compounds (100−150) needed to adequately support the choice of a particular meqnum is based more on a rule-of-thumb derived from personal experience but is also supported by the work of Nilakantan et al.[35] The examples presented here are based on a retrospective analysis of previously screened data sets, not descriptions of the method applied prospectively. This issue will also be addressed in our following papers.

Overall, we have presented a new method that utilizes inactive-based rule generation to improve screening efficiency through a combination of several factors. These include the generation of stopping rules to define when a portion of chemistry space has been sampled adequately, the deprioritization of regions of chemistry space that contribute little significant information, and the focusing of further screening onto regions of chemistry space that are more likely to yield structurally significant or structurally new compounds.

ACKNOWLEDGMENT

The authors would like to thank Dr. Mark Johnson for his helpful comments, which significantly clarified a number of points in the manuscript. The authors also wish to thank Chad Storer (Pharmacia) for the Perl script.

REFERENCES AND NOTES

(1) Fernandes, P. B. Moving into the Third Millennium after a Century of Screening. In *Handbook of Drug Screening, Drugs and Pharmaceutical Sciences*; Seethala, R., Fernandes, P. B., Eds.; Dekker: New York, 2001; Vol. 114, pp 1−4.
(2) Sundberg, S. A. High-Throughput and Ultra-high-Throughput Screening: Solution- and Cell-Based Approaches. *Curr. Opin. Biotechnol.* **2000**, *11*, 47−53.
(3) Davies, E. K.; Glick, M.; Harrison, K. N.; Richards, W. G. Pattern Recognition and Massively Distributed Computing. *J. Comput. Chem.* **2002**, *23* (16), 1544−1550.
(4) ChemNavigator: http://www.chemnavigator.com/cnc/chemInfo/iResearchLibrary.pdf.
(5) Lajiness, M. S. Dissimilarity-Based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.
(6) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.
(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(8) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738−746.
(9) Villar, H. O.; Koehler, R. T. Comments on the Design of Chemical Libraries for Screening. *Mol. Diversity* **2000**, *5*, 13−24.
(10) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest Neighbor and Cluster Analysis in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.
(11) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.
(12) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-like Trees. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1069−1079.
(13) Toledo-Sherman, L. M.; Chen, D. Q. High-Throughput Virtual Screening for Drug Discovery in Parallel. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 414−421.
(14) Andersson, P. M.; Linusson, A.; Wold, S.; Sjostrom, M.; Lundstedt, T.; Norden, B. Design of Small Molecule Libraries for Lead Exploration. In *Molecular Diversity in Drug Design*; Dean, P. M., Lewis, R. A., Eds.; Kluwer: Dordrecht, The Netherlands, 1999; pp 197−220.
(15) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary Quantitative Structure−Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(16) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(17) Izrailev, S.; Agrafiotis, D. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176−180.
(18) Hopfinger, A. J.; Duca, J. S. Extraction of Pharmacophore Information from High-Throughput Screens. *Curr. Opin. Biotechnol*. **2000**, *11*, 97−103

(19) Hecker, E. A.; Duraiswami, C.; Andrea, T. A.; Diller, D. J. Use of Catalyst Pharmacophore Models for Screening of Large Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1204−1211.
(20) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies: Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc*. **1984**, *106*, 7315−7321.
(21) Rosenkranz, H. S.; Cunningham, A. R. SAR Modeling of Unbalanced Data Sets. *SAR QSAR Environ. Res.* **2001**, *12*, 267−274.
(22) Thampatty, P.; Rosenkranz, H. S. SAR Modeling: Effect of Experimental Ambiguity. *Comb. Chem. High Throughput Screening* **2003**, *6*, 161−166.
(23) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Noisy High Throughput Screening Data Using a Naïve Bays Classifier. *J. Biomol. Screening*, in press.
(24) Gedeck, P.; Willett, P. Visual and Computational Analysis of Structure−Activity Relationships in High-Throughput Screening Data. *Curr. Opin. Chem. Biol*. **2001**, *5*, 389−395.
(25) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 28−35.
(26) Roberts, G.; Myatt, G. J.; Johnson W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.
(27) Xu, Y.-J.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 181−185.
(28) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics;* Kluwer: Dordrecht, The Neatherlands, 2003.
(29) Hastie, T.; Tibshirani, R.; Eisen, M. B.; Alizadeh, A.; Levy, R.; Staudt, L.; Chang, W. C.; Botstein, D.; Brown, P. "Gene Shaving" as a Model for Identifying Distinct Sets of Genes with Similar Expression Patterns. *Genome Biol.* **2001**, *1*, 1−21.
(30) Lajiness, M. Molecular Similarity Based Methods for Selecting Compounds for Screening. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science: New York, 1990; pp 299−316.
(31) Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; Spangler, D. P. Algorith5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1195−1204.
(32) Xu, Y.-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.
(33) Nilakantan, R.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. A Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic, *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 65−68.
(34) Johnson, M.; Xu, Y.-J. Structural Browsing Indices as High-Throughput SAR Analysis Tools. In *Chemical Data Analysis in the Large*; Hicks, M., Ed.: Logos Verlag: Berlin, 2000; pp 67−80.
(35) Software program available at: http://www.pannanugget.com/index.htm.
(36) Maggiora, G. M.; Johnson, M. A. Introduction to Similarity in Chemistry. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A.; Maggiora, G. M., Eds.; Wiley Interscience: New York, 1990; p 1.
(37) Brown, N.; Willett, P.; Wilton, D. J. Generation and Display of Activity-Weighted Chemical Hyperstructures. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 288−297.
(38) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 339−353.
(39) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1218−25.
(40) Hagadone, T. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 515−521.
(41) Nilakantan, R.; Immermann, F..; Haraki, K. A Novel Approach to Combinatorial Library Design. *Comb. Chem. High Throughput Screening* **2002**, *5*, 105−110.

CI030025S