Molecular Diversity, **4:** 1–22, 1999.
KLUWER/ESCOM

Advances in diversity profiling and combinatorial series design

Dimitris K. Agrafiotis*, James C. Myslik & F. Raymond Salemme
3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, PA 19341, U.S.A.

Key words: analog design, combinatorial chemistry, experimental design, high throughput screening, QSAR, series design, similarity

Summary

Rapid advances in synthetic and screening technology have recently enabled the simultaneous synthesis and biological evaluation of large chemical libraries containing hundreds to tens of thousands of compounds, using molecular diversity as a means to design and prioritize experiments. This paper reviews some of the most important computational work in the field of diversity profiling and combinatorial library design, with particular emphasis on methodology and applications. It is divided into four sections that address issues related to molecular representation, dimensionality reduction, compound selection, and visualization.

Introduction

There is only atoms and space. The rest is opinion.
Democritus

Diversity is a much sought after commodity these days. Governments, universities and corporations alike espouse the benefits of biological, cultural, intellectual and commercial diversity. The problem is that no two individuals or organizations can seem to agree upon a common definition for this term. What one person perceives a being diverse, another sees as being homogeneous. Maybe that's the point. Diversity, after all, is based upon human perception and is, therefore, inherently subjective. It is a quality rather than a quantity. However, with the advent of combinatorial chemistry and high-throughput screening as tools for massively parallel drug discovery, computational chemists have found themselves in need of quantitative measures that can be used to distinguish between good and mediocre experiments.

Historically, drug discovery has been based on a serial and systematic modification of chemical structure guided by the 'similar property principle', i.e. the assumption that structurally similar compounds tend to exhibit similar physicochemical and biological properties. New therapeutic agents are typically generated by identifying a lead compound, and creating variants of that compound in a systematic and directed fashion. The first phase of this process, known as *lead generation*, is carried out by random screening of large compound collections, such as natural product libraries, corporate banks, etc. The second, known as *lead optimization*, represents the rate-limiting step in drug discovery, and involves the elaboration of sufficient SAR around a lead compound and the refinement of its pharmacological profile. Prior to the arrival of combinatorial chemistry, this process involved a simple prioritization of synthetic targets based on preexisting structure-activity data, synthetic feasibility, experience, and intuition.

However, revolutionary advances in synthetic and screening technology have recently enabled the simultaneous synthesis and biological evaluation of large chemical libraries containing hundreds to tens of thousands of compounds. These tools haven't changed the fundamental way in which drugs are discovered, but they have changed the way in which chemists contemplate experiments. With the expansion of our knowledge base of solid and solution-phase chemistry and the continuous improvement of the underlying robotic hardware, combinatorial chemistry has moved beyond its traditional role as a source of compounds

* To whom correspondence should be addressed.

for mass screening, and is now routinely employed in lead optimization and SAR refinement. This has led to the conceptual division of combinatorial libraries into 1) *exploratory* or *universal* libraries which are target-independent and are designed to span a wide range of physicochemical and structural characteristics, and 2) *focused* or *directed* libraries which are biased against a specific target, structural class, or known pharmacophore.

Most work to date has focused on lead generation, using molecular diversity as a means to design and prioritize experiments. The need for careful experimental design became apparent as the initial euphoria that accompanied the advent of this exciting new technology gave way to the realization of its practical limitations. Indeed, it is clear that at least for now the number and types of molecules that can be synthesized using parallel synthesis still represents a small fraction of all the compounds of potential therapeutic interest, and that the quality of the biological response degrades rapidly with throughput. Thus, the synthesis and testing of a chemical library should ideally be planned in a way that renders maximum information about the underlying biological target. To achieve this goal, it is no longer sufficient to examine the properties of individual compounds, but it is also important to assess the *collective* quality and information content of our libraries. In this respect, molecular diversity can be viewed as one design strategy for maximizing the hit-rate of high-throughput screening experiments. It represents a generalization of the concept of molecular similarity from individuals to collections, and its effective use can reduce the redundancy and cost of experiments and substantially increase the odds of discovering new drugs. It is closely related to molecular similarity, structure-activity correlation and statistical series design, which are thoroughly reviewed in numerous texts [1,2]. The scale and complexity of the problem is daunting for all but the most gifted chemists, and makes it ideally suited to computation.

This paper reviews recent advances on computational aspects of molecular diversity, with particular emphasis on methodology and applications. There are probably as many different definitions of diversity as the number of researchers who are active in the field, and it is our belief that theoretically the subject borders on religion. However, many of the concepts described herein are now routinely employed in combinatorial library design, and the choice of methods has definitive and measurable practical implications. The article is divided into three main sections that address issues related to molecular representation, compound selection, and visualization. The reader is also referred to various reviews [3–5,70].

Molecular encoding

> *Rabbi Raditz of Poland was a very short rabbi with a long beard, who was thought to have inspired many pogroms with his sense of humor. One of his disciples asked,*
>
> *'Who did God like better – Moses or Abraham?'*
>
> *'Abraham,' the Zaddik said.*
>
> *'But Moses led the Israelites to the Promised Land,' said the disciple.*
>
> *'All right, so Moses,' the Zaddik answered.*
>
> *'I understand rabbi, it was a stupid question.'*
>
> *'Not only that, but you are stupid, your wife's a meeskeit, and if you don't get off my foot you are excommunicated.'*
>
> Woody Allen, Getting Even.

Conceptually, the problem of quantifying molecular diversity involves two parts: the first is the definition of chemical distance, and the second is the selection of a representative set of compounds from a (typically) much larger collection. This section reviews a number of methods that have been proposed to encode molecular structures in a way that is suitable for numerical processing. It is organized in three parts, which discuss descriptors derived from the topology, three-dimensional structure, and physicochemical and electronic properties of the molecules.

Two-dimensional descriptors

Molecular connectivity indices

When asked to describe a molecule, a chemist will instinctively reach for a pen and some paper and draw a picture. The picture shows how constituent atoms are interconnected to form the molecule (i.e. the topology of the molecule). Mathematically, this is equivalent to a molecular graph. Molecular connectivity or topological indices are numerical values calculated from certain invariants (characteristics) of a molecular graph [13,14] which encode features such as number of atoms, branching, ring structures, heteroatom content, and bond order. They are attractive for quantifying molecular diversity because they are inexpensive to compute, and have been validated through years of use in the field of structure-activity correlation. In particular, the widespread availability of Kier and Hall's Molconn-X program [15], which calculates indices based on connectivity, shape, sub-graph counts,

topological equivalence, electrotopological state and information content, has done much to promote the use of molecular connectivity indices.

Topological indices may be classified into four groups based on their logical derivation:

1) Those derived from the adjacency matrix: the total adjacency index, the Zagreb group indices, the Randic connectivity index, the Platt index, the compatibility code, and the largest eigenvalue index;
2) Those based on the topological distance matrix, including the Wiener index, the polarity number, the distance sum, the Altenburg polynomial, the mean square distance, the Hosoya index, and the distance polynomial;
3) Centric indices, including the generalized graph center;
4) Information-theoretic indices, including the Shannon index, the chromatic information index, the orbital information index, the topological information superindex, the electropy index, and the Merrifield and Simmons indices [13,14].

Binary descriptors

Binary descriptors include substructure keys and hashed fingerprints. Substructure keys encode molecular structures as bitstrings, each binary digit of which indicates the presence or absence of a selected structural feature or pattern. Typical target features might include the number of occurrences of a particular element (e.g. the presence of at least 1, 2 or 3 nitrogen atoms), electronic configurations or atom types (e.g. sp^2 nitrogen or aromatic carbon), common functional groups such as alcohols, amines etc., and ring systems. Features that are rare enough not to merit an individual bit, yet extremely important when they do occur, are assigned a common bit which is set if any one of the patterns is present in the target molecule (a 'disjunction'). Substructure keys were originally developed for rapid searching of large databases, but have also proven effective in similarity applications. Generating substructure keys is time-consuming, requiring a substructure search for each target pattern in every molecule in a database. However, once generated, they allow a database to be searched by means of Boolean operations upon the keys; a process performed very rapidly by digital computers.

Database designers tailor substructure keys to minimize molecule retrieval time. As such, the choice of encoded structural features tends to be application-specific. For example, keys employed in drug databases encode functional groups of particular interest to medicinal chemists, while those used in databases of organometallic compounds contain features related to metal-carbon bonding. Despite this specificity, substructure keys contain sufficient information about the molecular structures to permit meaningful similarity comparisons.

Like structural keys, hashed fingerprints are bitstrings derived directly from the connection table and were developed primarily for database searching. They differ from structural keys in that they do not depend on pre-selected structural fragments to perform the bit assignment. Instead, every pattern in the molecule up to a predefined path length is systematically enumerated. The resulting set of patterns serves as input to a hashing algorithm that turns 'on' a small number of bits at pseudo-random positions in the bitstring. Because the number of possible patterns far exceeds the length of the fingerprint, many patterns are mapped onto a single bit. In practice, this does not pose a problem for database searching. Every bit that is set in a fingerprint of the target pattern will also be set in that of a molecule which contains the pattern, making database screening deterministic and fast. For similarity comparisons, this is not the case. While two different molecules may have exactly the same fingerprint, the probability of this occurring is extremely small for all but the simplest cases. Experience has shown that the similarity between two fingerprints is a good indicator of the similarity between the two structures. Hashed fingerprints have the additional characteristic that as structures become more complex the density of encoded information increases. A number of studies have shown that fingerprints and substructure keys are equally effective for the purpose of diversity analysis (see section Descriptor Validation below).

While most of the descriptors require an enumerated structure (i.e. the full connection table of a molecule), for combinatorial libraries, this is not strictly necessary. Downs and Barnard [16] have recently presented an elegant method to compute molecular fingerprints based on reaction precursors, using techniques developed for Markush structure handling in chemical patents.

A number of similarity metrics have been proposed for binary descriptors [6]. The most frequently used ones are the normalized Hamming distance:

$$H = \frac{|\mathrm{XOR}(x, y)|}{N} \qquad (1)$$

where x and y are two binary sets (encoded molecules), XOR is the bitwise exclusive OR operation (a bit in the result is set if the corresponding bits in the two operands are different), and N is the number of bits in each set. The result, H, is a measure of the number of bits that are dissimilar in x and y; the Tanimoto or Jaccard coefficient:

$$T = \frac{|\mathrm{AND}(x, y)|}{|\mathrm{IOR}(x, y)|} \tag{2}$$

where AND is the bitwise AND operation (a bit in the result is set if both of the corresponding bits in the two operands are set) and IOR is the bitwise inclusive OR operation (a bit in the result is set if the either of corresponding bits in the two operands are set). The result, T, is a measure of the number of substructures shared by two molecules relative to the ones they *could* have in common; and the Dice coefficient:

$$T = \frac{2|\mathrm{AND}(x, y)|}{|x| + |y|} \tag{3}$$

Another popular metric is the Euclidean distance which, in the case of binary sets, can be recast in the form:

$$E = \sqrt{N - |\mathrm{XOR}(x, \mathrm{NOT}(y))|} \tag{4}$$

where NOT(x) denotes the binary complement of x, and the expression |XOR(x, NOT(y))| represents the number of bits that are identical in x and y (either 1's or 0's). The Euclidean distance is a good measure of similarity when the binary sets are relatively rich, and is mostly used in situations in which similarity is measured in a relative sense. Of all the indices defined above, Tanimoto is perhaps the most commonly used. As we will see below (section Descriptor Validation), these simple binary descriptors, when used with an appropriate clustering methodology, have been surprisingly successful in discriminating active from inactive compounds.

Molecular holograms

Molecular holograms are an extension of hashed fingerprints that are based on fragment counts rather than simple yes/no information alone. As such, they are encoded as relatively short vectors of integers rather than bitstrings. As a consequence, much more information about fragment patterns in the molecule is retained upon hashing. Holograms can also incorporate structural information that ordinary fingerprints can not, such as branching, cyclic structures, hybridization patterns and chirality [74]. They have proven quite useful in QSAR studies [75] and once a QSAR model is in hand, can be used to screen databases of compounds which are likely to be active.

Atom pairs and topological torsions

Atom pairs and topological torsions are two related types of descriptors which represent another attempt to eliminate the subjectivity inherent in substructure keys [17]. Atom pairs are patterns of the form $a_i - d - a_j$, where a_i and a_j are the types of atoms i and j, respectively, and d is the topological distance between the atoms (the number of bonds along the shortest path connecting these atoms). A molecule with N atoms has $N(N-1)/2$ atom pairs, although some of them may not be unique. Topological torsions take the form $a_i - a_j - a_k - a_m$, where i, j, k and m are sequentially bonded atoms, and a_i is again the type of the i-th atom. Originally, atoms were assigned types based on atomic number, number of neighbors and number of π-electrons. More recent studies have utilized physicochemical properties [18] and geometric features [19] as well. Physicochemical atom pairs include binding properties, atomic log$P$ contributions and partial atomic charges. Binding properties categorize atoms into seven classes: anions, cations, neutral hydrogen bond donors, neutral hydrogen bond acceptors, polar atoms, hydrophobic atoms, and other. Geometric atom pairs are similar to the regular atom pairs with the exception that the topological distance is replaced by the through-space distance of the corresponding atoms in some low-energy conformation of the target molecule. Geometric and topological atom pairs are equally effective in similarity searching, but the new generation of descriptors seem to perform worse than the original ones in their overall ability to discriminate biologically active from inactive compounds. As one would expect, which set does better than another varies greatly from probe to probe, and is very difficult to predict *a priori*.

The similarity between two structures described by atom pairs or topological torsions is measured by:

$$s(i, j) = \frac{\sum_{k=1}^{K} \min(f_{ik}, f_{jk})}{0.5\left[\sum_{k=1}^{K}(f_{ik} + \sum_{k=1}^{K} f_{jk}\right]} \tag{5}$$

where f_{ik} is the count of the k-th descriptor in the i-th structure, and K is the union of all unique descriptors in i and j. This index ranges from 0 to 1, with 1 indi-

cating complete identity and 0 indicating that the two structures have nothing in common.

Atom layers

Martin et al. [8] developed the concept of atom layers in an attempt to account for the topological distribution of chemical features around a combinatorial core that are believed to play a critical role in receptor binding. Atom layers are based upon the premise that atoms of a substituent which are close to the point of attachment to the core contribute differently to binding than those that are more distant. They are constructed by summing a given property over all atoms in a substituent at a given number of bonds distant from the attachment point. Properties considered by the Chiron group were atomic radius, acidity, basicity, the ability to serve a hydrogen bond donor or acceptor, and aromatic character. The similarity between two substituents was computed by comparing the corresponding atom layer tables element by element, and dividing the sum of the minimum by the sum of the maximum values in each cell.

2D autocorrelation vectors

Moreau [20] has proposed an autocorrelation function to encode the topology of a molecular graph:

$$A(d) = \sum_{i,j} p_i p_j \qquad (6)$$

where p_i and p_j represent the values of an atomic property at atoms i and j, respectively, and d is the topological distance between the two atoms measured in bonds along the shortest path. The function has the useful property that no matter how large and complex the molecule, it can be encoded in a fixed-length vector of small rank. Typically, only path lengths of 2 to 8 are considered. Atomic properties that have been encoded using this method include volume, electronegativity, hydrogen bonding character and hydrophobicity. Originally, a separate autocorrelation vector was computed for each property, and the resulting set was reduced into a smaller number of variables using principal component analysis.

Topological autocorrelation vectors were also used by Gasteiger [21] as input to a Kohonen network which successfully separated dopamine from benzodiazepine receptor agonists, even when these compounds were embedded in a large and diverse set of chemicals extracted from a commercial supplier catalog. Gasteiger [22] has also extended the concept to three dimensions, by introducing a spatial autocorrelation vector based on properties measured on the molecular surface. These spatial autocorrelation vectors were used to model the activity of 31 steroids against the corticosteroid binding globulin, and the cytosolic Ah receptor activity of 78 polyhalogenated aromatic compounds. These descriptors were also found to be effective in describing the diversity of combinatorial libraries, also through the use of Kohonen networks (see section Self-Organizing Maps below) [23].

B-Cut values

B-Cut values were developed by Pearlman [24] as an extension of Burden's concept of molecular identification numbers [25] to multiple dimensions. Burden represented the hydrogen-suppressed connection table of a molecule as a symmetric NxN matrix in which atomic numbers were placed in the diagonal elements and the off-diagonal elements were assigned a value of 0.1 times the nominal bond order if the corresponding atoms were bonded or 0.001 if not. An additional score of 0.01 was added to the (off-diagonal) terminal nodes. A molecular identification number was then calculated from the two lowest eigenvalues of the matrix. Rusinko and Lipkus used molecular identification numbers for similarity searching of a 60 000-membered subset of the CAS registry. They found this method compared well with results obtained from an established similarity searching procedure.

Pearlman broadened this approach to include properties deemed significant to protein-ligand binding. Three classes of matrices were constructed placing atomic charge, polarizability and hydrogen bonding characteristics in the diagonal elements and a combination of interatomic distances, overlaps, computed bond orders in the off-diagonal elements. A six-dimensional property space was then defined using the lowest and highest eigenvalues (B-Cut values) of a representative matrix from each of these three classes. Optimal combinations of on- and off-diagonal properties were selected by their ability to produce a uniform distribution of molecules in the property space, as determined by a χ^2 criterion. The resulting six-dimensional space is small enough to permit diversity analysis based on partitioning (binning). This is described in greater detail below (section Partitioning Techniques). Pearlman concluded that B-Cut values based solely on the connection table were proven satisfactory for most diversity profiling tasks.

Three-dimensional descriptors

3D structural keys

Geometric structural keys are direct analogues of topological (two-dimensional) substructure keys in three dimensions. They are binary sets tailored to minimize compound retrieval time from three-dimensional molecular databases [26,27]. Typically, the keys encode the Euclidean or angular distance between pairs of selected features such as atom types, centroids of aromatic rings, ring normals, or attachment points of functional groups. The distance between the members of the pair is divided into a fixed number of ranges and a bit is assigned for each range. In forming a key, if a molecule contains a pair of selected features, the distance between the members of the pair is calculated and the bit corresponding to the range into which the distance falls is set.

The success of 3D structural keys, or any three-dimensional descriptors, depends upon their ability to account for the possibility of multiple molecular conformations. Early implementations of three-dimensional molecular database systems stored molecules as single low-energy conformations, determined experimentally through X-ray crystallography or calculated using a fast structure-generation software package [28]. Later implementations employed conformational search procedures, but, in the interest of speed, these searches were relatively crude and did not rule out highly improbable conformations. Other problems associated with 3D structural keys are poor representation of shape and chirality, and the limit to the number of features that can be encoded in a finite length string. Shape and chirality are concepts that are undefined in the single dimension defined by a pair of features. To overcome these limitations, several groups have developed a related class of descriptors defined by sets of three or four selected features, known as 3D pharmacophore keys.

3D pharmacophore keys

Pharmacophore keys, introduced by Sheridan and coworkers [29], are 3D structural keys which incorporate features of functional importance to macromolecular recognition including hydrogen bond donors and acceptors, centers of positive charge, aromatic ring centers and hydrophobic centers. A pharmacophore is defined as a combination of 3 or 4 such loci, forming a triangle or tetrahedron, respectively, and is characterized by the set of distances between the loci. As with structural keys, the distances are divided into ranges, and a bit is assigned to each range.

As is the case for structural keys, pharmacophore keys can be readily extended to account for multiple conformations. Additionally, because pharmacophores are two- and three-dimensional objects, they are able to capture information on molecular shape and chirality. Three-point pharmacophore keys also lend themselves well to visualization *via* three-dimensional scatter plots (see section Visualization without Dimensionality Reduction below). Sheridan's original work has been extended by a number of groups, most notably those at Chemical Design [27], Rhone-Poulenc [30], and Abbott [10]. Davies and Briant [31] have employed pharmacophore keys for similarity/diversity selection using an iterative procedure that takes into account the flexibility of the compounds and the amount of overlap between their respective keys (see section Boolean Logic).

Molecular hashkeys

Defining and measuring the three-dimensional surface properties and similarities of molecules are also the key to MetaXen's approach for predicting biological properties of molecules as well as in determining the diversity of a population of molecules. Since determining pairwise surface similarity measurements for large sets of molecules is computationally prohibitive, Sage and co-workers have extrapolated the similarity measure to a novel representation of molecular surface properties which they called a molecular hashkey [76]. A molecular hashkey is a real-valued vector of fixed dimension, that is used to represent information about the surface properties of a molecule. The molecular hashkey is much smaller than a complete 3D surface representation of the molecule. The term 'hashkey' is borrowed from computer science, and represents a compact numerical representation of an object that is used to solve indexing problems by storing objects using their hashkeys as memory addresses. A molecular hashkey has the property that molecules with similar hashkeys will appear similar based on observation of their surfaces. Molecules with identical surface properties will have identical hashkeys, independent of the underlying atomic scaffolding. Given a molecule M, its N-dimensional hashkey (H_1, H_2, ..., H_N) is computed by calculating its molecular surface similarity to a set of N basis molecules. The basis molecules are in arbitrary fixed conformations. M is flexibly aligned to each B_i of the basis molecules ($B_{1...N}$) to maximize molecular similarity, and the best match yields

the similarity value that becomes H_i. The molecular hashkey has been used successfully in combination with machine-learning techniques for developing predictive models of biological properties of molecules. The hashkey technique has also been used to create diverse subsets of molecules from a starting population by maximizing the molecular hashkey distance between the molecules in the subset.

Physicochemical and electronic descriptors

Physicochemical properties

Physicochemical properties have long been used to develop structure-activity relationships. They quantify a large number of molecular characteristics known to determine the transport and binding of a drug to its target. It is natural, then, that the first attempts to quantify molecular diversity were based upon physicochemical properties. These properties can be calculated using standard molecular modeling and quantum mechanical packages. They include the number of filled orbitals, HOMO and LUMO energies, standard deviation of partial atomic charges and electron densities, dipole moment, ionization potential, heat of formation, total energy, molecular weight, octanol-water partition coefficient (logP), molar refractivity, van der Waals volume and surface area, and many others. Molecular property descriptors have been used for diversity profiling by Willett et al. [6,7], Martin et al. [8], Lewis et al. [9], Brown et al. [10,11], and many others. They have been extensively reviewed by Kubinyi [12].

Electronic fields

Cramer and co-workers [33] have developed descriptors for combinatorially generated molecules based upon steric field methods employed in 3D-QSAR [32]. Their procedure attempts to find a representative conformation for each substituent group pendant upon a particular point of variation of a combinatorial template. The process begins with a low energy conformation generated by a model-building routine, which is then fitted as a rigid body onto the combinatorial template using least-squares minimization. The torsional angles of the rotatable bonds within the substituent are then sequentially adjusted, starting from the bond closest to the template, using a simple set of topological precedence rules. Once aligned, the steric field of the substituent group is computed using a CoMFA-like approach. Two compounds may be compared, for example, by calculating the root-square-differences in steric field values summed over all lattice points in the CoMFA region.

These descriptors have been used to classify 736 commercially available thiols into 231 bioisosteric clusters, consistent with results obtained using molecules encoded as 2D fingerprints and compared with the Tanimoto similarity coefficient. While such an alignment procedure has utility in comparing members of a single combinatorial library, it is unclear how it could be applied to comparison of compounds which belong to different libraries, to heterogeneous compound collections, or to libraries having a variable template.

Affinity fingerprints

Affinity fingerprints represent an entirely different class of molecular descriptor than those previously presented. They are based upon the measured affinity of a molecule for a set of target proteins rather than the structure of a molecule itself. The resulting vector of affinities – the affinity fingerprint – may be used in the same way as structure-derived descriptors to discern similarities between molecules and to quantify molecular diversity. This functional approach was pioneered by Terrapin [34], based upon the company's fluorescent polarization high-throughput screening technology. The significance of the results depends critically upon the selection of an appropriate basis set of proteins. The set of proteins must be able to recognize a wide variety of organic compounds and its members binding specificities must be uncorrelated. A systematic analysis of several hundred candidates resulted in a set of 18 proteins, which are now used routinely at Terrapin for screening new compound collections.

Affinity fingerprints provide an empirical way to assess the molecular diversity of a chemical library. Kauvar [34], for example, suggests that an estimate of the diversity of a given set of compounds can be determined based on the maximum separation and most frequently occurring distance between two affinity vectors in the collection. Although this is a rather qualitative measure, the mathematical nature of the affinity fingerprint makes possible more quantitative measures of diversity (see section Compound Selection). From the standpoint of drug design, affinity fingerprints have one significant limitation: they do not have the ability to predict the affinity profile of an untested class of compounds. It is, however, conceivable that it may be possible to predict the biological profiles of tested compounds against new protein targets through traditional regression techniques.

Validation

While no generally accepted theoretical definition for molecular diversity exists, there is agreement on the criterion for success in choosing descriptors and diversity metrics: a successful choice should be able to discriminate between biologically active and inactive compounds. The most comprehensive study specifically designed to address this issue was reported by Brown and Martin [10] of Abbott Laboratories. What makes their contribution unique is the size of the data set used in their analysis. More than 20 000 structures were analyzed, including three different sets of compounds tested against monoamine oxidase and two other proprietary enzyme targets, and a collection of over 16 000 compounds that were tested over the years in Abbott's high-throughput screens. Seven types of descriptors (MACCS, Unity and Daylight fingerprints, Unity 3D rigid and flexible descriptors, and two pharmacophore descriptors developed at Abbott), and 4 different clustering methodologies (Jarvis-Patrick, Ward, group-average and Guenoche) were evaluated based upon their abilities to map active and inactive compounds to different regions of descriptor space. The results indicated that the two-dimensional descriptors (the fingerprints) were considerably more effective than the three-dimensional ones. Among the clustering techniques, Ward's hierarchical agglomerative algorithm prevailed. In a follow-up study, they found that this ranking held for the ability of these descriptors to predict individual receptor-ligand binding determinants such as hydrophobicity, dispersion, electrostatics, and steric and hydrogen bonding capability. These results were consistent with those reported previously by the Sheffield group [7].

Patterson et al. [35] reported an alternative method for validating descriptors based on the concept of a 'neighborhood radius'. Their approach was to plot differences in the value of a descriptor against those in biological activity for a set of related compounds. If the descriptor is to be useful as a measure of similarity, the resulting plot should exhibit a characteristic trapezoidal distribution revealing a 'neighborhood behavior' for that descriptor. The method was applied to 20 data sets, and 11 descriptors were ranked by performance. They concluded that 2D fingerprints and 3D CoMFA fields far outperformed physicochemical properties such as logP and molar refractivity, while topological descriptors such as connectivity indices, atom pairs and auto-correlation vectors fell in the middle of the spectrum. Interestingly, they also found that, for combinatorially generated molecules, 2D fingerprints based on the whole molecule performed worse than those based on the substituents alone. They attributed this to a 'diluting' effect due to the presence of an identical template in each molecule. However, Patterson's study considered one descriptor at a time, and did not account for the possibility of correlations between two or more descriptors.

Dimensionality reduction

The high-dimensional data representations that are commonplace in molecular diversity/similarity analyses pose a number of problems. Firstly, as the number of variables used to describe data increases, the likelihood that some of the variables are correlated dramatically increases. While certain applications are more sensitive to correlation than others, in general, redundant variables tend to bias the results. Secondly, the amount of the computational effort needed to perform the analysis increases in proportion to the number of dimensions. Finally, visualization of the results in a concise and intuitive manner rapidly becomes impossible.

Fortunately, most multivariate data in $\Re^d$ are almost never d-dimensional. That is, the *underlying structure* of the data is almost always of dimensionality lower than d. To simplify the analysis and representation of the data, it is often desirable to reduce the dimensionality of the space by eliminating dimensions that add very little to the overall picture. We must stress that none of the methods that will be discussed here guarantees to extract the most important features for the application at hand. There is always the possibility that some critical piece of information is left behind, buried under a pile of redundancies. Experience in many different application areas has shown that, in practice, this situation does not arise often.

This discussion will focus on three main techniques to perform the reduction: 1) principal component analysis, 2) factor analysis, and 3) multidimensional scaling. Other approaches such as nonlinear mapping and Kohonen networks are discussed in greater detail in the section Visualization below.

Principal component analysis

Principal component analysis (PCA) [8,40–41] takes as its input a set of vectors described by partially cross-correlated variables and transforms it into one described by a smaller number of orthogonal variables

(principal components) without a significant loss in the variance of the data. Principal components correspond to the eigenvectors of the covariance matrix, m_{ij}, a square symmetric matrix that contains the variances of the variables in its diagonal elements and the covariances in its off-diagonal elements:

$$m_{ij} = m_{ji} = \frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \mu_i)(x_{kj} - \mu_j) \quad (7)$$

where μ_i is the mean value of variable i:

$$\mu_i = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \quad (8)$$

and N is the number of observations in the data set. The eigenvalues of this matrix represent the variances of the principal components. PCA achieves a reduction in dimensionality by filtering out the principal components which contribute the least to the variance of the data (i.e. those with the smallest eigenvalues) until the variance reaches some pre-defined threshold, typically 90–95% of the original value. Finally, the original data are transformed using Equation (9):

$$\mathbf{x}' = \mathbf{V}^T\mathbf{x} \quad (9)$$

where $\mathbf{V}^T$ is the transpose of the filtered eigenvector matrix, $\mathbf{x}$ is the input vector in the original coordinate frame, and $\mathbf{x}'$ are the coordinates of that sample in the transformed frame. The components of $\mathbf{x}'$, therefore, are linear combinations of the original, cross-correlated variables.

The main advantage to PCA is that it makes no assumptions about the underlying probability distributions of the original variables. The primary disadvantage is that it is sensitive to outliers, missing data and poor correlations between variables due to poorly distributed variables.

Factor analysis

Factor analysis (FA) is a closely related technique that attempts to extract coherent subsets of variables that are relatively independent from one another [40]. It is often the case in science that the variable we are interested in is not directly observable. However, it is often possible to measure other quantities that reflect the underlying variable of interest. Factor analysis is an attempt to explain the correlations between variables in the form of underlying factors, which are themselves not directly observable, and which are thought to be representative of the underlying process that has created these correlations.

Factors are linear combinations of original variables. They may be associated with two or more of these variables (*common factors*) or with a single variable (*unique factors*). The specific association between the original variables and the derived factors is described in the form of *loadings*, which are derived from the magnitude of the eigenvalues of the covariance matrix. Factor loadings are inherently indeterminate. Rotation attempts to put these factors into a simple position, so that each variable is loaded highly on one factor, and all factor loadings are either large or near zero. A number of different rotation methods are available, including varimax, quartimax, and equimax. The varimax method maximizes the variance of the loadings, and is the most widely used.

On the surface, factor analysis and principal component analysis are very similar. Both rely on an eigenvalue analysis of the covariance matrix, and both use linear combinations of variables to explain a set of observations. However, in PCA the quantities of interest are the observed variables themselves; the combination of these variables is simply a means for simplifying their analysis and interpretation. Conversely, in factor analysis the observed variables are of little intrinsic value; what is of interest is the underlying factors.

Factor analysis has been used by Cummins et al. [42] to reduce a set of 61 molecular properties to 4 factors, which were then used to compare the diversity of 5 chemical databases (see section Partitioning Techniques below). It was also explored by Gibson et al. [41] in a comparative study of 100 different heterocyclic aromatic systems, but they concluded that FA did not reduce the complexity of the analysis, and did not offer any significant advantages over PCA.

Multi-dimensional scaling

Multi-dimensional scaling (MDS) [65,66] is a positional refinement technique that attempts to map a set of points in a high-dimensional space onto one of lesser dimensionality while preserving, as best as possible, the pairwise Euclidean distances, d_{ij}, between the points. Each iteration of the procedure consists of calculating the distances, δ_{ij}, between each pair of points in a lower-dimensional trial configuration and, using a steepest descent algorithm, shifting the positions of those points so as to create a new configuration characterized by a smaller sum-of-squares difference

between δ_{ij} and d_{ij}. Two commonly used objective functions are *Kruskal's stress*:

$$S = \sqrt{\frac{\sum_{i<j}(\delta_{ij} - d_{ij})^2}{\sum_{i<j}\delta_{ij}^2}} \qquad (10)$$

and Lingoes' *alienation coefficient*:

$$S = \sqrt{1 - \frac{\sum_{i<j}(\delta_{ij} \cdot d_{ij})^2}{\sum_{i<j} d_{ij}^2}} \qquad (11)$$

The procedure terminates when the change in the objective function between iterations falls below a user-defined threshold.

Using multi-dimensional scaling, the group at Chiron [8] has shown that the 2048-bit Daylight fingerprints associated with 721 commercially available primary amines could be reduced to only five continuous variables that reproduced all 260 000 original pairwise dissimilarities (distances) with a standard deviation of only 10%. Similarly, only seven dimensions were necessary to represent the 642 000 pairwise similarities among a set of 1133 carboxylic acids and acid chlorides to the same precision. Despite its successes, the substantial computational cost of traditional multidimensional scaling makes the technique inapplicable to large data sets, such as those encountered in combinatorial library designs.

Compound selection

> *'There is nothing wrong with shooting, just as long as the right people get shot.'*
>
> Clint Eastwood, Magnum Force.

Clustering

Clustering algorithms attempt to organize elements of a set into groups (clusters) based upon selected characteristics. Members of a cluster must be similar to one another (internally homogeneous) and dissimilar to members of other clusters (externally heterogeneous). Because of its long-standing application in determining molecular similarity, clustering was one of the first selection methods to be applied to diversity analysis [6]. In contrast to other selection algorithms, which are based on statistical theory, clustering is an entirely heuristic approach consisting of four principal steps. First, a set of molecular descriptors must be selected and scaled. Second, the distances between pairs of molecules in the collection are calculated and collected in a similarity matrix. Third, members are assigned to clusters by a set of user-defined criteria. Finally, the clustering is validated by visual inspection or statistical means.

Clustering algorithms may be classified as hierarchical or non-hierarchical based on the way in which the clusters are formed. The end-result of hierarchical clustering analysis is a tree, or dendrogram, the structure of which reflects the organization of all the members of the collection. The dendrogram may be created from the top down beginning with a single cluster which is recursively sub-divided into increasingly smaller groups until each member is a cluster unto itself (a 'singleton'). Alternatively, one could begin with singletons and work up the tree by combining clusters until all the members belong to a single group. Non-hierarchical clustering (also known as k-nearest-neighbor clustering) produces a set of clusters based on some user-defined criteria. The most commonly used member of this class, particularly for diversity applications, is the Jarvis-Patrick algorithm. This method begins by determining the k nearest neighbors of each member of the collection. Members are placed in the same cluster if they have a user-defined number of nearest neighbors in common. The major advantage of this method is its speed; its main disadvantage is its tendency to generate either too many singletons or too few very large clusters depending on the stringency of the clustering criteria.

Willett [7] has performed a systematic evaluation of four different clustering methodologies – the Ward and group-average hierarchical agglomerative methods, the minimum diameter polythetic hierarchical divisive method, and the Jarvis-Patrick nearest neighbor algorithm – for purposes of determining molecular similarity. The test set was comprised of 5982 compounds characterized by 13 molecular descriptors. The results were evaluated by means of simulated property prediction experiments. Willett concluded that any of the three hierarchical methods was preferred to the Jarvis-Patrick (non-hierarchical) algorithm. A subsequent study by Brown and Martin has confirmed these findings [10].

Regardless of the particular algorithm used to cluster the collection, a diverse set is typically created by selecting one member from each cluster (most commonly the centroid). The resulting set may be examined for possible colinearities. If it is found to be non-orthogonal, suspect compounds are replaced with other members of the same cluster, and the new

solution is re-evaluated. This cycle continues until a quasi-orthogonal set is identified.

Maximin

The maximin algorithm begins with a randomly selected member of a set of compounds and builds a maximally diverse subset, one compound at a time. At each step, the compound added to the subset is that which is farthest from its nearest neighbor in the subset. While this selection method is conceptually straightforward and easy to implement, it is impractical for use with large collections of compounds since the number of operations scales to the square of the size of the set. Maximin was first applied to diversity selection by Lajiness [45], and has been extended by others including Polinsky [46]. Hassan [43] and Agrafiotis [44] used the maximin criterion itself as a diversity metric that was optimized via simulated annealing.

Chapman [47] has recently reported a maximin algorithm which employs a diversity metric that explicitly accounts for multiple molecular conformations. Each molecule of a collection is subjected to an exhaustive search to identify its lowest energy conformations which, in turn, are aligned with the lowest energy conformations of every other molecule in the collection. Through rigid body rotations, the alignment procedure attempts to maximize a similarity coefficient which is a measure of the extent of steric and charge overlap between the two conformations. At each step of the procedure, the molecule that most increases the diversity of the selected subset, as computed using the following equation, is chosen:

$$D(M) = \sum_{m \in M} \left[\left(\sum_{c \in C(m)} \min_{c' \in C} (d(c, c')) \right) - T\Delta S(m) \right] \quad (12)$$

where M is the set of all compounds, $C(m)$ is the set of all conformers of compound m, and C the set of all conformers of all compounds. $T\Delta S(m)$ is an entropic term proportional to the number of rotatable bonds that penalizes highly flexible compounds.

Chapman applied this approach to two test collections, one of which consisted of naturally occurring amino acids and the other of 1371 commercially available carboxylic acids. In the case of the amino acids, the results were intuitive and suggested that the measure of similarity is indeed a reasonable one. In the case of the carboxylic acids, the selection compared favorably to random controls, and identified reagents that were quite diverse in terms of shape, size and functionality. While the method is intellectually robust and intuitive, it remains to be seen whether the markedly increased computational cost of taking multiple conformations into consideration is offset by any advantage over existing techniques.

Stepwise elimination

Taylor [48] has developed a selection method which sequentially eliminates members from the whole set rather than building up the diverse subset from a single compound. Starting with the symmetric $N \times N$ similarity matrix, the largest off-diagonal element (i.e. the most similar pair of compounds) is identified, and one of the pair of compounds associated with it is eliminated. This process continues until a single compound is left in the set. This algorithm then sorts the compounds, placing the most diverse molecules at the top of the list.

Cluster sampling

Cluster sampling is another nearest neighbor selection algorithm developed by Taylor [48] which, despite its name, does not explicitly partition a set of compounds into clusters. Using a minimum similarity threshold of 0.8, the method begins by generating a list of nearest neighbors for each compound in the set, which are then merged to form the nearest neighbor table (NNT) for the entire set. During each iteration, the procedure selects the compound which occurs most often in the NNT, which corresponds to the compound situated at the center of the most densely populated region (cluster) of property space. All the nearest neighbors of this molecule are then flagged as unavailable for subsequent cycles of selection. The procedure terminates when all the compounds in the set are either selected or flagged as unavailable. Both cluster sampling and stepwise elimination are intuitive and robust procedures, but scale to the square of N, which makes them impractical for large data sets.

Experimental design

In an attempt to provide a rational criterion for selecting a maximally diverse subset of amines and carboxylic acids for use as side chains and capping groups in N-substituted glycine peptoid combinatorial libraries, Martin et al. [8] developed a selection

procedure based upon an established method of statistical experimental design known as D-optimal design. D-optimal design strives to identify a subset of compounds which is both diverse (the inter-compound distance in property space is maximized) and orthogonal (the covariances are minimized). The method begins with the empty set or a set of pre-selected compounds. At each step, a compound is chosen which maximizes the determinant of the 'information matrix', $X^T X$, which is equivalent to the volume subtended by the subset in covariance space. The rows of the design matrix, X, index the compounds of the subset, and its columns index either individual properties or higher-order combinations of properties such as their squares, cubes or cross products. The procedure terminates when a pre-determined number of compounds have been selected. This approach is nearly identical to one developed previously by Marsili and Saller [49] for the purpose of selecting multivariate synthetic analogues.

Recent work by Hassan et al. [43] has indicated that this method tends to increase the rank of the selected subset at the expense of spread. They used a Monte-Carlo method to maximize a diversity objective function based on the D-optimal criterion. The resulting sets of compounds were biased toward the periphery of the property space. This bias was especially evident when the number of compounds selected far exceeded the dimensionality of the space. It must be noted that the design matrix used in this study did not include higher-order combinations of properties, which may partially account for the redundancy of the results.

Partitioning techniques

Most of the algorithms discussed to this point scale to the square of the number of molecules in the set considered, making their use impractical for large data sets. In attempts to reduce this computational complexity, a number of groups have investigated simple partitioning (binning) techniques. Each axis of a molecular property space is divided into equal segments, creating a honeycomb of multi-dimensional cells. Compounds are assigned to the cells based on their properties. Many diversity-related tasks are greatly simplified using this approach. For example, a diverse subset of compounds may be created by selecting a pre-determined number of molecules from each cell. Comparison of two compound collections reduces to comparing the number of molecules assigned to corresponding cells in each collection. However, as with any partitioning method, the challenge is to minimize the number of discrete cells without having them become so large as to be useless. Pearlman [24] and Cummins et al. [42] have addressed this problem by reducing the dimensionality of the property space.

Pearlman constructed a 6-dimensional space using eigenvectors corresponding to the largest and smallest eigenvalues of an optimized set of three matrices which encoded atomic charge, polarizability and hydrogen bonding characteristics (B-Cut values). The principal shortcoming of this method is that it lacks a straightforward physical interpretation. Why choose the extreme eigenvalues? Burden, upon whose work this approach is based [25], contests that the smallest eigenvalue contains contributions from every atom, and therefore reflects the topology of the entire molecule. In addition, there is ample evidence that an adequate representation of molecular diversity requires substantially more than six dimensions [8]. Cummins et al. used factor analysis to reduce a 61-dimensional property space to four factors which accounted for 90% of the variance of five chemical databases (CMC, MDDR, ACD, SPECS and the Wellcome Registry) containing in excess of 300 000 compounds. The diversity of each data set was computed as the fraction of the total volume occupied by that set, using a Riemann-style approach. A trimming procedure was also employed to eliminate outliers and focus the analysis on the more densely populated areas of the feature space. The resulting density functions for two of these factors are shown in Figure 1.

Stochastic techniques

All of the algorithms discussed thus far select a diverse subset of molecules from a larger collection, but none guarantee that the subset will be the most diverse possible of a given size. Agrafiotis [44,50] and Hassan et al. [43] have independently proposed using simulated annealing to select an optimal subset of compounds based on an objective function which measures the diversity of any conceivable set of compounds. Simulated annealing is a global, multivariate optimization technique based on the Metropolis Monte-Carlo search algorithm. The method, as it is applied to molecular diversity, starts with a randomly selected subset of molecules. At each step, it makes a small change in the composition of the subset (usually on the order of 1–10% of the constituent members) and evaluates the diversity of the new subset. If the new subset is more diverse than the old, it is retained and the cycle begins

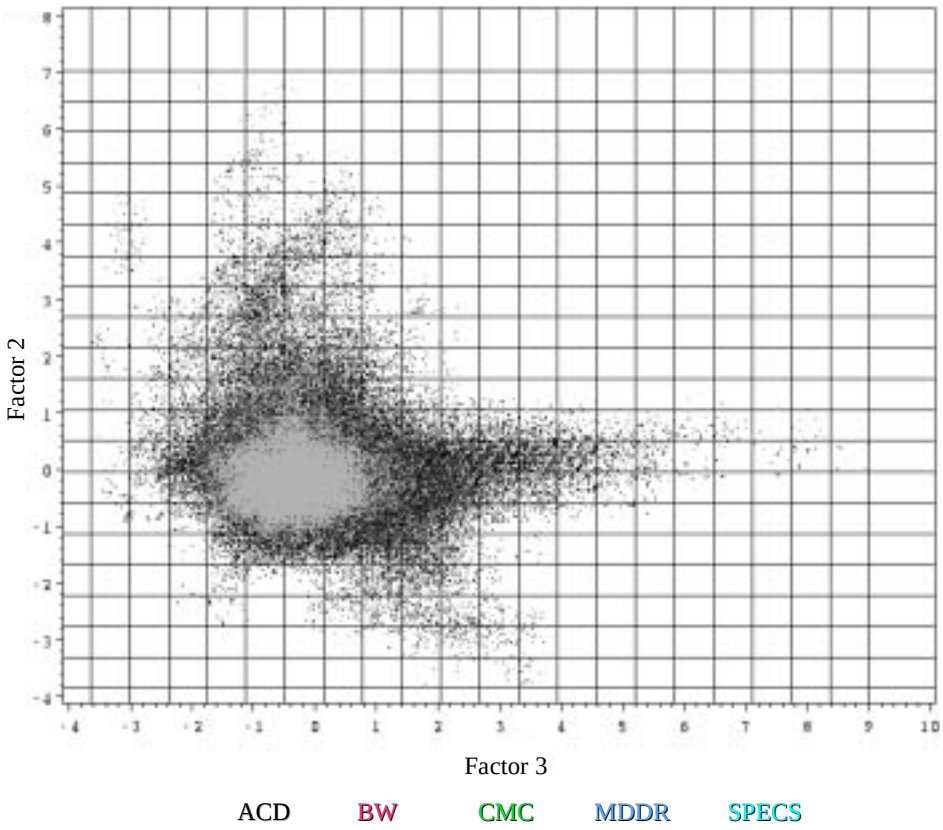

Figure 1. A comparison of the property distributions of five chemical databases using factor analysis.

anew. If it is less diverse, it is not rejected out-of-hand, but is retained with a probability that is inversely proportional to difference in diversity between the two subsets.

Hassan employed the maximin, power-sum and product functions as diversity metrics, while Agrafiotis' original implementation used maximin and a volumetric diversity measure of his own device. Although their methods were similar, the motivations of the two groups appear to be rather different. Hassan et al. wanted to compare the performance of different diversity metrics. Agrafiotis, on the other hand, was interested in developing a generalized selection method based on any number of disparate criteria (e.g. similarity and predicted activity as well as cost and availability of starting materials or reaction block design). This requirement was essential since the algorithms were to be used as part of an iterative drug discovery system known as DirectedDiversity®, in which the selection criteria would typically differ from iteration to iteration [51,52]. He later applied this algorithm to study other diversity metrics with surprising results [53,54] (see the sections on Information Theory and Cosine Coefficient below), and has also reported evolutionary and genetic variants of this general sampling approach.

Optimizable K-dissimilarity selection

K-Dissimilarity selection (OptiSim) is a generalization [71] of maximin which balances diversity against representativeness in the selection set. It starts by selecting a compound at random, then examines K other compounds chosen at random from the data set, excluding from consideration any which are too similar to the initial selection. From among these K compounds, the one which is most dissimilar to the initial selection is added to the selection set. At each iteration thereafter, a fresh sample of K candidates is compared to the compounds which have already been selected, and the most dissimilar among them is added to the selection set; all compounds are considered once before any candidate is considered twice. The published application employs the maximin dissimilarity criterion (the minimum pairwise dissimilarity of

the candidate to any compound already in the set) to identify the 'best' candidate from each subset, though other measures [2] can also be used.

The algorithm scales with $K*M^2$, where M is the number of compounds being selected. The smaller the sample size K, the more representativeness is favored in the selection set. As K approaches N, the total number of compounds in the data set, OptiSim reduces to maximin. At modest values of K (2–10), the balance between representativeness and diversity is very similar to that seen for selection based on hierarchical clustering [72]. Multiple passes with different random seeds are even more efficient (order $P*K*M^2$, for P passes of M selections each), and correspond to selecting multiple representatives from each cluster. OptiSim is commercially available from Tripos, Inc. as part of ChemEnlighten [73].

Vector analysis

A couple of groups have reported diversity methods based on an analysis of the spatial relationships of intramolecular functionalities. Boyd [56] reported a method, called HookSpace, that measures diversity based on the spatial distribution of distances between user-defined functional groups. In particular, each pair of functional groups in a given compound was aligned on the xy plane so that one of the groups was placed along the x axis with the head atom at the origin, and the other was positioned on the xy plane, with the head-to-tail vector pointing in the positive z direction. Once the alignment was complete, the position of the head atom of the second group on the xy plane was recorded. This process was repeated for every pair of functional groups in each structure, and for every structure in the database. The xy plane was then partitioned into a finite number of cells, and each cell recorded either the total number of functional groups, or the number of different functional groups at that position. This permitted diversity measurements by computing the percentage of non-vacant cells on the xy plane, similar to the method described by Pearlman. The authors used this approach to compare the structural diversity of the Available Chemicals Directory (ACD), the Cambridge Structural Database (CSD), and a benzodiazepam combinatorial library, using a theoretical reference space. They concluded that the ACD covered 85% of that space, whereas the CSD and the benzodiazepam library covered only 34 and 13% of the space, respectively. It is quite likely, however, that this difference reflects the different origins of the three-dimensional structures of these compounds (computed versus experimental), rather than the intrinsic functional and geometric diversity of the two databases.

In a related approach, Bartlett [57] presented a system that compared the diversity of different combinatorial templates using the angles between the bond vectors connecting the core to the substituent. The method followed the spirit of the Caveat approach, and the results were presented in a visual form.

Minimum spanning trees

At a recent conference, Ruppert and co-workers at Arris presented a novel method of computing molecular diversity based on spanning trees [77]. Their method, which they named IcePick, measures the difference between small molecules by comparing steric and polar features on their (flexible) three-dimensional surfaces. IcePick compares one molecule to another by 'turning it inside out' to form a pocket that perfectly fits around it (i.e. an ideal protein) and then flexibly docking the second molecule into the pocket. The fit is scored by comparing the protein accessible surface and the vector directions available for both hydrogen bond donation and acceptance; this is then averaged over the flexible conformational space of both molecules. From this measure of pairwise dissimilarity, IcePick computes the intrinsic diversity of a set or library of molecules using a 'spread' metric based on minimum spanning trees. The most diverse set is the one maximizing the weight of the minimum spanning tree. Since flexible three-dimensional docking can be computationally expensive, a statistical fingerprinting technique was developed to speed the diversity computation for large libraries.

Information theory

Lin [58] has proposed a diversity metric based upon the premise that maximizing the diversity of a subset of molecules is equivalent to maximizing its information content. The crux of the approach is the postulate that every collection of molecules is composed of a finite number of distinct species, or classes, of molecules, and that the ability to distinguish among these species can be described as a function of their mutual dissimilarity. The more distinguishable the species, the greater their information content of the collection. The diversity of the collection may then be quantified using Shannon's entropy formalism:

$$I = S_{\max} - S \tag{13}$$

where

$$S = -\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij} \ln p_{ij} \tag{14}$$

and N is the number of molecules in the collection, p_{ij} is the probability of finding the i-th molecule in the j-th species (given some function of their dissimilarity), and $S_{\max}$ is the maximum entropy of the system.

While the use of information theory seems like a natural choice, the actual implementation suffers from a number of disadvantages. In a recent article [53], we reported that a strict application of this approach produced extremely unbalanced designs, and clustered points at maximum separation along the diagonal of the feature space. We believe that this is due to the use of the wrong type of 'information', and to the implicit assumption that ideal designs should be equiprobable (i.e. that the pairwise intermolecular dissimilarities should be as uniform as possible). In a private communication, Lin suggested that our results could be an artifact of the similarity measure used in our study, but a detailed response has yet to appear in print. As of this writing, the debate is still open.

Cosine coefficient

Willett et al. [59] have noted that if the diversity of a set of compounds, $D(A)$, is defined as the complement of the mean pairwise intermolecular similarity:

$$D(A) = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\delta(i,j)}{N^2} \tag{15}$$

where $\sigma(i,j)$ is the similarity between compounds i and j, and N is the number of compounds in A, the expression can be reduced from one that scales with the square of the number of compounds in the set to one that scales linearly by using the cosine coefficient to evaluate the pairwise similarities. The cosine coefficient of similarity is defined as the cosine of the angle between two molecular property vectors:

$$\sigma(i,j) = \cos(i,j) = \frac{\sum_{k=1}^{K} m(i,k)m(j,k)}{\sqrt{\sum_{k=1}^{K} m(i,k)^2 \sum_{k=1}^{K} m(j,k)^2}} \tag{16}$$

where $m(i,k)$ is the k-th component of the i-th atomic property vector, $m(i)$, and K is the dimensionality of the space. Substituting Equation (16) for $\sigma(i,j)$, Equation (15) reduces to:

$$D(A) = 1 - \frac{\mathbf{a}_c \cdot \mathbf{a}_c}{N^2} \tag{17}$$

where

$$\mathbf{a}_c = \sum_{i=1}^{N} w(i)\mathbf{m}(i) \tag{18}$$

and the weights, $w(i)$, are given by:

$$w(i) = \frac{1}{\sqrt{\sum_{k=1}^{K} m(i,k)}} \tag{19}$$

Unfortunately, this dramatic improvement in performance comes at a significant price. Using the stochastic approach outlined above, Agrafiotis showed that the method has a tendency to over-sample the principal axes of the property space [54]. He postulated that this behavior is an artifact of the simple summation function used for the dissimilarity metric and the fact that the cosine coefficient only measures the angle between two property vectors and ignores their lengths, which are necessary to measure spread.

Visualization

> *The ancients built Valdrada on the shores of a lake, with houses all verandas one above the other, and high streets whose railed parapets look out over the water. Thus the traveler, arriving, sees two cities: one erect above the lake, and the other reflected, upside down. Nothing exists or happens in the one Valdrada that the other Valdrada does not repeat, because the city was so constructed that its every point would be reflected in its mirror, and the Valdrada down in the water contains not only all the flutings and juttings of the facades that rise above the lake, but also the rooms' interiors and ceilings and floors, the perspective of the halls, the mirrors of the wardrobes ...*
>
> *Valdrada's inhabitants know that each of their actions is, at once, that action and its mirror image, which possesses the special dignity of images, and this awareness prevents them from succumbing for a single moment to chance and forgetfulness.*
>
> Italo Calvino, Invisible Cities.

One of the most difficult challenges in data analysis is to be able to represent whatever complexities might be intrinsic to the data in a simple and intuitive form. In fact, one might argue that if the results of an analysis are unable to be conveyed to a target audience in a straightforward manner, the analysis, no matter how thorough, has failed. As will become evident, traditional methods of data visualization are inadequate to

represent the extremely large, high-dimensional data sets common to molecular diversity/similarity analyses. In order to minimize the complexity and shear number of individual plots needed to visualize this sort of data, one must attempt to reduce the dimensionality of the representation. The three techniques described below, self-organizing maps, multidimensional scaling and non-linear mapping, use different approaches to achieve dimensionality reduction, while preserving the topology of the original space. That is, points near each other in the high-dimensional space are also near each other in the low-dimensional space.

Visualization without dimensionality reduction

Multivariate data may be visualized up to three dimensions at a time through use of traditional techniques such as histograms and two- and three-dimensional scatter plots. Histograms provide a convenient means of analyzing one variable at a time, but do not reveal any relationships between variables. Scatter plots are more effective in this respect, but they become difficult to interpret if the density of points is high. In addition, if the number of variables exceeds all but a handful, the number of plots needed to represent the data rapidly becomes overwhelming. One particularly useful type of three-dimensional scatter plot is a pharmacophore plot, which is used to visualize the three-point pharmacophore keys (see previously) present in a molecule or set of molecules. Each pharmacophore is positioned in the plot according to its three inter-point distances, is labeled by a symbol which reflects its type, and is color-coded according to whether it contains 1, 2 or 3 identical loci. An example of such a plot generated using Chemical Design's Chem-X suite is shown in Figure 2.

Another interesting approach are the flower plots of Martin et al. [8] which represent aestetically pleasing variants of the traditional star diagrams used in multivariate analysis. Flower plots display all the properties associated with a single compound in one plot. They are circular bar graphs in which each 'petal' of the flower represents a molecular property or descriptor. The value of the property is reflected in the size of the petal and its sign determines whether the petal points outward (positive values) or inward (negative values). In addition, the central sphere is color-coded according to some additional property such a biological activity or similarity to a reference compound. Figure 3 shows the flower plot of tyramine, described by means of five chemical functionality descriptors, five shape descriptors, five atom-layer receptor recognition descriptors and the computed logP.

Because there is a one-to-one correspondence between the number of compounds and the number of plots, flower plots are impractical for visualizing large data sets. Flower plots are particularly useful, however, in assessing the distribution of properties across a small collection of compounds, for example, a set of reagents used in generating a combinatorial library.

Visualization with dimensionality reduction

Self-organizing maps

Self-organizing maps (SOM's) or Kohonen networks [64] belong to a class of neural networks known as competitive learning or self-organizing networks. They were originally developed to model the ability of the brain to store complex information as a reduced set of salient facts without loss of information about their interrelationships. High-dimensional data are mapped onto a two-dimensional rectangular or hexagonal lattice of neurons in such a way as to preserve the topology of the original space.

A Kohonen network is trained in the following manner: Each neuron, i, has an associated vector of weights $m_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{iN}]$, where N is the dimensionality of the original space. A randomly chosen training sample, $x = [\xi_1, \xi_2, \ldots, \xi_N]$, is presented to the network, and the weighting vectors of all the neurons are adapted to the input sample through use of a *neighborhood function* or *smoothing kernel*. A widely used kernel is given in Equation (20):

$$h_{ci}(t) = \alpha(t) \exp\left(-\frac{\|\mathbf{r}_c - \mathbf{r}_i\|^2}{2\sigma(t)^2}\right) \qquad (20)$$

where c is the neuron whose Euclidean distance to the input sample is the smallest, $\mathbf{r}_c$ and $\mathbf{r}_i$ are the respective locations of the c-th and i-th neurons on the lattice ($\mathbf{r}_c$, $\mathbf{r}_i \in \Re^2$), $\alpha(t)$ is the *learning rate*, and $\sigma(t)$ is the width of the function. This process is repeated until each training sample has been presented to the network, a phase referred to as a training epoch. Typically, many training epochs are necessary to complete the training. Upon completion, each neuron is sensitized to a particular region of the original space. Samples which fall within the same region, whether they were or were not included in the original training set, are mapped onto the same neuron.

Self-organizing maps were first used to visualize collections of molecules by Gasteiger et al. at the Uni-

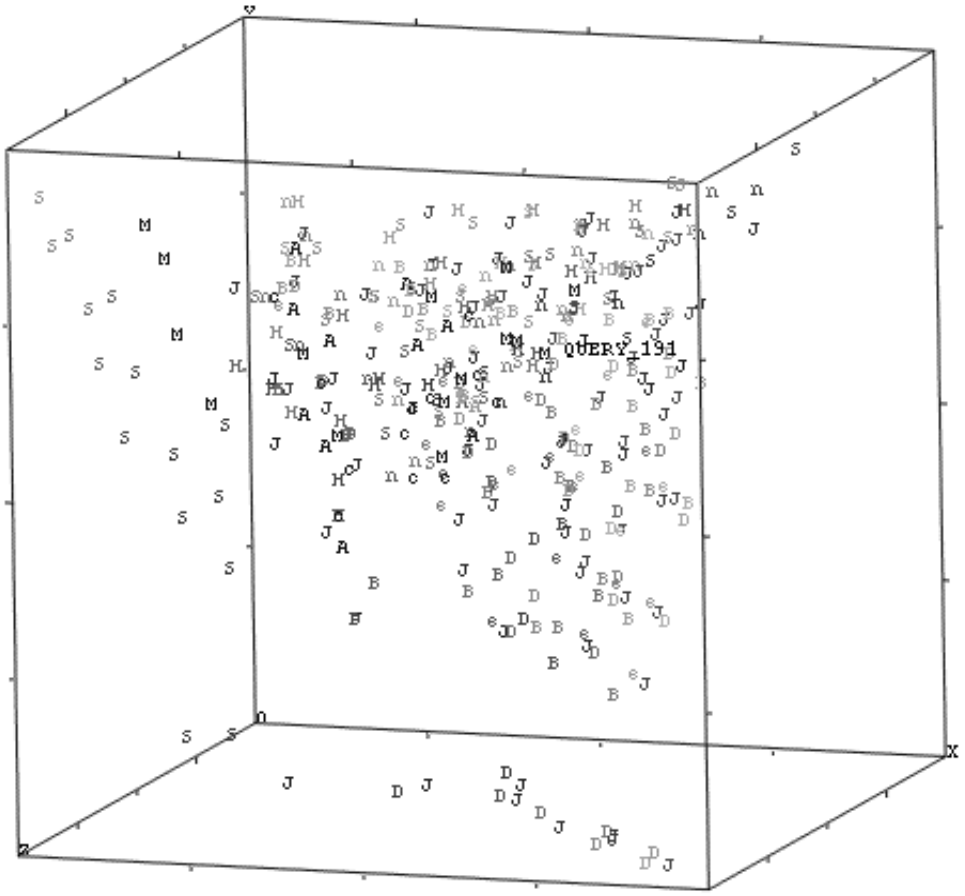

Figure 2. Pharmacophore plot generated by Chemical Design's ChemX software suite.

versity of Erlangen [23]. They validated the utility of the method by showing that it was able to spatially separate two combinatorial libraries designed by Rebek as potential trypsin inhibitors. The libraries were created by combining the tetra-substituted xanthene and cubane scaffolds shown in Figure 4 with a common set of 19 L-amino acids. Thus, while their substituents were structurally identical, the xanthene library arranged its substituents around a common plane, while those of the cubane library were arranged at the vertices of a tetrahedron. In addition, they showed that a third, adamantane-based library, which also oriented its substituents tetrahedrally, mapped to the same region as the cubane library.

Each molecule was described by a 12-dimensional spatial autocorrelation vector that encoded the electrostatic potential at its surface according to Equation (21):

$$A(d_l, d_u) = \frac{1}{N} \sum_{i,j} p_i p_j \tag{21}$$

where p_i and p_j are the values of the electrostatic potential at two randomly chosen points, i and j, on the molecular surface and N is the number of distance bins on the interval $[d_l, d_u]$ (12 in this case). Two 50×50 neuron Kohonen networks were trained: the first with the cubane and xanthene libraries, and the second with all three libraries. As is shown in Figure 5a, the cubane library (black cells) maps to a region distinct from that of the xanthene library (light gray cells). The overlap between the libraries amounts to only 3% of the total number of neurons, and is confined to the periphery of the cubane cluster. Figure 5b shows the map created by training the network with all three libraries. As expected due to their similar geometry, members of the cubane (black cells) and adamantane (dark gray cells) libraries mapped onto the same region of the network which was, again,

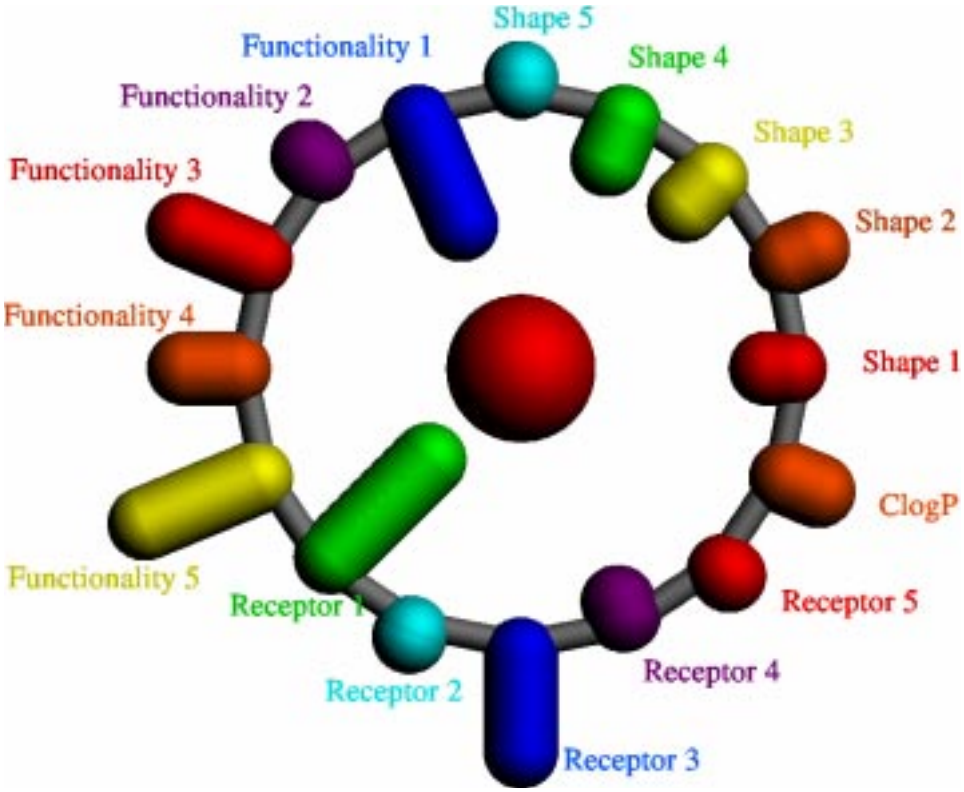

Figure 3. Flower plot of tyramine. There is one petal for each of the five chemical functionality descriptors, five shape descriptors, five receptor recognition descriptors and the computed logP.

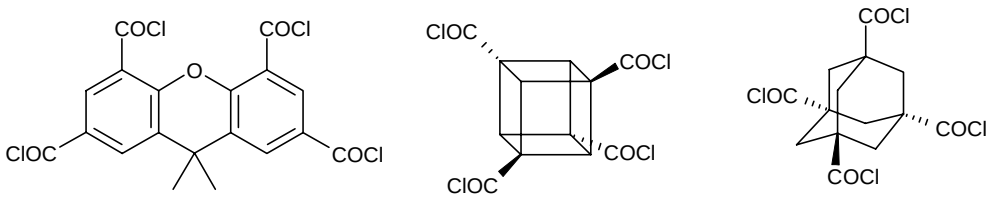

Figure 4. Combinatorial scaffolds used by Sadowski, Wagener, and Gasteiger: (left) xanthene, (middle) cubane, (right) adamantane.

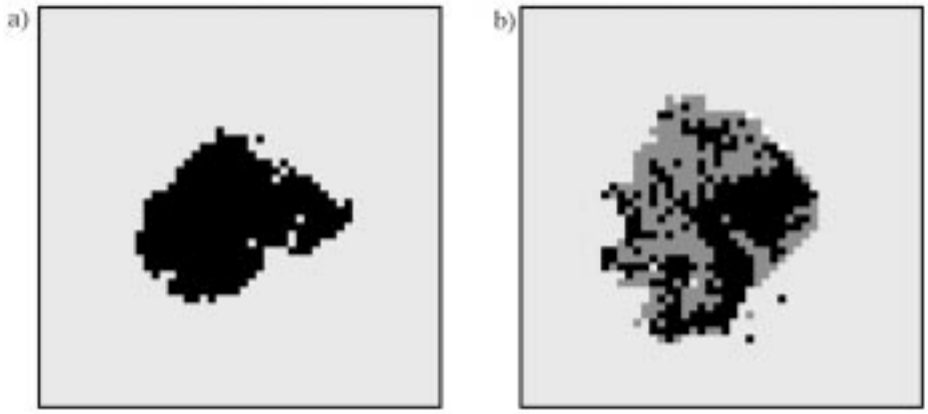

Figure 5. Self-organizing maps of (a) the xanthene (light gray) and cubane (black) libraries, and (b) the xanthene (light gray), cubane (black), and adamantane (dark gray) libraries.

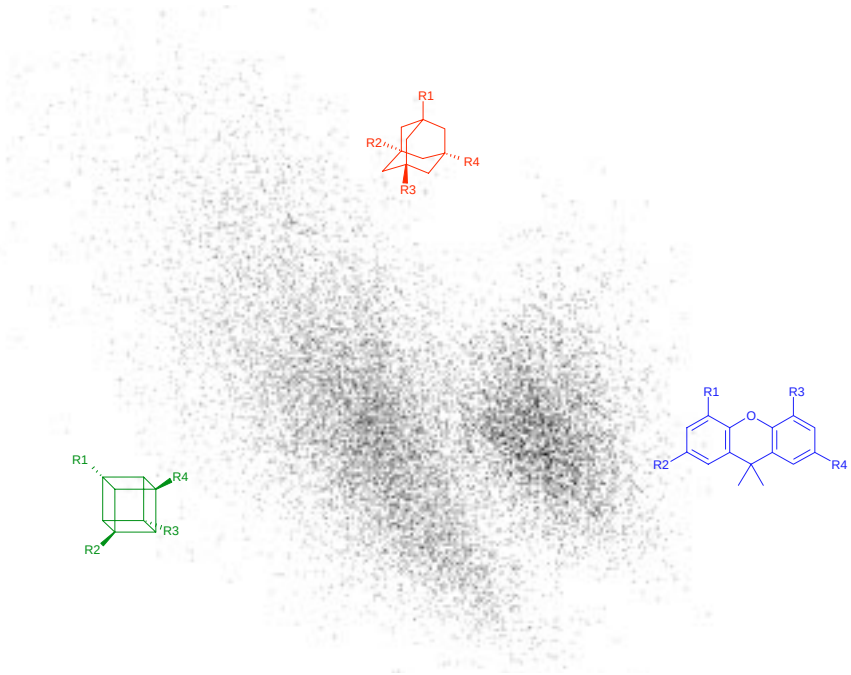

Figure 6. Sammon map of the xanthene, cubane, and adamantane libraries used by Gasteiger et al.

distinct from that of the xanthene library (light gray cells).

Given the simplicity of their output, SOM's can be very effective for visualizing and comparing chemical libraries, particularly when they are coupled with advanced, interactive graphical tools.

Non-linear maps

Non-linear mapping is a multivariate statistical technique that is closely related to multi-dimensional scaling [67]. Just like MDS, the objective is to approximate local geometric relationships on a two- or three-dimensional plot. The difference between MDS and non-linear mapping is in the minimization procedure. Sammon's algorithm is the most commonly used, but it too does not scale gracefully with the size of the data set. Self-organized non-linear mapping is a variant of Sammon's original algorithm that was developed by Agrafiotis [44,50] and is based on a self-organization principle reminiscent of Kohonen's SOM training algorithm. This method belongs to the family of non-metric algorithms, and is therefore applicable to a wide variety of input data. This is particularly useful when the (dis)similarity measure is not a true metric, i.e. it does not obey the distance postulates and, in particular, the triangle inequality (such as the Tanimoto coefficient). Although an 'exact' projection is only possible when the distance matrix is positive definite, meaningful projections can be obtained even when this criterion is not satisfied.

Non-linear maps were introduced by Agrafiotis to visualize protein sequence relationships in two dimensions [68], and were later employed as a means of visualizing and comparing large compound collections, represented by a set of molecular descriptors [44,50]. The advantage of non-linear maps over Kohonen networks is that they provide much greater detail about the individual compounds and their interrelationships. To provide a direct comparison between self-organized and non-linear maps, we applied the Sammon algorithm on the xanthene, cubane and adamantane libraries that were used by Gasteiger et al. in [23] (Figure 6). The projection was carried out using the same 12-dimensional auto-correlation descriptors and the Euclidean metric as a pairwise measure of dissimilarity. The resulting map is shown in Figure 6.

The map is sufficiently faithful, as manifested by a Sammon and Kruskal stress value of only 10 and 8%, respectively. It is evident that the Sammon map is not only capable of reproducing the sharp separation between the planar and tetrahedral systems that was observed in the self-organized maps (Figure 5), but also revealed a more subtle distinction between the cubane and adamantane libraries that was not captured

by the Kohonen network.

While the first application of this technique involved continuous molecular descriptors, the programs were later extended to include other molecular representations and molecular similarity metrics such as substructure keys, hashed fingerprints, Tanimoto coefficients etc. [69].

Conclusions

> *'You will, I am sure, agree with me that if page 534 finds us only in the second chapter, the length of the first one must have been really intolerable.'*
>
> Sherlock Holmes, The Valley of Fear.

The study of molecular diversity is an evolving field, fueled by rapid advances in experimental discovery and an urgent need for rigorous statistical experimental design. However, despite its conceptual simplicity, a rigorous mathematical definition remains elusive. Most approaches developed to date are rooted in the fields of molecular similarity and QSAR, and some have shown promise in increasing the hit rates of combinatorial chemistry experiments. On the other hand, there are many examples that suggest that diversity is serendipity in disguise. Validation is critical, but it can only come from comparison with appropriate control experiments which are hard to design and too expensive to execute. Although interest in molecular diversity will continue to grow, we believe that it will eventually become inextricably linked with structure-activity correlation and statistical series design, perhaps giving rise to a new unifying field devoted to the planning and analysis of massively parallel experiments.

References

1. Johnson, M. A. and Maggiora, G. M., Concepts and Applications of Molecular Similarity, Wiley, New York, NY, 1990.
2. Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, D.C., 1995.
3. Martin, E. J., Spellmeyer, D. C., Critchlow, R. E. Jr. and Blaney, J. M., *Does combinatorial chemistry obviate computer-aided drug design?*, In Lipkowitz, K. B. and Boyd, D. B. (Eds.) Reviews in Computational Chemistry, VCH, Weinheim, 1997, pp. 75–100.
4. Blaney, J. M. and Martin, E. J., *Computational approaches for combinatorial library design and molecular diversity analysis*, Curr. Biol., in press.
5. Martin, Y. C., Brown, R. D. and Bures, M. G., *Quantifying Diversity*, In Kerwin, J. F. and Gordon, E. M. (Eds.) Combinatorial Chemistry and Molecular Diversity, Wiley, New York, NY, in press.
6. Willett, P., Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, 1987.
7. Downs, G. M. and Willett, P., *Similarity searching and clustering of chemical-structure databases using molecular property data*, J. Chem. Inf. Comput. Sci., 34 (1994) 1094–1102.
8. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K. and Moos, W. H., *Measuring diversity: experimental design of combinatorial libraries for drug discovery*, J. Med. Chem., 38 (1995) 1431–1436.
9. Lewis, R., McLay, I. M. and Mason, J. S., *Chem. Des. Autom. News*, 10(4) (1995) 37–38.
10. Brown, R. D. and 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., 36 (1996) 572–584.
11. Brown, R. D. and Martin, Y. C., *The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding*, J. Chem. Inf. Comput. Sci., 37 (1997) 1–9.
12. Kubinyi, J., In Manhold, R., Krogsgaard-Larsen, P. and Timmermann, H. (Eds.) Methods and Principles in Medicinal Chemistry, Vol. 1, VCH, Weinheim, 1993, pp. 21–36.
13. Kier, L. B. and Hall, L. H., Molecular Connectivity in Structure-Activity Analysis, Wiley, New York, NY, 1986.
14. Charton, M. and Motoc, I. (Eds.), Steric Effects in Drug Design, Springer-Verlag, Heidelberg, 1983.
15. *Molconn-X*, Haney Associates, Mercer Island, WA.
16. Downs, G. M. and Barnard, J. M., *Techniques for generating descriptive fingerprints in combinatorial libraries*, J. Chem. Inf. Comput. Sci., 37 (1997) 59–61.
17. Carhart, R. E., Smith, D. H. and Venkataraghavan, R., *Atom pairs as molecular features in structure-activity studies: definition and application*, J. Chem. Inf. Comput. Sci., 25 (1985) 64–73.
18. Kearsley, S. K., Sallmack, S., Fluder, E. M., Andose, J. D., Mosley, R. T. and Sheridan, R. P., *Chemical similarity using physicochemical property descriptors*, J. Chem. Inf. Comput. Sci., 36 (1996) 118–127.
19. Sheridan, R. P., Miller, M. D., Underwood, D. J. and Kearsley, S. K., *Chemical similarity using geometric atom pair descriptors*, J. Chem. Inf. Comput. Sci., 36 (1996) 128–136.
20. Moreau, G. and Broto, P., *The autocorrelation of a topological structure: a new molecular descriptor*, Nouv. J. Chim., 4 (1980) 359–360.
21. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J., *Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists*, J. Chem. Inf. Comput. Sci., 36 (1996) 1205–1213.
22. Wagener, M., Sadowski, J. and Gasteiger, J., *Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks*, J. Am. Chem. Soc., 117 (1995) 7769–7775.
23. Sadowski, J., Wagener, M. and Gasteiger, J., Angew. Chem. Int. Ed. Engl., 34 (1996) 23–24.
24. Pearlman, R. S., *Novel software tools for addressing chemical diversity*, Network Science, 1996, June, http://www.awod.com/netsci/issues/.
25. Burden, F. R., *Molecular identification number for substructure searches*, J. Chem. Inf. Comput. Sci., 29 (1989) 225–227.
26. Unity Chemical Information Software, Tripos Associates, St. Louis, MO.
27. Murrall, N. W. and Davies, E. K., J. Chem. Inf. Comput. Sci.,

30 (1990) 312–316.
28. Sadowski, J., Gasteiger, J. and Klebe, G., *Comparison of automatic three-dimensional model builders using 639 X-ray structures*, J. Chem. Inf. Comput. Sci., 34 (1994) 1000–1008.
29. Sheridan, R. P., Nilikantan, R., Rusinko, A., Bauman, N., Haraki, K. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 29 (1989) 255–260.
30. Pickett, S., Mason, J. S. and McLay, I. M., *Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ)*, J. Chem. Inf. Comput. Sci., 36 (1996) 1214–1223.
31. Davies, E. K. and Briant, C., *Combinatorial chemistry library design using pharmacophore diversity*, Network Science, 1995, http://www.awod.com/netsci/issues/.
32. Kubinyi, H., 3D-QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993.
33. Cramer, R. D., Clark, R. D., Patterson, D. E. and Ferguson, A. M., *Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers*, J. Med. Chem., 39 (1996) 3060–3069.
34. Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, A., Bukar, R., Bauer, K. E., Dilley, H. and Rocke, D. M., *Predicting ligand binding to proteins by affinity fingerprinting*, Chem. Biol., 2 (1995) 107–118.
35. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. and Weinberger, L. E., *Neighborhood behavior: a useful concept for validation of molecular diversity descriptors*, J. Med. Chem., 39 (1996) 3049–3059.
36. Bellman, R. E., Adaptive Control Processes, Princeton University Press, Princeton, 1961.
37. Scott, D. W., Multivariate Density Estimation: Theory, Practice and Visualization, Wiley, New York, NY, 1992.
38. Wegman, E., *Hyperdimensional data analysis using parallel coordinates*, J. Ann. Statist., 41 (1970) 457–471.
39. Teig, S., Cambridge Healthtech Institute's Conference on Chemoinformatics, May 12–13, 1997, Arlington, VA.
40. Cooley, W. and Lohnes, P., Multivariate Data Analysis, Wiley, New York, NY, 1971.
41. Gibson, S., McGuire, R. and Rees, D. C., *Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments*, J. Med. Chem., 39 (1996) 4065–4072.
42. Cummins, D. J., Andrews, C. W., Bentley, J. A. and Cory, M., *Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds*, J. Chem. Inf. Comput. Sci., 36 (1996) 750–763.
43. Hassan, M., Bielawski, J. P., Hempel, J. C. and Waldman, M., *Optimization and visualization of molecular diversity of combinatorial libraries*, Mol. Div., 2 (1996) 64–74.
44. Agrafiotis, D. K., *Stochastic algorithms for maximizing molecular diversity*, J. Chem. Inf. Comput. Sci., 37 (1997) 841.
45. Lajiness, M. S., In Silipo, C. and Vittoria, A. (Eds.) QSAR: Rational Approaches to the Design of Bioactive Compounds, Elsevier, Amsterdam, 1991, pp. 201–204.
46. Polinsky, A., Feinstein, R. D., Shi, S. and Kuki, A., *LiBrain: software for automated design of exploratory and targeted combinatorial libraries*, In Chaiken, I. M. and Janda, K. D. (Eds.) Molecular Diversity and Combinatorial Chemistry, American Chemical Society, Washington DC, 1996, pp. 219–232.
47. Chapman, D., *The measurement of molecular diversity: a 3-dimensional approach*, J. Comput.-Aided Mol. Design, 10 (1996) 501–512.
48. Taylor, R., *Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals*, J. Chem. Inf. Comput. Sci., 35 (1995) 59–67.
49. Marsili, M. and Saller, H., *ANALOGS: a computer program for the design of multivariate sets of synthetic analogs*, J. Chem. Inf. Comput. Sci., 33 (1993) 266–269.
50. Agrafiotis, D. K., 3-rd Electronic Computational Chemistry Conference, http://hackberry.chem.niu.edu/ECCC3/paper48, 1996.
51. Agrafiotis, D. K., Bone, R. F., Salemme, F. R. and Soll, R. M., United States Patent 5,463,564, 1995.
52. Graybill, T. L., Agrafiotis, D. K., Bone, R., Illig, C. R., Jaeger, E. P., Locke, K. T., Lu, T., Salvino, J. M., Soll, R. M., Spurlino, J. C., Subasinghe, N., Tomczuk, B. E. and Salemme, F. R., *Enhancing the drug discovery process by integration of high-throughput chemistry and structure-based drug design*, In Chaiken, I. M. and Janda, K. D. (Eds.) Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery, American Chemical Society, Washington, DC, 1996, pp. 16–27.
53. Agrafiotis, D. K., *On the use of information theory for assessing molecular diversity*, J. Chem. Inf. Comput. Sci., 37 (1997) 576–580.
54. Agrafiotis, D. K. and Lobanov, V. S., *An efficient implementation of distance-based diversity measures based on k-d trees*, J. Chem. Inf. Comput. Sci., in press.
55. Shemetulsksis, N. E., Weininger, D., Blankley, C. J., Yang, J. J. and Humblet, C., *Stigmata: an algorithm to determine structural commonalities in diverse datasets*, J. Chem. Inf. Comput. Sci., 36 (1996) 862–871.
56. Boyd, S. M., Beverly, M., Norskov, L. and Hubbard, R. E., *Characterizing the geometrical diversity of functional groups in chemical* databases, J. Comput.-Aided Mol. Design, 9 (1995) 417–424.
57. Bartlett, P. A., Abstracts of Papers of the American Chemical Society, 1996, 211.
58. Lin, S. K., *Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy mixing*, Molecules, 1 (1996) 57–67.
59. Turner, D. B., Tyrrell, S. M. and Willett, P., *Rapid quantification of molecular diversity for selective database acquisition*, J. Chem. Inf. Comput. Sci., 37 (1997) 18–22.
60. Chernoff, H., *The use of faces to represent points in k-dimensional space graphically*, J. Am. Statist. Assoc., 68 (1973) 361–368.
61. Andrews, D. F., *Plots of high dimensional data*, Biometrics, 28 (1972) 125–136.
62. Fienberg, S. E., *Graphical methods in statistics*, Am. Statist., 33 (1979) 165–178.
63. Inselberg, A., *The plane with parallel coordinates*, The Visual Computer, 1 (1985) 69–91.
64. Kohonen, T., Self-Organizing Maps, Springer-Verlag, Heidelberg, 1996.
65. Torgerson, W. S., *Multi-dimensional scaling: I. Theory and method*, Psychometrika, 17 (1952) 401–419.
66. Kruskal, J. B., *Non-metric multi-dimensional scaling: a numerical method*, Phychometrika, 29 (1964) 115–129.
67. Sammon, J. W., *A non-linear mapping for data structure analysis*, IEEE Trans. Comp., C-18 (1969) 401–409.
68. Agrafiotis, D. K., *A new method for analyzing protein sequence relationships based on Sammon maps*, Protein Sci., 6 (1997) 287–293.

69. Agrafiotis, D. K. and Lobanov, V. S., unpublished results.
70. Agrafiotis, D. K., *Diversity of chemical libraries*, In Schleyer, P. v. R. (Ed.), Encyclopedia of Computational Chemistry, Wiley, New York, NY, 1998.
71. Clark, R. D., *OptiSim: an extended dissimilarity selection method for finding diverse representative subsets*, J. Chem. Inf. Comput. Sci., 37 (1997) 1181–1188.
72. Holliday, J. D. and Willett, P. J., *Definitions of dissimilarity for dissimilarity-based compound selection*, J. Biomol. Screening, 1 (1996) 145–151.
73. Tripos, Inc., St. Louis, MO.
74. Lowis, D., Tripos Technical Notes, 1 (1977) 2–7.
75. Nayeem, A., Heritage, T. and Hurst, T., IBC 6-th Annual Conference on Rational Drug Design, December 11–12, 1996, Coronado, CA.
76. Sage, C. R. and Ghuloum, A. M., IBC 6-th Annual Conference on Rational Drug Design, December 11–12, 1996, Coronado, CA.
77. Mount, J., Ruppert, J., Welch, W. and Jain, A., IBC 6-th Annual Conference on Rational Drug Design, December 11–12, 1996, Coronado, CA.