Hit-Directed Nearest-Neighbor Searching

Veerabahu Shanmugasundaram,* Gerald M. Maggiora,‡ and Michael S. Lajiness§

Structural & Computational Chemistry, Pharmacia Corporation, Kalamazoo, Michigan 49007

Received August 6, 2004

This work describes a practical strategy used at Pharmacia for identifying compounds for follow-up screening following an initial HTS campaign against targets where no 3-D structural information is available and preliminary SAR models do not exist. The approach explicitly takes into account different representations of chemistry space and identifies compounds for follow-up screening that are likely to provide the best overall coverage of the chemistry spaces considered. Specifically, the work employs hit-directed nearest-neighbor (HDNN) searching of compound databases based upon a set of "probe compounds" obtained as hits in the preliminary high-throughput screens. Four different molecular representations that generate nearly unique chemistry spaces are used. The representations include 3-D, 2-D, 2-D topological BCUTs (2-DT) and molecular fingerprints derived from substructural fragments. In the case of the BCUT representations the NN searching is distance based, while in the case of molecular fingerprints a similarity-based measure is used. Generally, the results obtained differ significantly among all four methods, that is, the sets of NN compounds have surprisingly little overlap. Moreover, in all of the four chemistry space representations, a minimum of 3- to 4-fold enrichment in actives over random screening is observed even though the actives identified in each of the sets of NNs are in large measure unique. These results suggest that use of multiple searches based upon a variety of molecular representations provides an effective way of identifying more hits in HDNN searches of chemistry spaces than can be realized with single searches.

1. Introduction

Today, the notion of performing nearest-neighbor (NN) searches to follow-up "hits" obtained in high-throughput screening (HTS) campaigns, so-called hit-directed nearest-neighbor (HDNN) searching, is almost second nature to scientists in the pharmaceutical industry.[1−5] Such NN searches can be accomplished using either the similarities of or distances between pairs of compounds located in a chemistry space as a measure of their "neighborness," namely whether a given molecule is a NN, a next-NN, a next-next-NN, etc., with respect to a given probe, "seed", or query molecule. In contrast to many other search methods such as those based upon substructure[6] and pharmacophore[7] queries, similarity- or distance-based NN searches take a more "subjective" approach in that they attempt to identify molecules that are "similar to" or "close to" the query molecule with respect to a set of molecular fragments, or to a set of geometric, electronic, physicochemical, or topological properties, to name a few. Moreover, depending upon the representation used, either 2-D or 3-D structural information can be encoded so that the similarities or distances will also reflect, either implicitly or explicitly, features associated with the 2-D or 3-D structures of the molecules. Comprehensive compilations covering virtually all classes of descriptors now exist.[8,9] In similarity- or distance-based HDNN searching there are no definitive answers as the similarity or distance values depend on both the type of molecular representation (vide supra) and similarity or distance measure used.[10]

In HDNN searching one need not know what parts of the molecule confer activity, as is the case in both substructure-based and pharmacophore-based searching. In both these cases a significant amount of quality assay data is required in order to develop substructure or pharmacophore-based models that faithfully capture the underlying structure−activity relationships (SAR). This is a decided disadvantage compared to HDNN searching, which does not require such specific molecular information, especially in the early phases of drug discovery where detailed structure−activity relationships are usually unavailable. Virtual screening ("high-throughput docking") of large, electronic compound collections offers an alternative approach that does not require a preliminary "structural model," but does require detailed 3-D structural information on both the ligand and protein target. While some success has been obtained in such studies, the methodology remains problematic today.[11,12]

It is well-known that chemistry spaces are representation dependent.[10,13] As a result, relationships among compounds in one chemistry space are not necessarily preserved in another chemistry space. Thus, *an intrinsic chemistry space does not exist*, and this has important consequences with regard to the distribution of compounds in these spaces. For example, it is entirely possible that clusters of compounds in one chemistry space may become uniformly spread out in another chemistry space and vice-versa. An important point for

* To whom correspondence should be addressed. Current address: Computer-Assisted Drug Discovery, Pfizer Global Research & Development, 2800 Plymouth Rd., Ann Arbor, MI 48105. Phone: 734-622-7131. Fax: 734-622-2782. E-mail: Veerabahu.Shanmugasundaram@pfizer.com.

‡ Current address: Department of Pharmacology & Toxicology, University of Arizona, College of Pharmacy, Tucson, AZ 85721.

§ Current address: Lilly Research Laboratories, Indianapolis, IN 46285.

10.1021/jm0493515 CCC: $30.25
Published on Web 12/16/2004

the work described here is that NN relationships may not be the same in the different spaces.

A well-known principle that is often used in searching for active compounds is that "similar compounds have similar activities".[1,14] While not uniformly true, due to the underlying differences in the nature of the activity landscapes,[15] it still holds in enough cases that similarity-based HDNN searching has become a well-accepted way of finding additional "interesting compounds" based on a known lead, hit, or series of hits.

These confounding factors, namely that different chemistry-space representations lead to different distributions of compounds and that significant violations of the similarity principle occur, have led to the realization that the quest for the *best* computational technique in NN searching of compound databases may be a futile exercise.[13] Several researchers[16,17] in the field have, hence, developed and applied a variety of novel computational tools to mitigate some of these representation-dependent and similarity or distance-biased views of chemistry space.

In the current work a *practical procedure* is described for identifying compounds for follow-up screening against targets for which 3-D structural information and preliminary SAR are unavailable. The procedure, which was developed and used at Pharmacia over the last several years,[18] explicitly takes account of the multiplicity of chemistry spaces in a way that identifies many more active compounds than are likely to be found searching a single chemistry space. The results reported here, which were gathered over a period of time and in collaboration with several therapeutic-area project teams, tend to indicate that HDNN searching over multiple chemistry spaces tends to select sets of compounds with very small overlaps. In addition, it will be shown that the enrichment of actives found in each of the chemistry spaces is, in most instances, roughly the same, a minimum of 3−4-fold above background for primary HTS.

The procedure employed in this work utilizes 3-D, 2-D, and 2-D topological (2-DT) BCUTs and a fragment-based molecular fingerprint method. BCUT descriptors were developed in the laboratory of Professor Robert Pearlman at the University of Texas[19] and implemented in the program DiverseSolutions (DVS),[20] which is used for the calculations reported in this work. BCUTs encode information about the electrostatic, hydrophobic, and hydrogen-bonding characteristics of molecules and are defined in a manner that incorporates distance information based upon through-bond or through-space interatomic distances and atomic properties relevant to intermolecular ligand-protein interactions. BCUTs have been repeatedly shown to be useful as descriptors for describing chemistry spaces.[21−26] BCUT values are calculated from matrices consisting of atomic properties as the diagonal elements, connectivity-related properties as the off-diagonal elements, and a scaling factor, which balances the two types of structural information. For example, 'bcut_haccept_S_invdist_000.500_R_H.bdf' refers to the highest eigenvalue (H) of a matrix formed after removing hydrogens (R), with fractional surface area-weighted (S) h-bond-acceptor-ability (haccept) on the diagonal and 0.5 (000.500) times the inverse-distance (invdist) as the off-diagonal elements. Different definitions for the off-diagonal elements differentiate 3-D from 2-D and 2-DT BCUTs. 3-D BCUTs use through-space distances between atoms as the off-diagonal elements, whereas 2-D BCUTs uses Burden numbers[27] as the off-diagonal elements and 2-DT BCUTs uses topological distances as the off-diagonal elements.[20]

The fragment-based molecular-fingerprint method was developed at Pharmacia,[28] but is similar in approach to most molecular fingerprint methods that are available today.[6] Each compound is represented by a 320-component binary vector that contains a combination of atom and bond count information and information on the presence or absence of substructural fragments. The overall procedure has been largely automated and was made available to the general scientific community through Pharmacia's cheminformatic software engine, ChemLink.[28] Nearest-neighbor searches carried out with BCUTs are based upon distance, while those carried out with molecular fingerprints are based upon Tanimoto similarity.[6]

2. Methods

2.1. General. Although references are made to the PRCC (Pharmacia Research Compound Collection), searches were also carried out, where appropriate, on commercial databases of interest to therapeutic-area project teams. Prefiltering of databases, if performed, was based on substructural filters or property filters in tune with our compound purchasing strategy.[29] In addition to the PRCC, ChemLink provides access to a large variety of chemical, structural (e.g., 2-D structure, structure alerts, etc.), and biological information related to the entire compound collection, which includes compounds for which there is no physical inventory (i.e., electronic structures only), compounds that are part of combinatorial libraries, and compounds that contain structural alerts, etc. An SDFile containing structures of compounds with chemical inventory was generated from ChemLink. Only the largest molecular component, which effectively eliminated counterions, was considered for each compound registered in the database.

2.2. BCUT Representations of Chemistry Space. The BCUT descriptors were calculated using DiverseSolutions 4.0.9.[20] A set of 73 standard BCUT descriptors (29 3-D hydrogen suppressed, 18 2-D hydrogen suppressed and 26 top-D (2-DT) hydrogen suppressed) were computed. Three BCUT chemistry spaces, 3-D, 2-D, and 2-DT, respectively, that best represent the PRCC were generated using the χ^2 algorithm implemented in DVS and used as references for all NN distance-based searching. The chemistry spaces used in this work include a six-dimensional 3-D BCUT chemistry space, a five-dimensional 2-D chemistry space, and a five-dimensional 2-DT chemistry space.[30−32]

2.3. Nearest-Neighbor Searching. Nearest-neighbor searching can be carried out using either distance or similarity as a means for identifying the neighboring compounds of a specific query/probe compound. For each given measure there are two ways to carry out the search: (1) by identifying a fixed number of neighbors ("number-based") with respect to either measure or (2) by identifying all neighbors within a given distance or similarity ("distance- or similarity-based") to the probe compound. List-based searches can be applied in those

cases where HDNN searching is carried out with respect to more than one probe compound. Rather than searching a given distance (similarity) or identifying a fixed number of compounds with respect to each probe, list-based searching identifies the set of compounds that are closest (most similar) to at least one of the compounds in the set of probe compounds (i.e., list). Thus, probes whose NNs are far removed, a low-density situation, are unlikely to furnish many compounds to the search set. On the other hand, probes that are located in regions of high-compound density will tend to contribute a significant number of compounds to the NN search set; such a procedure could be called "density biased." List-based searches can be either distance- (similarity) or number-based.

In list-based HDNN searches carried out in BCUT chemistry spaces with DVS, the number of compounds obtained, n_{NN}, typically varied from 50 to 500 depending on the screen. In all cases reported here, receptor-relevant subspaces[33] of the respective BCUT chemistry spaces could not be determined, as the number of well-characterized actives was too small. Thus, "native" 3-D, 2-D, and 2-DT BCUT chemistry spaces based upon the PRCC were used in all HDNN searches.

A list-based, density-biased procedure employing molecular fingerprints and Tanimoto similarity, as implemented in ChemLink, was also used to carry out the NN searches. The program Dfragall[34] computes all pairwise Tanimoto similarities and selects the k NNs closest to the set of active compounds (i.e., probes). Selections were made such that a compound was selected if it was close to "any" of the active compounds.

2.4. Aggregation Procedures. Combining Results from Multiple Chemistry Spaces. In both retrospective and prospective studies, different directed-screening procedures tend to yield different subsets of active compounds for the same biological target. Furthermore, a given procedure tends to work better on some targets than on others in ways that are difficult to predict a priori. Thus, based upon a substantial amount of screening data obtained from Pharmacia and from the arguments advanced earlier concerning the lack of invariance of different representations and their associated chemistry spaces,[10,13] it does not appear that any single approach to directed screening in general and to HDNN searching in particular can unequivocally identify compounds that are similar to active compounds obtained in screening studies. Thus, it is expected that *consensus search methods*, which tend to identify compounds obtained by a set of HDNN search methods, are unlikely to be appropriate in this work. The approach described in this paper falls under the rubric of *aggregation procedures.* As such, it does not utilize any data-fusion procedures such as sum-rank, mean-rank, or best-in-N fusion,[16] or consensus approaches such as those based upon conditional probabilities[35] or any machine learning boosting algorithms.[36] Rather, the results obtained by all of the searches are simply aggregated, a procedure akin to set-theoretical union. This procedure provides a practical, convenient, and rapid means, based upon HDNN searching, of identifying sets of compounds for follow-up screening. As will be seen the aggregation procedure described here is able to locate a far greater number of different types of active compounds from HDNN searches over multiple chemistry spaces than a comparable search over a single chemistry space. This occurs partly because intersection among the sets obtained by the various HDNN searches tends to be quite small (vide infra), which is not entirely surprising since, different representations can lead to dramatically different chemistry spaces. Nonetheless, it provides an unexpected bonus that makes aggregation procedures worthy of further evaluation. The present work is a preliminary attempt to carry out such an evaluation.

2.5 Aggregation Procedures. Mathematical Framework. The above scheme can be formulated mathematically as follows. Consider the "universe of compounds" U, which in the present work is the set of compounds available for screening in the PRCC. Select using, for example, random or diversity-based sampling, a large subset of compounds $\bar{U}$ from U, where the number of compounds in $\bar{U}$, given by $N(\bar{U})$, can be quite large, on the order of 50 000−100 000 to in some cases more than one million compounds. This constitutes the *primary screening set*. The compounds in $\bar{U}$ are then screened, yielding a set of actives, $\bar{U}^*$, where the following subset relationship, $U \supseteq \bar{U} \supseteq \bar{U}^*$, holds, although in practice all of the subsets are proper subsets.

The *background hit-rate* (in percent) for $\bar{U}$ is given approximately by

$$\mathscr{H}_{\text{background}}(\bar{U}) = \frac{\mathscr{N}(\bar{U}^*)}{\mathscr{N}(\bar{U})} \times 100 \qquad (1)$$

where $\mathscr{N}(\bar{U})$ and $\mathscr{N}(\bar{U}^*)$ equal, respectively, the number of compounds in the primary screening set and the number of actives ("hits") obtained in the primary screen. Since the compounds in $\bar{U}^*$ are the probe molecules, $\mathscr{N}(\bar{U}^*)$ is also equal to the number of probe molecules, which is the basis for the list-based HDNN searches use here to identify compounds for follow-up screening.

The HDNN searches are performed over the set of compounds that were not screened previously, $\bar{U}^c = U - \bar{U}$, where $\bar{U}^c$ is the complement of $\bar{U}$, which is equivalent to the difference set $U - \bar{U}$, that is the set of compounds in U that are not also in $\bar{U}$. This constitutes the *search set*. As discussed earlier, different representations of chemistry space yield different subsets of compounds from HDNN searches carried out with the same set of active/probe molecules, $\bar{U}^*$, over the same search set, $\bar{U}^c$. This can be expressed as the "NN function" f_{NN} that *maps* compounds from the search set $\bar{U}^c$, represented in their i-th molecular representation, into an appropriate subset S_i

$$f_{NN}: \bar{U}_i^c \rightarrow S_i \text{ for } i = 1,2,3,...,n_{\text{search}} \qquad (2)$$

where n_{search} is the number of HDNN searches performed and the number of compounds in each of the subsets approximately satisfies

$$\mathscr{N}(S_i) \approx n_{\text{total}}/n_{\text{search}} \text{ for } i = 1,2,3,...,n_{\text{search}} \qquad (3)$$

In the examples presented in this work the subsets S_i show surprisingly little "overlap" with each other; this means that the number of compounds in the intersection

of two sets is much less than the number in either of the sets, that is

$$\mathcal{N}(\mathrm{S}_i \cap \mathrm{S}_j) \ll \mathcal{N}(\mathrm{S}_i) \approx \mathcal{N}(\mathrm{S}_j) \tag{4}$$

for $i = 1, 2, 3,..., n_{\mathrm{search}}$ and $j > i$. This can be generalized to ternary and higher-order set intersections by noting that in the case of three subsets, for example,

$$\mathcal{N}(\mathrm{S}_i \cap \mathrm{S}_j \cap \mathrm{S}_k) < \begin{cases} \mathcal{N}(\mathrm{S}_i \cap \mathrm{S}_j) \\ (\mathcal{N}\mathrm{S}_i \cap \mathrm{S}_k) \\ (\mathcal{N}\mathrm{S}_j \cap \mathrm{S}_k) \end{cases} \tag{5}$$

The subsets are then "unioned"

$$\mathrm{S} = \mathrm{S}_1 \cup \mathrm{S}_2 \cup \cdots \cup \mathrm{S}_i \cup \cdots \cup \mathrm{S}_{n_{\mathrm{search}}} \tag{6}$$

which ensures that all of the compounds in S are *unique*. The compounds in S are then screened yielding a set of actives S*. In analogy to eq 1, the *net hit-rate* (in percent) for HDNN-based follow-up screening is given by

$$\mathcal{H}_{\mathrm{follow-up}}(\mathrm{S}) = \frac{\mathcal{N}(\mathrm{S}^*)}{\mathcal{N}(\mathrm{S})} \times 100 \tag{7}$$

and the *net enrichment* (in "statistical notation") is given by the ratio

$$\mathcal{E}(\mathrm{S}|\bar{\mathrm{U}}) = \frac{\mathcal{H}_{\mathrm{follow-up}}(\mathrm{S})}{\mathcal{H}_{\mathrm{background}}(\bar{\mathrm{U}})} \tag{8}$$

The qualifier "net" is used to emphasize that values computed by eqs 7 and 8 refer to sets of unique compounds screened, S, and unique actives derived from this set, S*.

It is possible and sometimes desirable to consider the enrichment gained for each subset obtained by HDNN searches. In this case, the background hit-rate remains unchanged but eqs 7 and 8 become, respectively,

$$\mathcal{H}_{\mathrm{follow-up}}(\mathrm{S}_i) = \frac{\mathcal{N}(\mathrm{S}_i^*)}{\mathcal{N}(\mathrm{S}_i)} \times 100 \tag{9}$$

and

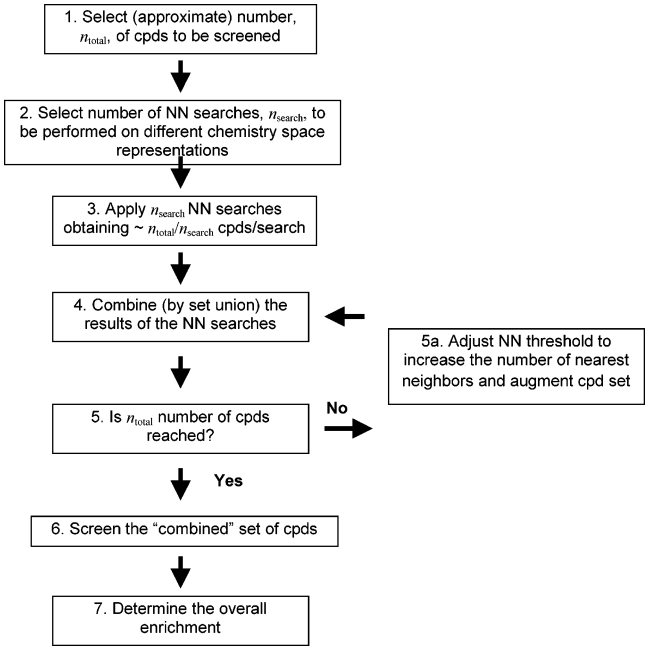
$$\mathcal{E}(\mathrm{S}_i|\bar{\mathrm{U}}) = \frac{\mathcal{H}_{\mathrm{follow-up}}(\mathrm{S}_i)}{\mathcal{H}_{\mathrm{background}}(\bar{\mathrm{U}})} \tag{10}$$

for $i = 1,2,3,...,n_{\mathrm{search}}$. Because of the set intersections

$$\mathcal{E}(\mathrm{S}|\bar{\mathrm{U}}) \le \sum_{i=1}^{n_{\mathrm{search}}} \mathcal{E}(\mathrm{S}_i|\bar{\mathrm{U}}) \tag{11}$$

2.6 Aggregation Procedures. Overall Scheme. The aggregation procedure employed in this work is based upon Scheme 1, which graphically summarizes the overall process. Choosing the total number of compounds to be screened, n_{total} (Step 1), depends on a number of factors such as compound availability, screening capacity and plans, needs of therapeutic-area projects, etc. Once this choice is made, the number of HDNN searches must be determined (Step 2), which depends on the number and type of chemistry spaces desired for exploration (N.B. this involves the same compound collection represented in a number of different ways). The HDNN searches are carried out (Step 3), and the resulting sets of compounds are combined (Step 4) using set-theoretic union, which ensures that the combined set contains only *unique* compounds. Generally, it is observed that the number of compounds common to two or more sets, as measured by the ratio of the number of compounds of their set intersections to the number of compounds of their set unions (N.B. that this measure is identical to the Tanimoto similarity measure used in the molecular-fingerprint-based NN searches carried out in this work) is surprisingly small. If the similarity among the sets is large, additional compounds can be added until the total number of compounds to be screened, n_{total}, is reached (Steps 5, 5a, 4, and 5). The compounds are then screened (Step 6) and the overall enrichment of hits is determined (Step 7).

Scheme 1. Aggregation Procedure. Overall Scheme



3. Results and Discussion

3.1. General Considerations. All the results reported here were gathered from *active* therapeutic-area projects when the assays were actually run and do not involve any retrospective analysis. For a variety of experimental reasons, such as plate-to-plate assay variability, variability in the signal of the control wells, differences in samples, solubility issues, artifacts in the detection technique, etc., the biologists who perform a particular assay are most likely the best informed to assess the data. Thus, in this work the designation of compounds as active was based solely on the judgment of the biologists who performed the assays. This distinction is important in that many times what is considered to be an active compound varies with who performs the experiment and who analyzes the data. In all studies reported here activity criteria have been applied consistently for both the primary screen and the follow-up screens.

Seven different types of assays are considered in this work, and these include (1) a bacterial enzyme target

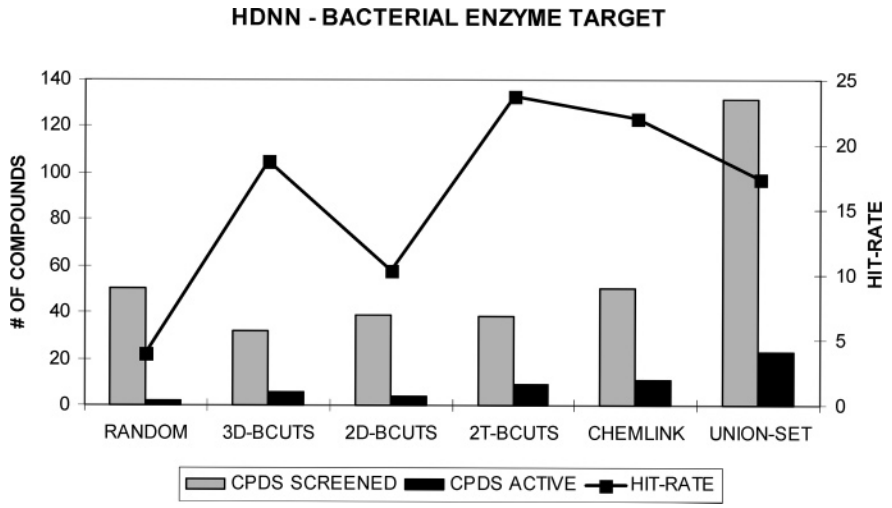


Figure 1. A graphical depiction of the number of compounds screened (grey bars) compared to the number of actives found (black bars) plotted with respect to *Y*-axis (LHS of figure). The hit-rate (black line) for sets (including the union-set) of NNs based on different chemistry space representations is also plotted with respect to the *Y*-axis (RHS of figure).

important in the recycling of certain protein synthesis factors, (2) a fungal enzyme target that is critical to a lipid biosynthesis pathway and is vital to the growth and viability of fungal cells, (3) a fungal whole cell assay that identifies compounds toxic to fungal cells, (4) a bacterial transporter assay that identifies compounds that inhibit the proper functioning of the transporter, (5) a CNS assay that characterizes the formation of Alzheimer's plaque, (6) a GPCR antagonist screen for treatment of anxiety, and (7) a panel of four different nuclear hormone receptors and several of their isoforms assayed for agonist and antagonist activity. In addition, the quality and type of data obtained for the assays varies as some of the assays are 'HTS-like' that provide percent inhibition values, while some are well-characterized concentration-dependent kinetic assays that provide IC_{50} or K_i values.

3.2. Bacterial Enzyme Target. This study provides a prototypical example of studies of HDNN searching and follow-up screening. In this study, an antibacterial therapeutic-area project team had evidence that the activity of a particular enzyme ensures the recycling of certain key protein-synthesis factors. Inhibiting this activity would lead to an accumulation of these factors, thereby impairing the initiation of translation, thus reducing the rate of protein synthesis to an extent that is deleterious to the bacterial cell.

Based on an assay developed in-house, a HTS campaign was undertaken. The compounds screened in the initial HTS, which totaled about 30 000 compounds, included a diverse subset of the PRCC obtained by a dissimilarity selection protocol,[34] recently purchased and plated commercially acquired compounds, and a focused, bacterial whole-cell active (bacteriocidal or bacteriostatic) library. After several levels of confirmatory testing and analysis, 12 'well-validated hits' were identified. The project team decided to carry out follow-up screening based upon these 'hits.' About 200 compounds were desired for the follow-up screen. Approximately 50 compounds were obtained by HDNN searches of each of four different chemistry-space representations of the PRCC generated, respectively, by 3-D, 2-D, and 2-DT BCUTs and by a set of fragment-based molecular fingerprints developed in-house.[28] The search was directed only toward that part of the PRCC not screened in the initial HTS, namely $\bar{U}^c$. Of the nearly 200 compounds obtained by the four HDNN searches, only 132 were available for screening. Thus, $\mathcal{N}(S) = 132$, and the number of compounds in each of the specific subsets, is given by $\mathcal{N}(S_{3-D}) = 32$, $\mathcal{N}(S_{2-D}) = 39$, $\mathcal{N}(S_{2-DT}) = 38$, and $\mathcal{N}(S_{MF}) = 50$, whose sum is given by $\mathcal{N}(S_{3-D}) + \mathcal{N}(S_{2-D}) + \mathcal{N}(S_{2-DT}) + \mathcal{N}(S_{MF}) = 159$. Of the compounds available for screening, only about 10−15% were common to more than one set, which is manifested by the near equality between the number of compounds in S and the sum over the number in each of the individual subsets. This is clearly a manifestation, as discussed in Sections 1 and 2, of the fact that different molecular representations generate different chemistry spaces with no guarantee that NN relationships will be preserved.

Of the 132 compounds screened, 23 were found to be active, that is $\mathcal{N}(S^*) = 23$. From eq 7 the net hit-rate, $\mathcal{H}_{follow-up}(S)$, is calculated to be 17%, yielding a net enrichment, $\mathcal{E}(S|U')$, of 4.35, which is the "true" measure of the overall enrichment obtained irrespective of overlaps among subsets. The number of actives found in each of the subsets is $\mathcal{N}(S^*_{3-D}) = 6$, $\mathcal{N}(S^*_{2-D}) = 4$, $\mathcal{N}(S^*_{2-DT}) = 9$, and $\mathcal{N}(S^*_{MF}) = 11$, which sum to 30, a number that is about 1.3 times as large as the number of unique actives, $\mathcal{N}(S^*) = 23$. This again shows that there is no appreciable overlap among the sets of actives. The corresponding subset and union-set hit-rates and enrichments, which can be calculated from eqs 9 and 10, respectively, are graphically depicted in Figure 1.

From the figure it is clear that the HDNN searches derived from HTS hits lead to measurable increases in the hit-rate over the baseline value of four percent observed in the primary screen, for all of the NN subsets from a low of 10.3% to a high of 23.7%. These values correspond to enrichments of 2.6 to 5.9, respectively, values that are in accord with the net hit-rate and net enrichment values given above. Whether the use of additional chemistry-space representations will continue to yield new compounds and similar enrichments is unknown at this time. Nevertheless, the above

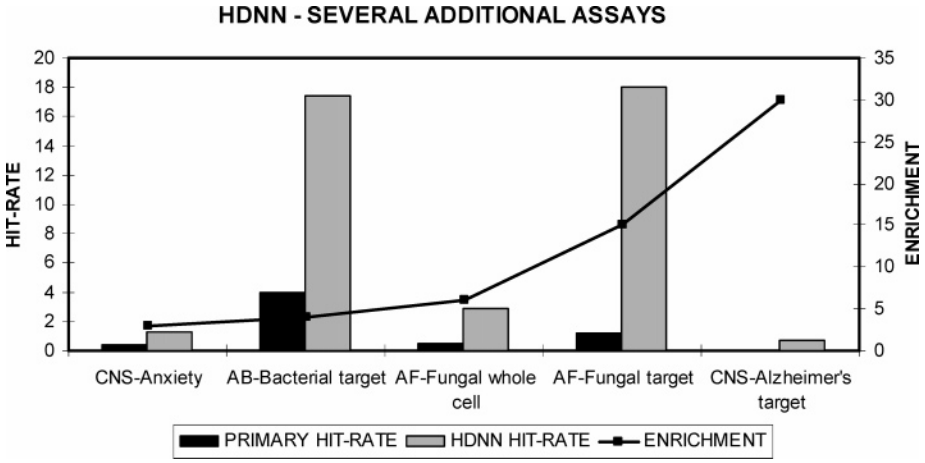


Figure 2. A graphical depiction of the enrichment of hit-rates observed in five different assays. Hit-rates are given along the *Y*-axis (LHS of figure) for initial screens of compounds (black bars) and for screens of compounds obtained by HDNN searches (grey bars). Enrichments are given on the *Y*-axis (RHS of figure) and are indicated by the black line.

example clearly demonstrates the advantage of exploring multiple chemistry-space representations for potential active compounds.

3.3. Several Additional Targets. Studies were carried out on a variety of other targets using the list-based, HDNN searches of multiple chemistry-spaces. Based on the type of assays that were developed in-house, the viability of setting up a HTS campaign, and the priority of the particular target in the therapeutic-area portfolio, different types of primary-screening campaigns were carried out. For instance, if a target is not amenable to HTS or is considered to be of lower priority to the therapeutic area under consideration, a subset-based screening approach is typically adopted. In such cases, a subset of compounds is usually obtained from the PRCC through a dissimilarity-based procedure[34] and used as the initial screening set. In many cases, additional subsets were derived in a similar manner from other compound collections such as those containing recently purchased compounds and those containing focused, target-based libraries. In aggregate these screens tend to ensure that a reasonably broad coverage of chemistry space has been obtained. Based upon the results obtained, HDNN searching is then performed against the remainder of the PRCC not screened initially, $\bar{U}^c$, to identify additional compounds for follow-up screening.

In cases of high-priority targets amenable to HTS, a screen of the entire PRCC is usually performed. In such cases, HDNN searching and follow-up screening is based entirely on either commercial sources, although additional passes through selected regions of the PRCC are sometimes carried out to identify borderline actives and/or false negatives. For the purposes of the current discussion, it does not matter what the source of the screened compounds is, what assay the compounds are screened against, what the goal of a particular screening campaign was, or how many initial seeds were used as probe compounds in an HDNN search. What matters is the ability of the HDNN search algorithms and the various chemistry-space representations to identify more active compounds than would be identified using any single chemistry-space representation and NN search algorithm.

The results obtained for targets (2) through (6) are summarized in Figure 2. Hit-rates for the primary screen varied from 0.02% for the CNS Alzheimer's target to 4% for the antibacterial target, while those for the corresponding follow-up screening of sets of compounds obtained by HDNN searching varied from 0.7% to 17%, respectively. The enrichment ratios varied from a minimum of 3-fold to a maximum of 30-fold and overlap between BCUT chemistry space nearest neighbors and ChemLink FP based similar compounds varied from 10 to 15%.

Again, it is clear from the figure that there is a considerable enrichment in the actives found by screening compounds obtained by HDNN searching, with values ranging from ~30 in the case of the CNS Alzheimer's target to a low of ~3 for a CNS anxiolytic target. In all cases, the value of the enrichment is always at least three times greater than background. This shows, not surprisingly, that HDNN searching appears capable of identifying additional sets of compounds for screening that contain a significantly higher fraction of active compounds than is the case for primary screening.

3.4. Nuclear Hormone Receptor Targets. A panel of four different nuclear hormone receptors (NHR_1, NHR_2, NHR_3, NHR_4) and two isoforms of NHR_4 were assayed for agonist and antagonist activity. The initial screening set of compounds, $\bar{U}$, was obtained from the PRCC using our 'standard' dissimilarity-selection protocol.[34] Some additional known NHR ligands were also included in the screening set. The set of molecules identified as hits (agonists and antagonists), $\bar{U}^*$, were then used as queries/probes in multiple HDNN searches of the unscreened portion of the PRCC, $\bar{U}^c$. As was done before, the four subsets of compounds obtained from the HDNN searches were aggregated by taking their set-theoretic union, $S = S_1 \cup S_2 \cup S_3 \cup S_4$, and the resulting set of compounds was screened yielding the "hit set," S*. The results are summarized in Figure 3 for each of the NHR assays. The figure graphically shows a comparison of the background HTS hit-rate, $\mathscr{H}_{\text{background}}(\bar{U})$, given by the dark bars, to the enhanced follow-up hit-rates of HDNN, $\mathscr{H}_{\text{follow-up}}(S)$, given by the light gray bars. In all cases, the results of follow-up screening are superior to those obtained in the primary HTS, and the

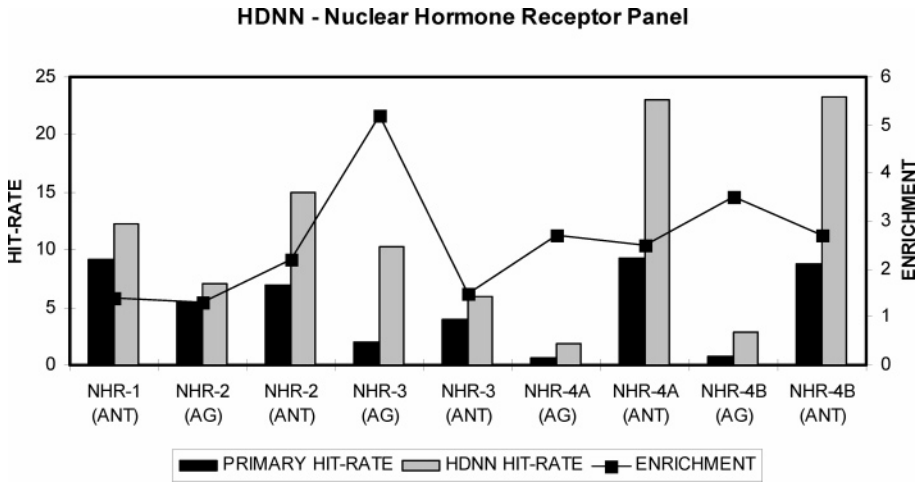


Figure 3. A graphical depiction of the enrichment of hit-rates observed in nine different NHR assays. Hit-rates are given along the *Y*-axis (LHS of figure) for initial screens of compounds (black bars) and for screens of compounds obtained by HDNN searches (grey bars). Enrichments are given on the *Y*-axis (RHS of figure) and are indicated by the black line.

overlap between the different methods was again very small. The hit-rates of the primary screen for the panel of assays varied from about one to nine percent and the hit-rates for the corresponding follow-up directed screen varied from about two to more than 20%, respectively, with enrichments in the range of from slightly more than one to more than five and an average enrichment of slightly more than 2.5. In the latter case of the largest enrichment, namely NHR_3, this was due to the discovery of a novel class of compounds that had potent agonist activity.

3.5. Iterative HDNN Searching and Screening. The question naturally arises as to whether a single follow-up screen of n_{total} compounds obtained by HDNN searching provides higher enrichment, as well as better coverage and more novel scaffolds than is obtained by a number, n_{search}, of follow-up screens of sets of approximately n_{total}/n_{search} compounds each obtained by HDNN searches of multiple chemistry-space representations of the PRCC or any other compound collection. If, rather than being radially dispersed about an initial, clustered set of actives as is typically assumed, the unknown actives lie along a "trend vector" in chemistry space, a single HDNN search could miss a significant number of potentially active compounds. Rather, even if the actives are found, such a strategy is likely to retrieve a large number of inactive compounds in order to find most actives, lowering the enrichment. Such an approach requires taking a very large search radius or very large sample. On the other hand, iterative HDNN searching can, in some cases, ameliorate this situation by sequentially moving through chemistry space in manner that is directed by the results of previous steps in the process. Various iterative or sequential screening strategies have been reported in the literature.[37] The application of these strategies vary depending on the hit or lead information available and on the quality of structural and biological information available.[37−41] Figure 4 depicts a single such scenario for a particular fungal cell-based assay. The results of two such cycles of iterative, list-based HDNN searching for a fungal cell assay following the first iteration of the primary screen are shown. As was the case in all of the studies presented in this work, multiple chemistry space NN

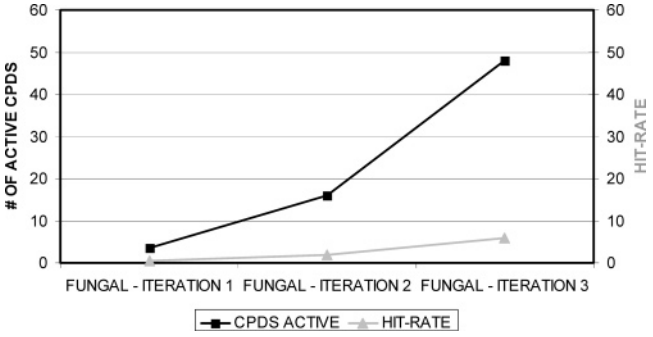


Figure 4. A graphical depiction of the results of an iterative searching and screening process for an antifungal target.

searches were carried out using the four (three BCUT and one molecular fingerprint-based chemistry space representations). About 800 compounds were screened in each of two successive iterations, yielding 16 actives for a hit-rate of about two percent and a 4.4-fold enrichment above the primary screen in the second iteration, and in the third iteration about 48 compounds were found to be active. This yielded a six percent hit-rate and a 13-fold enrichment above the primary screen. Interestingly, the stepwise approach used here lead to the identification of an interesting class of compounds that were both novel and active.

4. Summary and Conclusions

The present work describes applications of HDNN searching to active drug-discovery projects carried out at Pharmacia in order to better focus follow-up screening efforts. A persistent issue associated with HDNN searches is their dependence on the chemistry-space representation, as has been described in the literature by a number of authors (vide supra). An important consequence of this lack of invariance is that NN relationships are not generally preserved; that is, two compounds that are NNs in one chemistry-space representation may not even be close to another. Although it may be a bit unsettling, it is possible to use this lack of invariance to advantage through the use of HDNN searches of multiple chemistry-space representations of the same compound collection. The results described in this work clearly show that aggregating the results of

such HDNN searches of multiple chemistry-space representations provides a significant increase in the total number of *unique* compounds. Moreover, because the amount of overlap among the sets of compounds obtained in the HDNN searches is surprising small, on the order of only 10−15%, the number of compounds obtained in the HDNN searches appears to scale approximately to the number of chemistry-space representations employed, but more work will have to be done before this can be conclusively demonstrated. Importantly, there is significant enrichment, which is nearly constant, in hits over background in most cases. Further, the set of assays examined in this work covers a range of therapeutic-area projects, from CNS to antifungal to antibacterial and contains both cell-based and target-based assays, some of which are functional assays and some of which are binding assays. Thus it appears, at least from a pragmatic viewpoint, that the procedure described here provides a practical method of finding actives "surrounding" a set of hits obtained from a primary HTS.

Acknowledgment. The authors would like to acknowledge several discovery scientists from Pharmacia−Kalamazoo who have contributed immensely to this work, Tom Hagadone for providing the ChemLink connection, Martin Schulz for making the Dfragall program available, Christian Parker, Benjamin Turner, Larry Erickson, Paul Bonin, Larry Fitzgerald, Scott Knauer, Chris Chio, Brian Stockman, Daneen Hadden, Matt Peach, Howard Miller, Alice Laborde, and Mike Bienkowski for all the biological assays and screening results discussed in the paper. The authors would also like to acknowledge Prof. Bob Pearlman for several helpful and enjoyable discussions on BCUTs and DVS.

References

(1) Johnson, M. A., Maggiora, G. M., Eds. In *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(3) Dean, P. M., Ed.; In *Molecular Similarity in Drug Design;* Chapman & Hall: Glasgow, 1994.
(4) Vandrie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274−283.
(5) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.
(6) Leach, A. R.; Gillet, V. J. *An Introduction to Cheminformatics*; Kluwer Academic Publishers: Dordrecht, Neatherlands, 2003.
(7) Güner, O. F., Ed. In *Pharmacophore Perception, Development and Use in Drug Design*; International University Line: La Jolla, 2000.
(8) Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley−VCH: New York, 2000; vol. 11, p 667.
(9) http://www.disat.unimib.it/chm/.
(10) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In Methods in Molecular Biology vol. 275. *Chemoinformatics: Concepts, Methods and Tools for Drug Discovery*; Bajorath, J., Ed.; Humana Press: Totawa, New Jersey, 2004; pp 1−50.
(11) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(12) Lyne, Paul D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7* (20), 1047−1055.
(13) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(14) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(15) Shanmugasundaram, V.; Maggiora, G. M. Characterizing property and activity landscapes using an information-theoretic approach. *Abstracts of Papers*, 222nd American Chemical Society National Meeting, Chicago, IL, Aug 26−30, 2001; American Chemical Society: Washington, DC, 2001; CINF-032.
(16) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−552.
(17) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(18) Shanmugasundaram, V.; Maggiora, G. M. Hit-directed nearest neighbor searching. *Abstracts of Papers*, 227th American Chemical Society National Meeting, Anaheim, CA, Mar 28−Apr 1, 2004; American Chemical Society: Washington, DC, 2004; CINF-065.
(19) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(20) Pearlman, R. S. *DiverseSolutions User's Manual*; University of Texas, Austin, TX, 1995.
(21) Gao, H. Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402−407.
(22) Pirard, B.; Pickett, S. D. Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(23) Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(24) Beno, B. R.; Mason, J. S. The design of combinatorial libraries using properties and 3-D pharmacophore fingerprints. *Drug Discovery Today* **2001**, *6*, 251−258.
(25) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design and compound selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(26) Schnur, D. Design and diversity analysis of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.
(27) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(28) Hagadone, T. R.; Lajiness, M. S. 1993 Integrating Chemical Structures into an Extended Relational Database System. *Proceedings of the 2nd International Chemical Structures in Chemistry Conference*; Warr, W., Ed.; Springer: Berlin, Germany; pp 257−269.
(29) Maggiora, G. M.; Shanmugasundaram, V.; Lajiness, M. S.; Doman, T. N.; Schulz, M. W. A Practical Strategy for Directed Compound Acquisition. In *Cheminformatics Aspects in Drug Discovery*; Oprea, T., Ed.; Wiley-VCH: New York, 2004; pp 317−332.
(30) The six 3-D BCUTs that best represent the structural diversity of compounds contained in the PRCC are Bcut_gastchrg_S_invdist2_001.250_R_L, Bcut_gastchrg_S_invdist6_000.600_R_H, Bcut_haccept_S_invdist_000.600_R_H, Bcut_hdonor_S_invdist2_001.200_R_H, Bcut_tabpolar_S_invdist2_001.000_R_L and Bcut_tabpolar_S_invdist_000.500_R_H.
(31) The five 2-D BCUTs that best represent the structural diversity of compounds contained in the PRCC are Bcut_gastchrg_burden_000.100_R_H, Bcut_gastchrg_burden_000.100_R_L, Bcut_haccept_burden_000.900_R_H, Bcut_hdonor_burden_000.600_R_H and Bcut_tabpolar_burden_000.750_R_H.
(32) The five 2-DT BCUTs that best represent the structural diversity of compounds contained in the PRCC are Bcut_gastchrg_invtopd_000.800_R_L, Bcut_haccept_invtopd2_001.600_R_H, Bcut_hdonor_distDtopd_000.070_R_H, Bcut_tabpolar_invtopd2_000.950_R_H and Bcut_tabpolar_invtopd_003.000_R_L.
(33) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(34) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, 7/8, 65−84.
(35) Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional probability: A new fusion method for merging disparate virtual screening results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−609.
(36) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*; Springer-Verlag: New York, 2001.
(37) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882−894.
(38) Jenkins, J. L.; Kao, R. Y. T.; Shapiro, R. Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins: Structure, Funct. Genet.* **2003**, *50*, 81−93.

(39) van Rhee, A. M. Use of recursion forests in the sequential screening process: Consensus selection by multiple recursion trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941−948.
(40) Jones-Hertzog, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J. Pharmacol. Toxicol. Methods* **1999**, *42*, 207−215.
(41) Engels, M. F. M.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R. CerBeruS: A system supporting the sequential screening process. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 241−245.

JM0493515