Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery

David T. Stanton, Timothy W. Morris, Siddhartha Roychoudhury, and Christian N. Parker*

Procter & Gamble Pharmaceuticals Health Care Research Center, 8700 Mason-Montgomery Road, Mason, Ohio 45040-9462

Received June 6, 1998

High throughput screening (HTS) programs based on diverse collections of compounds can rapidly identify leads for potential drug candidates. In cases where the compound collection is truly diverse, one may only identify a few compounds of interest. However, where a large number of hits are identified, it becomes necessary to examine the structures to determine the true number of compound classes involved so that follow-up studies may be conducted as efficiently as possible. In this case, cluster analysis is applied to determine the structural relationship among HTS hits. To efficiently expand around the region of the hit (or a class of hits) in chemical space, we have applied nearest neighbors analysis[1] to select additional compounds from collections of a large number of commercial vendors, achieving an average hit rate in excess of 15%. Applying these techniques in a number of different cases, we obtained results that are useful for subsequent investigations of hits from HTS and other relevant molecular structures from the literature.

INTRODUCTION

The objective of any screening program is the identification of suitable hits to fuel an efficient drug discovery program. This paper will discuss how measures of molecular similarity have been used during many of the early stages of the drug discovery process to (a) improve the efficiency of screening by evaluating the similarity of hits identified, in order to determine the actual number of different classes of hits that have been identified, thus saving time and money by following up on selected representatives from each class of HTS hits, (b) aid in the identification of additional potentially active compounds, and (c) develop a more complete picture of the chemistry space available around any particular hit. This in turn will also provide a preliminary understanding of the structure-activity relationship a set of compounds has for a given target. The overall goal is to produce sufficient data for each class of hits so that one can make intelligent choices regarding which of these classes should be considered for additional follow-up work. The objective is not only to obtain hits but also to provide relevant SAR information to guide further lead expansion.

Fortunately, a number of techniques that facilitate these approaches are included as part of the computational tools available for molecular diversity analysis. One requirement of these tools that is critical to these processes is that they employ accurate measures of molecular structure. A related requirement is the ability to evaluate these molecular descriptors in order to select those that produce a multidimensional chemistry space that is general enough to be used to study large and diverse sets of chemical structures, while also providing the ability to perceive subtle differences between highly similar structures. With these requirements met, the next critical need is the ability to rapidly evaluate the large numbers of databases ($N > 15$) of commercially available compounds that include a very large number of compounds (total $>$ 500 000). Once this functionality is available, one can perform a number of different types of analyses that have been found to improve the efficiency and effectiveness of our new drug discovery program. This paper outlines a number of examples of how these molecular diversity related techniques have been successfully applied.

EXPERIMENTAL SECTION

General. The molecular structure descriptors used for all portions of this work were taken from the set of BCUT metrics calculated using the DiverseSolutions software package (Version 1.2.1) obtained from Prof. R. S. Pearlman.[1] The DiverseSolutions package was run using a Silicon Graphics Indigo-II workstation (R10000 processor), with Extreme Graphics, running under the IRIX operating system (version 6.2).

Cluster Analysis. Hierarchical cluster analysis was carried out using Minitab for Windows[2] and using a set of six BCUT metrics[3] which form a particular chemistry space.[4] All the BCUT values were autoscaled (mean-centered and variance normalized) before proceeding with clustering. The squared Euclidean distance was used as the similarity metric, and complete linkage[5] was used as the cluster-fusion method. The dendrograms resulting from the cluster analysis were examined visually to determine the appropriate level of similarity to act as the cut-point of the graph.

Nearest-Neighbors Searches. Nearest-neighbors (NN) searches were carried out using the procedure provided for that purpose in the DiverseSelector package.[4] The chemistry-space (the set of BCUT metrics) used was the same one used for cluster analysis. The databases used for the NN searches were obtained from commercial vendors of small organic compounds (e.g., Specs/BioSpecs, Maybridge, Chembridge,

* Corresponding author e-mail address: parker.cn@pg.com.

10.1021/ci9801015 CCC: $18.00
Published on Web 01/05/1999

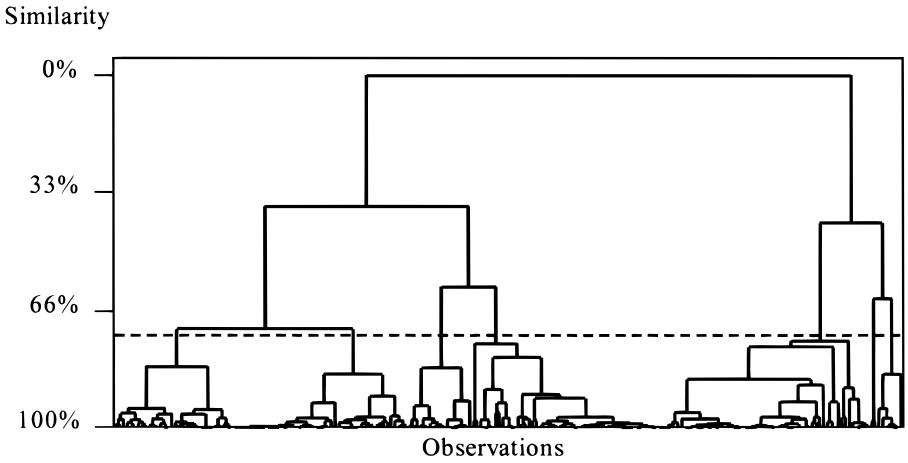


Figure 1. A dendrogram illustrating the results of cluster analysis of a set of 212 hits found to be active in HTS using a single point inhibition assay. The cut-point yielding the seven subsets shown in the plot was selected on the basis of visual examination.

etc.) in MACCS SD format.[6] In all cases, the structures of query compounds were added to copies of the vendor original databases before calculation of the BCUT metrics. Typically, the 20−30 closest neighbors to a given query were selected from each database for subsequent purchase and screening. Initially, this number was chosen arbitrarily as there was no known relationship between the molecular similarity of compounds and their biological similarity.

Antibacterial Activity. All these results describing antibacterial activity were obtained as Minimal Inhibitory Concentration (MIC) data for aerobic organisms as determined using the standard NCCLS microbroth dilution method[7] using Mueller Hinton broth and an innoculum size of 10^5 colony forming units per milliliter.

RESULTS AND DISCUSSION

Evaluating Hits from Screening Using Cluster Analysis. The concept of screening a diverse set of compounds that cover as many different chemical structure templates as possible has been eloquently described by scientists from Tripos/PanLabs[8] and has been used to generate "optimally" diverse libraries of compounds. Several approaches have been developed to describe compounds using molecular structure descriptors (e.g., molecular fingerprints[9]) or even surrogate biological descriptors of molecular structure (e.g., the TRAP technology of Terrapin Technologies Inc.[10]). All of these methods provide the researcher with the ability to build or select a large and structurally diverse collection of compounds to facilitate the identification of screening hits that can be further progressed to identify potential drug candidates. High throughput screening of large numbers of compounds ($N > 100\,000$) is likely to generate a large number of hits. While the hit rate of any assay can be controlled by altering the assay conditions used or by setting more rigorous criteria (e.g., by calling for 80% inhibition rather than 50% or by reducing the test concentration), it may still be appropriate to take a broad look at the hits. Even though significant effort may have been expended in ensuring the diversity of the screening library, HTS can still result in a large number of hits. In such a case, it is not unreasonable to assume that the individual hits do not each represent a unique structural class, and it is entirely possible that many hits are simply representatives of a class of compounds. We have found that application of cluster analysis methods provides a rapid and simple way to reduce a large set of HTS hits to a smaller more manageable number of structural classes. This approach has the effect of significantly reducing the number of "hits" that have to be considered for second-tier screening and allows the available resources (screening, follow-up synthesis, etc.) to be used in a more efficient manner.

This application of cluster analysis is illustrated using the following two examples. The first example involves the examination of 212 hits identified from among a large combinatorial library ($N > 10\,000$) using an assay designed to identify compounds with antibacterial activity. This library was synthesized to be used for general HTS applications and was not focused on any particular target assay. The goal of the computational effort in this case was to determine if the collection of hits identified represented a set of 212 unique classes of compounds, or if some could be grouped to make follow-up more efficient. The results of the hierarchical cluster analysis are shown graphically in Figure 1. An examination of the dendrogram suggested that a reasonable cutoff point for the graph would yield seven subsets of structures. Follow-up studies were then conducted on representative structures from each of the subsets involved. As a result, the number of actual "hits" to be considered was reduced 30-fold, allowing for more efficient use of screening and follow-up resources.

Another application of this approach is illustrated in Figure 2. In this example, cluster analysis was used to examine a collection of 35 structurally confirmed hits from HTS involving a focused combinatorial library designed for a specific assay. Three separate classes of compounds were clearly visible in the resulting dendrogram. Separate follow-up studies were then made using representatives from each class. It was interesting to note that the average antibacterial activity observed for these compounds seemed to correlate with the group classification (i.e., with the activity of group 1 > activity of group 2 > activity of group 3). Cluster analysis has helped to focus follow-up work on representatives of each class of compound very early in the process, resulting in a reduction in the amount of follow-up analyses required by an order of magnitude, while still allowing a

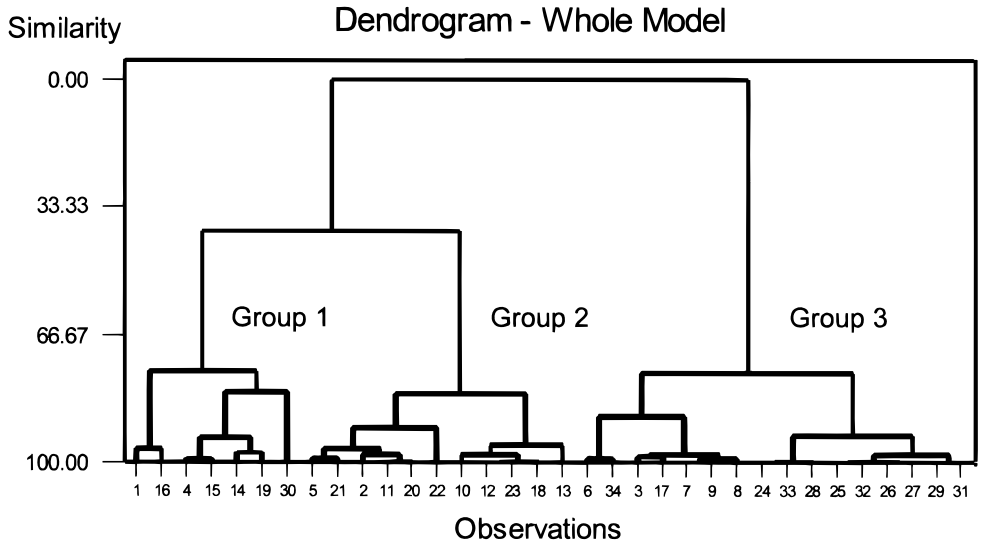


Figure 2. Results of the cluster analysis of a focused library of compounds showing the identification of three separate subsets of compounds.

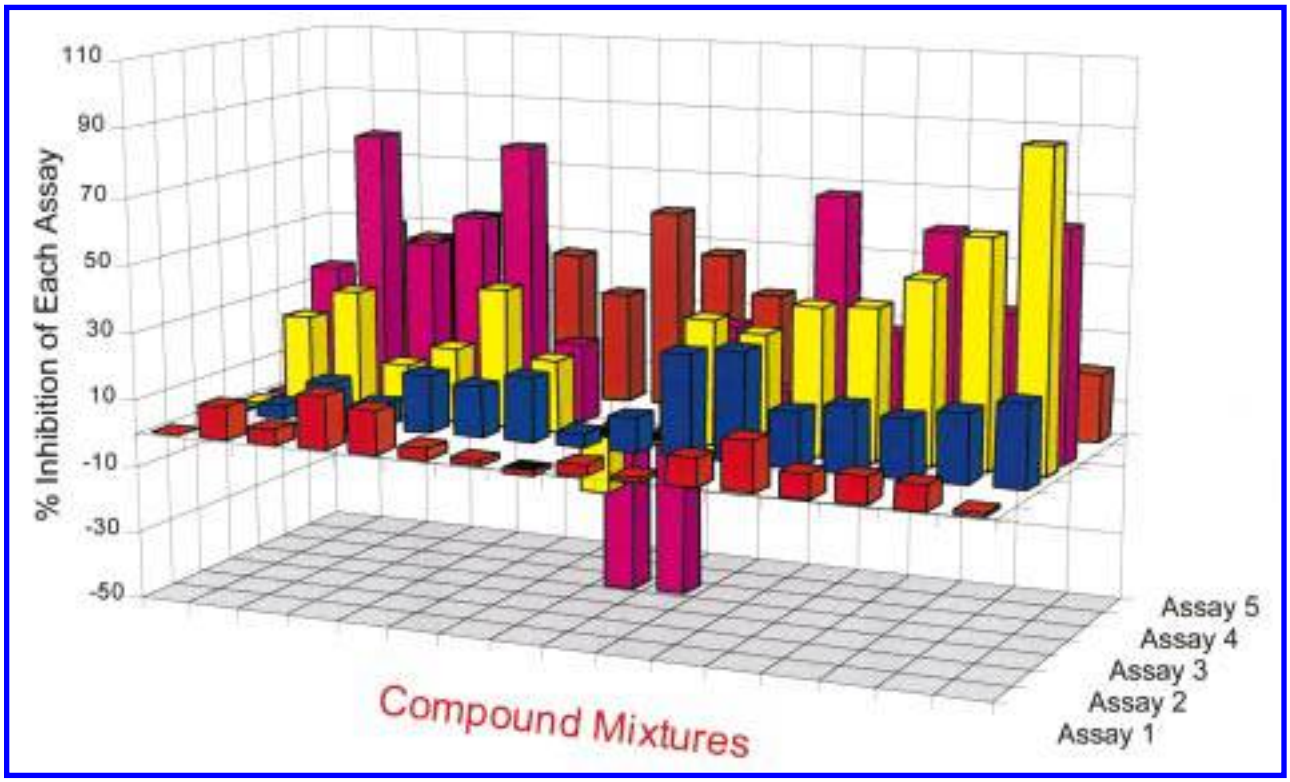


Figure 3. Screening results (HTS) obtained for a set of 14 different compound mixtures from five different anti-infective assays.

full evaluation of these classes of compounds in subsequent assays.

Cluster analysis has also shown utility in the detection of false-positives among HTS hits. This has been especially true when the source of compounds is combinatorial synthesis. In these cases, one may observe that the structural similarity among the hits is low (many small clusters) and that the list of hits appears to include a large number of structurally unique compounds. However, we have observed that, in such cases, the activity detected in screening is often the result of a common side-product that is active. Thus, the computational methods also assist in a quality control capacity for compound purity.

In addition, computational methods may also serve a quality control capacity in selecting relevant biological assays which can identify different classes of hits. For example, in the case of anti-infectives, it is relatively easy to screen for inhibition of bacterial cell growth, inhibition of bacterial metabolic pathways or isolated enzymes, known to be essential for cell growth. However, for these different assays to be worth performing they should have the potential to generate different classes of hits, all of which could lead to novel antibacterial compounds. For example, Figure 3 shows the results from screening a set of 14 compound mixtures with five different assays, all focused toward identifying novel antibacterial compounds. The results demonstrate that while these assays are all designed to identify antibacterial compounds, they do identify different hits, which will then act as the starting point for further expansion. The availability of a diverse set of compounds then gives the biologist a

Table 1. Antibacterial Activity of the Five Most Active Compounds Identified as Nearest Neighbors to PG 5730591[a]

	compd no. and antibacterial activity (MIC) in μg/mL					
bacterial species	PGE 2278007	PGE 3961402	PGE 4085286	PGE 4483603	PGE 5711080	PGE 5730591
S. aureus	0.12	4	4	8	4	8
S. aureus	0.25	2	4	4	4	8
E. faecium	0.12	4	16	64	8	8
E. faecalis	0.12	4	16	32	16	8
S. pneumoniae	</=0.06	2	8	8	4	4
S. pneumoniae	</=0.06	2	4	8	4	4
E. coli	64	>128	>64	>64	>128	>128
K. pneumoniae	4	1	8	>64	8	4
Euclidean distance from original hit	2.38	2.09	1.66	1.46	0.95	0

[a] This shows the activity of a compound (PGE# 5730591) identified by random screening of the P&GP compound repository. Also listed is the activity of the six best hits, identified from 210 compounds, ordered from seven different commercial vendors and the Euclidean distance these compounds were found from the original hit.

means to determine if the assays they are using will allow the identification of different and distinct classes of compounds.

Using Molecular Similarity in HTS Hit Expansion and Follow-Up. Once hits have been identified from a screening program, it is possible that the compounds of interest may have been identified from a portion of the library that is sparsely populated. Alternatively, the hit may come from a class of compounds with a common scaffold and which is already patented by a competitor or is known to have undesirable properties. However, since similarity is not scaffold-dependent in the sense of pharmacophore presentation, it should be possible to use similarity measures of molecular structure to identify other active molecules resulting from different scaffolds. To use a lock and key analogy, the lock does not care if the key is made from steel, copper, bronze, or composite alloy just as long as the key fits the lock, displays the correct tumblers, and is strong enough to force the lock, to allow the lock to be opened (or simply, structure dictates function).

With the large numbers of compound vendors now available, it is possible to identify large numbers of compounds with similarity to any given hit and then purchase and test them to determine the extent of chemistry space surrounding any hit to be explored. The goal in this case is to find and screen sets of compounds that are similar with regard to both structure and activity. Using these data, it is possible to start building a preliminary understanding of the structure−activity relationship (SAR) involved. This preliminary SAR information can then be used to focus further hit expansion and directed synthetic effort. We have found that the process of nearest neighbor (NN) analysis can be used effectively to identify sets of potentially active compounds which are similar to a given active query (a hit). An example of this approach is given in Table 1, which shows the activity of a compound (PGE# 5730591) identified by random screening of the P&GP compound repository. This table also lists the activity of five related hits, identified from 210 compounds, ordered from seven different commercial vendors and each compound's Euclidean distance from the original query. This example highlights the point that while nearest neighbor analysis will allow the identification of similar active compounds, the pattern of activity shown by these compounds will also help to identify a vector of increasing activity to guide further synthetic efforts. This shows that it is possible to identify related and active compounds starting from just one active, thus helping to identify structure activity relationships, and all without having committed valuable internal synthetic chemistry efforts to follow up on that hit. This figure also highlights the point that the original hit will most probably not represent the most active compound present in a given "island" of activity.

Table 2. Summary Table of 11 Nearest Neighbor Searches[a]

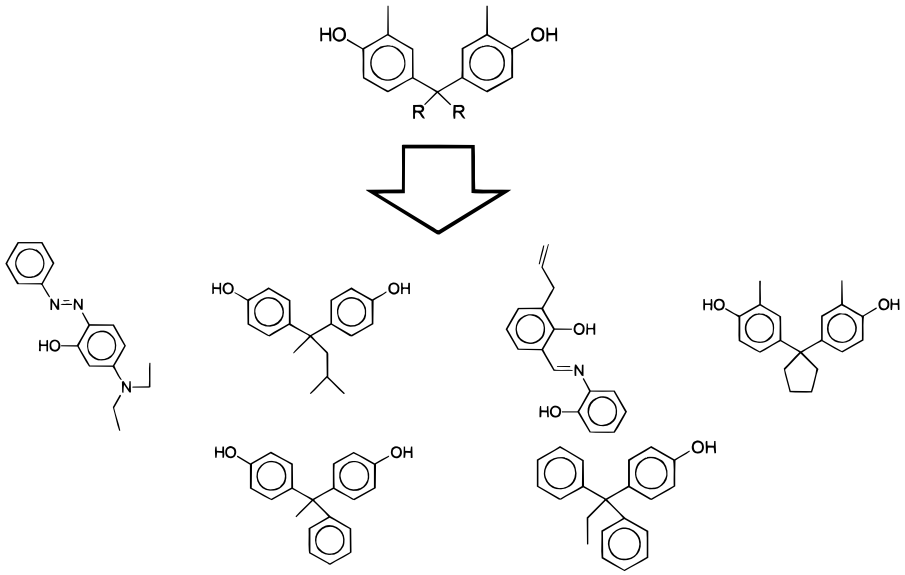
example	no. of compds tested	no. of compds showing MICs of </= 16μg/mL	hit rate as a % of compds tested
1	41	4	10
2	39	4	10
3	58	15	26
4	66	7	10
5	37	6	15
6	54	6	11
7	58	4	6
8	56	7	12
9	56	20	36
10	46	7	15
11	35	1	3

[a] This table gives the results of 11 different nearest neighbor searches for compounds identified and purchased from commercial vendors from hits having antibacterial activity. Both the original queries and the additional hits demonstrated MICs of 16 μg/mL or less.

This strategy can be extended to using actives identified by other researchers as a starting point for the identification of NN which may represent novel leads. An example of this is given in Figure 4, where an active originally described by J. Domagala[11] of Parke-Davis against a target very similar to ours was used as the starting point of a NN search. From a collection of only 80 compounds obtained from just one vendor, we identified a number of compounds with activity of <= 50 μM. As illustrated in Figure 4b, many of these compounds showed antibacterial activity, and yet some of these compounds displayed many differences in their basic scaffold, giving an indication of the structural requirements for activity. So while nearest neighbor analysis will help to generate more leads, these compounds will also help identify a vector to guide further synthetic efforts.

The results illustrated in Figure 4 are just one example of how this approach has been applied. This strategy has consistently generated collections of compounds which give hit rates well above that observed when using the same assays to screen random collections of diverse compounds. Table 2 gives the results of testing nearly 1000 compounds

a Biaryl nearest neighbor analysis



b **Nearest neighbor results**

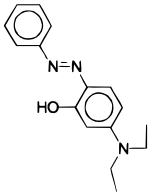
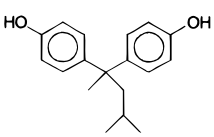
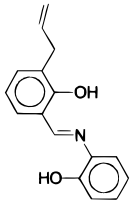
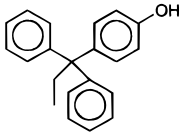
Compound	Species	Strain	MIC (ug/ml)
	*S. aureus*	MI246	>125
	S. pneumoniae	STP6301	>125
	B. catarrhalis	BC2	>125
	*S. aureus*	MI246	<=0.3
	S. pneumoniae	STP6301	<=0.3
	B. catarrhalis	BC2	<=0.3
	*S. aureus*	MI246	2
	S. pneumoniae	STP6301	0.5
	B. catarrhalis	BC2	<=0.3
	*S. aureus*	MI246	0.6
	S. pneumoniae	STP6301	<=0.3
	B. catarrhalis	BC2	<=0.3

Figure 4. (a) Results of a nearest neighbor search from just one vendor for a compound originally described by J. Domagala.[11] (b) Structures and screening results for four of the compounds that were identified as nearest neighbors as shown in part (a).

identified in 11 separate nearest neighbor searches. Even the lowest hit rate of 3% is at least 30 times higher than the hit rate experienced in random screening for this type of assay (typically between 0.01 and 0.1%). The results described in Table 2 are illustrated in Figure 5, by plotting the hit frequency for nearest neighbors compared to the "nearest neighbor" (Euclidean) distance of each compound compared to its respective query. These data are represented in two ways: first, Figure 5a shows the number of compounds identified and obtained at each Euclidean distance from a hit and the number of active compounds found, and second, Figure 5b then represents this as the percentage of compounds obtained and tested which were found to be active. A number of observations can be made from these graphs. First, the gap found in the region of the shortest NN distances is a result of the difficulty in obtaining compounds that are

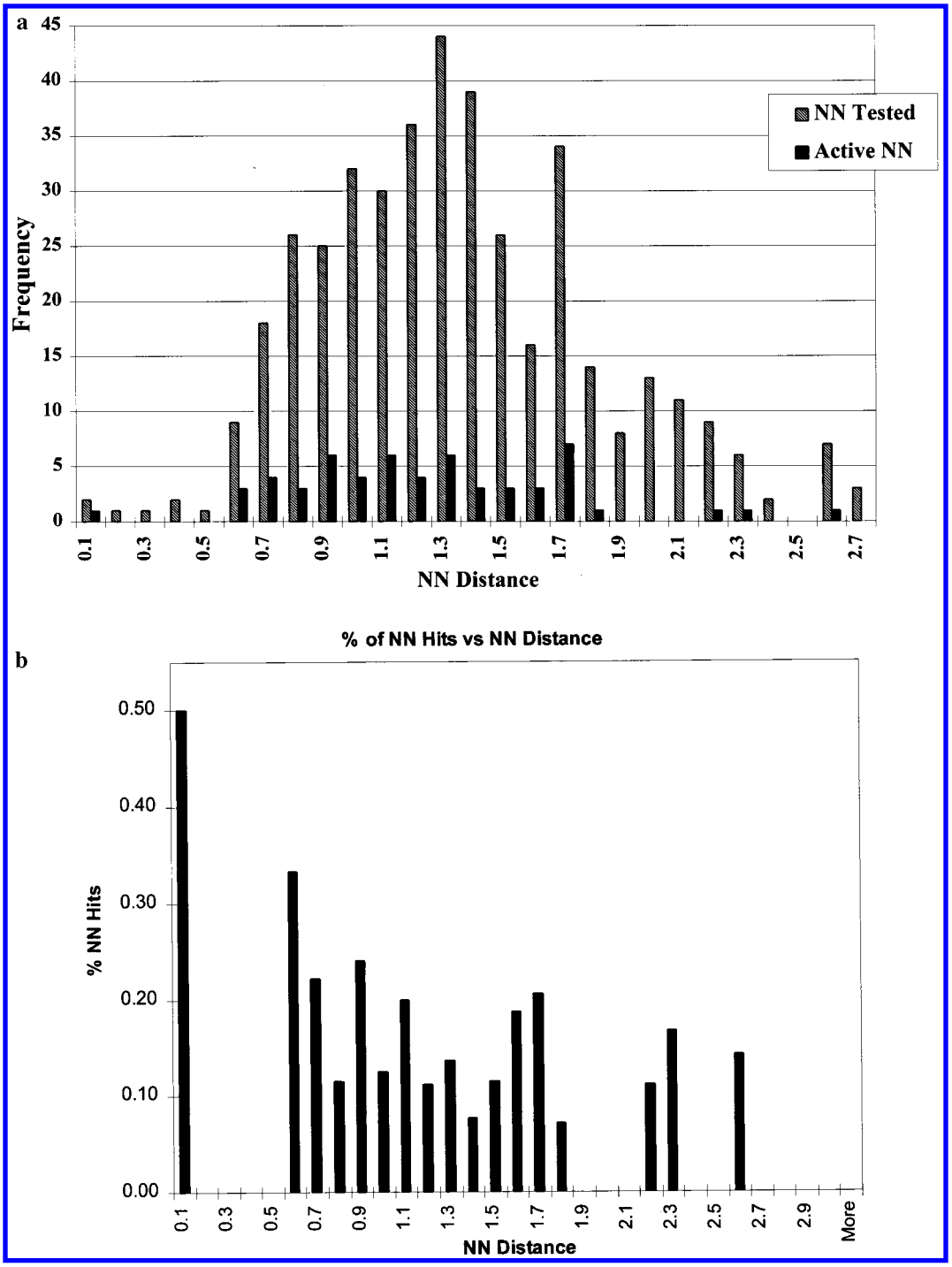


Figure 5. (a) Shows the number of compounds received for screening and the number of those that were found to be active at a given Euclidean distance from the original query (hit). Logistic regression was performed on these data and was found to be described by the following relationship: p(of obtaining a hit) = exp(−0.9786 − 0.6624*NNDistance)/[1 + exp(−0.9786 − 0.6624*NNDistance)]. Part (b) shows the data presented in part a with the data presented as the % of NN compounds showing activity at each Euclidean distance away from the original query (hit).

highly similar to the original hit. The overall average hit rate for these searches was found to be about 16%. However, it is also clear that the hit rate drops off dramatically with increasing NN distances from the original query. This type of result is expected given that the similarity to the original active hit was that much less. Analysis of these data using logistic regression showed that this drop-off is statistically significant at a confidence limit of 95% (p-value = 0.0394). This analysis also showed that one could be confident of obtaining a 20% hit rate by considering only the available compounds within a NN distance of 1.9 or less. The probability of finding a hit at a given distance (level of similarity) from the query structure is described by the following relationship

$$p(\text{of obtaining a hit}) = \exp(-0.9786 - 0.6624*\text{NNDist})/[1 + \exp(-0.9786 - 0.6624*\text{NNDist})]$$

where NNDist is the Euclidean distance between a near-neighbor and the query structure.

It should be noted that this relationship is relevant only in the context of the chemistry space we have employed, although there is no reason to expect that similar results would not be obtained using other validated chemistry space definitions.

CONCLUSIONS

The results described here illustrate a number of different instances where the application of diversity-related computational techniques have been applied to improve the process of random screening by including rational approaches. Computational measures of diversity offer the opportunity to improve the efficiency of early screening: first, by allowing diverse, representative collections of compounds to be identified for screening, and second, by offering a benchmark against which to compare the diversity of targets being used to screen these compounds.

However, it may not always be possible to submit an assay to a program of a comprehensive HTS screening. This may be due to the complexity or the cost associated with the assay. In such cases it is essential to find ways to focus screening efforts on the most productive areas of chemical space. Such efforts to focus screening could include literature based approaches where descriptions of inhibitors to targets similar to the one of interest can be used to identify other similar compounds, possibly "leapfrogging" a project into studying new compound classes. By extension, it may be possible to initiate focused screening efforts using NN analysis of substrates, ligands, or even suggestions from *de novo* drug design attempts for a novel target. For example, if the substrate specificity of a protease is known, searches to identify compounds similar to the amino acids either side of the cleaved peptide bond may identify hits. If ligands, such as peptides or small molecules, are known to bind a target, these can be used as the starting point for similarity searches. By extension of the previous examples, similarity measures can also be used to test hypotheses generated using rational drug design methods. If a crystal structure or a putative receptor binding pocket can be generated which describe the functional groups required for ligand binding, this can be used to generate a hypothetical molecule. Such hypothetical molecules could be generated and then synthesized by a motivated medicinal chemist. However, similarity searches of commercially available compounds can be used to quickly evaluate the validity of these hypotheses and may even generate a number of different structural scaffolds for the medicinal chemist to pick from for further expansion.

Once hits have been identified by HTS, computational techniques are also required to help efficiently evaluate the diversity of a large collection of hits. By using cluster analysis, it is possible to significantly improve the effectiveness and efficiency of the subsequent follow-up process. Additionally, this technique helps to focus the follow-up process on those sets of compounds that will provide information that can be used to build an understanding of the SAR involved. This in turn improves the ability of project chemists to design and synthesize new compounds leading to development leads, and hopefully, new drugs.

ACKNOWLEDGMENT

The authors would like to thank B. Kuzmak for help in performing the logistic regression on the data presented in Figure 5. The authors would also like to thank Prof. R. S. Pearlman for providing access to the DiverseSolutions software package and for supplying the chemistry-space definition used throughout most of this work.

REFERENCES AND NOTES

(1) DiverseSolutions User's Manual, Version 1.2.1; Laboratory of Molecular Graphics and Theoretical Modeling, College of Pharmacy, University of Texas at Austin: Austin, TX 78712.
(2) Minitab for Windows, Release 10; Minitab, Inc.: State College, PA.
(3) Pearlman, R. S.; Smith, K. M. In *3D-QSAR and Drug Design: Recent Advances*; Kubinyi, H., Martin, Y., Folkers, G., Eds.; Kluwer Academic: Dordrecht, Netherlands, 1997; pp 339−353.
(4) Pearlman, R. S., personal communication. The chemistry space was based on the examination of databases of 100 000−250 000 structures of druglike molecules. The BCUT metrics that formed the chemistry space definition were calculated using DiverseSelector (version 1.0.3) and included the following specific metrics: bcut_gastchrg_burden_0.01_R_H.bmf, bcut_gastchrg_burden_0.01_R_L.bmf, bcut_haccept_burden_10.00_R_H.bmf, bcut_hdonor_burden_0.10_R_H.bmf, bcut_tabpolar_burden_0.01_R_H.bmf, and bcut_tabpolar_burden_1.00_R_L.bmf.
(5) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1988; p 560.
(6) Dalby, A.; Nourse, J. G.; Hounshell, D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Lauffer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 244−255.
(7) National Committee for Clinical Laboratory Standards: Methods for Dilution Antimicrobial Suceptibility Tests for Bacteria that Grow Aerobically. Coc No. M7-A2, NCCLS, Villanova, PA, 1990.
(8) 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.
(9) Turner, D. B.; Tyrrell, S. M.; Willette, P. Rapid Quantification of Molecular Diversity for Selective Database. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 18−22.
(10) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A. E.; Bukar, R. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol*. **1995**, *2*, 107−118.
(11) Domagala, J.; Alessi, D.; Gracheck, S.; Huang, L.; Huband, M.; Johnson, G.; Olsen, E.; Shapiro, M.; Singh, R.; Somg, Y.; Van Bogelen, R.; Vo, D.; Wold, S. Bacterial two component signaling as a therapeutic target in drug design: Inhibition of NRII by Diphenolic-Methanes. Presented at the second International Antibacterial Discovery Summit, 1997; Princeton, NJ.

CI9801015