One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation

Steven L. Dixon* and Kenneth M. Merz, Jr.‡

Accelrys, Box 5350, Princeton, New Jersey 08543

Received March 29, 2001

Drug discovery research is increasingly dedicated to biological screening on a massive scale, which seems to imply a basic rejection of many computer-assisted techniques originally designed to add rationality to the early stages of discovery. While ever-faster and more clever 3D methodologies continue to be developed and rejected as alternatives to indiscriminant screening, simpler tools based on 2D structure have carved a stable niche in the high-throughput paradigm of drug discovery. Their staying power is due in no small part to simplicity, ease of use, and demonstrated ability to explain structure−activity data. This observation led us to wonder whether an even simpler view of structure might offer an advantage over existing 2D and 3D methods. Accordingly, we introduce 1D representations of chemical structure, which are generated by collapsing a 3D molecular model or a 2D chemical graph onto a single coordinate of atomic positions. Atoms along this coordinate are differentiated according to elemental type, hybridization, and connectivity. By aligning 1D representations to match up identical atom types, a measure of overall structural similarity is afforded. In extensive structure−activity validation tests, 1D similarities consistently outperform both Daylight 2D fingerprints and *Cerius*^{2} pharmacophore fingerprints, suggesting that this new, simple means of representing and comparing structures may offer a significant advantage over existing tried-and-true methods.

Introduction

Modern drug discovery, once frequently reliant on computer-assisted methods of rational design, is increasingly identified as a somewhat irrational confluence of genomics, combinatorial chemistry, and high-throughput screening. Chemists and biologists have responded to the challenges posed by this intersection of disciplines, and experimental data are being accumulated at a rate that far exceeds the capacity of many computational methods once touted as viable means of accelerating the drug discovery process. For computer-assisted approaches aimed at lead discovery, emphasis has begun to shift somewhat away from areas such as de novo design and more toward developing ways to leverage huge internal warehouses of data. So-called *data mining* methods are in increasing demand, and the challenge is to unravel mysteries locked in mountains of biological screening data, thus identifying cost-effective alternatives to the pervasive practice of screening every compound against every target.

In many respects, data mining is a repackaging of tools and practices that have been around for years, and there is little doubt that this ostensibly abstract field still relies on the ability to search a chemical library (real or virtual) for compounds that either resemble known actives or satisfy some hypothesis derived therefrom. A primary issue of contention in this regard is the appropriate representation of chemical structure. Interactions between a drug and its receptor are most correctly modeled using three-dimensional (3D) information, and there are any number of powerful 3D searching techniques designed around this principle, including pharmacophore matching,[1−4] fast docking,[5,6] and structural similarity calculations based on molecular fields and surface properties.[7−11] Despite the growth and successes in these areas, simple 2D representations of structure continue to be used routinely by modelers and medicinal chemists alike. With the aid of in-place database software, only minimal time and effort is required to carry out a substructure search or to perform similarity calculations using 2D fragment descriptors.[12] Moreover, the ability of these simple approaches to explain structure−activity data is well established.[12−16] Thus in a climate where 3D methodologies are often viewed as holding the greatest promise, Occam's razor still governs many of the actual choices that are made.

Working on this premise of simplicity, the question arises as to whether a 1D view of chemical structure offers any advantage over existing 2D and 3D methods. More specifically, can a molecule be collapsed to one dimension, and can this representation be used to develop, for example, a superior measure of overall similarity? SMILES[17,18] strings and Sybyl Line Notation[19] are well-known ways of encoding molecular structure using 1D representations, but they are designed for compact storage of information rather than as tools for defining molecular similarity scales.

A more promising direction in which to proceed is suggested by pioneering work in the field of amino acid sequence alignment. Needleman and Wunsch[20] provided

* To whom correspondence should be addressed. Phone: (609) 452-3729. Fax: (609) 919-6155. E-mail: sdixon@accelrys.com.

‡ Current address: Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16802.

10.1021/jm010137f CCC: $20.00
Published on Web 10/13/2001

a practical means of measuring similarity between proteins based solely on the types and order of amino acids that appear along their backbones. While proteins are clearly 3D entities, the success of sequence-based analyses indicates that a great deal of valuable information can be encoded using simple 1D strings, and that sequence homology (i.e., 1D similarity) is sufficient to predict many complex behaviors that proteins share.

In translating the principles of protein sequence homology to small molecules, it is natural to substitute atoms for amino acids, but there is no clear counterpart to the protein backbone and hence the correct sequence of atoms. This issue is resolved through the use of multidimensional scaling,[21] which is a technique that may be employed to map the atoms of a structure onto a 1D coordinate such that the distances among atoms are preserved in an optimal sense. A molecule is thus represented by a set of atomic code strings (i.e., atom types) and their corresponding positions along a single coordinate axis. Molecular similarity is then obtained by formally sliding one set of atomic strings past another, until an alignment is found which provides maximal overlap of matching atom types.

Some parallels may be drawn between our 1D methodology and work carried out by Robinson and coworkers,[22,23] who use a nonlinear mapping technique to generate 2D pixel-based images of 3D structures. Digital image processing techniques are then used to rapidly align pairs of 2D structures to achieve maximal overlap of occupied pixels. Their method offers a significant computational advantage over full 3D similarity calculations, but continuous rotational freedom must still be addressed. One-dimensional similarity as defined here removes all rotational freedom, with the exception of a possible 180° phase difference. Moreover, atoms are differentiated according to type, so a rich array of information is encoded in the 1D representations.

Design issues aside, the most important aspect of developing any such methodology is the careful investigation of its potential to actually speed up the drug discovery process. In doing so, a variety of intuitive and relevant validation tests should be carried out, and a sufficient number of data sets should be examined to provide reasonable sampling of the classes of compounds and biological targets that are of current therapeutic interest. In the present paper, we examine numerous data sets encompassing a wide range of targets, and we consider several different approaches to validation. Tests are designed to demonstrate both the relevance of 1D representations and potential ways in which they can be used as tools to accelerate drug discovery.

One-Dimensional Representations

Points in high-dimensional space can be mapped to any lower number of dimensions through a technique known as multidimensional scaling (MDS).[21] There are various tools that fall under the MDS blanket, but all of them are designed to carry out this mapping while preserving distances among the points in some optimal fashion. In the present case, the points to be mapped are the atoms in a molecule, and the "high" dimensional representation could come from a 3D model, or it may simply be a 2D graph, where no real geometric information is supplied. At the outset we need only know the distances d_{ij} between each pair of atoms i,j. This distance corresponds either to the usual 3D Euclidean definition or, in the case of the 2D graph, to the number of bonds in the shortest path connecting atoms i and j.

Starting with some estimate x_i^{1D} of each 1D atomic coordinate, the corresponding 1D distances are defined:

$$d_{ij}^{1D} = |x_i^{1D} - x_j^{1D}| = \sqrt{(x_j^{1D} - x_j^{1D})^2} \tag{1}$$

The MDS procedure involves iterative adjustment of the 1D coordinates so that the distances d_{ij}^{1D} best approximate their 3D or 2D counterparts d_{ij} according to some goodness-of-fit measure E^{1D}. We use a simple sum-of-squared errors approach, which is equivalent to Kruskal's Stress:[24]

$$E^{1D} = \frac{\sum_{i>j=1}^{n} (d_{ij}^{1D} - d_{ij})^2}{\sum_{i>j=1}^{n} d_{ij}^2}; \quad n = \text{number of atoms} \tag{2}$$

A BFGS[25] procedure is used to minimize the error, which requires first derivatives of E^{1D} with respect to each 1D coordinate:

$$\frac{\partial E^{1D}}{\partial x_k^{1D}} = [2/\sum_{i>j=1}^{n} d_{ij}^2]\left[\sum_{i=1}^{n}\left(\frac{d_{ik}^{1D} - d_{ik}}{d_{ik}^{1D}}\right)(x_k^{1D} - x_i^{1D})\right] \tag{3}$$

Second derivatives are estimated implicitly as the minimization proceeds, and convergence to a minimum is usually achieved in *order*(n) optimization cycles.

As noted previously, one must have some starting set of 1D coordinates, and these are most readily obtained by projecting the 3D or 2D structure onto a "primary" axis of the molecule. We choose this axis to coincide with the dominant eigenvector of an appropriate Gram matrix **G** for the 3D or 2D structure. In the 3D case, explicit atomic coordinates (x_i, y_i, z_i) are known, and a 3×3 Gram matrix may be constructed:

$$\mathbf{G} = \sum_{i=1}^{n}\begin{bmatrix} (x_i - x_0)(x_i - x_0) & (x_i - x_0)(y_i - y_0) & (x_i - x_0)(z_i - z_0) \\ (y_i - y_0)(x_i - x_0) & (y_i - y_0)(y_i - y_0) & (y_i - y_0)(z_i - z_0) \\ (z_i - z_0)(x_i - x_0) & (z_i - z_0)(y_i - y_0) & (z_i - z_0)(z_i - z_0) \end{bmatrix} \tag{4}$$

Here, (x_0, y_0, z_0) is the centroid of the 3D points, so that x_0, y_0, and z_0 are average coordinate values among the n atoms along the corresponding Cartesian axes. This approach is nothing more than a principal components analysis (PCA) applied to three variables **x**, **y**, and **z**, with the n atoms representing the data points. The dominant eigenvector is thus the axis in 3D space along which the data exhibit maximum variance, i.e., maximum sum-of-squared distances from the centroid. We refer to it as a primary axis to distinguish it from the *principal axes* that one obtains from the moment of inertia tensor, which is a matrix that depends on both the coordinates and the masses of the atoms.

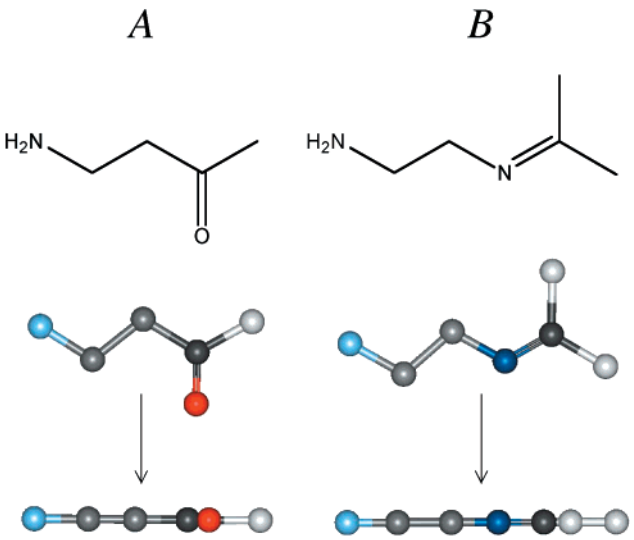


Figure 1. Three-dimensional structures and one-dimensional representations for two simple molecules. Atoms are color coded to differentiate elemental type, hybridization state, and degree of connectivity.

When only a 2D graph of the molecule is available, explicit 2D atomic coordinates are not provided and there is no corresponding 2 × 2 Gram matrix. Instead, a 2D → 1D embedding scheme borrowing ideas from the field of distance geometry[26,27] is used to construct an $n \times n$ matrix whose dominant eigenvector ultimately affords the projected 1D coordinates. This indirect approach, which is summarized in an appendix, requires only the 2D distances d_{ij} and not explicit 2D coordinates. It is, in fact, a generalized method of determining projected 1D coordinates, so that application of this embedding scheme to a set of 3D distances yields the same 1D coordinates that one would obtain from the 3D PCA approach above.

Once the primary axis and initial coordinates are defined, the BFGS optimization procedure makes relatively minor adjustments, typically decreasing the error in the 1D distances by about 20%. Note that the entire process is deterministic in nature, so a given 3D or 2D structure will give rise to a unique 1D representation.

Figure 1 shows two simple structures and their corresponding 1D representations generated by 3D → 1D projection onto the primary axis followed by BFGS optimization. In this example, only non-hydrogen atoms are treated and they are color-coded to distinguish different elements, hybridization states, and degrees of connectivity. When creating a database of 1D structures, these features are conveniently encoded by way of atomic character strings, for example, "C_3H2CH" indicates a carbon ("C_") with sp^3 hybridization ("3"), two attached hydrogens ("H2") and membership in a chain ("CH"), as opposed to a ring.

Molecule *A* contains 15 unique interatomic distances, while molecule *B* contains 21. The sets of 1D distances differ, in an RMS sense, from their target 3D distances by 0.443 and 0.461 Å, respectively. Note that distances are not reproduced well for atoms whose line of connection is orthogonal to the primary axis, for example, the C═O bond in molecule *A*. This inherent weakness in reduced dimensional representations becomes most pronounced for star-shaped and/or spherically shaped molecules, which are identified with excessive branching in the structure and conformational effects that cause the structure to fold back on itself. Robinson et al.[22] have addressed this issue in some detail, and they quantify the effect with a *spherosity* coefficient.

While it must be conceded that, to varying degrees, information is lost or distorted when a structure is mapped to a 1D coordinate, this does not imply that 1D representations are useful only for long, straight molecules. Frequently, the level of distortion provides important information about the overall shape of a molecule. For example, a "round" molecule (i.e., star-shaped or spherically shaped) will typically yield a 1D image wherein the atoms are densely distributed along a coordinate that spans roughly the diameter of the source structure. By contrast, a long, straight molecule with the same number and types of atoms will give rise to a far more extended 1D coordinate, so that the 1D images are as vastly different as their original structures.

Of course making comparisons *among* round molecules may pose difficulties because the direction of the primary axis can vary drastically from one molecule to the next. Mathematically speaking, a round molecule tends to yield a Gram matrix wherein two or more eigenvalues are very nearly the same size, so the eigenvector identified as the dominant one may change with only a slight change in the structure. This ambiguity can be resolved if the molecules being compared have some common (nonround) core that can be used to define a consistent Gram matrix whose dominant eigenvector is essentially the same for all of the molecules. Fortunately, most candidate molecules in the field of drug discovery tend to be more elongated than round, so no special treatment is required to assign the primary axis. Obviously, a given data set can be analyzed for its "roundness" to determine how faithfully the 1D representations reflect the original structures, but that information does not necessarily foretell how useful the 1D representations will be. Ultimately, extensive validation using biologically relevant data sets provides an indication as to whether 1D representations can be applied with success in the field of drug discovery.

One-Dimensional Similarity

In sequence alignment of proteins,[20] dynamic programming[28] is used to pair up identical amino acid residues, with interruptions in the chains, as necessary, to increase the number of matches. For 1D similarity of molecules, we use a related but somewhat different approach to match up atoms of the same type. Figure 2 illustrates the situation for the same two molecules used in Figure 1. Here, atoms are represented by rectangular cells of unit area centered on their respective 1D coordinate positions. The width of the cells indicates how faithfully the 1D representations reflect the original 3D or 2D structures. Across broad classes of molecules, we have observed that the RMS errors in 1D distances are almost always less than 1.0 Å (3D) or one bond unit (2D), so *dcell* is normally set equal to 1.0. To determine the 1D similarity, molecule *B* is aligned at various offsets with respect to molecule *A*, and the total overlap area S_{AB} between rectangular regions of the same type is computed. The largest overlap area that can be

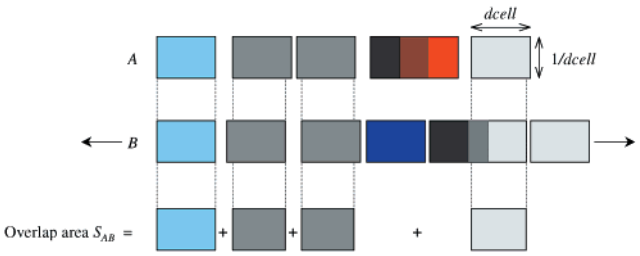


Figure 2. Calculation of the 1D overlap area between molecules *A* and *B* (from Figure 1). Each atom is represented by a rectangular region of unit area which is centered at the assigned 1D coordinate for that atom. In this example, the two molecules are aligned at their left-most atoms, and overlapping areas of the same type are tallied, giving a total overlap of slightly less than 4.0.

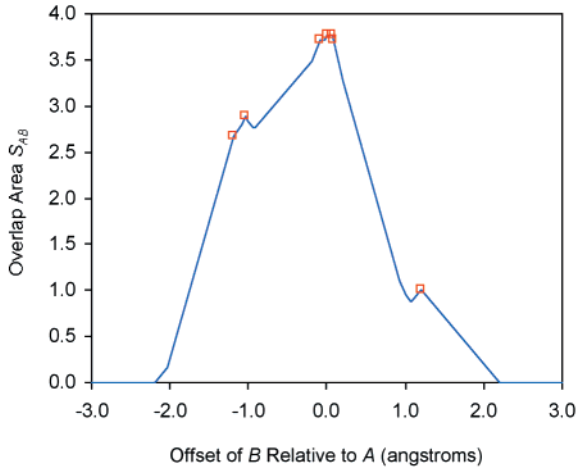


Figure 3. Variation of the 1D overlap area as molecules *A* and *B* (from Figures 1 and 2) are allowed to slide past each other in a continuous fashion. Red squares mark points where two atoms of the same type come into alignment.

achieved, S_{AB}^{max}, is combined with normalization factors from aligning each molecule with itself to define a 1D similarity function on the interval [0,1]:

$$Sim_{AB} = \frac{S_{AB}^{max}}{\sqrt{S_{AA}^{max} S_{BB}^{max}}} \quad (5)$$

Finding the optimal alignment is of course at the heart of the similarity calculation, and Figure 3 illustrates how the overlap area changes as the 1D representations of molecules *A* and *B* are allowed to slide past each other in a continuous fashion. Because the objects being overlapped are rectangles, the curve is simply a collection of line segments, with a change in slope occurring each time two atoms of the same type *begin* to overlap, *reach a peak* in their overlap, or *cease* to overlap. Red boxes mark the instances wherein two atoms of the same type reach their peak overlap, and it is only at these points that the curve can exhibit a negative change in slope. Since every local maximum on the curve is followed immediately by a negative change in slope, we are guaranteed that the global maximum will occur when some pair of atoms of the same type is perfectly aligned. Thus overlap areas need only be computed at these alignment points in order to determine S_{AB}^{max}. Because there could be a phase difference of 180° in the 1D coordinate systems, a second alignment procedure is carried out with molecule *B* "flipped" by 180° relative to its original orientation. This requires nothing more than algebraic negation of the 1D coordinates, followed by realignment on the same pairs of atoms.

For typical drug-like molecules containing 30−40 non-hydrogen atoms, a single 1D similarity calculation as described above requires 0.004−0.006 s of CPU time on a modern Unix workstation (SGI Origin 2000, 270 MHz R12000). It is possible to speed up the calculation by a factor of 2 or more by employing a rapid preprocessing step which automatically eliminates alignments that give rise to comparatively small overlap areas.

As shown in Figure 4a, 1D representations for molecules *A* and *B* are positioned at a small number of *template-based* offsets which are multiples of the cell width parameter. Starting at a given template offset, all matching atom types within half a cell width of each other are identified, Figure 4b. Aligning *A* and *B* on all of these pairs and doing full overlap calculations would offer no time savings, so instead, *partial overlap areas* are computed independently for each template cell of *B*. These small-scale calculations focus on just the atoms of a particular template cell, and they require mini-overlap calculations at the extreme right and left movements (±*dcell*/2) and at all intermediate positions which align atoms of the same type. By considering all of these alignments, a maximum partial overlap for each template cell may be found. Because this process involves independent movement of the individual template cells rather than rigid movement of the entire molecule, the sum of the maximum partial overlaps defines a strict upper bound for the full *A*,*B* overlap. This upper bound applies to all possible rigid alignments of *A* and *B* that are within ±*dcell*/2 of the associated template offset. Molecule *B* is then shifted to the next template offset, and the procedure is repeated.

After upper bounds for all necessary template offsets are determined, the corresponding areas are sorted from high to low, and full overlap calculations are initiated in an order that follows the sorted upper bounds. Thus starting at the template offset with the highest upper bound, full overlap areas are computed for *A* and *B*, at all atom-based alignments that require molecule *B* to be shifted by no more than ±*dcell*/2. During this process, the largest full overlap encountered is always stored. After these atom-based alignments are completed, the template offset with the second highest upper bound is considered, and so forth. At some point, the remaining upper bounds will be lower than the largest full overlap previously encountered. When this occurs, all subsequent template-based offsets and their associated atom-based alignments may be safely ignored.

By eliminating unnecessary alignments, similarity calculations on molecules containing 30−40 non-hydrogen atoms require only about 0.002−0.003 CPU seconds, which represents a 2-fold reduction in the computation time. For larger compounds (>40 non-hydrogen atoms), the speedup is more pronounced, often greater than a factor of 4. While this modest improvement may not be fully appreciated for small data sets, it is certainly noticeable when computing similarities across combinatorial libraries containing tens of thousands of com-

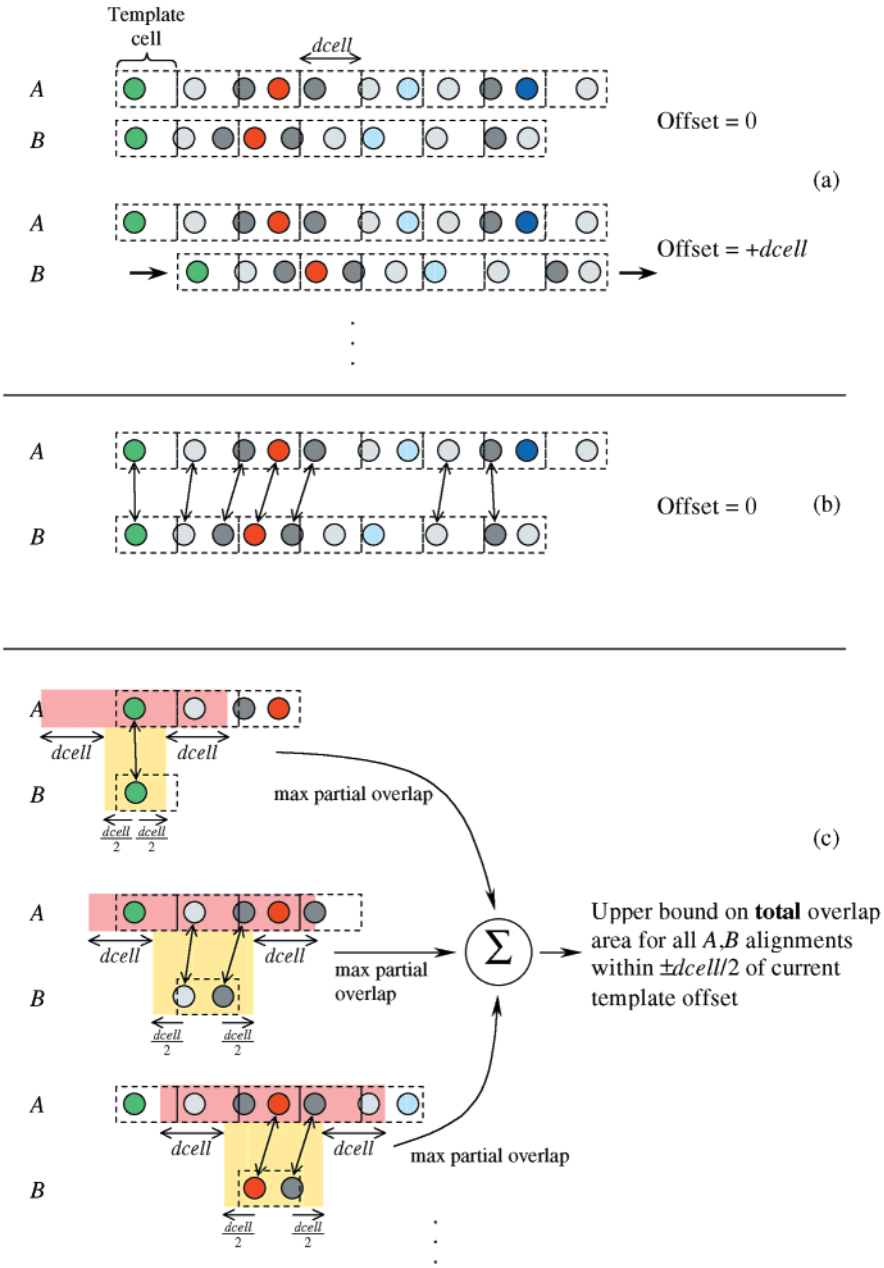


Figure 4. Template-based scheme used to eliminate unnecessary atom-based alignments. (a) A regular grid of template cells is locked onto each 1D representation, and molecule *B* is positioned at a series of offsets (0, ±*dcell*, ±2·*dcell*, ...) relative to molecule *A*. (b) At a given template offset, all atoms of the same type separated by no more than *dcell*/2 are identified. (c) Starting at the same template offset, partial overlap areas are computed for each template cell of *B*. These are calculated as the cell is positioned at the extreme right and left limits of movement (±*dcell*/2) and at all intermediate positions which align matching atom types. Yellow shaded regions indicate the range of movement of each template cell, and the pink-shaded regions identify the atoms of molecule *A* which have the potential to contribute to the partial overlaps. The largest partial overlap areas found for the respective template cells of *B* are summed to arrive at an upper bound on the total *A*,*B* overlap area. This upper bound applies to all atom-based alignments that are within ±*dcell*/2 of the current template offset.

pounds, many of which are larger than 40 non-hydrogen atoms.

Conformational Effects

When 1D similarities are derived from 3D structure, there will of course be conformational effects and, therefore, some level of noise introduced into the problem. The degree to which 1D similarities can vary depends on the flexibility of a molecule and the extent to which this flexibility changes both the direction of the primary axis and the distribution of atoms along the axis. The mere fact that 1D similarities vary with conformation does not necessarily constitute a defect in the method, but it is important to know how much 3D variation can be tolerated and whether a given similarity value exceeds background noise.

Figure 5 illustrates how conformational changes affect 1D similarities for some molecules that appear in recent *J. Med. Chem.* publications.[29a−d] The *Catalyst* program[30] was used to generate 50 low-energy conformations for each molecule, with ranges in energy up to 20 kcal. For a given molecule, 1D similarities between all pairs of conformers were computed, along with RMS differences in their 3D coordinates. Each set of 1D similarities was

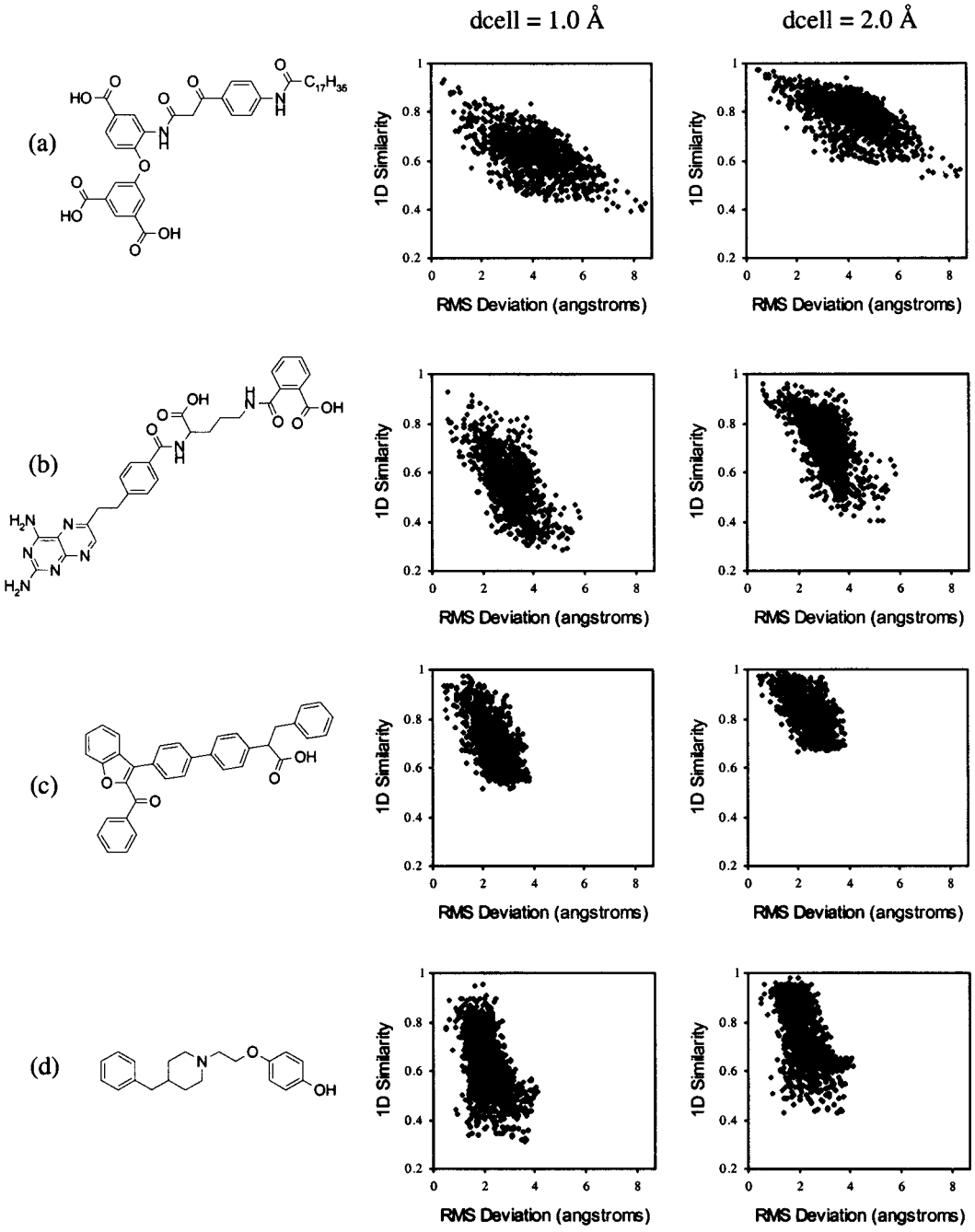


Figure 5. Effects of conformation on 1D similarity. The *Catalyst* program was used to generate 50 conformations for each molecule, and 1D similarities were calculated between all pairs of conformers, using two different values of the cell width parameter. Each 1D similarity is plotted against the corresponding RMS deviation in the atomic coordinates for that pair of conformers.

calculated using two different values of the cell width parameter to demonstrate the effect of "smearing out" atomic distributions along the 1D coordinate. Note that this approach is for illustration purposes only, and it is not intended to capture all of the information that would be obtained from a full conformational analysis.

Not surprisingly, there is a general trend toward lower 1D similarities as the 3D structures become increasingly dissimilar in an RMS sense, but the rate at which similarities drop depends on the overall size of the structure. Larger, elongated molecules tend to show a more gradual drop, and this is primarily because 1D similarities reflect the *relative* impact of conformational changes on overall structure. In other words, an RMS deviation of 2.0 Å is not as significant a change for a molecule that is 20 Å long as it is for a molecule that is only 10 Å long. One might infer from this that 3D models for small molecules need to be generated more carefully, but in going from the relatively large molecule (Figure 5a) to the small molecule (Figure 5d), there does not appear to be a consistent trend toward lower 1D similarities among all conformations generated.

Comparing results for different values of *dcell*, it is apparent that an increase in this parameter leads to a shift toward higher 1D similarities, and this is simply because wider cells exhibit nonzero overlap at greater interatomic offsets. In effect, then, smearing out the atoms decreases the sensitivity to conformational changes. At the same time, however, a poorer signal-

to-noise ratio is created because 1D similarities are less sensitive to genuine structural differences between *distinct* compounds.

While we employ single conformer models throughout the rest of this paper, there are alternative approaches that directly address the conformational issue. For example, one might generate a set of conformers for each molecule, as above, then compute all pairs of similarities between conformers from different molecules. An overall 1D similarity could then be defined as the maximum similarity observed between any two conformers, or perhaps as the average similarity observed by matching up each conformer from one molecule with the most similar conformer from the other molecule. Obviously, this sort of approach could become unwieldy for extremely large collections of compounds, unless the number of conformers generated is sufficiently small.

Validation Studies

Biological Activity. Throughout the course of the validation tests, all activity data that was originally reported on a concentration scale was converted to a negative base 10 logarithmic scale, e.g., $K_i \rightarrow -\log(K_i) \equiv pK_i$, where K_i is expressed in units of mol/L. Thus an increase of one unit on the pK_i activity scale corresponds to a 10-fold reduction in the K_i value.

Molecular Descriptors. One-dimensional representations were generated for non-hydrogen atoms only, starting with either 3D coordinates or 2D topological distances. These two procedures are henceforth referred to as 3D → 1D and 2D → 1D, respectively. Unless otherwise noted (see steroids below), initial 3D coordinates were obtained by using the default energy minimization procedure in *Catalyst*. The cell width parameter was set to 1.0 for all 1D similarity calculations, and intercompound distances were calculated as 1 − similarity.

For comparison, parallel sets of validation tests were run using both Daylight 2D fingerprints[18] and 3D pharmacophore fingerprints from the *Cerius*^{2} program.[31] Two-dimensional fingerprints were generated using standard Daylight protocol, i.e., paths of length 0−7 were considered in the initial fragmentation phase, and fingerprints for a given data set were folded in half until the average number of "on" bits was at least 30% of the total bits. Three-dimensional pharmacophore fingerprints were created using *Catalyst* conformational databases containing up to 100 conformers per molecule. Pharmacophores were defined using the complete set of *Cerius*^{2} features: negative charge, positive charge, negative ionizable, positive ionizable, H-bond acceptor, H-bond donor, aromatic ring, and hydrophobic group. Feature distances were binned at a resolution of 2.0 Å with a minimum separation of 1.0 Å. While *Cerius*^{2} has the ability to generate fingerprints from either three-point or four-point pharmacophores, we examined only the three-point variety, primarily because four-point fingerprints are far more expensive to calculate and store and therefore do not provide a practical means of searching potentially large chemical libraries. Tanimoto coefficient (TC)[32] was used as the similarity measure for both the Daylight and Cerius^2 fingerprints, and intercompound distances were calculated as 1-TC.

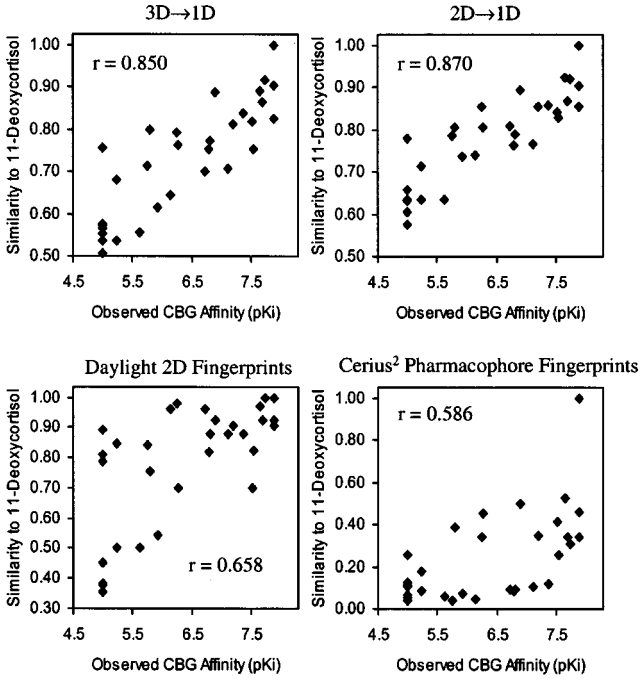


Figure 6. Correlation of similarity with binding affinity to corticosteroid binding globulin (CBG). The similarity of each compound to the tightest CBG binder (11-deoxycortisol) is plotted against the compound's own affinity to CBG.

Steroid Activity Correlation. Since the introduction of CoMFA,[33] the steroid data set has become a standard means of testing new methodologies, especially those which incorporate 3D information. Conformational freedom is very restricted in these molecules, so the most troublesome distraction in 3D calculations is largely removed. The lack of such complicating factors should, however, temper enthusiasm derived from obtaining a strong steroid QSAR, and it is probably wise to limit the complexity of validation tests that involve these compounds. Accordingly, we simply examine the correlation between the measured affinity to corticosteroid binding globulin (CBG) and the similarity of each compound to 11-deoxycortisol, the strongest binder.

Structures and activity data for the augmented set of 31 steroids were obtained directly from the Gasteiger Web site,[34] and the supplied structures were submitted to each descriptor generation protocol, as described previously. No further modeling of the 3D coordinates was done before generating 1D representations, but *Catalyst* conformers were still generated for the 3D pharmacophore fingerprints.

Figure 6 summarizes the similarity−activity correlations for each method of measuring similarity, and it is apparent that the 1D similarity scales are most strongly related to activity. In both the 3D → 1D and 2D → 1D cases, the correlation coefficient is at least 0.85, and this compares quite well with results from high-level parametric models obtained by fitting to the CBG activity data.[33,35−38] Note that the strength of the correlation weakens as the apparent complexity and/or dimensionality of the molecular representation increases: r(2D → 1D) > r(3D → 1D) > r(2D fingerprints) > r(pharmacophore fingerprints). This ranking implies that more advanced methods of comparing structures may offer no significant advantage in the case of the steroids, possibly because of their high rigidity and homology.

Table 1. Summary of Results from Neighborhood Validation Studies[a]

		3D → 1D			2D → 1D			Daylight 2D FPs			Cerius[2] pharmacophore FPs			Unity 2D FPs[c]	
data set[b]	ncmpd	D_{LRT}/D_{tot}	χ^2	χ^2_d	D_{LRT}/D_{tot}	χ^2	χ^2_d	D_{LRT}/D_{tot}	χ^2	χ^2_d	D_{LRT}/D_{tot}	χ^2	χ^2_d	D_{LRT}/D_{tot}	χ^2
1. Uehling	9	1.59	7.80	0.421	1.70	9.59	0.317	1.83	12.50	0.296	1.89	14.22	0.023	1.55	6.22
2. Strupczewski	34	1.51	94.94	0.189	1.38	54.79	0.136	1.41	62.95	0.294	1.93	246.29	0.000	1.41	59.61
4. Garratt1	10	1.66	11.94	0.967	1.22	1.56	0.468	1.35	3.66	0.406	1.33	3.63	0.177	1.07	0.19
5. Garratt2	14	1.56	14.81	0.074	1.43	10.87	1.103	1.46	10.48	0.098	1.80	30.20	0.128	1.08	0.50
6. Heyl	11	1.08	0.26	0.000	1.02	0.01	0.001	1.52	7.74	0.002	1.50	9.20	0.064	1.01	0.00
7. Cristalli	32	1.37	45.80	2.107	1.31	31.07	1.476	1.48	69.99	1.145	1.74	147.46	0.130	1.31	30.27
8. Stevenson	5	1.40	0.80	0.000	1.33	0.73	0.003	1.19	0.24	0.039	1.93	4.47	0.106	1.07	0.04
9. Doherty	6	1.36	1.45	0.176	1.51	2.60	0.744	1.06	0.06	0.030	1.67	3.75	0.051	1.06	0.04
10. Penning	13	1.14	1.27	0.014	1.06	0.22	0.000	1.64	16.56	0.000	1.88	31.10	0.046	1.53	12.73
11. Lewis	7	1.44	2.67	0.241	1.40	2.18	0.197	1.32	1.59	0.070	1.51	3.50	0.043	1.01	0.00
12. Krystek	30	1.42	52.73	0.272	1.51	69.36	0.062	1.17	8.85	0.000	1.82	167.51	0.129	1.23	16.31
13. Yokoyama1	13	1.16	1.66	0.077	1.14	1.40	0.141	1.38	7.11	0.065	1.81	26.59	0.136	1.01	0.00
14. Yokoyama2	12	1.25	3.13	0.164	1.13	0.85	0.008	1.73	17.45	0.283	1.84	23.94	0.052	1.70	16.03
15. Svensson	13	1.27	3.76	0.000	1.18	1.94	0.000	1.24	2.86	0.165	1.62	14.77	0.000	1.02	0.02
16. Tsutsumi	13	1.17	1.81	0.164	1.06	0.29	0.002	1.40	7.02	0.026	1.90	32.89	0.462	1.58	14.35
17. Chang	34	1.47	80.38	0.430	1.43	65.65	0.116	0.99	0.06	0.001	1.62	128.77	0.040	1.13	8.36
18. Rosowsky	10	1.75	13.32	1.923	1.56	8.26	1.128	1.48	6.56	0.747	1.49	6.69	0.056	1.01	0.00
19. Thompson	8	1.39	2.63	0.024	1.24	1.14	0.054	1.07	0.12	0.160	1.68	6.84	0.130	1.17	0.68
20. Depreux	26	1.67	79.59	4.379	1.28	16.88	1.476	1.16	5.87	0.201	1.53	56.17	0.741	1.21	8.61
averages		1.40	22.14	0.612	1.31	14.70	0.391	1.36	12.72	0.212	1.71	49.89	0.132	1.22	9.16

[a] D_{LRT}/D_{tot} = (LRT point density)/(total point density); χ^2 is the statistical test defined in ref 39; χ^2_d is the corrected value defined in this paper. [b] See ref 39 for complete citations and further details about the data sets. Set 3 (Siddiqi) was omitted because activity data was of a mixed nature. [c] Results as reported in ref 39. χ^2_d values were not computed because Unity fingerprint distances were not available.

Neighborhood Behavior. The second set of validation tests employed Patterson's "neighborhood behavior" method,[39] which was introduced as a means of evaluating molecular diversity descriptors. This approach is based on the notion that a descriptor should not measure two structures to be highly similar when there is a large difference in their biological activity values with respect to a given target. To determine whether and to what degree a descriptor obeys this principle, the absolute differences $|\Delta act_{ij}|$ in biological activity are plotted against the corresponding distances d_{ij} in descriptor space for all unique pairs of compounds i,j in a data set. Each distance is used to define a neighborhood, which corresponds to a triangular region on the plot with vertexes (0,0), (d_{ij},0), (d_{ij}, max $|\Delta act_{ij}|$), where max $|\Delta act_{ij}|$ is the largest absolute activity difference observed among all pairs of compounds. The neighborhood containing the highest density of points (number of points per area) is combined with the rectangular region to the right of it, creating a lower right trapezoid (LRT) which, for a valid descriptor, will contain a higher density of points than the remaining upper left triangular (ULT) region of the plot. Overall validity is characterized by the ratio of the point density in the LRT to the average point density across the full rectangular region (LRT ∪ ULT). Maximum validity is observed when all of the points lie in the lower right *triangular* half of the plot, in which case the ratio is 2.0.

A χ^2 value is computed to measure the statistical significance of the buildup of point density in the LRT:

$$\chi^2 = \frac{(N_{LRT} - n_{LRT})^2}{n_{LRT}} \tag{6}$$

Here, N_{LRT} is the number of points actually observed in the LRT, and n_{LRT} is the number that would be expected if all of the points were distributed randomly and uniformly across the entire rectangular region. This random expectation value can be calculated from the ratio of the LRT area to the total area:

$$n_{LRT} = \frac{A_{LRT}}{A_{LRT} + A_{ULT}} N_{tot} \tag{7}$$

where N_{tot} is the total number of points in the plot.

Table 1 summarizes results for 19 of the 20 data sets analyzed by Patterson et al. Set 3 (Siddiqi) was omitted because of the presence of mixed activity data.[40] We have included in Table 1 results reported by Patterson et al. for Unity 2D fingerprints, which exhibited the strongest neighborhood behavior of any whole-molecule descriptor they analyzed. A side-chain-only version of the Unity fingerprints did perform better, as did Cramer's topomeric descriptors,[10] but these approaches analyze only the portion of the molecule that varies, and they are restricted to data sets where changes occur only within a single R group. Such descriptors are not readily amenable to comparisons between arbitrarily selected structures and are therefore difficult to implement in broad applications.

Before examining Table 1 in its entirety, we pause to note that some caution should be exercised when interpreting results from this validation technique. Figure 7 contains neighborhood plots for data set number 2 (Strupczewski), which is one of the two largest collections tested by Patterson. The optimal neighborhood line is drawn in each plot, and the corresponding point density ratio and χ^2 value are reported. Patterson's ideal plot would show a "fanning out" of points under the ULT/LRT bisecting line, so that as one moves to the right in the plot, there would be a distinct pattern of progressively wider ranges in the biological activity differences. From a visual inspection, none of the four

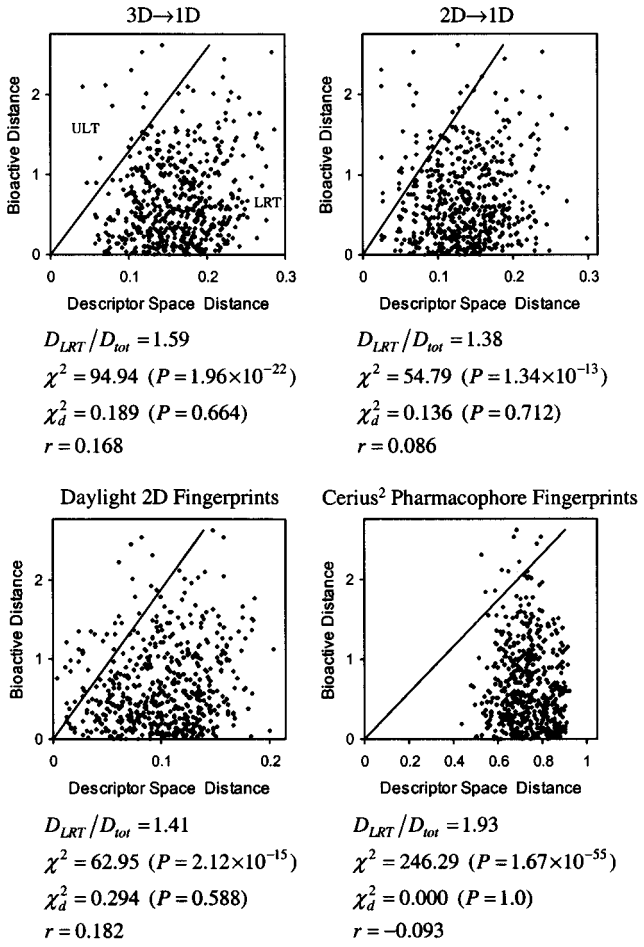


Figure 7. Neighborhood behavior plots for data set 2 (Strupczewski). Patterson's original validity measures D_{LRT}/D_{tot} and χ^2 indicate that all four sets of descriptors have moderate to strong neighborhood behavior. However, the distribution-corrected statistic χ^2_d implies much weaker neighborhood behavior, and the validity rankings based on this measure follow the same order as the correlation r between absolute activity differences and descriptor space distances.

plots in Figure 7 displays a marked tendency to fan out in this manner, yet the point density ratios and χ^2 values indicate moderate to strong neighborhood behavior, with pharmacophore fingerprints yielding what would appear to be staggeringly high validity. To varying degrees, these validation results are clearly at odds with common sense.

Pharmacophore fingerprints, which we believe to produce the most misleading results, exhibit a narrow range of unusually large distances, and this makes the distribution of points intrinsically more likely to be located in *any* LRT that can be defined by drawing a line from the origin to some position at the top of the plot. The exaggerated behavior associated with these descriptors may not have been anticipated nor observed by Patterson,[39] but even the 1D representations and 2D fingerprints receive undeservedly high validity ratings that are directly attributable to statistical properties of their distance distributions, irrespective of any relationship to activity. While Patterson noted that some spurious neighborhood enhancements could occur, no indication was given as to what extent this might be occurring in general. We have found that in just about every case the density ratio and the χ^2 statistic overestimate any genuine neighborhood enhancement and that certain descriptors with no neighborhood validity can in fact be measured to have extremely high validity.

One way to detect and correct for this phenomenon is to employ an adjusted χ^2 value which properly accounts for the width and center of each distance distribution. Rather than assuming a uniform distribution of points when computing n_{LRT} (i.e., eq 7), one can simply scramble the activity values among the data set members, regenerate the $|\Delta act_{ij}|$ vs d_{ij} plot, then count the number of scrambled points that fall in the original LRT.[41] In this way, n_{LRT} automatically accounts for the statistical properties of the distance distribution, yielding the correct null hypothesis for the χ^2 test. For each descriptor and data set, we have run this parallel experiment 100 times to arrive at an average value for n_{LRT}, which can then be inserted into eq 6 to yield a *distribution corrected* statistic, χ^2_d.

Referring again to Figure 7, we see that the χ^2_d values tell quite a different tale from the other measures of neighborhood validity. First of all, this statistical measure is much smaller than its uncorrected analogue χ^2, indicating a more conservative (and realistic) estimate of the significance of the LRT point densities. The probabilities P that the χ^2_d values could have occurred by random chance are also much higher than those associated with χ^2. Note that pharmacophore fingerprints are found to have absolutely no neighborhood validity with respect to this data set, a conclusion that is consistent with the narrow, right-skewed distribution of points observed for this descriptor. Daylight fingerprints exhibit the highest χ^2_d significance, and the corresponding plot does appear to show the most pronounced fanning out of points, though the pattern is not particularly strong.

To further support the conclusions implied by χ^2_d, we also include in Figure 7 the correlation coefficient r between the absolute activity differences and the descriptor distances. Though this does not measure neighborhood behavior in the same way, it does indicate whether there is some relationship between $|\Delta act_{ij}|$ and d_{ij}. Note that the correlations are weak in every case, but they follow the exact same order as χ^2_d, with pharmacophore fingerprints even exhibiting a slightly negative correlation.

For comparison purposes, neighborhood plots with more ideal behavior are shown in Figure 8. Data set number 7 (Cristalli) yields χ^2_d values of greater than 1.0 for every descriptor except pharmacophore fingerprints. As before, these corrected statistics are much smaller than their χ^2 counterparts, but they are among the highest χ^2_d values observed for the neighborhood data sets. Correlation coefficients again follow the same order as the χ^2_d statistic, and the strengths of the correlations are correspondingly higher than in Figure 7. From a visual perspective, the neighborhood plots do reveal a more pronounced fanning out of the points, which is perhaps more evident upon direct comparison to the almost Gaussian-shaped patterns of Figure 7.

Armed with a more robust measure of neighborhood validity, we now return to Table 1 and examine the entire collection of 19 data sets. According to the density ratios and χ^2 values, pharmacophore fingerprints show,

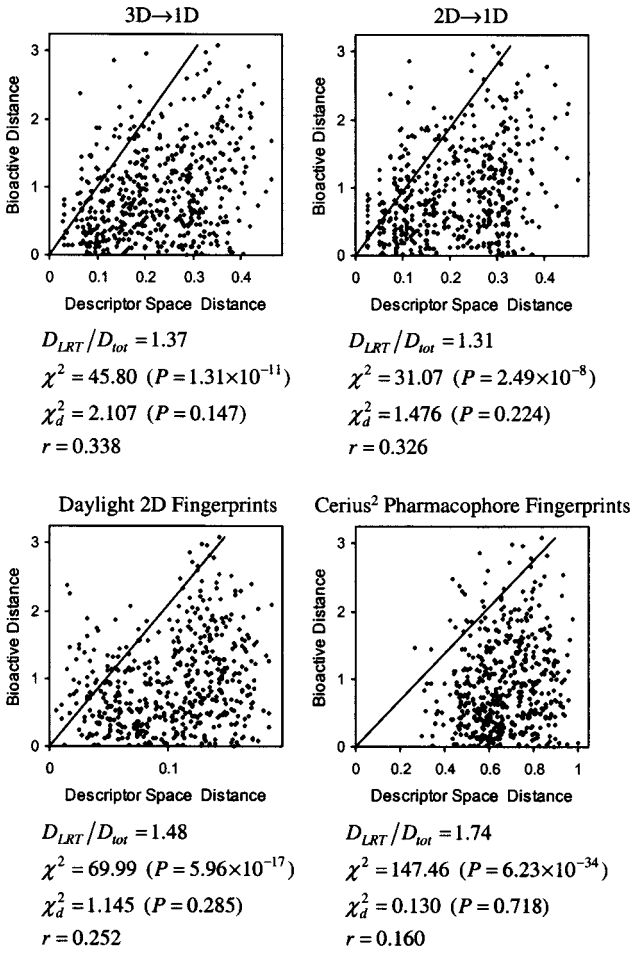


Figure 8. Neighborhood behavior plots for data set 7 (Cristalli). The χ^2_d statistic, while still much smaller than χ^2, indicates significantly stronger neighborhood behavior than observed previously for data set 2. Correlation coefficients r once again rank validity in the same order as χ^2_d, and their values are also higher than observed for data set 2.

on average, the strongest neighborhood behavior. However, as observed in previous examples, the corrected statistic χ^2_d indicates that these descriptors are the least valid. If we discount pharmacophore fingerprints, we see that the 3D → 1D distances yield the most pronounced neighborhood behavior by all measures, and 2D → 1D distances rank second according to both χ^2 and χ^2_d. Daylight fingerprints appear to be more valid than Unity fingerprints, but it is difficult to say with certainty since we were unable to determine the χ^2_d statistic for the latter.

Aside from the reversal of the pharmacophore fingerprint validity, the overall conclusions based on χ^2_d generally agree with those derived from Patterson's original measures. Note, however, that the individual performances generally do not indicate high levels of statistical significance; only the 3D → 1D method applied to data set 20 produced a χ^2_d value greater than 3.84, which corresponds to the 95% confidence level. Nevertheless, it must be remembered that even a moderate level of statistical significance still implies an advantage over random screening, where the discovery stage hit rate is often 1/10000 or lower.

***k*-Nearest Neighbors Activity Prediction.** *k*-Nearest neighbors[42] (kNN) is a pattern recognition technique wherein a prediction is made for an unknown object based on information from the *k* objects in a training set that are most similar to the unknown. In QSAR-related applications, kNN frequently involves predicting the class or category into which a compound is most likely to fall, for example, active versus inactive. For purposes of descriptor validation, however, this binary division can be too subjective, and we prefer to use a kNN design that involves a continuous activity scale. In this case, the most straightforward approach is to predict the activity of each compound to be the average activity observed for its *k* nearest neighbors. Zheng and Tropsha[43] have used this technique to select variables for QSAR analysis, and we adopt it here as a means of descriptor validation.

An optimal value of *k* is determined for each data set/descriptor combination based on jack-knifed activity predictions. In the present set of tests, the value of *k* which yields the highest correlation *r* between predicted and observed activities is selected. One could just as well use the RMS error in the predictions, or the coefficient of determination R^2, but these measures can penalize predictions which accurately reflect the rank ordering of activities, unless the predicted activity range corresponds to the observed range. Using a mean approximation necessarily contracts the range of predicted activity values, so RMS and R^2 are less suitable than the simple correlation *r*. Note that while Zheng and Tropsha[43] restrict the optimum value of *k* to lie between 1 and 5, we place no such limits on *k* in order to avoid introducing any possible external bias into the validation process. Allowing *k* to become larger than five generally did not lead to any significant improvements in correlation coefficients, but at the same time, no method was penalized for requiring larger values of *k* in order to arrive at optimal predictions.

Table 2 summarizes the 10 data sets used in the kNN validation studies. The first five sets were chosen from among recent issues of *J. Med. Chem.*, with the following criteria: (1) multiple activity determinations with high precision (<15% variation on average); (2) at least 3 orders of magnitude range in the activity values; and (3) a large number of compounds by *J. Med. Chem.* standards. The remaining compounds in Table 2 correspond to the five largest data sets analyzed in the neighborhood validation studies. There is some similarity between the endothelin antagonists of sets 5 and 8, but the former collection is much larger and considerably more diverse structurally. Altogether, the 10 data sets provide reasonably broad coverage of both compound and target space, with data derived from a variety of experiments, including radioactive binding assays, in vivo measurement of antipsychotic behavior, and cytotoxicity to cancer cells.

Table 3 summarizes the kNN validation results. There is a fairly general consensus among the various descriptors as to which data sets are easiest to model and which are most difficult. Not surprisingly, the performance for set 6 (Strupczewski) was poor across the board, and this result was also seen in the neighborhood validation studies when the χ^2_d statistic was examined. The underlying SAR for this in vivo data appears to be very elusive indeed. For most data sets, the observed correlation coefficients are not particularly

Table 2. Summary of Data Sets Used in kNN Validation Studies

data set[a]	ncmpd	compound class	activity profile
1. Huang	37	β-carbolines	$GABA_A$ α1 receptor affinity
2. Vacher	68	6-substituted-2-pyridinylmethylamines	serotonin 5-HT_{1A} antagonism
3. Gamage	41	bis(acridine-4-carboxamides)	topoisomerase I/II inhibition
4. Deutsch	40	2-(aminoethyl)-3-phenylbicyclo alkanes	dopamine transport complex affinity
5. Murugesan	73	biarylsulfonamides	endothelin ET_A antagonism
6. Strupczewski	34	3-[[(aryloxy)alkyl]piperidinyl]-1,2-benzisoxazoles	in vivo antipsychotic D_2/5-HT_2 antagonism
7. Cristalli	32	adenosines	adenosine A_{2a} receptor affinity
8. Krystek	30	arylsulfonamides	endothelin ET_A antagonism
9. Chang	34	1,2,4-triazolinone biphenylsulfonamides	angiotensin II AT_1 receptor affinity
10. Depreux	26	*N*-naphthylethyl amides	melatonin receptor affinity

[a] References and notations for data sets: 1. Huang, Q.; He, X.; Ma, C.; Liu, R.; Yu, S.; Dayer, C. A.; Wenger, G. R. *J. Med. Chem.* **2000**, *43*, 71−95. Compounds from Table 6, excluding **70** and **73** due to identical structures but different affinities, and excluding **96**−**98**, **102**, **104**, **105** due to specification of only a bound on K_i. Activity = pK_i for α1 affinity. 2. Vacher, B.; Bonnaud, B.; Funes, P.; Jubault, N.; Koek, W.; Assié, M.; Cosi, C. *J. Med. Chem.* **1998**, *41*, 5070−5083. Compounds from Tables 2−5, excluding duplicate entries, 8-OH-DPAT, and excluding **21** and **75** due to bounded K_i. Activity = pK_i for 5-HT_{1A}. 3. Gamage, S. A.; Spicer, J. A.; Atwell, G. J.; Finlay, G. J.; Baguley, B. C.; Denny, W. A. *J. Med. Chem.* **1999**, *42*, 2383−2393. Compounds from Table 1, excluding **9a**, amsacrine, and doxorubicin. Activity = pIC_{50} for Murine Lewis Lung carcinoma (LL). 4. Deutsch, H. M.; Collard, D. M.; Zhang, L.; Burnham, K. S.; Deshpande, A. K.; Holtzman, S. G.; Schweri, M. M. *J. Med. Chem.* **1999**, *42*, 882−895. Compounds from Table 1, excluding (−)-cocaine and WIN 35,428. Activity = pIC_{50} for inhibition of [^{3}H]WIN 35,428 binding. 5. Murugesan, N.; et al. *J. Med. Chem.* **1998**, *41*, 5198−5218. Compounds from Tables 2−6, excluding duplicate entries, BQ-123, BMS 182874, and excluding **40a** due to bounded K_i. Activity = pK_i for ET_A. 6. Strupczewski, J. T.; et al. *J. Med. Chem.* **1995**, *38*, 1119−1131. Compounds from Tables 2 and 3 with R_1 = 6-F, n = 3, X = O, and excluding **56** and **57** due to unspecified chirality. Activity = pED_{50} for inhibition of apomorphine-induced climbing (ED_{50} converted from mg/kg to mol/kg). 7. Cristalli, G.; et al. *J. Med. Chem.* **1995**, *38*, 1462−1472. Compounds from Table 1, excluding CGS 21680, NECA, CCPA, and compound **20**. Activity = pK_i for binding to rat striatum A_{2a}. 8. Krystek, S. R.; Hunt, J. T.; Stein, P. D.; Stouch, T. R. *J. Med. Chem.* **1995**, *38*, 659−668. Compounds from Table 1, excluding **16**, **17**, and **33**−**36**. Activity = pIC_{50} for ET_A. 9. Chang, L. L.; et al. *J. Med. Chem.* **1994**, *37*, 4464−4478. Compounds from Table 1 with R^3 = (2-Cl)C_6H_4. Activity = pIC_{50} for AT_1 (rabbit aorta). 10. Depreux, P.; et al. *J. Med. Chem.* **1994**, *37*, 3231−3239. Compounds from Table 1 with R_1 = 7-OCH_3, R_2 = H, x=1, R_3 = H. Activity = pK_D.

Table 3. Summary of Results from KNN Validation Studies

		3D → 1D		2D → 1D		Daylight 2D FPs		Cerius[2] pharmacophore FPs	
data set	ncmpd	*k*	*r*	*k*	*r*	*k*	*r*	*k*	*r*
1. Huang	37	5	0.3489	5	0.4175	9	0.4920	2	0.5002
2. Vacher	68	3	0.5585	3	0.4819	13	0.4238	3	0.4134
3. Gamage	41	4	0.5957	1	0.5935	2	0.4464	3	0.5125
4. Deutsch	40	14	0.8778	1	0.7679	10	0.8025	2	0.7838
5. Murugesan	73	5	0.6286	2	0.5781	6	0.4376	2	0.5101
6. Strupczewski	34	7	0.1543	0[a]	0.0000	11	0.1490	12	0.1885
7. Cristalli	32	20	0.5376	19	0.5250	8	0.6629	10	0.4363
8. Krystek	30	1	0.7152	12	0.5868	1	0.1951	8	0.5298
9. Chang	34	12	0.5338	2	0.4366	5	0.4968	2	0.3480
10. Depreux	26	12	0.8584	1	0.8435	1	0.6529	3	0.7737
averages:			0.5809		0.5231		0.4759		0.4996

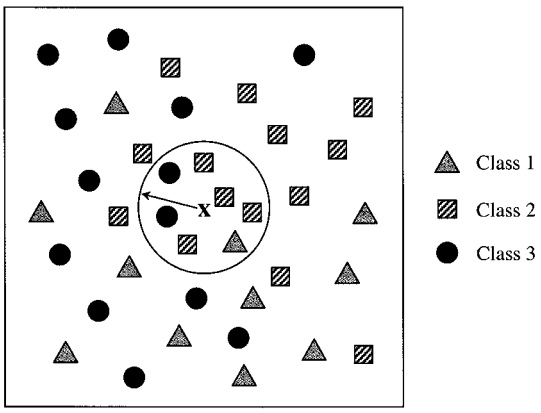
[a] No value of *k* performed better than predicting the activity to be the same for all compounds.

high, so this technique cannot be relied upon to produce a strong QSAR consistently. Nevertheless, it does allow a determination of the general active/inactive discriminating abilities of each similarity measure. On the basis of the average correlation coefficients, the overall rankings for kNN validity are 3D → 1D > 2D → 1D > pharmacophore fingerprints > 2D fingerprints. Once again we see that 1D similarities perform quite well compared to similarities based on explicit 2D and 3D molecular representations.

Prediction of Target Class. The final set of validation tests were aimed at determining each descriptor's ability to distinguish ligands according to their purported biological targets. This is in stark contrast to previous tests, which focused on fine scales of similarity and activity within congeneric series of compounds. Validation on a more coarse scale addresses issues related to chemical library searching; specifically, given one or more compounds known to be active against a particular target, is it possible to identify new actives in a diverse library, without extracting an undue number of false positives? Having access to a descriptor that recognizes critical structural differences between ligands of different targets is clearly an advantage in this situation, because compounds which could not possibly bind to the target of interest will tend to be disregarded.

When presented with a collection of ligands and their corresponding targets, the ideal descriptor would of course separate the ligand classes into distinct clusters, but, unfortunately, this sort of behavior is rarely observed. A less demanding requirement is that the population of compounds within a certain similarity radius of a given ligand be *statistically enriched* with ligands of that same class. In this way, a typical library search that is focused around known actives will identify additional actives at a rate that is enhanced compared to random selection.

Figure 9 illustrates how this idea is used to design a validation experiment. Here, object **x** is compared to a training set of objects that fall into three different categories. Within this collection, the overall class populations N_{bulk} determine random probabilities P_{bulk} for membership in each of the three classes. The analogous quantities N_{sim} and P_{sim} are derived from only those training set objects which lie within the indicated



	N_{bulk}	P_{bulk}	N_{sim}	P_{sim}	P_{sim}/P_{bulk}
Class 1	11	0.289	1	0.167	0.576
Class 2	14	0.368	4	0.571	1.551
Class 3	13	0.342	2	0.333	0.974

Figure 9. Classification of an object **x** by sampling within a similarity radius. For a training set of objects that fall into three different categories, the overall class populations N_{bulk} determine random probabilities P_{bulk} for membership in each class. N_{sim} and P_{sim} are based on only those training set objects which lie within the similarity radius of **x**. The ratios P_{sim}/P_{bulk} indicate that class 2 is statistically enriched within the similarity radius, and thus object **x** would be predicted to be in class 2.

similarity radius of **x**. On the basis of the ratios P_{sim}/P_{bulk}, we see that class 2 is statistically enriched within the similarity radius, and thus object **x** would be predicted to be in class 2. Note that predictions are not simply based on which class appears most frequently within the similarity radius; rather, they are determined by the class that exhibits the greatest *relative increase* in probability inside the similarity radius. For validation purposes, the true class of **x** would be known beforehand, so the accuracy of this and other predictions assesses the ability of the descriptor to distinguish compounds in a way that facilitates chemical library searching.

Validation tests proceed much as they did in the kNN studies, with each compound in a collection being held out for prediction, using the remaining compounds as a training set. The process is repeated using different values of the similarity radius, until an optimal radius is found which yields the best overall jack-knifed classifications. The quality of the predictions can be measured in various ways, perhaps the most obvious being the total number of correct classifications. However, this tends to bias predictions toward classes that are most heavily represented in the training set, potentially leading to poor predictions for compounds that fall into less populous classes. To alleviate this problem, prediction quality Q_p is defined as follows:

$$Q_p = \frac{1}{nclass} \sum_{i=1}^{nclass} \frac{N_{correct}(i)}{N_{bulk}(i)} \qquad (8)$$

Here, the number of correct jack-knifed predictions for each class i is normalized with respect to the number of compounds $N_{bulk}(i)$ contained within that class. Q_p has a maximum value of 1.0 and indicates, on average, the fraction of compounds *within each class* that are classified correctly. The similarity radius giving rise to the largest value of Q_p is judged to be optimal.

Table 4. Summary of CMC[a] Compounds Used in Classification Studies

class	number
ACE inhibitors	30
α adrenergics	23
β adrenergics	66
dopaminergics	39
H_2 antihistamines	24
serotonergics	37

[a] Comprehensive Medicinal Chemistry Database: *ISIS* CMC-3D version 1999.1.

Table 5. Average Intraclass and Interclass Tanimoto Similarities Based on Daylight Fingerprints

	ACE	α	β	dopa.	H_2	serot.
ACE	0.5721	0.4029	0.3736	0.4052	0.3472	0.4022
α	0.4029	0.3867	0.3571	0.3882	0.3467	0.3808
β	0.3736	0.3571	0.4980	0.3627	0.3240	0.3508
dopa.	0.4052	0.3882	0.3627	0.4127	0.3541	0.4000
H_2	0.3472	0.3467	0.3240	0.3541	0.4097	0.3506
serot.	0.4022	0.3808	0.3508	0.4000	0.3506	0.3969

A data set was assembled by searching the Comprehensive Medicinal Chemistry database (CMC-3D[44]) for various categories of ligands defined according to biological target. Compounds were considered for inclusion if the target was clearly identified in the "class" field of the database, and there was no cross-reactivity indicated toward another distinct target. Table 4 summarizes the composition of the data set that was retained for classification studies. Where possible, class definitions were made at receptor subtype levels, i.e., α- and β-adrenergics and H_2 antihistamines. It was not practical to subdivide dopaminergics and serotonergics because only a handful of compounds in the database are reported to have high selectivity for any one subtype. Certain families of targets were avoided, for example, steroidal receptors, because the ligands tend to be so homologous that classification is trivial. Cholinergics were also not considered because their population in the CMC database is almost as large as the combined populations in Table 4.

Table 5 contains average Daylight 2D similarities between pairs of ligands within each class and from one class to the next. In most cases, ligands that bind to a given target are not significantly more similar to each other than they are to ligands that bind to other targets. This suggests that the individual families of ligands probably do not form distinct clusters, but again, this is not assumed to be necessary for the present set of validation tests. It is worth noting that the intraclass similarities here are generally much lower than those of the previous congeneric series of compounds. Average Daylight 2D similarities within the data sets of Tables 1 and 2 are typically greater than 0.7, and some are as high as 0.9.

Results of the jack-knifed classifications are provided in Table 6. ACE inhibitors, β-adrenergics, and H_2 antihistamines appear to be classified reasonably well regardless of how similarity is computed. Problems arise, however, for the other ligand classes, especially serotonergics, which are misclassified by 2D fingerprints

Table 6. Summary of Results from Classification Studies

		number of correct classifications			
class	no. of ligands	3D → 1D	2D → 1D	Daylight 2D FPs	Cerius[2] pharm. FPs
ACE inhibitors	30	30	29	29	30
α adrenergics	23	13	17	11	14
β adrenergics	66	60	59	58	53
dopaminergics	39	31	31	21	29
H_2 antihistamines	24	21	22	21	21
serotonergics	37	20	17	9	12
totals	219	175	175	149	159
similarity radius		0.444	0.461	0.544	0.111
prediction quality		0.781	0.795	0.663	0.726

more than three-fourths of the time, and about two-thirds of the time when pharmacophore fingerprints are used. We note that the misclassified serotonergics were most often predicted to be dopaminergics, which is not surprising considering the cross-reactivity so frequently observed between 5-HT and D_2/D_4 receptor ligands. The CMC database did not indicate any cross-reactivity for these compounds, but there is obviously a general tendency for high similarity between ligands of these two receptor families. In contrast to the 2D and 3D fingerprints, both sets of 1D descriptors correctly classify the majority of the serotonergics, as well as every other family of ligands, yielding identical results for the total number of correct classifications. The 2D → 1D scheme receives a higher prediction quality, however, because it performs significantly better for the comparatively small collection of α-adrenergics. On the basis of the Q_p measure, the overall validity rankings are as follows: 2D → 1D > 3D → 1D > pharmacophore fingerprints > 2D fingerprints.

Conclusions

A 3D or 2D structure can be collapsed onto a 1D coordinate to create a novel and useful way of depicting molecules. With atoms differentiated according to type, these 1D representations of structure may be rapidly aligned, much like protein sequences, to provide a measure of overall similarity between molecules. Similarity calculations can be done at a rate of several hundred per second, so queries within real combinatorial libraries generally require only minutes. With trivial parallelization, this can be sped up to any desired degree.

Extensive validation tests show that 1D representations, whether derived from 3D models or chemical graphs, perform better than 2D hashed fingerprints and 3D pharmacophore fingerprints in a wide variety of situations. This somewhat surprising result suggests that an exceedingly simple model of chemical structure may offer one of the best means of searching chemical libraries and analyzing structure−activity data.

Acknowledgment. We are grateful to Dr. Marvin Waldman for many helpful discussions and advice regarding the 2D → 1D scheme. We also acknowledge the considerable intellectual input from Paul Bartlett, Clark Still, and other members of our scientific advisory board.

Appendix: 2D → 1D Embedding Scheme

When 1D representations are derived from a chemical graph, each 2D distance d_{ij} is simply the number of bonds along the shortest path connecting atoms i and j. Explicit 2D coordinates are not available and, moreover, because the 2D distances are not Euclidean, there frequently is no realizable structure that exhibits true point-to-point interatomic distances that are identical to the path lengths from the chemical graph. Despite these difficulties, it is possible to define a primary axis and obtain an initial 1D projection by borrowing some concepts from the field of distance geometry.[26,27] While the equations and formulas presented here are grounded in Euclidean geometry, they can still be applied formally to the non-Euclidean case to arrive at a high-quality 1D estimate.

One critical piece of the puzzle involves definition of a centroid using only the distances among the atoms. In the absence of explicit atomic coordinates, one cannot directly determine the location of the centroid itself but it is possible to calculate the distance between each atom and the centroid:[26]

$$d_{0i}^2 = \frac{1}{n}\sum_{\mu=1}^{n} d_{i\mu}^2 - \frac{1}{n^2}\sum_{\nu>\mu=1}^{n} d_{\mu\nu}^2 \qquad \text{(A1)}$$

When the distances d_{ij} are not Euclidean, as in the 2D → 1D case, there may be instances wherein a few of the d_{0i}^2 values are actually negative. This nonphysical situation occurs only for atoms very close to the centroid, so the negative terms are quite small in absolute value. Despite the apparent mathematical paradox, no special treatment is required in these cases and the estimated 1D coordinates that are ultimately obtained are still quite reasonable.

Again, while specific atom locations are not known, one can define, for operational purposes, the set of vectors $\{\mathbf{p}_{01}, ..., \mathbf{p}_{0n}\}$ which connect the centroid to each atom in the structure. The lengths of these vectors are given in eq A1, and the law of cosines can be applied to the angle θ_{ij} between any pair of vectors $\mathbf{p}_{0i}$, $\mathbf{p}_{0j}$:

$$d_{0i}d_{0j}\cos(\theta_{ij}) = \frac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2) \qquad \text{(A2)}$$

The reader should recognize the left-hand side of eq A2 as being the scalar or dot product $\mathbf{p}_{0i} \cdot \mathbf{p}_{0j}$. In anticipation of future use, we make the following definition:

$$\mathbf{p}_{0i} \cdot \mathbf{p}_{0j} = \frac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2) \equiv G_{ij} \qquad \text{(A3)}$$

We are now in a position to formally derive the method used to obtain the initial 1D estimate from distance information alone. This derivation differs somewhat from those given in the distance geometry literature,[26,27] as it does not assume at the outset that the eigenvectors of **G** will provide the initial coordinates. Rather, we show how **G** naturally arises in the solution of an optimization problem aimed specifically at identifying the primary axis. The objective, then, is to find a unit vector **v** which originates at the centroid of the molecule and runs along the primary axis we seek. If **v** is in fact the primary axis, then the scalar projections $\{\mathbf{p}_{01} \cdot \mathbf{v}, ..., \mathbf{p}_{0n} \cdot \mathbf{v}\}$ will be the initial 1D coordinates and they

should exhibit a maximum sum-of-squares $SS(\mathbf{v})$:

$$SS(\mathbf{v}) \equiv \sum_{i=1}^{n}(\mathbf{p}_{0i} \cdot \mathbf{v})^2 \rightarrow \text{a maximum} \qquad \text{(A4)}$$

With no explicit Cartesian coordinate system, the only means we have of defining $\mathbf{v}$ is in terms of the vectors $\{\mathbf{p}_{01}, ..., \mathbf{p}_{0n}\}$, and thus we write

$$\mathbf{v} \equiv \sum_{j=1}^{n} v_j \mathbf{p}_{0j} \qquad \text{(A5)}$$

where the unknown scalars $\{v_1, ..., v_n\}$ are to be determined. These scalars should be chosen according to eq A4, with the restriction that $\mathbf{v}$ have unit length:

$$\text{Maximize } SS(\mathbf{v}) \text{ with respect to } (v_1, ..., v_n)$$
$$\text{Subject to } \mathbf{v} \cdot \mathbf{v} = 1 \qquad \text{(A6)}$$

This is a constrained optimization problem that may be solved by employing Lagrange's method of undetermined multipliers.[45] Accordingly, a function L is constructed which contains both the sum-of-squared projections $SS(\mathbf{v})$ and a second term that incorporates the constraint:

$$L(\mathbf{v}, \lambda) \equiv SS(\mathbf{v}) - \lambda(\mathbf{v} \cdot \mathbf{v} - 1) \qquad \text{(A7)}$$

The value of the Lagrange multiplier λ is left unspecified until the correct family of solutions $\mathbf{v}$ is determined.

Proceeding with standard techniques of constrained optimization, the partial derivatives of L with respect to $(v_1, ..., v_n)$ are set to zero, treating the Lagrange multiplier λ as a constant:

$$0 = \frac{\partial L}{\partial v_k} = \frac{\partial SS(\mathbf{v})}{\partial v_k} - \lambda\frac{\partial(\mathbf{v} \cdot \mathbf{v} - 1)}{\partial v_k} \quad \text{for } k = 1, ..., n \qquad \text{(A8)}$$

We note that

$$SS(\mathbf{v}) = \sum_{i=1}^{n}(\mathbf{p}_{0i} \cdot \mathbf{v})^2 = \sum_{i=1}^{n}[\mathbf{p}_{0i} \cdot (\sum_{j=1}^{n} v_j \mathbf{p}_{0j})]^2 = \sum_{i=1}^{n}[\sum_{j=1}^{n} v_j \mathbf{p}_{0i} \cdot \mathbf{p}_{0j}]^2 = \sum_{i=1}^{n}[\sum_{j=1}^{n} v_j G_{ij}]^2 \qquad \text{(A9)}$$

and

$$\mathbf{v} \cdot \mathbf{v} = (\sum_{i=1}^{n} v_i \mathbf{p}_{0i}) \cdot (\sum_{j=1}^{n} v_j \mathbf{p}_{0j}) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \mathbf{p}_{0i} \cdot \mathbf{p}_{0j} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j G_{ij} \qquad \text{(A10)}$$

Expanding eq A8 in terms of eqs A9 and A10, we obtain for $k = 1, ..., n$

$$0 = \frac{\partial L}{\partial v_k} = 2\sum_{i=1}^{n}\sum_{j=1}^{n} G_{ik} G_{ij} v_j - 2\lambda\sum_{i=1}^{n} G_{ik} v_i \qquad \text{(A11)}$$

From eq A3 it is evident that $G_{ik} = G_{ki}$, so eq A11 may be rewritten in the following matrix−vector form:

$$\mathbf{GGv} = \lambda\mathbf{Gv} \qquad \text{(A12)}$$

By making the substitution $\mathbf{y} = \mathbf{Gv}$, a standard eigenproblem is obtained:

$$\mathbf{Gy} = \lambda\mathbf{y} \qquad \text{(A13)}$$

Any of the n eigenvectors of $\mathbf{G}$ will afford an extremum in L, but, as we will show later, the eigenvector with the largest eigenvalue will ultimately yield the greatest value for $SS(\mathbf{v})$. We label this solution $(\mathbf{y}^{\langle 1\rangle}, \lambda^{\langle 1\rangle})$ and note that the dominance of $\lambda^{\langle 1\rangle}$ allows the use of a simple power iteration[46] for determining just this dominant eigenvector−eigenvalue pair, thus avoiding a costly $n \times n$ matrix diagonalization. The vector obtained will not automatically have the correct length—this will be inferred shortly from the constraint on the length of $\mathbf{v}$.

The corresponding solution to eq A12 is obtained by back-substitution:

$$\mathbf{Gv}^{\langle 1\rangle} = \mathbf{y}^{\langle 1\rangle} \qquad \text{(A14)}$$

It is trivial to show that $\mathbf{v}^{\langle 1\rangle} = \mathbf{y}^{\langle 1\rangle}/\lambda^{\langle 1\rangle}$ is a solution to eq A14:

$$\mathbf{G}\left(\frac{\mathbf{y}^{\langle 1\rangle}}{\lambda^{\langle 1\rangle}}\right) = \frac{\lambda^{\langle 1\rangle}\mathbf{y}^{\langle 1\rangle}}{\lambda^{\langle 1\rangle}} = \mathbf{y}^{\langle 1\rangle} \qquad \text{(A15)}$$

The constraint on the length of $\mathbf{v}^{\langle 1\rangle}$ may now be enforced to arrive at the correct length for $\mathbf{y}^{\langle 1\rangle}$:

$$1 = \mathbf{v}^{\langle 1\rangle} \cdot \mathbf{v}^{\langle 1\rangle} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i^{\langle 1\rangle} v_j^{\langle 1\rangle} G_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{y_i^{\langle 1\rangle}}{\lambda^{\langle 1\rangle}}\right)\left(\frac{y_j^{\langle 1\rangle}}{\lambda^{\langle 1\rangle}}\right) G_{ij} = \sum_{i=1}^{n}\left(\frac{y_i^{\langle 1\rangle}}{\lambda^{\langle 1\rangle}}\right) y_i^{\langle 1\rangle} \qquad \text{(A16)}$$

Hence, we have the following requirement:

$$\sum_{i=1}^{n}(y_i^{\langle 1\rangle})^2 = \lambda^{\langle 1\rangle} \qquad \text{(A17)}$$

This result may be combined with eq A9 to show that the eigenvector with the largest associated eigenvalue does in fact yield the maximum sum-of-squared projections:

$$SS(\mathbf{v}^{\langle 1\rangle}) = \sum_{i=1}^{n}(\sum_{j=1}^{n} v_j^{\langle 1\rangle} G_{ij})^2 = \sum_{i=1}^{n}(y_i^{\langle 1\rangle})^2 = \lambda^{\langle 1\rangle} \qquad \text{(A18)}$$

Finally, we show that the initial 1D coordinates we seek are simply the entries of $\mathbf{y}^{\langle 1\rangle}$:

$$x_i^{\text{1D}}\ (\text{initial}) = \mathbf{p}_{0i} \cdot \mathbf{v}^{\langle 1\rangle} = \sum_{j=1}^{n} v_j^{\langle 1\rangle}(\mathbf{p}_{0i} \cdot \mathbf{p}_{0j}) = \sum_{j=1}^{n} v_j^{\langle 1\rangle} G_{ij} = y_i^{\langle 1\rangle} \qquad \text{(A19)}$$

References

(1) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An Integrated Tool for Computer-Assisted Molecular Design and Pharmacophore Recognition from Geometric, Steric, and Substructure Searching of Three-Dimensional Molecular Structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225−251.

(2) Güner, O. F.; Henry, D. R.; Pearlman, R. S. Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 101−109.
(3) Clark, D. E.; Jones, G.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.
(4) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1307.
(5) Sun, Y.; Ewing, T. J. A.; Skillman, A. G.; Kuntz, I. D. CombiDOCK: Structure-Based Combinatorial Docking and Library Design. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 597−604.
(6) Makino, S.; Ewing, T. J. A.; Kuntz, I. D. DREAM++: Flexible Docking Program for Virtual Combinatorial Libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(7) Carbo, R.; Arnau, M.; Leyda, L. How Similar Is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(8) Meyer, A. Y.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 427−439.
(9) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(10) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(11) Jain, A. N. Morphological Similarity: A 3D Molecular Similarity Method Correlated with Protein−Ligand Recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199−213.
(12) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(13) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(14) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules. *J. Med. Chem.* **1998**, *41*, 3314−3324.
(15) Hunt, P. A. QSAR Using 2D Descriptors and TRIPOS' SIMCA. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 453−467.
(16) Dixon, S. L.; Villar, H. O. Investigation of Classification Methods for the Prediction of Activity in Diverse Chemical Libraries.. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 533−545.
(17) Weininger, D. SMILES, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(18) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 370, Mission Veijo, CA 92691.
(19) TRIPOS, Inc., 1699 S. Hanley Road, St. Louis, MO 63144.
(20) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(21) Borg, I.; Groenen, P. *Modern Multidimensional Scaling*; Springer-Verlag: New York, 1997.
(22) Robinson, D. D.; Barlow, T. W.; Richards, W. G. Reduced Dimensional Representations of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939−942.
(23) Robinson, D. D.; Barlow, T. W.; Richards, W. G. The Utilization of Reduced Dimensional Representations of Molecular Structure for Rapid Molecular Similarity Calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 943−950.
(24) Kruskal, J. B. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrica* **1964**, *29*, 115−129.
(25) Fletcher, R. *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization*; Wiley: New York, 1980.
(26) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. The Theory and Practice of Distance Geometry. *Bull. Math. Biol.* **1983**, *45*, 665−720.
(27) Havel, T. F. Distance Geometry. In *Encyclopedia of NMR*; Grant, D. M., Harris, R. K., Eds.; Wiley: New York, 1996.
(28) Cooper, L.; Cooper, M. W. *Introduction to Dynamic Programming*; Pergamon Press: New York, 1981.
(29) Hiramatsu, Y.; Tsukida, T.; Nakai, Y.; Inoue, Y.; Kondo, H. Study on Selectin Blocker. 8. Lead Discovery of a Non-Sugar Antagonist Using a 3D-Pharmacophore Model. *J. Med. Chem.* **2000**, *43*, 1476−1483. (b) Rosowsky, A.; Wright, J. E.; Vaidya, C. M.; Rorsch, R. A.; Bader, H. Analogues of the Potent Nonpolyglutamatable Antifolate N^{α}-(4-Amino-4-deoxypteroyl)-N^{δ}-hemiphthaloyl-L-ornithine (PT523) with Modifications in the Side Chain, *p*-Aminobenzoyl Moiety, or 9,10-Bridge: Synthesis and in Vitro Antitumor Activity. *J. Med. Chem.* **2000**, *43*, 1620−1634. (c) Malamas, M. S.; Sredy, J.; Moxham, C.; Katz, A.; Xu, W.; McDevitt, R.; Adebayo, F. O.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Taylor, J. R. Novel Benzofuran and Benzothiophene Biphenyls as Inhibitors of Protein Tyrosine Phosphatase 1B with Antihyperglycemic Properties. *J. Med. Chem.* **2000**, *43*, 1293−1310. (d) Schelkun, R. M.; Yuen, P.; Serpa, K.; Meltzer, L. T.; Wise, L. D. Subtype-Selective *N*-Methyl-D-aspartate Receptor Antagonists, Benzimidazalone and Hydantoin as Phenol Replacements. *J. Med. Chem.* **2000**, *43*, 1892−1897.
(30) *Catalyst 4.5*; Molecular Simulations, Inc., 9685 Scranton Road, San Diego, CA 9212.
(31) *Cerius² 4.0*; Molecular Simulations, Inc., 9685 Scranton Road, San Diego, CA 92121.
(32) Willet, P.; Winterman, V. A. Comparison of some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1985**, *5*, 18−25.
(33) Cramer, R. D.; Patterson, D. E.; Bruce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(34) *Dataset of 31 Steroids Binding to the Corticosteroid Binding Globulin (CBG) Receptor*; http://www2.ccc.uni-erlangen.de/services/steroids/.
(35) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(36) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure−Activity Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(37) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic *Ah* Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(38) Robert, D. Amat, L. Carbó-Dorca, R. Three-Dimensional Quantitative Structure−Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steroid Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(39) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D. Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(40) In the Siddiqi paper, K_i values were reported for some compounds, while percent displacement was reported for others. Patterson notes that missing K_i values were estimated from percent displacement, but we are not aware of how this can be done in a reliable fashion.
(41) Note that Patterson did perform some parallel randomized experiments in which the true descriptor distances were replaced with distances computed from a uniform random variable. While this is a good way of generating a plot with no neighborhood enhancement, it does not provide an obvious means of correcting for the statistical bias that may be present in a given set of distances.
(42) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; Wiley: New York, 1986.
(43) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the *k*-Nearest-Neighbor Principle. *J. Chem. Inf. Comput.* Sci. **2000**, *40*, 185−194.
(44) *ISIS/Base CMC-3D 1999.1*; MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
(45) Margenau, H.; Murphy, G. M. *The Mathematics of Physics and Chemistry*; Van Nostrand: Princeton, NJ, 1956.
(46) Golub, G. H.; Van Loan, C. F. *Matrix Computations*, The Johns Hopkins University Press: Baltimore, MD, 1983.

JM010137F