J-CAMD 374

# The measurement of molecular diversity: A three-dimensional approach

David Chapman

*Afferent Systems Inc., 442A Collingwood Street, San Francisco, CA 94114, U.S.A.*

Received 23 April 1996
Accepted 27 June 1996

*Keywords:* Combinatorial chemistry; Library screening; Ligand–protein binding; Molecular similarity; Molecular shape; Scoring function

**Summary**

This paper describes a method for selecting a small, highly diverse subset from a large pool of molecules. The method has been employed in the design of combinatorial synthetic libraries for use in high-throughput screening for pharmaceutical lead generation. It computes diversity in terms of the main factors relevant to ligand–protein binding, namely the three-dimensional arrangement of steric bulk and of polar functionalities and molecular entropy. The method was used to select a set of 20 carboxylates suitable for use as side-chain precursors in a polyamine-based library. The method depends on estimates of various physical–chemical parameters involved in ligand–protein binding; experiments examined the sensitivity of the method to these parameters. This paper compares the diversity of randomly and rationally selected side-chain sets; the results suggest that careful design of synthetic combinatorial libraries may increase their effectiveness several-fold.

## Introduction

Compound library screening is an important step in drug discovery. Typically, thousands of small molecules are assayed for relevant biological activity, and the most active are selected as leads for optimization. The thousands of potential ligands may be natural products, commercially available synthetics, proprietary synthetics, or the results of combinatorial synthesis. The recent development of robotic high-throughput screening techniques has significantly increased the effectiveness and decreased the cost of screening, resulting in earlier, more potent leads. Traditionally, screening was considered a random process. In the absence of any knowledge of the protein target, one molecule is as good as any other, so long as it is not obviously toxic or highly reactive. It has also long been thought, however, that screening works because the molecules screened are *diverse*. To use assay resources optimally, one would like to screen molecules as dissimilar as possible, so that each might bind a different, minimally overlapping set of possible targets.

Until very recently, 'diversity' was a vague concept that had not been quantified. A rigorous, quantitative definition of diversity might, however, allow more effective screening by enabling a rational selection of molecules to assay, even in the absence of any information about the target site.

Rational design of screening libraries would be particularly valuable in choosing side chains for combinatorial libraries. Recently developed combinatorial methods use parallelism to allow the synthesis of many molecules simultaneously, which makes it possible to generate libraries of thousands or millions of distinct molecules in a few weeks [1]. Typically, combinatorial libraries consist of a central 'scaffold' molecule substituted at several sites with all possible combinations of a set of side chains. If, for example, there are 20 side chains and three attachment sites, generating all possible substitutions will yield $20 \times 20 \times 20 = 8000$ distinct compounds. Increases in the diversity of the set of side chains result in exponential increases in diversity in the resulting compound library as a function of the number of attachment sites. Because only a few side chains are used, the stakes of the rational selection problem are raised tremendously.

A group at Chiron Corporation [2] has recently described a measure of molecular diversity and an algorithm for selecting a diverse set of side chains. This algorithm computes, for each candidate side chain, a set of properties similar to those used in 2D QSAR. These include bulk molecular properties (weight, number of atoms and bonds, number of distinct elements, and estimated lipophilicity); the Molconn-X topological indices; Daylight fingerprint Tanimoto coefficients [3], which encode information about the presence or absence of functional groups and substructure fragments; and a set of 'atom layer properties' which sum, for the set of atoms at a given number of bonds from the side-chain attachment point, either the atomic radius or a binary variable indicating whether the atom is an acid, a base, an H-bond donor or acceptor, or aromatic. Statistical techniques then select a small set of side chains that maximize dissimilarity relative to these properties. Similar techniques have subsequently been used by other groups [4].

Experience has shown that the use of three-dimensional molecular descriptions can dramatically improve QSAR performance [5]. The reason is that these descriptions are more directly related to the physical properties that determine whether a molecule will bind to a given protein site than are 2D properties. While it is often possible to find a small set of 2D variables that correlate with binding, the reason for this correlation is often not clear, and it is often not possible to find such correlations. Further, most of the variables will not only not correlate with binding, but will contribute noise. Therefore, maximizing diversity with respect to 2D parameters may not provide maximal diversity in binding.

This paper describes an alternative method for maximizing the diversity of a set of molecules (or side chains) which is rooted in known principal determinants of molecular binding: the three-dimensional shapes of molecules, the locations of their polar functionalities, and their entropies. Diversity in terms of these properties should correlate well with diversity in binding to protein sites. Since these properties depend on molecular conformation, all low-energy conformers of each molecule must be examined. A conformational search is required to find these conformers. (Boyd et al. [6] have described a diversity method that uses three-dimensional information, specifically the distances between functional groups. However, their program works from only a single conformation for each molecule, and looks only at a small set of functional groups (such as *N*-methyl acetamide) chosen by the user for a specific application.)

The next section presents the diversity measure and an algorithm for choosing small diverse sets of molecules out of a large pool of candidates. The Results section describes the application of this method to choosing a set of diverse carboxylates, used as side-chain precursors for a combinatorial library, and, as a control, to examining the natural amino acid side chains. The Discussion section is devoted largely to suggestions for further research.

## Methods

### *Overview*

The method is based on the idea that the diversity of a set of molecules should be measured by the number of different protein target sites they can bind, weighted by how well they bind each site. This section presents the conceptual basis of the method; details are given in subsequent sections.

The shape of a molecule is one of the primary determinants of whether it binds a given site, and how well. However, most molecules are flexible and do not have a unique shape. Therefore, the program must examine the range of shapes each molecule can take on by performing a conformational search, which yields for each a set of conformers. The program uses the DGTSS systematic search method of Lozano-Perez (unpublished results) to generate starting conformations and the MM2 force field as augmented in BATCHMIN [7] to minimize them. The DGTSS method divides torsion space into a grid of equally spaced buckets and uses distance geometry [8] to find, in each bucket, a conformation that violates no van der Waals atom intersection constraints. Energy minimization used the Broyden–Fletcher–Goldfarb–Shanno technique [9] with a tolerance of $10^{-4}$. All conformers differing by at least 0.05 Å rms atomic deviation and within 5.0 kcal of the minimum-energy conformer were retained.

Given two conformers (of the same or of different molecules) the algorithm computes a measure of *dissimilarity* in terms of the factors relevant to binding. Specifically, this dissimilarity measure has a steric and a polar component. Conformations are rated sterically in terms of the amount of bulk in each that fails to overlap atoms in the other. They are rated for polar dissimilarity in terms of the distance from each polar functionality in one conformer to the nearest functionality of the same polarity in the other.

This dissimilarity metric is meaningful only relative to a specific relative alignment of the compared conformers. The algorithm computes an alignment that attempts to maximize similarity, using a continuous function optimization method. That is, the two conformers are allowed to rotate relative to each other until they find an alignment that overlaps steric bulk and polar functionalities as well as possible. This alignment process models the physical alignment process that occurs in ligand binding; the ligand will orient itself as well as possible to fit a target site.

To evaluate the diversity of a set of molecules, the algorithm must consider all the conformers of each molecule in relationship to all the conformers of all the other molecules. It computes a measure that combines the pairwise dissimilarities of all the conformers to yield a diversity score. Using this measure, a selection algorithm

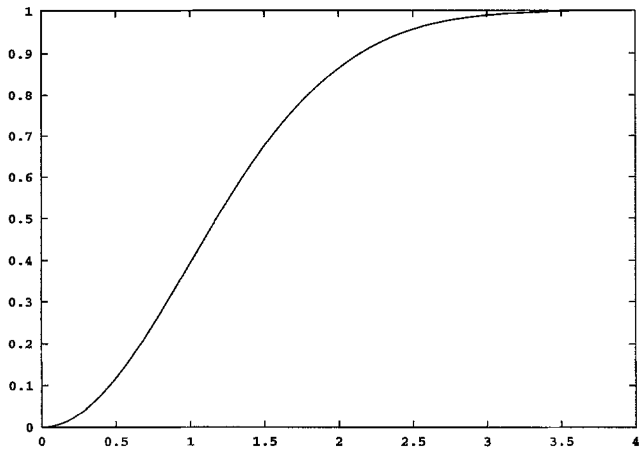

Fig. 1. The 'soft threshold' function w(1−exp(−d$^2$/t)) plotted with w= 1.0, t=2.0.

chooses a small, highly diverse set of molecules from a large pool of candidates.

The algorithms were implemented in Lucid Common Lisp on a four-processor Silicon Graphics Challenge Series computer server.

### *Dissimilarity of poses*

Let us use the word 'pose' to denote a conformation in a particular alignment. The innermost routine computes a dissimilarity score for two poses. This score takes into account the dissimilarity of the poses in terms of their steric, hydrogen-bonding, and electrostatic features.

There is a considerable literature on three-dimensional measures of molecular similarity [10], almost all of which actually considers pose similarity. (That is, it assumes that a single conformation and alignment is given for each molecule.) The diversity computation requires a pose similarity metric that can be evaluated very quickly, because, as we shall see, it has to evaluate the similarity of many millions of pose pairs. Because the algorithm must sacrifice some accuracy for speed, it takes a less sophisticated approach than many in the literature.

The method approximates steric overlap with interatomic distance. That is, for each atom in each pose it computes the distance to the nearest atom in the other pose and takes this distance as indicative of a degree of atomic nonoverlap. (This ignores variations in atomic radius.) Conceptually, it then sums these distances to yield an overall measure of steric dissimilarity. It measures dissimilarity in polar functionalities analogously, in terms of the distance from each partially or formally charged group in the one pose to the nearest group of the same sign in the other pose. Atoms were assigned donor, acceptor, and charged status using a system of a few dozen rules based on known polar functionalities.

As stated, this method will award a dissimilarity contribution proportional to the displacement of, for example, a hydrogen-bond donor group between the two poses. Obviously, this is unrealistic, since a donor that is displaced more than a few Å simply cannot bind to the same acceptor in the target site. As a result, no consistently greater difference in binding will be incurred as a result of further shifts. Therefore, a 'soft threshold' is applied to the interatomic distances before summing them. This threshold function has the form w(1 − exp(−d$^2$/t)), where d is the distance (Fig. 1). This is an S-shaped weighting function that is approximately linear for distances less than about t and approaches asymptotically an upper bound of w, so that the dissimilarity of large substitutions is bounded. The parameter t corresponds to the distance at which ligand–protein interactions drop close to zero. There are good estimates for this distance; however, these assume that the target site is rigid, and so do not account for protein flexibility. The method uses t = 2.0 Å for all interaction types; although this value is quite arbitrary, the experiments described later show that it can be halved or doubled without substantially affecting the results. The weighting function is continuous, which we will see in the next section is required by the alignment procedure. Applying this function yields a metric most similar to the Gaussian-overlap similarity measure of Good and Richards [11].

The effects of steric similarity, partially charged functionalities, and formally charged functionalities are weighted according to the results of Böhm [12], who estimated the relative strength of these factors based on fitting crystallographic ligand–protein structures with known affinities. (This provides values for w.) Specifically, values are taken from his scoring function #3, which makes use of only interatomic distances and not hydrogen-bond angles. The inclusion of hydrogen-bond angles would have added significantly to the computational cost, and Böhm found that omitting them did not dramatically degrade the results. He estimated the strength of steric interactions to be 1.5 kJ atom$^{-1}$ mol$^{-1}$ (assuming an average solvent-accessible surface area of 10.0 Å$^2$ atom$^{-1}$), that of hydrogen bonds to be 3.0 kJ bond$^{-1}$ mol$^{-1}$, and that of electrostatic interactions to be 6.8 kJ bond$^{-1}$ mol$^{-1}$. Böhm did not separately consider the weaker hydrogen bonds formed by halogens and by sulfur; in the work described in this paper, these were arbitrarily assigned half the strength of the more common nitrogen–oxygen hydrogen bonds. There are many other, widely varying estimates of the strengths of steric and hydrogen-bonding effects in the literature. However, the experiments described below found that varying the weighting parameters across a broad range had surprisingly little effect on the results.

### *Alignment of conformers*

Given two conformers, the method must derive a dissimilarity score by finding an appropriate relative alignment and applying the pose dissimilarity measure. One could of course make the dissimilarity very large by altogether avoiding overlap of the two conformers; the opposite, overlapping them as well as possible in order to minimize their dissimilarity, is the goal. We can think of this as asking how well one conformer can mimic the other. A pair of perfect mimics would bind the exact same set of target sites, and so would contribute no more diversity than either conformer individually.

The applications reported here selected sets of combinatorial library side-chain precursors rather than whole molecules for screening. Although any relative alignment of whole molecules is legal, in the alignment of side chains one should only consider alignments that are consistent with the way in which they would be attached to a scaffold. Fortunately, this constraint decreased the computational cost of alignment.

The constraint is implemented by designating in each candidate side-chain precursor a 'handle atom', the atom that would be bonded to the scaffold, and superimposing the handle atoms of the two conformers being compared. This leaves the conformer free to rotate about three axes, as with a ball joint; translations are not permitted. The rationale for this is explained further in the Discussion section.

Optimal alignments can be found using any of several standard continuous function minimization techniques, because the dissimilarity measure is continuous with respect to changes in alignment. The method uses gradient descent to minimize dissimilarity with respect to these three rotational pose parameters. In effect, it applies a synthetic force to the conformers that is proportional to the dissimilarity and that tugs them into a minimally dissimilar alignment. The specific gradient method used is steepest descent, with minimization judged to have converged when the magnitude of the gradient falls below 0.01 or when the decrease in dissimilarity falls below 0.01 units.

To give good results, gradient descent requires reasonably good starting points, lest it get stuck in local minima. The program uses two starting points for each pair of conformers, applies gradient descent to each, and chooses the better alignment. These starting points intuitively correspond to treating the two conformers as though they were flat and superimposing the planes they define. Specifically, for each conformer the program finds the atom furthest from the handle, and then the plane passing through this atom and through the handle that minimizes the rms distance of all the other atoms to it. It then superimposes the handles of the two conformers, places the two atoms most distant from the handles as close to each other as possible, and then rotates the planes to superimpose them. There are two such rotations, at 180° from each other; they define the two starting alignments for gradient descent.

*Diversity measure*

In screening a set of molecules M, one assays the set C of all the conformers of all the molecules. The method defines the diversity of M in terms of the pairwise dissimilarities of the set C. We may think of each conformer in C contributing to the diversity of M by the extent that it differs from other members of C. So the method considers, for each conformer, the least dissimilar conformer (which may be of the same or a different molecule), and sums the resulting minimum dissimilarities. To a first approximation,

$$\text{diversity}(\text{M}) = \sum_{\text{m}\in\text{M}} \sum_{\text{c}\in\text{C}(\text{m})} \min_{\text{c}'\in\text{C}} (\text{dissimilarity}(\text{c},\text{c}'))$$

where C(m) is the set of conformers of molecule m, and c and c' are distinct.

This definition accounts for only the enthalpic contribution to binding. Highly flexible molecules have more conformers to sum over and so will generally contribute more diversity. It is true that flexible molecules can bind to more different sites than rigid ones, but one may not want to select only highly flexible molecules, because they have a high entropic cost for binding due to conformational fixation. Accordingly, the method applies an entropic penalty term:

$$\text{diversity}(\text{M}) = \sum_{\text{m}\in\text{M}} \left[ \left( \sum_{\text{c}\in\text{C}(\text{m})} \min_{\text{c}'\in\text{C}} (\text{dissimilarity}(\text{c},\text{c}')) \right) - \text{T}\Delta\text{S}(\text{m}) \right]$$

The algorithm estimates TΔS(m), using the results of Böhm, to be 1.2 kJ per rotatable bond in m. (Of course, this value depends on the assay temperature, and is effectively averaged over the assays used in Böhm's data set. Presumably, however, it represents roughly biological temperature.) This accounts for only the fixation entropy; a better measure would take into account desolvation entropy as well. A weighted logarithm of the number of low-energy conformers might provide a better estimate of fixation cost than the weighted number of rotatable bonds. The number of low-energy conformers is available in the program described in this paper, but was not used by Böhm.

*Selection algorithm*

The goal of the selection algorithm is, given a large pool of candidate molecules, to select a highly diverse subset of a given size. The algorithm constructs this subset by repeated addition of the molecule that will add the most diversity to those selected so far (Fig. 2). The first molecule chosen is that which would add the most diversity to the null set, i.e. the molecule with the greatest 'intramolecular' diversity among its conformers. (Alternatively, an initial set of molecules may be supplied by the user, and added to using this algorithm.) The second molecule is chosen by evaluating the diversity that would be added by each remaining candidate to the existing set and selecting the one that would add the most.

This algorithm is termed a 'greedy' algorithm because it always chooses the candidate that looks best individually, without considering the effect of that choice on subsequent choices. Greedy algorithms are not guaranteed to yield an optimal final set, but are computationally efficient and in practice generally approach optimality [13]. Finding the truly optimal set is an NP-complete problem, and is therefore provably computationally infeasible, as it would require the consideration of exponentially many subsets of the desired size.

The second formula for diversity in the previous section is still not complete. Under this definition, each conformation of a newly added molecule adds diversity in proportion to its dissimilarity from the most similar conformation in the expanded set. Now consider adding a molecule with two very similar conformations to a set of other molecules, which are all very different from the new molecule. There would be very little added diversity, because each of the two new conformations would be most similar to each other, and the molecule as a whole would add diversity only equal to twice this small dissimilarity (once for each of the two conformations). The computation of Fig. 2 corrects this problem by crediting the new molecule for its difference from the existing ones by, in effect, adding its conformations to the existing ones one by one, starting with the one most similar to the existing ones. The first conformation is compared only with those of the previously selected molecules, and is added to the set **examined**; the second conformation is compared with the previously selected molecules and also with the first conformation. Thus, the total diversity added will be the dissimilarity of the first conformation from the most similar conformation among the previously selected molecules, plus the dissimilarity of the two conformations of the new molecule.

The algorithm as described in Fig. 2 requires a comparison on each step of all the conformers of each candidate with all the conformers of the previously selected molecules and with each other (to account for intramolecular diversity). This quadratic cost is potentially very expensive. Two optimizations keep the cost manageable: (i) it caches the results of all comparisons, so that any given pair need never be compared more than once; (ii) it avoids reevaluating most molecules on each addition cycle. The second optimization depends on the observation that when a molecule is added to the set, the diversity that would be added by each of the remaining candidates either decreases (if some conformer of the newly selected molecule is less dissimilar to some conformer of the candidate than any conformer previously in the set) or stays the same. Thus, the diversity that would have been added by the candidate to the set before the addition of the new member is an upper bound on the diversity that would be added to the newly enlarged set.

To exploit this fact, the program keeps a sorted list of all the candidates together with an upper bound on the diversity they could add to the set being constructed (Fig. 3). To select the molecule that would add the most diversity, the algorithm examines the first molecule on the list

```
define select(number, candidates):
  selected ← {}
  repeat number times:
    for each candidate molecule m
      examined ← {}
      for each conformation c of m
      min_selected(c) ← the minimum
        over c' in all conformations
        of all members of selected
        of dissimilarity(c,c')
    for each conformation c of m,
        sorted in ascending order by min_selected(c)
      add c to examined
      min_examined(c) ← the minimum
        over c' in examined
        of dissimilarity(c,c')
      min(c) ← the lesser of min_selected(c) and min_examined(c)
    added_diversity(m) ← (the sum, for c in C(m), of min(c)) -
      (constant * number of rotatable bonds in m)
    selected ← selected plus the candidate with largest added-diversity
return selected
```

Fig. 2. Pseudocode for the selection algorithm. Variables are boldfaced.

```
define optimized_select(number, candidates):
  selected ← {}
  for each m in candidates, upper_bound(m) ← infinity
  repeat number times:
    loop:
      m ← the first member of candidates
      if added_diversity(m) > upper_bound(the second member of candidates)
          then selected ← selected plus m
            candidates ← rest of candidates
            exit loop
          else
            reinsert m in candidates in the appropriate place
            with added_diversity(m) as its new upper bound
            continue looping
return selected
```

Fig. 3. Pseudocode for the optimized selection algorithm.

and recomputes the diversity it would add to the current set. If this number is greater than that of the bound on the second molecule on the list, the first molecule must in fact be the best, because the bound on the second molecule is the highest remaining upper bound. If the bound on the second molecule is greater than the actual measure for the first molecule, the first molecule is moved in the list, with its actual value, to its correct sorted position. Then the algorithm begins again with the new first molecule (previously the second). In practice, the algorithm typically only needs to reevaluate a few molecules on each cycle, thereby saving a huge amount of work.

Even with these optimizations, the selection of a set of 20 molecules from a pool of about a thousand (described in the next section) involves millions of comparisons of conformers (and many more millions of comparisons of poses during alignment).

## Results

To verify that the dissimilarity computation gives reasonable results, it was first applied to the natural amino acid side chains. Then, the greedy algorithm was used to select a diverse set of carboxylates. The diversity of this set was compared with randomly chosen sets. Finally, a variety of experiments examined both the effect of varying numerical parameters and the performance of the alignment method.

### *The natural amino acid side chains*

By applying the dissimilarity measure of this paper to the natural amino acid side chains, a familiar set, one can verify that its results are intuitive. The dissimilarity between two amino acids was defined as the minimum dissimilarity over all their conformers. The results appear in Table 1. The underlined entries in each row denote the most similar distinct side chain. All of these seem to comport with intuition. The two smallest, A and G, are much more similar to each other than to any others; the next most similar is V, the next smallest. The greasy nonaromatic side chains, V, L, I, and M, form a block, with the smaller three (V, L, and I) more similar to each other than to M. The small hydroxy and thiol amino acids S, C, and T are also grouped together, and are more like the similarly sized but nonpolar groups V, L, I, M than like any others. F is best approximated by M (which is also greasy and about the same size). It is also relatively similar to V, L, and I, and to Y, from which it differs by only a hydroxyl. Y and W, the uncharged aromatic side chains other than F, are best approximated by F. The two amides (N and Q) and the two acids (D and E) are most similar to each other. The positively charged side chains (H, K, and R) go together, with the nonaromatics most similar to each other. The next most similar to H is W, which is also aromatic and has a nitrogen in a similar position. Q is almost as similar to K as R is, differing only in the third digit; although its nitrogen is uncharged, it is a better steric fit, and has only one unmatched polar functionality (its oxygen), where R has two unmatched uncharged nitrogens.

### *Selection of diverse carboxylates*

This experiment selected a set of carboxylates to serve as side-chain precursors for a combinatorial library based on a differentially protected polyamine scaffold and condensation chemistry. These carboxylates were selected from the computer-readable Available Chemicals Directory (ACD) of 143 116 commercially available molecules. (The ACD database is a product of MDL Information Systems Inc., San Leandro, CA, U.S.A. The experiments were carried out using the 92.1 release.)

The ACD was first screened for molecules that had a single carboxy group, no more than 20 non-hydrogen atoms, and none of a set of structural features that might render them toxic or synthetically unsuitable for incorporation into the library. The size cutoff (20 atoms) was chosen on the basis of the target size for the products of combinatorial synthesis.

These criteria were met by 4961 molecules. This was too many to process given the available computational resources, so this set was filtered topologically, eliminating molecules that differed by a single atomic substitution. Two molecules were considered 'neighbors' if, ignoring hydrogens and bond order, they were identical, or identical except that a single atom was changed to a different element, or identical except that one had an additional atom that was bonded to only one other non-hydrogen atom. Pairs of neighboring molecules are sufficiently similar that it is unlikely that both would end up in a small diverse set. Successively eliminating the molecule with the greatest number of neighbors from the set, until there were no neighboring pairs remaining, resulted in a pool of 2238 molecules.

From these, all that had more than five rotatable bonds (not counting bonds in rings of six or fewer atoms) were eliminated, and the remainder were conformationally searched. Eliminating excessively flexible molecules was required to make conformational search practical; even so, about three CPU weeks are required for this many molecules. For the same reason, flexible rings were treated as rigid and not searched. From the resulting set, all those with more than 30 conformers were eliminated, yielding 1371 molecules.

From the 1371 molecules, 20 were selected using the algorithm of this paper (Fig. 4). There are a great many factors affecting toxicity and synthetic suitability, and not all could practically be incorporated into the initial screen. Therefore, the algorithm ran in an interactive mode in which it would, after tentatively selecting each next molecule from the candidate pool, ask the user whether it was suitable, and find the next best alternate if not. The run time for selection is about one day.

The algorithm is completely insensitive to atom type and to molecular connectivity, except insofar as these affect the distribution of polar functionalities and three-dimensional shape. Therefore, there is no guarantee that the resulting set will be diverse in the more familiar sense of having a wide variety of chemical groups and ring structures. Nevertheless, this set does contain all the elements allowed (H, C, N, O, P, S, F, Cl, Br, and I) except bromine. (All molecules containing other elements were excluded on the grounds that they would likely be toxic or synthetically problematic.) The set contains 10 different ring systems, both aromatic and aliphatic, and molecules with single, double, and triple bonds.

It is curious that so many (18/20 = 90%) of the molecules contain aromatic rings. The algorithm has no specific bias toward aromatics, and only 27% of the molecules in the ACD contain aromatic rings. However, 62% of the carboxylates in the ACD are aromatic, as are 82% of the set of carboxylates with 30 or fewer conformers from which the 20 were chosen. The difference between 82% and 90% represents a difference of less than two molecules and is probably not significant; in runs selecting diverse molecules with other coupling functionalities, the algorithm has shown an apparent bias *against* aromatic molecules.

The only molecules with anionic side chains are the

TABLE 1
COMPARISON OF NATURAL AMINO ACID SIDE CHAINS

| | G | A | V | L | I | M | S | C | T | F | Y | H | W | N | Q | D | E | K | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 0.0 | 1.5 | 4.5 | 6.0 | 6.0 | 7.5 | 9.0 | 6.0 | 10.5 | 10.5 | 18.0 | 18.8 | 18.0 | 12.0 | 13.5 | 15.8 | 17.3 | 14.3 | 23.3 |
| A | 1.5 | 0.0 | 2.2 | 3.8 | 3.8 | 5.4 | 6.8 | 4.2 | 8.1 | 8.2 | 15.7 | 16.5 | 15.7 | 10.3 | 11.7 | 14.1 | 15.7 | 12.2 | 21.1 |
| V | 4.5 | 2.2 | 0.0 | 3.3 | 1.2 | 4.9 | 7.2 | 4.0 | 7.3 | 7.2 | 15.0 | 15.2 | 14.3 | 8.9 | 10.1 | 12.6 | 14.3 | 11.2 | 18.8 |
| L | 6.0 | 3.8 | 3.3 | 0.0 | 2.4 | 4.0 | 8.4 | 5.3 | 9.2 | 5.2 | 12.6 | 12.5 | 10.8 | 7.4 | 9.5 | 11.3 | 12.6 | 10.2 | 18.1 |
| I | 6.0 | 3.8 | 1.2 | 2.4 | 0.0 | 4.0 | 8.5 | 5.3 | 8.6 | 5.7 | 13.2 | 14.0 | 12.4 | 7.9 | 9.2 | 11.6 | 12.6 | 10.4 | 17.4 |
| M | 7.5 | 5.4 | 4.9 | 4.0 | 4.0 | 0.0 | 8.0 | 4.5 | 9.7 | 4.7 | 11.9 | 13.6 | 10.7 | 6.3 | 6.7 | 7.5 | 7.6 | 9.3 | 17.6 |
| S | 9.0 | 6.8 | 7.2 | 8.4 | 8.5 | 8.0 | 0.0 | 3.2 | 1.3 | 12.4 | 19.8 | 17.6 | 19.1 | 10.4 | 12.7 | 13.3 | 15.1 | 13.3 | 20.8 |
| C | 6.0 | 4.2 | 4.0 | 5.3 | 5.3 | 4.5 | 3.2 | 0.0 | 4.3 | 9.1 | 16.6 | 13.9 | 15.5 | 7.5 | 9.8 | 9.6 | 12.3 | 11.4 | 17.0 |
| T | 10.5 | 8.1 | 7.3 | 9.2 | 8.6 | 9.7 | 1.3 | 4.3 | 0.0 | 13.2 | 20.6 | 18.7 | 20.2 | 11.2 | 12.0 | 14.3 | 14.9 | 12.5 | 19.5 |
| F | 10.5 | 8.2 | 7.2 | 5.2 | 5.7 | 4.7 | 12.4 | 9.1 | 13.2 | 0.0 | 7.0 | 12.9 | 10.3 | 9.5 | 10.6 | 15.9 | 13.9 | 10.5 | 17.1 |
| Y | 18.0 | 15.7 | 15.0 | 12.6 | 13.2 | 11.9 | 19.8 | 16.6 | 20.6 | 7.0 | 0.0 | 19.8 | 16.3 | 16.9 | 14.8 | 20.7 | 16.3 | 12.5 | 16.9 |
| H | 18.8 | 16.5 | 15.2 | 12.5 | 14.0 | 13.6 | 17.6 | 13.9 | 18.7 | 12.9 | 19.8 | 0.0 | 10.9 | 13.9 | 13.0 | 21.7 | 23.7 | 10.5 | 12.1 |
| W | 18.0 | 15.7 | 14.3 | 10.8 | 12.4 | 10.7 | 19.1 | 15.5 | 20.2 | 10.3 | 16.3 | 10.9 | 0.0 | 13.6 | 12.4 | 20.1 | 18.6 | 12.5 | 17.6 |
| N | 12.0 | 10.3 | 8.9 | 7.4 | 7.9 | 6.3 | 10.4 | 7.5 | 11.2 | 9.5 | 16.9 | 13.9 | 13.6 | 0.0 | 3.4 | 9.6 | 11.9 | 10.0 | 17.2 |
| Q | 13.5 | 11.7 | 10.1 | 9.5 | 9.2 | 6.7 | 12.7 | 9.8 | 12.0 | 10.6 | 14.8 | 13.0 | 12.4 | 3.4 | 0.0 | 11.4 | 9.8 | 8.9 | 15.4 |
| D | 15.8 | 14.1 | 12.6 | 11.3 | 11.6 | 7.5 | 13.3 | 9.6 | 14.3 | 15.9 | 20.7 | 21.7 | 20.1 | 9.6 | 11.4 | 0.0 | 6.3 | 19.3 | 27.9 |
| E | 17.3 | 15.7 | 14.3 | 12.6 | 12.6 | 7.6 | 15.1 | 12.3 | 14.9 | 13.9 | 16.3 | 23.7 | 18.6 | 11.9 | 9.8 | 6.3 | 0.0 | 19.0 | 25.6 |
| K | 14.3 | 12.2 | 11.2 | 10.2 | 10.4 | 9.3 | 13.3 | 11.4 | 12.5 | 10.5 | 12.5 | 10.5 | 12.5 | 10.0 | 8.9 | 19.3 | 19.0 | 0.0 | 8.9 |
| R | 23.3 | 21.1 | 18.8 | 18.1 | 17.4 | 17.6 | 20.8 | 17.0 | 19.5 | 17.1 | 16.9 | 12.1 | 17.6 | 17.2 | 15.4 | 27.9 | 25.6 | 8.9 | 0.0 |

Entries are the minimum dissimilarity across conformers of the two molecules. The smallest value in each row is underlined. Proline was omitted because it does not form a single side chain, as the other amino acids do. Histidine was protonated.

two phosphorus compounds **126103** and **50080**. All nitro compounds and dicarboxylates were excluded due to toxicity risks or incompatibility with the synthetic protocol. This leaves only sulfur and phosphorus anions.

All of the molecules selected are relatively large; the smallest, **50080**, has 11 non-hydrogen atoms, and most are closer to the upper limit of 20. The algorithm has no specific bias in favor of larger molecules; the dissimilarity measure is symmetric, so in considering a pair of molecules, one of which is significantly larger than the other, they will be 'rewarded' equally for their failure to overlap. We will see below that the algorithm has a probably unwarranted bias toward flexible molecules; it is probably as a consequence of this bias that no very small molecules were chosen, since small molecules cannot be as flexible as larger ones.

*Evaluation of random sets*

To evaluate the possible advantage of using the rational selection method, the diversity measure was applied to five randomly selected sets of carboxylates. The set of Fig. 4, selected to be maximally diverse, has a total measured diversity of 8769 units. The random libraries had total diversities of 4332, 4645, 5059, 5182, and 5433, for an average of 4930 units. This average is 56% of the diversity of the set of Fig. 4. The set of Fig. 4 achieves a greater diversity (5058 units) in only its first 10 compounds.

These results must be interpreted carefully, since there are reasons both to think it may over- and understate the advantage of the method over random selection. Most importantly, the algorithm tries to optimize its own diversity measure, which is the very measure used to evaluate

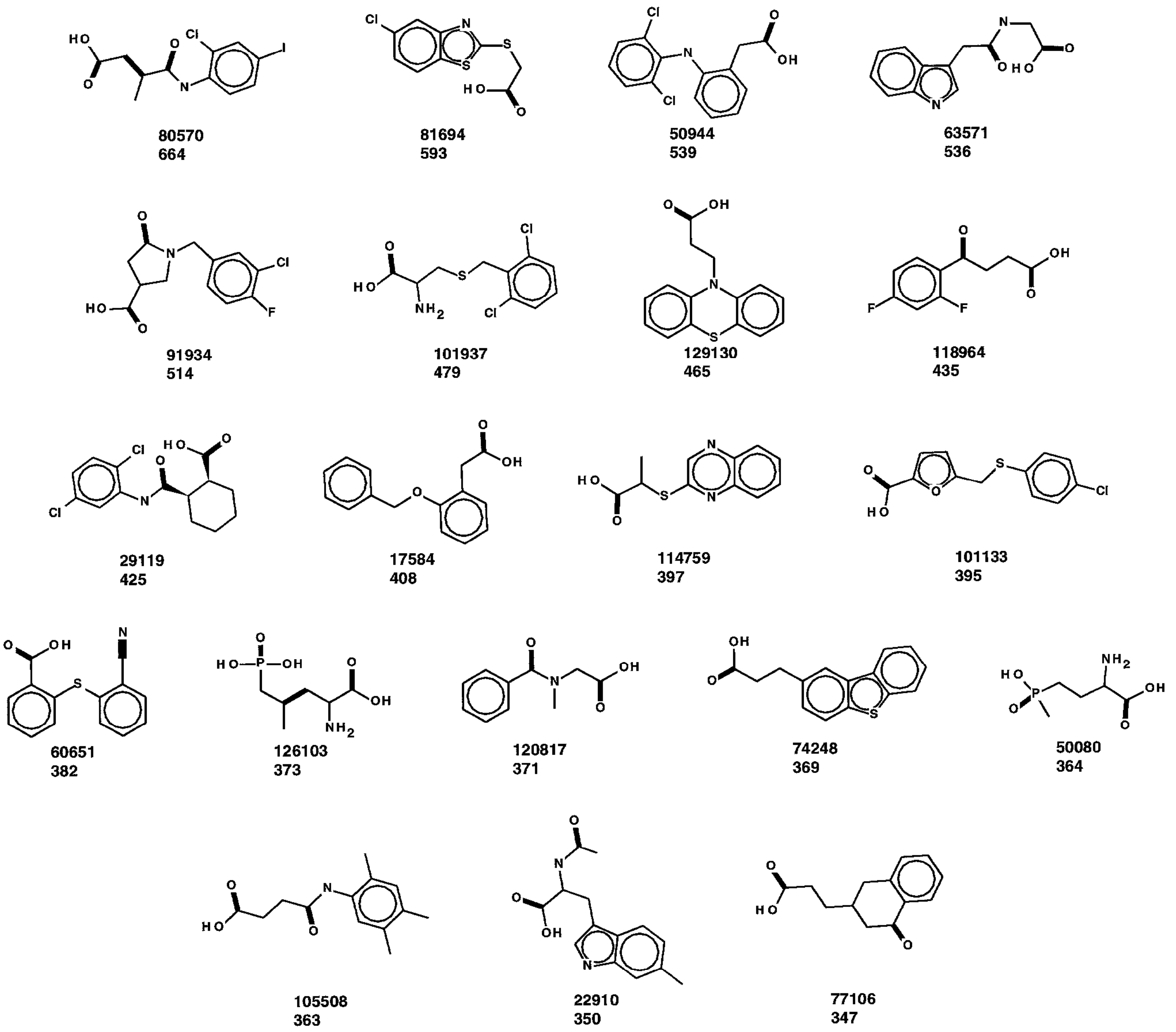

Fig. 4. The 20 selected carboxylates, shown in the order selected. The first number beneath each molecule is its ACD identifier; the second is the amount of diversity it added to the previously selected molecules, rounded to the nearest unit.

TABLE 2
SENSITIVITY OF RESULTS TO w PARAMETERS

| $w_{polar}$ | $w_{charged}$ | PRCC | Mean rank change |
|---|---|---|---|
| 1.5 | 3.0 | 0.92 | 6.23 |
| 2.25 | 4.5 | 0.96 | 3.00 |
| 2.75 | 5.5 | 0.98 | 1.73 |
| **3.0** | **6.8** | **[1.0]** | **[0.0]** |
| 5.0 | 10.0 | 0.94 | 4.38 |
| 9.0 | 18.0 | 0.89 | 8.13 |

Values for the weighting of uncharged polar and charged interactions are shown; the value for steric interactions was held constant at 1.5. The boldfaced entries represent default values.

the random sets. To the extent that the measure is inaccurate, the algorithm will optimize the wrong thing and this comparison experiment will overstate the true diversity of the selected set relative to that of the random sets. (The actual range of diversities achievable may be less than that measured by the algorithm. In this case, even a truly optimally diverse set of molecules might be less of an improvement over a random set than these experiments suggest. Of course, the opposite may also be true.) In other words, the experiment controls only for the possibility that the rational selection algorithm does not significantly improve on random selection, relative to the given diversity measure; it does not control for the possibility that the diversity measure itself is incorrect.

On the other hand, the random sets were chosen from the set of 1371, which had already been topologically filtered to remove molecules differing only in bond order or by single-atom substitutions. This suggests that the results may understate the advantage. Also, to control for the bias of the method towards molecules with more conformers, the random selection used in this experiment forced a choice of molecules that have the same numbers of conformers as do the molecules in Fig. 4. The molecules that added the most diversity in the selected set had 26 or 27 conformers. Since there were not many molecules in the set of 1371 that had 26 or 27 conformers, the random sets tended to overlap significantly with the selected set. In effect, forcing the number of conformers to match gave the random selection process 'hints' about how to optimize diversity and the 'randomly' chosen sets were more diverse than they would have been in an ideal experiment.

*Parameter sensitivity*

This section describes experiments that examined the effect of varying several numerical parameters, i.e., those taken from Böhm (the relative strengths of various types of interactions and of the entropic cost of fixation) and the distance threshold parameter t.

The relative strength of polar and steric interactions has been debated extensively in the literature [14]. The first experiment varied the ratio of the relevant w parameters over a sixfold range. To measure the effect of these parameters, the algorithm was applied to a random subset of carboxylates from the set of 1371 used in the selection experiment reported above. The intramolecular diversities of these molecules were recomputed for each parameter setting. The relative strengths of ionic and hydrogen-bonding interactions were held constant at 2.0; this is approximately correct and there are few enough formally charged groups that the results are unlikely to depend strongly on the ratio. To control for the algorithm's bias toward molecules with many conformers, 100 were selected that all had the same number of conformers (nine).

Table 2 shows the results. Because the computed diversity is roughly proportional to the w values, a comparison of the relative magnitudes of the results is not meaningful; instead the table shows the effect of the parameter changes on the rank ordering of the molecules as sorted by their intramolecular diversity. For each parameter setting, it shows the pair rank correlation coefficient (PRCC) of the results relative to the default values. The PRCC [15], a measure closely related to Kendall's $\tau$ [16], is defined as the probability, given two randomly chosen members of a set, that they are ranked the same way in both orderings. The table shows that, even with the polar w values increased threefold, there is an 89% probability that the ranking of two randomly chosen molecules was unchanged. The table also shows the mean change in rank as a result of a parameter change. For example, halving the strength of polar interactions results in a mean rank change of 6.23, meaning that, on average, each molecule slid about six places in the 100-molecule-long list. The entries in the table corresponding to the largest and smallest w parameter ratios are almost certainly greater and smaller than the actual physical ratio. These results suggest that the algorithm is relatively insensitive to the values of the w parameters.

Similar experiments examine the effect of changing the distance threshold t (Table 3). The true value of t depends on both the rate at which intermolecular forces drop off as a function of distance and on the extent to which a ligand and its binding site will flex to accommodate each other. The latter effect (which probably dominates) is difficult to estimate. Recall that the implementation takes t = 2.0 Å; this value was chosen arbitrarily. However, the table shows that decreasing it to 1.0 Å or increasing it to as much as 5.0 Å has surprisingly little effect on the results.

The most problematic parameter is the entropic cost of fixation, per rotatable bond. Table 4 examines the effect of increasing the entropic parameter. (It is based on a different control set of 100 molecules chosen at random from the set of 1371. This control set includes molecules with varying numbers of torsions, and therefore varying numbers of conformations; otherwise, changing the entropic parameter would have no effect.) Estimates of the entropic cost of fixing a single torsion vary by more than

TABLE 3
SENSITIVITY OF RESULTS TO THE t PARAMETER

| t | PRCC | Mean rank change |
|---|---|---|
| 1.0 | 0.94 | 4.12 |
| **2.0** | **[1.0]** | **[0.0]** |
| 3.0 | 0.97 | 2.71 |
| 5.0 | 0.94 | 4.77 |

The t parameter is the 'soft threshold' distance at which the strength of interactions drops close to zero. The boldfaced entries are default values.

a factor of 2, but Böhm's value of 1.5 kJ cannot be very wrong. Molecules with five or fewer rotatable bonds cannot have a fixation cost of more than, say, 20 kJ. However, as we saw above, some have an intramolecular diversity of several hundred units, dwarfing the entropic term. Not surprisingly, then, the table shows that the entropic cost must be raised to clearly nonphysical values in order to substantially affect the results. The next section discusses this further.

Overall, the insensitivity of the algorithm to parameter changes gives greater confidence in its meaningfulness, and may decrease concern about the physical inaccuracies in the dissimilarity function.

## Discussion

*Empirical tests*

Although the algorithms presented here have a rational physical basis, and although the results make intuitive sense, one may still be skeptical that they will behave as expected in practice. How could one test them empirically?

We can measure the quality of a library by screening it against a statistically large enough group of assays and looking at the number and potency of the hits. The ideal experiment, then, would synthesize several combinatorial libraries on the same scaffold with varying side chains and assay them against a statistically meaningful set of targets. One library would be designed to maximize diversity, one to minimize it, and a few would be chosen at random. We could then see how much improvement a selection algorithm offers over random choice, if any.

Such a prospective experiment would, however, be very expensive, requiring the synthesis and assay of enormous numbers of compounds specifically chosen to be of indifferent quality. One could, as an alternative, perform retrospective experiments by analyzing data from screenings carried out for other purposes. One could take from a set of previously selected molecules various equally sized subsets: a minimally diverse subset, a maximally diverse subset, and several random subsets. One could then look at the quality of hits in these sublibraries. Again, the prediction is that in a statistically large enough set of screening data, the maximally diverse subsets will have more and better hits than the random and minimally diverse subsets. Although such experiments are planned, enough data to make them statistically meaningful are not yet available.

*Alignment and conformational analysis*

To evaluate the diversity added by a candidate molecule to a set of other molecules, one must ask (i) to what class of sites the molecule will bind, and (ii) how well other molecules already in the set will bind to those same sites. To answer the first question, the method finds all low-energy conformers of the candidate molecule, which form a set of shapes complementary to the sites it will bind. To answer the second question, the method asks how well conformers of the other molecules can mimic those of the candidate. This involves aligning them as well as possible to the candidate conformers.

The alignment method of this paper allows for three rotational degrees of freedom, about the handle atom at which a side chain is coupled to the scaffold. This is not physically correct. If we suppose the scaffold itself is rigid, the correct formulation would allow only one degree of freedom, which is the torsion about the bond joining the side chain to the scaffold. However, some scaffolds are quite flexible, so that differences between side chains can be accommodated by scaffold flexing. Consider, for example, the aspartic and glutamic acid side chains, which differ by a methylene in length before their carboxy groups. Were these side chains attached to a rigid scaffold, they would bind quite differently. Were they attached to a very flexible scaffold, however, and if we suppose that the binding of the ligand as a whole does not fully constrain the conformation of the scaffold, the aspartic and glutamic acid side chains might bind virtually to the same set of cationic sites. The scaffold would flex to reach further toward the cation in the case of glutamic acid, and would shift the attachment site away from the cation in the case of aspartic acid, which is one methylene shorter. In a library based on a rigid scaffold, then, it might make sense to use both these side chains, whereas they are liable to be redundant on a flexible scaffold.

The algorithm does not yet incorporate a principled method for addressing scaffold flexibility. Conformational

TABLE 4
SENSITIVITY OF RESULTS TO THE ENTROPIC COST PARAMETER

| Cost | PRCC | Mean rank change |
|---|---|---|
| **1.5** | **[1.0]** | **[0.0]** |
| 2.4 | 0.99 | 0.98 |
| 3.5 | 0.98 | 1.7 |
| 10.0 | 0.94 | 4.86 |
| 50.0 | 0.75 | 18.3 |

Values for the entropic cost are given in kJ/rotatable bond/mol. The boldfaced entries are default values.

TABLE 5
NUMBER OF CONFORMERS OF SELECTED CARBOXYLATES

| Molecule | Conformers | Molecule | Conformers |
|---|---|---|---|
| **80570** | 27 | **114759** | 25 |
| **81694** | 27 | **101133** | 24 |
| **50944** | 27 | **60651** | 24 |
| **63571** | 26 | **126103** | 23 |
| **91934** | 27 | **120817** | 30 |
| **101937** | 18 | **74248** | 26 |
| **129130** | 26 | **50080** | 24 |
| **118964** | 24 | **105508** | 21 |
| **29119** | 14 | **22910** | 21 |
| **17584** | 29 | **77106** | 25 |

Molecules are listed in the order selected.

analysis of the scaffold might form the basis of such a method. In the meantime, allowing rotations about the handle atom provides an intuitively simple method that also represents a compromise between a single torsional degree of freedom and the full six degrees of freedom mathematically possible.

Available computational resources and a quadratic cost forced limitation of the carboxylate side-chain precursors to 30 conformers. Since the scaffold to which they are attached is quite flexible, this restriction seems reasonable. However, current work is examining ways to lift the restriction by extrapolating the diversity measure from a subsample of the set of conformers.

*Entropy of fixation*

As mentioned previously, the diversity metric favors flexible molecules, because it sums over conformers. Table 5 shows the number of conformers for each chosen molecule. Recall that the number of allowable conformers was capped at 30. Although there were not 20 30-conformer molecules to choose among, the algorithm could have selected a set consisting only of molecules with 30, 29, or 28 conformers. It did not do so; the average number of conformers per selected molecule is 24.4, and the minimum is 14. Therefore, we see that the preference is not absolute by any means. Still, there is a marked bias; the correction for the entropy of fixation does not adequately balance the enthalpic contribution to diversity. Why not?

First, it may indeed be that flexible side chains are preferable in screening. They will, after all, bind to a greater variety of sites, and so are more likely to bind to the site being assayed. More empirical data are needed to know how this effect trades off against the entropy of fixation. Intuitively, though, one may suspect that they roughly balance, so that there is no strong preference for either flexible or inflexible molecules.

That the method described here diverges from this balance is probably due to a treatment of enthalpy that is not physically correct. It measures fixation entropy in units of energy, but the 'enthalpic' component of diversity – the measure of dissimilarity in steric and polar functionality – is not in units of energy. Although the method weights these effects in kJ, a sum of dissimilarities in kJ does not yield units of kJ; this sum does not have a direct physical interpretation. (A tempting fix might be to charge the entropic cost against each conformer. This, however, is also not physically correct, and in practice results in a strong bias toward rigid molecules. It also results in large negative added diversity values, which make no sense; diversity must be nondecreasing as molecules are added to a set.)

To reconcile enthalpic and entropic effects, one must measure both in the same units, which must be units of energy. This will yield a total diversity measure in units of energy. What would this mean? We might regard the diversity of a library as the probabilistically expected binding energy of its best molecule when it is screened against an unknown target. In other words, we could imagine screening the library against all possible protein sites and taking its diversity to be the average of the sites of the binding energy of the best hit on each site. Current work is developing a method for estimating the enthalpic contribution to this value.

## Conclusions

The combination of combinatorial library synthesis and high-throughput screening holds the promise of dramatically decreasing the time and cost required to generate high-quality leads for drug discovery. This paper presents a means for rationally designing combinatorial libraries in order to increase their likely effectiveness.

The methods presented here, like the similar 'scoring' functions used in docking [12], are grounded in the known physical chemistry of ligand–protein interactions. This physical basis provides intuitive and theoretical justification for the approach. Although the method incorporates several assumptions and approximations that have no direct physical basis, sensitivity experiments suggest that these are not critical. However, the methods must be validated experimentally by the analysis of statistically meaningful quantities of screening data, before one can be fully confident in them. Currently, several nonpeptidic combinatorial libraries using side chains chosen with the algorithms presented here have been synthesized and screened against a variety of targets, but as yet there are not enough data to quantitate the value of the method.

How much can a rational selection of combinatorial library side chains help? Although only empirical evidence can answer this question definitively, experiments comparing optimal and random sets suggest that the method may improve side-chain diversity by a factor of 2. In other words, diversity equivalent to that derived without using rational design may be achieved with a set of side chains half as large. This becomes significant in the synthesis and screening of libraries of trimers or tetramers, in which halving the number of side chains may make it possible to achieve equivalent results with 1/8th or 1/16th the effort.

## Acknowledgements

I thank Nathan Collins, Lutz Giebel, Jim Hauske, Ajay Jain, Chuck Johnson, Teri Klein, Tomas Lozano-Perez, Mike Ross, Mike Venuti, and Peng Zhou for helpful discussions, and Mike Ross, Mike Venuti, John Walker, and especially Jim Hauske for supporting this research.

## References

1 a. Gallop, M.A., Barret, R.W., Dower, W.J., Fodor, S.P.A. and Gordon, E.M., J. Med. Chem., 37 (1994) 1233.
b. Gordon, E.M., Barret, R.W., Dower, W.J., Fodor, S.P.A. and Gallop, M.A., J. Med. Chem., 37 (1994) 1386.
2 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
3 James, C.A. and Weininger, D., Daylight, 4.41 Theory Manual, Daylight Chemical Information Systems Inc., Irvine, CA, U.S.A., 1995.
4 Shemetulskis, N.E., Dubar, J.B., Dunbar, B.W., Moreland, D.W. and Humblet, C., J. Comput.-Aided Mol. Design, 9 (1995) 407.
5 Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
6 Boyd, S.M., Beverly, M., Norskov, L. and Hubbard, R.E., J. Comput.-Aided Mol. Design, 9 (1995) 417.
7 Chang, G., Guida, W.C. and Still, W.C., J. Am. Chem. Soc., 111 (1989) 4379.
8 Crippen, G.M. and Havel, T.F., J. Chem. Inf. Comput. Sci., 30 (1990) 222.
9 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in C, 2nd ed., Cambridge University Press, Cambridge, U.K., 1992.
10 Johnson, M.A. and Maggiora, G.M., Concepts and Applications of Molecular Similarity, Wiley, New York, NY, U.S.A., 1990.
11 Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., 33 (1993) 112.
12 Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
13 Garey, M.R. and Johnson, D.S., Computers and Intractability, Freeman, New York, NY, U.S.A., 1979.
14 a. Burley, S.K. and Petsko, G.A., Adv. Protein Chem., 39 (1988) 125.
b. Sharp, K.A., Nicholls, A., Friedman, R. and Honig, B., Biochemistry, 30 (1991) 9686.
c. Shirley, B.A., Stanssens, P., Hahn, U. and Pace, C.N., Biochemistry, 31 (1992) 725.
d. Williams, D.H., Searle, M.S., Mackay, J.P., Gerhard, U. and Mapleston, R.A., Proc. Natl. Acad. Sci. USA, 90 (1993) 1172.
15 Jain, A.N., Harris, N.L. and Park, J.Y., J. Med. Chem., 38 (1995) 1295.
16 Snedecor, G.W. and Cochran, W.G., Statistical Methods, Iowa State University Press, Ames, IA, U.S.A., 1989.