Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting

Steven L. Dixon*,† and Hugo O. Villar‡

Telik, Inc., 750 Gateway Blvd., South San Francisco, California 94080

Received June 6, 1998

The Similarity Principle provides the conceptual framework behind most modern approaches to library sampling and design. However, it is often the case that compounds which appear to be very similar structurally may in fact exhibit quite different activities toward a given target. Conversely, some targets recognize a wide variety of molecules and thus bind compounds that have markedly different structures. Affinity fingerprints largely overcome the difficulties associated with selecting compounds on the basis of structure alone. By describing each compound in terms of its binding affinity to a set of functionally dissimilar proteins, fundamental factors relevant to binding and biological activity are automatically encoded. We demonstrate how affinity fingerprints may be used in conjunction with simple algorithms to select active-enriched diverse training sets and to efficiently extract the most active compounds from a large library.

INTRODUCTION

High throughput screening (HTS) techniques are now used routinely in the pharmaceutical industry to assay the entire contents of large corporate libraries (>100 000 compounds) against biological targets. While this brute-force approach to lead generation certainly has its place in the field of drug discovery, it is not practical to adapt to the HTS format every new target of potential biological importance. The steady stream of "low throughput" targets being produced by genomics research obligates the continued development of library sampling techniques which can find low μM hits by screening relatively small numbers of compounds.

When there is no a priori knowledge regarding the structure of active compounds, the generally accepted procedure is to screen a diverse subset of the overall library, then examine compounds which are structurally similar to any promising leads. Within a given library, the success of this *rational sampling* approach is heavily dependent upon the makeup of the initial diverse subset, and, to some extent, on the way in which the similarity searching is done.

Once the algorithmic designs for diversity and similarity have been specified, the remaining determinant of success is the bioactive relevance of the descriptors used to characterize each compound. Descriptors which are devoid of information that is relevant to target activity cannot be expected to enhance the success rate of rational sampling over that of random sampling. While this point may seem obvious, it deserves special consideration as compounds are represented in chemical spaces of higher and higher dimensionality.[1,2] Without proper selection of descriptors, simply increasing the number of dimensions may tend to obscure information provided by the bioactively relevant subset. Descriptor selection itself is complicated by the fact that it is difficult to know which elements of structure are relevant without benefit of an existing QSAR model.

One way of addressing these issues is to redefine our notions of compound space. The goal is to minimize the number of dimensions and maximize the amount of bioactively relevant information provided. We describe here the selection of diverse and focused screening libraries based on a bioactive profile of each compound. These profiles, which we term "affinity fingerprints",[3] are based on the idea that commonalities exist among the binding sites of certain proteins and that these shared characteristics are manifested by statistical correlations in binding affinity data. To the extent that binding sites resemble one another, affinity fingerprints encode information that is directly relevant to bioactivity. We demonstrate how these bioactive profiles of compounds compare to a typical set of structural fingerprints known as the ISIS MOLSKEYS.[4]

AFFINITY FINGERPRINTS

Details have been published elsewhere[3] regarding the experimental measurement of affinity fingerprints and the general characteristics of the reference proteins, so we provide only an outline here. Briefly, affinity fingerprints are determined using high throughput competitive binding assays, wherein each compound in our library is screened against a panel of functionally dissimilar proteins. An IC_{50} value is determined in each assay, and the binding affinity is defined as $-\log_{10}(IC_{50})$ or pIC_{50}. The set of binding affinities measured for a single compound across the entire panel of proteins is termed the affinity fingerprint.

Ideally, proteins are selected for the panel on the basis of statistical criteria, the most important being orthogonality to other panel members. At the same time, proteins must also provide a minimum level of information about the library, and we generally require that greater than 20% of the compounds bind with an IC_{50} value below 100 μM. These two criteria produce a panel that is fairly small (<20 proteins), yet highly informative.

Orthogonality of a given reference protein is measured most conveniently in terms of its multicollinearity to the other proteins in the panel. A protein is obviously not worthy of

* Author to whom correspondence should be addressed.
† E-mail: sdixon@telik.com.
‡ E-mail: hugo@telik.com.

10.1021/ci980105+ CCC: $15.00
Published on Web 09/25/1998

inclusion in the panel if its set of binding affinities can be accurately fit as a linear combination of the binding values from the other proteins. The use of any particular multicollinearity threshold is somewhat arbitrary, but the difference between the measured and fitted binding values should certainly not approach the precision of the binding assay. In reproducibility experiments, we have found the pIC_{50} values from temporally separated HTS runs to correlate at usually no higher than $r = 0.9$. This serves as a strict upper bound for the R value in multilinear fits of the binding values for potential new panel members. In practice, multicollinearities in any panel we have ever used have ranged from about $R = 0.35$ to $R = 0.75$.

It is important to note that when a reference panel is fairly small, say, fewer than 10 proteins, then it is relatively easy to find new proteins which satisfy the $R < 0.9$ criterion. As the panel becomes larger, however, it becomes increasingly difficult to find informative proteins that cannot be fit in terms of the existing panel. We believe this phenomenon to be a reflection of panel completeness. This is not meant to imply that such a panel will be able to accurately predict affinities for every other conceivable protein. Some proteins are so selective that perhaps only one compound in 10 000 will bind with an IC_{50} below 10 μM. These types of proteins would be considered highly orthogonal to just about any finite panel, but they provide so little chemical information about the vast majority of compounds, that their inclusion in the panel is not justified.

Aside from statistical issues, there are of course practical considerations surrounding the composition of the panel. Protein availability, consistency, and stability, and whether or not a robust high throughput assay can be developed are factors that come into play. In some instances, a protein may cease to be available in sufficiently high quantity or quality, and it must be removed from the panel for all future fingerprinting. Thus, over time, the number of reference proteins we have used has fluctuated. The primary panel used in this investigation contained 16 proteins which were selected from a pool of several hundred according to the statistical and practical criteria just discussed. Some examples are presented with proteins that are not members of this primary panel, but which are under consideration for future panels.

Affinity fingerprints, like conventional QSAR descriptors, simply provide a means of characterizing or describing compounds in multidimensional space. Unlike structural descriptors, however, affinity fingerprints automatically tell us whether a compound has some or all of the features that are essential for favorable interaction with each of a wide variety of binding sites. This is a very powerful tool for rational sampling because most protein targets will share some binding site characteristics with one or more proteins in a sufficiently diverse panel. Figure 1 illustrates this principle at work. For a set of 200 compounds with diverse structures (average pairwise Tanimoto similarity = 0.309 based on ISIS MOLSKEYS), the measured binding affinities against human serum albumin are accurately represented as a linear combination of binding affinities from three other proteins. Any notion of accuracy is always somewhat subjective, but it must be remembered that this protein *surrogate model* is not based upon a series of structural analogues but rather on a set of compounds selected for diversity across an entire library. Figure 2 contains a representative sample (average pairwise similarity = 0.305) of compounds from the model, and they are seen to contain a wide variety of backbones and chemical functionalities.

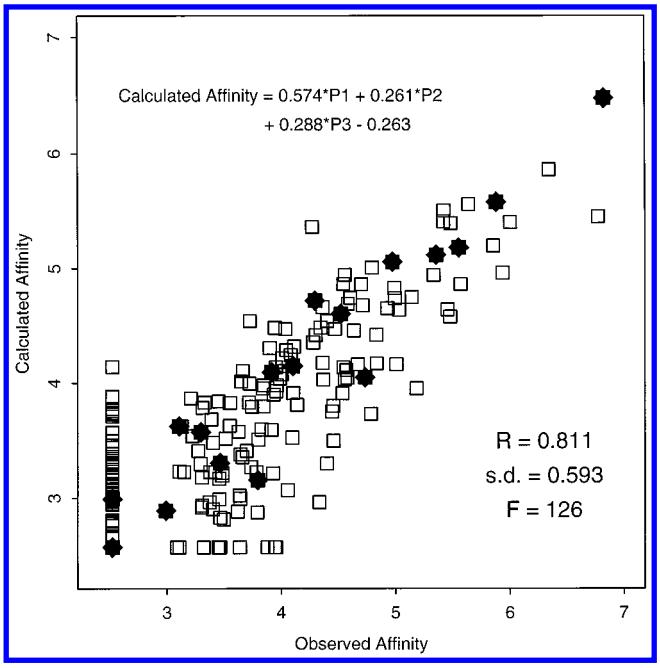

Figure 1. Protein surrogate model for human serum albumin. For a set of 200 structurally diverse compounds, the binding affinities (pIC_{50} values) measured for human serum albumin are approximately represented as a linear combination of binding affinities from three other proteins. Compounds associated with solid points are shown in Figure 2.

The ability to construct such surrogate models may seem at odds with our usual notions about proteins, i.e., that they only recognize a small number of compounds which align perfectly into a unique binding site. However, it is not unusual for a high affinity ligand of one protein to bind strongly to other proteins. Indeed, a lack of specificity is the downfall of many promising lead compounds in the drug discovery process. Also note that a great deal of the information that underlies a protein surrogate model comes from compounds which are not high affinity and thus have only a subset of the features that are necessary to bind strongly to the target protein. However, this is the exact sort of information that is required in order to carry out a search that starts with moderate affinity compounds and ultimately locates high affinity compounds.

MOLECULAR DIVERSITY

The scientific literature is increasingly populated by books and articles that address a wide range of issues surrounding the field of molecular diversity.[5,6] Topics include the choice of which molecular descriptors to employ,[2,7,15] the proper means of selecting diversity,[8−10,16] and reducing the dimensionality of compound space.[6,9,11] Since the way in which each compound is represented ultimately limits the success of all subsequent procedures, we begin the discussion with this fundamental and critical issue.

A number of important investigations[2,14,15,17] have focused on the selection of molecular descriptors that are able to distinguish compounds on the basis of biological activity. The datasets analyzed have usually been comprised of

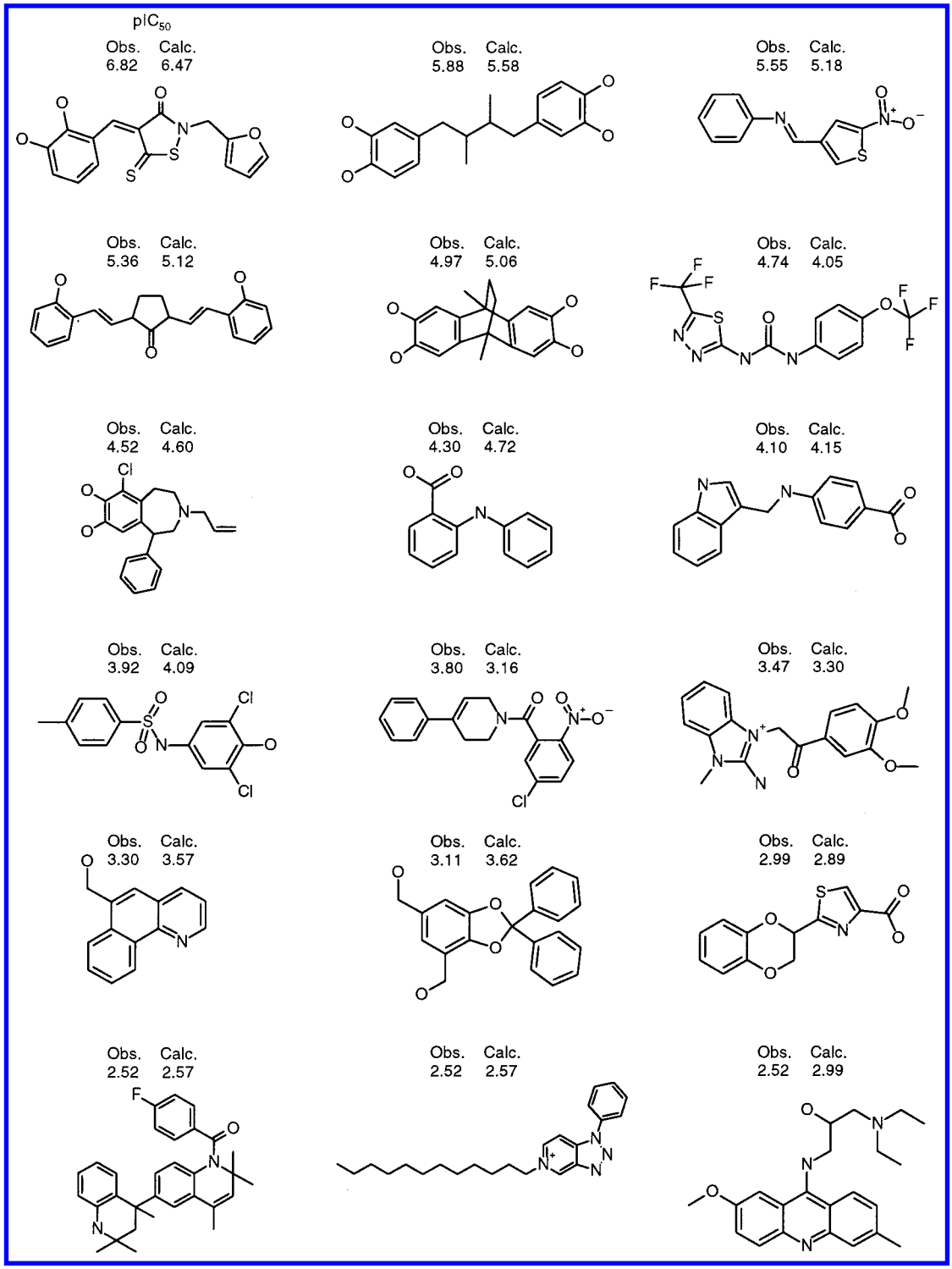

Figure 2. Sample compounds from human serum albumin surrogate model.

compounds that are selected either according to known activity against one or more targets or from libraries generated as a result of SAR studies around active hits. These datasets certainly encompass a great deal of bioactive diversity, and they are appropriate for demonstrating various properties of molecular descriptors. However, since the collections have tended to be biased toward the targets of interest, they contain a higher percentage of actives than would normally be discovered in a typical HTS of a large corporate library. Most published analyses have not focused on genuine HTS data, where the active hit rate is usually far less than 1%. As reported by Brown and Martin,[17] descriptors which are able to distinguish actives from inactives in small, biased datasets do much less well when applied to HTS data from larger, unbiased libraries. Since we are ultimately concerned with the discovery of new drugs in a practical setting, it is important to consider molecular diversity in the context of real libraries, where only a tiny fraction of the compounds will exhibit high activity.

One aspect of compound representation which deserves some discussion is to what extent molecular diversity as viewed by chemists and as calculated by structural descriptors actually resembles diversity as viewed by a biological target. Figure 3 illustrates how structural diversity and bioactive

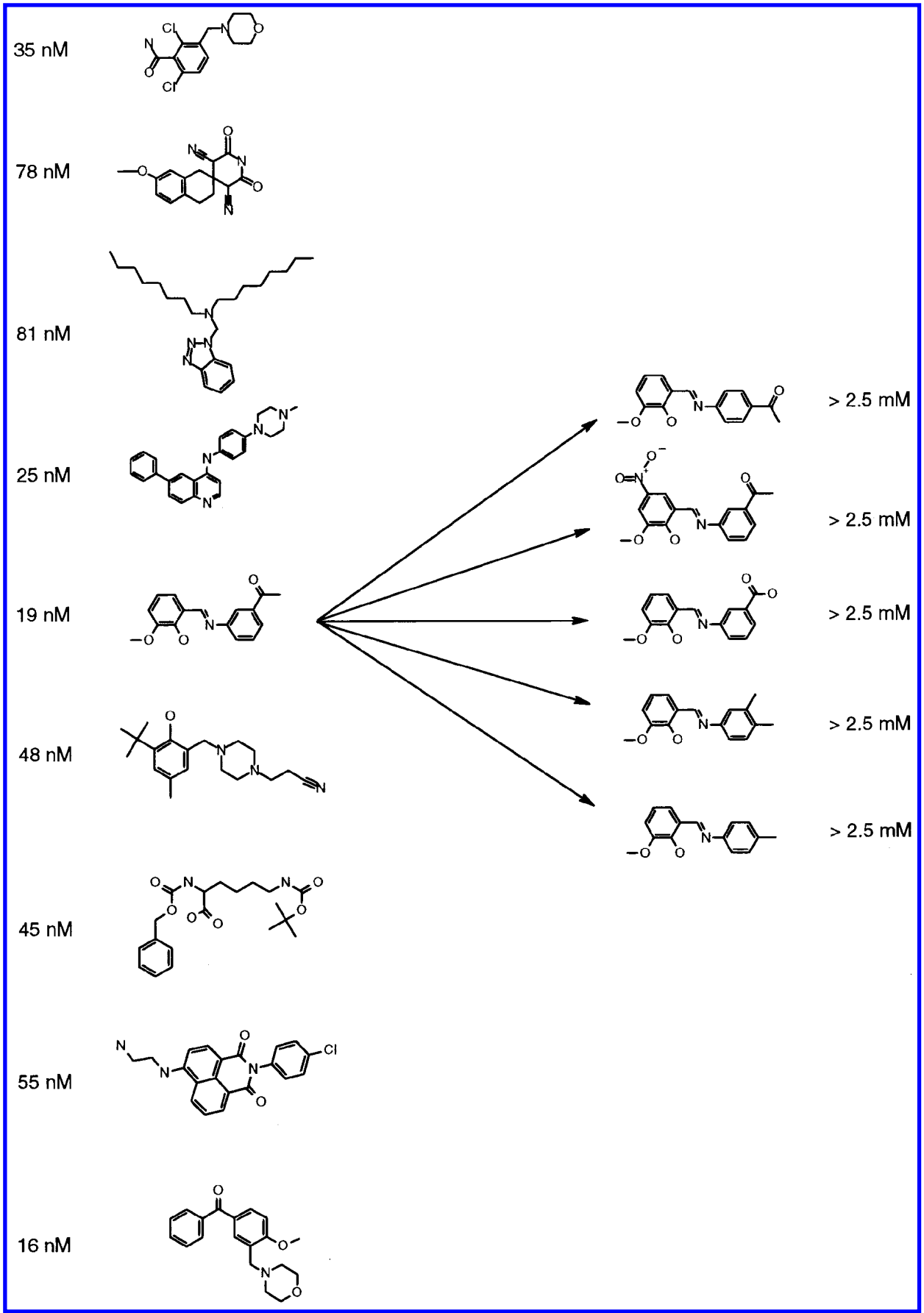

Figure 3. A counterintuitive example of structural diversity and its relationship to binding affinity for a member of the reference panel. Here, binding affinity is indicated by the IC_{50} value from a competitive binding assay.

diversity can sometimes seem to be at odds. For one of our reference proteins, IC_{50} values and structures for several high affinity compounds are shown. This series of nM hits clearly exhibits a considerable amount of structural diversity, with

an average pairwise Tanimoto similarity of only 0.421 based on the ISIS MOLSKEYS. Though they all look different to the eye of a chemist, they all in fact bind quite strongly, implying a high degree of similarity from the protein's perspective. Paradoxically, when small structural modifications are made to one of these compounds (average similarity to hit = 0.873), the affinity drops by more than five orders of magnitude. From the perspective of the protein, then, these compounds are all quite different from the 19 nM hit they so closely resemble structurally.

Examples such as these are not difficult to find after a library has been assayed against several proteins. Nevertheless, it is very perplexing that high affinity can be preserved while leaping across chemical families, yet it can be destroyed altogether by one small structural change. The generally accepted explanation is based on a pharmacophore concept, i.e., that all of the high affinity compounds must in fact possess the correct combination and orientation of groups to interact favorably with the binding site on the protein. One small change can remove one of these essential elements and a great deal of the affinity.

Once a compound with these essential features is identified, it is often possible to explore the chemical space around it and develop a QSAR that satisfactorily explains the variations in affinity. In general, though, the relationship is only valid for compounds sufficiently similar to those used in developing the model. This is because the descriptors are frequently just measuring small differences among compounds that all share some common backbone or scaffold which provides an appropriate framework for attachment of groups that can lead to high affinity. In effect, it is holding constant a myriad of other factors that govern affinity to a particular binding site. If this active backbone is replaced by another, then the descriptors may not carry over to a QSAR built around the new backbone.

This discussion is directly relevant to the issue of molecular diversity, because many structural parameters, while extremely effective in explaining differences in biological activity among compounds that are referenced to some restricted template, do not necessarily give meaningful comparisons of activity among things that have gross structural differences. It is very difficult to conceive of structural descriptors that can reliably predict activity for a given target across the range of compounds present in a typical corporate library. If there were such descriptors, then HTS would not be nearly so widespread as it is. The lack of library-wide QSARs is a result of our inability to model the range of thermodynamic processes involved in binding between a protein and an essentially unrestricted collection of compounds. Yet in order to select subsets of compounds that exhibit significantly more bioactive diversity than random sampling, one needs descriptors which correlate with activity in this global sense. The implication is that unless one has access to such parameters, then focusing on many of the finer points of molecular diversity may have a limited impact on the amount of bioactive diversity present in the compounds selected.

This is not meant to imply that conventional molecular diversity is a waste of time. In many instances, a library has been biased in favor of certain classes of targets, so decisions about diversity should consider this information whenever a new target in one of these classes is encountered. And, of course, there are certain types of structural features in small molecules that should generally be avoided because they are associated with undesirable chemical and pharmacokinetic properties. Overall design issues can also have an impact, as there are extreme cases where a diversity algorithm does not give reasonable coverage of the space it samples.[10,12] But in the absence of library-wide knowledge of biological activity, there is little reason to believe that one subset of compounds selected in an unbiased fashion will be significantly more prolific than another when it comes to generating leads from large libraries that have no particular bias toward the target of interest.

DIVERSITY ALGORITHMS

We now turn our attention to the issue of algorithms for compound selection and focus on two simple but significantly different diversity designs: one which selects compounds that are distributed in an approximately uniform fashion throughout space and one which samples compounds only from the edges of space. These two subsetting approaches, which we shall refer to as spread and edge, were chosen to see whether or not drastic differences in design really make any difference, and also whether deliberately selecting outliers, i.e., edge compounds, would have a deleterious effect on rational sampling.

In spread design, the goal is to select a subset of compounds **S** that fills the chosen descriptor space with minimal redundancy. The approach we adopt involves picking subset members that are as far away as possible, on average, from their nearest neighbors. Accordingly, the objective function to maximize for spread diversity was defined as

$$O_{\text{spread}} = \sum_{i \in \mathbf{S}} \text{MIN}(d_{ij}:\ j \in \mathbf{S}, j \neq i) \qquad (1)$$

In the case of affinity fingerprints, d_{ij} represents the Euclidean distance between compounds i and j. For binary structure keys, d_{ij} is simply one minus the Tanimoto similarity, which, for a string of n bits, is given by

$$d_{ij} = 1 - \sum_{k=1,n} \text{bit}_{ik}\text{bit}_{jk} / \sum_{k=1,n} (\text{bit}_{ik}^{\ 2} + \text{bit}_{jk}^{\ 2} - \text{bit}_{ik}\text{bit}_{jk}) \qquad (2)$$

A simple stochastic procedure is used to maximize the objective function. Starting with a randomly chosen subset, the two compounds with the smallest pairwise distance are identified. Of those two, the one which is closer to some other compound in **S** is flagged for ejection. This flagged compound is exchanged for one that is outside of **S** if the exchange will bring about an overall increase in the objective function. A series of these pairwise exchanges is made until no further increase in O_{spread} can be achieved. At this point, a new random subset is selected, and the procedure is repeated. After several random restarts, the collection of compounds with the highest associated objective function is retained.

Note that we are not using a Monte Carlo simulated annealing technique, genetic algorithm, or any other purported global optimization method, just a greedy exchange criterion with random restarts. We maintain that a genuine global optimization method for such problems has yet to be

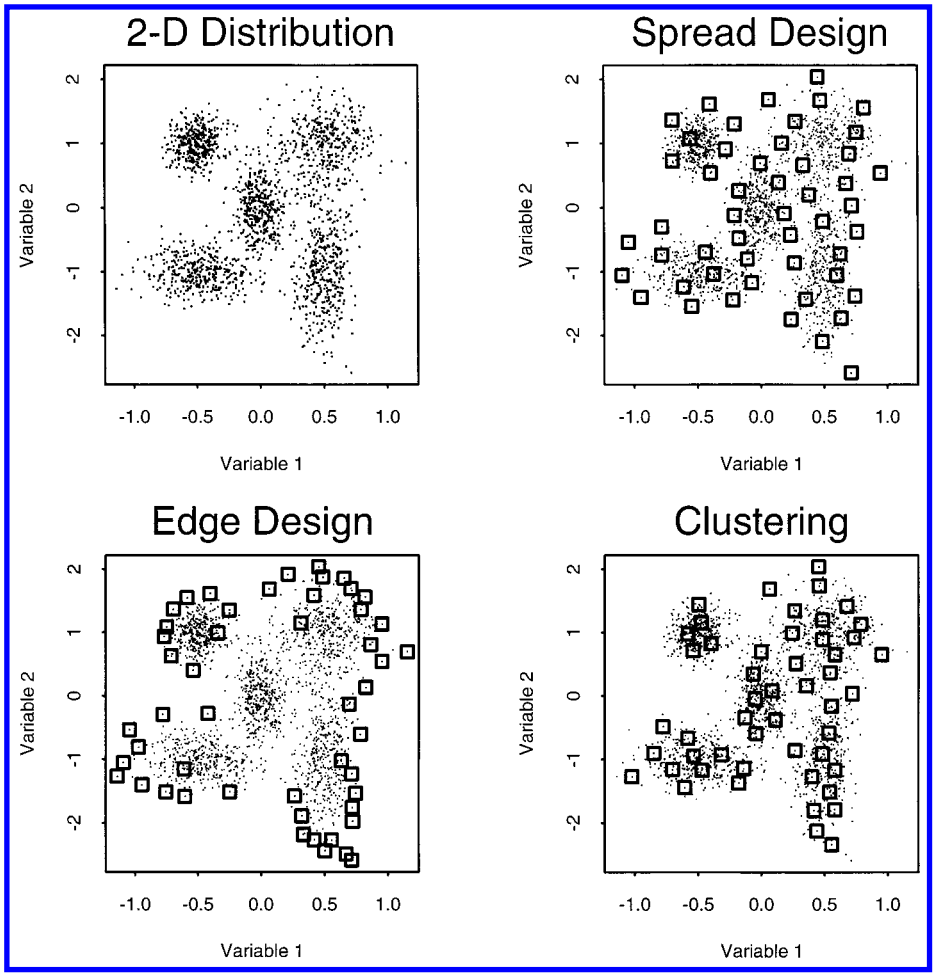

Figure 4. Illustration of spread and edge diversity designs using a two-dimensional distribution of Gaussian random data points. Cluster centroids from hierarchical agglomerative clustering are included for comparison.

developed, and one stochastic approach with a mechanism for local optimization can be made to perform about as well as another. The effectiveness of these "global" methods in locating satisfactory optima lies not so much in the subtle ways in which they go from totally arbitrary to locally optimal, but more in the opportunity that they afford to sample a wide range of randomly generated configurations. In simulated annealing, for example, this aspect is controlled by the cooling schedule. In our algorithm, it is controlled by the number of random restarts.

While the spread design seeks to maximize the average distance between each subset member and its nearest neighbor, the edge design attempts to maximize the average distance between each subset member and *all the remaining compounds* in **S**,

$$O_{\text{edge}} = \sum_{i,j \in \mathbf{S}} [d_{ij} - d_{av}(d_{av}/d_{ij})] \qquad (3)$$

This expression contains a penalty term, $-1/d_{ij}$, that prevents any two highly similar compounds from being selected. The average pairwise distance d_{av} observed over the entire library is used to construct a reasonable scaling factor for the penalty term. It is of course expensive to compute d_{av} for extremely large libraries, but a randomly chosen subset of 1000 compounds is usually sufficient to give a reliable estimate of this quantity.

Using the same type of stochastic approach described earlier, a series of pairwise exchanges is made in order to increase the objective function. The only difference is that the compound flagged for ejection is the one which exhibits the smallest average distance to the other subset members.

Both of these diversity algorithms are extremely simple to implement, and their computational expenses scale only linearly with the size of the overall library. They do exhibit quadratic scaling with respect to the size of **S**, but this does not become a serious drawback unless very large subsets are desired. The intended use here is for testing in a low throughput mode, so typically only 50−100 compounds would be selected, and quadratic scaling is not an issue.

Figure 4 illustrates how these selection methods behave when applied to a set of 2000 synthetically generated points in 2-D space. Here, a subset of 50 points (enclosed by boxes) was selected using each diversity algorithm. For comparison, results from hierarchical agglomerative clustering with complete linkage[13] are also included. In this case, 50 clusters were generated, and the point closest to each centroid was selected for the subset. Clustering bears some resemblance to spread in overall appearance, but the former is seen to be affected somewhat more by variations in the density of points. This is certainly not a criticism, but hierarchical agglomerative clustering does become prohibitively expensive for large libraries, regardless of the subset size. Note that the spread and edge subsets are quite different, so they should provide a good demonstration of the effect of diversity design on rational sampling.

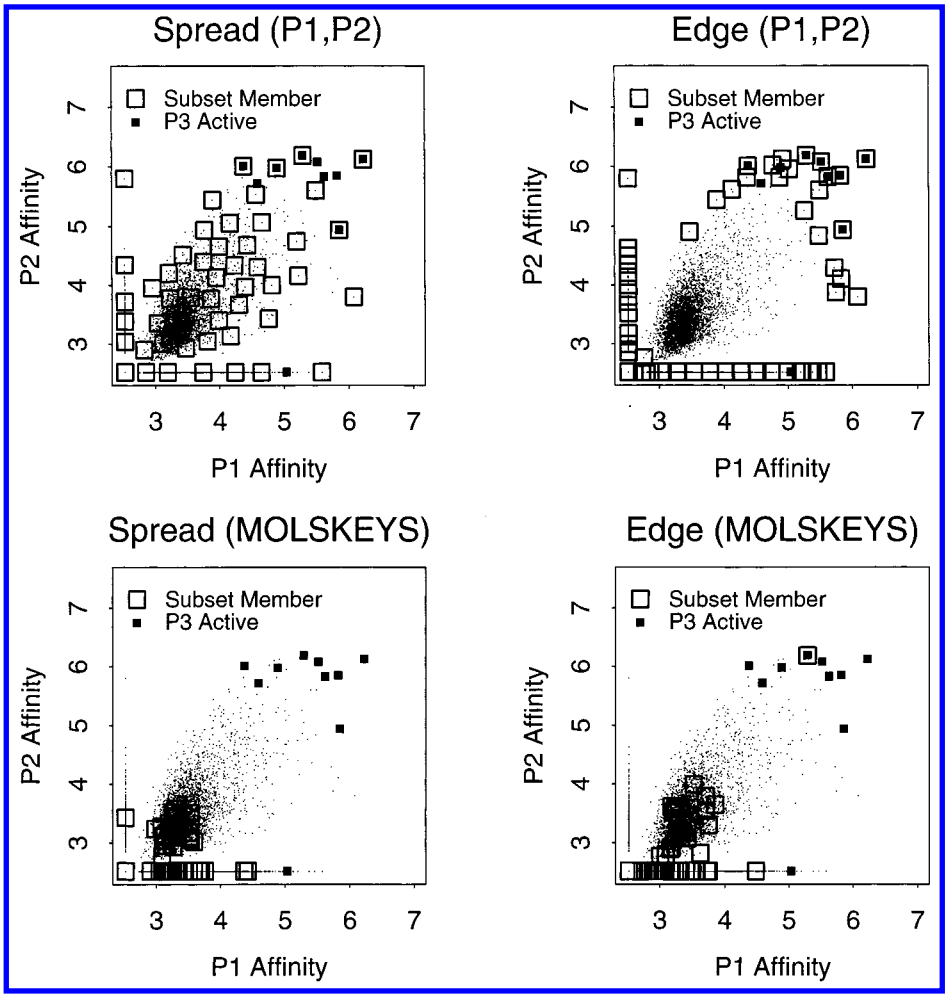

Figure 5. Affinity space representation of bioactively diverse (P1, P2) and structurally diverse (MOLSKEYS) subsets selected from a library of 8000 compounds. P3 is a protein which is statistically related to the affinity fingerprint proteins P1 and P2.

BIOACTIVE DIVERSITY EXAMPLES

As stated earlier, when one has access to molecular descriptors that correlate with target activity across an entire library, then the subtle issues surrounding diversity become more relevant. Distances in descriptor space then have a direct bearing on the distribution of activities, so it is possible to control, to some extent, the bioactive diversity of a subset.

Figure 5 illustrates the results of a diversity exercise carried out on 8000 compounds from our library. Here, P1 and P2 are proteins that comprise a 2-D affinity fingerprint, and P3 is a third "target" protein that exhibits a multicollinearity of $R = 0.75$ with P1 and P2. This is a stronger relationship than would normally exist between most targets and the reference panel, but these proteins were selected to provide a clear demonstration of the ability of bioactively relevant descriptors to select a greater number of compounds with high activity against the target.

For this exercise, active compounds on P3 were defined to be those with IC_{50} values below 1 μM. Using both edge and spread designs, subsets of 50 compounds were selected in the 2-D affinity fingerprint space (P1,P2), and in the 166-bit structural fingerprint space of the ISIS MOLSKEYS. All data is plotted in P1,P2 space to demonstrate the obvious edge-oriented pattern of P3 active compounds as well as the lack of correlation between diversity in the MOLSKEYS structural space and diversity in the space of affinity fingerprints.

Table 1. Summary of Average Properties of 1000 Randomly Selected Library Compounds

property	av value
MW	278
log P	2.59
no. of rings	2.18
diameter (in bonds)	9.90
no. of H-bond donors	1.52
no. of H-bond acceptors	5.19
no. of hydrogens	15.3
no. of carbons	13.9
no. of nitrogens	1.91
no. of oxygens	2.52

The edge algorithm applied with P1 and P2 locates seven of ten P3 active compounds, while the spread algorithm finds five of ten. Note that the success of the spread algorithm is essentially a consequence of its tendency to sample some compounds from the edge. When the MOLSKEYS are used to select compounds, the edge technique generates a subset which appears to be slightly more diverse in P1,P2 space than when spread is used, and the MOLSKEYS edge design selects one P3 active compound.

It should be apparent from Figure 5 that the majority of the 8000 compounds fall in a densely populated region of relatively low affinity on both proteins, and this region is rather distant from the compounds that bind most strongly to P3. The situation is somewhat exaggerated because of the deliberate choice of a target that is related to P1 and P2,

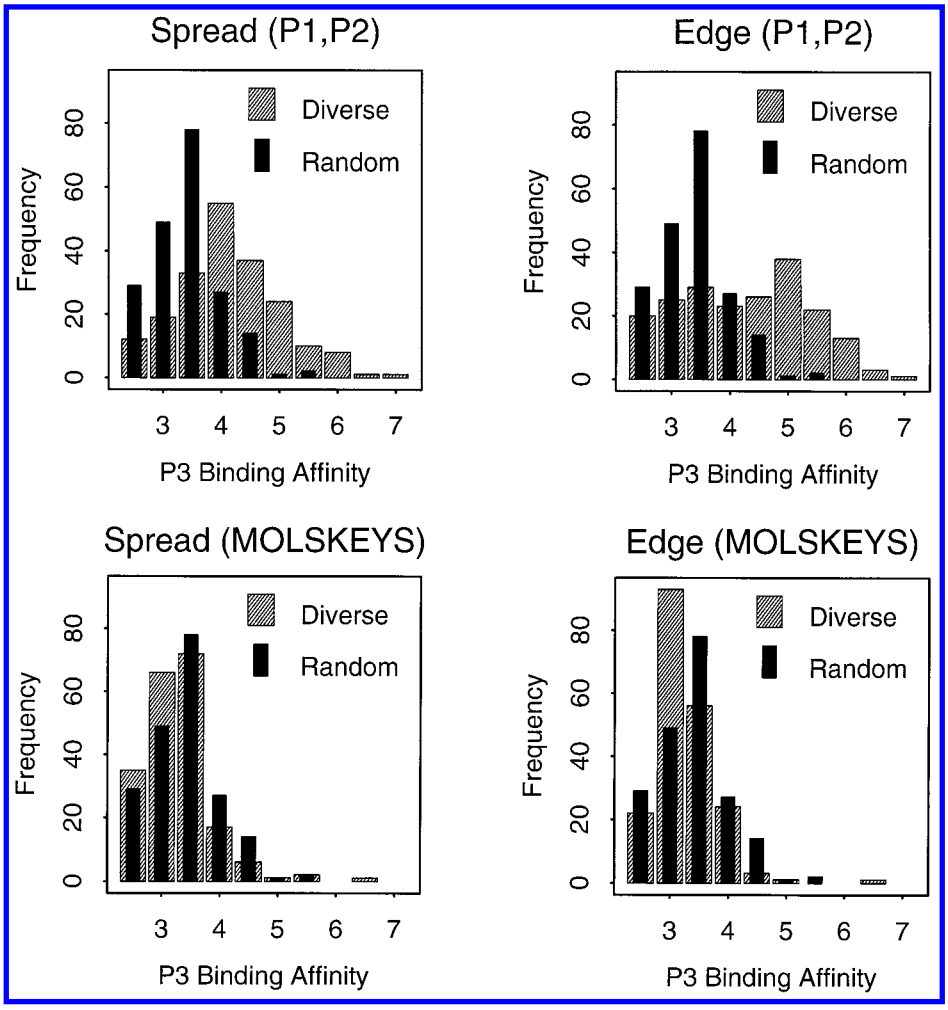

Figure 6. Frequency distributions of binding affinities are used to illustrate the bioactive diversity, with respect to protein P3, of compound subsets selected according to diversity in affinity fingerprint space (P1, P2) and structural space (MOLSKEYS).

but it illustrates an underlying reason why affinity fingerprints can be so powerful. Although the compounds with high affinity for a particular target may not always cluster in one small region of affinity fingerprint space, they do tend to bind strongly to at least one member of the reference panel and are thus distinct from the concentrated mass of compounds that have low affinity. This sort of separation is difficult if not impossible to achieve with ordinary structural descriptors, simply because the low affinity compounds are so diverse structurally.

Continuing with this example, Figure 6 summarizes the frequency distributions of P3 binding affinities for subsets of compounds selected as before. In these cases, however, the subset size was increased from 50 to 200 so as to obtain smoother statistics. For comparison, a single set of 200 compounds was selected randomly, and its distribution is overlaid with that of each diverse subset.

Compared to random selection, both of the P1,P2 subsets show a significant shift in the distribution toward higher affinities. The edge design actually exhibits a near uniform distribution over about three orders of magnitude in concentration. Intuitively speaking, this result is perhaps more along the lines of what would be expected with spread diversity. However, it must be remembered that the natural distribution of affinities is essentially a skewed bell shape, so the only way to achieve a uniformly distributed sample is to significantly bias the selection toward higher affinities.

By contrast, structurally diverse subsets selected using the MOLSKEYS appear to offer little or no advantage over random sampling as far as P3 is concerned. The spread design compounds are distributed very much like the random subset, with the only difference being that random selection appears to result in slightly higher average affinity. The edge design differs more significantly from random than does spread, but these differences are confined primarily to the region of low affinity compounds.

RATIONAL SAMPLING

After a diverse subset of compounds has been screened against a target, the activity data obtained from this *training set* may be used to select focused libraries for subsequent examination. Ideally, a series of small, focused blocks of compounds is screened, with new information being incorporated at the end of each block in order to fine tune the search for active compounds. This rational sampling approach to lead generation and optimization is summarized in Figure 7.

Each focused block is nothing more than a collection of compounds which are expected, or at least hoped, to be active. Upon screening, most of these compounds will usually turn out to be inactive, but with proper design of the focused block, the compounds should show a higher level of activity, on average, than the large library from which

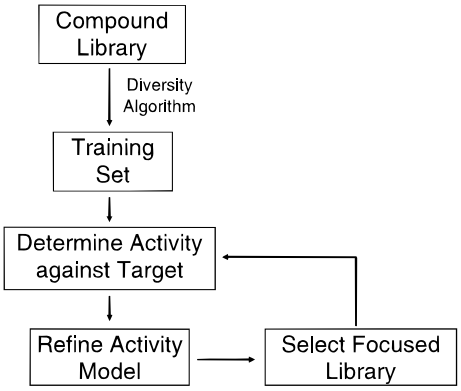

Figure 7. Flowchart summary of rational sampling methodology.

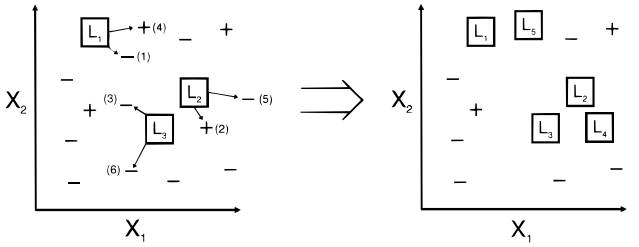

Figure 8. Illustration of the iterative search procedure used in nearest neighbors rational sampling.

they were selected. The choice of compounds may be based on an outright model of activity that can be applied to the remaining unscreened portion of the library, or it may be based on an implicit model, which assumes a neighborhood behavior of activity.[14] In this paper, we focus on the latter approach and employ a straightforward nearest neighbors search around lead compounds in order to locate additional high activity compounds.

Figure 8 illustrates the iterative search procedure used in nearest neighbors rational sampling. Initial leads are simply the handful of most active compounds uncovered in the training set screen, and they may or may not be actives. However, each time an active compound is discovered in the focused screen, it is incorporated into the list of leads. A search that cycles repetitively through all the leads is used so that the focused library is not confined to one region of compound space, and so that a disproportionate amount of time is not spent screening analogues of a lead that has no actives nearby.

Screening in blocks not only allows the search to be expanded around new leads but also affords the opportunity to determine which descriptors are contributing relevant information and which are not. Distances in compound space should reflect, as much as possible, the relative positions of compounds in activity space. One way of accomplishing this is to use what we call *activity-biased scaling*. A given descriptor x_k is weighted according to how strongly it has been observed to correlate with activity over the set of compounds that has already been screened,

$$x_k \rightarrow |r_k| \cdot x_k \tag{4}$$

Scaling by the correlation coefficient r_k automatically removes dimensions that appear to have no overall relationship with activity.

We now present the results of a number of rational sampling exercises applied to two protein targets: glutathione S-transferase P1-1 (GST P1-1) and papain. The library contained 20 000 compounds, some properties of which are summarized in Table 1. These compounds were obtained from various vendors, through collaborations with other companies, and from synthetic work for internal projects. A more detailed description of the types of compounds in our library is given in ref 3.

For each target, the 10 highest affinity compounds from the library were defined as active. This corresponded to compounds with IC_{50} values below 30 nM in the case of GST P1-1 and 250 nM for papain. The use of two different thresholds for activity is actually a good test of the robustness of any rational sampling approach, since one generally expects a higher degree of structural similarity among actives which bind with higher affinity.

Diverse training sets were comprised of 50 compounds, and each focused block contained 100. Unless otherwise noted, the three highest affinity compounds from the training set were used as initial leads for the focused blocks. Affinity fingerprints were generated from a panel of 16 proteins, which were selected independently from the targets. Although the entire 20 000-compound library was screened against these targets, the present exercises utilized no information beyond what was discovered as the rational sampling proceeded. Figures 9 and 10 summarize for GST P1-1 and papain, respectively, the results of three different approaches to rational sampling using affinity fingerprints. These exercises illustrate the effect of varying the training set design and the effect of using activity-biased scaling. Each curve is actually an average of ten separate rational sampling experiments generated using ten different training sets. Since there is a stochastic element associated with the diversity algorithms, a statistical average was used in order to achieve a more robust representation of the two designs.

Both targets exhibit the same basic trends with regard to the different rational sampling approaches. The most efficient searches utilize an edge design for the training set followed by activity-biased scaling for the generation of focused libraries. Spread diversity without activity-biased scaling provides the least satisfactory results. The advantage of using an edge-diverse training set is a direct result of the increased tendency of these compounds to exhibit high affinity for any target that shares binding site characteristics with the reference proteins. Activity-biased scaling further enhances the search by amplifying any underlying statistical relationships.

In terms of outright performance, affinity fingerprints are seen to be extremely efficient at extracting the highest affinity compounds from the 20 000-member library. Only 850 compounds need to be examined in order to identify all ten actives for GST P1-1. This 850-member subset is thus 23 times more enriched in active compounds than the library as a whole. Perhaps even more impressive is the fact that eight of ten actives for papain are identified after screening the training set and only two focused libraries of 100 compounds. The success rate drops off dramatically after this point, but for the first 250 compounds, the activity enrichment factor is 64.

The previous rational sampling exercises all involved the use of three initial lead compounds. Figure 11 illustrates the effect of varying the number of leads within the framework of an edge-diverse training set and activity-biased

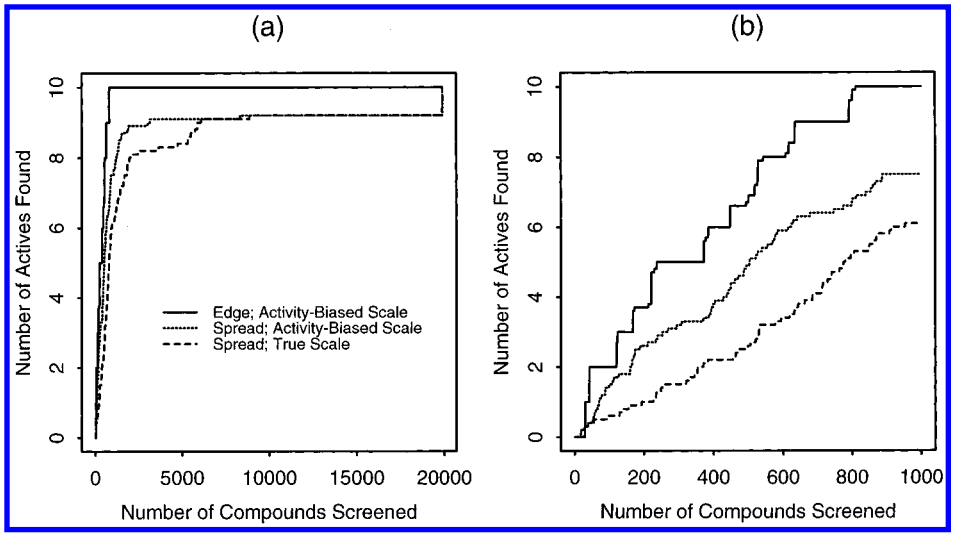

Figure 9. Rational sampling with affinity fingerprints. Effect of training set design and activity-biased scaling for the target protein GST P1-1: (a) full library and (b) the first 1000 compounds selected.

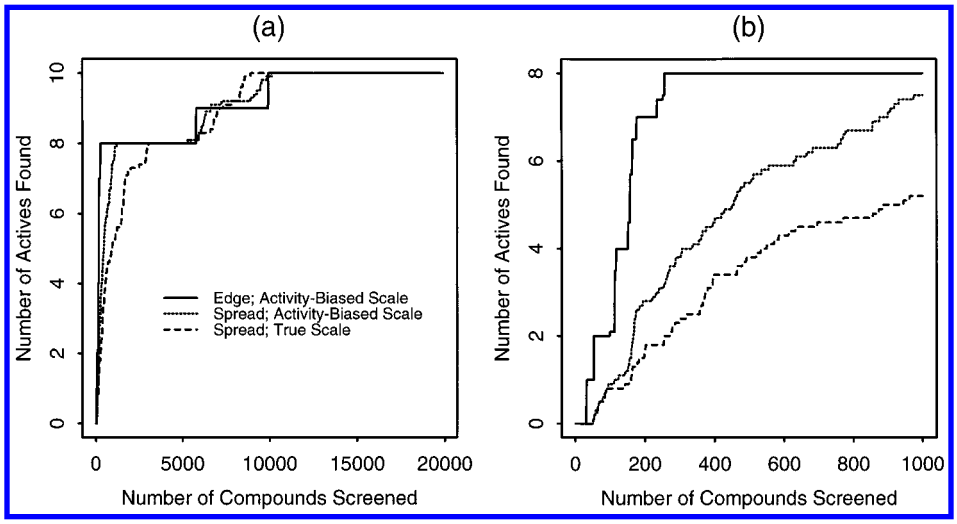

Figure 10. Rational sampling with affinity fingerprints. Effect of training set design and activity-biased scaling for the target protein papain: (a) full library and (b) the first 1000 compounds selected.

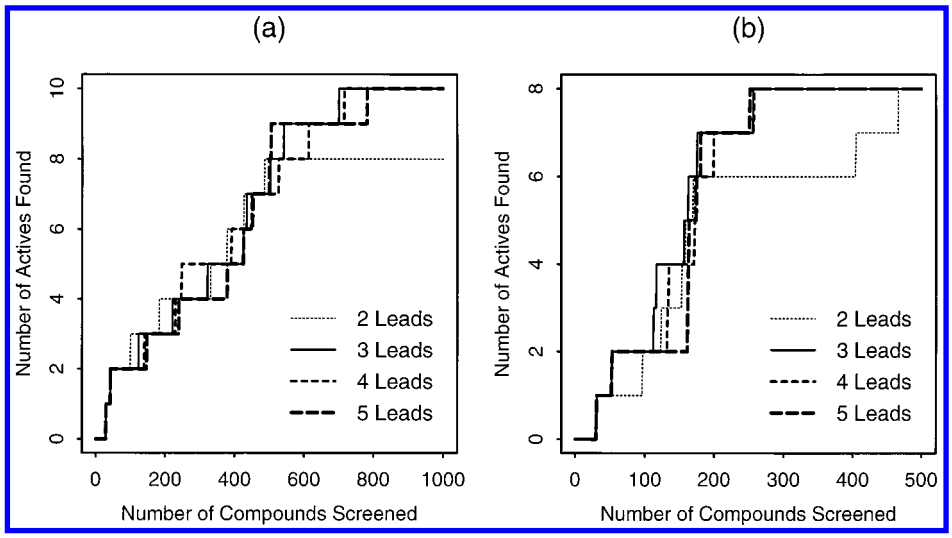

Figure 11. Effect of varying the number of lead compounds in affinity fingerprint rational sampling: (a) GST P1-1 and (b) papain.

focused screening. The results of using a single lead compound are omitted for clarity, and because the performance lagged far behind those of multiple leads. Overall, the choice of three leads appears to be a reasonable compromise between the competing factors of diversity and potency. If the number of leads is too small, then there may not be sufficient diversity within the focused libraries to cover the space occupied by the active compounds. If too many

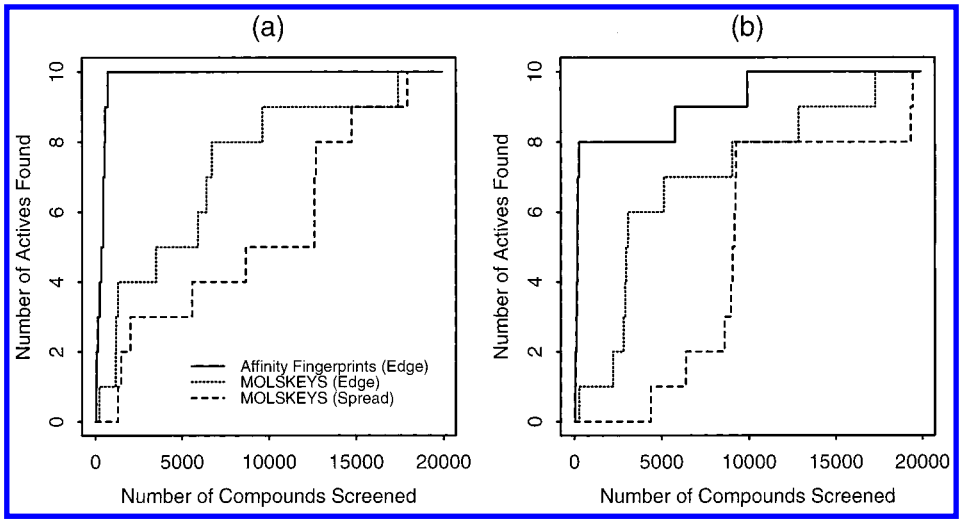

Figure 12. Head-to-head comparisons of affinity fingerprints and binary substructure keys for rational sampling: (a) GST P1-1 and (b) papain.

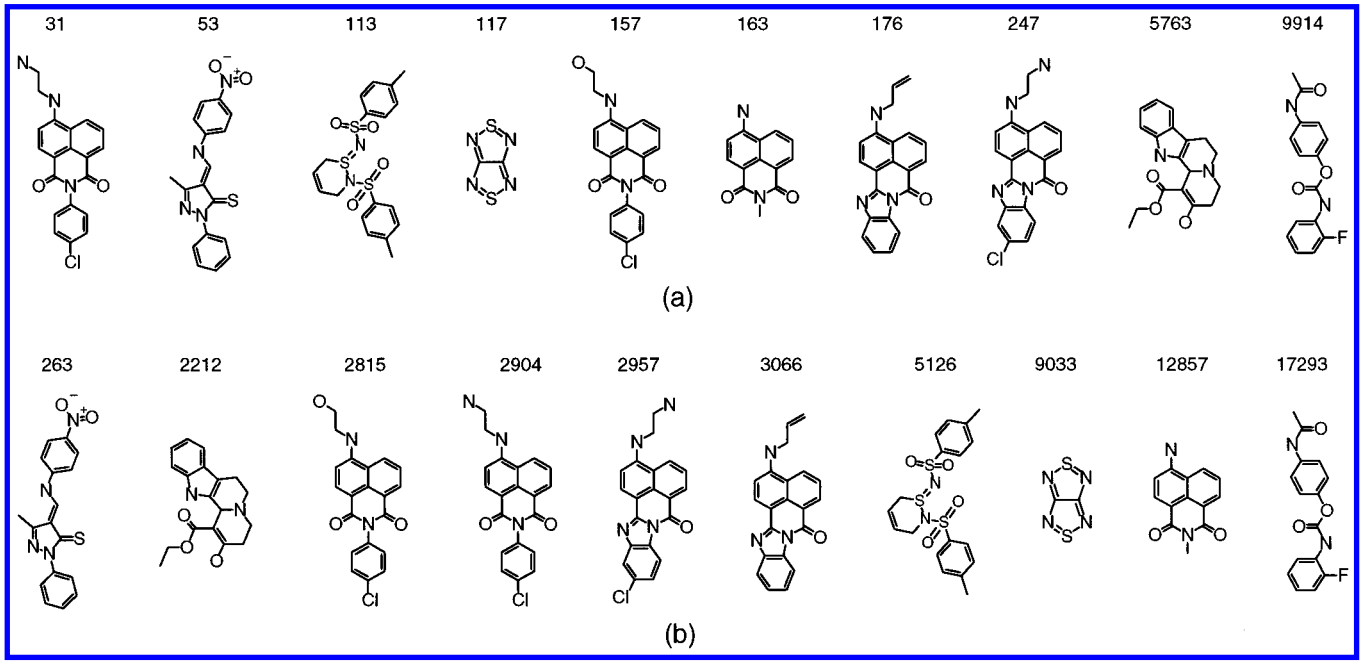

Figure 13. Rational sampling sequence of hits for papain: (a) affinity fingerprints and (b) MOLSKEYS. The cumulative number of compounds screened to find each active is indicated. These results correspond to the edge training sets in Figure 12b.

leads are selected, then lower potency compounds may end up being incorporated, and this could result in unproductive searching in bioactively irrelevant regions of compound space.

Finally, in Figure 12, head-to-head comparisons between affinity fingerprints and the MOLSKEYS are presented. Results from both spread and edge training set designs are included for the MOLSKEYS, and activity-biased scaling is used throughout. The overall success rate of the MOLSKEYS is really no better than random sampling when a spread-diverse training set is used. Performance is significantly improved, however, by the use of an edge design, but the hit rate is still far behind that of affinity fingerprints.

There are regions in the MOLSKEYS curves where several active structural analogs are found in rapid succession. The sequence of hits in Figure 13 reveals that the MOLSKEYS identify four structurally similar papain actives during the course of screening only 252 compounds. However, the existence of multiple chemical families within the sets of active compounds for both targets causes this type of success to be sporadic. Thus, as employed here, the MOLSKEYS do not appear to offer the kind of information that is required in order to develop library-wide models of activity. If the different bioactive chemical families were properly represented in the MOLSKEYS training sets, then these descriptors would probably be much more effective rational sampling tools with respect to these targets.

CONCLUSIONS

An affinity fingerprint is a bioactive profile of a compound which is obtained by measuring its binding affinity against a reference panel of diverse proteins. These biologically-based descriptors provide a direct measure of whether or not a compound possesses features that are essential for binding in a wide variety of environments. Because proteins share certain binding site characteristics, the information encoded by affinity fingerprints is relevant to the biology of any target that is sufficiently similar to any of the panel proteins.

As a tool for molecular diversity, affinity fingerprints show a pronounced ability to select subsets of compounds which exhibit higher than average activity against targets that are related to proteins in the panel. Use of an edge design to select diversity appears to offer a significant advantage over a uniform distribution of compounds in affinity fingerprint space. This is a consequence of the fact that targets which share binding site characteristics with the panel proteins will tend to share some of the same high affinity ligands. These ligands, in turn, appear on the outer edges of affinity fingerprint space.

Rational sampling is an iterative procedure for locating the most active compounds within an existing library. It begins with selection on the basis of diversity and proceeds with a series of focused libraries concentrated around the most promising lead compounds. Affinity fingerprints are shown to be extremely effective in this approach to lead generation and are particularly valuable when the target has not been adapted for high throughput screening. For the examples shown here, the MOLSKEYS structural fingerprints appear to be less effective for rational sampling, partially because the training sets, while structurally diverse, do not sufficiently cover the different chemical families encompassed by the active compounds.

REFERENCES AND NOTES

(1) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(2) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(3) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.

(4) *ISIS/Base 2.1.3*; Molecular Design Ltd.: 14600 Catalina Street, San Leandro, CA 94577.

(5) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying Diversity. In *Combinatorial Chemistry and Molecular Diversity*; Kerwin, J. F., Gordon, E. M., Eds.; John Wiley & Sons: New York, 1997.

(6) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(7) Chapman, D. The Measurement of Molecular Diversity: A Three-Dimensional Approach. *J. Comput.-Aided Mol. Design* **1996**, *10*, 501−512.

(8) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.

(9) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(10) Lin, S. K. Molecular Diversity Assessment: Logarithmic Relationships of Information and Species Diversity and Logarithmic Relations of Entropy and Indistinguishability after Rejection of Gibbs Paradox of Entropy and Mixing. *Molecules* **1996**, *1*, 57−67.

(11) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Chemistry Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(12) Agrafiotis, D. K. On the Use of Information Theory for Assessing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576−580.

(13) Murtagh, F. *Multidimensional Clustering Algorithms*; Physica-Verlag: Heidelberg, 1985; Vol. 4.

(14) 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.

(15) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discov. Design* **1997**, *7/8*, 31−49.

(16) Willett, P. Computational Tools for the Analysis of Molecular Diversity. *Perspect. Drug Discov. Design* **1997**, *7/8*, 1−11.

(17) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

CI980105+