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)

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

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

May 3, 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 measures the dissimilarity between two molecules based on their 3D surface accessible features, taking into account conformational flexibility. Then, the intrinsic diversity of an entire set of molecules is calculated from a *spanning tree* over the dissimilarities. We compare our dissimilarity measure against traditional 2D topological approaches, and compare the spanning tree diversity measure against commonly-used variance techniques. The method has proven easy to implement and 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]. The selection of reactants to be used (from the large universe of possible materials) is an important problem in combinatorial chemistry. Our focus in this paper is the selection of a maximally diverse set of reactants. We break this problem into two pieces: a pairwise dissimilarity (or molecular distance) measure and a novel concept of the diversity of a set of molecules.

Our computational approach is based on 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 as used in Hammerhead [6]. Both of these techniques have been proven effective in biological test systems [5, 7, 6, 8].

1.1 The Diversity Problem

Given a large collection of molecules meeting any number of pre-specified conditions such as cost, availability, ease of synthesis or predicted activity, the diversity problem is to select a smaller collection of molecules that best represents the larger set while avoiding redundancy or duplication.

Our purpose in screening is to find novel ligands that non-covalently bind to a protein. We assume that any lead can be expanded around (e.g. with directed libraries [9]), so two similar leads have little more value than one lead in early stage screening. So, we want to

maximize the probability of finding at least one lead. It is intuitive that one should do this by picking the molecules to be as dissimilar to each other as possible. This intuition can be justified. If one can not increase the probability of any individual molecule being a lead one can increase the probability an ensemble contains a lead by decreasing the probability that the ensemble contains multiple leads (as long as odds of any individual molecule in the ensemble being a lead are not decreased).

1.2 Pairwise Dissimilarity

Molecular docking [6, 10, 11, 12, 13] and surface based binding site models [14, 7, 15, 16, 17] have proven to be good predictors of biological activity. We adapt some of these ideas to yield our notion of dissimilarity. This measure is based on the explicit comparison of surface accessible steric and polar features (derived from simultaneous conformational analysis of pairs of molecules). This measure is compared to a common fragment/topology based fingerprint system. We demonstrate an implementation with sufficient throughput and flexibility to support a large combinatorial chemistry laboratory. We further show that these scores are "stable:" knowing many of these scores allows one to infer additional scores (without re-running the scoring function). It is also likely that these scores directly correlate to other surface mediated biological properties and so may be useful in their prediction [15].

Our dissimilarity measure (called the *IcePick* measure and denoted $\mathrm{d}(a, b)$) directly models several interactions important to specific binding of ligands to proteins. This

measure is intended to have the property that if $\mathrm{d}(a, b)$ is large then it is less likely that a and b have similar binding behavior (with respect to an unknown receptor) and if $\mathrm{d}(a, b)$ is small then it is likely that a and b have similar binding behavior. The measure is used in place of an arbitrary set of pharmacophores. This differs from other approaches [18, 19, 24, 25] in that we are not trying to maximize the number of different values of different features (logP, mass, number of rings) or cluster data, but to directly design a selection of molecules that are simultaneously novel with respect to each other. It is hoped that this method stays closer to the problem of maximizing the chance of finding a hit.

This notion of pairwise dissimilarity follows *Compass* [5, 7]. *Compass* encodes molecules based on their surface features (hydrophobic surface area, hydrogen bond donors, hydrogen bond acceptors). These representations are then used as targets for a flexible docking program. We compare one molecule to another by "turning it inside out" to form a pocket that perfectly fits around it (i.e. an ideal protein) and then flexibly docking the second molecule into the pocket. The docking is scored by comparing the protein accessible surface and the vector directions available for both hydrogen bond donation and acceptance. This process is averaged over a representative set of low energy conformations of the first molecule to get an overall score. This score measures how well the second molecule, as a flexible entity, can imitate typical conformations of the first.

An important property of our dissimilarity measure is that it is "stable." Even though computing *IcePick* dissimilarity involves multiple flexible dockings we find the score is well behaved in that if one knows both the distances from x to a pre-selected set of

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 any additional knowledge about the molecules x and y). This allows one to significantly reduce the number of dockings performed and represents a dramatic speedup of the algorithm.

1.3 Set Diversity

We compute the diversity of an entire set of molecules using a structure called a "spanning tree" [20]. Spanning tree methods have been used previously for chemical clustering [21, 22]; in this application they are used to estimate the total amount of novelty in binding modes represented by a set of molecules. We propose using the weight of the minimum weight spanning tree as the dissimilarity score of a set of molecules. Spanning trees, unlike traditional variance or additive measures, are well suited to eliminate near duplicate selections. The spanning tree method estimates the intrinsic diversity of a set; it is not limited to comparing relative diversities of sets.

Given all of the pairwise dissimilarities for a set of molecules, the diversity of the set is defined to be the sum of the edge weights of the minimum weight spanning tree drawn on the set. For a collection of n molecules: $M_1, M_2, \cdots, M_n$ a spanning tree is a collection of $n - 1$ pairs of molecules that connects all of the molecules (indirectly) with each other. For example, with $n = 4$ the 3 pairs $(M_1, M_2), (M_1, M_3), (M_1, M_4)$ form a spanning tree (e.g. we say M_2 is connected to M_4 because we could move from M_2 to M_1 and then 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 weight of a particular spanning tree is the sum of the $n-1$ listed pair-dissimilarities (or edges) in the tree (i.e. the pair (M_1, M_2) denotes the dissimilarity $\mathrm{d}(M_1, M_2)$). A minimum weight spanning tree is just a spanning tree that has the least possible weight. The spanning tree method has the desirable properties that it is good at eliminating duplicate behavior and it counts each instance of novelty only once. We feel this compares favorably to variance type methods (see Section 3.7) which can over-count novelty, and covering methods that require prodigious amounts of additional information (such as a list of all relevant pharmacophores).

We point out that a structure called a Steiner Tree is similar to a minimum weight spanning tree and has the additional important property that it is monotone (its measure of diversity never goes down when items are added). Because of the difficulty of computing Steiner Trees we use the spanning tree measure, despite its lack of monotonicity.

2 Methods

2.1 Pairwise Dissimilarity

The first component of our system is the pairwise dissimilarity measure. Given multiple low energy conformations of molecules we flexibly dock them into each other's *Compass* representation to determine how dissimilar they are. The conformations used are from a proprietary Axys program *Twitch* that performs a sparse conformational analysis under a Dreiding force-field [26], though any representative set of conformations would be

adequate.

The form of this component is:

Input: Two sets of low energy conformations of two molecules A and B.

Output: The dissimilarity measure of the two molecules, in the range 0-1, where 0 means identical.

To compute how well molecule B imitates molecule A we sample A's conformations and compute how well B flexibly imitates each one. The average of all these matchings is how well molecule B is able to imitate an average conformation of molecule A. The similarity of molecule A and molecule B is B's ability to imitate A on average plus A's ability to imitate B on average. The pairwise dissimilarity is 1.0 minus the similarity.

We note that our measure is not, in general, a metric (but this is a technical consideration).

The flexible dockings required for the dissimilarity score can be performed either in free space (with a number of starting orientations per conformation) or with the common moieties of each reactant held in an enforced correspondence. The per conformation dissimilarity score is described in more detail in [14, 7, 6]. What is being calculated (for each given orientation and conformation) is the expected difference in distances from many "feature probes" in space to ligand A and their corresponding distances to ligand B (aligned onto A). The feature probes are arranged in spherical shells about the ligands. In addition

to this steric term, two polar terms are computed for each probe: "distance and direction to nearest polar feature" for both hydrogen bond donors and hydrogen bond acceptors. These three terms encode the difference in surface presentations of two molecules.

This measure is quick to compute and its gradient exists almost everywhere. The gradient allows the use of standard smooth global optimization techniques to find simultaneous orientation and conformation parameters that minimize the difference between the two molecules. (During this flexible fitting, low-energy conformations and correct chiralities are maintained.) We use the Broyden-Fletcher-Goldfarb-Shanno gradient optimization method [27]. Most of the time required to evaluate dissimilarity is spent performing these dockings.

While in our application each conformation is a different low energy conformation of a given molecule—we note that the algorithm can automatically handle mixtures (e.g. racemic mixtures, tautomers or different molecules) when scoring. To score how well one mixture imitates another, one supplies a set of molecules and conformations that is thought to well represent the mixture as the set A. To score optically pure compounds one uses only conformations with the desired chirality.

When using a small number of these conformations (typically 10 to 100) 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 [28] (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 on a single DEC Alpha.

2.2 Diversity Definition

The second component of our system is the spanning tree method for scoring the diversity of a collection of molecules. The method uses our pairwise dissimilarity measure. Given molecules $M_1, M_2, \cdots, M_k$ and the k^2 dissimilarities ($d_{i,j}$ = dissimilarity of M_i and M_j) we define the "diversity of the set $M_1, M_2, \cdots, M_k$" to be the total edge weights of the minimum weight spanning tree drawn on the complete graph with vertices $1, 2, \cdots, k$ and edge weights $d_{i,j}$.

The spanning tree algorithm has the form:

Input: A set of k molecules and a function $\mathrm{d}(\cdot, \cdot)$ that computes the pairwise dissimilarity of molecules.

Output: The diversity score: the weight of the minimum weight spanning tree drawn on the graph with nodes $1, 2, \cdots, k$ and edge weights $d_{i,j}$.

This algorithm is used as the "score" subroutine for the optimization method given below. The spanning tree calculation itself is done efficiently using Kruskal's algorithm [20].

2.3 Optimization

Input: A set of n molecules, the size k of the desired diversity set, and a function called "$\mathrm{score}()$" that returns the diversity score of a set of molecules.

Output: A set of k molecules with a high diversity score.

What we have described up to now is a system that given a set returns a diversity score. It remains to find a set with a high score. We describe a simple "swap-one" optimization heuristic for this purpose.

We do not guarantee that the selection picks a set that maximizes the weight of the minimum weight spanning tree because the selection problem is computationally intractable (it can encode independent set detection [29] and there are no known algorithms for problems this expressive that simultaneously guarantee accuracy and speed).

Note that representations of the actual molecules being optimized are not needed by this algorithm—it works solely with the dissimilarity information in the input.

In our application we are using

$$\mathrm{score}(M_1, M_2, \cdots, M_k) = \text{weight of MWST}\ , \tag{1}$$

where "weight of MWST" denotes the total edge weight of the minimum weight spanning tree drawn through the complete graph with vertices $1, 2, \cdots, k$ and edge weights $d_{i,j}$ (weight = total weight of edges in tree). One could also make information other than the

spanning tree score available to the optimization algorithm.

The algorithm starts with a random subset of k molecules, and then uses a "swap-one" approach to improve the diversity score. It scans through the entire list of allowed molecules. If any one of these would increase the set-score, it replaces the least-favored molecule in the current set. The scanning process is repeated until no further improvements are found.

Other optimization techniques, such as the greedy method or simulated annealing, could be used, but initial experiments have not indicated any worthwhile advantage in using such techniques.

A nice consequence 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 passes 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 savings is quite significant: if *IcePick* converges in 5 passes 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 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.

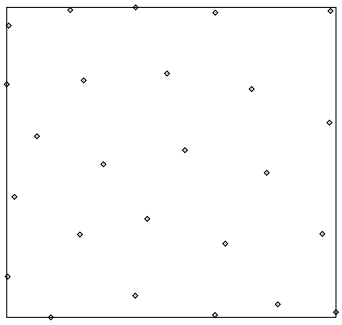

Figure 1: 25 points selected from the unit square

We do not suggest that molecular dissimilarity data is distributed as in this example, but note that Agrafiotis has used this as a kind of "sanity-test" [30]. 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 which maximimally 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.

Recently, Waldman, Li and Hassan have independently rediscovered our use of spanning trees as set-scores (instead of as clustering tools) [23].

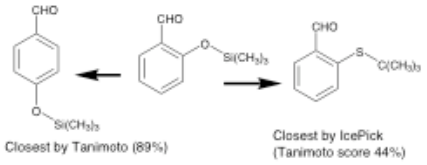

Figure 2: Substitutions with topological and IcePick similarity measures.

3 Results and Discussion

3.1 Substitution Search

Our first example is chosen to be simple. Consider the three molecules depicted in Figure 2. Suppose we had a set of molecules including the chemically undesirable molecule 2-(trimethyl silyloxy)benzaldehyde. An ideal replacement would have a similar structure without the offending silicon. Both Daylight's Merlin tool [31] and IcePick were given the same subset of molecules from the Available Chemicals Directory [32] from which to choose a substitute. (The subset consisted of about 40,000 molecules that had been loaded into the IcePick system over the previous year.) The chemical similarity engine found in Merlin suggests salicylaldehyde 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*.

Daylight's topological similarity system (like many others) is designed as an all purpose piece of software. It is intended to solve the molecular similarity problem for

all applications, including predicting chemical reactions. Molecules are summarized by determining all local neighborhoods (also called fragments, 2D, or topological features) of each atom. Dissimilarity of two molecules is computed by determining what fraction of each other's local neighborhoods they agree on. This count is then "Tanimoto" normalized by dividing by the total number of different local neighborhood types found in the two molecules.

Under such a topological measure the 2-(tri-methyl silyloxy)benzaldehyde to salicylaldehyde 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 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, a topological system that considers silicon and carbon as being structurally similar might suggest the same substitution that IcePick does.

3.2 Comparison with Topological Dissimilarity

We wish to show that the previous example is not unique. 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 [32].

1,000 of the possible pairs of molecules that could be chosen from the set of 2,000 molecules were selected. For each of these pairs both the topological distance and *IcePick* dissimilarity were computed and the result is a point 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 compresses the x-range and makes the trend even worse. Hamming distance (or squared Euclidean distance) also worsens the correlation.

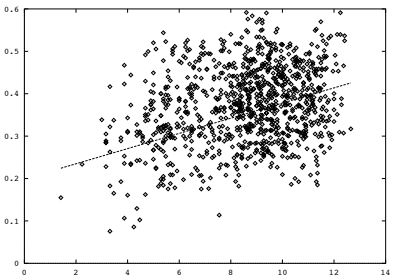

Figure 3: Topological Features versus *IcePick* dissimilarity

3.3 Classifying 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 arginine as the nearest analogue to lysine, which agrees with common observations [33]. 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 [34]. The *IcePick* selection of size 4 does not include representatives of all 4 classes (because *IcePick* considers steric factors in addition to polar moieties).

3.4 Minimum Spanning Trees

Another example illustrates *IcePick*'s spanning tree method. Consider Figure 4. The four disks represent four idealized molecules and the dissimilarities are as drawn in the figure (i.e. disks drawn near each other are similar and disks drawn far apart are dissimilar). The lines drawn are a minimum weight spanning tree. Under our system the diversity of the set would be the sum of the lengths of the three drawn edges. *IcePick* automatically recognizes (without need of a clustering algorithm to pre-process the data) that almost all of the diversity of this set would come from the one long edge.

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

Figure 4: Minimum spanning tree for four ideal molecules

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 given by how far apart the figures are drawn. Each of the five rows in this figure depicts the minimum weight spanning tree drawn between a set of points in two clusters. All that varies from row to row is the distance between the centers of the two clusters. Our system initially scores the diversity of the set as being the diversity in the set of squares plus the diversity in the set of triangles plus the dissimilarity of the nearest square to the nearest triangle.

As the distance between the two clusters is decreased the diversity score decreases similarly, until the two clusters touch (or in some sense are no longer two clusters). 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.

It is important to remember that the spanning tree method implies all of these calculations automatically. It does not require a separate clustering algorithm to identify clusters or depend on complicated software. All of the behavior exhibited here follows

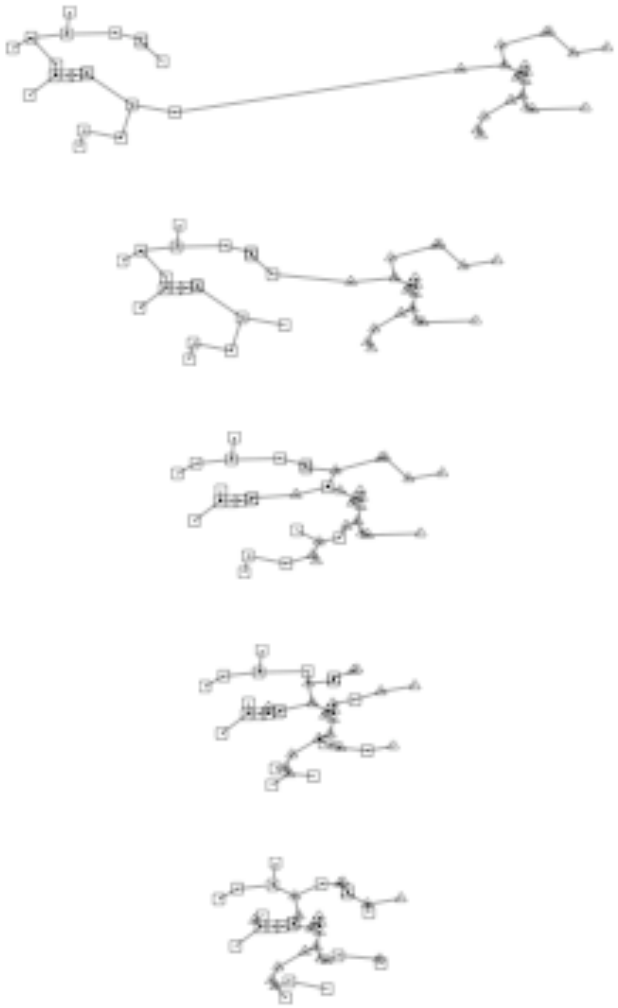

Figure 5: Two clusters drawn at 5 different distances

from the definition of a minimum weight spanning tree.

3.5 Practical Experience

A final experience, related to illustrate that the method is fast enough to be used in practice: *IcePick* has been in routine use at Axys since November of 1996. In this time it has routinely solved problems such as picking 10 to 40 diverse side-chains from a set of 200 to 3,000 possibilities in 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

Inferred dissimilarities are both a systematic speed-up for *IcePick* and an important direction for future applications. The idea is similar to "affinity fingerprinting" [35]. The concept is: if one knew, for a given molecule, its assay value for many assays, then one would (in a biological sense) know every thing 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.

Briem and Kuntz [36] suggest a similar approach using UCSF Dock. A molecule is encoded as its computed binding affinity to a large number of proteins. Further predictions about the molecule (such as will it dock into a novel protein) can then be performed by applying standard machine learning/classification algorithms on the vector of Dock-data.

A similar effect is known for our 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 A

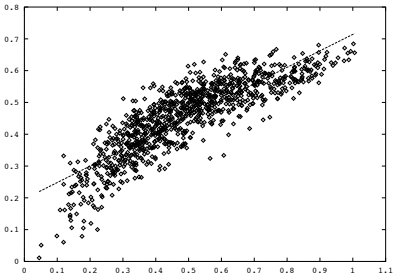

Figure 6: Inferred dissimilarity $\gamma(A, B)$ versus computed dissimilarity $\mathrm{d}(A, B)$

and B two arbitrary molecules let $\mathrm{d}(A, B)$ be the *IcePick* dissimilarity of A and B. Then we have found the following distance geometry [37] based method of encoding molecules as vectors in $\mathbb{R}^{t-1}$ to be useful.

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

$$||x_i - x_j|| \approx \mathrm{d}(M_i, M_j) \text{ for all } i, j \ . \tag{2}$$

This can be done by a minimization algorithm or using a matrix method such as Cholesky decomposition [27]. Then the molecule A is encoded by finding a point $x_A \in \mathbb{R}^{t-1}$ that minimizes the expression

$$\sum_{i=1}^{t} \left(||x_i - x_A||^2 - \mathrm{d}(M_i, A)^2 \right)^2 \ . \tag{3}$$

The new approximate dissimilarity function is

$$\gamma(A, B) = ||x_A - x_B|| \ . \tag{4}$$

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 be a very substantial savings (typical values: $n = 1,500, k = 22, t = 32$ so only 48,000 dissimilarity calculations are needed to score the entire set, which is less than 1/23rd of the total possible computations needed). It must be noted that t is basically a "speed vs. accuracy" control and the value should not be set too low.

The ability to find a set of points in $\mathbb{R}^{t-1}$ that well represents our 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 meaningful interpretation. The Johnson–Lindenstrauss theorem [38] 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 we can find good representations with dimension lower than d until we have enough molecules so that $n > 2^{d/c}$. This effectively masks d until we have an enormous amount of data to a very high accuracy.

3.7 Variance Measures

In this paper we refer to some 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 would be defining the diversity of a set to be the sum of all the squared distances of the pairs of molecules in the set [39].

Even though the superficial form of the formulas used in these measures seem to indicate 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 is due to cancellations well known in statistics.

These cancellations 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 useless molecules in such a way that they do not significantly move the center and thus 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).

To demonstrate the problem with variance we work an example.

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:

$$\text{diversity}(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} \tau(M_i, M_j) \tag{5}$$

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

For instance, for squared-Euclidean distance:

$$\tau(M_i, M_j) = \sum_{k=1}^{d} \left((M_i)_k - (M_j)_k\right)^2 \tag{6}$$

one could re-arrange and speed up the calculation by the identity:

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=1}^{n} \tau(M_i, M_j) && (7)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{d} \left((M_i)_k - (M_j)_k\right)^2 && (8)\\
&= 2n \sum_{i=1}^{n} \sum_{k=1}^{d} \left((M_i)_k - \frac{1}{n} \sum_{j=1}^{n} (M_j)_k\right)^2 && (9)\\
&= 2n \sum_{i=1}^{n} \tau\left(M_i, \frac{1}{n} \sum_{j=1}^{n} M_j\right), && (10)
\end{aligned}$$

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 is shown for the "cosine coefficient" in [40], though the authors do not draw the same conclusion as given here.

4 Conclusions

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.

Flexible surface feature models (models that depend on performing multiple ligand/hypothesis 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 form opinions as to what degree one molecule can imitate the various 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 in that it is able to predict itself. It is anticipated that these dissimilarity scores will be able to predict other biological activity (such as solubility or biological transport).

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 and Chris Phelan of Axys for their support and discussions.

References

[1] M. PLUNKETT AND J. ELLMAN, Combinatorial chemistry and new drugs, *Scientific American*, vol. 276, pp. 68–73, April 1997.

[2] J. ELLMAN, B. STODDARD, AND J. WELLS, Combinatorial thinking in chemistry and biology, *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 2779–2782, April 1997.

[3] X. WILLIARD, I. POP, L. HORVATH, R. BAUDELLE, P. MELNYK, B. DEPREZ, AND A. TARTAR, Combinatorial chemistry: a rational approach to chemical diversity, *Eur. J. Med. Chem.*, vol. 31, pp. 87–98, 1996.

[4] R. N. ZUCKERMANN, E. J. MARTIN, D. C. SPELLMEYER, G. B. STAUBER, K. R. SHOEMAKER, J. M. KERR, G. M. FIGLIOZZI, D. A. GOFF, M. A. SIANI, R. J. SIMON, S. C. BANVILLE, E. G. BROWN, L. WANG, L. S. RICHTER, AND W. H. MOOS, Discovery of nanomolar ligands for 7-transmembrane G-protein-coupled receptors from a diverse N-(substituted)glycine peptoid library, *J. Med. Chem.*, vol. 37, no. 17, pp. 2678–2685, 1994.

[5] A. N. JAIN, T. G. DIETTERICH, R. H. LATHROP, D. CHAPMAN, R. E. CRITCHLOW JR., B. E. BAUER, T. A. WEBSTER, AND T. LOZANO-PEREZ, Compass: A shape-based machine learning tool for drug design, *Journal of Computer-Aided Molecular Design*, vol. 8, pp. 635–652, 1994.

[6] W. WELCH, J. RUPPERT, AND A. N. JAIN, Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites, *Chemistry & Biology*, vol. 3, pp. 449–462, June 1996.

[7] A. N. JAIN, N. L. HARRIS, AND J. Y. PARK, Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the $5-HT_{1A}$ receptor, *Journal of Medicinal Chemistry*, vol. 38, no. 8, pp. 1295–1308, 1995.

[8] W. WELCH, J. RUPPERT, T. KLEIN, A. JAIN, C. SAGE, R. STROUD, AND T. STOUT, Discovery of novel inhibitors of thymidilate synthase using flexible docking, *Submitted for publication*, April 1998.

[9] E. K. KICK, D. C. ROE, G. SKILLMAN, G. LIU, T. J. EWING, Y. SUN, I. D. KUNTZ, AND J. A. ELLMAN, Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D, *Chemistry & Biology*, vol. 4, pp. 297–307, April 1997.

[10] I. KUNTZ, J. BLANEY, S. OATLEY, R. LANGRIDGE, AND T. FERRIN, A geometric approach to macromolecule-ligand interaction, *J. Mol. Biol.*, vol. 161, pp. 269–288, 1982.

[11] A. LEACH AND I. KUNTZ, Conformational analysis of flexible ligands in macromolecular receptor sites, *J. Comp. Chem.*, vol. 13, pp. 730–748, 1992.

[12] M. MILLER, R. SHERIDAN, S. KEARSLEY, AND D. UNDERWOOD, FLOG: a system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure, *J. Comput. Aided. Mol. Des.*, vol. 8, pp. 153–174, 1994.

[13] M. RAREY, B. KRAMER, AND T. LENGAUER, Time-efficient docking of flexible ligands into active sites of proteins, in *ISMB-95 Proceedings*, no. 3 in International Conference on Intelligent Systems for Molecular Biology, pp. 300–308, AAAI Press, Menlo Park, CA, 1995.

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

[15] K. PALM, K. LUTHMAN, A.-L. UNGELL, G. STRANDLUND, AND P. ARTURSSON, Correlation of drug absorption with molecular surface properties, *Journal of Pharmaceutical Sciences*, vol. 85, pp. 32–39, January 1996.

[16] A. M. DOWEYKO, Three-dimensional pharmacophores from binding data, *J. Med. Chem*, vol. 37, no. 12, pp. 1769–1778, 1994.

[17] F. R. SALEMME, J. SPURLINO, AND R. BONE, Serendipity meets precision: the integration of structure-based drug design and combinatorial chemistry for efficient drug discovery, *Structure*, vol. 5, pp. 319–324, March 1997.

[18] D. E. PATTERSON, R. D. CRAMER, A. M. FERGUSON, R. D. CLARK, AND L. E. WEINBERGER, Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors, *J. Med. Chem.*, vol. 39, pp. 3049–3059, 1996.

[19] S. GIBSON, R. MCGUIRE, AND D. C. REES, Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments, *J. Med. Chem.*, vol. 39, pp. 4065–4072, 1996.

[20] T. CORMEN, C. LEISERSON, AND R. RIVEST, *Introduction to Algorithms*. McGraw Hill, 1990.

[21] Y. MIYASHITA, Y. TAKAHASHI, Y. YOTSUI, H. ABE, AND S. SASAKI, Clustering Molecules on the Basis of Antibacterial Spectra or Physiochemical Properties, Anal. Chem. Acta., 133, pp. 614–624, 1981.

[22] G. RITTER, AND T. ISENHOUR, Minimal spanning tree clustering of gas chromatographic liquid phases, Computers and Chemistry, Vol. 1, pp. 145–153, 1977.

[23] M. WALDMAN, Personal communication March, 1998.

[24] D. J. CUMMINS, C. W. ANDREWS, J. A. BENTLEY, AND M. CORY, Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds, *J. Chem. Inf. Comput. Sci.*, vol. 36, pp. 750–763, 1996.

[25] N. E. SHEMETULSKIS, J. B. DUNBAR JR., B. W. DUNBAR, D. W. MORELAND, AND C. HUMBLET, Enhancing the diversity of a corporate database using chemical database clustering and analysis, *Journal of Computer-Aided Molecular Design*, vol. 9, pp. 407–416, 1995.

[26] S.L. MAYO, B.D. OLAFSON, W.A. GODDARD, Dreiding: A generic Force Field for Molecular Simulations, *J. Phys. Chem*, 94, 8897, 1990.

[27] W. PRESS, B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING, *Numerical Recipes in C*. Cambridge, 1991.

[28] P. NELSON, GDBM, `http://www.mit.edu:8001/afs /athena.mit.edu/project/gnu/src/g/gdbm-1.7.3/`.

[29] M. GAREY AND D. JOHNSON, *Computers and Intractability, A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979.

[30] D. AGRAFIOTIS Stochastic Algorithms for Maximizing Molecular Diversity, *J. Chem. Inf. Comput. Sci.*, vol. 37, no. 5, pp. 841-851, 1997.

[31] C. JAMES, D. WEININGER, AND J. DELANY, Daylight theory manual. `http://www.daylight.com/dayhtml/doc /theory/theory.toc.html`.

[32] MDL INFORMATION SYSTEMS INC SAN LEANDRO CA, Available chemicals directory, `http://www.mdli.com/prod/suppdb.html`.

[33] D. BORDO, AND P. ARGOS, Suggestions for safe residue substitutions in site-directed mutagenesis, J. Mol. Biol. 217, 1991, pp. 721-729.

[34] B. ALBERTS, D. BRAY, J. LEWIS, M. RAFF, K. ROBERTS, J.D. WATSON, Molecular Biology of the Cell, Garland Publishing, NY, NY 1983.

[35] L. M. KAUVAR, D. L. HIGGINS, H. O. VILLAR, J. R. SPORTSMAN, A. ENGQVIST-GOLDSTEIN, R. BUKAR, K. E. BAUER, H. DILLEY, AND D. M. ROCKE, Predicting ligand binding to proteins by affinity fingerprinting, *Chem. & Biol.*, vol. 2, pp. 107–118, 1995.

[36] H. BRIEM AND I. D. KUNTZ, Molecular similarity based on dock-generated fingerprints, *Journal of Medicinal Chemistry*, vol. 39, no. 17, pp. 3401–3408, 1996.

[37] G. CRIPPEN AND T. HAVEL, *Distance Geometry and Molecular Conformation*. Research Studies Press Ltd., 1988.

[38] NATHAN LINIAL, ERAN LONDON, YURI RABINOVICH, *The geometry of graphs and some of its algorithmic applications*. 35th Annual Symposium on Foundations of Computer Science, IEEE, 1995, pp. 577–591.

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

[40] D. TURNER, S. TYRREL, AND P. WILLETT, Rapid quantification of molecular diversity for selective database acquisition, *J. of Chemical Information and Computer Sciences*, vol. 37, pp. 18–22, November 1997.