Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research

Richard D. Cramer,* David E. Patterson, Robert D. Clark, Farhad Soltanshahi, and Michael S. Lawless

Tripos, Inc., 1699 S. Hanley Road, St. Louis, Missouri 63144

Received March 2, 1998

Virtual compound libraries, descriptions of all of the structures that might be produced by specified transformations involving specified reagents, are especially useful in molecular discovery when suitably fast and relevant searching techniques are available. Issues to be considered include fundamental data structures, neighborhood searching principles, useful searching approaches and techniques, library definition and construction, algorithmic details of library comparison, and user interfaces.

INTRODUCTION

In the iterative make-and-test process which has long characterized most molecular discovery research, ultimate success depends greatly on how effectively one selects each next candidate molecule. Such decisions have traditionally been made by the individual chemists who would be responsible for synthesis. In recent decades, additional guidance might come from computational specialists skilled in the modeling of the physical chemistry of molecular interactions or in the statistical analysis of extant structure−property data. However, the rates of the make-and-test processes, the laboratory experiments, have until recently been slow enough that the molecule selection process itself affected discovery outcomes *only* by its effectiveness, not by how long each decision took.

The "combinatorial chemistry/high-throughput screening revolution", with both synthesis and testing becoming automated, miniaturized, and "parallelized", has major implications for molecular selection as well. Chemists who might once have chosen one or two new synthetic targets each week will soon be expected to make hundreds of such choices every day, while considering many thousands of new structure−activity observations. Yet the effectiveness of molecule selection remains just as important to molecular discovery, inasmuch as the initial combinatorial fantasy of simply "testing everything" remains fantasy. One response has been an eruption of interest in "Virtual Compound Libraries", representations of *all* of the molecules that might easily be obtained by the combinatorial chemistry resources of a discovery organization. Such a virtual library by definition constitutes a general starting point for molecule selection. *However, to be useful in molecule selection, a virtual library must be searchable in timely and relevant ways.*

To our knowledge, no previous descriptions of virtual library technologies have appeared in the literature, so the description that follows is necessarily based on our own experience. This experience does comprise several man decades in software development and usage, from the products Legion[1] and Selector[1] to ChemSpace,[1] which we believe to be the most advanced virtual library technology extant today, but which is not commercially available. Although most of the details of ChemSpace are proprietary, voluminous, and of little general interest, we will draw primarily on this recent and intensive work to propose generalizations about the scope, the uses, and the consequent design requirements of virtual library technologies.

WHAT IS A VIRTUAL LIBRARY?

Essentially, a virtual library may be regarded as a reaction library, within which each reaction description has been extended, first to specify the reagents ("building blocks") that may undergo that reaction, and thence to the products resulting from all possible combinations of those reagents. Figure 1 shows this schematically. The reaction A + B ⟹ C, at the upper left, becomes a component in a virtual library, first by specifying the particular reactants in each class, A1 .. A*i* and B1 .. B*j*, shown at the upper right, and then by forming the array of all the $i * j$ products resulting from every combination of those reactants.

Clipping Rules. Note that the reactants in a virtual library must always change into a different structural form as they become constituents of a reaction product. The change may be as simple as deleting atoms and forming a single bond between the resulting fragments, as in amide bond formation. Another acylation reaction involves a more complex "clipping rule",[2] the formation of ureas by amination of an isocyanate (R′NC(═O)NHR ⟹ R′NH + RN═C═O). At a higher level of complexity, the regio- and stereoisomeric subtleties of the Diels−Alder reaction may best be expressed in terms of the transference of all attached fragments and atoms from the carbon skeletons of butadiene and ethylene to the appropriate free valences of a carbon skeleton of cyclohexene. Obviously, every reaction definition within a virtual library must include its own "clipping rules", in a generic form that is automatically and unambiguously applicable to any appropriate reactant. Typically these rules are applied at "library creation time" so that the "reagents" in a virtual library component actually have already been converted into their "product-fragment" structural forms.

10.1021/ci9800209 CCC: $15.00
Published on Web 07/17/1998

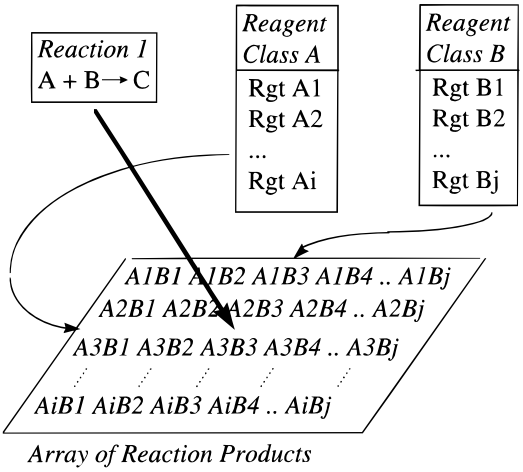

Figure 1. Schematic representation of an entry in a virtual library. Each entry consists of a reaction description, one or more specifically enumerated sets of reagents, and representations of all the combinatorial products.

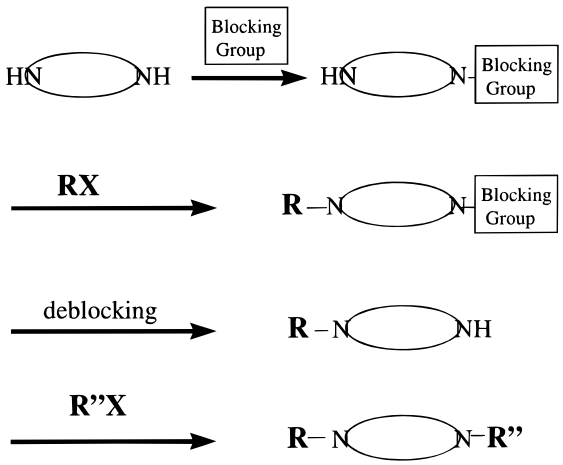

Figure 2. The sequence of synthetic operations that forms any product in the diamine virtual library. The starting point is any of 1700 commercially available diamines (defined as a structure with two nitrogen atoms, each of which are basic and attached to a hydrogen), which here is represented schematically by an ellipse. In the first and most problematic step, one of the nitrogen atoms is made temporarily unreactive, or "blocked", by a group of atoms which can later be removed. The second step introduces a first side chain, labelled R. Approximately 26 700 variations of R are available, comprising various acylating agents, selectively displaceable halides, and carbonyls susceptible to reductive amination. Then the blocking group is removed, and finally a second side-chain R″, drawn from the same pool as R, completes one of the 1700 × 26 700 × 26 700 (10^{12}) distinctive products in the diamine library. Figures 9, 10, 11, and 12 show chemical structures for a few members of this library.

These resulting structures are often called "**clipped reagents**".

Relations between Virtual and Combinatorial Libraries. Figure 1 of course represents an experimentally realized combinatorial library just as well as it does a virtual library. But the large variety of reagents offered commercially makes it almost certain that any actual combinatorial library will be a small subset of its potentially corresponding virtual library. Indeed, one of the more pragmatic but most immediately appreciated uses of virtual libraries is to track the membership of successively synthesized actual libraries within some virtual library.

Usually a synthesized combinatorial library represents a **full** sub-array of its virtual library, in that every reagent chosen from class A has been combined (at least in principle) with every reagent chosen from class B. Such sublibraries can be recorded simply by specifying the selected reagent sets. However there is an increasing emphasis on libraries that represent **sparse** arrays, in which each reagent selected from class A may be combined with a different subset of reagents from class B. Recording these libraries requires that each individual reaction product be specified, which has design implications discussed further later.

It should also be clear that there is no requirement that the reactions in a virtual library be implemented combinatorially, and further that the apparently single "reaction" in a virtual library may represent in reality an arbitrarily long sequence of reactions, or else a completely unrealizable "virtual reaction". It might also be noted that "Markush" structures, exemplified in many patent claims, differ from our definition of a virtual library reaction only in the nature of the "clipped reagent" lists. In a fully constructed, searchable virtual library, these lists are complete enumerations associated with specific synthetic processes, whereas in a typical Markush claim these lists will be as open-ended as legal precedence permits.

In principle, some of the reactants in a virtual library might be taken from the implicit product list of another virtual library. As a simple example, in an amide library, one source of amine reagents might be a reference to the products of a nitration + reduction sequence. However, searching techniques for such "indirect" virtual library products are more challenging to create.

Size of Virtual Libraries. Just how many product structures may be represented by an individual reaction in a contemporary virtual library? Let us consider a particularly prolific reaction scheme involving three reagent classes. Its central reagent class "diamines" contain two basic and derivatizable nitrogens. One nitrogen is to be blocked (selectively, if the diamine is unsymmetric and a single product is intended, which may in practice be very difficult). The second is derivatized with a reagent R (where R is monoreactive, and in this example either an acylating reagent, a cleanly reactive halide, or an imine-forming carbonyl). Then, the first nitrogen is deblocked and derivatized with a second reagent R″ taken from the same R set. This sequence is summarized in Figure 2.

The 1996 ACD (Available Chemicals Directory)[3] offers 1750 varieties of HNXNH and 26 700 varieties of R, yielding a total of 1700 × 26 700 × 26 700 or a bit more than 10^{12} potential products. A significant fraction, over 10%, of drug molecules in clinical use, might be found therein. It may be noted that 10^{12} also represents:

- 2 million typical large pharma compound inventories
- 50 000 × world's cumulative chemical literature[4]
- 30 000 years of testing at 100,000 compounds/day.

The impossibility of "making and testing everything", so as to experimentally overwhelm any need for the step of molecule selection, seems obvious.

The design implication of such very large numbers of product molecules is that *the product molecules within a virtual library should not be represented explicitly*. If 100 bytes are needed to store each product structure, the disk

requirements for an explicit diamine library would be of the order of 10 000 gigabytes, and the library creation time at a rate of 10 000 structures/s would be ~3 years. It then follows that, if the products are not to be explicit, the selection of product molecules from a virtual library must emphasize operations on "clipped reagents". Put differently, molecular descriptors whose values may be calculated in an almost additive, or otherwise "decomposable", way from the descriptor values of its constitutive clipped reagents seem mandatory when initially selecting product molecules from a virtual library. Within our diamine library, for example, the "decomposed" descriptors would need to be computed only for the 26 700 + 1700 = 28 400 unique clipped reagents to enable selection among the entire million of millions of potential reaction products. Perhaps needless to add, the resulting product descriptor value estimates must also be sufficiently relevant and accurate for the selections to be meaningful, but fortunately, as detailed next, many such usefully decomposable descriptors do exist.

SEARCHING VIRTUAL LIBRARIES: UNDERLYING CONCEPTS

As emphasized in the Introduction, virtual libraries are useful only if individual molecules can practicably be selected from them. Most existing computational approaches to molecular selection could be described as "hypothesis based", where the hypothesis might be either a receptor or other physical model or else a structure−property relation. These hypotheses take time and often specialized expertise to develop. But the new experimental technologies are demanding instead that each iteration of molecular selection be much faster, which make new and different computational approaches very attractive. The new approaches we are strongly emphasizing here are based on "molecular similarity" (or its more fashionable complement "molecular diversity").

Because it is similarity (or more often improvement) in some useful property that molecular discovery projects seek, a discussion of similarity-based virtual library searching must begin by considering, "under what circumstances *does* molecular descriptor similarity imply, for example, biological property similarity?"

Neighborhood Behavior. The fundamental axiom of molecular similarity/diversity as applied to molecular discovery is often called the "neighborhood principle" - the proposition that molecular "descriptor spaces" exist such that molecules in the same local region ("neighborhood") of such a descriptor space tend to have similar values of a desired physical property. Several qualitative points can be made to help illuminate this proposition:

- It is very conservative. Without it no rational strategy for selecting molecules can exist.
- It is supported by the experience (and the continuing employment!) of synthetic chemists in drug discovery. Their responsibilities are to select and make molecules "near" other molecules of interest (however each understands "near"), in the expectation, frequently fulfilled, of discovering equal or better physical properties. The very disappointment when replacement of −H by $-CH_3$ occasionally "kills biological activity" testifies to our underlying expectation that "similar" compounds should have similar properties.
- It is not universal. Different types of molecular descriptor spaces can have quite different "neighborhood behaviors" for different molecular properties. For example, molecular weight is one of the most important descriptors in the physical separation of molecules. Yet molecular weight has no detectable neighborhood behavior with respect to most biological properties. If a lead structure has a molecular weight of, for example, 364.2, few would limit lead optimization to structures with molecular weights in some range such as 344.2 to 384.2, nor expect most compounds from arbitrary chemical series to be active if their molecular weight happens to fall within that range.

How is neighborhood behavior to be used in selecting molecules? An analogy, that of exploring for useful islands by sailing around a large uncharted ocean, may be helpful, because the geographical descriptors of latitude and longitude have very well-defined neighborhood behavior for the property of surface elevation. In this situation, the selection of a molecule for testing corresponds to the captain climbing up to the masthead and sweeping the horizon. If no land could be seen, there would be little point in looking again (for land) until the ship had sailed some tens of kilometers further. The analogous phase of drug discovery research is called "lead (molecule) discovery", a "lead" corresponding to a sighting of an island. A second phase of geographical exploration begins when a promising island is found, perhaps even a new continent. The first landfall becomes a starting point for a detailed examination of the immediate geographic neighborhood, or "lead followup", persisting as long as land, resources, and hope of gain continue. The final phase in drug discovery is called "lead optimization", when synthetic chemists make small changes in the "lead molecule", thereby tweaking some local structure−property relation as an explorer might climb a jungle mountain, seeking both a superior compound for commercialization and also an understanding of structure−property relationships desirable for patent protection.

As suggested by the explorer analogy, similarity concepts suggest the following fundamental strategies when selecting molecules for testing from a virtual library:

- If no example of a desirable molecule exists (lead discovery phase), search as widely as possible, by selecting for testing sets of molecules that are not too closely spaced, in terms of any descriptors known to have a neighborhood behavior.
- Whenever an example of a desirable molecule exists (lead follow-up or "explosion" phase), select other molecules within its neighborhood. Obtain and test those molecules. Iterate, by performing additional neighborhood selections on the most promising of the newly obtained molecules, as many times as desired. (As promising data accumulate, also seek "lead optimization" opportunities for developing and exploiting the more directed and conventional hypothesis-based molecule selection.)

Of course the explorer analogy is imperfect. Geographic space has only two dimensions, which are constant, well-defined, and easily traversed. In contrast, there are thousands of potentially relevant dimensions of molecular descriptors. Selectivity of drug action is a result of each receptor distinguishing its own preferred ligand properties, and thus the best neighborhood behavior that can be hoped for is probabilistic, not deterministic. Distances between com-

pounds depend greatly on the descriptor space used to map them, such that any "activity island" will exist only within specific descriptor spaces. Traversing any descriptor space requires discrete molecules, so that it may be difficult or impossible to envision molecules having particular desirable combinations of descriptors. Yet to repeat, without any neighborhood behavior no rational basis for molecule selection exists, and medicinal chemical experience provides reassurance that some neighborhood behaviors do exist (structures that look similar to a chemist tend to share biological properties).

Neighborhood Behavior for Biological Properties. Thus, before these two neighborhood strategies should be used in molecular discovery to select molecules, two questions ought to be confronted:

•What computable descriptors can be "validated" as indeed conferring a neighborhood behavior for the desired molecular property?

•How big are these neighborhoods—how far apart may two molecules be, in terms of these computable descriptors having the neighborhood behavior, and still be similar in their desired molecular property?

We have previously introduced some general techniques for answering these two questions,[5] and used these techniques to evaluate the neighborhood behavior for biological properties of several classes of computable molecular descriptors.[5,6] To summarize this work, note that an ideal "neighborhood behavior" in a descriptor means that a small change in a value of that descriptor should never produce a large change in the biological property. (When no islands are visible from the masthead, the explorer should be highly confident that no islands exist.) Therefore, to validate a descriptor, an experimental data set is chosen and plots made of all possible absolute *differences* (intercompound distances) in that descriptor value vs the resulting absolute *differences* in biological property. For descriptors having the desirable neighborhood behavior, these plots will have a low density of points in their upper left-hand quadrant (equivalent to infrequent instances of small descriptor changes producing large changes in biological potency, or the explorer failing to sight an island). Figure 3 shows contrasting examples of two such plots for one data set, the left-hand plot showing a consistent neighborhood behavior by the Tanimoto two-dimensional (2D) fingerprint descriptor for this data set, and the right-hand plot the absence of neighborhood behavior by the "random number" descriptor, which is serving as a negative control descriptor. From the left-hand plot, a "neighborhood radius" can also be estimated, if one assumes the bioequivalence of two compounds in initial screening if their potency differs by no more than two log units, by reading as shown from the 2.0 log units difference on the Y-axis, through a slope line constructed to optimize the low point density area of the graph, down to the corresponding descriptor difference value.

Plots considering all differences in each of 11 classes of computable descriptors along the X-axis were constructed for 20 series of structure−activity data chosen randomly from the recent medicinal chemistry literature. We must caution that this data source surely biases the *overall* result toward "too good" neighborhood behavior. (Most published data sets comprise "related" molecules—molecules that are neighbors of each other in the eyes of a synthetic chemist—and most have measurable activity. Related molecules that do not tend to have the same biological activities will seldom be described in publications.) However, this overall bias does not seem to affect the major conclusion of the study, the observation of large *relative* differences in neighborhood behaviors among the 11 descriptor classes.

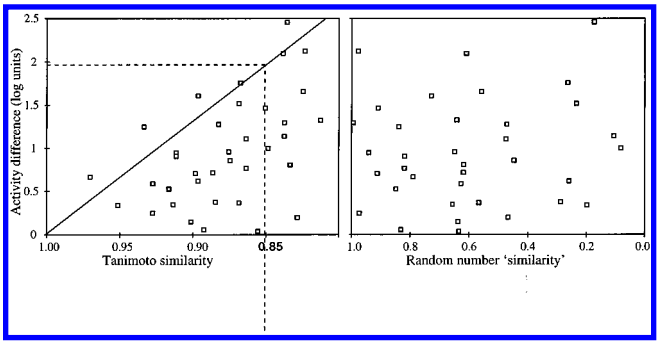

Figure 3. Two contrasting examples of plots for quantifying the "neighborhood behavior" of molecular descriptors. Each point in a plot corresponds to a *pair* of compounds, its X coordinate showing their absolute difference (distance) in a descriptor value and its Y coordinate their distance in a property value. The lower the density of such points in the upper left-hand region of the plot, the better the "neighborhood behavior" of its descriptor. The left-hand plot typifies the excellent neighborhood behavior observed for the Tanimoto 2D-fingerprint descriptor, whereas the right-hand plot shows a negative control descriptor, the same data set with a random number assigned to each compound, which of course has no neighborhood behavior. (Data taken from Uehling, D. E., et al., *J. Med. Chem.* **1995**, *38*, 1106, Table 2.) Construction lines added to the left-hand plot show how a "neighborhood radius" is established. A diagonal line through the origin is found that divides the plot into regions of most contrasting point density. A desired radius of "biological similarity" (compounds differing by this potency or less are bioequivalent) is marked on the Y axis (here 2.0 log units). Through this point, a horizontal line is drawn, and then from the intersection of this line with the diagonal a vertical line is dropped. The intersection of the vertical line with the X-axis is the "neighborhood radius" for this descriptor with this data set.

Two descriptor classes stood out as having the most general neighborhood behavior, that is, an appropriate and statistically significant point distribution was detected in >75% of the 20 data sets studied.

•"2D Fingerprints", sequences of bits ("yes/no" descriptors), coding for the presence of various kinds of bonded atom sequences within a particular molecule. For example, a particular bit might code for CNCCO and another for C(=O)OH. Such bit sequences were of course developed as keys, for rapid retrieval of particular structures from databases. The "fingerprint similarity" between two molecules is usually expressed as the Tanimoto coefficient, which ranges between 0.0 and 1.0, and in words is calculated as: the number of bits = 1 common to both the molecules, divided by the number of bits = 1 in either of the molecules. Molecule pairs that have a Tanimoto coefficient of ~0.85 or greater appear to a synthetic chemist as "similar". Given the historical success of medicinal chemistry, it is perhaps not surprising that several other groups in addition to us reported that such pairs of molecules also tend to have similar biological properties;[7] in our words, to be in one another's neighborhoods. (We observed this very high frequency of neighborhood behavior for fingerprints when the fingerprints coded only those fragments known to differ within a series, those within side chains. A decreased neighborhood behavior of fingerprints, when all nonvarying molecular fragments

were coded as well, was traced to the "folding" of fingerprints; that is, some of the bits that usefully distinguished side-chain fragments were all being set to 1 by other fragments common to all molecules. This phenomenon may account for an improved performance of MACCS keys, which are unique, compared with Unity or Daylight fingerprints, which are "folded" by defaults.[7])

•"Topomeric fields",[8] the gridded steric and/or hydrogen-binding fields of a molecular fragment, when aligned in a particular rule-determined "topomeric" conformation/configuration and orientation. The "topomeric distance" between two molecular fragments is defined as the square root of the squared field differences summed over the thousand-odd grid points representing the field. The "topomeric distance" between two molecules is defined to be the square roots of the sum of squared topomeric distances between their comparable fragments. In comparing multivalent fragments, differences in the relative positions of the fragment connections are also included, with a very high relative weight. For this descriptor we found a neighborhood radius such that molecules whose steric fields differ by the effect of a $-CH_2-$ group or less (equivalent to replacing $-H$ by $-CH_3$) tend to differ in biological potencies by two orders of concentration or less. Classical medicinal chemistry also recognizes this phenomenon, in describing structures so similarly shaped as bioisosteric.[9]

The other descriptor classes we evaluated were log *P*, molecular weight (MW), connectivity indices, and internal strain energy, which were not distinguishable from the negative control, in showing significant neighborhood behavior in 10% or fewer of the 20 data sets; 2D fingerprints on whole molecules, atom-pair fingerprints and autocorrelation vectors, showing neighborhood behavior in about half the data sets; and a hydrogen-bonding "field" of the topomeric conformation, for which 10 of the 11 applicable data sets possessed a neighborhood behavior. It is noteworthy that there was a very strong tendency for the descriptors with a high dimensionality to exhibit the most frequent neighborhood behavior.

Not only are "fingerprints" and "topomers" the descriptors that we found to exhibit the most consistent neighborhood behavior for biological properties, but they also allow very fast neighborhood searching in virtual libraries. The differences between two product molecules in terms of topomers, and to a first approximation fingerprints, are indeed "decomposable", that is, calculable by combining the differences between corresponding sets of clipped reagents, without having to build the individual product molecules, and the differences themselves are very rapidly calculated.

Combinatorial Docking. Probably the most active area of virtual library, research, a natural outgrowth of structure-based drug design, is "combinatorial docking". The receptor structure is required, preferably including or at least inferring, as a bound ligand, some product structure from a virtual library. Each of the clipped reagents in that virtual library is then docked into an appropriately specified region of the receptor cavity. The product molecules then selected for synthesis would typically be the full array, which included all the high-scoring clipped reagents. ("Decomposability" is usually not assumed; that is, at least some of the product molecules will usually be built computationally to ensure that the docking conformations of the fragments are compatible with each other.) This approach has shown great promise in at least one exploratory study,[10] producing a much larger proportion of more potent inhibitors than did the control, a full array of products selected by the "Chiron procedure" (described in the next paragraph). On the other hand, the required receptor structure will not be available for the majority of drug discovery opportunities. The docking scores of all the candidate clipped reagents also need to be recomputed for each new receptor to be studied, which is probably not a serious practical problem, but interesting as a characteristic that distinguishes such a hypothesis-based molecule selection approach from the similarity approaches that we have emphasized.

Other Descriptors. Considerable ingenuity is evident in the creation of many of the other notable descriptor classes that have recently been introduced for selection in molecular discovery. Many of these descriptors have been intended more for selection within existing or commercially available compound collections than selection within virtual libraries. Pioneering groups at Chiron both produced the first combinatorial library from synthetic building blocks (peptoids) and also were the first to articulate the need for selection from what we here call "virtual libraries" as well as a strategy for doing so.[11] Their strategy applied classical experimental design statistics to a somewhat *ad hoc* descriptor space including log *P*. MW, 2D fingerprints, and other proprietary parameters intended to capture three-dimensional (3D) information. Pearlman[12] advocates "BCUT" parameters, the highest and lowest eigenvalues of property-weighted adjacency graphs. Several groups are promoting "pharmacophoric" descriptors,[13,14] typically 2D fingerprint-like entities in which the individual bits represent whether a molecule can achieve a particular geometric arrangement of generalized functionalities believed likely to be critical in receptor binding. Various internal systems used by several large molecular discovery organizations seem to be stressing "atom-pair"-based fingerprints originated by Carhart[15], which we have found also to have a moderately consistent neighborhood behavior, as already mentioned.

SEARCHING VIRTUAL LIBRARIES

In addition to the variety of molecular descriptors just discussed, and the contrasting searching objectives of "lead discovery" and "lead optimization", approaches for similarity-based molecule selection can also be contrasted along two other different, somewhat interdependent axes. With respect to Figure 1, which of the entities shown, reactants or products, are to be searched? What searching techniques are to be used? The answers have design implications for virtual library software.

Entities to be Searched. The first axis to be considered is the virtual database entities to be searched. In both principle and practice, there are two possibilities. **Reagent-based selection**, certainly the most often used approach, and the only approach compatible with library synthesis/decoding strategies, such as "mix and split" or "positional scanning", prescribes the following: select only reagent molecules, then synthesize full combinatorial libraries of products. The other possibility is **product-based selection**, which, as suggested by the name, implies the following: select directly among the (far more numerous) products, then work backward to

the list of reagent quantities needed for synthesis. In general, reagent-based selection is much faster and the result is easier to execute in the laboratory, whereas product-based selection provides more information per molecule synthesized, because of lower redundancy. Note that it is very difficult to design full matrices in which all compounds are outside of each others' neighborhoods, because, as may be visualized from Figure 1, traversing any row or column of a full matrix sweeps out sets of molecules that are at least half structurally identical. And, even when successfully designed, such a nonredundant full matrix would be very small and so necessarily omit much of the "descriptor space" that the full virtual library is capable of occupying. Thus substantial information gains, from 35% to 50%,[16−17] result from substituting a "best product-based" design for a "best reagent-based" design, both considering the same molecular descriptors and producing the same number of compounds. However, the superiority of the "product-based" design must be balanced against its increased synthetic cost and therefore is greatest when the resulting library is intended for screening, repeatedly, in many different primary assays.[18]

These considerations suggest that the virtual library designer should anticipate a need to represent and manipulate product subsets as well as reactant subsets. The resulting challenge is the large potential size of these product subsets. For example, if a search yields only one hit per million products in the diamine virtual library, the resulting subset will still reference a million products, many more structures than are found in all but the very largest corporate compound inventories! Clearly a very compact notational form is needed for such large hitlists. The product set representation used within ChemSpace is a bit array describing an entire virtual library entry, the selection or absence of an individual product being indicated by a 0 or a 1 at a particular location. Standard image compression technology affords a compact disk file representation for such "bitset" search results, and their explicitly set-like nature also facilitates repetitive searching with additional constraints. On the other hand, such a terse way of describing individual products requires considerable contextual support, to ensure that the structure and related descriptors referenced by any single bit can always be retrieved, regardless for example of any updating of reagent lists that may have occurred in the interim.

Distance-Based Searching Algorithms. The other classification axis is the searching algorithm itself. Several sophisticated searching algorithms have been developed, especially for generating "diversity designs" as general screening libraries from corporate compound inventories. We will discuss these other algorithms later in a section on the comparison of virtual libraries. However, in selecting compounds from within virtual libraries, we have found direct and simple intercompound distance algorithms to be highly effective.

Algorithmically, similarity-based lead follow-up is a simple "neighborhood search": "Step through the candidates, keeping those which are close enough to the lead". Stepping through virtual libraries will be especially fast whenever the searching descriptors have the "decomposability" already mentioned, because one can start by eliminating all clipped reagents that *by themselves* are too far from the corresponding fragment(s) in the lead structure. This tactic is most productive when the radius of the requested neighborhood search is small compared with the average interfragment distance, and in our experience typical searches meet this criterion, at least for the fingerprint and topomer descriptors that ChemSpace neighborhood searching most emphasizes.

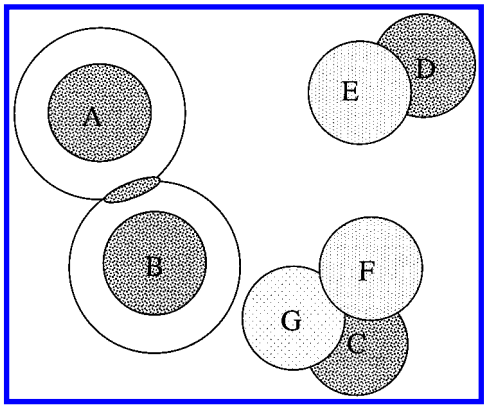

Figure 4. Venn diagram illustrating how a prospective library for synthesis might be assembled by combining hitlists from nine separate virtual library searches. It is assumed that seven hits, named A through G. have recently been discovered. Secondary testing of suggests that hits A and B are the most desirable, whereas E, F. and G have unacceptable properties. These results suggested that the best candidates for a first round of follow-up synthesis would be neighbors of A, B. C, and D, without being neighbors of E, F, or G. Also desirable would be structures that were not too far from both A and B. To identify such structures, "standard radius" neighborhood searches are performed around each of the seven hits, and two more searches with a larger radius were performed around A and B. Each circle represents the "descriptor space" swept out by one of these nine searches. Compounds identified within any of the more strongly shaded areas become the recommended library for synthesis.

A typical "diversity design" algorithm used to produce a lead discovery library would be: "Select a molecule. Eliminate all its neighbors from further consideration. Iterate these two steps until no molecules remain for selection." Such an algorithm provides complete freedom in the "select a molecule" step and can thus be customized in many ways. However, because such an algorithm involves many cycles of neighborhood searching, it is much slower.

Combining Search Results. These two basic operations, along with Boolean operations on the results, can be combined in various ways. For example, assume seven interesting structures, three of which have undesirable behavior in a secondary assay (see Figure 4). A target library then might be selected as the union of the product molecules within a small radius of any of the four superior structures, less the union of product molecules within a small radius of the three active but undesirable structures, and perhaps also including all those product molecules that simultaneously lie within a larger search radius of at least two of the four superior structures. If these combined neighborhood search operations produce too many product molecules for synthesis, a diversity design with a smaller-than-neighborhood radius may be performed, to maximize the chance that the synthesized library will reveal the descriptor regions where the most biologically interesting of the unsynthesized compounds are to be found.

Pragmatic Search Criteria. We have focused on similarity-based searching of virtual libraries, as a powerful tool for rapidly identifying the biologically interesting molecules, but there also are important pragmatic criteria, not based on

similarity, when selecting product molecules. Are all the needed reagents obtainable in a timely way (inventory, supplier, and price information)? Are the product molecules unlikely to show nonspecific toxicity in the biological assays (absence of well-known "toxophoric" fragments)? Are any bioactive molecules found also likely to have the pharmacodynamic characteristics typical of a clinically useful therapeutic agent (appropriate ranges of molecular weight, log *P*, counts of hydrogen-bonding atoms)? For all of these descriptor classes, the descriptor values of the product molecules are, to a satisfactory approximation, rapidly computable in a "decomposable", usually additive, fashion from the descriptor values of the reactant-derived fragments, so including these criteria as part of virtual library searching is straightforward.

Conventional Searching. Perhaps surprisingly, in our experience conventional searching has only rarely been useful to perform conventional "substructure" searching on virtual libraries. This rare usage has led us to think about the user's real purpose in conventional substructure searching of structural databases, which usually is to identify existing molecules likely to have some useful property. Existing substructure search technology forces the user to express this question by hypothesizing some generalized substructure as a "pharmacophoric fragment" and searching for compounds containing that fragment. The similarity-based searching we are emphasizing seems to provide a more direct and powerful means to the same end.

Nevertheless, depending on the neighborhood descriptor, similarity-based searching can also face an analogous "query remapping" challenge. The user's actual desire seems to be something like: "I am entering this query structure because it has some desirable property (hit in an assay, prominently cited in a competitive patent, ..). Find me structures likely to have a similar property, and I really prefer not to need to know about how this searching is done." For similarity searching, the user's query structure must then be mapped to a set of descriptors, which both have neighborhood behavior and that also are suitable for rapid comparison with a stored set of descriptors. For 2D fingerprints, this mapping is trivial (albeit with virtual databases there *is* a challenge in organizing the stored fingerprints of fragments for rapid comparison with the query-generated fingerprint, but that is beyond the scope of this discussion). But for topomeric (shape) comparisons, the stored virtual library descriptors to be searched are actually sets of shapes of fragments, to be assembled in accordance with various specific reaction schemes. The searching process must then begin with a successful mapping (fragmentation) of the query structure onto the assembly instructions for a particular library. Furthermore, to be most useful, this mapping process should not be limited to an identity in features critical for synthesis but must instead recognize bioisosteric features. For example, it is desirable that a query structure that contains no nitrogen be capable of retrieving topologically identical products from the aforementioned diamine library; for example, products differing from the query only in that a query structure carbon or oxygen matches a product structure nitrogen.

Exact-match searching is a common operation with conventional structural databases, the underlying question usually being, while registering or perhaps purchasing a new compound, "Is this query structure unique?" With virtual libraries, there are many possible ways of defining uniqueness: Is the library identical to another? Is the library a complete or partial subset of another? Are any compounds in this library found in another library? Is this compound in any virtual library? The last of these definitions is the most clearly useful, and it is also the most easily answered.

BUILDING VIRTUAL LIBRARIES

Before a virtual library can be searched, it must be built. As implied by Figure 1 and the surrounding discussion, the reaction itself must be described, the appropriate reagents somehow identified (usually by searching some compendium such as the ACD), the product fragments derived from the reagents via "clipping rules", and finally any searchable descriptors of the resulting clipped reactants must be calculated.

With reagent lists and reaction requirements typically being very dynamic, it is highly desirable that such virtual library maintenance operations be wholly automated. Such structural notations as SMILES or SLN[19] are useful and robust ingredients for specifying the various inputs to automatic maintenance, and we also experience a continuing need for refinements and extensions of these notations and the algorithms that manipulate them.

Choosing the "Core" Structure. When adding a new reaction to a virtual library, there is often some flexibility about its fundamental description. Consider simple amide bond formation; C(═O)N ⇒ C(═O)X + NH. What should a "common core" of this reaction, if any, be? Two obvious and equally acceptable possibilities are as follows: there is no common core other than the amide bond itself, or the C(═O)N portion is found in every product so that should be the common core. The "clipping rules" for the two possibilities will differ. In particular, the rules for the second possibility will produce two, probably disconnected, fragments from the amine component (contrast, as amine reactants, ammonia, diethylamine, and piperidine). The decision between such possibilities is usually determined by the descriptors that will be used for neighborhood (or other) searching. Specifically, we represent amides using the first possibility shown, because topomeric searching is most powerful whenever a reaction is described as the formation of an acyclic bond between the clipped reagents. On the other hand, for fingerprint estimation with amides we also take advantage of the C(═O)N structural commonality explicit in the second common core possibility. For similar reasons our contrasting practice for representing ring-forming reactions such as the Diels−Alder is to consider the product core to be *all* the common ring atoms, each product then having the two distinct sets of ten variable attachments contributed by each reactant.

Generalizing Synthetic Organic Chemistry. Another set of issues that arises with ring-forming reactions such as the Diels−Alder, involving two new bonds, are regioisomerism and stereoisomerism. Such reaction libraries are characterized by any single set of reagents producing an indeterminate mix of actual products taken from a completely determinate set of potential products. (Different issues arise when regio and/or stereoselectivity are determined—or postulated—for every combination of reagents in the virtual library, but such

situations are rare.) It is assumed that the searcher would like to retrieve every reagent set for which any of the potential products satisfies his query. In principle, this goal can be achieved by selectively expanding either the query or the set of stored products, to allow detection of any possible match. Within ChemSpace, the implementation of topomeric (shape) searching of the Diels−Alder virtual library expediently involves both of these expansions.

Organic chemistry is notorious for its fickleness. The intense "rehearsal" or "validation" effort that typically precedes the actual preparation of a combinatorial library usually uncovers little-known limitations of even the most widely used synthetic reaction. Yet additional experimentation, from a temperature change to protection of the interfering groups, can often induce successful combination of even the most obstinate combination of reagents, if the incentive is great enough. Because such know-how is so dynamic, we have found that reagent selection for a virtual library most usefully involves two or more layered procedures. First, whenever the reaction is built or updated, *all* formally acceptable reagents should be included. Then, as additional restrictions specific to a reaction and perhaps to particular reaction conditions are discovered and codified, these may be applied to clipped reagent lists or, in urgent cases, even to the products, effectively at search time, to restrict the product structures as desired. A further possibility is to flag the individual reagents that contain moieties which make them "acid sensitive", "carcinogenic", and so forth. It can then be left for the searcher to decide whether to accept all possible products, to limit output to the products from reagents an expert has deemed to be valid, or to eliminate products that require reagents having a specified undesirable character.

COMPARING VIRTUAL LIBRARIES

Molecular discovery organizations frequently need to evaluate new sources of molecules, either already realized, as when a collection of compounds becomes available for addition to the corporate collection, or potentially realizable, as when a new synthesis is contemplated, which can then be expressed as a virtual library and compared with other virtual libraries or with the corporate collection. Thus, the development of new techniques for "library comparison" is a very active area of diversity research. Here we review recent conceptual advances among a variety of approaches, at least as applicable to the corporate collection as to the comparison of virtual libraries.

Counting Compounds. One way to assess the diversity of a realized library is simply to count the number of distinct compounds in it, to calculate the respective costs per compound, or to apply some combination of these two criteria.[20] Generally, such a count is discounted by the number of compounds in the candidate library that are already in hand, for example, in the corporate database. When the libraries in question are virtual, however, the number of compounds "contained" within each library loses most of its relevance, if not most of its meaning. As already noted, a thoroughly enumerated combinatorial can easily overwhelm a chemist's ability to synthesize more than a fraction of its constituent compounds. For such libraries, any comparison drawn on a "per compound" basis will be practically meaningless. Instead, one is generally interested in how *extensive* a region of biochemically relevant diversity space is covered by the library.

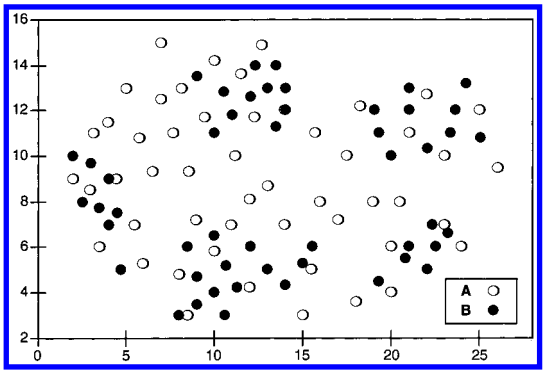

Figure 5. Illustrative diagram of two kinds of distributions within libraries. One (A) is more or less homogeneously distributed through the space covered. The other (B) is comprised of five distinct clusters. Units in the graph are arbitrary.

Comparing Scopes of Victual Libraries. Thus, the next step in evaluating a realized library is in terms of range alone (i.e., in terms of the minimum and maximum values attained across some set of descriptors). This approach finds its extreme manifestation in so-called D-optimally designed realized libraries,[11] where only low-dimensional descriptors can be used. Such ranges are calculable as limits for suitably additive descriptors, even for combinatorial virtual libraries. Subsequent experience has demonstrated that how thoroughly the included regions of the descriptor space are covered is as important or more so[21] as is how evenly the compounds in the library are distributed across the covered regions of descriptor space. This result is a directly a consequence of the fact that strong neighborhood behavior falls off quite sharply with decreasing similarity.[5,7,8] It follows that how *intensive* the coverage of a structural subspace is can be just as important as how *extensive* the coverage is, or even more so.

Consider, for example, two libraries that are quite similar in scope, yet each of which has its constituent compounds quite differently distributed (Figure 5). The compounds represented by the open symbols (A) are spread more or less homogeneously throughout the occupied structural space. The compounds represented by the closed symbols (B), on the other hand, are clumped together; five included subregions are sampled heavily, perhaps to the point of redundancy, whereas regions between clusters are represented sparsely or not at all. For lead discovery, the more evenly dispersed, less redundant library A will almost always be preferred, so long as most of the covered area lies in a potentially relevant (reasonably drug-like) region of structural space (see the discussions of reagent filtering and reaction selection already presented). It may be, however, that one knows in advance that one of the densely sampled regions is of particular interest; this is typically the case in lead optimization. In that event, the more heterogeneous library B may be preferred.

Note, however, that the economic value of a prospective library increases sharply as its similarity to known actives falls. Hence, some level of broadly based sampling is appropriate even in very focused development programs. This type of sampling becomes especially relevant when one considers the fact that proximity with respect to descriptors

that exhibit good neighborhood behavior is a *sufficient* condition for similarity in biochemical activity, but generally is not a *necessary* one;[22] hence, multiple, locally dense archipelagos of activity are to be expected, rather than a single contiguous "continent".

Partition Analysis. One way to assess the thoroughness and evenness of distribution simultaneously is to partition low-dimensional descriptor spaces into cells small enough for neighborhood behavior to be seen, then count the number of occupied cells for each candidate library or calculate the pairwise correlation between bin counts for different libraries. This has been described using classical quantitative structure−activity relationship (QSAR) descriptors such as calculated octanol/water partition coefficients and molecular refractivities, as well as using descriptors derived for this purpose.[14,23] Although partitioning has intuitive appeal, there is generally a strong tendency for central cells to be much more densely populated than are peripheral ones. This fact, together with the knowledge that the descriptors tend to be correlated with one another, limits the practical resolution one can achieve with this approach. Non-uniform bin sizes can be used to address this problem, but not always very satisfactorily. Nonetheless, partitioning can be very fast and very useful, particularly if care is taken to identify more or less mutually orthogonal descriptors (e.g., BCUT descriptors[12]) and to maximize dispersion among cells.[24] Comparisons between quite large datasets can be handled straightforwardly,[24] though it seems unlikely that any one set of descriptors can be found that is suitable for all libraries. If a partition analysis across BCUT parameters indicates that library A is distinct from B, one may feel confident that this conclusion will, in fact, extend to many other descriptor spaces of interest as well.[12]

Unfortunately, partitioning is inherently impractical for the very high-dimensional metrics for which the neighborhood behavior has been shown to hold most generally, because even a very modest resolution along each axis produces an astronomical number of cells, almost all of which are empty or contain just one element. Many cells may, in fact, be empty for *any* realizable sublibrary. Consider 2D substructural fingerprints: many, probably most, of the possible fingerprints defined by a Tanimoto radius of 0.15 (similarity coefficient of 0.85 or more) around penicillin, for example, probably do not correspond to a set of fragments attainable by any stable chemical structure.

Nearest-Neighbor Analysis. Many of these difficulties can be dealt with by shifting the focus of analysis from long-range relationships to local ones, that is, typically those between each compound and a nearest neighbor.[20] In its fullest expression, this approach produces a nearest-neighbor similarity profile in which the frequency of occurrence is plotted as a function of nearest neighbor distance. Examples of such a similarity profile, obtained by applying the *dbcmpr* executable in Selector[1] to structured combinatorial pyrimidine and cyclohexane libraries, which have been described in detail elsewhere[25] are shown in Figure 6. Each library has a different core structure, but the 2244 substituent patterns are common to both. Here, each is compared to a 6000-compound pyridine library that was designed to be partially homologous to the other two. As is common for fingerprint analyses, similarity is quantitatively assessed in terms of Tanimoto coefficients.[26]

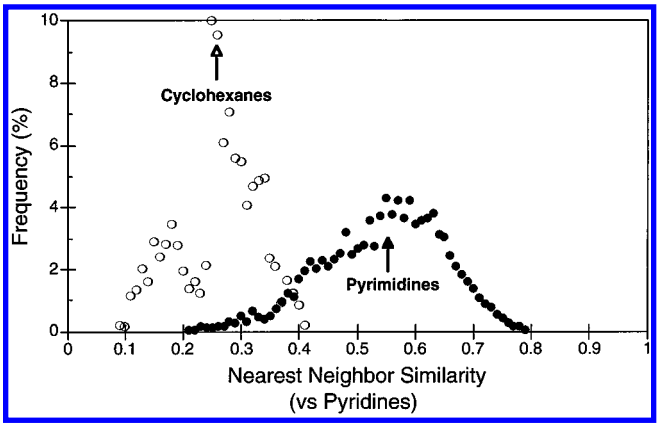

Figure 6. Nearest-neighbor similarity profiles for 2224-compound combinatorial libraries that share a common distribution of substitutions, but which have cyclohexane (○) or pyrimidine (●) core structures. The nearest neighbor similarity was determined for each library with respect to a reference library of 6000 pyridines, and the results were binned across intervals 0.01 in width to generate the histograms shown. The mean nearest neighbor similarities indicated by the open and closed arrows are 0.27 ± 0.07 and 0.55 ± 0.10. Detailed structures are given elsewhere.[25]

Displacement of the profile to the left, towards less similar nearest neighbors, indicates less overlap in coverage between the libraries in question. As expected the pyrimidine library is substantially more similar to the pyridine reference library than is the cyclohexane library. Note that it makes a difference which library is taken as the reference, because the nearest neighbors relationship need not be reciprocal. Such nearest-neighbor profiles can be very informative when the candidate libraries in question are comprised of similar numbers of compounds and have been elaborated to similar degrees. They can illuminate general underlying relationships: the substructure in the cyclohexane profile in Figure 6, for example, reflects in part the fundamental differences in symmetry between the cyclohexane and pyridine cores.

These profiles emphasize local similarity, which is more consistent with our understanding of neighborhood behavior than is using the means of *all* pairwise similarities between compounds; indeed, the mean similarity approach[27] cannot, in general, distinguish clumped from evenly dispersed libraries. The respective means of all pairwise distances for sets A and B in Figure 5, for example, are 108 and 116 units, whereas the mean nearest-neighbor distances are 1.75 and 1.13 units.

Nearest-neighbor similarity analysis can be extended to give an absolute measure of diversity by comparing a library to itself rather than to some external reference. There are two problems with interpreting such a self-similarity analysis, however. One is that no single "goodness number", such as the mean nor the median, fully captures its character. The other is that the meaning of the profile is rather perversely dependent on the size of the dataset: a subset of a library is generally less self-similar than is the library itself because removing compounds can only increase nearest neighbor distances. This phenomenon is illustrated in Figure 7 for a randomly drawn 1200-compound subset (open circles) of the pyridine library (dark circles). The second problem is that nearest-neighbor analysis is too exclusively focused on local relationships for application to the virtual library comparisons of interest to us here.

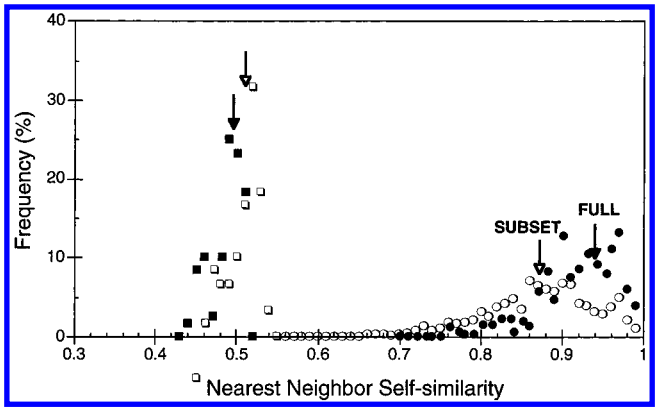

Figure 7. Self-similarity profiles for the full library of 6000 pyridines (●) and for a sublibrary of 1200 compounds selected from it (○); the mean nearest neighbor self-similarities indicated by the closed and open arrows are 0.94 ± 0.04 and 0.88 ± 0.07, respectively. Self-similarity profiles are also shown for characteristic 60 compound subsets drawn from the full library (■) and from the sublibrary (□); the mean self-similarities are 0.494 ± 0.02 and 0.516 ± 0.02, respectively.

A useful alternative is to draw a subset of intermediate, fixed, size from each library using maximal dissimilarity selection[28] and run a self-similarity analysis on that. This procedure is akin to measuring the area of a room by spreading a fixed number of pennies as far apart as possible, then measuring the distance between pennies. The size of the maximally diverse subset selected is dictated by the resolution desired: it should be less than twice the dimensionality of the underlying descriptor, which is the maximum possible resolution of a binary metric, but more than the number of underlying clusters expected. The differences between sets A and B in Figure 5, for example, will only be apparent for characteristic subsets containing at least six compounds.

Given these constraints, a relatively sharp peak is generally obtained in the self-similarity profile for the characteristic subset (square symbols in Figure 7). The position of this peak has a well-determined, **characteristic value** for each library and represents a limiting value for characteristic sets drawn from subsets of that library. In addition, the expected value is the same for all random subsets of a given size that might be drawn from the parent combinatorial(s) (the exact value will vary slightly due to stochastic sampling effects and near-symmetries of distribution in most libraries).

Direct maximal dissimilarity selection may not be practical for many full virtual libraries of interest because it would require too much (albeit not complete) enumeration and is on the order of N^2 in time. Fortunately, it can be approximated to any desired degree by generating a random subset and running the requisite calculations on a characteristic subset of that. This procedure is illustrated in Figure 8 for the mixed virtual library obtained by combining the pyridines with 500 compounds drawn from the cyclohexane library and 100 compounds from the pyrimidine library. For the sake of conciseness, the mean nearest-neighbor self-similarity for a characteristic set of m compounds is designated $1 - \delta*(m)$ in Figure 8. [Note that $\delta*(m)$ itself is most appropriately a Soergel distance (1 − Tanimoto),[26] rather than a similarity, which is done for generality across metrics and to maintain consistency with earlier usage.[25]] Note that the mixed library is less self-similar ("larger") than

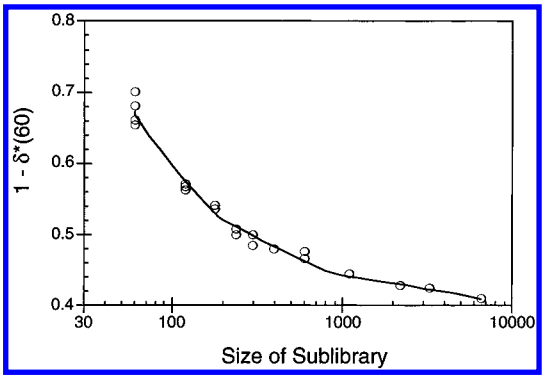

Figure 8. Sublibraries of varying size were created by randomly sampling a parent library comprised of 6000 pyridines, 500 cyclohexanes, and 100 pyrimidines; detailed structures are presented elsewhere.[25] Mean nearest neighbor self-similarities [$1 - \delta*(60)$] were then computed for characteristic subsets of 60 compounds drawn from each sublibrary using the QuickSel option in Selector.[1]

is the pyridine library alone; the respective values for $1 - \delta*(60)$ are 0.409 and 0.494. The cyclohexane and pyrimidine sublibraries contribute relatively few compounds numerically, yet they provide a significant complement to those in the pyridine library. This result is gratifyingly consistent with intuitive notions of what diversity should mean when considered with neighborhood descriptors in mind.

Such analyses do not, of course, indicate whether two libraries cover the same or distinct spaces. That information can most efficiently be obtained by running pairwise comparisons between diverse, representative subsets drawn from each virtual library. Such subsets can be obtained using optimizable K-dissimilarity selection[27] in ChemEnlighten,[1] then calculating nearest-neighbor similarity profiles between them by applying *dbcmpr* in Selector.[1] Here diversity alone is not enough because each subset should be representative as well. This makes OptiSim[1] selection particularly well suited to the task.

USER INTERFACES

A virtual library implies two major tasks, searching existing libraries and building new libraries. These tasks are expected to be performed by different types of users, and thus have rather different user interface requirements.

Searching Interfaces. Searching existing virtual libraries is expected to become an organization-wide activity, and so today only one presentation medium seems worth considering, the corporate Intranet, more specifically the Web. The major challenge afforded by a Web-based interface to such a noninteractive computation as virtual database searching is maintaining user context. The user does not want to (and indeed cannot reliably) tie up his local browser for hours or more with an ongoing search process. For most users, the idea of a named "batch process" is well established for managing such sustained activities. The difficulty is that Web protocols do not themselves naturally support the batch concept of data or processes "belonging" to a named user. However, with effort, this limitation can be overcome. A typical search session might involve query generation and search submission, search process management, viewing and manipulating search results, and possibly transmitting search results to colleagues. A few comments about each of these tasks are in order.

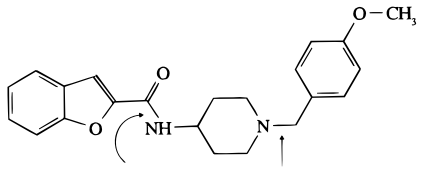

Figure 9. The structure of a candidate antiarrhythmic agent submitted as a ChemSpace query, to identify within the diamine library of Figure 2 the structures that are most likely to exhibit similar biological properties. Two arrows point to the bonds that would be formed by the diamine reaction sequence.

Entering and Performing Queries. The traditional query generation tool is a structure "sketcher". With today's enormous information flow, another useful starting point might be a list of compound identification (ID) numbers or "regids". The user also needs to specify the database to be searched. Because one of the major objectives in similarity-based searching of virtual libraries is to identify molecules with similar properties but different structures, it seems clear that the user will usually want to search large collections of reactions rather than individual reactions. Also, we already presented a variety of search techniques, each with its own parameters and many being used in combination. To avoid intimidating the occasional user, to provide some degree of search standardization within an organization, and to minimize overuse of shared computing resources, a "canned search" facility limiting the options available to the general community has proved useful. Finally, like any other lengthy noninteractive process, virtual library searching requires job management tools. The user must be able to control the job (base file/directory) name, check on progress and possible error conditions, and stop unwanted jobs.

Viewing Search Results. The primary output of a molecule selection process must be chemical structures. An attractive aspect of the output from a virtual compound library, often more than compensating for its possibly large volume, is that the structures are naturally organized by reaction. A summary page can contain a series of entries, one for each reaction within which any product "hits" were found. Such an entry would include the number of total structures and of variations at each site for the reaction, accompanied either by the core structure marked at the positions of variation, or more compactly as a structureless but sortable line of text. Clicking on a selected set of entries allows "drill down", *via* a new viewer, to display the possible structural variations at one or more sites. Also valuable are facilities permitting Boolean operations such as AND'ing, OR'ing, and subtraction on commensurate hitlists.

Building Interfaces. Building new libraries is a much more complex activity. In general, the data input facilities should allow for the entry of all needed information to: describe any arbitrary structural transformation; perform appropriate reagent searches, apply "clipping rules" to the search results; and calculate those descriptors of the clipped reagents that may be required for any supported type of virtual database searching. Then the actual library building process needs to be dispatched and monitored. There are many problems in designing a user-friendly interface for these tasks, with perhaps the most challenging being the large

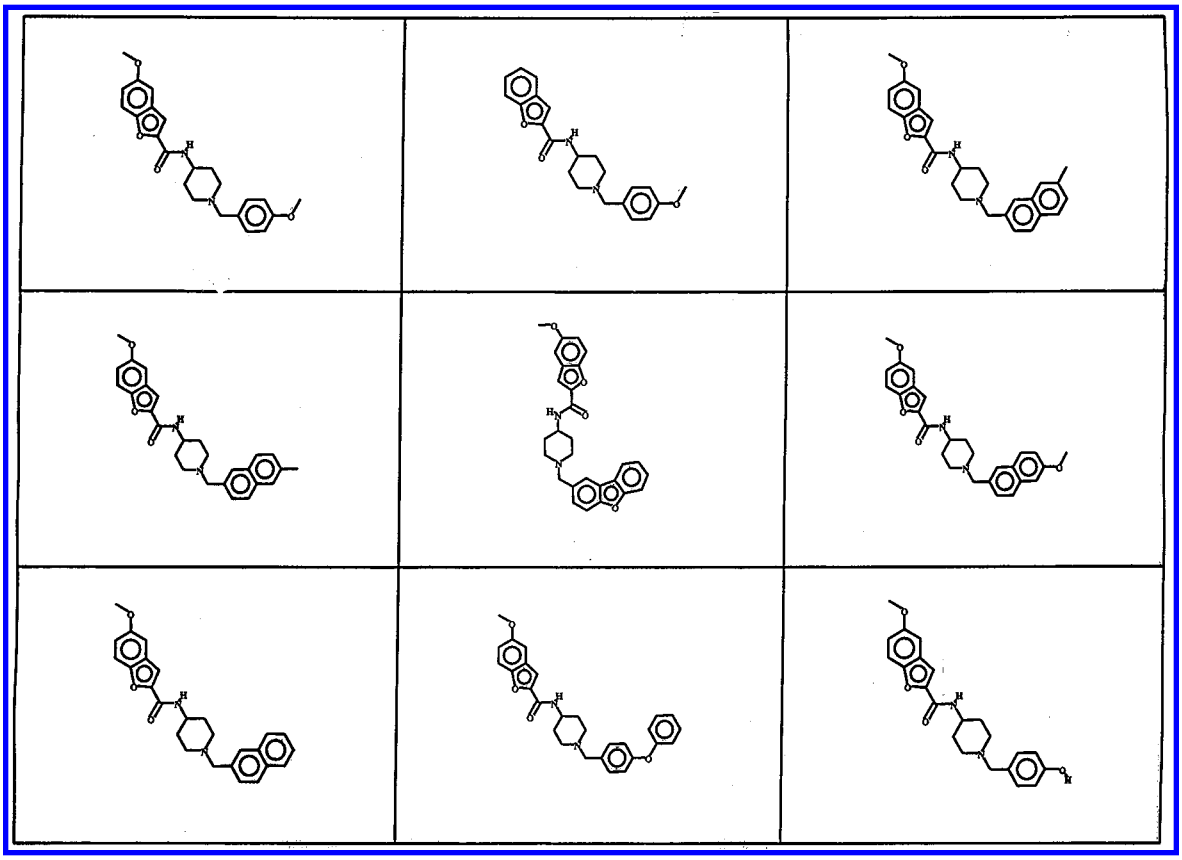

Figure 10. The nine structures most similar in fragment composition to the one in Figure 9, of 15 000+ structures retrieved from the diamine library summarized in Figure 2. The similarity descriptor is the Tanimoto coefficient of 2D fingerprints. As would be expected if the virtual library references suitable reagents, the query structure of Figure 9 is indeed (one of) the most similar structure(s) retrieved, at the top middle of this Figure.

number of unfamiliar and often ad hoc concepts that must be learned before any but the simplest virtual library can be described.

APPLICATIONS

Our experience suggests that a set of selected molecules must meet two criteria for the selection tool to be judged reliable and useful:

•Most of the structures found must be *intuitively appropriate*, perhaps even trivially so, to an experienced chemist. This result raises confidence that the searching process is trustworthy, whatever "magic" may be invoked.

•Some of the structures found must be innovative, or *nonobvious*. There is little value in a searching technology that retrieves only information that is already known to the user.

Even better, some structures found should be judged both innovative and appropriate.

Another practical goal is accessibility. Particularly when a "lead structure" is first discovered, before any other structure−activity data can be known, lead follow-up structures should emphasize the most available candidates, that is, those synthesizable in a very few steps from commercially offered reagents.

Examples of Virtual Library Search Results. In this spirit, we present some typical ChemSpace neighborhood search results. The query structure, shown in Figure 9, was a compound in the 100 000-member Optiverse screening library, which we noticed because of its accidental structural similarity to an antiarrhythmic lead described in a Searle patent application. Two independent neighborhood searches of the aforementioned trillion-member diamine library were done, each using one of the descriptors we had found to have the most consistent neighborhood behavior, 2D fingerprints, which retrieve structures composed of similar fragments, and topomeric fields, which retrieve structures of similar shape. The search ran overnight for the 2D-fingerprint neighborhood search and yielded >15 000 hits. The topomeric or bioisosterism neighborhood search took an hour or so, retrieving ∼24 000 structures. Both searches were performed on a typical Silicon Graphics workstation (R4000 series).

The structures from the 2D-fingerprint search were sorted by descending Tanimoto coefficient relative to the query structure, with only the nine most similar molecules being shown in Figure 10. Here the query structure itself appears as the second structure retrieved, in a tie with the first place compound, which differs from the query structure only by an additional methoxy group that evidently did not set any additional bits in its fingerprint. Inspection of the other compounds shows their very similar fragment composition. However, many of these "closest fingerprint neighbors" are much larger than the query, comprising multiple occurrences of the same fragment sets. To avoid such cases, an upper bound on molecular weight or count of heavy atoms may usefully be combined with 2D-fingerprint searching.

The structures shown in Figures 11 and 12 are representative of the entire set retrieved by the topomeric neighborhood search for the query structure in Figure 9. Figure 11 shows products formed by performing the same reactions at the same nitrogens, whereas the products in Figure 12 result from the many other possible pairings of diamine derivatizations. Each figure comprises two vertical lists, with a list showing

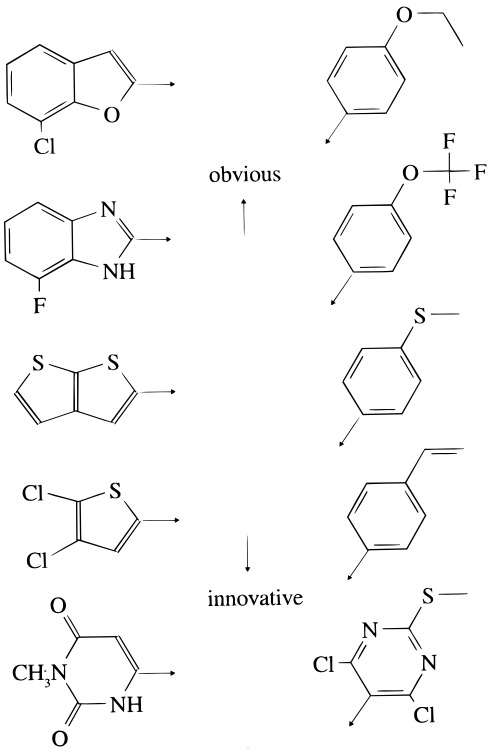

Figure 11. Structures of 10 representative shape-similar fragments, components from various of the roughly 120 structures of those within the diamine library of Figure 2 that were both shape-similar to, and also produced by exactly the same reaction sequence as, the structure in Figure 9. Fragments in the left hand column correspond to the benzofuran ring in Figure 9, while fragments in the right hand column correspond to the *p*-methoxyphenyl group. Only commercially offered reactants were considered. Each of these columns is ordered, subjectively, so that fragments at the bottom are less likely to be obvious to an experienced chemist.

the range of similarly shaped variations at a particular site in the query molecule; a top block of Figure 12 shows bioisosteric variations of the 4-aminopiperidine core. Within Figure 11, the two vertical lists have also been ordered, subjectively, from the most obvious structures at the top to the most innovative at the bottom. For example, in the left-hand list, a benzofuran moiety is first substituted with small atoms, then replaced by other 5,6-fused heterocycles, and finally replaced by other heterocyclic rings substituted to confer overall shape similarities. The 2D structural diversity of this numerically small group of candidates is worth emphasizing, as suggestive of a promising "series-hopping" capability, useful in circumventing toxicity or pharmacodynamic misbehavior, or in the competition to discover molecules that appear similar enough to the receptor as a promising rival structure, but different enough to the patent examiner. The interpretability associated with this shape descriptor should also be noted, in that the structure−activity data resulting from making and testing these compounds would provide immediate evidence of any regions and types of shape change that would be most likely to improve activity, in a second round of synthesis.

FUTURE DIRECTIONS

Because of the critical requirement for molecule selection from virtual libraries to be rapid and effective, frequent improvements in searching descriptors and technologies are

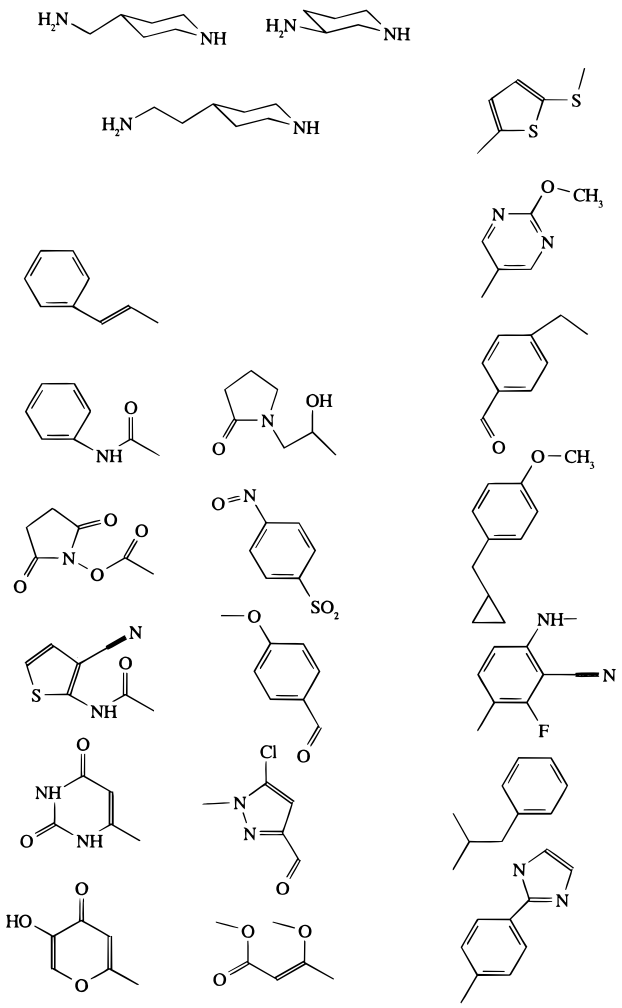

Figure 12. Structures of 22 representative shape-similar fragments, components from various of the ~24 000 structures of those within the diamine library of Figure 2 that were both shape-similar to and also produced by a reaction sequence somehow different from the structure in Figure 9. Fragments in the left-hand block correspond to the benzofuran-2-acyl moiety in Figure 9, fragments in the right hand column correspond to the *p*-methoxybenzyl group, and fragments in the block at top correspond to the 4-aminopiperidine "core". Only commercially offered reactants were considered.

likely. Currently our own highest priority goal is to increase the frequency of "lead hopping" (finding bioisosteres of a query molecule in synthetically unrelated virtual libraries) by adding a greater variety of reactions and reagents. A particular challenge is to devise data representations and techniques that will enable efficient shape-based searching of the products of multistep sequences. For example, if there are not enough suitably shaped reagents available "off the shelf", how might more be synthesized?

We have discussed virtual libraries for molecule selection mostly with respect to the earliest stages of discovery research, lead discovery and follow-up, when the structural candidates are most numerous. However, molecule selection remains an important issue throughout molecular discovery programs, so that the impact of these technologies could be much broader. For example, the close relationship between a virtual library entry and a Markush patent claim suggests the potential of these technologies for strengthening patent positions, both offensively and defensively. Or, as data relating structural changes to metabolic, toxic and pharmacokinetic effects become available for molecular selection within virtual library systems and thereby to early decision making, the success rate in all subsequent steps should improve.

There seems to be a growing tendency for discovery organizations to consider as their sources of screening compounds not only the traditional compound collection, but also the virtual libraries readily accessible to their chemists and synthesizers. (The differences in delivery times between these two sources are disappearing.) Yet the memberships of the virtual libraries readily available to different organizations is converging, because the same reagents and reactions amenable to automated synthesis are mostly available to everyone. These conflicting tendencies suggest that organizations that want to maintain the proprietary positions essential to any high cost molecular discovery will begin to develop inventories of proprietary reagents and reactions, with virtual libraries playing an obvious role.

"Thinking in multidimensional descriptor space" is not a familiar or comfortable experience for most people, so the development of appropriate interactive tools for navigating virtual libraries is desirable. The locality of neighborhood behavior suggests that the emphasis in simplification, necessary in reducing thousands of descriptor dimensions down to a visualizable two or three, should be on preserving short-range distances rather than the long-range distances characteristic of most statistical analyses.

CONCLUSIONS

Some form of virtual library technology can help in preventing the "high throughput" productivity revolution from being impeded by the slow speed of unsupported human decision making. We believe that the marriage of neighborhood searching with virtual libraries deserves to become accepted as a major advance in the effectiveness, as well as the speed, of the selection processes so critical to success in molecular discovery research.

ACKNOWLEDGMENT

We especially thank Allan Ferguson, Robert Glen, Peter Hecht, Stefan Guessregen, Laurence Weinberger, Peter Willett, Eric Martin, Mark Hermsmeier, and David Floyd for helpful discussions and support as the ChemSpace technology has been developed.

REFERENCES AND NOTES

(1) Legion, Selector, and ChemSpace are registered trademarks of Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144. Optiverse is a registered trademark of MDS PanLabs, Bothell, WA.

(2) Leland, B. A.; Christie, B. D.; Nourse, J. G.; Grier, D. L.; Carhart, R. E.; Maffett, T.; Welford, S. M.; Smith, D. H. Managing the Combinatorial Explosion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 62−70.

(3) This compendium of all commercially offered compounds may be obtained from MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.

(4) Chemical Abstracts Service reported 17 417 623 registered structures on 2/12/98.

(5) 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.

(6) 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.

(7) 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.
(8) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(9) Clark, R. D.; Ferguson, A. M.; Cramer, R. D. Bioisosterism and Molecular Diversity. In *3D QSAR and Drug Design: Volume 2*; Kubinyi, H., Martin, Y. C., Folkers, G., Eds.; Kluwer: Dordrecht, The Netherlands, 1998; pp 213−224.
(10) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J. A.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-Based Design and Combinatorial Chemistry Yield Nanomolar Inhibitors of Cathepsin-D. *Chem. Biol.* **1997**, *4*, 297−307.
(11) 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.
(12) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *In 3D QSAR and Drug Design: Volume 2*; Kubinyi, H., Martin, Y. C., Folkers, G., Eds.; Kluwer: Dordrecht, The Netherlands, 1998; pp 339−353.
(13) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New perspectives in lead generation II: evaluating molecular diversity. *Drug Disc. Today* **1996**, *1*, 71−78.
(14) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.
(15) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *24*, 64−73.
(16) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(17) Unreported studies performed at Tripos.
(18) Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screening* **1996**, *1*, 65−73.
(19) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. J. *Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.
(20) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using MDL keys as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(21) Martin, E. J.; Critchlow, R. E.; The well-tailored library. Beyond mere diversity. *ACS Abstract* CINF 003, 213th ACS National Meeting, 1998.
(22) Clark, R. D.; Cramer, R. D. Taming the combinatorial centipede. *CHEMTECH* **1997**, *27*, 24−30.
(23) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(24) Studies done with DiverseSolutions, a program created by R. S. Pearlman and K. M. Smith at the Laboratory for Molecular Graphics and Theoretical Modeling, College of Pharmacy, University of Texas, Austin, and commercially distributed by Tripos, Inc.
(25) Clark, R. D. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(26) Gower, J. C. Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1985; Vol. 5, pp 397−405.
(27) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(28) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity methods. In *QSAR: Quantitative Structure-Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss, Inc.: New York, 1989; pp 173−176.

CI9800209