Drugs of the Future 1998, 23(8): 885-895

CCC: 0377-8282/98

Software for chemical diversity in the context of accelerated drug discovery

R.S. Pearlman*[*] and K.M. Smith

Laboratory for Molecular Graphics and Theoretical Modeling, College of Pharmacy, University of Texas, Austin TX 78712, USA. [*]Correspondence

CONTENTS

Introduction

The traditional approach to drug discovery typically involved synthesizing and testing 8,000-12,000 compounds for each new drug reaching the marketplace. The traditional "discovery phase" of the overall drug development effort typically took 3-5 years. Computer-assisted drug design (CADD) software was developed to accelerate the discovery phase by helping medicinal chemists reduce the number of low-affinity compounds they synthesized while, simultaneously, helping to design ligands which might interact more favorably with a given receptor.

Recent advances in combinatorial synthetic organic chemistry coupled with advances in laboratory robotics now make it easy to synthesize 12,000 compounds (or many more) in a single week. Similarly, advances in molecular biology now enable the expression of large quantities of relatively pure receptors, and advances in biotesting methodologies now make it easy to screen 12,000 compounds (or many more) in a single week. As a result, lead generation and lead exploration strategies have undergone a remarkable change throughout the pharmaceutical and agrochemical industry.

Ironically, rather than simply freeing us from the time constraints of far slower (traditional) synthesis and testing, combinatorial chemistry and high throughput screening have introduced a new set of resource-related constraints. We still need the traditional CADD software tools to accelerate the lead exploration phase of drug discovery but we now need a new set of software tools to help manage the potentially overwhelming number of compounds a company can now synthesize and/or acquire.

Diversity-related tasks

Although the concept of chemical diversity has been intuitively considered by chemists for many years, the advent of combinatorial chemistry and high-throughput screening have focused unprecedented attention on the need for efficient software tools to address a variety of diversity-related tasks. Perhaps the most fundamental task related to chemical diversity is that of selecting a diverse subset of compounds from a much larger population of compounds. The obvious objective of that task is to identify a subset which best represents the full range of chemical diversity present in the larger population to avoid the time and expense of either synthesizing "redundant" compounds or screening "redundant" compounds. However, in addition to simple subset selection, recent practical experience has revealed other equally (possibly more) important diversity-related tasks which must also be addressed in pharmaceutical and agrochemical industry.

High-throughput screening (HTS) can be an effective approach to lead discovery but is obviously limited by the structural diversity of compounds being screened. What if that population does not include representatives of one or more chemical classes or pharmacophores? Identifting "diversity voids" or "missing diversity" is an important task and, obviously, filling in diversity voids with compounds from various sources is equally important. It is also important to be able to recognize and choose among the many compounds which might fill a particular diversity void. These tasks become increasingly important as the number of combinatorially synthesizable compounds increases with advances in combinatorial chemical methods and as the number of commercially available compounds increases. Similarly, comparing diversities of alternative compound libraries is another important diversity-related task.

In addition to simple diverse subset selection, it is often desirable to select a subset chosen not only to provide structural diversity but also to satisfy one or more nonstructural criteria or "biases." For example, compound availability and/or physical properties may be important when selecting a subset for HTS purposes. Reagent cost and/or reagent usage frequency (to make most effective use of limited robotic resources) may be important when deciding which compounds to actually synthesize out of a very large range of compounds synthetically accessible through combinatorial chemical methods. Clearly, nonstructurally biased subset selection will yield subsets with somewhat less structural diversity than a subset chosen simply to maximize structural diversity, but practical considerations often make biased subset selection a very important diversity-related task.

In order to address these and other diversity-related tasks, we must first consider how chemical structures can be described for chemical diversity purposes. We shall refer to such descriptors as "metrics" of a "chemistry-space." An extremely important but often overlooked diversity-related task is that of choosing the chemistry-space metrics which best represent the structural diversity of a given population of compounds. For example, combinatorially generated populations or populations chosen to be similar to a particular active ("lead") compound are inherently less diverse than other more randomly assembled populations. Thus, it is quite reasonable to expect that metrics specifically tailored to focus on the limited diversity of such "focused populations" will provide some advantages over metrics which were developed to best represent the broad range of diversity found in "nonfocused populations." Last, but certainly not least, the notion of considering alternative chemistry-space metrics reminds us of the need for a rational approach for validating chemistry-space metrics – an important and often misunderstood diversity-related task.

Chemistry-space concepts and diversity-related algorithms

What do we mean by "chemical diversity"? The notions of chemical similarity, dissimilarity and, consequently, "diversity" are all related to the distance between chemical compounds positioned in some multidimensional "chemistry-space," the axes of which are the structure-related " chemistry-space metrics" mentioned above. In order to be a well-defined vector space, our chemistry-space axes must be orthonormal (mutually orthogonal, uncorrelated and normalized). We must also define a method for computing a true distance (one which satisfies the triangle inequality) within that space. The "diversity" of compounds positioned in chemistry-space is intuitively related to the intercompound distance as measured in that space.

Whereas the dimensionality of our physical world is predefined as 3, the dimensionality of a chemistry-space (as well as the definition of axes) can be chosen to best represent the diversity of a given population of compounds. Most software for addressing chemical diversity uses some form of "molecular fingerprint" to describe each compound in a population. Fingerprints are bit-strings (sequences of 1s and 0s) representing the answers to yes/no questions about the presence or absence of various substructural features within the molecular structure of a given compound. Although not often discussed in such terms, each bit represents an axis in a multidimensional chemistry-space. Each axis could have either of two values: 0 or 1. Fingerprints represent very high-dimensional chemistry-spaces: typically a few hundred or thousands of bits.

Fingerprints were developed as a means of addressing molecular similarity. The well known Tanimoto similarity index, T, ranges between 0.0 and 1.0, compares the bits set to 1 (indicating the presence of a given substructure) in the fingerprints of two compounds and has proven useful for finding similar compounds, within very large databases of chemical structures. Thus, it is typically assumed that the "Tanimoto dissimilarity", (1-T), represents a useful measure of distance within such high-dimensional spaces. Distance-based diversity algorithms consider only the distances between compounds in chemistry-space and are used to select diverse subsets by choosing compounds guaranteed to be distant from other compounds in the selected subset. Practical experience and various validation studies (*e.g.*, Brown and Martin [1]) indicate that such high-dimensional, distance-based, diversity-related algorithms are, indeed, useful for simple diverse subset selection. However, the following criticisms should be considered:

1. The "questions" (presence or absence of particular substructures) corresponding to the bits of fingerprints were specifically developed to identify compounds similar to one another. They were not developed to focus on differences which would constitute a structurally diverse subset.

2. Although some software permits users to redefine the "questions" corresponding to the bits of a fingerprint, this is rarely (if ever) done in actual practice. Thus, we use fingerprints developed to find similar compounds in diverse populations to select dissimilar (diverse) compounds not only from diverse populations but also from focused populations.

3. The Tanimoto similarity index was developed to gauge similarity, not dissimilarity. That is, if T(A,B) and T(A,C) (the Tanimoto similarities between compounds A, B and C) are 0.9 and 0.8, compounds A and B are probably more structurally similar than compounds A and C. However, it is not at all certain that if (1-T(X,Y)) and (1-T(X,Z)) are 0.9 and 0.8, that compounds X and Y are more dissimilar (diverse) than compounds X and Z.

4. Although the Tanimoto similarity index appears useful for the similarity purposes for which it was designed, (1-T) is not a valid measure of distance since it does not obey the triangle inequality. Thus, it appears that distance-based diversity (and, possibly, similarity) algorithms might be improved by using either the Euclidean

distance or the Hamming distance as suggested by Pearlman (2, 3).

5. Pearlman has also shown that most fingerprint spaces are nonorthogonal. That is, the settings of some bits are highly correlated with the settings of other bits. This further compromises the validity of distances computed (by any method) in fingerprint spaces.

Despite these criticisms, high-dimensional distance-based diversity algorithms appear to work satisfactorily for simple diverse subset selection, the most fundamental of the diverse-related tasks. However, as the term implies, distance-based algorithms consider only intercompound distances, not the absolute positions of compounds in chemistry-space. As a result, distance-based algorithms are inherently limited and are ill-suited for many of the other diversity-related tasks. For example, locating diversity voids in chemistry-space is essentially impossible since distance-based algorithms do not reference location.

In contrast, by dividing each axis of a multidimensional space into “bins”, cell-based diversity algorithms partition chemistry-space into a lattice of multidimensional hypercubes and, thereby, consider not only intercompound distance but also absolute position of compounds in chemistry-space. As will be illustrated below, this additional information makes cell-based diversity algorithms much more powerful and easily applicable to all of the diversity-related tasks mentioned in the preceding section. However, whereas distance-based algorithms can be applied in either a high-dimensional or low-dimensional representation of chemistry-space, cell-based algorithms can only be applied in low-dimensional chemistry-spaces. (For example, a 1000-bit fingerprint corresponds to a 1000-dimensional chemistry-space which would be partitioned into 2^{1000} cells – an astronomically large number of cells almost all of which would contain no compounds.)

In order to take advantage of the power and utility of cell-based diversity algorithms, we must identify low-dimensional metrics from which to construct a less than 10-dimensional chemistry-space satisfactory for diversity purposes. There have been numerous attempts to use “traditional” molecular descriptors (*e.g.*, molecular weight, shape factors, estimated logP, surface area, dipole moment, HOMO-LUMO gap, *etc.*) as the axes of a low-dimensional chemistry-space. There are three basic reasons for which these efforts have not proven particularly useful.

1. Many of the “traditional” descriptors are highly correlated; the axes of a vector space should be orthogonal (uncorrelated).

2. Some traditional descriptors (*e.g.*, logP and pK_a) are strongly related to drug transport or pharmacokinetics but are very weakly related to receptor affinity or activity as measured in most screening-based drug discovery efforts.

3. The traditional descriptors are whole-molecule descriptors which convey very little information about the details of molecular substructural differences which are the basis of structural diversity.

The first problem above could be addressed to a limited extent by using principal components of the “traditional” descriptors as the axes but the second and third more fundamental problems would remain. The advantages of cell-based methods can not be realized unless some nontraditional chemistry metrics can be found which enable the definition of a meaningful low-dimensional chemistry-space.

BCUT values: novel low-dimensional chemistry-space metrics

In 1989, Burden (4) suggested that a “molecular ID number” could be defined in terms of the two lowest eigenvalues of a matrix representing the hydrogen-suppressed connection table of the molecule. More specifically, Burden suggested putting the atomic numbers on the diagonal of the matrix. Off-diagonal matrix elements were assigned values of 0.1 times the nominal bond type if the two atoms are bonded and 0.001 if the two atoms are not bonded. He also added 0.01 to the off-diagonal elements representing “leaf edges” in the molecular graph (*i.e.*, terminal bonds to the last atom in a chain). In suggesting that structurally similar compounds would be near each other in an ID-ordered list, Burden was actually proposing a 1-dimensional chemistry-space. Since fingerprint-based similarity searching methods were just becoming available for the medium sized databases (under 0.5 million compounds) found in pharmaceutical and agrochemical industry, Burden’s seemingly far-fetched suggestion was generally ignored.

In 1993, eager to find some sort of “similarity searching method” applicable to the Chemical Abstracts Service (CAS) Registry File of approximately 12 million structures, Rusinko and Lipkus (5) applied Burden’s suggestion to a test database of 60,000 compounds. The results were relatively poor compared to fingerprint-based similarity searching methods but much better than expected. They also experimented with the notion of assigning a constant value to all diagonal matrix elements or a constant value for all bonded off-diagonal elements but, in each case, were using the lowest eigenvalue of a single matrix to define a 1-dimensional chemistry-space.

Based on Burden’s (B) original suggestion and CAS’s (C) “validation” of that suggestion, Pearlman at the University of Texas (UT) added the following very significant extensions which resulted in what we now refer to as the BCUT approach (2, 3, 6).

1. Given that a 1-dimensional chemistry-space showed some signs of promise, a similarly defined multidimensional chemistry-space should be even more promising. This is easily accomplished by using more than one matrix to represent each compound.

2. Mathematical analysis reveals that all eigenvalues of such matrices contain information related to molecular structure. The lowest and highest eigenvalues reflect the most different information (are least correlated). Considering both the lowest and highest eigenvalues

provides another mechanism to extend Burden's original suggestion to a multidimensional space.

3. Pharmaceutical and agrochemical researchers are interested in structural diversity with respect to the way in which compounds might interact with a bioreceptor. Since atomic number has almost no bearing on the strength of intermolecular interactions, much more relevant metrics can be defined by putting more relevant atomic properties on the diagonals of four "classes" of BCUT matrices: atomic charges, polarizabilities, H-bond donor- and acceptor-abilities corresponding to the electrostatic, dispersion and H-bonding modes of bimolecular interaction.

4. Burden's suggestion of using nominal bond-type information for the off-diagonal elements of the matrices was very good and should be retained. However, using CONCORD (7) to generate 3D structures opens the possibility of putting various functions of interatomic distance on the off-diagonals and, thereby, defining metrics which encode information about the 3D structure.

5. Another approach to incorporating aspects of 3D structure is to use atomic surface areas to weight the atomic properties placed on the diagonals.

6. Noting that the matrices contain atomic properties on the diagonals and connectivity information on the off-diagonals, there is clearly a need for a scaling factor to provide the proper balance of the two types of information.

Given the large number of possible combinations of diagonal, off-diagonal and scaling factor choices, it is clear that some method is needed to enable the rational choice of that particular combination of BCUT values (eigenvalues) which forms the chemistry-space which best represents the structural diversity of a given population of compounds. Many combinations can be quickly eliminated by requiring that the axes of the chemistry-space be mutually orthogonal. For example, different charge-related values (*e.g.*, Gasteiger-Marsili charges, AM1 charges, AM1 densities, *etc.*) all convey the same fundamental information and, therefore, will be intercorrelated. On the other hand, with the exception of the H-bond-ability matrices[1], both the highest and lowest eigenvalues should be relevant and turn out to be relatively uncorrelated. Often, a 6-dimensional chemistry-space (two charge-BCUTs, two polarizability-BCUTs and two H-bond-BCUTs) yields the best chemistry-space for a given population. Sometimes, the H-bond-acceptor- and charge-BCUTs are correlated, yielding a 5-dimensional chemistry-space. Pearlman and Smith (2, 3, 6) developed a powerful "auto-choose" algorithm which automatically determines both the best dimensionality of the chemistry-space and best choice of exactly which metrics best represent the structural diversity of a given population of compounds.

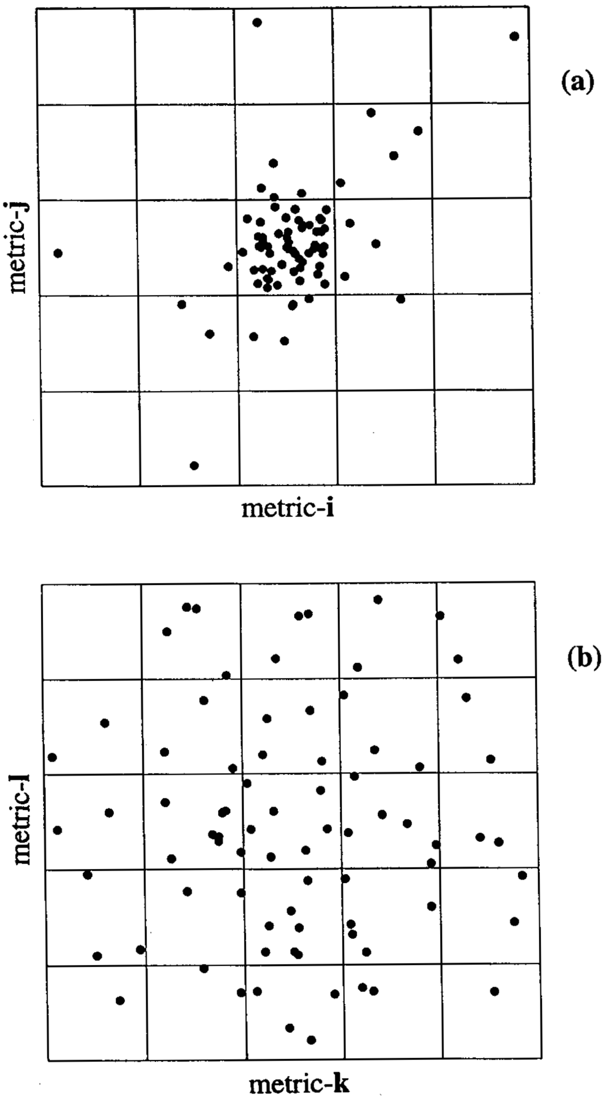

Fig. 1. (a) A cartoon representation of a non-optimal 2-dimensional chemistry-space showing poorly distributed compounds. (b) A representation of a better 2-dimensional chemistry-space showing more evenly distributed compounds.

The rationale for the auto-choose algorithm can most easily be explained by reference to Figure 1, which depicts two "cartoon" representations of the same population of compounds distributed in two different 2-dimensional chemistry-spaces. Recall that the most fundamental of diversity-related tasks is that of rational, structure-based diverse subset selection. Suppose that one were asked to select a subset of 25 diverse compounds. The chemistry-space defined by axes **i** and **j** in Figure 1a does a poor job of distinguishing one compound from another and positions most of the compounds in the central "cell". Since most other cells are empty, we are forced to choose many compounds from that central cell at random. Clearly, this would be a very poor choice of chemistry-space axes for this population of compounds since it would provide little or no advantage over a

[1] The lowest eigenvalues of these matrices contain "information" about atoms in the molecule which are neither H-bond donors or acceptors. Since all non-H-bonding atoms have the same zero value of H-bond ability, the lowest eigenvalues of these matrices convey relatively little useful information.

random choice made without reference to any chemistry-space considerations. In contrast, the chemistry-space defined by axes **k** and **l** in Figure 1b does a much better job of distinguishing compounds based on structural diversity and enables the selection of a subset of 24 diverse compounds (satisfactorily close to the desired size of 25) by simply choosing one compound from each of the 24 occupied cells. Note that by choosing compounds as close to the center of each occupied cell as possible, we have described a very natural cell-based subset selection algorithm which yields a set of compounds which "covers" the full range of diversity represented by the population and includes compounds which are mutually distant from one another. Clearly, the chemistry-space yielding the most uniform distribution of compounds would best suit our purposes. However, real-world populations will not be distributed in a perfectly uniform fashion. Valence and steric considerations limit the continuity of structures which can be achieved and, more obviously, the history of discovery efforts at a given company will result in non-uniform corporate databases reflecting those focused efforts. However, the χ-squared statistic provides a measure of how well one distribution matches another. Thus, minimizing the χ-squared statistic can reveal which combination of metrics yields a distribution of compounds closest to the desired uniform distribution. Simultaneously, by considering a range of dimensionalities, this χ-squared approach also reveals the dimensionality of the chemistry-space which best represents the diversity of the given population of compounds. Clearly, in order to include as much structure distinguishing information as possible, the algorithm will choose the highest possible dimensionality which does not result in correlated (nonorthogonal) axes.[2]

Note that the χ-squared approach yields the combination of metrics which best represents the diversity of a given population of compounds. For "truly diverse" populations of compounds, we are not surprised to find the same (or similar) "universal chemistry-space" definition being reported by different users. In contrast, populations resulting from generating different combinatorial libraries should be expected to occupy different and individually less diverse regions of "universal chemistry-space." Clearly, the χ-squared approach enables us to tailor a chemistry-space to best represent a focused population – an important diversity-related task listed above.

It should also be noted that this auto-choose, χ-squared approach can be applied not only to combinations of BCUT values but to combinations involving any other low-dimensional chemistry-space metrics. Although experience to date strongly supports the use of BCUT values as metrics, the DiverseSolutions software developed by Pearlman and Smith (8) encourages the user to consider his own metrics in addition to BCUT values. However, this must be done with extreme caution for the following somewhat ironic reason. Imagine assigning random numbers to each compound of a large population and then considering those numbers as a potential axis of a chemistry-space. Since the random numbers would be uniformly distributed over the population of compounds, the χ-squared approach would perceive this "metric" (and other similarly random "metrics") as good choices as axes of a chemistry-space. This brings us, rather dramatically, to the need to validate the choice of metrics used to define a chemistryspace.

Validation of chemistry-space metrics

Obviously, chemistry-space metrics which are merely random numbers with no relation to structure would be of no use for diversity-related tasks or any other purpose. How can we demonstrate that a given set of metrics is actually reflecting differences in molecular structure and, thereby, validate those metrics for use in addressing chemical diversity-related tasks? Perhaps the most intuitive approach to metric validation is to use the metrics as QSAR descriptors, *i.e.*, establish a linear regression equation relating the metrics to the experimentally measured "activities" of a set of compounds. Let us imagine that we can put aside the differences between receptor affinity and actual activity (differences due to transport issues or secondary processes in the cascade of events between initial receptor binding and eventual pharmacological effect). Since it would be impossible to establish a statistically significant regression based on meaningless, random numbers, demonstrating such a regression would be proof that the metrics are not random numbers but true indicators of chemical structure. After auto-choosing a 6-dimensional BCUT chemistry-space to best represent the diversity of their entire corporate database, Weintraub and Demeter (9) used those 6 BCUT values to regress the log IC_{50} values measured for 800 ligands at the benzodiazepine site of the GABA-A receptor and obtained a PLS model essentially as good as one they previously obtained based upon 70 classical QSAR descriptors.

While the results of Weintraub and Demeter certainly confirmed the validity of BCUT values as chemistry-space metrics, those results are, unfortunately but not unexpectedly, quite rare! As will be illustrated below, chemistry-space metrics are not QSAR descriptors and rarely yield regressions as good as that reported by Weintraub and Demeter. Chemistry-space metrics are intended to position compounds in a structure-based chemistry-space. QSAR descriptors are intended to provide quantitative estimates of bioactivity. Chemistry-space metrics are intended to reflect (in a necessarily crude manner) all features of molecular structure. In contrast, QSAR descriptors are specifically chosen to reflect (as accurately as possible) only those features of molecular structure which have been found relevant for a particular bioactivity. It is well known that QSARs will give unreliable estimates of activity when applied to compounds contain-

[2] Correlated axes would result in some highly populated cells along a diagonal of the chemistry-space and corresponding empty cells surrounding that diagonal. This would yield a high X-squared and the chemistry-space would be rejected.

ing structural features not present in the training set used to identify the "relevant" features characterized by the QSAR descriptors. We certainly should not expect QSARs based on chemistry-space metrics to do any better since (i) the metrics were not intended for this purpose and (ii) as will be explained below, position in chemistry-space is not quantitatively related to activity.

How then should chemistry-space metrics be validated? Pearlman and Smith (2, 6, 10) have presented a simple yet novel approach to metric validation which they refer to as activity-seeded, structure-based clustering. Unlike typical clustering algorithms (based on structure alone) which can be used for a variety of tasks, this algorithm requires activity data (preferably quantitative data) for a set of compounds and is intended only for the diversity-related task of validating chemistry-space metrics. Given a set of active compounds which all bind to a given receptor in the same way, it is certainly reasonable to expect that those active compounds should be positioned near each other in a small region of chemistry-space if the chemistry-space metrics are valid. The activity-seeded, structure-based clustering algorithm provides a method for directly testing that expectation in the typical case in which the chemistry-space dimensionality is greater than 3 and, thus, simple visual inspection of the distribution of active compounds is difficult or impossible. The algorithm consists of the following procedure:

1. Choose a unit-cluster radius: a small distance in the chemistry-space to be validated.

2. Center a sphere of that radius on the most active compound in the validation set.

3. Assign other active compounds located within that sphere to that "unit-cluster".

4. Center another sphere on the next most active compound not already assigned to some unit-cluster.

5. Repeat steps 3 and 4 until all active compounds have been assigned to some unit-cluster.

6. "Coalesce" adjoining (overlapping) unit-clusters and record the number of unit-cluster spheres per coalesced-cluster.

The algorithm can be implemented as an **O**(N) process and, thus, is extremely fast. More significantly, the algorithm can be used to validate all types of chemistry-space definitions including other (non-BCUT), low-dimensional chemistry-spaces and those based on high-dimensional fingerprints as well. When used in a cell-based context, the unit-cluster radius is typically chosen to yield a tiny hypersphere of volume equal to that of a single hypercubic cell reflecting the "resolution" corresponding to a user-specified number of bins/axis (see below). In any case, the total number of unit-cluster spheres contained in all coalesced-clusters provides an upper bound on the volume of chemistry-space required to contain all the active compounds.

Using the activity-seeded, structure-based clustering algorithm, Pearlman and Deanda (11) have performed a number of validation studies. For example, after auto-choosing the BCUT chemistry-space which best represents the diversity of compounds in MDL's MDDR database (12), they computed the positions of 191 relatively diverse ACE inhibitors in that chemistry-space. The 191 inhibitors were culled from the primary literature (13-20). Measured activities (-log IC_{50}) were reported for all compounds and spanned the range 5.24-9.64. The 74 most active compounds (top 40%) had activities in the range 7.85-9.64 and were identified as "highly active" compounds. If the BCUT values used as chemistry-space metrics were random numbers or quantities unrelated to structure and intermolecular interaction, the active compounds would be randomly distributed throughout chemistry-space. However, using the activity-seeded, structure-based clustering algorithm, they found that the 74 "highly active" compounds are all contained by just 3 coalesced-clusters occupying less than 0.02% of the entire chemistry-space and less than 0.19% of occupied chemistry-space. Significantly, the 3 clusters were close to each other; the largest intercluster distance being just 3.2R where R is the unit-cluster radius.

It is instructive to consider the analogous results obtained using all 191 compounds (including the 117 "poorly active" compounds). Once again, the active compounds were all clustered relatively near each other but they occupied a much larger volume of chemistry-space than that occupied by just the 74 "highly active" compounds. This result is entirely consistent with expectations. There can be many different structures which exhibit poor to modest activities. In contrast, there are relatively fewer structures which exhibit high activities. This fact may be easier to appreciate by considering the notion of making structural modifications of a very highly active compound. There may be a few modifications which preserve high activity but there are far more modifications which reduce or even completely destroy activity.

The fact that poorly to modestly active compounds are spread over larger regions of chemistry-space than highly active compounds illustrates one reason for which QSAR is an invalid approach to metric validation and also illustrates that chemistry-space metrics cannot (and should not) be used as QSAR descriptors. Compounds with significantly different structures would be positioned at widely distant points in chemistry-space but could exhibit low or moderate bioactivities (or affinities) of exactly equal magnitudes. The converse illustrates a second reason for which QSAR is an invalid approach to metric validation and also illustrates that chemistry-space metrics cannot (and should not) be used as QSAR descriptors. For example, adding just a single methylene unit to the middle of the -CH_2CH_2OH side-chain of some highly active compound could completely destroy the activity if the propyl-hydroxy derivative no longer fits into the receptor. Thus, two highly similar compounds (which any valid metrics would place very near each other in chemistry-space) could have entirely dissimilar activities. Clearly, neither QSAR nor any other approach based on the assumption of a quantitative relationship between activity and precise position in chemistry-space will be a valid approach to metric validation. On the other hand, the activity-seeded, structure-based clustering approach clearly indicates whether a given set of metrics places compounds active against the same receptor in the same small region of chemistry-space and, thus, provides a rational basis for metric validation.

We will revisit the notion of metric validation when we discuss the computer graphic visualization of "receptor-relevant subspaces" towards the end of this article.

Using low-dimensional metrics for diversity-related tasks

Once one has determined which metrics define the chemistry-space which best represents the diversity within a given population of compounds, one is then able to use various cell-based algorithms to address all of the other diversity-related tasks mentioned at the beginning of this article. Recall that such cell-based algorithms can exploit not only the knowledge of inter-compound distances but also the knowledge of absolute positions of compounds in chemistry-space. Structurally similar compounds are positioned near each other in chemistry-space and, thus, are found "clustered" in the same or neighboring cells.

A chemistry-space, like any other vector-space, must be comprised of normalized axes (so that a distance of, say, 4 units in one direction is equivalent to a distance of 4 units in any other direction). Thus, the "cells" are hypercubes resulting from dividing each of the normalized axes of a chemistry-space into equal numbers of evenly spaced "bins." The number of bins/axis is directly related to the "resolution" with which one examines the distribution of compounds across chemistry-space and is inversely related to the apparent "occupancy" of that chemistry-space. For example, if 250,000 compounds are distributed in some 6-dimensional chemistry-space and each axis is "divided" into just one single bin, all 250,000 compounds would be contained in just one single cell. The occupancy (number of occupied cells divided by total number of cells) would be 100% but the resolution would, obviously, be uselessly low. If each axis were divided into 20 bins, there would be 20^6 = 64,000,000 tiny cells. In this case, the occupancy would be extremely low and the resolution would be uselessly high: most cells would be empty and even very similar compounds could be in different, non-neighboring cells. Cell-based algorithms for some tasks automatically choose the number of bins/axis most appropriate for that task and population. Other tasks require that the user decide on the resolution (see below). Recalling that typical populations of compounds are not uniformly distributed, experience has shown that choosing the number of bins/axis which yields roughly 12-16% occupancy provides an appropriate level of resolution for most purposes.

Simple and biased diverse subset selection

Our explanation of the χ-squared approach to auto-choosing the metrics of a chemistry-space also illustrated the essence of the natural, cell-based approach to diverse subset selection. As implied in that illustration, we recommend selecting one compound from each occupied cell although the DiverseSolutions software also allows one to sample each cell in proportion to its occupancy or by selecting up to some fixed number of compounds per cell. Once the user has specified the sampling protocol (number per cell) and the size of the desired subset, the software automatically finds the number of bins/axis which yields the number of occupied cells required to provide a subset closest in size to that requested. Cell-based algorithms are extremely fast and are especially well-suited to handling very large populations of compounds. Even if the software must make three of four guesses before finding the best number of bins/axis, selecting a subset of 50,000 structurally diverse compounds from a population of 0.5 million would take approximately 10 cpu seconds on a modest workstation.

By selecting compounds nearest the center of each cell, we can avoid choosing compounds near each other but just barely on opposite sides of a plane separating two cells. In other words, by selecting compounds nearest the center of each cell, we are selecting a subset of maximal structural diversity, *i.e.*, simple diverse subset selection.

Biased subset selection can easily be accomplished by allowing the user to construct a modified selection rule. In other words, rather than choosing the compound nearest the center of a given cell, one can arrange to choose the compound which provides the best (user-specified) compromise between distance from center and one or more nonstructural properties. For example, given a choice between two compounds from the same small region (cell) of chemistry-space, availability (price, quantity on hand, *etc.*) might certainly be important considerations for assembling a subset for general screening purposes. Recalling that logP is a poor chemistry-space metric but, nevertheless, quite important for bioactivity, choosing compounds from each cell closest to some particular "ideal" logP could be advantageous.

Biased subset selection can also be used to improve the efficiency and economy of combinatorial library synthesis. Imagine that 1000 A-type and 1000 B-type reactants could be used to make 1,000,000 AB-type products but that just 10,000 products are desired for screening purposes. Selecting and reacting 100 diverse As and 100 diverse Bs offers obvious practical advantages but, clearly, does not yield as diverse a set of 10,000 products as could have been selected from the complete set of 1,000,000 products. Simple diverse subset selection from all the products would undoubtedly result in the need to use many more than 100 of each type of reactant. By keeping track of the frequency with which each reactant is used in the products being selected, and by specifying the dimensions of the plates used for the syntheses (*e.g.*, typical 8x12 = 96-well plate), DiverseSolutions (8) used in conjunction with CombinDBMaker (for combinatorial database generation [21]) enables the user to specify a selection rule which chooses compounds providing the best (user-specified) compromise between distance from center of cell and economy.

Identifying and filling in diversity voids

In order to address the possibility of finding leads to bioactive compounds in regions of chemistry-space not covered by their current collection of compounds, pharmaceutical and agrochemical companies allocate a certain fraction of their resources to compound acquisition programs: purchasing, trading for or synthesizing additional compounds for screening. Practical considerations (cost, screening capacity, *etc.*) limit the number of compounds companies choose to acquire.

Identifying and filling in diversity voids is trivially simple using cell-based algorithms. Obviously, “empty” cells represent regions of missing diversity. “Empty” can be defined to mean either that the cell contains no compounds or that it contains less than some user-specified number. When identifying the diversity voids in a given population of compounds, the number of empty cells will depend not only on how those compounds are distributed but also upon the “resolution” at which the “search” for empty cells is performed. Whereas the number of bins per axis can be chosen by the software during subset selection, the user must choose the number of bins/axis which will yield a number of diversity voids consistent with those practical considerations which limit the number of compounds his company chooses to acquire (see below).

Since cell-based algorithrns (unlike distance-based algorithms) reference absolute compound coordinates in chemistry-space, filling in diversity voids is also trivially simple using cell-based methods: compounds (from some secondary population) are acquired if they would occupy a previously “empty” cell in the chemistry-space containing the primary population. Of course, this entails precomputing the coordinates (metrics) of the secondary population in the same chemistry-space as that used to contain the primary population. Since we know exactly which cell would be filled by each candidate compound, we can easily bias our choice of fill-in compounds using the same sort of nonstructural criteria as discussed for biased subset selection. The user can also specify how many compounds he wants to add to “empty” cells. DiverseSolutions then presents the list of compounds to acquire in various formats to facilitate purchase decisions (*e.g.*, compounds which would ensure at least 1 compound in each “empty” cell, compounds which would ensure at least 2 compounds in each “empty” cell, *etc.*).

Finding the diversity voids in a population of 0.5 million compounds typically takes less than 5 cpu seconds on a modest workstation. Filling in those voids (to the extent possible) from a library of 50,000 compounds typically takes less than 4 cpu seconds. Thus, the user can easily experiment with several values of bins/axis and several filling in protocols.

It is worth noting that the filling in process does not require any information about the compounds contained in the primary population. All that is required is the definition of the chemistry-space (*i.e.*, name and range of the metric corresponding to each axis), the number of bins/axis and the cell numbers of the “empty” cells. Thus, without revealing the compounds in its proprietary database, company-X can enable company-Y to identify company-Y compounds which would fill in company-X’s missing diversity.

Comparing the diversities of two or more populations

Occasionally, it may be useful to compare the diversities of two (or more) populations of compounds — perhaps alternative third-party libraries one could purchase or alternative combinatorial libraries one could synthesize to augment the diversity of a corporate database. Distance-based approaches merely allow the comparison of statistics related to nearest-neighbor distances within the two populations. Such statistics provide no information regarding the redundancy of compounds contained in both populations or even the extent to which the regions covered by the populations overlap in chemistry-space.

In contrast, a cell-based approach provides an extremely rapid answer to the fundamental, pragmatic questions at the heart of the population-comparison issue: if population-A and population-B are alternative libraries and population-X is a corporate database,

1. how many population-A compounds fill voids in population-X?
2. how many population-B compounds fill voids in population-X?
3. how many population-A compounds fill voids in population-B?
4. how many population-B compounds fill voids in population-A?

Questions 1 and 2 are easily answered by identifying the voids in population-X and hypothetically using both populations-A and -B to fill those voids. Similarly, questions 3 and 4 can be answered simultaneously by a “compare diversities” algorithm which, essentially, performs two find-voids and fill-in tasks at the same time. In addressing questions 1 and 2, it is natural to use the chemistry-space which best represents the diversity of population-X. In addressing questions 3 and 4, one might use a chemistry-space previously defined for some related population or use a chemistry-space defined to best represent the union of the A and B populations.

Dimensional reduction: receptor-relevant subspaces

Pictures can be worth “a thousand words”. This fact has driven many people to resort to cartoon illustrations depicting “islands” of active compounds or combinatorial libraries in abstract, wishfully contrived, 2-dimensional “chemistry spaces”. Obviously, these cartoons are merely intended to illustrate broad concepts but, frankly, may often present a very misleading impression of how real compounds or libraries are really positioned in the real chemistry-space.

Other researchers have resorted to applying various mathematical techniques (*e.g.*, Sammon maps, Kohonen maps, principle components, *etc.*) to reduce the dimensionality of higher dimensional spaces down to 2 or 3 dimensions just so they can display computer graphic images. However, there is a very serious problem associated with any such mathematically based approach to dimensional reduction: the inevitable loss of potentially important information. That is, if 1000 bits are required to adequately describe structural diversity in a high-dimensional fingerprint space or if 6 BCUTs are required to adequately describe structural diversity in a low-dimensional space, the mathematical transformation of 1000 or 6 dimensions down to 2 or 3 must, inevitably, discard information which was deemed necessary for the adequate description of the structural diversity of those compounds.

Pearlman and Smith (10) have recently described a novel yet simple approach to dimensional reduction for specific purposes which minimizes the aforementioned loss of potentially important information. This approach is best appreciated by reference to a simple analogy. Consider the list of car descriptors needed to describe all the possible features available to all car buyers. The list would include descriptors such as brand, price, color, 2-door *vs.* 4-door, length, engine size, transmission type, cup holder location, *etc.* All of these descriptors would be of potential interest to some car buyers. However, a particular buyer might care about only three of those descriptors, *e.g.*, color, size and cup holder location. Those descriptors would be "relevant" to that particular car buyer. The other descriptors would be irrelevant to that particular buyer but, obviously, might be relevant to some other car buyer. If, for example, the particular car buyer prefers silverish cars, all cars for which he or she "shows affinity" will be very similar (tightly clustered) with respect to color but could be dissimilar (poorly clustered) with respect to brand.

Each of the (typically) 6 BCUT metrics chosen by the χ-squared algorithm (described above) represent descriptors which might be relevant to some receptor. However, a particular receptor might care about just 2 or 3 (or more) of those descriptors. Given a number of compounds for which a particular receptor has high affinity, we can identify the receptor-relevant subspace for that receptor by identifying the axes (metrics) along which the active compounds are tightly clustered. Axes (metrics) which are irrelevant to this particular receptor will show poor clustering of the active compounds. The algorithm for identifying receptor-relevant metrics has been implemented in the DiverseSolutions software.

To illustrate the utility of this algorithm, Pearlman and Smith (10) identified a 3-dimensional ACE-receptor-relevant subspace within the 6-dimensional chemistry-space which best represents the structural diversity of all compounds in MDL's MDDR database of drugs and bioactive compounds. Imagine that the MDDR compounds are actually the compounds in some corporate database. Figure 2 lists the 6 BCUT metrics which comprise the MDDR chemistry-space. By determining which of those metrics best cluster the aforementioned 74 highly active ACE inhibitors, it was found that 3 of the 6 metrics – those related to negative charge, H-bond donor-ability and H-bond acceptor-ability – are apparently relevant to the ACE receptor while the other 3 are not. Figure 2 lists the metrics in order of their ACE-receptor-relevance. Figure 3 depicts a cartoon representation of ACE inhibitor binding as proposed by Wyvratt and Patchett (16). Note that the 3 ACE-receptor-relevant metrics appear to be consistent with the proposed binding model.

bcut_gastchrg_S_invdist_1.00_R_L
bcut_hdonor_S_invdist_0.30_R_H
bcut_haccept_S_invdist_0.50_R_H
bcut_gastchrg_S_invdist2_0.08_R_H
bcut_tabpolar_S_invdist2_1.00_R_L
bcut_tabpolar_S_invdist_0.70_R_H

Fig. 2. The 6 BCUT metrics which best represent the structural diversity of compounds contained in the MDDR database. The three highlighted metrics comprise the receptor-relevant subspace identified for the ACE receptor.

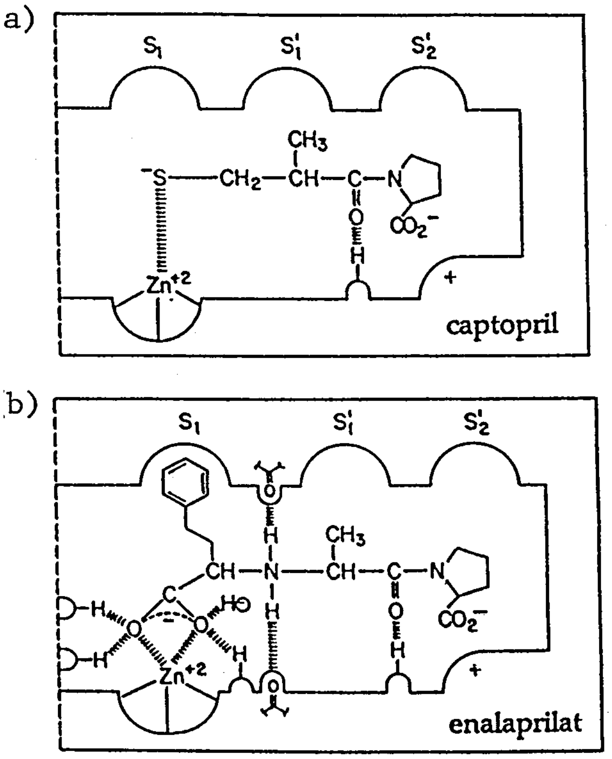

Fig. 3. Cartoon representations of (a) captopril and (b) enalaprilat binding to the ACE receptor as proposed by Wyvratt and Patchett (16). Note that the importance of negative charge and H-bond donor- and acceptor-abilities in the proposed binding models are consistent with the receptor-relevant axes identified in Figure 2.

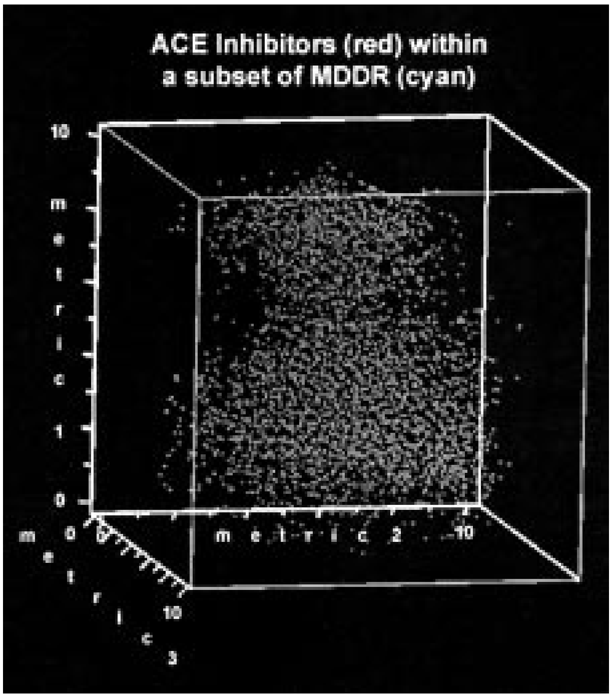

Fig. 4. 74 ACE inhibitors tightly clustered within the 3-dimensional ACE-receptor-relevant subspace of the MDDR (imaginary corporate database) chemistry-space. Only a 5% diverse subset of the total MDDR population is shown so that the actives can be seen within the 3-dimensional "cloud" of MDDR compounds.

Figure 4 shows the 74 ACE inhibitors tightly clustered within the 3-dimensional ACE-receptor-relevant subspace of the MDDR (imaginary corporate database) chemistry-space. In order to see the actives through the "cloud" of roughly 70,000 MDDR compounds, we have shown the positions of just a 5% diverse subset of the total MDDR population. Figure 5 shows the 74 ACE inhibitors tightly clustered within a 2-dimensional projection of the 3-dimensional subspace; all MDDR compounds are shown. Also shown in Figure 5 are the regions of chemistry-space covered by two hypothetical combinatorial libraries. Clearly, determining the receptor-relevant subspace enables computer graphic displays which, in turn, enable visual metric validation (do the metrics really cluster active compounds?), enable visual assessment of diversity (do the imaginary corporate database compounds adequately cover the region of chemistry-space containing active compounds?) and enable visual comparison of compound libraries (to what extent do libraries overlap with each other, with the corporate database and with the region of chemistry-space containing active compounds?). See the recent publication by Schnur (22) for additional examples.

While computer graphic visualization is certainly quite useful, determining the receptor-relevant subspace enables something even more important. It enables the calculation of receptor-relevant distances. Recall the car buyer analogy. What if, for example, the particular car buyer had previously owned a silver Chevrolet and a silver Oldsmobile – both produced by General Motors. A friend helping this particular car buyer choose the next car might think it a waste of time to look at Fords or Hondas unless the friend realized that "distance" (dissimilarity) with respect to brand is an irrelevant component of intercar distance for this particular car buyer. Similarly, imagine that you are helping someone choose which compounds to test (or synthesize) in a second round of screening based on distances from the hits discovered in the first round of screening. If the intercompound distance includes receptor-irrelevant components, compounds which appear to be distant from the known active compounds might not be considered worthy of screening, whereas if the distance was computed based only on receptor-relevant axes, some of those same compounds would appear to be close to the first round leads and, thus, worth screening.

Summary

If properly constructed, high-dimensional (fingerprint) and low-dimensional metrics can provide equally valid representations of chemistry-space for chemical diversity purposes. High-dimensional metrics offer the advantage of providing substantial detail regarding the topological aspects of molecular substructure but suffer the disadvantage that they can be used only for distance-based algorithms for addressing the various diversity-related tasks encountered in pharmaceutical and agrochemical

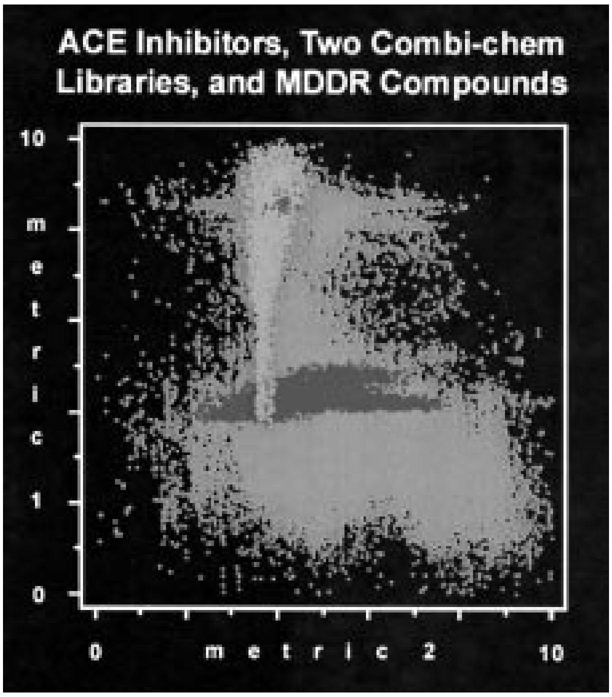

Fig. 5. 74 ACE inhibitors tightly clustered within a 2-dimensional projection of the 3-dimensional subspace. All MDDR compounds are shown. Also shown are the regions of chemistry-space covered by two hypothetical combinatorial libraries.

industry. Low-dimensional metrics offer the advantage of enabling the use of either distance-based or cell-based algorithms which are much better suited for most diversity-related tasks. Traditional molecular descriptors are often cross-correlated, provide little or no substructural information and, thus, are poor choices as low-dimensional chemistry-space metrics. BCUT values constitute a novel class of molecular descriptors which not only encode substructural topological (or topographical) information but also encode atom-based information relevant to the strength of ligand-receptor interaction.

Algorithms have been developed for choosing those low-dimensional metrics which best represent the diversity of a given population of compounds, validating the chosen metrics, and performing all of the diversity-related tasks such as subset selection, library design, compound acquisition and library comparison. Moreover, many diversity-related tasks are greatly facilitated by reference to a receptor-relevant subspace.

No set of metrics – whether they be high-dimensional fingerprints or low-dimensional BCUT values – can possibly describe either the essentially limitless diversity of all possible chemical structures or the subtlety of all structural differences which result in different pharmaceutical or agrochemical activities. All metrics are imperfect. However, despite their imperfections, chemistry-space metrics have proven useful for “improving the odds” and guiding various decisions which must be made during the course of drug and agrochemical discovery efforts. Ongoing research will make software for chemical diversity-related tasks even more useful in the future.

References

1. 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 Comp Sci 1996, 36: 572-84.

2. Pearlman, R.S. DiverseSolutions User’s Manual. University of Texas, Austin, TX, 1995.

3. Pearlman, R.S., Smith, K.M. *Novel software tools for chemical diversity.* Perspect Drug Discov Des 1998, 9-11: 339-53.

4. Burden, F.R. *Molecular identification number for substructure searches.* J Chem Inf Comput Sci 1989, 29: 225-7.

5. Rusinko, A. III, Lipkus, A.H. Unpublished results obtained at Chemical Abstracts Service, Columbus, OH.

6. Pearlman, R.S., Smith, K.M. *Novel metrics and validation of metrics for chemical diversity.* In: Rational Molecular Design in Drug Research. Liljefors, T., Jorgensen, F.S., Krogsgaard-Larsen, P. (Eds.). Munksgaard: Copenhagen 1998, 165-82.

7. CONCORD was developed by R.S. Pearlman, A. Rusinko, J.M. Skell and R. Balducci at the University of Texas, Austin, TX and is distributed by Tripos, Inc., St. Louis, MO.

8. DiverseSolutions was developed by R.S. Pearlman and K.M. Smith at the University of Texas, Austin, TX and is distributed by Tripos, Inc., St. Louis, MO.

9. Weintraub, H.J.R., Demeter, D.A. Personal communication.

10. Pearlman, R.S., Smith, K.M., *Metric validation and receptor-relevant subspace concepts.* J Chem Inf Comput Sci 1998, in press.

11. Pearlman, R.S., Deanda, F. Manuscript in preparation.

12. Modern Drug Data Report database is distributed by Molecular Design Ltd. Information Systems, San Leandro, CA.

13. Sweet, C.S., Ulm E.H., Gross, D.M., Vassil T.C., Stone C.A., *A new class of angiotensin-converting enzyme inhibitors.* Nature 1980, 288: 280-3.

14. Suh, J.T., Skiles, J.W., Williams, B.E. et al. *Angiotensin-converting enzyme inhibitors. New orally active antihypertensive (mercaptoalkanoyl) and [(acylthio)alkanoyl]glycine derivatives.* J Med Chem 1985, 28: 57-66.

15. Menard, P.R., Suh, J.T., Jones, H. et al. *Angiotensin-converting enzyme inhibitors. (Mercaptoaroyl)amino acids.* J Med Chem 1985, 28: 328-32.

16. Wyvratt M.J., Patchett A.A. *Recent developments in the design of angiotensin-converting enzyme inhibitors.* Med Res Rev 1985, 5: 483-531.

17. Karanewsky D.S., Badia, M.C., Cushman, D.W. et al. *(Phosphinyloxy)acyl amino acid inhibitors of angiotensin-converting enzyme (ACE). 1. Discovery of (S)-1-[6amino-2-[[hydroxy(4-phenylbutyl)phosphinyl]oxy]-1-oxohexyl-L-proline novel orally active inhibitor of ACE.* J Med Chem 1988, 31: 204-12.

18. Yanagisawa, H., Ishihara, S., Ando, A. et al. *Angiotensin-converting enzyme inhibitors. 2. Perhydroazepin-2-one derivatives.* J Med Chem 1988, 31: 422-8.

19. Krapcho, J., Turk C., Cushman, D.W. et al. *Angiotensin-converting enzyme inhibitors. Mercaptan, carboxyalkyl dipeptide, and phosphinic acid inhibitors incorporating 4-substituted prolines.* J Med Chem 1988, 31: 1148-60.

20. Karanewsky D.S., Badia, M.C., Cushman, D.W. et al. *(Phosphinyloxy)acyl amino acid inhibitors of angiotensin-converting enzyme. 2. Terminal amino acid analogues of (S)-1-[6-amino-2-[[hydroxy(4-phenylbutyl) phosphinyl]oxy]-l-oxohexyl]-L-proline.* J Med Chem 1990, 33: 1459-69.

21. CombinDBMaker was developed by R.S. Pearlman and E.L. Stewart at the University of Texas, Austin, TX and is available from the authors.

22. Schnur, D. *Computer aided design and diversity analysis of large combinatorial libraries.* J Chem Inf Comput Sci 1998, in press.