IcePick: A flexible surface based system for molecular diversity (draft/approved for release by Axys 11-6-97, reviewed J. Med. Chem 3-19-98 ref. JM970775R, revised 6-1-98, 8-7-98, 11-9-98)

JOHN MOUNT*†, JIM RUPPERT, WILL WELCH AND AJAY JAIN‡

Axys Pharmaceuticals, 180 Kimball Way, South San Francisco, CA 94080.

November 9, 1998

*Corresponding author.

†Present address: CombiChem Inc., 1804 Embarcadero Rd., Suite 201, Palo Alto, CA 94303.

‡Present address: Iconix Pharmaceuticals, 850 Maude Ave., Mountain View, CA 94043.

Abstract

IcePick is a system for computationally selecting diverse sets of molecules. It computes the dissimilarity of the surface-accessible features of two molecules, taking into account conformational flexibility. Then, the intrinsic diversity of an entire set of molecules is calculated from a *spanning tree* over the pairwise dissimilarities. *IcePick*'s dissimilarity measure is compared against traditional 2D topological approaches, and the spanning tree diversity measure is compared against commonly used variance techniques. The method has proven easy to implement, and is fast enough to be used in selection of reactants for numerous production-sized combinatorial libraries.

1 Introduction

Combinatorial chemistry is now a standard tool in drug development [1, 2, 3, 4]. Directed combinatorial chemistry involves the synthesis of families of analogs of active compounds, perhaps targeted towards a protein's x-ray crystal structure. Unbiased screening, in contrast, requires large libraries of compounds for assay against targets that may be unknown at the time of synthesis. The goal of the unbiased strategy is to maximize the probability that the library contains novel ligands for a broad variety of targets. Such a library should contain a diverse set of molecules, to minimize redundancy while providing a representative sample of the full range of molecules that could theoretically be made. This manuscript describes a computational approach towards the design of such libraries.

Given a large collection of molecules meeting any number of pre-specified conditions such as cost, availability, ease of synthesis or global physicochemical profile, the molecular diversity problem involves the selection of a smaller collection that best represents the larger set. The diversity problem is typically broken into two parts: "pairwise dissimilarity" and "set diversity". Pairwise dissimilarity measures the molecular distance between two molecules. Set diversity requires the selection of a diverse suset of molecules, and typically uses the pairwise measure in computing the overall diversity of a larger set.

IcePick's computation of molecular dissimilarity is based upon two previously validated techniques for representing and comparing molecules. The first technique is the shape-based binding site model used in *Compass* [5]. The second technique involves fast flexible conformational search in the presence of a binding site model, similar to the *Hammerhead* docking method [6]. Both of these techniques have been proven effective in biological test systems [5, 6, 7, 8].

Many other common approaches to molecular diversity [9, 10, 11, 12] attempt to to maximize the number of different values of global features such as logP, mass, or number of rings. *IcePick* considers the steric and electronic properties of 3D molecular surfaces, taking into account molecular flexibility. It is hoped that by selecting diverse molecular surfaces, *IcePick* more directly attacks the problem of maximizing the chance of finding leads in a variety of protein binding screens.

For the computation of set diversity, we introduce a new approach that uses "minimum weight spanning trees". A minimum weight spanning tree is the shortest way of interconnecting a set of points. For our molecular diversity application, it is used to give a measure of the overall "spread", or diversity, of a set of molecules. Because of this, the spanning tree method can estimate the intrinsic diversity of a set; it is not limited to comparing relative diversities of sets. Additionally, it eliminates the near-duplicate selections that are sometimes made by commonly-used variance or additive measures.

2 Methods

2.1 Pairwise Dissimilarity

The *IcePick* dissimilarity measure is based on the explicit comparison of surface accessible steric and polar features, as derived from simultaneous 3D conformational analysis of pairs of molecules. A representative set of low energy conformations of each molecule is generated. Each molecule is then flexibly docked into the other's molecular surface representation. This docking attempts to maximize the alignment of surface features such as hydrophobic surface area, hydrogen bond donors, and hydrogen bond acceptors. The dissimilarity is then computed from the set of optimized alignments.

The starting set of conformations is generated using random sampling and force field minimization. The goal is to generate a small set (< 15) of low-energy conformations that are low energy and representative" of the overall conformational profile of each molecule. The set is created as follows. Starting with a single minimized conformation, a large family of unminimized conformations is created by: 1) inverting chiral centers as appropriate, 2) substituting common conformations of flexible rings from a precomputed library, and 3) rotating torsion bonds, sampled at up to three angles. If this set is too large (> 100), then up to 100 members are sampled at random. These are all energy-minimized. Those with energies larger than 125 percent of the minimum energy are eliminated. The remainder are RMS filtered to choose up to 15 of the most "different".

This set of conformations is intended to be representative of the conformational range of a molecule while sparse enough for efficient computation. The set of conformations can contain either pure or mixed enantiomers as appropriate Chiralities are not altered during the subsequent alignments. We have not examined alternative conformation generation strategies in detail, though we believe that any method which produces a reasonably representative set of starting conformations will produce qualitatively similar overall results. In particular, we chose the DREIDING force field [13] for its convenience and efficiency; other force fields would be acceptable.

To estimate how well molecule B imitates molecule A, *IcePick* computes how well B flexibly imitates each precomputed conformation of A. The average of all these alignments is taken as a measure of how well molecule B is able to imitate an average conformation of molecule A. The overall similarity of A and B is calculated as the average of B's ability to imitate A and A's ability to imitate B. A flexible molecule is often able to imitate a rigid molecule while the rigid molecule is rarely able to imitate the flexible one, so these two terms can differ. We average A's ability to imitate B with B's ability to imitate A to get a symmetric score. Overall, the similarity measure is a blend of the best-fit and and average-fit between the two molecules. The similarity ranges from $0 - 1$, where 1 means identical. The distance, or *dissimilarity*, between A and B is 1.0 minus the similarity. We refer to the dissimilarity measure as the *IcePick* measure and denote it $\mathrm{d}(A, B)$). We note that our measure is not, in general, a metric, but this is a technical consideration.

The per-conformation dissimilarity score is computed using the flexible docking techniques of *Hammerhead* described in [6] and the molecular surface scoring of *Compass* [7, 14]. We briefly review the relevant portions of those approaches here.

Similarity between two molecules held in a *fixed* alignment is measured using their surface-accessible-features as measured by *Compass*Å molecule's surface is probed using two spherical shells of *probe points* surround the molecule at radii of 6 Å and 9 Å. There are 42 probe points spread evenly over each sphere. Three *features* are computed at each probe position: hydrophobicity, hydrogen bond donation, and hydrogen bond acceptance. This gives a total of 252 features. Each feature is assigned a value equal to the distance between that probe point and the nearest atom of the molecule possessing the appropriate characteristic (donor, acceptor, hydrophobic). Positively charged atoms are treated as hydrogen bond donors, negatively charged as acceptors. These features are computed for both of the molecules being compared. The difference in value of corresponding features on the two molecules represents the dissimilarity in their surfaces. In the case of polar features, preferred hydrogen bond orientations are also computed, and the magnitude of the angle between the orientations increases the difference between features. These 252 differences are smoothed slightly using a Gaussian weighting function, and then are summed and rescaled to give a dissimilarity score within a 0-1 range. This measure is defined for molecules in a fixed alignment, but it can also be used to optimize a given alignment. The features are differentiable almost everywhere, which means that gradients can be computed, and standard optimization techniques can be used to find simultaneous orientation and conformation parameters that minimize the difference between the two molecules. As was the case in *Compass*, *IcePick* uses a brief optimization step to "polish" the similarity score. This step varies the position, orientation, and torsion bonds of one molecule.

A coarse conformational search is applied in order to get a set of fixed alignments for the above similarity calculation. This requires finding the best possible alignment of one molecule to the other's surface representation. We have adapted the *Hammerhead* molecular docker's search engine to perform this task. For a fixed conformation of molecule A, molecule B undergoes an extensive flexible search. From a single starting conformation of B, many conformations are generated by a combination of: random orientation, rotation of torsion bonds, and substitution of precomputed ring conformations. During this process, self-penetrations are avoided, and chirality is maintained. The similarity between A and B is then computed for each conformation of B, and the optimization described above is applied to the most similar alignments.

Alternatively, these flexible dockings can be constrained by superimposing the shared moieties of A and B, and keeping them fixed. This is useful for selecting a diversity set of sidechains that will all join to the same fixed location on a scaffold. In this case, the docking time is reduced by an order of magnitude. Note that our similarity search uses the similarity score as an objective function and uses the conformational energy as a soft constraint, by disallowing high-energy conformations. A combined approach could use a weighted sum of the similarity and conformational energy. We have not explored this possibility.

Our system has mostly been used for reactant selection, where common moieties determine how the reactants would attach to common scaffolds. We have not used the system for database mining or other large similarity searches. We also note that molecules of substantially different sizes will usually score as being fairly dissimilar, as the small one will pull most of the features in and the large one will push most of the features out. In some cases, partial similarity can be achieved when part of each molecule's surface aligns exactly, at the expense of a complete mismatch along the remaining surface.

A dissimilarity calculation typically takes about 40 seconds on a single DEC Alpha. For efficiency, all dissimilarities calculated are stored in the GNU GDBM database [15], so that no dissimilarity is ever calculated twice. Currently, approximately 1/2 million dissimilarity results are stored in our GNU GDBM database. This represents almost 1 CPU year of computation.

2.2 Diversity a Set of Molecules

Having presented the *IcePick* method for computing the dissimilarity between two molecules, we now show how to use this method to build a system that selects diverse sets of molecules. This *set diversity* algorithm satisfies our overall goal of designing diverse combinatorial libraries.

IcePick computes the diversity of an entire set of molecules using a structure called a "minimum weight spanning tree" [16]. Loosely speaking, a minimum weight spanning tree, or MWST, is the shortest way to indirectly interconnect a set of points. We will use the size of the MWST as a measure of diversity, because when the MWST is large, the points are very "spread out".

In the molecular diversity setting, we define the minimum weight spanning tree as follows. For a collection of n molecules: $M_1, M_2, \cdots, M_n$, a *spanning tree* is any collection of $n-1$ pairs of molecules that connects all of the molecules (indirectly) with each other. For example, with $n = 4$, the 3 edges $(M_1, M_2), (M_1, M_3), (M_1, M_4)$ form a spanning tree. In this example, we say M_2 is connected to M_4 because there is an edge from M_2

to M_1 and one from M_1 to M_4. The 3 pairs $(M_1, M_2), (M_2, M_3), (M_3, M_4)$ form another spanning tree. The pairwise dissimilarity $\mathrm{d}(M_1, M_2)$ is assigned as the *weight* of edge (M_1, M_2). Then the weight of a particular spanning tree is the sum of its $n - 1$ edge weights. A *minimum weight spanning tree* is a spanning tree with the least possible weight. The *diversity* of a set of molecules is the weight of the minimum weight spanning tree over that set. The minimum weight spanning tree calculation can be done efficiently using Kruskal's algorithm[16, 17].

A few technical notes about minimum weight spanning trees: While easy to compute, the weight of the minimum weight spanning tree over a set of points is not monotonic. Its measure of diversity can increase *or* decrease when items are added. A structure called a Steiner tree is similar to a minimum weight spanning tree and the weight of the minimum Steiner tree is monotonic. However, computing the Steiner tree based score of a set of points is considered computationally intractable [18], and finding the minimum would be even more difficult. Because of the difficulty of computing Steiner trees, we use the spanning tree measure, despite its lack of monotonicity. Spanning tree methods have been used previously for chemical clustering [19, 20]; recently, their use as diversity set scores has been rediscovered [21].

So far, we have described a system that computes a diversity score for a set of molecules. The goal is to find a small set with a high diversity score. *IcePick* does this with a simple "swap-one" optimization heuristic. The algorithm starts with a random subset of k molecules, and then attempts to improve the set's diversity score by swapping in new molecules. All of the molecules under consideration are placed in a list in arbitrary order and each molecule A in the list is compared against every molecule B in the set to see if swapping A for B improves the set's score. Molecule A then replaces the molecule B that yields the largest increase (if there is such a B). This replacement, including the selection of the most favorable molecule B to remove from the current set, can be performed in

the time to compute a minimum weight spanning tree on the current set plus the potential new molecule. After the substition, the scan continues for new replacement molecules A, without returning to the

beginning of the list. The complete scan of the list is repeated until no further improvements are possible. The resulting set is "locally optimal", because no single molecule swap can improve the score.

A nice consequence of this method is that it helps minimize the number of dissimilarities computed. Suppose one wanted to select k molecules from a set of n molecules. There are $n(n-1)/2$ dissimilarities implied by the set of n molecules. However, local optimality of the set can be tested by looking at approximately kn edges. Typically the number of scans has been about 5, which requires the computation of only $5kn$ edges (an advantage when $k < n/10$). For a typical application such as "pick 22 amines out of a set of 1,500 amines", the speedup is quite significant: if *IcePick* converges in 5 scans, then no more than 165,000 dissimilarities are used by *IcePick* even though the total set of 1,500 molecules determines 1,124,250 dissimilarities, so fewer than 1/6th of the possible dissimilarity computations were needed.

Figure 1: 25 points selected from the unit square

Figure 1 shows the selection of 25 points from 1000 points distributed uniformly in the unit square using the spanning tree set-measure and our swap-one optimization method. This simple example has previously been used as a diversity "sanity-test" [22]. Our swap-one algorithm was compared to a standard greedy algorithm in this simplified setting. The greedy method builds a MWST one point at a time by repeatedly adding the point that maximally increases the weight of the MWST. The swap-one optimization method takes significantly longer than the greedy method, but produces a selection with a 9% lower (better) score. In this setting, commonly used variance methods, discussed in Section 3.7, select points far from the set's centroid, and hence would pick clusters of points from the corners of the square. However, in a more realistic, high-dimensional setting, a larger portion of the space is “near the periphery”, so the difference in the variance algorithm’s behavior would be less pronounced.

The swap-one heuristic finds good diversity sets, but it is not guaranteed to always find a globally optimal solution. Finding an optimal solution under the spanning-tree measure is considered computationally intractable, since it can encode independent set detection [18], and there are no known algorithms for this class of problems that simultaneously guarantee accuracy and speed. Other optimization techniques, such as the greedy method [17, 23] or simulated annealing [22, 24], could be used, but initial experiments have not indicated any substantial advantage in using such techniques.

3 Results and Discussion

We now present the results of running *IcePick* on a variety of test cases and discuss other approaches and related issues. First, *IcePick*’s 3D surface-based pairwise dissimilarity measure is compared with 2D topological measures. Then the minimum weight spanning tree method for set diversity is tested and compared to variance-based approaches. We also discuss an extension that allows rapid inference of *IcePick* dissimilarities without explicitly computing them.

3.1 Substitution Search

Our first example compares *IcePick*’s pairwise dissimilarity measure against a topological approach in the simple problem of searching for a substitute molecule that is similar

Figure 2: Substitutions with topological and IcePick similarity measures.

to a given molecule. Whereas *IcePick* dissimilarities involve the 3D flexible shapes of the molecules, topological approaches such as that of Daylight's Merlin tool [25] do not consider shape or conformational analysis. Molecules are summarized by determining all local neighborhoods (also called fragments, 2D, or topological features) of each atom. The similarity of two molecules is determined by the number of shared local neighborhoods. This count is then "Tanimoto" normalized by dividing by the total number of different local neighborhood types found in the two molecules. This topological method does not abstract away chemical information, because it is intended to solve the molecular similarity problem for many applications, including predicting chemical reactions.

Consider the three molecules depicted in Figure 2. Suppose a set of molecules included the chemically undesirable molecule 2-(tri-methyl silyloxy)benzaldehyde. An ideal replacement would have a similar structure without the offending silicon. Merlin and *IcePick* were given a set of 10,000 molecules from the Available Chemicals Directory [26] from which to choose a substitute. The chemical similarity engine found in Merlin suggests 4-(tri-methyl silyloxy)benzaldehyde as a structurally near replacement. *IcePick* suggests the strong *structural* analog 2-(tert-butylthio)benzaldehyde shown in Figure 2. Incidentally, this substitute has an almost identical molecular weight—even though this is *not* a feature considered by *IcePick*.

Under Merlin's topological measure, the 2-(tri-methyl silyloxy)benzaldehyde to 4-(tri-methyl silyloxy)benzaldehyde substitution is conservative. As much as possible of the molecule was retained (up to the offending silicon), and the rest removed without any replacement. *IcePick*'s surface-based measure, on the other hand, replaces 2-(tri-methyl silyloxy)benzaldehyde with 2-(tert-butylthio)benzaldehyde. It is chosen not because the chemical diagrams look similar, but because for *every* conformation of 2-(tri-methyl silyloxy)benzaldehyde, there is a nearly identical conformation of 2-(tert-butylthio)benzaldehyde, and vice versa. Of course, different topological systems with different descriptors such as [27, 28] might suggest the same substitution that *IcePick* does in this simple example.

3.2 Comparison with Topological Dissimilarity

We now compare *IcePick* dissimilarity with topological dissimilarity for 1,000 pairs of molecules, and show that the two measures are not well correlated. This provides quantitative evidence that they are computing truly different properties (though it does not make a qualitative statement about which is "better").

Figure 3 plots *IcePick* versus 2D (topological) dissimilarity and shows that they are not strongly correlated. Approximately 2,000 random primary amines were selected from the Available Chemicals Directory [26]. 1,000 pairs of these molecules were selected. For each pair, both the topological distance and *IcePick* dissimilarity were computed and plotted on the graph in Figure 3. In this graph, each point is one of the pairs of molecules, the x-coordinate is the square root of the number of local neighborhoods the pair differs in (or the Euclidean distance), and the y-coordinate is the *IcePick* dissimilarity. The linear correlation coefficient (Pearson r) is below $.4$, indicating that *IcePick* computes something different than the topological system. The Tanimoto normalization (which would alter the x-coordinates so all the points are in the interval $[0, 1]$) has been left out as it non-uniformly compressed the x-range and made the trend even worse. Hamming distance (or squared Euclidean distance) also worsens the correlation.

3.3 Diverse Subset Selection: Amino Acids

Another simple example examines the classification of the 20 natural amino acids. The dissimilarity measure can be used to suggest substitutions. For instance, it selects

Figure 3: Topological Features versus *IcePick* dissimilarity

arginine as the nearest analogue to lysine, which agrees with common observations [29]. Furthermore, *IcePick*'s pick of a diversity set of size 5 from the 20 natural amino acids is: arginine, aspartic acid, glutamine, tyrosine, proline. This selection hits all 4 classes of the typical amino acid classifications: basic, acidic, uncharged polar (twice) and non-polar [30]. The *IcePick* selection of size 4 does not include representatives of all 4 classes, because *IcePick* considers shape and steric factors in addition to polar factors.

3.4 Minimum Weight Spanning Trees

The amino acid test case above is reasonably simple because there are only twenty amino acids, that have built-in diversity. For larger, more practical problems, the data is often

Figure 4: Minimum weight spanning tree for four ideal molecules

unfavorably and non-uniformly clustered. This makes diversity selection more difficult, but is also the setting in which diversity selection is desirable. The next example shows how the minimum weight spanning tree method automatically handles non-uniform data.

Consider Figure 4, in which the four disks represent four molecules, and the dissimilarities are indicated by distance (i.e. disks drawn near each other are similar and disks drawn far apart are dissimilar). The lines drawn form a minimum weight spanning tree. The *IcePick* diversity score is the sum of the lengths of the three drawn edges. *IcePick* automatically recognizes (without using a clustering algorithm to pre-process the data) that almost all of the diversity of this set is due to the one long edge.

A variance based system (Section 3.7) would add the lengths of all 6 possible edges

Figure 5: Two clusters drawn at 5 different distances

in the diagram (including 4 long edges). In such a variance based system, one could add significant diversity to the set by adding a near duplicate of any of the 4 molecules, whereas under the spanning tree method a near duplicate molecule never adds a significant amount of diversity.

A more complicated example is found in Figure 5. In this diagram, all the squares and triangles represent idealized molecules with dissimilarities again indicated by distance. Each of the five rows in this figure depicts the minimum weight spanning tree drawn between a set of points in two clusters. Only the inter-cluster distance varies between rows. *IcePick* initially scores the diversity the whole set as the diversity within the squares plus the diversity within the triangles plus the dissimilarity between the nearest square and triangle.

As the distance between the two clusters is decreased the diversity score decreases similarly, until the two clusters join. At this point, the spanning tree drawn through the squares starts taking shortcuts through the spanning tree drawn through the triangles. *IcePick* determines that the diversity score is significantly below the sum of the diversities of the two original clusters.

Note that only the spanning tree method has been used in these examples. Any clustering occurs implicitly, without the use of a separate clustering algorithm or pre-processing.

3.5 Practical Experience

The *IcePick* approach is robust and fast enough for use in a practical setting. *IcePick* has been in routine use at Axys since November of 1996. Since then, it has routinely solved problems such as picking 10 to 40 diverse sidechains from a set of 200 to 3,000 possibilities, with runtimes of 4 to 5 days. These selections have been used as the reactants in a number of combinatorial libraries that now total over 50,000 diverse compounds.

3.6 Inferred Dissimilarities

An important property of the *IcePick* dissimilarity measure is that it is “stable.” This means that if many dissimilarities are known, new dissimilarities can be accurately estimated without explicitly computing them. This greatly speeds up the diversity algorithm. If

Figure 6: Inferred dissimilarity $\gamma(A, B)$ versus computed dissimilarity $\mathrm{d}(A, B)$

one has computed the *IcePick* distances from x to a pre-selected set of sentinel molecules $M_1, M_2, \cdots, M_k$, and the distances from y to the same set $M_1, M_2, \cdots, M_k$, then one can predict $\mathrm{d}(x, y)$, without re-running the *IcePick* algorithm or using any additional knowledge about the molecules x and y.

The idea is similar to “affinity fingerprinting” [31], [32]. The concept is: if one knew, for a given molecule, its assay value for many assays, then one would (in a biological sense) know everything there was to be known about the molecule. For example, one could make a crude prediction of the molecule’s behavior in a new assay using assays thought to be most similar to the new assay. Thus, a molecule is itself represented by its list of assay results.

A similar effect is known for the *IcePick* molecular dissimilarity score. Let $\{M_1, M_2, \cdots, M_t\}$ be a selection of t molecules chosen to be diverse (either by *IcePick* or by hand). For two arbitrary molecules A and B, let $\mathrm{d}(A, B)$ be the *IcePick* dissimilarity of A and B. Then we have found the following distance geometry [33] based method of encoding molecules as vectors in $(t-1)$-dimensional space $\mathbb{R}^{t-1}$ to be useful.

First t vectors $x_1, x_2, \cdots, x_t \in \mathbb{R}^{t-1}$ are chosen such that

$$||x_i - x_j|| \approx \mathrm{d}(M_i, M_j) \text{ for all } i, j \ . \tag{1}$$

This can be done by a minimization algorithm, or using a matrix method such as Cholesky decomposition [34]. Then the molecule A is encoded by finding a point $x_A \in \mathbb{R}^{t-1}$ that minimizes the expression

$$\sum_{i=1}^{t} \left(||x_i - x_A||^2 - \mathrm{d}(M_i, A)^2 \right)^2 \ . \tag{2}$$

The new approximate dissimilarity function is

$$\gamma(A, B) = ||x_A - x_B|| \ . \tag{3}$$

For 1,000 random pairs of molecules (Figure 6), a graph of $\gamma(A, B)$ (using $t = 50$ "basis molecules") versus $\mathrm{d}(A, B)$, shows a Pearson linear correlation coefficient of about $.85$. This is why we refer to *IcePick* dissimilarities as being stable.

This allows us to choose k molecules out of n looking at only tn edges (when a basis of size t is used). This can again yield a very substantial savings. For instance, with $n = 1,500, k = 22, t = 32$, only 48,000 dissimilarity calculations are needed, or less than

1/23rd of the total possible computations needed. The basis set size t represents a "speed vs. accuracy" control that should not be set too low.

The ability to find a set of points in $\mathbb{R}^{t-1}$ that well represents *IcePick* dissimilarity data as pairwise distances naturally leads one to ask if there is a minimal dimension, d, such that such a representation exists. Also, one would like to know if this dimension has a physical interpretation. The Johnson–Lindenstrauss theorem [35] states that it is not possible to determine the dimension d without a truly enormous amount of very accurate data. The Johnson–Lindenstrauss theorem implies that if there is a good representation of the dissimilarity data between n molecules as distances between n points in $\mathbb{R}^d$ (for any d), then there is a good approximate representation of the dissimilarity data in $\mathbb{R}^{c \log_2 n}$, independent of d (where c is a small constant, not given here, independent of n and d). So even if d is the correct minimal dimension for the problem, one can find good representations with dimension lower than d until there are more than $2^{d/c}$ molecules. This effectively masks the natural dimension d unless there is an enormous amount of highly accurate data.

3.7 Variance Measures

In this paper we refer to some common diversity measures as "variance measures." We call a method a "variance measure" if its overall purpose is to measure gross spread and its calculation is analogous to the computation of a variance. A common example of such a measure is defining the diversity of a set to be the sum of all the squared distances between pairs of molecules [36].

Even though the superficial form of the formulas used in these measures seems to imply that they are a function of all of the pair dissimilarities between molecules, we show that these measures depend only on each molecule's distance from the center of the set. This a variation of the fact that the expectation $E(x - Ex)^2 = Ex^2 - (Ex)^2$, which is well known in statistics [37].

These cancellations are known to provide an opportunity to significantly speed up the computation of variance-based diversity measures [38]. However, these cancellations also show that one can increase the variance measure of a set by adding duplicate or near duplicate molecules to the set. This is accomplished by adding the redundant molecules in such a way that they do not significantly move the center. Then all of the original molecules are still scored as before, and the new molecules contribute additional score. Since duplicate molecules add no real utility, this behavior is a weakness of variance type measures.

In contrast, the spanning tree method never increases its score when duplicate molecules are added. This is because the spanning tree algorithm ensures that a molecule's contribution to the diversity measure depends most on the molecules nearest it (and not on some abstract center).

The following example illustrates this problem of variance measures. Given a set of molecules S encoded as n vectors in $\mathbb{R}^d$, $\{M_1, M_2, \cdots, M_n\}$, a suggested total diversity of the set could be:

$$\mathrm{diversity}(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} \tau(M_i, M_j) \tag{4}$$

where is $\tau(M_i, M_j)$ is a distance function.

For instance, for squared-Euclidean distance:

$$\tau(M_i, M_j) = \sum_{k=1}^{d} ((M_i)_k - (M_j)_k)^2 \tag{5}$$

one could rearrange and speed up the calculation by the identity

$$\begin{aligned} & \sum_{i=1}^{n} \sum_{j=1}^{n} \tau(M_i, M_j) && (6) \\ = \; & \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{d} ((M_i)_k - (M_j)_k)^2 && (7) \\ = \; & 2n \sum_{i=1}^{n} \sum_{k=1}^{d} \left((M_i)_k - \frac{1}{n} \sum_{j=1}^{n} (M_j)_k \right)^2 && (8) \\ = \; & 2n \sum_{i=1}^{n} \tau \left(M_i, \frac{1}{n} \sum_{j=1}^{n} M_j \right), && (9) \end{aligned}$$

which reveals that such a diversity measure is only a function of the samples' distance from the center $\frac{1}{n} \sum_{j=1}^{n} M_j$, and not really a function of all the intermolecular dissimilarities. A similar effect has been shown for the "cosine coefficient" [39], though that work does not draw the same conclusion as given here.

The above observation on variance measures means that there are data sets where a variance system will unnecessarily pick duplicate points. This problem is serious, but not unique to variance measures. An "Optimal-D" design based system [40] will choose duplicate molecules if the duplicates are available in multiplicities that are near the D-design weights. The spanning tree system presented here will itself pick duplicate points if this is the only way to avoid picking a point from a Steiner tree, however (unlike the systems mentioned above), it does not increase its score when this happens.

4 Conclusions

Flexible surface feature models that depend on performing multiple molecular dockings are fast enough for practical use. We have presented a flexible surface based system for molecular diversity, designed to choose reactants for combinatorial chemistry. The system uses proven methods from structural drug design to estimate how well one molecule can imitate the conformations of another. These conformations are encoded into approximations of binding modes and include surface accessible steric and polar features. Flexibility of molecules is handled by a flexible docking procedure and averaging over multiple conformations. The weak correlation shown between 2D (or topological) indices and presented surface features indicate that the two notions encode fundamentally different information. The flexible surface feature dissimilarity system is stable for self-prediction. It seems likely that the dissimilarity measurements might correlate with other surface mediated biological properties, and so may be useful in their prediction [41].

The spanning tree system for assigning diversity scores to sets using only dissimilarity data is both novel and useful. The method provides an efficient automated method for evaluating the diversity of sets of molecules from dissimilarity data (without direct reference to any underlying features). It works well on known examples.

The *IcePick* system has selected reactants for combinatorial libraries from a database of 10,000 compounds and assisted in the design of a suite of libraries resulting in the production of over 50,000 diverse compounds at Axys.

5 Acknowledgments

We thank Guy Breitenbucher, Nathan Collins, Chuck Johnson, Doug Livingston, Bob Mcdowell and Chris Phelan of Axys for their support and discussions.

References

[1] Plunkett, M.; Ellman, J. Combinatorial chemistry and new drugs. *Scientific American* **1997**, 276, 68–73.

[2] Ellman, J.; Stoddard, B.; Wells, J. Combinatorial thinking in chemistry and biology. *Proc. Natl. Acad. Sci. USA* **1997**, 94, 2779–2782.

[3] Williard, X.; Pop, I.; Horvath, L.; Baudelle, R.; Melnyk, P.; Deprez, B.; Tartar, A. Combinatorial chemistry: a rational approach to chemical diversity. *Eur. J. Med. Chem.* **1996**, 31, 87–98.

[4] Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Richter, L. S.; Moos, W. H. Discovery of

nanomolar ligands for 7-transmembrane G-protein-coupled receptors from a diverse N-(substituted)glycine peptoid library. *J. Med. Chem.* **1994**, 37, 2678–2685.

[5] Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow Jr., R. E.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: A shape-based machine learning tool for drug design. *Journal of Computer-Aided Molecular Design* **1994**, 8, 635–652.

[6] Welch, W. ; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry & Biology* **1996**, 3, 449–462.

[7] Jain, A. N.; Harris, N. L.; Park,J. Y. Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the $5 - HT_{1A}$ receptor. *Journal of Medicinal Chemistry* **1995**, 38, 1295–1308.

[8] Welch, W.; Ruppert, J.; Klein, T.; Jain, A.; Sage, C.; Stroud, R.; Stout, T. Discovery of novel inhibitors of thymidylate synthase using flexible docking. *Submitted for publication* **1998**.

[9] 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.

[10] Gibson, S.; McGuire, R.; Rees, D. C. Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments. *J. Med. Chem.* **1996**, 39, 4065–4072.

[11] 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.

[12] Shemetulskis, N. E.;Dunbar Jr., J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design* **1995**, 9, 407–416.

[13] Mayo, S.L.; Olafson, B.D.; Goddard, W.A. DREIDING: A generic Force Field for Molecular Simulations. *J. Phys. Chem* **1990** 94, pp 88–97.

[14] Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *Journal of Computer-Aided Molecular Design* **1996** 10, 427–440.

[15] Nelson, P. *GDBM* Free Software Foundation 59 Temple Place - Suite 330 Boston, MA 02111-1307.

[16] Cook, W. J.; Cunningham, W. H.; Pulleyblank, W. R.; Schrijver, A. *Combinatorial Optimization*. **1998**.

[17] Kruskal, J.B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. American Math Soc.* **1956**, 7, 48–50.

[18] Garey, M.; Johnson, D. *Computers and Intractability, A Guide to the Theory of NP-Completeness*. **1979**.

[19] Miyashita, Y.; Takahashi, Y.; Yotsui, Y.; Abe, H.; Sasaki, S. Clustering Molecules on the Basis of Antibacterial Spectra or Physiochemical Properties. *Anal. Chem. Acta.* **1981**, 133, 614–624.

[20] Ritter, G.; Isenhour, T. Minimal spanning tree clustering of gas chromatographic liquid phases, *Computers and Chemistry* **1977**, 1, 145–153.

[21] Waldman, M.; *Personal communication* **1998**.

[22] Agrafiotis, D. Stochastic Algorithms for Maximizing Molecular Diversity, *J. Chem. Inf. Comput. Sci.* **1997**, 37, 841–851.

[23] Rado, R. A note on independence functions. *Proceedings of the London Mathematical Society* **1957**, 7, 300–320.

[24] Metropolis, N.; Rosenbluth A. W.; Rosenbluth, M. N.; Teller, M. N.; Teller, E. Equation of state calculations by fast computing machines, *J. Chem. Phys.* **1953**, 21, 1087–1092.

[25] James, C. A.; Weininger, D.; Delany, J. Daylight theory manual. *Daylight Chemical Information Systems Inc.* **1995**, Irvine, CA.

[26] *Available chemicals directory* MDL Information Systems Inc., 14600 Catalina Street, San Leandro Ca, 94577.

[27] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definitions and applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64–73.

[28] Nilakantan, R; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *CJACS* **1987**, 27, 82–5.

[29] Bordo, D.; Argos, P. Suggestions for safe residue substitutions in site-directed mutagenesis. *J. Mol. Biol.* **1991**, 217, 721–729.

[30] Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J.D. *Molecular Biology of the Cell* **1983**.

[31] Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. & Biol.* **1995**, 2, 107–118.

[32] Briem, H.; Kuntz, I. D. Molecular similarity based on dock-generated fingerprints. *Journal of Medicinal Chemistry* **1996**, 39, 3401–3408.

[33] Crippen, G.; Havel, T. *Distance Geometry and Molecular Conformation* **1988**.

[34] Press, W. ; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes in C.* **1991**.

[35] Linial, N.; London, E.; Rabinovich, Y. The geometry of graphs and some of its algorithmic applications. *35th Annual Symposium on Foundations of Computer Science, IEEE* **1995**, 577–591.

[36] Adding distances instead of squared distances is similar- but squared distances form a simpler example.

[37] McClave, J. T.; Dietrich II, F. H. *Statistics* **1991**, 5th edition, 50–52.

[38] Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases. *Quant. Struct.-Act. Relat.* bf 1995, 14, 501–506.

[39] Turner, D., Tyrrel, S.; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. of Chemical Information and Computer Sciences*, **1997**, 37, 18–22.

[40] Atkinson, A.C.; Donev, A.N. *Optimum Experimental Designs* **1992**.

[41] Palm, K.; Luthman, K.; Ungell, A.L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *Journal of Pharmaceutical Sciences* **1996**, 85, 32–39.