A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries

Dimitris K. Agrafiotis[†]

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, Pennsylvania 19341

Received July 11, 2000

We describe a novel diversity metric for use in the design of combinatorial chemistry and high-throughput screening experiments. The method estimates the cumulative probability distribution of intermolecular dissimilarities in the collection of interest and then measures the deviation of that distribution from the respective distribution of a uniform sample using the Kolmogorov−Smirnov statistic. The distinct advantage of this approach is that the cumulative distribution can be easily estimated using probability sampling and does not require exhaustive enumeration of all pairwise distances in the data set. The function is intuitive, very fast to compute, does not depend on the size of the collection, and can be used to perform diversity estimates on both global and local scale. More importantly, it allows meaningful comparison of data sets of different cardinality and is not affected by the curse of dimensionality, which plagues many other diversity indices. The advantages of this approach are demonstrated using examples from the combinatorial chemistry literature.

INTRODUCTION

The measurement of molecular diversity has become an issue of heated debate in recent years.[1−4] Although the concept is widely employed in the design of combinatorial libraries, it has been surprisingly difficult to define, both chemically and mathematically. This controversy has been fueled by a series of so-called “validation” studies, which advocate that diversity measures should be assessed on the basis of their ability to increase the hit rate of high-throughput screening experiments. While no one argues that relevance is an important factor, the root of the problem stems from a lingering confusion between two largely unrelated concepts: molecular diversity and experimental design. The former is an attempt to increase the number of chemical and structural features presented by a finite sample of molecules, whereas the latter is related to the art of contemplating experiments, and represents a vague and multifaceted process that involves a heterogeneous mix of chemistry, mathematics, experience, and intuition.

Perhaps the only undisputed aspect of this problem is the need for algorithmic efficiency. Diversity functions are notoriously hard to compute, and their use becomes particularly problematic with large, high-dimensional data sets. In general, diversity metrics fall into three broad categories: (1) *distance-based* methods, which express diversity as a function of the pairwise molecular dissimilarities defined through measurement or computation; (2) *cell-based* methods, which define it in terms of the occupancy of a finite number of cells that represent disjoint regions of chemical space; (3) *variance-based* methods, which are based on the degree of correlation between the molecules’ pertinent features. In their vast majority, these metrics encode the ability of a given set of compounds to sample chemical space in an even and unbiased manner, and are used to produce space-filling designs that minimize the size of unexplored regions known as “diversity voids”.

Perhaps the most common distance-based diversity measures are the minimum intermolecular dissimilarity,[5]

$$D(C) = \min_{i<j} d_{ij} = \min_{i} \min_{j \neq i} d_{ij} \qquad (1)$$

and the average nearest neighbor distance,[6]

$$D(C) = \frac{1}{N} \sum_{i} \min_{j \neq i} d_{ij} \qquad (2)$$

The main disadvantage of these and other related functions, such as the power-sum,[7] product,[7] minimum spanning tree,[8] etc., is their quadratic dependence on the number of compounds in *C*, which renders them virtually useless for the analysis of large collections. We have recently shown that when the dimensionality of the space is relatively small (<10), the nearest neighbor computation in eqs 1 and 2 can be carried out in an efficient manner using a combinatorial data structure known as a *k*-dimensional (or *k*−*d*) tree, resulting in an algorithm of $O(N \log N)$.[6] This algorithm achieves computational efficiency by first organizing all the points in *C* into a *k*−*d* tree and then performing a nearest neighbor search for each point in the set using a branch-and-bound approach. While the method is very efficient in low-dimensional spaces, it becomes less so as the dimensionality of the space increases and eventually degrades to a quadratic-order algorithm, with the additional overhead of constructing and traversing the tree.

Cell-based methods divide chemical space into hyperrectangular regions and measure the occupancy of the resulting cells.[9−11] They are typically of $O(N)$, and since they encode absolute position in space, they are very useful for detecting and exploiting diversity voids. However, these methods are sensitive to resolution and are applicable to data sets of very modest dimensionality (typically up to 5−6 dimensions).

The motivation for developing the present approach stemmed from the need to devise a diversity function that

[†] Tel: (610) 458-6045. Fax: (610) 458-8249. E-mail: dimitris@3dp.com.

10.1021/ci000091j CCC: $20.00
Published on Web 12/14/2000

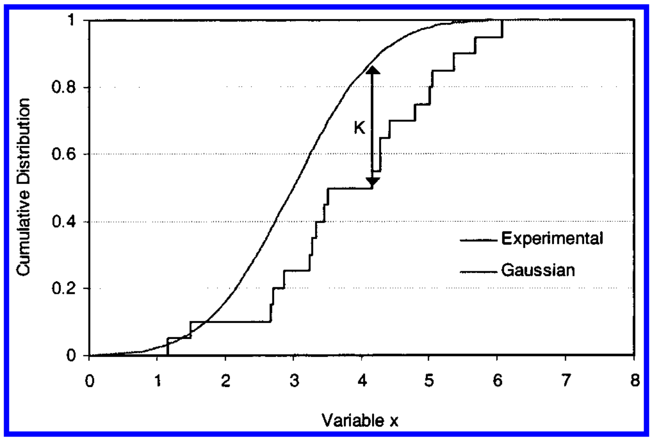

Figure 1. Komogorov−Smirnov statistic for comparing two cumulative distributions.

captures the notion of spread, is fast to compute, scales favorably with the number of compounds in the design, does not fall prey to dimensionality, and can be used to compare collections of different cardinality. The method is based on the fundamental assumption that an optimally diverse sample is one that is uniformly distributed in the property space it is designed to explore. Diversity is quantified by estimating the cumulative probability distribution of intermolecular dissimilarities in the collection of interest and then measuring the deviation of that distribution from the respective distribution of a uniform sample. Departure from uniformity induces sampling redundancy and formation of clusters and thus results in diminishing diversity. The method offers several important advantages over conventional methodologies and can be particularly useful in subset selection where the diversity function needs to be evaluated for a large number of candidate designs.

METHODS

Diversity Index. The proposed diversity index measures the extent to which the probability distribution of intermolecular dissimilarities in the collection of interest deviates from the respective distribution of a uniform sample. This difference is quantified using the Kolmogorov−Smirnov (K−S) statistic.[12] The K−S test is applicable to unbinned distributions that are functions of a single independent variable and is defined as the maximum value of the absolute difference between two cumulative distribution functions:

$$K = \max_{-\infty < x < \infty} |S_1(x) - S_2(x)| \qquad (3)$$

Here $S_1(x)$ and $S_2(x)$ are estimators of the cumulative distribution functions of the actual probability distributions from which they are drawn. For a set of N points x_i, $i = 1$, ..., N, $S(x)$ represents the fraction of data points to the left of a given value x (inclusive). The method is illustrated in Figure 1. A similar formula is used to compare a particular data set's $S(x)$ to a known cumulative distribution function, $P(x)$:

$$K = \max_{-\infty < x < \infty} |S(x) - P(x)| \qquad (4)$$

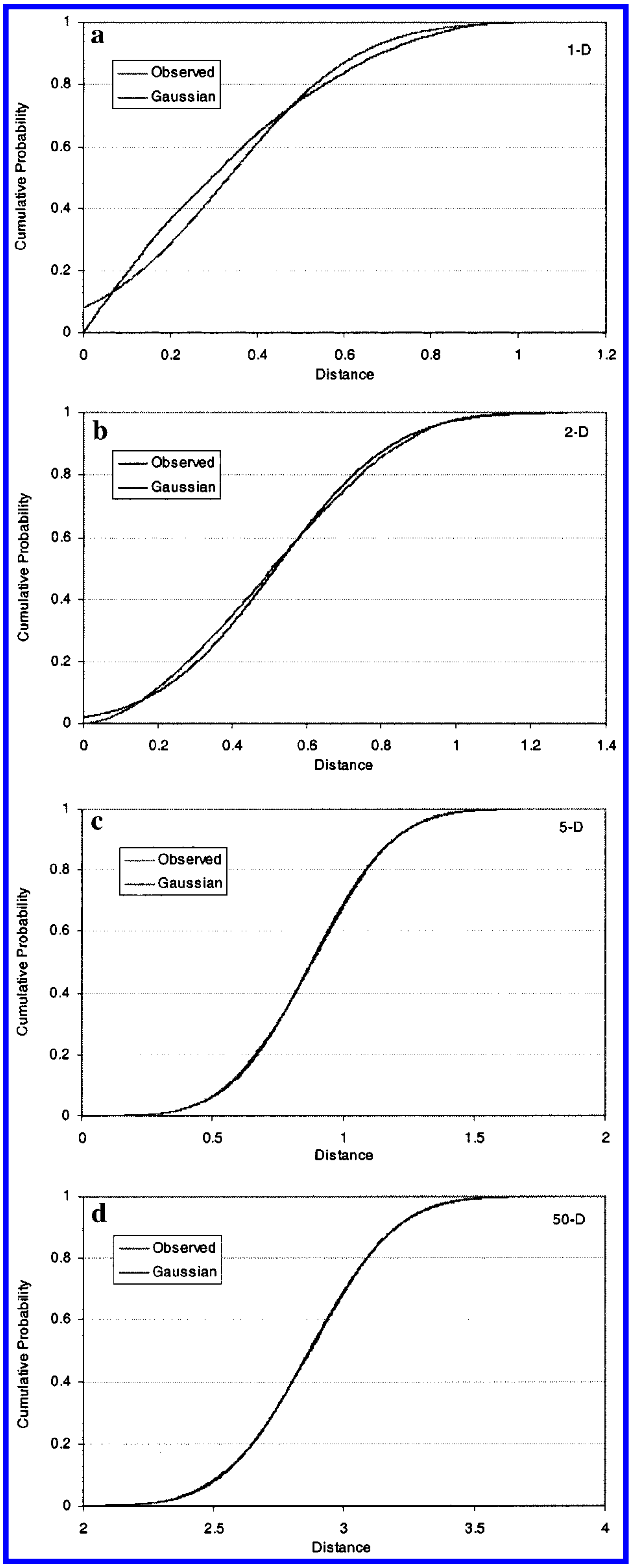

Figure 2. Comparison of the actual cumulative distributions and their Gaussian approximations for a set of points uniformly distributed in the unit hypercube: (a) 2-D; (b) 2-D; (c) 5-D; (d) 50-D hypercube.

Unlike the more commonly used χ^2 test, the Kolmogorov−Smirnov statistic does not require binning of the data, which is arbitrary and leads to loss of information. More importantly, the function is very fast to compute since it involves sorting the data in ascending order, followed by a linear scan to identify the maximum difference from the user-defined

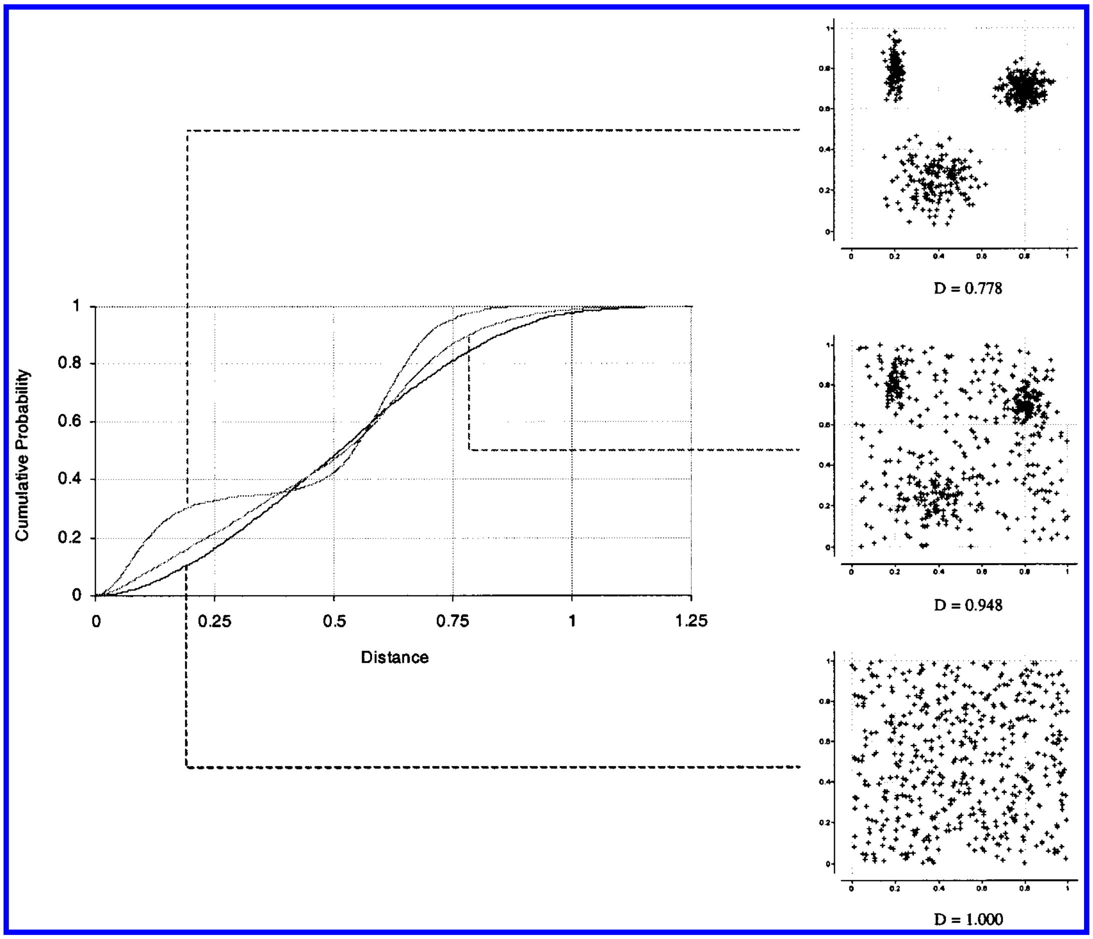

Figure 3. Effect of clustering on the shape of the cumulative probability distribution for three hypothetical 2-dimensional data sets.

cumulative distribution function (or a simultaneous scan of the two data sets in the case of two cumulative distributions). Speed of computation is particularly important in the context of library design, where the diversity function needs to be evaluated for potentially thousands of candidate designs (see below).

Since K is a measure of dissimilarity, the diversity of a given collection of compounds, C, is defined as

$$D = 1 - K(S(x),P(x)) \qquad (5)$$

where $S(x)$ is the cumulative probability distribution of intermolecular dissimilarities in C, $P(x)$ is the respective distribution of a uniform sample, and $K(S(x),P(x))$ is their difference defined by eq 4. D takes values in the interval [0, 1] with 1 indicating complete identity (i.e. maximum diversity) and 0 indicating that the two distributions have nothing in common. The latter is obtained when all the compounds in the collection coalesce into a single compact cluster. Note that the coefficient described above does not explicitly depend on the number of compounds in the collection, which makes it ideal for comparing the diversity of data sets of different cardinality.

The significance level of a particular value of K is a function of K and the number of data points, N. This function is relatively slow to compute, but when N is constant, it is a monotonic function of K. Since all we want is to determine which experimental distribution is closer to the "ideal" distribution $P(x)$, the significance level need not be computed.

Computational Details. All programs were implemented in the C++ programming language and are part of the DirectedDiversity[13] software suite. They are based on 3-Dimensional Pharmaceuticals' Mt++ class library[14] and are designed to run on all Posix-compliant Unix and Windows platforms. Parallel execution on systems with multiple CPUs is supported through the multithreading classes of Mt++. All calculations were carried out on a Dell workstation equipped with two 800 MHz Pentium III Intel processors running Windows 2000 Professional.

RESULTS AND DISCUSSION

Uniform Distributions. It is known from probability theory that the distances between uniformly distributed points should converge to a Gaussian distribution as the dimensionality of the space increases. This follows directly from the central limit theorem, which states that the distribution of the sum of a large number of independent, identically distributed variables is approximately normal. It turns out that, for most practical purposes, this is a very accurate estimation and the error is not even measurable. For simple distances such as the L_1 and squared L_2 norms, the mean and standard deviation of the distribution can be easily derived from the length, L, of the hypercube and the dimensionality, d, of the input space. For example, in the case of the L_1 norm (Manhattan distance), the expectation value for the absolute difference between two independent random variables uniformly distributed in the interval [0, L] is given by

$$
\begin{aligned}
E(|x-y|) &= 2\int_0^L \int_0^x (x-y)p_{xy}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= 2\int_0^L \int_0^x (x-y)p_{xy}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= 2\int_0^L \int_0^x (x-y)p_x(x)p_y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= 2\int_0^L \int_0^x (x-y)\,\frac{11}{LL}\,\mathrm{d}y\,\mathrm{d}x \\
&= \frac{2}{L^2}\int_0^L \frac{x^2}{2}\,\mathrm{d}x \\
&= \frac{L}{3} \qquad (6)
\end{aligned}
$$

where $p_{xy}(x,y)$ is the joint probability distribution of the two random variables with probability distributions $p_x(x)$ and $p_y(y)$, respectively. Since both x and y are independent and uniformly distributed in [0, L], it holds that $p_x(x) = p_y(y) = 1/L$ and $p_{xy}(x,y) = p_x(x)\cdot p_y(y) = 1/L^2$. Thus, the average Manhattan distance in a d-dimensional hypercube is $dL/3$. (The standard deviation is derived in a similar manner.)

Euclidean distributions are also asymptotically normal, but the analytical derivation of the mean and standard deviation is more problematic. The original cumulative distributions of Euclidean distances along with their fitted Gaussians for a unit hypercube with 1, 2, 5, and 50 dimensions are shown in Figure 2. The distribution is essentially normal at $d = 2$ and perfectly so at $d \geq 5$. Significant deviation from normality is observed at $d = 1$ (Figure 2a), as well as other hyperrectangular spaces whose sides differ greatly in length. Since in a typical application it may not be easy to determine the exact shape of these cumulative distribution functions, we have chosen to estimate their parameters by Monte Carlo sampling. This method is robust, expedient, and requires no a priori knowledge of the general functional form of the cumulative distribution.

Effect of Clustering. To illustrate the effect of clustering on the shape of the probability distribution of pairwise distances, we constructed three artificial data sets of 1000 points confined in a 2-dimensional square of unit length. The first consists of 1000 points drawn from a uniform distribution and, according to our definition, represents a maximally diverse set. The other two consist of 1000 samples drawn from three independent Gaussian clusters (two spherical and one ellipsoid) placed at random positions in the box and a hybrid design comprised of 500 points selected from each of these two distributions. The Gaussian clusters were derived using one-dimensional normal random deviates generated with the Box−Muller algorithm,[15] followed by centering around the respective cluster centers and rejection if the point exceeded the boundaries of the unit square. The three data sets are shown in Figure 3, along with their respective cumulative distributions. As clustering intensifies, there is an increase in the frequency of short- (intracluster) and midrange (intercluster) separations, which causes a sharp rise at the corresponding positions in the probability distribution. This difference is reflected in the diversity index (eq 5), which was 1.00, 0.95, and 0.78 for the three data sets, respectively.

Effect of Sample Size. By focusing attention on probability distributions as opposed to individual dissimilarities, our method can take advantage of probability sampling to sharply reduce the computational effort required to evaluate the diversity function. Probability sampling is based on the notion that a small number of randomly chosen members of a given population will tend to have the same characteristics, and in the same proportion, with the population as a whole. Indeed, the probability distributions in eq 5 can be accurately estimated by evaluating only a tiny fraction of pairwise distances, and this is true regardless of the size, dimensionality, and intrinsic clustering of the data space. Consider again the uniform 2-dimensional data set described in the previous paragraph. The cumulative distributions derived from 5 different uniform random samples of 100, 1000, and 10 000 distances are shown in Figure 4a−c, respectively. It is clear that variability decreases rapidly with sample size, and a few thousand distances are sufficient to accurately estimate the probability function.

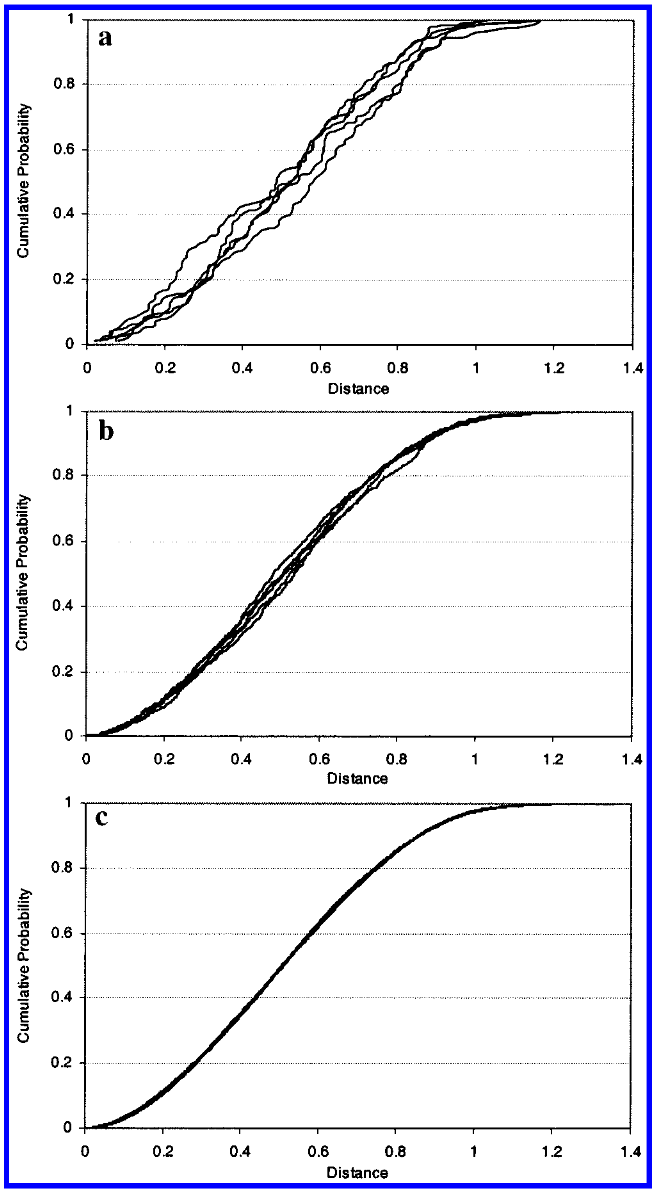

Figure 4. Variability in the estimation of the cumulative probability distribution for a set of points uniformly distributed in the unit square as a function of the number of distances sampled: (a) 100 distances; (b) 1000 distances; (c) 10 000 distances.

Library Design. We now examine the use of this metric in the context of compound selection and library design. The

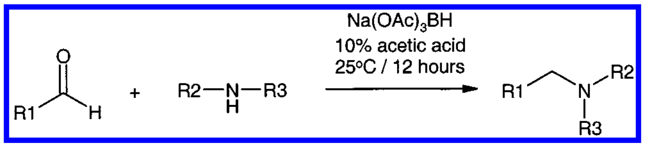

Figure 5. Reaction scheme for the reductive amination library.

data set used in this study is based on the reductive amination reaction (Figure 5) and is utilized for the construction of structurally diverse druglike molecules with useful pharmacological properties, particularly in the GPCR superfamily.[16] For demonstration purposes, 300 primary and secondary amines and 300 aldehydes were selected at random from the Available Chemicals Directory[17] and were used to generate a virtual library of 90 000 products using the library enumeration classes of the DirectedDiversity suite.[13] Each compound in the 90 000-membered library was characterized by an established set of 117 topological descriptors,[18] which were subsequently normalized and decorrelated using principal component analysis, resulting in an orthogonal set of 23 latent variables which accounted for 99% of the total variance in the data. To simplify the analysis and interpretation of results, this 23-dimensional data set was further reduced to 2 and 3 dimensions using a fast nonlinear mapping algorithm developed by our group.[19−22] The resulting nonlinear maps had a Kruskal stress of 0.188 (2-D) and 0.142 (3-D) and are shown in Figure 6a and Figure 7a, respectively.

To reduce the complexity of the search problem, subset selection was carried out in the form of arrays (or full arrays)

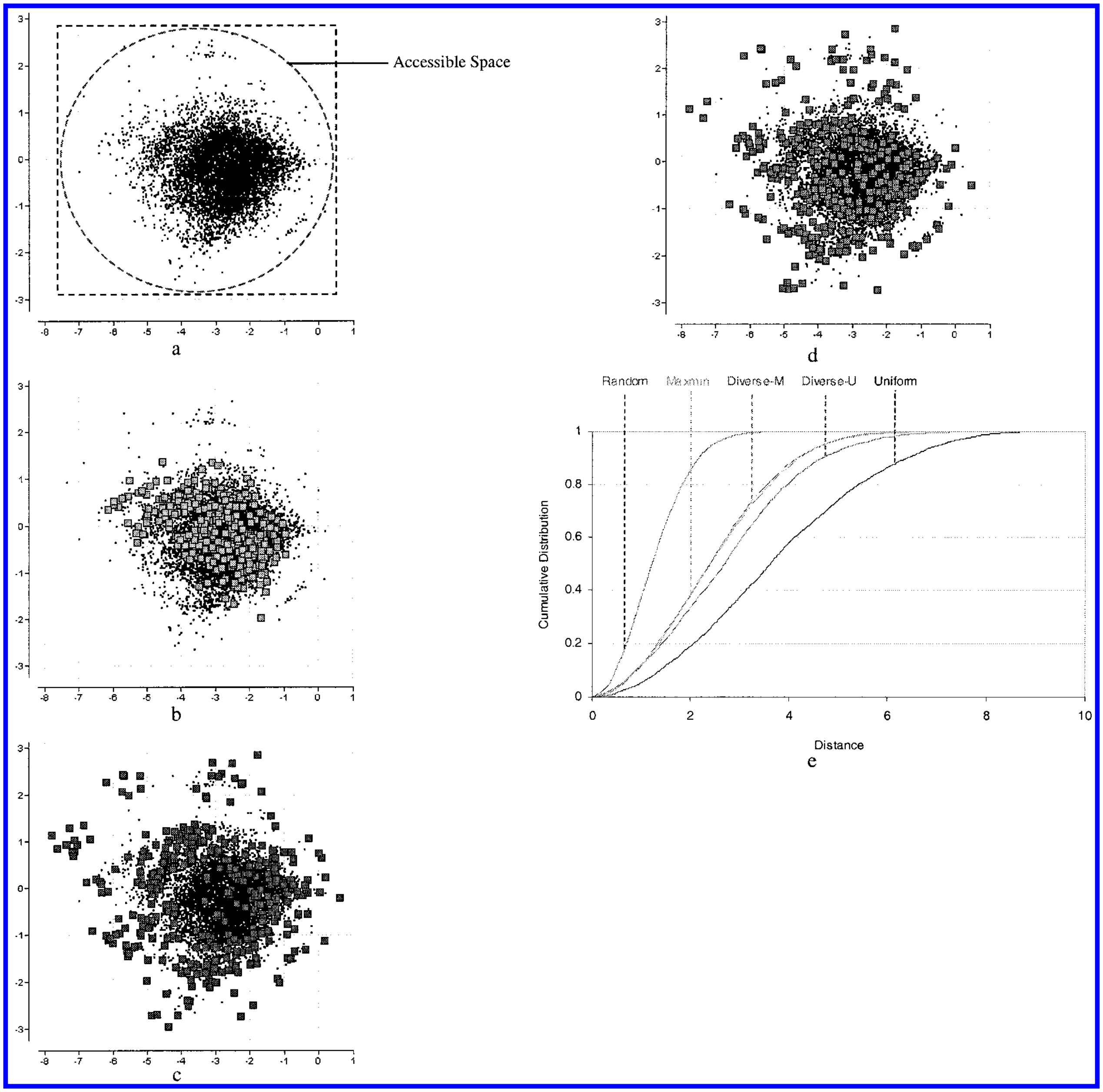

Figure 6. Array selections from the reductive amination library based on the coordinates of the compounds on the 2-D nonlinear map: (a) entire data set; (b) random 20 × 20 array; (c) 20 × 20 array that maximizes similarity to a uniform distribution; (d) 20 × 20 array that maximizes similarity to the distribution of 400 points selected with the maxmin algorithm; (e) respective cumulative probability distributions. In (e), *Random* represents the distribution of 400 compounds chosen at random from the virtual library, *Uniform* represents the distribution of a sample of points uniformly distributed in the rectangle defined by the virtual library, *Maxmin* represents the distribution of 400 points selected with the maxmin algorithm, *Diverse-U* represents the distribution of the array obtained by attempting to match the *Uniform* distribution, and *Diverse-M* represents the distribution of the array obtained by attempting to match the *Maxmin* distribution.

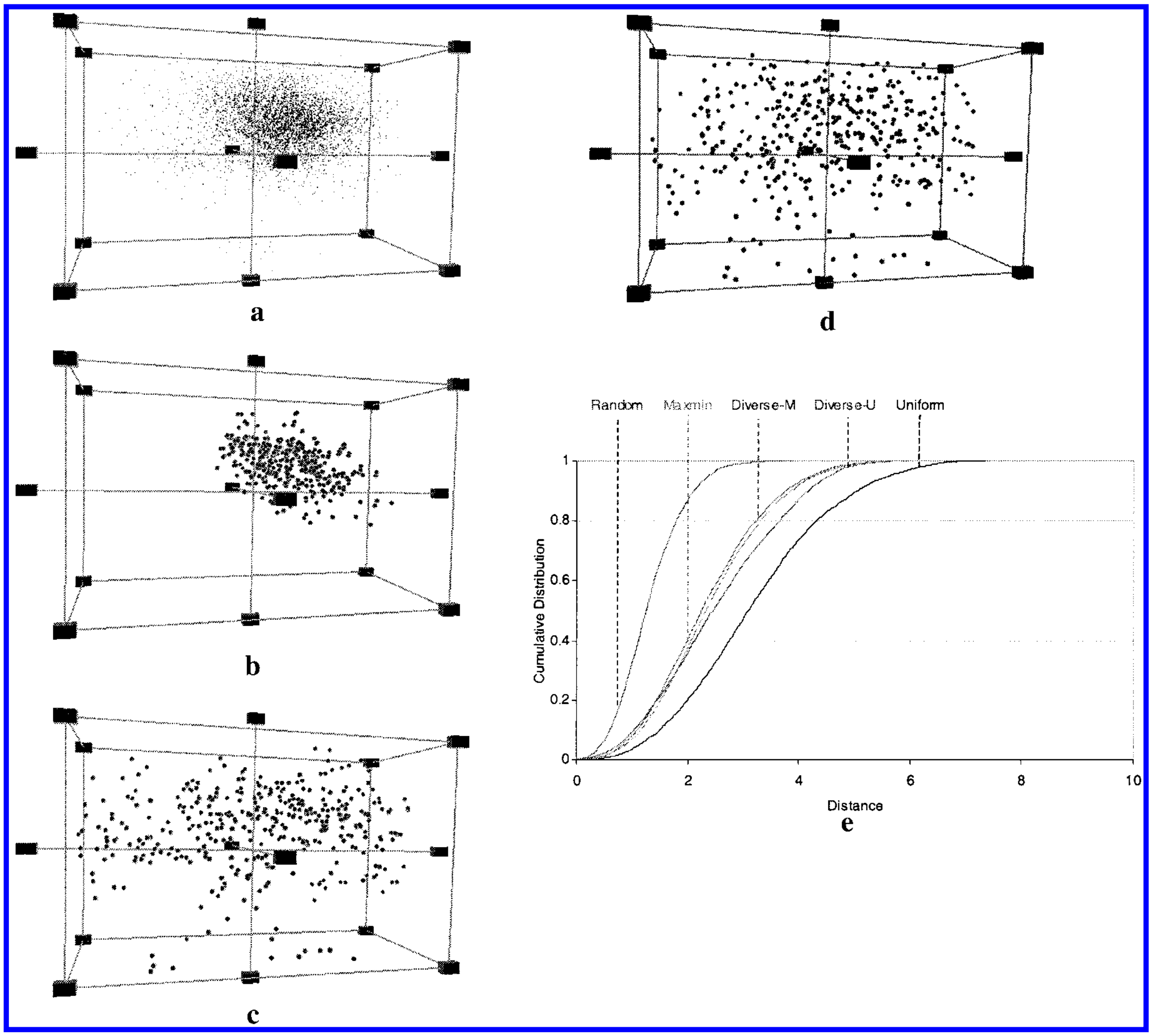

Figure 7. Array selections from the reductive amination library based on the coordinates of the compounds on the 3-D nonlinear map: (a) entire data set; (b) random 20 × 20 array; (c) 20 × 20 array that maximizes similarity to a uniform distribution; (d) 20 × 20 array that maximizes similarity to the distribution of 400 points selected with the maxmin algorithm; (e) respective cumulative probability distributions. See Figure 6 for a description of the labels in (e).

using the simulated annealing algorithm described in refs 23−25. The selection was aimed at identifying the most diverse 20 × 20 array as determined by eq 5, i.e., the array whose distribution of distances approximates as closely as possible that of a uniform sample. As shown in Figure 6c, the selection based on the coordinates of the compounds on the 2-D nonlinear map leads to a somewhat troubling result: while the samples are significantly more dispersed than a random selection (Figure 6b), they tend to concentrate on the periphery of the data space, leaving the core relatively empty. This effect can be easily understood by looking at the scatterplot in Figure 6a and the corresponding cumulative distributions in Figure 6e. Indeed, the library occupies approximately one-fourth to one-third of the pertinent property space, making the target distribution an unattainable goal (the reader should be reminded that the target uniform distribution is estimated by generating pairs of points at random positions in the property box defined by [$x_{\min}$, $x_{\max}$] and [$y_{\min}$, $y_{\max}$] (dashed square in Figure 6a) and recording their distances). The absence of samples in the extremes of that box severely restricts the fraction of points at large separations, which dominate the value of the K−S statistic. To match the target distribution, the optimization algorithm attempts to shift the cumulative distribution to the right by increasing the fraction of large distances in the design. This is best achieved by arranging the points away from the center into some sort of an extended annulus.

Although the effect is not as discernible in 3 dimensions (i.e. when the distances are evaluated on the basis of the coordinates of the compounds on the 3-D nonlinear map; see Figure 7), it does become more pronounced with increasing dimensionality. It is known, for example, that the fraction of the volume of a *d*-dimensional hypercube contained within the inscribed hypersphere is given by[26]

$$F(d) = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \qquad (7)$$

For $d = 1-7$, *F(d)* is 1, 0.785, 0.524, 0.308, 0.164, 0.081, and 0.037, respectively, which means that in high-dimensional spaces the center of the hypercube becomes insignificant and most of its volume is concentrated near its corners. This is one of the many manifestations of the *curse of dimensionality*, a term which is often used to refer to the sparsity of data in higher dimensions.

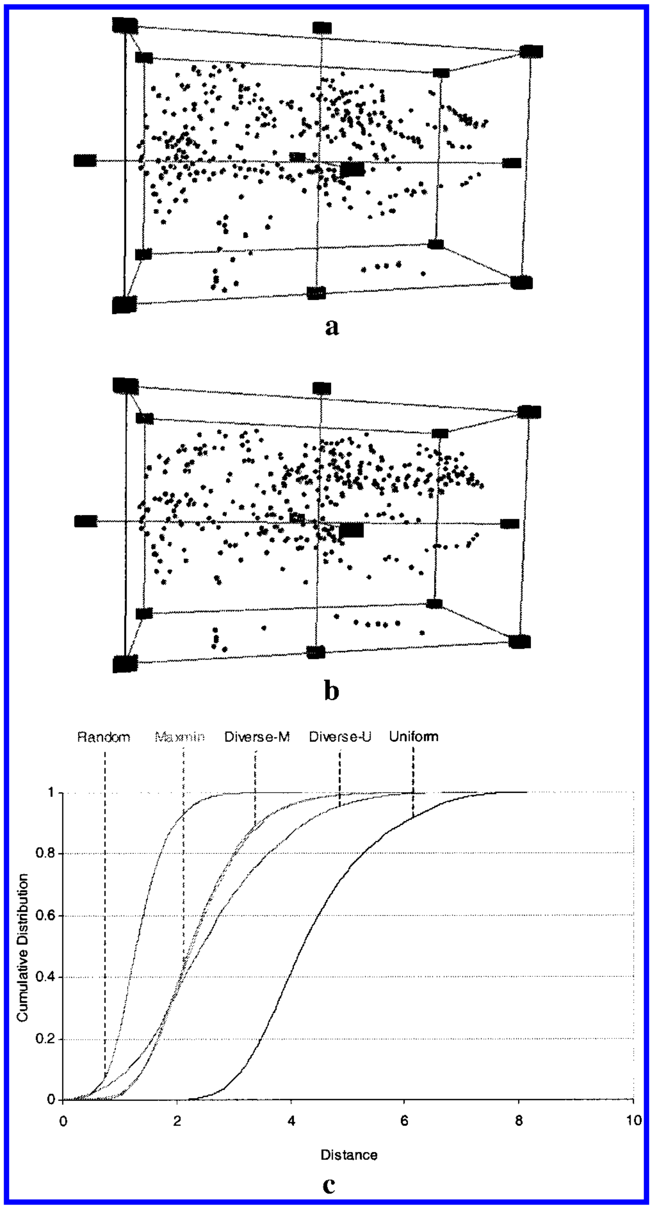

Figure 8. Array selections from the reductive amination library based on all 23 principal components highlighted on the 3-D nonlinear map: (a) 20 × 20 array that maximizes similarity to a uniform distribution; (b) 20 × 20 array that maximizes similarity to the distribution of 400 points selected with the maxmin algorithm; (c) respective cumulative probability distributions. See Figure 6 for a description of the labels in (c).

A simple way to eliminate this problem is to restrict attention to the region of space *accessible* by a particular library (dashed circle in Figure 6a). One can devise several strategies to achieve this goal; the approach taken here is to extract a small representative sample of the virtual library and use a conventional algorithm such as maxmin[5] to identify a maximally diverse subset of points, which are then used to estimate the target probability distribution. Although it may at first seem odd that the selection of a maximally diverse set would require knowledge of another maximally diverse set, the reader should be reminded that the initial selection need only be performed once and the problem is much smaller in scale (this calculation typically requires a few seconds on a modern personal computer). The resulting designs in 2, 3, and 23 dimensions are illustrated in Figures 6d, 7d, and 8b, and the corresponding distributions, in Figures 6e, 7e, and 8c, respectively. In these plots, *Random* represents the distribution of 400 compounds chosen at random from the virtual library, *Uniform* represents the distribution of a sample of points uniformly distributed in the rectangle defined by the virtual library, *Maxmin* represents the distribution of 400 points selected with the maxmin algorithm, *Diverse-U* represents the distribution of the array obtained by attempting to match the *Uniform* distribution, and *Diverse-M* represents the distribution of the array obtained by attempting to match the *Maxmin* distribution. By minimization of the influence of large distances, the cumulative distribution is shifted to the left, and the selection does not exhibit any preference for the outer regions of the property space, nor does it manifest any significant clustering that reflects the density distribution of the parent collection. A look at the reagents that comprise the optimal 20 × 20 array based on the coordinates of the compounds on the nonlinear map (Figure 9) confirms that the selection is indeed diverse, as it consists of building blocks containing a wide variety of atom types, connectivity patterns, ring systems, and functional groups.

The most compelling reason for the use of this metric is, of course, the fact that execution time is essentially constant with respect to the size of the collection being analyzed. This becomes particularly important for large, high-dimensional data sets where cell-based approaches are impractical and distance-based methods become quadratic. For example, the time required to complete the selection of a 20 × 20 array in the full 23-dimensional principal component space using 30 temperature cycles and 1000 sampling steps/cycle (i.e. a total of 30 000 function evaluations) was 8 CPU min on an 800 MHz Pentium III Intel processor, and this time is roughly the same regardless of the size of the array being optimized. Conversely, the time required to perform the same selection using the mean nearest-neighbor distance[6] was 73 min, and that time scales to the square of the number of compounds selected. This differential is smaller in lower dimensions where algorithmic techniques such as $k-d$ trees[6] can reduce the complexity of nearest-neighbor detection, but the problem eventually manifests itself when the size of the design becomes very large.

CONCLUSIONS

The method presented herein was designed for the analysis of large data sets and is particularly useful in the context of subset selection where the diversity function needs to be evaluated hundreds to thousands of times in the course of optimization. It is the only algorithm of its kind whose execution time does not increase with the size of the data set and offers the ability to compare collections of different cardinality. Of course, the method is only as good as the underlying parameters used to define chemical space and the extent of their relevance to the problem at hand. Its use does not increase or decrease the hit rate of high-throughput screening experiments nor does it improve the likelihood of identifying sustainable leads. It merely simplifies quantitation and by doing so extends diversity profiling to a class of problems that are intractable with conventional methodologies.

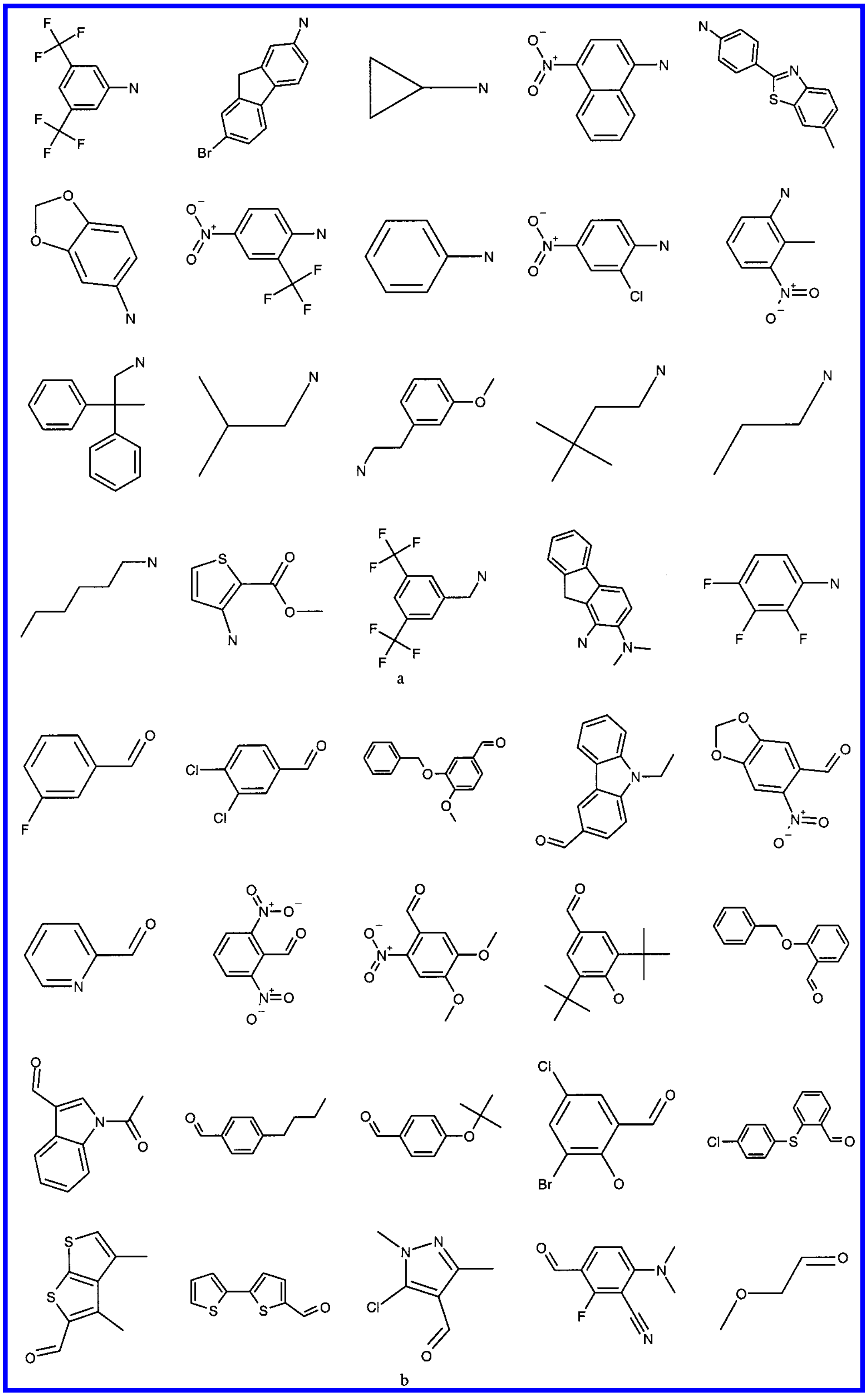

Figure 9. Reagents comprising the optimal 20 × 20 array based on the coordinates of the compounds on the 2-D nonlinear map and the distribution of 400 points selected with the maxmin algorithm (see Figure 6d): (a) amines; (b) aldehydes.

ACKNOWLEDGMENT

The author thanks Dr. Dmitrii N. Rassokhin, Dr. Victor S. Lobanov, and Dr. Sergei Izrailev of 3-Dimensional Pharmaceuticals, Inc., for many useful discussions and Dr. Raymond F. Salemme for his insightful comments and support of this work.

REFERENCES AND NOTES

(1) Agrafiotis, D. K. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley and Sons: Chichester, U.K., 1998; pp 742−761.

(2) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. *Mol. Diversity* **1999**, *4* (1), 1−22.

(3) Agrafiotis, D. K.; Lobanov, V. S.; Rassokhin, D. N.; Izrailev, S. In *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, Germany, 2000; pp 265−300.

(4) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E. Jr.; Blaney, J. M. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: Weinheim, Germany, 1997; Vol. 10.

(5) Lajiness, M. S. In *QSAR: Rational Aproaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 201−204.

(6) Agrafiotis, D. K.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci*. **1999**, *39* (1), 51−58.

(7) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. *Mol. Diversity* **1996**, *2*, 64−74.

(8) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. *J. Med. Chem.* **1999**, *42*, 60−66.

(9) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(10) Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.*, in press.

(11) Jamois, E. A.; Hassan, M.; Waldman, M. *J. Chem. Inf. Comput. Sci.*, in press.

(12) von Mises, R. *Mathematical Theory of Probability and Statistics*; Academic Press: New York, 1997.

(13) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. U.S. Patent 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; 5,901,069, 1999.

(14) Copyright © 3-Dimensional Pharmaceuticals, Inc., 1994−2000.

(15) Box, G. E. P.; Muller, M. E.; Marsaglia, G. *Ann.. Math. Stat.* **1958**, *28*, 610.

(16) Dhanoa, D. S.; Gupta, V.; Sapienza, A.; Soll, R. M. Presented at the American Chemical Society National Meeting, Anaheim, CA, 1999; Poster 26.

(17) MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.

(18) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.

(19) Agrafiotis, D. K. *Protein Sci.* **1997**, *6*, 287−293.

(20) Agrafiotis, D. K.; Lobanov, V. S. *J. Chem. Info. Comput. Sci*. **2000**, *40*, 1356−1362.

(21) Rassokhin, D. N.; Lobanov, V. S.; Agrafiotis, D. K. *J. Comput. Chem.*, in press.

(22) Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. *J. Comput. Chem.*, in press.

(23) Agrafiotis, D. K. *J. Chem. Info. Comput. Sci.* **1997**, *37*, 841−851.

(24) Agrafiotis, D. K. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (3), 576−580.

(25) Rassokhin, D. N.; Agrafiotis, D. K. *J. Mol. Graphics Model*. **2000**, *18* (4−5), 370−384.

(26) Scott, D. W. *Multivariate density estimation: theory, practice and visualization*; Wiley: New York, 1992.

CI000091J