Expedited Articles

Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery

Eric J. Martin,* Jeffrey M. Blaney, Michael A. Siani, David C. Spellmeyer, Alex K. Wong, and Walter H. Moos

Chiron Corp., 4560 Horton Street, Emeryville, California 94608-2916

Received December 30, 1994[⊗]

Screening synthetic combinatorial libraries, such as mixtures of oligo(*N*-substituted)glycines, facilitates rapid drug lead discovery and optimization by vastly increasing the number of candidate molecules made and tested. Discovery efficiency and productivity can be further improved by using experimental design to maximize molecular diversity for a given library size or to bias the library with key features for a specific receptor. We describe new methods to quantify molecular diversity using descriptors that characterize lipophilicity, shape and branching, chemical functionality, and specific binding features. Experimental design methods select sets of side chains that are diverse in these properties, and "flower plots" allow the diversity to be graphically compared. We also quantify the overall diversity accessible to different families of combinatorial chemistry.

Introduction

Synthesizing and screening combinatorial mixtures of novel organic compounds is emerging as an increasingly important new technology for drug discovery.[1] For example, oligo(*N*-substituted)glycine "peptoids" (NSGs) are synthetic oligomers with a peptide backbone but with side chains attached at the amide nitrogen[2] instead of the α-carbon. Screening target-biased NSG libraries yielded novel, nanomolar antagonists to the α_1-adrenergic and μ-opiate receptors.[3] NSGs are synthesized by a solid-phase "submonomer" route which can incorporate the side chains from any of over 1000 readily available amines.[4] Coupled with well over 1000 possible amino-terminal capping groups, this abundance of side chains can yield over 10^{12} possible capped trimers, permitting highly diverse NSG libraries without resorting to high molecular weights that would limit oral bioavailability. However, this enormous complexity carries a price. While one can readily synthesize and screen all 160 000 possible tetrapeptides of the 20 coded amino acids, it is impractical to make and test trillions of possible capped "tripeptoids". A strategy for designing small combinatorial subsets of this vast potential library that will still effectively discover potent ligands would further accelerate drug lead discovery. This paper describes a general methodology, applicable to peptoids or other modular chemistries, that facilitates the design of subsets for synthesis and testing. It introduces a general approach for reducing high-dimensional discontinuous descriptors, such as bit strings or tables, to low-dimensional continuous descriptors suitable for visualization, experimental design, and other mathematical manipulations. Examples will be given from a set of 721 primary amines suitable as peptoid side chains and a set of 1133 carboxylic acids suitable for capping groups.

Efficiently designed combinatorial libraries for general screening of new, structurally uncharacterized receptors should minimize redundancy by employing a basis set of components with very diverse structures. Conversely, highly focused screening to optimize a lead or incorporate other available structural information should employ monomers with functionality similar to those in known ligands. Intermediate "biased-diversity" strategies should mix some monomers containing pharmacophoric features with others that are highly diverse, presenting the specific features to the receptor in myriad geometries and chemical environments. Our method quantifies similarity between monomers by computing a set of 15–20 properties for each side chain, so the distance between property vectors reflects the similarity between the monomers. Experimental design methods minimize or maximize these distances, thus selecting maximally similar or dissimilar monomer sets. The method also provides an explicit mechanism to incorporate "chemical intuition" into the design.

Our approach quantifies and compares molecular diversity between different types of libraries, an important step in answering the critical question: "How much diversity is sufficient?". We also describe a novel and simple graphical approach, "flower plots", for describing and comparing multidimensional diversity.

Methods

Monomer Pools. The publicly contributed program fcdtothor[5] was used to convert the ACD[6] and SPECS[7] databases of commercially available chemicals into a THOR database.[8] It contained 4517 aliphatic primary amines suitable for side chains in submonomer peptoid synthesis. A Daylight toolkit[9] program was written to filter these by price (≤$4/g), quantity available (>10 g), and lack of toxic or reactive chemical features. Some aromatic primary amines which coupled efficiently were also included, giving a total set of 721 primary amines. Similarly, 12 152 carboxylic acids and acid chlorides were filtered to give a set of 1133 potential capping groups.

Descriptors for Experimental Design. A general method for combinatorial library design should work with virtually any potential chemical building blocks; thus, easily calculated properties are preferred. We computed 15–20 descriptors that characterize lipophilicity, shape and branching, chemical functionality, and receptor recognition features.

[⊗] Abstract published in *Advance ACS Abstracts*, April 1, 1995.

0022-2623/95/1838-1431$09.00/0

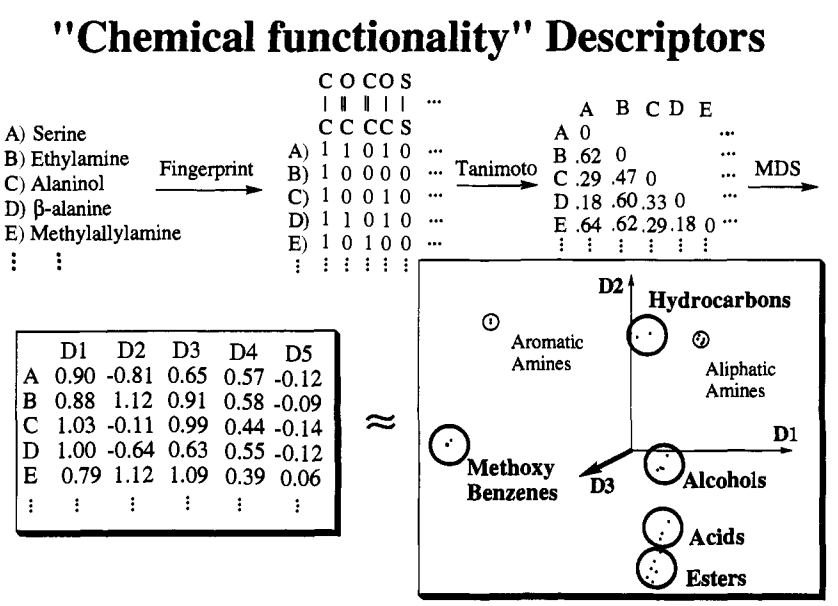


Figure 1. For each of 721 primary amines, a 2048-bit "fingerprint" string is computed where a 1 is set to indicate the presence of each chemical fragment up to seven bonds long. 1−Tanimoto coefficient for each pair of fingerprints gives a matrix of dissimilarities. Multidimensional scaling yields a set of five orthogonal "chemical functionality" properties, such that the Euclidean distance between property vectors for each pair of monomers reproduces the dissimilarity matrix with a relative standard deviation of 10%. The first three descriptors have been plotted for a set of 27 amines which have similar values in the remaining two dimensions. This shows the clustering by chemical functionality.

(1) **Lipophilicity.** Lipophilicity is described as estimated octanol/water partition coefficient. It is calculated using fragment-based methods with the commercially available CLOGP,[10,11] LOGKOW,[12] and HINT[13] programs or is estimated by comparison to experimental values for analogous compounds from the Pomona92 database.[14]

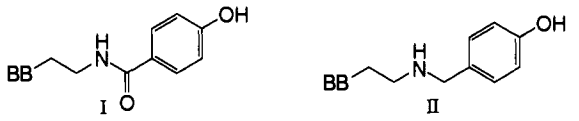
(2) **Topological Indices.** Overall side chain shape, flexibility, branching, and arrangement of cycles are characterized by topological indices[15] calculated with the commercially available program Molconn-X.[16] Topological indices have frequently been used to measure two-dimensional shape similarity.[17] We calculated 70 connectivity indices, seven shape indices, molecular weight, number of elements, nonhydrogen atoms, and bonds. These 81 descriptors were then reduced with principal components (PC) analysis using SAS PROC PRINCOMP.[18] For the set of 721 amines, PC analysis reduced the 81 descriptors to five latent variables that explained 86% of the variance. Similarly, the first five PCs explained 87% of the variance for the set of 1133 carboxylic acids. Inspection confirmed that compounds with similar sets of these five descriptors appear similar in overall two-dimensional shape, whereas compounds with very different sets of shape descriptors appear diverse.

(3) **Chemical Functionality Descriptors.** Another aspect of molecular similarity reflects the kinds of chemical functional groups represented in the monomers. Chemical database search keys enumerate the various chemical fragments in a molecule. For example, Daylight "fingerprint" routines search a molecule for all substructures up to seven bonds long and set one bit in a 2048-bit string for each fragment found.[19] The Tanimoto coefficient, which measures similarity between two bit strings,[20] has been applied to these fingerprints to measure chemical similarity for database searching and clustering.[21] Figure 1 illustrates how we used these similarities as the starting point to derive a small number of continuous descriptors suitable for experimental design. A Daylight toolkit program calculated the dissimilarity between monomers as 1−Tanimoto coefficient for each pair of fingerprints. This gives a symmetrical $N \times N$ dissimilarity matrix with $(N^2 + N)/2$ unique values, where N is the number of monomers. These intermonomer dissimilarities, which vary from zero for identical monomers to one for maximally dissimilar side chains, can be regarded as distances in a Cartesian space of unknown dimension. Multidimensional scaling (MDS) determines low-dimensional Cartesian coordinates for every monomer which best reproduce, simultaneously, the entire set of inter-side chain dissimilarities.[22] For the set of 721 amines, SAS PROC MDS[23] reduced the 2048-bit fingerprints to just five continuous variables that reproduce all 260 000 original dissimilarities with a relative standard deviation of just 10%. Seven dimensions were required to reproduce the 642 000 pairwise similarities among the 1133 carboxylic acids to the same precision. The calculations required 7 h on an IBM RS6000 580 computer. Since monomers with similar values for these latent variables contain similar chemical fragments, we called these dimensions "chemical functionality" descriptors.

(4) **Receptor Recognition Descriptors.** "Atom layer" properties were developed to describe the distribution, throughout the side chain, of chemical features that contribute to specific intermolecular interactions in receptor binding. These descriptors account for the directionality of the side chain, *i.e.*, that atoms near the peptoid backbone may contribute to binding differently than those that are more remote. As such, they require a unique atom, such as a point of backbone attachment or a pharmacophoric feature. They also incorporate isosterism, *e.g.*, that one acidic functionality can often substitute for another. A Daylight toolkit program characterized each non-hydrogen atom by six properties: radius and whether it is acidic, basic, an H-bond donor, an H-bond acceptor, and/or aromatic. Within a side chain, all atoms a given bond count distance from the backbone comprise an "atom layer". Figure 2 shows how each of the six atom properties were summed for all atoms within each layer (for 15 layers) to make a table of 6 columns (properties) by 15 rows (layers). For each pair of side chains, the maximum and minimum values were determined for each pair of corresponding cells in their respective atom layer tables. The sum of the minima divided by the sum of the maxima gives "$\Sigma_{min}/\Sigma_{max}$" similarity between that pair of side chains. These values, which vary from 0 to 1, can be considered fuzzy logic analogs of Tanimoto similarity and equal the Tanimoto coefficient for binary data. As with the chemical database fingerprint similarities, a dissimilarity matrix was generated and MDS was applied. For both the 721 amines and the 1133 carboxylic acids, five dimensions sufficed to reproduce the original dissimilarities with a relative standard deviation of 10%. Other atomic properties, such as partial charge or atomic hydrophobicity, could be included.

(5) **Scaling.** This completes the 16 properties calculated

Receptor Recognition Similarity Based on Atom Layer Tables



Atom Layer Table for I

	Radius	Acid	Base	HBD	HBA	Ar
1	1.9	0	0	0	0	0
2	1.9	0	0	0	0	0
3	1.8	0	**0**	1	1	0
4	1.9	0	0	0	0	0
5	**3.5**	0	0	0	1	1
6	3.8	0	0	0	0	**2**
7	3.8	0	0	0	0	2
8	1.9	0	0	0	0	1
9	1.7	1	0	**0**	1	0
10	0	0	0	0	0	0

Atom Layer Table for II

	Radius	Acid	Base	HBD	HBA	Ar
1	1.9	0	0	0	0	0
2	1.9	0	0	0	0	0
3	1.8	0	1	1	**0**	0
4	1.9	0	0	0	0	0
5	**1.9**	0	0	0	**0**	1
6	3.8	0	0	0	0	2
7	3.8	0	0	0	0	2
8	1.9	0	0	0	0	1
9	1.7	**0**	0	1	1	0
10	0	0	0	0	0	0

$\Sigma min / \Sigma max = 28.6/35.2 = 0.81$

Figure 2. "Atom layer" table for a side chain made by summing each of several properties for all non-hydrogen atoms a given bond count distance from the oligomer backbone (BB), including: radii, acids, bases, H-bond donors (HBD), H-bond acceptors (HBA), and aromatics (Ar). Two tables are compared element by element, and the sum of the minima divided by the sum of the maxima gives the similarity between the side chains. Although compound **II** differs from **I** only by the removal of one atom, there are six changes in the tables (in bold), since the neutral amide HBA in **I** becomes a basic amine in **II**, and the phenolic oxygen, which had been acidic due to the electron-withdrawing carbonyl in **I**, is a neutral HBD without it in **II**. (Rows 11–15 were not shown as they contain all zeros. Note that elements that are zero in both tables do not affect the calculated similarity.)

for each amine monomer (or 18 for each carboxylate capping group since they required two more chemical functionality descriptors). log *P* was scaled to unit variance and a mean of 0. The remaining descriptors that describe each monomer fall into 3 sets: five "shape and branching" PCs from topological indices, five (or for the carboxylates, seven) "chemical functionality" MDS dimensions from database fingerprints, and five "receptor recognition" MDS dimensions from atom layer properties. Within each set of five, the descriptor values were centered about the mean and identically scaled to give unit standard deviation for the "first" component (*i.e.*, with the largest Eigenvalue, which explains the most variance). Thus, log *P* and the first PC and MDS dimensions received equal weight in the designs, and the higher PCs and MDS dimensions received increasingly less weight.

Experimental Design. The primary motivation behind monomer characterization is to facilitate experimental design. The current Chiron mixture production robots are designed to use up to 36 monomers in each position of a modular combinatorial synthesis.[24] The final task in the design, therefore, is to accommodate the robot design by choosing small subsets of amines and carboxylates from the pools of 721 or 1133 that maximize or minimize the similarity between members. Each amine submonomer or carboxylate capping group was represented by a vector of the 16 or 18 properties (or by the top 12–14 PCs from those properties). Sets of monomers similar to a lead were chosen simply by rank ordering every member of the pool by the Euclidean distance from a selected side chain in the lead. Finding dissimilar sets is more difficult. In particular, we often wanted to design a "biased-diversity" set, by including some particular monomers based on a pharmacophore hypothesis or other criteria, and then complete the rest of the set with a small number of additional monomers from the full pool of 721 (or 1133) that are mutually diverse. This is accomplished with "D-optimal" design,[25] using the commercially available JMP software.[26]

D-optimal design chooses subsets from a large fixed pool that maximize the determinant of the "information matrix" $|\mathbf{X}'\mathbf{X}|$ for (in this case) a quadratic design matrix, **X**. This minimizes the determinant of the inverse, which is the variance of the parameter estimates for a cubic model. The rows of **X** are the monomers, and the columns are the properties and their squares. Roughly speaking, in order to determine accurate parameter estimates to a quadratic response surface, the D-optimal design algorithm chooses small subsets of points from the large pools of 721 or 1133 that are well spread out and nearly orthogonal in property space; *i.e.*, they are diverse. On the basis of an existing structure–activity relationship (SAR) or other information, some monomers can be preselected for inclusion in the set. The D-optimal design algorithm will then select additional monomers which best complement those, completing a design of a specified size with maximal overall diversity.

Results and Discussion

Sample Design. A sample monomer set for a biased combinatorial library is shown in Figure 3A. The top row of structures is tyramine and its five closest analogs. (The side chain from tyramine is found in several low-nanomolar NSG peptoid ligands for both the α_1-adrenergic and μ-opiate receptors.[3]) The lower two rows were chosen by D-optimal design to complete a diverse set of 18 from the pool of 721 amines.

D-optimal algorithms are extremely fast; a trial design can be modified by excluding candidates from the total pool or including new preselected side chains, guiding the generation of new designs interactively. After many such cycles, a set can be achieved which is both chemically reasonable and statistically diverse. This ability to interactively steer the evolution of the design allows for the essential marriage between chemical intuition and statistical rigor.

The use of MDS to convert similarities to a Euclidean property space, followed by an optimal design procedure, has several advantages over the alternative approach of performing cluster analysis directly on the similarities and then choosing one compound from each cluster. It allows diverse properties from many sources to be combined. Collinearity analysis of the properties reveals the dimensionality of the data, which in turn helps indicate how many monomers are needed to supply what degree of coverage. Even if properties are correlated, principal components analysis can be performed at a final step. Imposing a grid on a Euclidean property space reveals the number and size of unrepresented regions and whether a given monomer is near an "edge" of the space. Compounds selected from clusters (centroids or otherwise) generally do not produce a balanced, orthogonal design. In many clustering methods, the centroid of one cluster can actually lie within another cluster, so the "most representative member" of the cluster would never be chosen. In short, extracting a latent property space allows one to apply all of the tools of Euclidean geometry to the design and evaluation of monomer sets. Cluster analysis, however, does have the important advantage that it can be applied to much larger data sets.

Flower Plots. "Flower plots" were developed to simultaneously display all 16 properties for each individual amine monomer. This graphical tool allows the similarity within a monomer set to be visually evaluated and aids in the interpretation of structure–activity relationships. Flower plots are bar graphs in which the "*x*-axis" has been wrapped in a circle. They were generated with a modified version of M. Connolly's

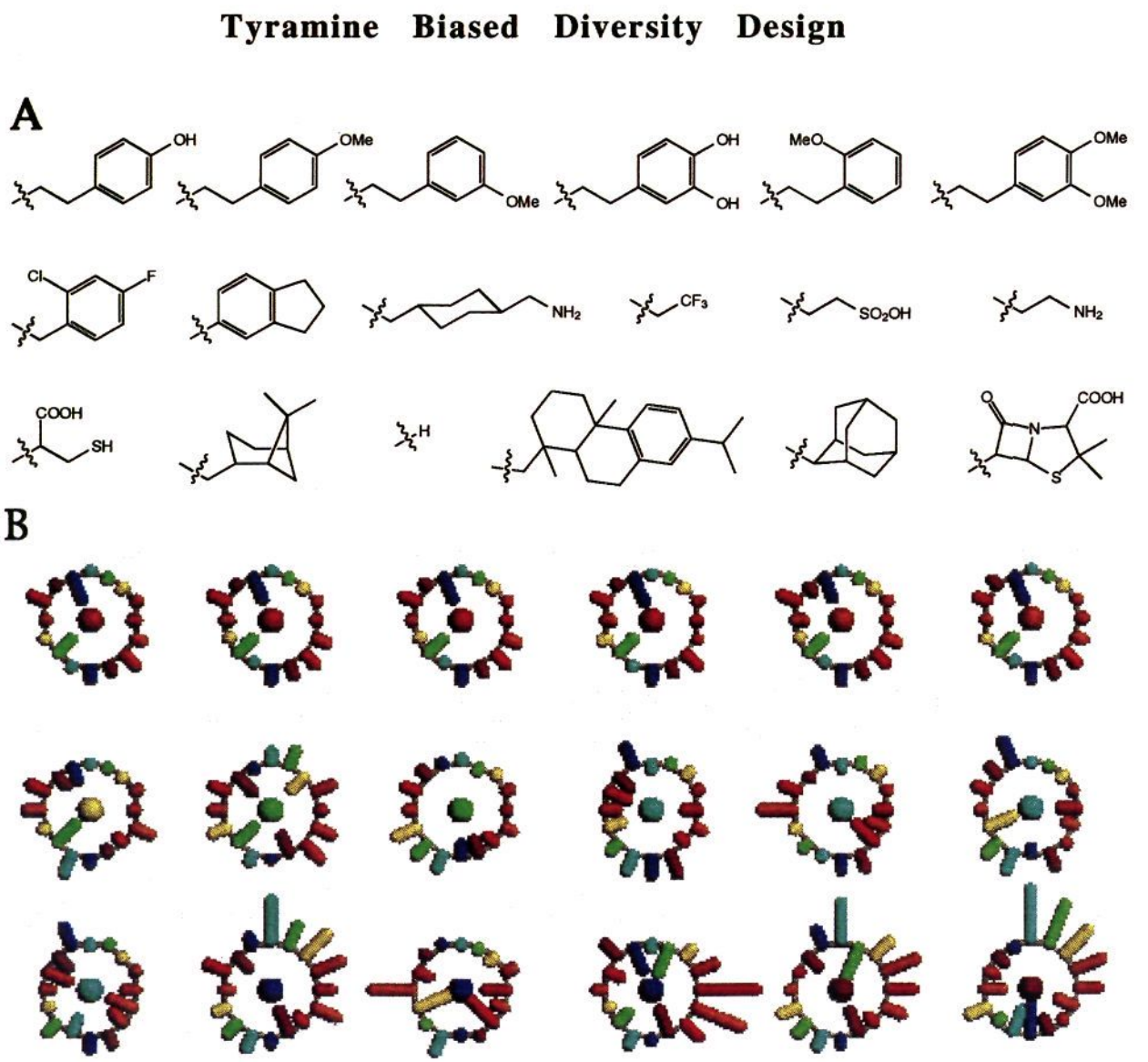


Figure 3. Structures and "flower plots" of 18 side chains from a biased NSG peptoid combinatorial library based on the tyramine submonomer. (A) The top row side chains are from tyramine and its five closest available analogs. The 12 side chains in the lower rows were chosen by D-optimal design from a pool of 721 amines. (B) Corresponding flower plots each represent all 16 properties for a single side chain. Petals for positive values point outward, and negative petals point toward the center. The radius is 3 standard deviations, and the center has been colored by similarity to tyramine.

MSP software.[27] Figure 3B shows the flower plots for the biased-diversity set of peptoid side chains based on the tyramine monomer. Each property is assigned to a "petal", with positive petals pointing outward and negative petals pointing toward the center. The radius is 3 standard deviations for the first dimension, and the center circle is colored in spectrum order to indicate an additional property such as biological activity or, in this case, similarity to the reference structure tyramine. The top row of tyramine analogs has nearly identical flower plots. The lower two rows of flower plots for diverse structures appear dissimilar. The flower plots can show how diversely the 721 amines are distributed through the computed property space and how widely the members of a small, statistically chosen subset sample this space. They cannot, of course, prove that the subset is diverse in any "absolute" sense, independent of the calculated properties.

Collinearity Analysis. For the 721 amines, only the first database fingerprint MDS dimension and the first atom layer MDS dimension are correlated, with $r^2 = 0.67$ and variance inflation factors (VIFs) of 9.6 and 10.8, respectively. VIFs, which are the diagonal elements of the inverse correlation matrix, measure multicollinearity.[22] Values vary from 1 for an orthogonal variable to infinity for a completely redundant descriptor. All other pairwise correlations are low ($r^2 \leq 0.4$), and the low VIFs (≤ 3) for the remaining 14 descriptors indicate low multicollinearities as well. Principal components analysis showed that the first PC explained only 15% of the variance for the 16 properties, and 12 PCs were needed to explain 95% of the variance. For the larger set of carboxylates, there were no high ($r^2 > 0.4$) pairwise correlations. However, as with the amines, the VIFs for the first database fingerprint MDS dimension and the first atom layer MDS dimension were fairly high

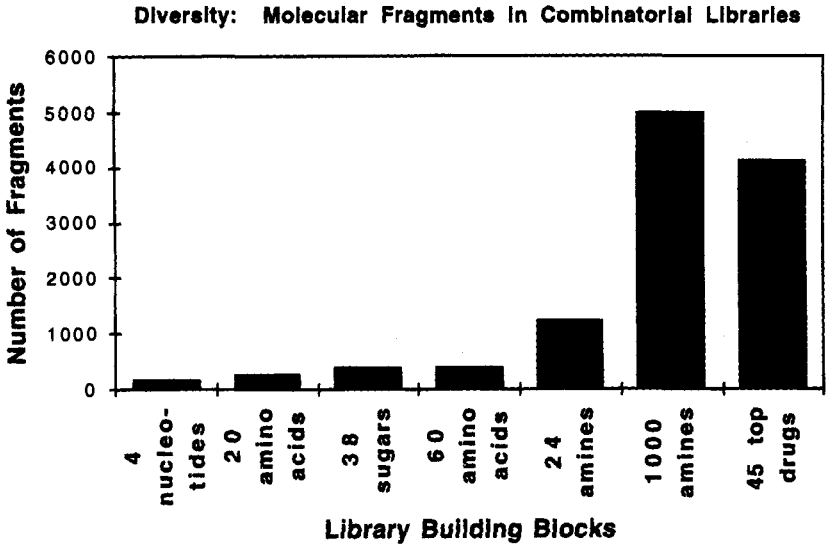


Figure 4. "Chemical functional group diversity" measured as the number of molecular fragments of up to seven bonds that comprise various chemical diversity libraries. NSG peptoids made from primary amines are far more diverse than any of the biopolymer libraries and slightly more diverse than the set of 45 best selling small molecule drugs from 1991.

(9.4 and 8.9, respectively), indicating multicollinearities. The first PC explained only 16% of the variance for 18 properties, and 14 PCs were needed to explain 95% of the variance. Thus, although these descriptors are not strictly orthogonal, there is little redundancy among them.

Comparing Diversity Among Libraries. In addition to serving as the starting point for the chemical functionality descriptors used in experimental design, the Daylight database fingerprints can be used as a measure of the "chemical functional group diversity" spanned by a library. Recall that the fingerprint routines search through a molecular structure to find all of its substructures of up to seven bonds. A hashing algorithm randomly assigns each substructure to one bit in a 2048-bit string. The logical OR of the all the fingerprints in a library of molecules is termed the "library fingerprint", with bits set for all the chemical fragments represented in the entire library. Thus, the more bits set in the library fingerprint, the greater its chemical functional group diversity.

There is a one-to-many relationship between fingerprint bits and fragments. The first fragment found will always set an unset bit. However, if one-half of the bits have already been set, then there is only a 50% chance that the hashing algorithm will assign the next new fragment to a previously unset bit. If the library fingerprint is already 90% set, then, on average, 10 new fragments are required to set just one additional bit. The number of fragments (f) required to set a given number of number of bits (b) is approximately $f = 2048\text{-}(\ln(2048/(2048 - b)))$. Eventually, all of the bits will be set, and the library fingerprint will be "saturated". A 2048-bit library fingerprint saturates at about 15 000 fragments.

Figure 4 compares the number of such fragments comprising several types of libraries. Combinatorial oligonucleotide libraries made from four bases, regardless of length, are all composed of fewer than 200 fragments. Peptide libraries made from the 20 coded amino acids are composed of fewer than 300 fragments. Although the chemistry may not yet exist to make such a library, a mixture of all possible oligosaccharides made from 38 commercially available sugars, including branched oligomers, would include only about 400 fragments. Peptide libraries made from an augmented set of 60 amino acids would contain just over 400 fragments. NSG peptoid libraries made from all combinations of the 721 amines in this study encompass over 5000 fragments, giving them 1 order of magnitude higher chemical functional group diversity than any of the biopolymer libraries. Even a small library from a D-optimally designed set of 24 amines contains 3 times as many fragments as the best biopolymer libraries. As a sobering bench mark, however, Figure 4 also shows that the 45 best selling small molecule drugs from 1991 contained over 4000 fragments, far more than any of the biopolymer libraries and almost as many as the NSGs. Medicinal chemists have employed not only a wide variety of chemical functionality to solve their drug design problems but also a high density of functional groups compared to simple oligomeric compounds. The average size of the small molecule drugs is slightly less than that of the NSG trimers and much less than typical biopolymer libraries. If "diversity density" is characterized as the number of unique fragments in a library divided by the average molecular weight, the existing small molecule drugs actually have slightly greater diversity density than the NSG libraries and much higher diversity density than the peptides, oligosaccharides, or oligonucleotides. Such analyses can be used to indicate whether to continue exploring a particular series of compounds or to move on to a new family of combinatorial chemistry. While the density of structural fragments is not a complete or absolute measure of molecular diversity, particularly in large surface interactions between biopolymers, it seems like a useful one for assessing libraries of low molecular weight compounds likely to produce orally active drugs. It will be a challenge to combinatorial synthesis technology to develop new low molecular weight libraries that are as diverse as the small organic compounds that make up the current generation of drugs.

Conclusion

Synthesizing and testing combinatorial libraries need not be "brute force" screening or "irrational", purely random drug discovery. Principles of sound experimental design now routinely employed in traditional medicinal chemistry can be applied to the design of combinatorial mixtures as well. Using structural descriptors and statistical techniques, monomers can be chosen to maximize a library's diversity or to bias a library toward certain features while keeping other features dissimilar. When testing mixtures, maximizing diversity not only minimizes redundancy to increase efficiency but also improves the chance that activity in a potent mixture is due to a few highly active unique compounds, rather than a large number of similar compounds with only moderate potency. While the method was illustrated for selecting amines and carboxylic acids for capped peptoid libraries, it is general and has been employed for other modular chemistries as well.

While these methods are now routinely used to design NSGs and other combinatorial libraries, they still represent only the initial efforts to apply computational methods to modular synthesis and screening. Current work includes experimental validation of these theoretical design approaches, using our growing database of combinatorial library structure–activity data. Studies are also underway to characterize three-dimensional shape similarity for flexible side chains, characterize and design new scaffolds for additional families of modular chemistry, and design, *de novo*, libraries to bind a receptor or enzyme site of known structure. With sufficient ingenuity, many computational techniques currently employed in "rational drug design" should also be adaptable to "rational library design" for combinatorial screening.

References

(1) Moos, W. H.; Green, G. D.; Pavia, M. R. Recent Advances in the Generation of Molecular Diversity. *Annu. Rep. Med. Chem.* **1993**, *28*, 315.

(2) Simon, R. J.; Kania, R. S.; Zuckermann, R. N.; Huebner, V. D.; Jewell, D. A.; Banville, S.; Ng, S.; Wang, L.; Rosenberg, S.; Marlowe, C. K.; Spellmeyer, D. C.; Tan, R.; Frankel, A. D.; Santi, D. V.; Cohen, F. E.; Bartlett, P. A. Peptoids: A Modular Approach to Drug Discovery. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9367–9371.

(3) Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Moos, W. H. Discovery of Nanomolar Ligands for 7-Transmembrane G-Protein Coupled Receptors from a Diverse (N-Substituted)-Glycine Peptoid Library. *J. Med. Chem.* **1994**, *37*, 2678–2685.

(4) Zuckermann, R. N.; Kerr, J. M.; Kent, S. B. H.; Moos, W. H. Efficient method for the preparation of peptoids [oligo(N-substituted glycines)] by submonomer solid-phase synthesis. *J. Am. Chem. Soc.* **1992**, *114*, 10646–10647.

(5) Siani, M. A. fcdtothor. *Daylight directory of publically contributed programs*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994; Vol. 4.34. Note that this program also converts price and quantity information to U.S. dollars/g to assist filtering the reagents by cost.

(6) MDL. *Available Chemicals Directory*; Molecular Design Ltd.: San Leandro, CA, 1993.

(7) Brandon/SPECS. *Brandon/SPECS Database and Compound Collection*; Brandon Assoc.: Merrimack, NH, 1993.

(8) Weininger, D. *Thor*; Daylight Chemical Information Systems: Irvine, CA, 1993.

(9) James, C. A.; Weininger, D.; Scofield, J. *Daylight Toolkit Programmer's Guide*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.

(10) Leo, A. Calculating log Poct from Structures. *Chem. Rev.* **1993**, *93*, 1281.

(11) Weininger, D. *CLOGP*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.

(12) Howard, P.; Meylan, W. *LOGKOW*; Syracuse Research Corp.: Syracuse, NY, 1993.

(13) Kellogg, G. E.; Joshi, G. S.; Abraham, D. J. *Med. Chem. Res.* **1992**, *1*, 444.

(14) Leo, A. *Pomona 92 Database*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1992.

(15) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In *Reviews in Computational Chemistry;* Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers, Inc.: New York, NY, 1991; Vol. 2; pp 367–421.

(16) Hall, L. H. *Molconn-X*, Version 1.0; Hall Assoc.: Quincy, MA, 1991.

(17) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, NY, 1990; pp 99–117.

(18) SAS. *SAS/STAT User's Guide Vol. 2*, 6th ed.; SAS Institute Inc.: Cary, NC, 1990.

(19) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.

(20) Willett, P. *Similarity and Clustering in Chemical Information Systems*; John Wiley & Sons: New York, NY, 1987; pp 54.

(21) Blaney, J. M. Similarity Searching and Clustering. Presented at the Medchem User Group Meeting, Pomona, CA, 1989.

(22) Catterjee, S.; Price, B. *Regression Analysis by Example*; John Wiley and Sons: New York, NY, 1977.

(23) SAS, Technical Report P-229, 6.10; SAS Institute Inc.: Cary, NC, 1992.

(24) Zuckermann, R. N.; Kerr, J. M.; Siani, M. A.; Banville, S. C. Design, Construction and Application of a Fully Automated Equimolar Peptide Mixture Synthesizer. *Int. J. Pept. Protein Res.* **1992**, *40*, 497–506.

(25) Federov, V. V. *Theory of Optimal Experiments*; Academic Press: New York, 1972.

(26) SAS. *JMP*, Version 3; SAS Institute Inc.: Cary, NC, 1992.

(27) Connolly, M. L. The molecular surface package. *J. Mol. Graph.* **1993**, *11*, 139–141.

JM940872+