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, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599.

A strategy for rational design of targeted combinatorial libraries is described. The aim of this method is to select a subset of available building blocks for the library synthesis that are most likely to be present in active compounds. Building blocks that are used in the underlying combinatorial chemical reaction are randomly assembled to produce virtual combinatorial libraries. Individual library compounds are represented by various chemical descriptors. Stochastic algorithms (simulated annealing, genetic algorithms, neural net methods) are used to search the potentially large structural space of virtual chemical libraries in order to identify compounds similar to lead compound(-s). The selection of a virtual molecule as a candidate for the targeted library is based either on its chemical similarity to a biologically active probe or on its biological activity predicted from a pre-constructed QSAR equation. 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 compound(-s). This method is illustrated herein by rational design of the library with bradykinin potentiating activity. Twenty eight bradykinin potentiating pentapeptides were used as a training set for the development of a QSAR equation, and, alternatively, two active pentapeptides, VEWAK and VKWAP, were used as probe molecules. In each case, the frequency distribution of amino acids in the top 100 peptides suggested by the method resembles the frequency distribution of amino acids found in the active peptides. The results obtained after GA optimization also compared favorably with those obtained by the exhaustive analysis of all possible 3.2 millions pentapeptides.

1 Introduction

Rapid development of combinatorial chemistry and high throughput screening techniques in recent years has provided a powerful alternative to traditional approaches for lead generation and optimization. In traditional medicinal chemistry, these processes frequently involve purification and identification of bioactive ingredients of natural, marine, or fermentation products or random screening of synthetic compounds. This is often followed by a series of painstaking chemical modification or total synthesis of promising lead compounds, which are tested in adequate bioassays. On the contrary, combinatorial chemistry involves systematic assembly of a set of "building blocks" to generate a large library of chemically different molecules which are screened simultaneously in various bioassays.[1,2] In the case of targeted library design, the lead identification and optimization then becomes generating libraries with structurally diverse compounds

which are similar to a lead compound; the underlying assumption is that structurally similar compounds should exhibit similar biological activities. Conversely, structurally dissimilar compounds should exhibit very diverse biological activity profiles; thus the goal of the diverse library design is to generate libraries with maximum chemical diversity of the composing compounds.[3]

In many practical cases, the exhaustive synthesis and evaluation of combinatorial libraries becomes prohibitively expensive, time consuming, or redundant.[4] Herein, we describe a new approach to rational design of targeted chemical libraries called FOCUS-2D. This approach uses various descriptors of chemical structures, e.g. topological descriptors[5] and employs stochastic search algorithms and chemical similarity functions including, if available, a pre-constructed Quantitative Structure-Activity Relationship (QSAR) as a means of selecting virtual library compounds with high predicted biological activity. We describe an application of this methodology to rational design of a targeted library with bradykinin (BK) potentiating activity. 28 BK potentiating pentapeptides[6,7] were used as a training set to develop a QSAR equation that was employed to predict the bioactivity of virtual library peptides. Alternatively, two active pentapeptides, VEWAK and VKWAP, were used as similarity probe molecules. We show that amino acids suggested by FOCUS-2D as preferred building blocks are actually found most frequently in known active BK peptides. We also show that the results obtained with FOCUS-2D compare favorably with those obtained after an exhaustive analysis of all 3.2 million pentapeptides.

2. Computational details.

Biological Activity

The log relative activity index (RAI) values of bradykinin potentiating pentapeptides were used as dependent variables. The activity of VESSK was set to 1.0, and all other activities were expressed relative to this activity. The detailed description of the assay as well as the calculation of relative activity index values were described in the original publications.[6,7]

General details.

Figure 1 shows the schematic diagram of targeted combinatorial library design using FOCUS-2D which consists of description, evaluation, and optimization steps. Molconn-X program[5] was used to generate topological descriptors (indices) for pentapeptides. These descriptors have been developed by Kier and Hall on the basis of chemical graph theory (e.g.,[8]). Programs implemented in FOCUS-2D as well as

genetic algorithms-partial least squares (GA-PLS) routine for QSAR developed earlier[9] were written in C programming language. The descriptor variables were autoscaled prior to PLS[10,11] and GA-PLS[9] calculations. All calculations were done on the IBM RS6000 workstation (Model 340).

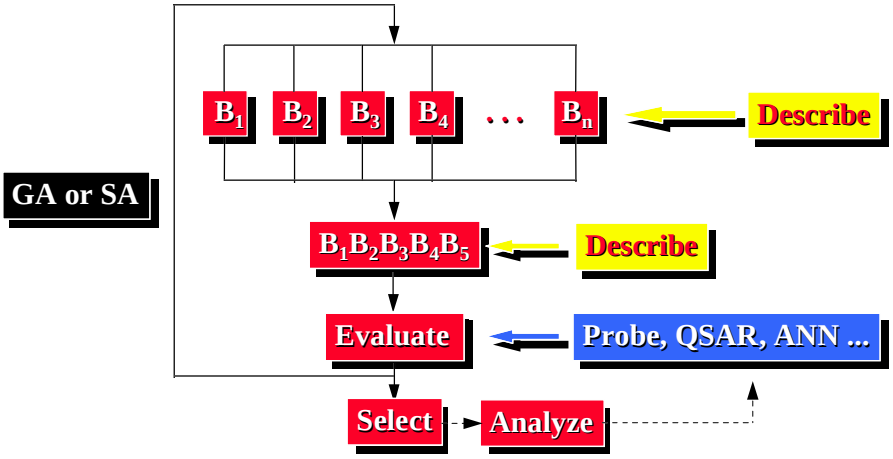


Figure 1. The schematic diagram of FOCUS-2D library design.

Structure description. The description step employs two different protocols where virtual compounds (in our example, pentapeptides, which are represented as $B_1B_2B_3B_4B_5$ in Figure 1) can be described either by topological indices or by a combination of physico-chemical descriptors, generated for each amino acid. The topological indices of assembled pentapeptide were calculated using the Molconn-X program.[5] The MOLCONN format,[5] which is the standard input file format for Molconn-X, was used to input the structure of each peptide: atom-id, the number of hydrogens connected, atom type, and atom-ids of all other heteroatoms are listed in a connection table separated by a comma for each heteroatom of the peptide. Amino acids were pre-described in this way, and the connection tables of selected amino acids were combined as necessary to construct the input file for Molconn-X.[5]

Alternatively, we have employed several amino acid based descriptors, including Z_1, Z_2, and Z_3 descriptors (related to hydrophilicity, bulk, and electronic properties of individual amino acids, respectively) reported by Hellberg *et al.*[6] and isotropic surface area (ISA) and electronic charge index (ECI) descriptors reported by Collantes and Dunn.[12]In this case, virtual pentapeptides are encoded in the form of a string of descriptor values. Each string consists of 15 descriptor values (five

blocks of three descriptors per amino acid) when using Z descriptors, or 10 descriptor values (five blocks of two) when using ISA-ECI descriptors.

Evaluation of the Virtual Peptide Library. The evaluation step employs different protocols to assess the fitness of each virtual pentapeptide. The fitness can be evaluated either by a peptide's chemical similarity to the biologically active peptide (probe), by the value of its biological activity predicted from a pre-constructed QSAR equation (inverse QSAR), or by other mapping techniques such as artificial neural network (not implemented yet).[13,14,15] The similarity of a peptide under evaluation to a biologically active probe is measured by the modified Euclidean distance between the two molecules calculated with the following equation:

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

where d_{ij} is the Euclidean distance between any pair of compounds *i* and *j*, M is the number of descriptors, and X_{ik} represents *k-th* descriptor.

As an alternative measure of fitness, the activity of the peptide under evaluation is predicted from the QSAR equation obtained using 28 pentapeptides as a training set. For peptides encoded using ISA-ECI and Z_1-Z_2-Z_3 descriptors, partial least squares (PLS)[10,11] and cross-validation[11] methods were used to construct QSAR equations (Table 1). For peptides encoded using topological indices, we used a novel QSAR method recently developed in our laboratory,[9] which utilizes genetic algorithms and PLS (GA-PLS; see short description below) (Table 1).

Library Optimization. In order to identify potentially active compounds, FOCUS-2D employs stochastic optimization methods such as Simulated Annealing (SA)[15,16,17] and Genetic Algorithms (GA).[18,19,20] The latter algorithm is employed here. Genetic algorithms implement two key concepts important in evolution: natural selection and sexual reproduction. For our optimization problem, these two concepts roughly translate into an iterative process which includes generation of a peptide population, evaluation of each peptide member of the population, mixing amino acids of members through crossover and mutations, and replacing low fitting members with high fitting offsprings to optimize the population. The detailed description of the optimization process is as follows.

Initially, a population of 100 peptides is randomly generated and encoded using topological indices or amino acid dependent physico-chemical descriptors, Z_1-Z_2-Z_3 or ISA-ECI. The fitness of each peptide is evaluated either by its chemical

similarity to a biologically active probe or by its biological activity predicted from a pre-constructed QSAR equation. Two parent peptides are chosen using the roulette wheel selection method (i. e., high fitting parents are more likely to be selected). Two offspring peptides are generated by a crossover (i. e., two randomly chosen peptides exchange their fragments) and mutations (i. e., a randomly chosen amino acid in an offspring is changed to any of 19 remaining amino acids). The fitness of the offspring peptides is then evaluated and compared with those of the parent peptides, and two lowest scoring peptides are eliminated. This process is repeated for 2000 times to evolve the population.

Table 1. Summary of statistics.

	PLS		GA-PLS		
	ISA-ECI[a]	Z_1-Z_2-Z_3[b]	Topological Indices[c]		
# of crossovers	0	0	0	2000	10000
# of compounds	28	28	28	28	28
# of variables	10	15	160	45	23
ONC[d]	3	2	1	2	5
Q^2 [e]	0.725	0.633	0.367	0.533	0.845
SDEP[f]	0.410	0.464	0.598	0.524	0.322
Fitness[g]	0.702	0.619	0.367	0.515	0.818
RSD of the X matrix[h]	0.886	0.818	0.381	0.134	0.195
SDEE[i]	0.313	0.315	0.544	0.466	0.260
R^2	0.840	0.831	0.476	0.630	0.899
F values	42.020	61.355	23.575	21.289	38.984

[a]ISA-ECI (n = 28, k = 3). [b]Z_1-Z_2-Z_3 (n = 28, k = 2). [c]Topological indices: (n = 28, k = 1) for 0 crossover; (n = 28, k = 2) for 2,000 crossovers; and (n = 28, k = 5) for 10,000 crossovers. [d]The optimal number of components. [e]Cross-validated R^2. [f]Standard error of prediction. [g][1 - (n - 1)(1 - q^2)/(n - c). [h]The residual SD of the X matrix. [i]Standard error of estimate.

GA-PLS Method for QSAR. The algorithm of the GA-PLS method[9] is implemented as follows. **Step 1**. The descriptors are generated using one of the available methods. **Step 2**. All applicable descriptors are enumerated arbitrarily, and this enumeration is maintained throughout the whole analysis. A population of 100 different random combinations of these descriptors is generated. In order to apply GA methodology, each combination is considered as a parent. Each parent represents a binary string of digits, either one or zero; the length of each string is

the same and is equal to the total number of descriptors (indices). The value of one implies that the corresponding descriptor is included for the parent, and zero means that the descriptor is excluded. **Step 3**. Using each parent combination of descriptors, a QSAR equation is generated for the whole dataset using the PLS algorithm; thus for each parent an initial value of q^2 is obtained. The [1 - (n - 1)(1 - q^2)/(n - c)] value, where q^2 is cross-validated r^2, n is the number of compounds, and c is the optimal number of components, is then used as the fitting function to guide GA. **Step 4.** Two parents are selected randomly based on the roulette wheel selection method (i. e., high fitting parents are more likely to be selected). **Step 5**. The population is evolved by performing a crossover between two randomly selected parents which produces two offsprings. **Step 6**. Each offspring is subjected to a random single-point mutation, i.e. randomly selected one (or zero) is changed to zero (or one). **Step 8.** The fitness of each offspring is evaluated as described above (cf. Step 4). **Step 9**. If the resulting offsprings are characterized by a higher value of fitness function, then they replace parents; otherwise, parents are kept. **Step 9.** Steps 4 - 8 are repeated until a predefined maximum number of crossovers are reached.

3. Results

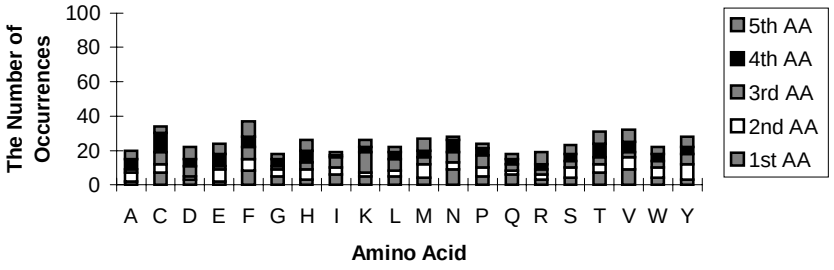
Generation of QSAR Models. The 28 bradykinin potentiating pentapeptides were included in the training set to generate QSAR equations using the GA-PLS method. The two most active compounds, VEWAK and VKWAP, were excluded from the training set. The calculated log RAI values compared favorably with the experimental data (data not shown). The equations correctly predicted them to have higher activities compared to activities of compounds in the training set (the log RAI values of 1.79, 1.48, and 1.47 were obtained for VEWAK using ISA-ECI, Z_1-Z_2-Z_3, and topological indices, respectively, and the log RAI values of 1.80, 1.74, and 1.95 were obtained for VKWAP using ISA-ECI, Z_1-Z_2-Z_3, and topological indices as descriptors, respectively).

The statistics obtained from the PLS regression analyses and the GA-PLS method applied to the training set using ISA-ECI, Z_1-Z_2-Z_3, and topological indices are shown in Table 1. In order to test the reliability of the prediction using pre-constructed QSAR equations with these descriptors, we incorporated the modified "degree of fit" condition. According to this condition, if RSD of dependent variables of a virtual peptide is less than the RSD of the X matrix of the training set, the predicted values are considered to be reliable. If this condition is not met the log RAI of the virtual peptide is not predicted or set to a low log RAI number to

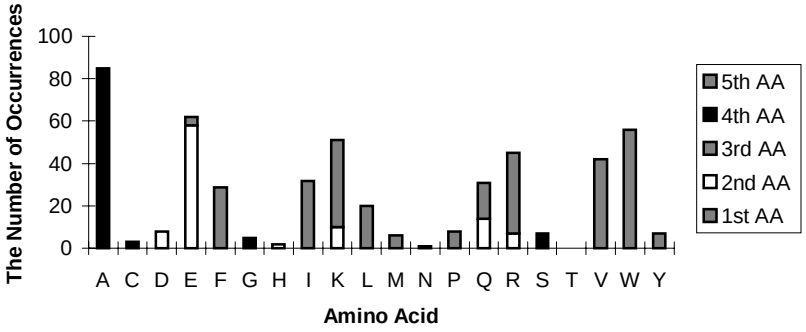
avoid selecting it. The condition does not allow the Focus-2D program to over-extrapolate. Since the number of peptides in the training set is very small compared to theoretical number of different pentapeptides (3.2 million), the extrapolation of QSAR relationship should be done very carefully in small increments, and the "degree of fit" condition implemented here allows us to do this. The RSD values (of the X matrix of the training set) of 0.886, 0.818, and 0.195 were obtained for ISA-ECI, Z_1-Z_2-Z_3, and topological indices description methods, respectively and used to test the reliability of the prediction (Table 1).

FOCUS-2D with ISA-ECI and Z_1-Z_2-Z_3 Description Methods. The distributions of amino acids in the initial and final populations, i.e., before and after Focus-2D, as well as after an exhaustive search using ISA-ECI and Z_1-Z_2-Z_3 amino acid based descriptors, were obtained. For brevity, the results using Z descriptors only are shown in Figures 2-3. The x- and y-axes of three bar graphs shown in each figure represent single letter coded amino acid names and the number of occurrences, respectively. The position of amino acid in a pentapeptide is described by different patterns.

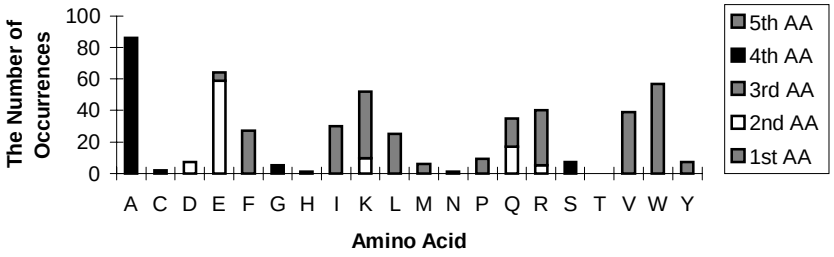
FOCUS-2D Using Similarity Probes VEWAK and VKWAP. As described in the computational details section, FOCUS-2D initially creates a population of 100 pentapeptides randomly. The random distribution of amino acids is important to ensure the unbiased evolution of the population; ideally the fraction of each amino acid in the initial population should be exactly the same. Figure 2a and 3a show the amino acid composition of the initial population before the FOCUS-2D was applied. These initial populations were then evolved with GA using VEWAK as the similarity probe. The amino acid composition of the final populations obtained after 2000 crossovers are shown in Figures 2b and 3b. Amino acids V, E, W, A, and K found in the probe are represented well in the population, and the preferred position of each amino acid is correctly identified. In addition, other selected amino acids largely include those that are chemically similar to amino acids found in the probe. In order to test whether the GA optimization method is sufficiently effective in searching through possible structure space, an exhaustive analysis of the whole population of 3.2 million pentapeptides was performed using both descriptors, and top 100 peptides most similar to VEWAK, were identified. The amino acid composition of the population containing these peptides is shown in Figures 2c and 3c. The resulting frequency distributions are very similar to those obtained with FOCUS-2D (cf. Figures 2b and 3b). Similar results were obtained with VKWAP as a probe.



(a)

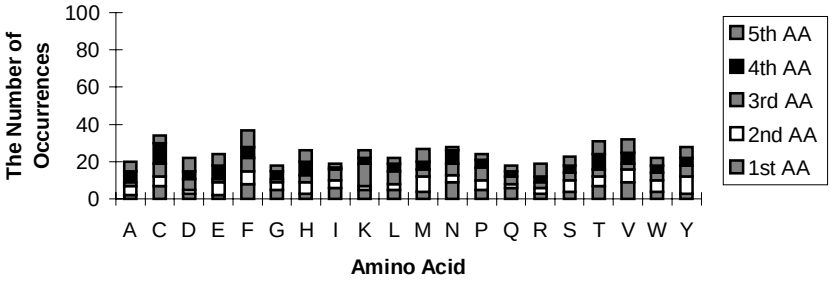


(b)

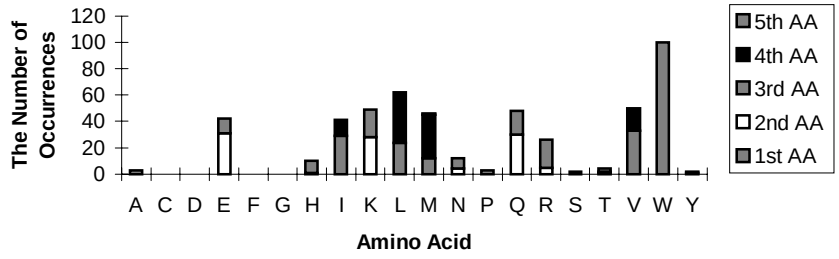


(c)

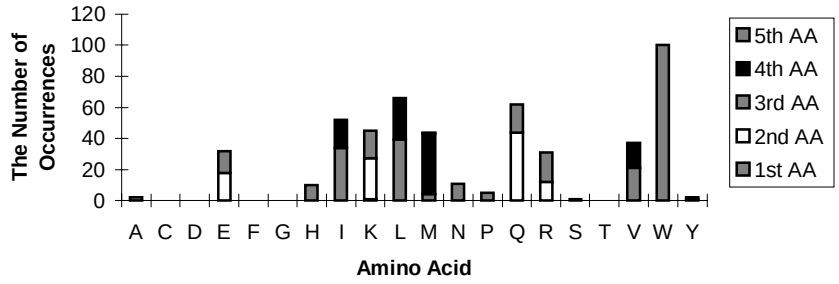
Figure 2. FOCUS-2D using Z_1-Z_2-Z_3 description method and VEWAK as the similarity probe: (a) initial population; (b) final population after FOCUS-2D; and (c) final population after the exhaustive search.



(a)



(b)



(c)

Figure 3. FOCUS-2D using Z_1-Z_2-Z_3 description method and a QSAR equation: (a) initial population; (b) final population after FOCUS-2D; and (c) final population after the exhaustive search.

FOCUS-2D Using QSAR Equation (Inverse QSAR). The results obtained with FOCUS-2D and a QSAR based prediction as the evaluation method are shown in Figures 3 for Z_1-Z_2-Z_3 descriptors. Again the populations before (Figure 3a) and after (Figures 3b) FOCUS-2D as well as the population after the exhaustive search (Figure 3c) are shown. The populations after FOCUS-2D and the exhaustive search were once again very similar to each other. With Z_1-Z_2-Z_3 descriptors, FOCUS-2D analysis selected amino acids E, I, K, L, M, Q, R, V, and W. Interestingly, these selected amino acids include most of those found in two most active pentapeptides, VEWAK and VKWAP, and the actual spatial positions of these amino acids are correctly identified: the first and fourth positions for V; the second and fifth positions for E; the third position for W; and the second and fifth positions for K.

4. Discussion

Combinatorial chemistry has emerged as a powerful approach in medicinal chemistry, providing researchers with a vast variety of chemical functionalities and assisting them in the identification and optimization of lead compounds. FOCUS-2D has been developed to enhance the rational design of chemical libraries. This method utilizes the existing SAR information in order to identify virtual library compounds with potentially high biological activity, and the building blocks frequently found in these virtual libraries are proposed to be used in targeted library synthesis. The current implementation of the program includes two different description (building block based and whole molecule based) and evaluation (similarity probe, QSAR prediction) protocols that are used along with either GA (this paper) or SA optimization methods. The key aspects of the algorithm are described in Figure 1.

In order to test this methodology, we have selected 30 bradykinin potentiating pentapeptides as a training set to design targeted library with bradykinin activity. Selection of a peptide data set was based on the fact that there are almost no published non-peptide combinatorial chemical libraries which contain SAR information.[21] In contrast, there is a large number of peptide datasets for which the experimental SAR information is available. An additional advantage of using a peptide dataset is that there are only 20 naturally occurring amino acids (building blocks) and the experimental approaches to peptide library synthesis are well developed.

As one of the ways to guide GA based selection process, the similarity of a virtually synthesized peptide to one of two active peptides, VEWAK or VKWAP,

was measured by its Euclidean distance to the probe. The results obtained with this fitting function show that those amino acids found in the similarity probe are indeed present in the final population as the dominant amino acids with their positions correctly identified most of the time (Figures 2-3). The identification of preferred positions of amino acids strongly depended on the types of descriptors used. The number of different suggested positions for each amino acid was less for the amino acid dependent descriptors than for the topological descriptors. This was somewhat expected since topological indices describe a peptide as a whole, so the identity of the amino acid in each position is described implicitly whereas the amino acid dependent descriptors encode the identity explicitly.

As discussed above, we have considered both amino acid and the whole molecule based descriptors. One major advantage of topological indices, as well as any whole molecule based descriptors, over amino acid based descriptors is that topological indices can also describe non-peptides. This is important point because peptides similar to a non-peptide probe or, alternatively, non-peptides similar to a peptide probe can be identified as well. Furthermore, a large number of QSAR studies available in literature can be used to direct combinatorial chemical library synthesis.

An obviously positive result of this work is that it proved the effectiveness of the GA optimization method. In all cases when using both two types of amino acid based descriptors and the inverse QSAR prediction method, the results of stochastic search were comparable to those obtained after an exhaustive search (cf. Figures 2-3): in each case, the amino acid composition of the final population obtained from FOCUS-2D was very similar to that obtained from the exhaustive search.

To the best of our knowledge, no experimental targeted library with bradykinin potentiating activity has been described in the literature yet. Thus, the present study provides practical suggestions for the rational design of such a library. Our predictions summarized in Figures 2-3 can be validated by the practical design and evaluation of the BK library(-ies). This experimental evaluation will also help us to determine the most adequate descriptors among three different types used in this work.

5. Acknowledgments

This work was supported in part by PHS grant MH 40537 and Center grants HD03310 and MH33127. WZ acknowledges the 1996 Award from Chemical Structure Association Trust and the graduate assistantship from EPA/UNC

Toxicology Research Program, Training Agreement #T901915, with the Curriculum in Toxicology, UNC-CH.

6. References

1. Gallop, M. A.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233-1251.
2. Gordon, E. M.; Barret, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. *J. Med. Chem.* **1994**, *37*, 1385-1401.
3. Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
4. Sheridan, R. P.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci*. **1995**, 35, 310-320.
5. MOLCONN-X version 2.0, Hall Associates Consulting, Quincy, MA.
6. Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.* **1987**, *30*, 1126-1135.
7. Ufkes, J. G. R.; Visser, B. J.; Heuver, G.; Van Der Meer, C. *Eur. J. Pharm.* **1978**, *50*, 119-122.
8. Hall, L. H.; Kier, L. B. In Reviews in Computational Chemistry II; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Publishers, **1991**, pp 367-422.
9. Cho, S. J.; Cummins, D.; Bentley, J.; Andrews, C. W.; Tropsha, A. *J. Comp. Aided Mol. Design.* submitted.
10. Dunn, W. J. III; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131-137.
11. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967
12. Collantes, E. R.; Dunn, W. J. III. *J. Med. Chem.* **1995**, *38*, 2705-2713.
13. Tetko, I. V.; Luik, A. I.; Poda, G. I. *J. Med. Chem.* **1993**, *36*, 811-814.
14. Ajay. A. *J. Med. Chem.* **1993**, *36*, 3565-3571.
15. So, S. S.; Richards, W. G. *J. Med. Chem.* **1992**, *35*, 3201-3207.
16. Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. *Technometrics* **1986**, *28*, 209-217.
17. Kalivas, J. H.; Sutter, J. M.; Roberts, N. *Anal. Chem.* **1989**, *61*, 2024-2030.
18. Goldberg, D. E. Genetic Algorithm in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, MA, 1989.
19. Holland, J. H. Genetic Algorithms. *Scientific American* **1992**, *267*, 66-72.
20. Forrest, S. *Science* **1993**, *261*, 872-878.
21. Zuckermann, R. N et al. *J. Med. Chem.* **1994**, *37*, 2678-2685.