A Fractal Approach for Selecting an Appropriate Bin Size for Cell-Based Diversity Estimation

Dimitris K. Agrafiotis* and Dmitrii N. Rassokhin

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

Received September 3, 2001

A novel approach for selecting an appropriate bin size for cell-based diversity assessment is presented. The method measures the sensitivity of the diversity index as a function of grid resolution, using a box-counting algorithm that is reminiscent of those used in fractal analysis. It is shown that the relative variance of the diversity score (sum of squared cell occupancies) of several commonly used molecular descriptor sets exhibits a bell-shaped distribution, whose exact characteristics depend on the distribution of the data set, the number of points considered, and the dimensionality of the feature space. The peak of this distribution represents the optimal bin size for a given data set and sample size. Although box counting can be performed in an algorithmically efficient manner, the ability of cell-based methods to distinguish between subsets of different spread falls sharply with dimensionality, and the method becomes useless beyond a few dimensions.

I. INTRODUCTION

Molecular diversity continues to attract significant interest in the combinatorial chemistry and high-throughput communities.[1−4] Despite an increase in the sophistication and involvement of diversity profiling techniques in library design and compound acquisition, the concept has been difficult to define in both a chemical, and mathematical sense. Diversity functions are very 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 pairwise molecular dissimilarities, (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, and (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".

Cell-based methods divide chemical space into hyperrectangular regions and measure the occupancy of the resulting cells.[5−7] They are intuitive, programmatically simple, and fast to compute, and can be particularly useful in subset selection where the diversity function needs to be evaluated for a large number of candidate designs. This has been a major advantage over distance-based methods, which usually scale quadratically with the size of the data set, and are thus unsuitable for the analysis of large collections. Although algorithmic improvements such as the use of *k*-d trees can reduce the complexity of nearest neighbor detection, which lies at the heart of many distance-based metrics,[8] the scaling problem eventually manifests itself in data spaces of nontrivial complexity. Other, more recent probabilistic approaches address the performance issue very effectively but are generally more difficult to interpret.[9] An additional advantage of cell-based techniques is that they encode absolute position in space and thus provide a natural way of detecting and exploiting diversity voids. Several cell-based diversity measures have been proposed in the literature, ranging from simple occupancy counts to entropy measures, χ^2 values, and other metrics.[7]

However, two factors that have often been overlooked are the effects of dimensionality and grid resolution. These methods are becoming increasingly popular through the availability of the DiverseSolutions software suite developed by Pearlman,[10] but the choice of resolution is still guided primarily by intuition and lacks any theoretical or empirical support. In this work, we present a systematic analysis of several typical chemical data sets and propose a simple technique for identifying a suitable bin size for cell-based diversity estimation using an algorithm inspired from the field of fractal analysis. We demonstrate that the relative variance of the diversity score as a function of resolution exhibits a characteristic bell shape that depends on the size, distribution, and dimensionality of the data set under consideration, and whose maximum represents the optimum resolution for a given data set.

II. METHODS

Algorithm. The method described herein attempts to identify the grid resolution at which the diversity of different point sets can be discriminated more easily. This is accomplished by generating random subsets of *k* objects from the collection of interest, measuring their diversity at several grid resolutions, and identifying the resolution at which the relative variance of the diversity score over all random subsets assumes its maximum value.

In this study, the diversity of a *d*-dimensional point set *T* embedded in a grid of hypercubic cells of width *r* is defined

* Corresponding author phone: (610)458-6045; fax: (610)458-8249; e-mail: dimitris@3dp.com.

10.1021/ci010314l CCC: $22.00
Published on Web 01/05/2002

as a function of the sum of squared occupancies of all the cells in the grid, $S(T, r)$

$$D(T,r) = -S(T,r) = -\sum_i C_{ri}^2 \qquad (1)$$

where C_{ri} denotes the occupancy of the ith cell (i.e. the number of points occupying the cell). This quantity is widely employed in fractal analysis and is used to compute the correlation fractal dimension, D_2, of self-similar data sets. D_2 is defined as the derivative of the logarithm of $S(T, r)$ with respect to the logarithm of the cell width, r

$$D_2(T) = \frac{d \log S(T,r)}{d \log r} \qquad r \in [r_1, r_2] \qquad (2)$$

where $[r_1, r_2]$ is the range of scales for which the data set exhibits self-similarity. Thus, the fractal dimension, D_2, can be computed by plotting $S(T, r)$ for different values of r and measuring the slope of the linear part of the resulting log−log curve. Although in the present paper we are not interested in the fractal dimension itself, we have a similar need to measure $S(T, r)$ at different grid resolutions, a problem that has been addressed several times in the fractal literature.[11−13]

Since the number of cells in the grid increases exponentially with dimensionality, a naïve approach involving direct indexing into a multidimensional array is impractical for all but the simplest cases. Traina et al.[13] recently presented an efficient algorithm involving a hierarchical grid structure, where each level has a radius half the size of the previous level ($r = 1, 1/2, 1/4, 1/8$, etc). The algorithm starts with the unit cell ($r = 1$) and iteratively subdivides it into 2^d subcells of equal size. This process is repeated for every populated cell until the desired depth (resolution) is reached. The advantage of this approach is that it involves a single pass through the data set, and only occupied cells need to be stored and explicitly analyzed during each iteration. The algorithm is $O(n \cdot d \cdot k)$ where n is the number of data points, d is the dimensionality of the data, and k is the number of resolutions scanned, i.e., the number of points in the log−log plot of $S(T, r)$ vs r.

Although Traina's approach is very appealing, its implementation can be very wasteful for high-dimensional data spaces. Their algorithm is based on a data structure similar to a multidimensional quad-tree, where each cell is represented by (1) its occupancy, C_{ri}, and (2) a vector of pointers to the 2^d subcells in the next level of the hierarchy covered by that cell. Although these pointers are null if the respective subcells are unoccupied, for large d, the size of the pointer array becomes prohibitively large and the algorithm collapses.

Our solution to this problem is to store in each cell only the pointers to the occupied subcells, along with a binary set that reflects the relative position of that cell with respect to its parent. The binary set, which is also referred to as *position mask*, consists of d bits, where the value of each bit signifies whether the cell occupies the lower (0) or upper (1) half in the respective dimension. Its purpose is to uniquely and compactly identify each subcell within a given cell and serve as an indexing key for fast retrieval. The pointers to the occupied subcells are stored in a map or a sorted collection, which typically exhibit $O(\log n)$ retrieval times (here implemented as a sorted pointer vector using binary search for insertion and retrieval).

The data structure is built by examining each point in turn and recursively determining into which of the subcells it falls at any given level of the hierarchy. As each new point, $\boldsymbol{x}$, is being examined at a particular resolution, a decision must be made as to whether the subcell in which this point resides has already been visited. This is done by computing the position mask, $\boldsymbol{p}$, of the current point according to eq 3

$$\mathbf{p}_i = \begin{cases} 0 \text{ if } \mathbf{x}_i \le (\mathbf{c}_{\min,i} + \mathbf{c}_{\max,i})/2 \\ 1 \text{ if } \mathbf{x}_i > (\mathbf{c}_{\min,i} + \mathbf{c}_{\max,i})/2 \end{cases} \qquad i = 1, 2, ..., d \qquad (3)$$

where $\boldsymbol{x}_i$ is the ith coordinate of the current point, and $\boldsymbol{c}_{min,i}$, and $\boldsymbol{c}_{max,i}$ are the lower and upper boundaries of the current cell $\boldsymbol{c}$ in the ith dimension, and performing a fast search in the sorted collection of existing subcells using $\boldsymbol{p}$ as an index. If a subcell with the same position mask is found, it becomes the current cell, its counter is incremented, and the algorithm continues with the next level. If the cell is not found, a new cell is allocated and added to the collection of subcells of the parent cell, and its counter is set to 1. Once all the points are processed, the tree is traversed in a top-down fashion, and the occupancies of the cells are cumulated to compute the value of S at each resolution. While this method requires some minimal additional work to locate the occupied subcells, it has minimal memory requirements and can be used with data sets of any dimensionality.

This process is repeated for many different k-subsets of an n-point set, the relative variance, V, of the sum of squared occupancies is plotted against r, and the resolution for which this variance is maximized is chosen for any subsequent diversity profiling or compound selection tasks. V is defined as

$$V = \frac{\sigma(\log S)}{\mu(\log S)} \qquad (4)$$

where $\mu(\log S)$ and $\sigma(\log S)$ represent the mean and the variance of $\log S$, respectively.

Computational Details. All computations were carried out using proprietary software written in the C++ programming language and based on 3-Dimensional Pharmaceuticals' Mt++ class library.[14] These programs are part of the DirectedDiversity[15] software suite and were designed to run on all POSIX-compliant Unix and Windows platforms. The calculations were performed on a Dell Inspiron 7500 laptop computer equipped with a 733 MHz Intel Pentium III processor running Windows 2000.

III. RESULTS AND DISCUSSION

The sum-of-squared-occupancies in eq 1 measures the evenness of the distribution of a given set of points among the grid cells and decreases with increasing diversity. Indeed, $S(T, r)$ is at a minimum when every point occupies its own cell and at a maximum when all the points coalesce into a single cell. Our method is based on the observation that even the most differently distributed data sets become increasingly indistinguishable at very coarse and very fine resolutions. At $r = 1$, every point occupies the same cell, and S is equal

to k^2, where k is the number of points in the sample. On the other extreme, as $r \rightarrow 0$, each point tends to occupy its own cell, and S assumes an asymptotic value that depends on the number of duplicates. Clearly, the method becomes more discriminative at intermediate values of r, but finding the optimum resolution is impossible without some knowledge of the general probability distribution from which the data is drawn.

Although the function in eq 1 is negative definite and therefore somewhat unusual, it is a "true" diversity function in the sense that its value increases with increasing diversity. Like many other cell-based diversity measures that have appeared in the literature,[7] it makes the simplistic assumption that pairs of compounds located in adjacent cells are no more different than pairs of compounds located in the most remote cells, regardless of the nature of the property space and grid resolution. In most cases, this is a very poor assumption, particularly when the dimensionality of the space is high. The problem can be easily ameliorated by replacing the counts in eq 1 with local densities obtained, for example, with a Parzen density estimator. Although our method is general and can be used with any measure of this type, the reader should be aware that the specific results may differ.

Our analysis is based on one artificial and two chemical data sets that we routinely use to test new diversity profiling and library design algorithms. The first set consists of 10 000 random points uniformly distributed in a unit hypercube of 2, 4, 8, and 16 dimensions, whereas the last two sets consist of a 90 000-member combinatorial library based on the reductive amination reaction (see refs 16 and 17) and 214 820 commercially available reagents from the 1998 release of the Available Chemicals Directory.[18] Each compound in these data sets was characterized by three different sets of descriptors: (1) Kier-Hall topological indices (KH), which include 117 molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstić indices, and topological state indices,[19,20] (2) Ghose-Crippen atom counts (GC), which represent an extended set of 144 atom types used by the ALOGP method for the prediction of hydrophobic properties of small organic compounds,[21] and (3) ISIS substructure keys (IK), which represent 166-dimensional binary vectors where each bit encodes the presence or absence of a particular structural feature in the molecule. The bit assignment is based on the fragment dictionary used in the ISIS chemical database management system and is often referred to as the ISIS *public keys*. To eliminate redundancy and study the effect of dimensionality, the three sets of descriptors were independently normalized and decorrelated using principal component analysis (PCA). For each data set, the first 2, 4, 8, and 16 principal components were extracted and were used to represent four different chemical spaces of increasing dimensionality. We should point out that PCA is not the best possible choice for reducing the dimensionality of binary fingerprints. This is done here merely for convenience. A much better approach would be to choose a similarity measure, construct the pairwise similarity matrix of the compounds under investigation, and then nonlinearly map them into a latent space of the desired dimensionality. We have recently presented several very effective algorithms for performing this transformation that are applicable to massive data sets, of both conventional[22] and combinatorial[23] nature.

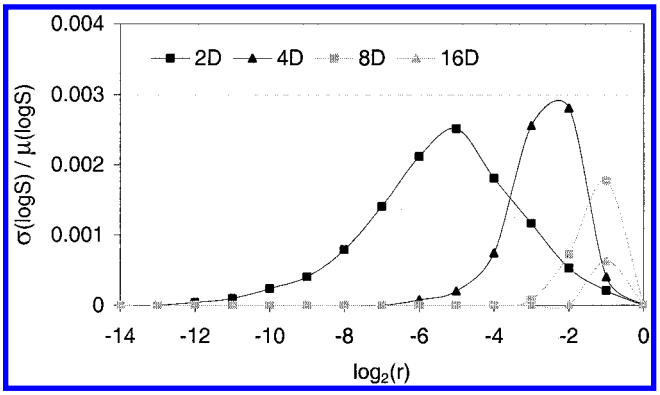

Figure 1. Plot of relative variance (V) vs the base 2 logarithm of the cell width (r) for 10 000 random points uniformly distributed in a unit hypercube of 2, 4, 8, and 16 dimensions.

Figure 1 shows the dependence of V on grid resolution and dimensionality for the random data. For two dimensions, V shows a maximum at $r = 1/32$, but this maximum falls sharply with increasing dimensions. Even in four dimensions, a mere four bins per dimension appear to represent the optimum resolution, and that number drops to 2 in 8 and 16 dimensions.

Similar trends, though less extreme, are also seen with the molecular descriptor data. To study the effect of density, we constructed two sets of V plots derived from random samples of 100 and 10 000 points, respectively. These plots for the reductive amination library are shown in Figures 2 and 3. In all cases, the V plots exhibit a characteristic bell shape with a relatively well-defined maximum, but the exact form of these plots and the location of the maximum depends on the nature of the descriptors, the number of dimensions, and the size of the sample. Invariably, as the sample size increases, the maximum is shifted to higher resolutions, which suggests that cell-based methods may not be suitable for comparing collections of very different cardinality. As was the case with the uniform data, the maximum becomes more sharply defined with increasing dimensions, which is another manifestation of the *curse of dimensionality*,[24] i.e., the sparsity of data in higher dimensions.

These trends are not limited to combinatorial libraries but are also true for more diverse collections of compounds such as the ACD. However, when applied to the ACD, the V plot (Figure 4) revealed an interesting feature: the variance did not diminish to zero at high resolutions but rather plateaued around a value of 0.01. This reflects the degeneracy of the data, i.e., the presence of points with identical descriptor (PC) values. This is typical of graph-based and fragment count descriptors, but the problem can be easily alleviated by introducing a small amount of random noise to the coordinates prior to binning. The resulting plots, shown in Figures 5 and 6, have a similar shape, but in all cases the maxima are shifted to substantially higher resolutions compared to the amination library.

So, how does this method help one in deciding the optimal bin size for cell-based diversity estimation? For example, one of the reviewers suggested that the analysis of random subsets may not be relevant in this context, since the users of these methods are never interested in random designs. The reviewer argued that a better criterion would be the difference between the best (most diverse) selection and a random one, particularly in high-dimensional spaces where

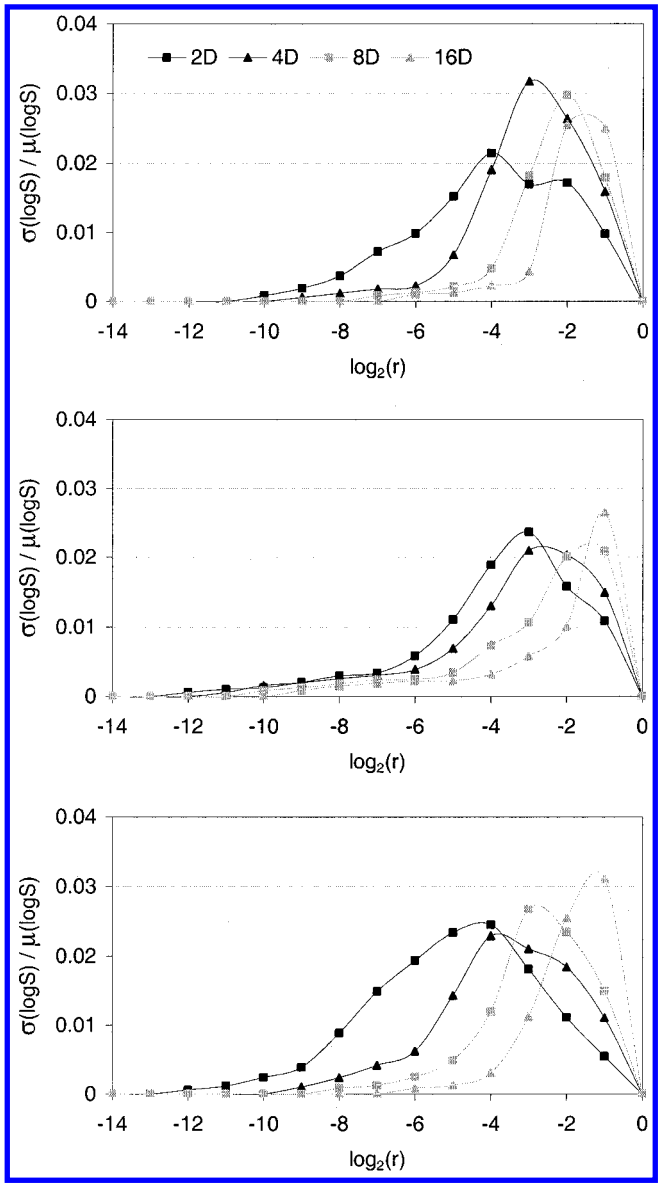

Figure 2. Plot of relative variance (*V*) vs the base 2 logarithm of the cell width (*r*) for 100 compounds randomly chosen from the amination library using the 2, 4, 8, and 16 principal components that accounted for most of the variance in the (a) Kier-Hall, (b) Ghose-Crippen, and (c) Isis Keys descriptor sets.

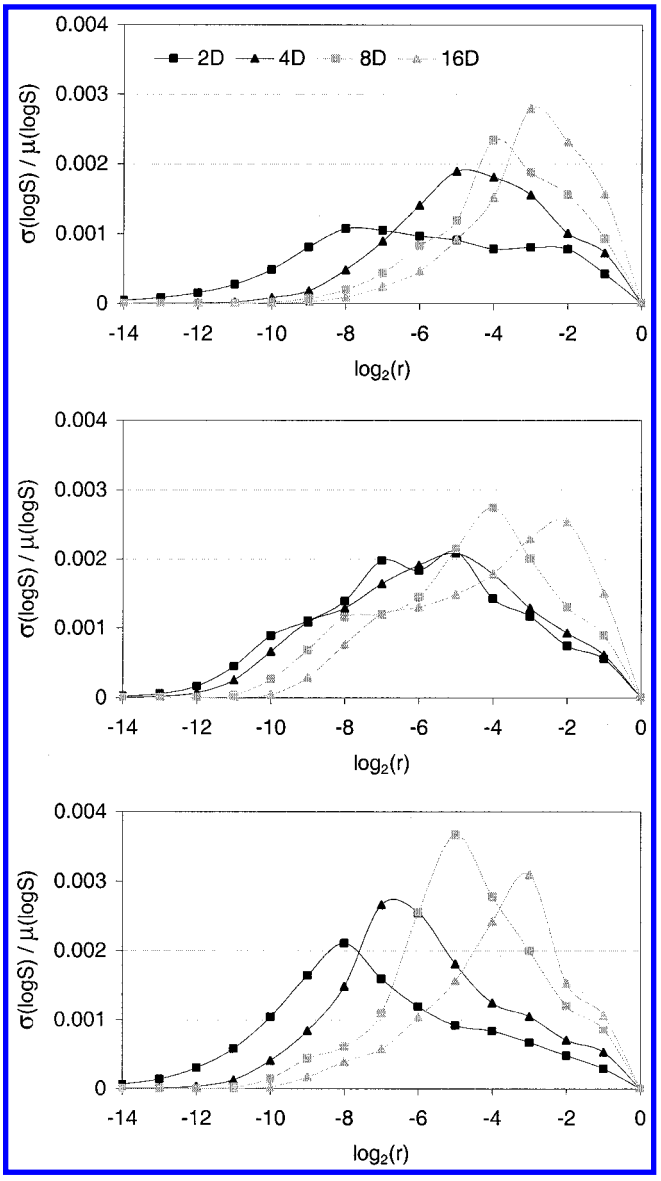

Figure 3. Plot of relative variance (*V*) vs the base 2 logarithm of the cell width (*r*) for 10 000 compounds randomly chosen from the amination library using the 2, 4, 8, and 16 principal components that accounted for most of the variance in the (a) Kier-Hall, (b) Ghose-Crippen, and (c) Isis Keys descriptor sets.

the distribution of pairwise distances becomes very narrow. In such cases, most random designs would score roughly the same, but the difference between random selections and diverse ones would be very substantial because of the long distribution tails. This is actually reflected in the *V* plots, which show very small relative variances for every selection containing 10 000 compounds.

There are two problems with this reasoning. The first is that diversity is rarely the only objective in combinatorial library design. Most often, diversity is combined with other design criteria such as the drug-likeness of the selected compounds, the cost and availability of the starting materials, etc.[16] Thus, a maximally diverse subset would be a good reference point only in a relatively small number of applications. But more importantly, determining that maximally diverse subset would necessitate that the best resolution already be known, leading to a circular argument. In fact, it is unclear that a single reference subset would be sufficient

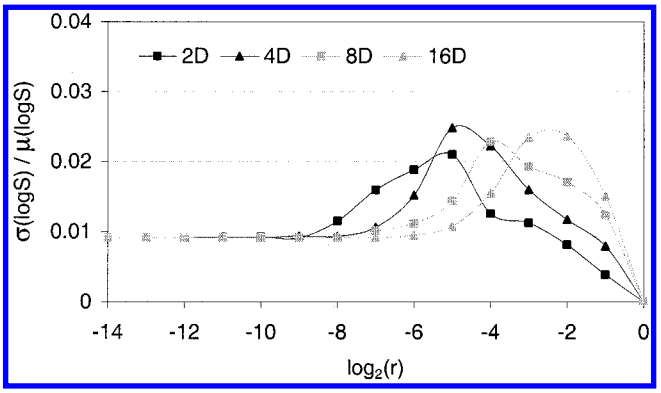

Figure 4. Plot of relative variance (*V*) vs the base 2 logarithm of the cell width (*r*) for 100 compounds randomly chosen from the ACD using the 2, 4, 8, and 16 principal components that accounted for most of the variance in the Kier-Hall descriptor set.

for that purpose. This approach taken here attempts to find a resolution that can, *on average*, discriminate more ef-

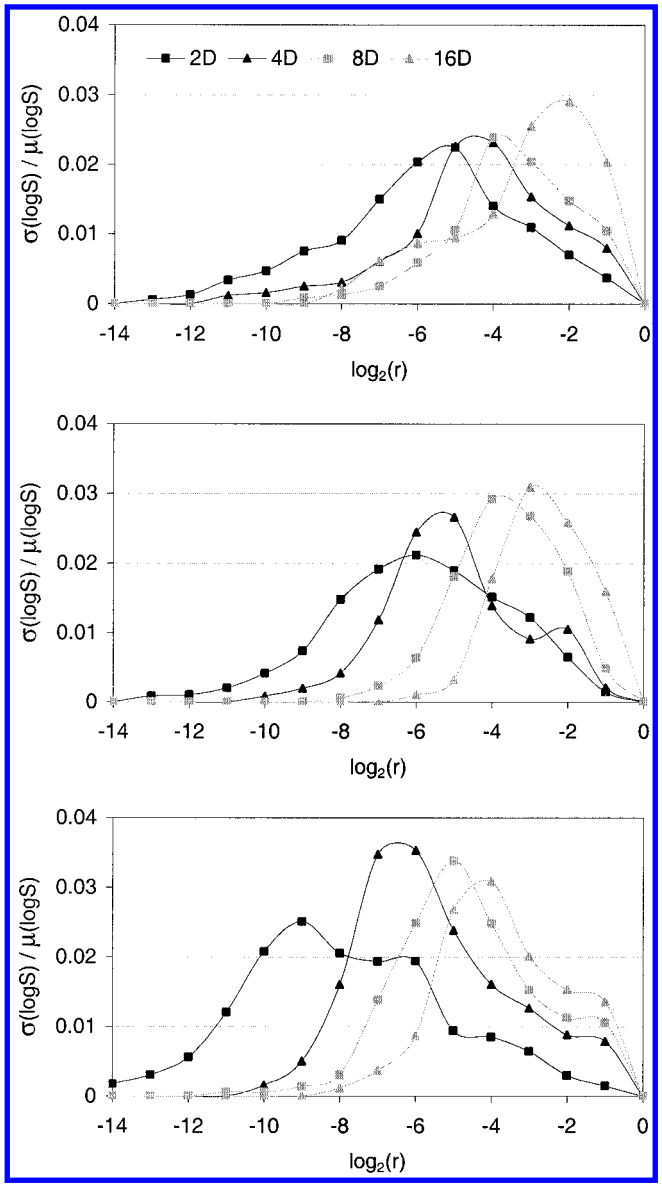

Figure 5. Plot of relative variance (*V*) vs the base 2 logarithm of the cell width (*r*) for 100 compounds randomly chosen from the ACD using the 2, 4, 8, and 16 principal components that accounted for most of the variance in the (a) Kier-Hall, (b) Ghose-Crippen, and (c) Isis Keys descriptor sets, after introducing a small amount of random noise in the PC coordinates of each compound.

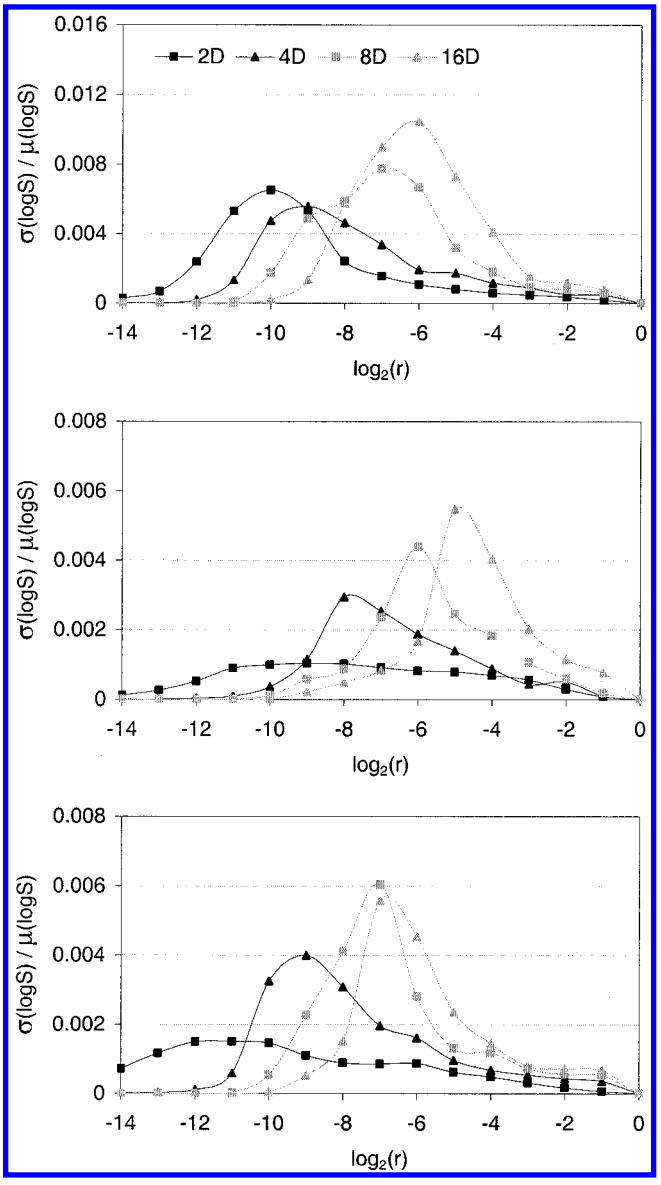

Figure 6. Plot of relative variance (*V*) vs the base 2 logarithm of the cell width (*r*) for 10 000 compounds randomly chosen from the ACD using the 2, 4, 8, and 16 principal components that accounted for most of the variance in the (a) Kier-Hall, (b) Ghose-Crippen, and (c) Isis Keys descriptor sets, after introducing a small amount of random noise in the PC coordinates of each compound.

fectively between two subsets of comparable cardinality. To determine the optimum bin size for a particular compound selection task, the user need only generate a sufficient number of random subsets from the collection of interest (where the size of the random subsets is the same as, or at least comparable to, the number of compounds selected), plot the relative variance of the diversity score at several grid resolutions, determine the resolution where this variance is at a maximum, and use that resolution to perform the actual selection. This procedure can be easily implemented as a preprocessing step prior to compound selection.

Finally, we should point out that rectangular grids are the simplest but not necessarily the best lattice models for diversity profiling. Channel coding theory describes efficient lattices in various dimensions, with fast decoding algorithms for determining which cell a particular point falls into given its coordinates. These methods can produce the same amount of space coverage with a fraction of the cells required by rectangular grids and could in theory lead to improved performance. Although our recursive algorithm could be easily tailored to work with nonrectangular lattices, rectangular ones do not require special decoding schemes and therefore are programmatically trivial to implement and can be crafted in a way that does not require explicit enumeration and storage of every cell in the grid.

IV. CONCLUSIONS

The results presented above suggest that cell-based diversity estimation is not as straightforward as one might expect from intuition. Identifying the most appropriate resolution depends critically on the nature of the descriptors, the dimensionality of the feature space, the distribution of the data set, and the size of the collection being analyzed.

Here, we presented a fast and programmatically simple approach for estimating an appropriate bin size for any given problem prior to the diversity profiling task. Traversing the entire 90 000-member amination library in 16 dimensions (vide supra) to 15 levels of recursion (i.e. to a cell width of 2^{-15}) required a mere tenth of a CPU second on an Dell Inspiron 7500 laptop computer equipped with a 733 MHz Intel Pentium III processor, which means that the method is applicable to most data sets encountered in compound acquisition and combinatorial library design.

ACKNOWLEDGMENT

The authors would like to thank Dr. Raymond F. Salemme for his insightful comments and support of this work.

REFERENCES AND NOTES

(1) 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, 1997; Vol. 10.
(2) 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, 1998; pp 742−761.
(3) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. *Mol. Divers.* **1999**, *4*(1), 1−22.
(4) 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, 2000; pp 265−300.
(5) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(6) Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(7) Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.
(8) Agrafiotis, D. K.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 51−58.
(9) Agrafiotis, D. K.; Rassokhin, D. N. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159−167.
(10) Pearlman, R. S. DiverseSolutions, University of Texas at Austin.
(11) Schroeder, M. *Fractals, Chaos, Power Laws*; W. H. Freeman and Company: New York, 1991.
(12) Belussi, A.; Faloutsos, C. In *21st International Conference On Very Large Databases (VLDB)*; Zurich, Switzerland, 1995.
(13) Traina, C., Jr.; Traina, A.; Wu, L.; Faloutsos, C. *Fast feature selection using the fractal dimension*; XV Brazilian Symposium on Databases (SBBD), Paraiba, Brazil, 2000.
(14) Copyright 3-Dimensional Pharmaceuticals, Inc., 1994−2001.
(15) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. United States Patents 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; and 5,901,069, 1999.
(16) Rassokhin, D. N.; Agrafiotis, D. K. *J. Mol. Graphics Mod.* **2000**, *18*, 370−384.
(17) Agrafiotis, D. K. *IBM J. Res. Develop.* **2001**, *45*, 545−566.
(18) MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.
(19) Hall, L. H.; Kier, L. B. The Molecular connectivity chi indexes and kappa shape indexes in structure−property relations. In *Reviews of Computational Chemistry*; Boyd, D. B., Lipkowitz, K. B., Eds.; VCH Publishers: 1991; Chapter 9, pp 367−422.
(20) Bonchev, D.; Trinajstić, N. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(21) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(22) Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. *J. Comput. Chem.* **2001**, *22*, 488−500.
(23) Lobanov, V. S.; Agrafiotis, D. K. *J. Comput. Chem.* **2001**, *22*, 1712−1722.
(24) Bellman, R. E. *Adaptive Control Processes*; Princeton University Press: Princeton, 1961.

CI010314L