Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries

Weifan Zheng, Sung Jin Cho, and Alexander Tropsha*

The Laboratory for Molecular Modeling, School of Pharmacy,
University of North Carolina, Chapel Hill, North Carolina 27599-7360

Received March 31, 1997

We describe a new computational approach, called Focus-2D, to the rational design of targeted combinatorial chemical libraries. This approach is based on the hypothesis that structurally similar compounds display similar biological activity profiles. Building blocks that are used in a combinatorial chemical synthesis are randomly assembled to produce virtual library compounds. Individual library compounds are represented by Kier−Hall topological descriptors. Molecular similarities between compounds are evaluated quantitatively by modified pairwise Euclidean distances in multidimensional descriptor space. Simulated annealing is used to search the potentially large structural space of virtual chemical libraries to identify compounds similar to lead molecules. Frequency analysis of building block composition of selected virtual compounds identifies building blocks that can be used in combinatorial synthesis of chemical libraries with high similarity to the lead molecules. We show that this method correctly identifies building blocks found in active peptoids with adrenergic or opioid activities.

INTRODUCTION

The experimental process of drug discovery has been influenced in recent years by the advances of combinatorial organic synthesis and high throughput screening of chemical libraries.[1,2] The general idea of these approaches is to assay all structurally diverse compounds synthesized via the combination of available building blocks. The two common trends in combinatorial chemistry include the synthesis of chemical libraries for broad or targeted biological screening. Although these methods have been highly efficient, it still remains impossible in most cases to synthesize and assay all library compounds in a sufficiently short period of time. Thus, from the practical standpoint, it is necessary to select a limited number of building blocks on a rational basis for both broad and targeted screening projects. However, the criteria for the selection should be different in the two cases.

To minimize the experimental efforts for a broad screening project, one would like to synthesize libraries with a minimum number of maximally diverse chemical compounds. Adequate computational methods should be developed that could exhaustively search, if feasible, or randomly sample the whole chemical structural space covered by virtual libraries and select building blocks that would afford combinatorial synthesis of representative and nonredundant chemical libraries.

On the contrary, in the case of a targeted screening project, one would like to synthesize a library of chemical structures with a desired biological activity. Rational design of targeted libraries can be achieved by taking advantage of the available information about the target, such as three-dimensional (3D) structures of target enzymes and/or chemical structures of known lead compounds. Thus, computational library design in this case should be directed toward finding compounds with stereochemical complementarity to the active site of the target and/or compounds that are structurally similar to the available lead molecules.

Computational analysis of chemical diversity in the context of combinatorial library design should address several issues. First, chemical structures should be characterized by calculable molecular descriptors that provide quantitative representation of chemical structures. Second, special measures should be developed on the basis of these descriptors to quantify structural similarity between pairs of molecules. Finally, adequate methods should be established for the effficient sampling of huge combinatorial structural space of chemical libraries.

Many descriptors of chemical structures have been developed over the years, primarily for use in quantitative structure-activity relationships (QSAR). The examples of such descriptors include various measurable physicochemical parameters such as partition coefficients, molar refractivities, and various quantum mechanical quantities such as HOMO and LUMO energies and electrostatic potentials. However, in chemical library design, compounds are yet to be synthesized and, therefore, they can not be characterized by experimental parameters *a priori.* Although quantum mechanical quantities can be calculated for the virtual library compounds, the high computational cost makes this approach impractical. One possible computationally efficient way for the representation of chemical structures is to calculate various topological[3] or topographical descriptors[4] based on two-dimensional (2D) or 3D representations of chemical structures, respectively. The relative merit of various 2D and 3D descriptors in terms of diversity analysis was discussed in recent papers.[5,6]

Several groups have been developing computational approaches to combinatorial chemical library design. Sheridan

* To whom correspondence should be addressed.

S0095-2338(97)00095-4 CCC: $15.00
Published on Web 03/04/1998

and Kearsley[7] used genetic algorithms, atom pair descriptors, and trend vectors as similarity metrics for targeted library design. Martin et al.[8] described their work on "flower diagram" for diversity analysis. More reports[9−11] appeared recently addressing various aspects of molecular diversity analysis in the context of chemical library design and database mining.

Herein, we propose a new method for targeted library design that we call Focus-2D (the suffix is used because the descriptors currently used by the program are obtained from a 2D topological description of molecules). (This method was first discussed at the 211th American Chemical Society Meeting.[12]) Focus-2D employs several different strategies for the effective sampling of virtual chemical libraries to select compounds expected to be active. The implementation described herein uses modified Euclidean distance in multidimensional descriptor space as a measure of chemical similarity to a lead molecule and a simulated annealing algorithm for sampling combinatorial structural space. In the accompanying paper,[13] we also describe an implementation that uses genetic algorithms for sampling and a pre-constructed QSAR equation for the prediction of biological activities of virtual library compounds.

We also describe an application of Focus-2D to the design of a tripeptoid library. The experimental work on this library was described by Zuckermann et al.[14] who have shown that a few members of the library had high affinities for α_1-adrenergic or μ-opiate receptors. The results of that work were used as a test case to evaluate the effectiveness of Focus-2D. We showed that when a peptoid was used as a lead compound, the program converged to finding the peptoid itself very rapidly and it also identified building blocks found in other known active peptoids. When *met*-enkephalin was used as a lead compound, Focus-2D suggested almost all building blocks found in peptoids with opioid activity. When morphine was used as a lead compound, Focus-2D also suggested several building blocks that were found in active opioid peptoids. Standard hypothesis testing indicated that the results of Focus-2D were statistically significant. We suggest that Focus-2D may serve as a useful tool in the rational design of targeted chemical libraries.

COMPUTATIONAL DETAILS

General Design of Focus-2D. Focus-2D has been designed to select building blocks that can be used to construct targeted libraries with a high content of compounds similar to a lead molecule (or several lead compounds). This approach is based on the assumption that similar compounds have similar biological activities. Thus, it implements the principle of analogue design that is widely employed in traditional medicinal chemistry. The virtual library compounds are generated by a random combination of available building blocks based on the underlying combinatorial chemical reaction, and the resulting virtual compounds are represented by Kier−Hall topological descriptors[15] (other descriptors can be implemented as well). Focus-2D samples the structural space of virtual libraries using simulated annealing (SA) or genetic algorithms (GA) and attempts to maximize similarity between computationally "synthesized" molecules and the lead molecule. Finally, frequency distribution analysis of building blocks found in the molecules with the highest similarity to a lead molecule is performed, and the building blocks found more frequently than random expectation are suggested as candidates for combinatorial synthesis.

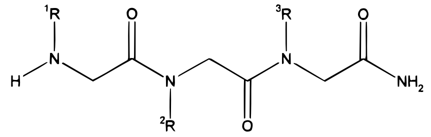

Figure 1. Markush structure of tripeptoids. ^{1}R, ^{2}R and ^{3}R are positions where various side chains (building blocks) are attached.

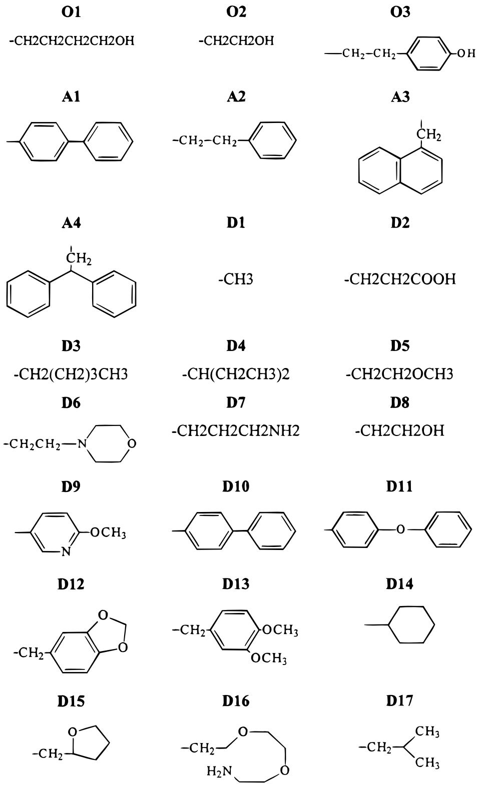

Figure 2. The chemical structure of building blocks.

Building Blocks. Computational methods described in this work were tested with a tripeptoid combinatorial library described by Zuckermann et al.[14] These authors described chemical structures of 24 amines used as building blocks for the peptoid synthesis. Three primary amines of different chemical nature (ethyl-, propyl-, and isopropylamine, which are referred to below as amine 1, amine 2, and amine 3, respectively) were included as additional building blocks for this computational experiment. The common Markush structure of tripeptoids is shown in Figure 1 where ^{1}R, ^{2}R, and ^{3}R are the alkyl portions of primary amines used as building blocks. The structures of the building blocks are shown in Figure 2, and we followed the abbreviations used in the original publication.[14]

Generation of Virtual Library Compounds. A general "computational synthesizer" of virtual molecules implemented in our program uses the common structural skeleton (Markush structure) of the underlying reaction products as the template. The program reads in building blocks represented in the form of connection tables and attaches them to the Markush structure. Thus, virtual molecules are generated that are also represented in the form of connection tables. These connection tables are used by the MolConnX program[15] to generate topological descriptors. This computational synthesizer algorithm is sufficiently generic so that various combinatorial chemical reactions can be implemented easily in Focus-2D.

In this experiment, building blocks were randomly combined to generate the virtual library compounds. Therefore, we allowed any building block to occupy any position. This strategy was different from the experimental approach of Zuckermann et al.[14] who did not allow building blocks of the same type to occupy two or three positions in tripeptoids simultaneously (e.g., AAD or DDA would not be synthesized).

Quantitative Analysis of Molecular Similarity. Each sampled molecule of the virtual library was represented by a set of topological descriptors calculated with the MolConnX program.[15] Several descriptors were eliminated because of their dependency on the arbitrary numbering of atoms in the molecule. The remaining 312 descriptors were used to characterize the chemical identity of each molecule in the library.

Molecular similarity between any two molecules was measured by the Euclidean distance between them in M-dimensional descriptor space (where M is the total number of applicable descriptors). Thus, each compound can be represented by a point in this space with the coordinates X_1, X_2, ..., X_k,...X_M, where X_ks are the absolute values of individual descriptors. The Euclidean distance $d_{i,j}$ between two points i and j (which correspond to compounds i and j) in M-dimensional space can be calculated as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2} \tag{1}$$

In principle, this distance can be considered as a quantitative measure of molecular similarity. However, the contribution of individual descriptors to this distance may be misrepresented because of the different absolute range of values for each descriptor. To scale contributions from individual descriptors to the total value of the distance, the difference between each pair of descriptors X_{ik} and X_{jk} was divided by the average of the two-descriptor values. Thus, the modified Euclidean distance used in this work was calculated as follows:

$$d'_{i,j} = \sqrt{\sum_{k=1}^{M} \left(\frac{X_{ik} - X_{jk}}{(X_{ik} + X_{jk})/2} \right)^2} \tag{1'}$$

Sampling of the Virtual Combinatorial Structural Space. For a small collection of building blocks, there is only a small number of compounds that can be synthesized. In such cases, a straightforward exhaustive search of the structural space can be applied. However, in general case of huge combinatorial structural space, the exhaustive search may become computationally very expensive. Focus-2D implements two most efficient stochastic search methods, GA and SA. These algorithms are used to optimize the selection of building blocks for the synthesis of virtual library compounds. The selection can be based either on their similarities to a lead molecule or on their activities predicted from a preconstructed QSAR. The implementation of GA- and QSAR-based optimization is described in the accompanying paper.[13] The general flow chart of Focus-2D is shown in Figure 3.

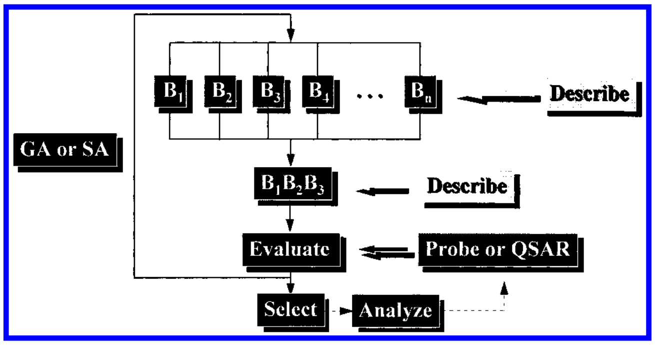

Figure 3. The flow chart of Focus-2D.

The idea of SA is to simulate the physical process called annealing, in which a system is heated to a high temperature and then is gradually lowered to a preset temperature value (e.g., room temperature). During this process, the system samples possible configurations according to Boltzmann distribution. At equilibrium, low energy states will be mostly populated. The first implementation of the SA procedure was described by Metropolis,[16] followed by the development of more general mathematical optimization method.[17] The implementation of SA in Focus-2D is as follows.

1. A lead molecule is selected and its topological descriptors are generated with the MolconnX program.[15]

2. A virtual chemical structure ($B_jB_jB_k$) is generated by random combination of any three of the available building blocks B_1, B_2, ..., B_n, and its topological descriptors are generated.

3. The dissimilarity value between this structure and that of the lead compound is calculated based on eq 1′ and is denoted as D_{curr}.

4. A new virtual structure is generated by random substitution of any one or any two building blocks in the current structure by one or two different building blocks randomly selected from the available pool, and its topological descriptors are generated.

5. The dissimilarity value between the new structure and that of the lead compound is calculated from eq 1′ and is denoted as D_{new}.

6. If $D_{\mathrm{new}} < D_{\mathrm{curr}}$, the new structure is accepted and used to replace the current structure. If $D_{\mathrm{new}} > D_{\mathrm{curr}}$, the new structure is accepted only if the following Metropolis criterion is satisfied; that is,

$$\mathrm{rnd} < \mathrm{e}^{-(D_{\mathrm{new}} - D_{\mathrm{curr}})/T} \tag{2}$$

where rnd is a random number uniformly distributed between 0 and 1, and T is a parameter analogous to the temperature in Boltzmann distribution law.

7. Steps 4–6 are repeated until the termination condition is satisfied. The temperature lowering scheme and the termination condition used in this work were adopted from Sun et al.[17] Thus, every time a new structure is accepted or a preset number of successive structure generation steps does not lead to a better structure, the temperature is lowered by 10% (the default initial temperature is 1000). The calculations are terminated when either the current temperature of simulations is lowered to the value of $T = 10^{-6}$ or the ratio

between the current temperature and the temperature corresponding to the best structure found (i.e., the structure with the highest similarity to the lead compound) is equal to 10^{-6}.

Analysis of Top Candidates. The SA-guided sampling of the virtual chemical library produces a set of compounds with the highest similarities to the lead molecule. The resulting set is then analyzed in terms of relative frequency of each building block, f_i, which is calculated as:

$$f_i = \frac{N_i}{N_t} \tag{3}$$

where N_i and N_t are the number of occurrences for a building block i and the total number of building blocks in the top scoring compounds, respectively. The value of f_i is compared with the expected frequency (i.e., if the selection were random), which is calculated as 1/27 (because the total number of building blocks was 27). The building blocks with frequency of occurrence higher than random expectations are considered as candidates for combinatorial synthesis of the targeted library.

Statistical Significance of Rational Selection of Building Blocks. The purpose of this analysis was to examine whether the number of building blocks correctly identified by Focus-2D was higher than what would be expected on the basis of random selection. We employed a standard hypothesis testing approach.[18] Two alternative hypotheses were formulated: H_0: $h = \mu$; H_1: $h > \mu$, where μ is the average number of *hits* (i.e., building blocks found in the reported active peptoids) obtained by random selection and h is the number of hits identified by the Focus-2D method. Thus, the null hypothesis H_0 stated that the Focus-2D-based selection was not significantly different from random selection and the alternative hypothesis H_1 assumed that Focus-2D performed significantly better than random selection. The decision making was based on the standard one-tail test that involved the following procedure:

1. Determine the average number of hits (μ) and its standard deviation (σ) for random selection via computer simulation as follows. A subset of i building blocks, where i is the number of hits identified by Focus-2D for a particular probe, was selected randomly 10 000 times from the pool of 27 available building blocks and the number of hits in each selection was identified. The values of μ and σ were calculated for the whole sampling.

2. Calculate Z score for the number of hits found by Focus-2D, where $Z = (h - \mu)/\sigma$.

3. Compare the Z score with the tabular critical values of Z_c at different levels of significance (α) to determine the level at which H_0 should be rejected. The frequently used α values and the corresponding critical values Z_c for one-tail test are given in Table 2. If the Z score is higher than a certain Z_c, one concludes, at the level of significance corresponding to that Z_c, that H_0 should be rejected and, therefore, H_1 should be accepted. This implies that the result obtained by Focus-2D is better than that obtained by random selection at the given level of significance.

RESULTS AND DISCUSSION

We have applied Focus-2D to the design of a tripeptoid library developed by Zuckermann et al.[14] The structures of several active peptoids found by Zuckermann et al.[14] are shown in Table 1. CHIR2279, CHIR2283, and CHIR2276 are high-affinity ligands for the α_1-adrenergic receptor, whereas CHIR4531, CHIR4534, and CHIR4537 were found to have high affinity for μ-opiate receptor. We now discuss the results of our computational experiments using different hypothetical lead compounds.

Table 1. Structures of Peptoids with α_1-Adrenergic and μ-Opiate Activity[a]

compound ID	structure	K_i (nM)
CHIR2279	O3-D10-A2	5[b]
CHIR2283	O3-D11-A2	140[b]
CHIR2276	O3 -D10 -A1	310[b]
CHIR4531	A4-D12-O3	6[c]
CHIR4534	A4-D3-O3	46[c]
CHIR4537	A4-D13-O3	31[c]

[a] Data as reported in Zuckermann et al.[14] [b] Data for α_1-adrenergic receptor binding. [c] Data for μ-opiate receptor binding.

Table 2. Frequently Used α Values and the Corresponding Critical Values of Z_c for One-Tail Test[a]

α	Z_c
0.10	1.28
0.05	1.64
0.01	2.33

[a] Reference 18.

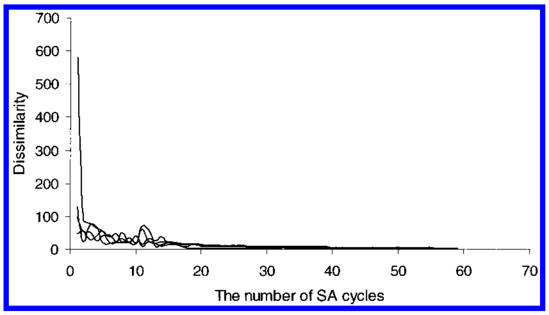

Figure 4. The SA trajectories of four independent Focus-2D runs using CHIR2276 as a "lead" compound.

Peptoid as a Lead Compound. This experiment was intended to examine the convergence and computational efficiency of Focus-2D; that is, its ability to find the hypothetical lead compound that belongs to the same structural family as the products of the underlying combinatorial chemical synthesis. CHIR2276, a peptoid with adrenergic activity (Table 1), was chosen on a purely random basis as a hypothetical lead compound. The trajectories of four independent Focus-2D runs using different random seeds are shown in Figure 4. Three of these trajectories converged to zero dissimilarity value (i.e., the lead compound itself was found) after 30, 48, and 57 SA cycles, respectively, and the fourth trajectory apparently required a longer time to converge to the same value. However, all four trajectories converged to the same low dissimilarity value very rapidly after ~20 SA cycles. The fact that the search was successful in rapidly finding the lead compound is not surprising; it just illustrates the effectiveness of the SA sampling as implemented in Focus-2D. This result also suggests that in practice one should conduct several runs with different random seeds to secure the convergence of the results.

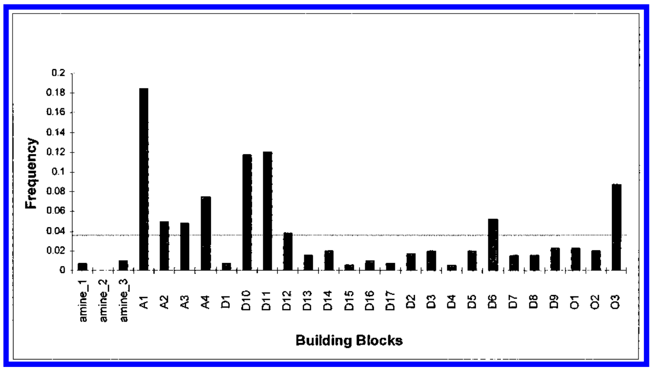

Figure 5. The averaged frequency distribution of building blocks in the final population of 50 compounds most similar to CHIR2276. The position of the dashed line corresponds to the expected (random) frequency of occurrence for every building block.

We have combined the results of all four runs and calculated the average frequency distribution (Figure 5) of building blocks in the top 50 compounds (i.e., those with the highest similarity to CHIR2276) obtained from all four trajectories. The most frequently occurring building blocks included A1, D10 (which is identical to A1), O3, D11, and A4 (cf. Figure 2). Building blocks with frequency of occurrence above random expectation also included A2, A3, and D6. Thus, these should be the building blocks of choice for the combinatorial synthesis of compounds similar to CHIR2276. The identification of the first two building blocks is not particularly surprising, given the fact that A1 (D10) and O3 are the actual components of CHIR2276. However, it is interesting that we also identified A2, which was found in the other two reported adrenergic peptoids, CHIR 2279 and CHIR2283, and D11, which was found only in the latter peptoid (cf. Table 1). Thus, the program correctly identified all building blocks found in the reported adrenergic peptoids. An obvious suggestion can be made that peptoids containing A3 and D6 along with the building blocks found in the reported adrenergic peptoids may also have adrenergic activity; this prediction can be tested experimentally.

The results of this experiment show that the information obtained in the course of combinatorial chemical synthesis of a limited size library can then be used in the rational design of new chemical libraries generated with the same combinatorial chemical reaction. We believe that the application of this iterative design strategy in combinatorial chemical synthesis would help to increase the efficiency of the lead identification process.

Peptide as a Lead Compound. The main objective of this experiment was to demonstrate that a peptide lead compound could be used in rational design of a nonpeptide library. This practice is actually common in the drug discovery process because most peptides can not be used as orally available drugs because of a high degree of biodegradation.

One of the natural opiates, *met*-enkephalin, was used as a hypothetical lead compound. The trajectories of four Focus-2D runs are shown in Figure 6. Although the starting points were different, all trajectories converged quite rapidly. Three of the four runs converged to the same dissimilarity value of ~25 at least after 40 SA cycles. The dissimilarity value for the fourth trajectory after 40 SA cycles was almost the same.

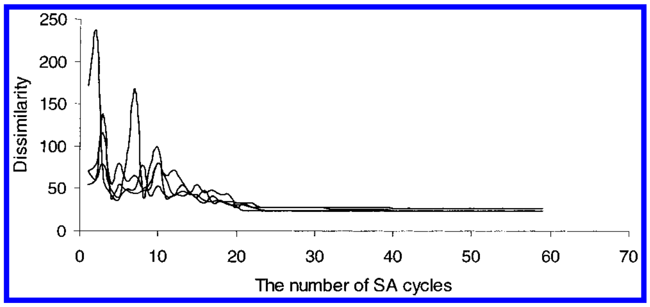

Figure 6. The SA trajectories of four independent Focus-2D runs using *met*-enkephalin as a "lead" compound.

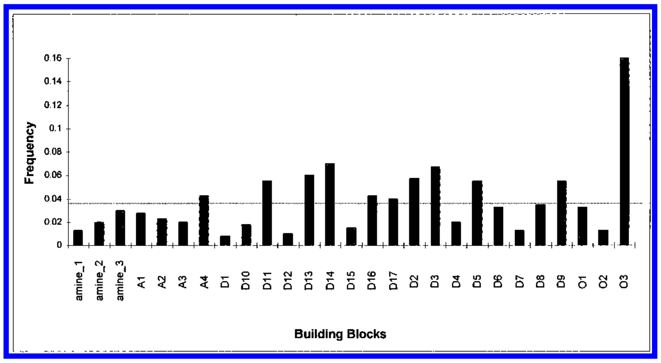

Figure 7. The averaged frequency distribution of building blocks in the final population of 50 compounds most similar to *met*-enkephalin. The position of the dashed line corresponds to the expected (random) frequency of occurrence for every building block.

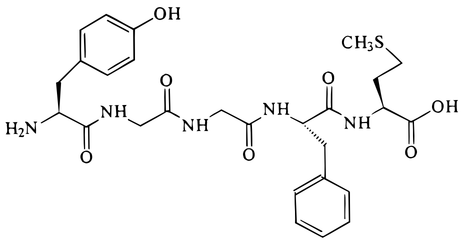

Figure 8. Chemical structure of *met*-enkephalin.

The averaged frequency distribution based on all four runs is shown in Figure 7. Based on this result, O3 had the highest frequency, and the frequencies of A4, D11, D13, D14, D16, D2, D3, D5, and D9 were also above random expectation. Apparently, O3 appeared in all the reported active peptoids with opioid activity (cf. Table 1). Comparison of the structure of *met*-enkephalin (Figure 8) with that of O3 indicated that O3 is similar to the side chain of tyrosine, which is the *N*-terminal residue of *met*-enkephalin. Among other building blocks found more frequently than random expectation, A4, D3, and D13 are present in the reported opioid peptoids (cf. Table 1). Thus, Focus-2D correctly identified four out of five building blocks found in the active peptoids (we did not identify D12). In addition to these building blocks, Focus-2D also selected D2, D3, D5, D9, D11, D14, and D16. Once again, this prediction could be tested experimentally. We also note that the universal nature of topological descriptors (applicable to molecules of different chemical classes) affords application of Focus-2D to the design of nonpeptide chemical libraries using peptides as lead compounds.

Morphine as a Lead Compound. This experiment represents a scenario when an organic lead compound is

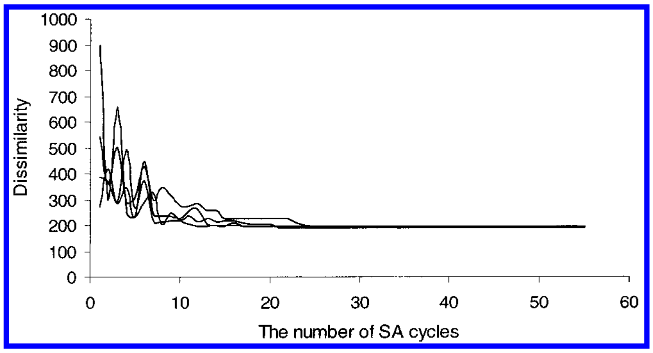

Figure 9. The SA trajectories of four independent Focus-2D runs using morphine as a “lead” compound.

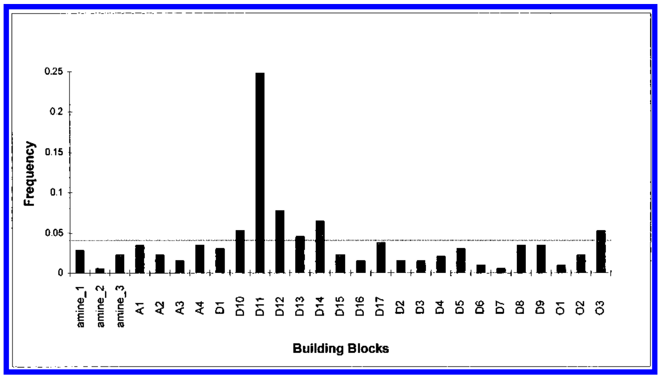

Figure 10. The averaged frequency distribution of building blocks in the final population of 50 compounds most similar to morphine. The position of the dashed line corresponds to the expected (random) frequency of occurrence for every building block.

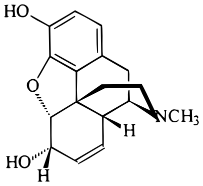

Figure 11. Chemical structure of morphine.

Table 3. Assessment of the Statistical Significance of Rational Selection of Building Blocks for a Peptoid Library with Opiate-Like Activity[a]

	Focus-2D			Random Selection	
lead molecule	no. of hits[b]	Z score	level of significance, α	average no. of hits, μ	standard deviation, σ
morphine	2 (5)	1.11	0.13	1.08	0.83
met-enkephalin	4 (10)	1.81	< 0.05	2.19	1.00
morphine and *met*-enkephalin	5 (12)	2.36	< 0.01	2.59	1.02

[a] Five different building blocks were found in the active opioid peptoids (cf. Table 1). [b] The number in parentheses is the total number of building blocks suggested by Focus-2D.

available. We chose morphine, a known opiate receptor ligand of nonpeptide chemical nature, as a hypothetical lead compound. The trajectories of four Focus-2D runs are shown in Figure 9. Again, starting from different random points, the four trajectories converged to the same or very comparable dissimilarity values after only 25 SA cycles. The averaged frequency distribution based on all runs is shown in Figure 10. The most frequent building block was D11. Building blocks D10, D12, D14, and O3 were less frequent, but all were above random expectation.

Comparison between the structures of D11 and morphine (Figure 11) makes it obvious that D11 is similar to a substructure of morphine. This result indicates that Kier–Hall topological indices can recognize similar compounds in terms of substructure similarity. Among other building blocks found more frequently than random expectation, D12 appears in CHIR4531 and O3 is found in all reported active peptoids (Table 1). Thus, in this test, we correctly identified two out of five building blocks found in the active opioid peptoids, and this result was statistically significant (*vide infra*). It is interesting that with morphine as a lead, we identified D12 (found in CHIR4531) that we missed when *met*-enkephalin was used as a probe. Thus, with two lead molecules of nonpeptoid nature, we could identify all five building blocks found experimentally in active opioid peptoids. The statistical significance of Focus-2D selection in this case is discussed next.

Assessment of Statistical Significance. The statistical hypothesis testing was performed for three different situations: using morphine as a probe, using *met*-enkephalin as a probe, and using both molecules as a composite probe. The results of the hypothesis testing are summarized in Table 3.

When *met*-enkephalin was used as a probe, 10 building blocks were selected by Focus-2D (cf. Figure 7). Accordingly, 10 building blocks were also chosen in the random selection experiment. The Z score was 1.81; this suggested (cf. Table 2) with 95% confidence that Focus-2D predictions were better than random selection of building blocks.

With morphine as a probe, Focus-2D identified five building blocks with frequencies of occurrence higher than the random expectation (cf. Figure 10). Thus, five building blocks were also selected randomly as described in Computational Details. The Z score was 1.11, which suggested (cf. Table 2) that Focus-2D was better than random with 87% confidence.

Finally, when both *met*-enkephalin and morphine were considered as a composite probe, 12 building blocks were suggested by Focus-2D (cf. Figures 7 and 10) and, therefore, the same number was targeted in the corresponding random sampling experiment. The Z score was 2.36, indicating that Focus-2D performed better than random sampling at >99% confidence.

The results of our statistical analysis clearly indicate that in all cases the rational selection was significantly better than random. The results obtained using *met*-enkephalin as a probe were better than those obtained using morphine probe. The fact that morphine and peptoids are chemically different, whereas peptides (*met*-enkephalin) and peptoids are rather similar, could probably explain this result. Perhaps, the application of different types of descriptors, (e.g., atom pairs[20]) could improve the situation, and we are currently investigating this option. On the other hand, the composite probe was clearly better than each of the individual probes, raising the statistical confidence level from 87% (morphine) or 95% (*met*-enkephalin) to >99%. These results suggest that the use of composite probes (if available) should be encouraged in rational library design.

Measures of Molecular Similarity. It is a normal practice in pattern recognition and cluster analysis that descriptors are auto-scaled or range-scaled to eliminate the size effect caused by different absolute values of individual descriptors.[19] Principal component analysis (PCA) is used frequently to reduce the dimensionality of the descriptor space. In this study, the distances were computed between a lead molecule and only those compounds sampled by Focus-2D. As a result, the calculation of the average and standard deviation of the descriptors for the whole population (i.e., all compounds) was not applicable and, therefore, the auto-scaling procedure was not possible. For the same reason, we did not use range-scaling, which required that the minimum and maximum values of each descriptor be calculated. PCA was not applicable either because it also required the analysis of the whole population. Thus, in our research, the size effect for each individual descriptor has been taken care of by dividing the difference between two descriptor values by the average of these descriptor values for the two structures under comparison (cf. eq 1′).

Euclidean distance measure has been widely used in clustering algorithms and similarity search programs.[4] However, this is not the only possible measure of chemical similarity. The general design of Focus-2D allows the use of other similarity measures such as Tanimoto coefficient and cosine coefficient.[4] These different distance metrics shall be explored and compared in future studies.

Rational Selection by Focus-2D versus Exhaustive Experimental Evaluation. The design of a peptoid library with nonpeptoid leads such as *met*-enkephalin and morphine represents an important practical instance of a rational library design. Indeed, the first case discussed in this paper (i.e., the design of a peptoid library using a peptoid as a lead) should be viewed more as a successful exercise to test the convergence and computational efficiency of Focus-2D. On the other hand, the use of *met*-enkephalin and morphine as leads could have been attempted, in principle, prior to the experimental synthesis and biological testing of the peptoid library developed by Zuckermann et al.[14] Our results indicate that Focus-2D proposed 12 building blocks on a rational basis (when a composite probe was used), which included all five building blocks found in the three reported active opioid peptoids (cf. Table 1 and Figures 7 and 10). Simple evaluation shows that if all combinations of building blocks were explored in a true sense of combinatorial chemical synthesis, as many as $24^3 = 13\,824$ compounds would have to be synthesized and tested. On the other hand, if the experiments were limited to using only 12 building blocks suggested by Focus-2D, only $12^3 = 1728$ compounds would have to be synthesized. Our results show that the same active compounds would be a part of this smaller library and therefore they would be identified. Thus, if the suggestions made by Focus-2D, using nonpeptoid leads, were accepted prior to the synthesis and testing, the total number of compounds subjected to experimental screening would be reduced by almost an order of magnitude. We believe that this illustrates the success of the rational approach implemented in Focus-2D as a means to reduce the amount of experimental efforts.

CONCLUSIONS

We have developed a Focus-2D approach to rational library design using a topological index based measure of molecular similarity measure and simulated annealing as the optimization tool. A similar approach to targeted library design was recently described by Sheridan and Kearsley.[7] However, they used different descriptors, different searching algorithms, and did not analyze their final building block selection in terms of statistical significance. Thus, our report, although similar in spirit extends, beyond the approaches discussed by Sheridan and Kearsley.

The successful application of the Focus-2D to the design of a peptoid library using lead compounds of different chemical nature suggests that Kier−Hall topological indices may serve as adequate descriptors to represent chemical identity. We conclude that Focus-2D can facilitate rational design of targeted chemical libraries in those cases when a lead molecule or several lead molecules are available.

Molecular recognition process obviously involves both ligands and receptors. In general, molecules that are viewed as similar by one target receptor could be dissimilar in the view of another receptor. This consideration describes a natural limitation of the chemical analogue design approach. The relative nature of molecular similarity with respect to biological function implies its intrinsic fuzziness. Further exploration of different descriptors and similarity metrics such as atom-pair,[20] *E*-state topological indices,[21] and 3D geometrical pair[22] should help in finding the most appropriate (or complementary) measures for the quantification of molecular similarity.

Special Note. All programs described in this paper can be obtained from the authors upon request.

ACKNOWLEDGMENT

The authors gratefully acknowledge Professors L. B. Kier and L. H. Hall for their interest in this work and for providing the MOLCONN-X program. This work was supported in part by PHS grant MH 40537 and Center grants HD03310 and MH33127. W. Z. acknowledges the 1996 Award from the Chemical Structure Association Trust and the graduate assistantship from EPA/UNC Toxicology Research Program, Training Agreement #T901915, with the Curriculum in Toxicology, UNC-CH.

REFERENCES AND NOTES

(1) Gallop, M. A.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(2) Gordon, E. M.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(3) Hall, L. H.; Kier, L. B. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: Deerfield Beach, FL, 1991; pp 367−422.

(4) Downs, G. M.; Willett, P. In *Reviews in Computational Chemistry 7*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: Deerfield Beach, FL, 1996; pp 1−66.

(5) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(6) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validating Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(7) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm To Suggest Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(8) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of

Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(9) Polinski, A.; Feinstein, R. D.; Shi, S.; Kuki, A. LiBrain: Software for Automated Design of Exploratory and Targeted Combinatorial Libraries. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Chaiken, I. M.; Janda, K. D., Eds.; ACS Conference Proceeding Series, 1996; pp 219−232.

(10) Turner, D. B.; Tyrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.

(11) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(12) Tropsha, A.; Zheng, W.; Cho, S. J. Application of Topological Indices in Rational Design of Combinatorial Chemical Libraries. *Book of Abstracts, 211th ACS National Meeting, New Orleans, LA, March* 22−28; American Chemical Society: Washington, D.C., 1996; ClNF-068.

(13) Cho, S. J.; Zheng, W.; Tropsha, A. Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 259−268.

(14) Zuckermann, R. N.; Martin, E. J.; Spellmeyer, D. C.; Stauber, G. B.; Shoemaker, K. R.; Kerr, J. M.; Figliozzi, G. M.; Goff, D. A.; Siani, M. A.; Simon, R. J.; Banville, S. C.; Brown, E. G.; Wang, L.; Richter, L. S.; 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.

(15) *MOLCONN-X version 2.0*; Hall Associates Consulting: Quincy, MA.

(16) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(17) Sun, L-X.; Xie, Y-L.; Song, X-H.; Wang, J-H.; Yu, R-Q. Cluster Analysis by Simulated Annealing. *Computers Chem.* **1994**, *18*, 103−108.

(18) Gilbert, N. *Statistics.* W. B. Sounders: Philadelphia, PA, 1976.

(19) Kaufman, L.; Rousseeuw, P. J. *Finding Groups In Data: An Introduction To Cluster Analysis*; Wiley: New York, 1989.

(20) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(21) Hall, L. H., Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. J *Chem. Inf. Comput. Sci.* **1995**, *35*, 1074.

(22) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.

CI970095X