Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches

Sung Jin Cho, Weifan Zheng, 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

Received March 31, 1997

We have developed a novel strategy for rational design of targeted peptide libraries. The goal of this method is to select a subset of natural amino acids that are most likely to be present in active peptides for the synthesis of library. Two different protocols are employed where chemical structures of peptides are described either by topological indices or by a combination of physicochemical descriptors for individual amino acids. The selection of a peptide 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 preconstructed quantitative structure−activity (QSAR) equation. The optimization of the library is achieved by means of genetic algorithms (GA). This method was tested 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 million pentapeptides.

INTRODUCTION

Rapid development of combinatorial chemistry and high throughput screening techniques in recent years has provided a powerful alternative to traditional approaches to lead generation and optimization. In traditional medicinal chemistry, these processes frequently involve purlfication and identification of bioactive ingredients of natural, marine, or fermentation products or random screening of synthetic compounds. These processes are often followed by a series of painstaking chemical modification or total synthesis of promising lead compounds that 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 that are screened 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 that 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] Recently, we have initiated the development of computational approaches aimed at rational design of combinatorial libraries for both targeted and diverse screening. In the first paper of this series,[5] we introduced our new approach to rational design of targeted chemical libraries called Focus-2D. This approach uses various descriptors of chemical structures (e.g., topological descriptors[6]) and employs stochastic search algorithms and chemical similarity functions to optimize the selection of building blocks. In the previous paper,[5] we discussed the implementation of simulated annealing (SA) optimization method and a chemical similarity function based on Euclidean distances between compounds in multidimensional descriptor space. We show in this paper that as an alternative to the Euclidean distance-based similarity function, Focus-2D can use a preconstructed quantitative structure−activity relationship (QSAR) as a means of selecting virtual library compounds with high predicted biological activity. We refer to this approach as the inverse QSAR method. (This approach was first discussed at the 211th American Chemical Society meeting.[7]) We also show that genetic algorithms (GA) can be used as an alternative to SA for the optimization of building block selection.

To test this methodology, we designed a virtual targeted library with bradykinin (BK)-like activity. We used 28 BK-potentiating pentapeptides[8,9] as a training set for the development of QSAR equations that were used to estimate 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 were 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.

* To whom correspondence should be addressed.

S0095-2338(97)00094-2 CCC: $15.00
Published on Web 03/04/1998

Table 1. Actual, Calculated, and Predicted Log Relative Activity Index Values of 30 Bradykinin Potentiating Pentapeptides[8,9]

		description method					
		ISA-ECI[14]		Z_1-Z_2-Z_3[8]		topological indices[6]	
peptide	actual	calculated[a]	predicted[b]	calculated[a]	predicted[b]	calculated[a]	predicted[b]
VESSK	0.00	0.13	—[c]	0.04	—	0.11	—
VESAK	0.28	0.26	—	0.23	—	0.13	—
VEASK	0.20	0.04	—	0.07	—	0.19	—
VEAAK	0.51	0.17	—	0.26	—	0.23	—
VKAAK	0.11	−0.01	—	0.48	—	0.40	—
VEAAP	0.18	0.35	—	0.29	—	0.52	—
VEHAK	1.53	0.74	—	0.73	—	1.45	—
VAAAK	−0.10	−0.04	—	0.10	—	−0.12	—
GEAAK	−0.52	−0.33	—	−0.12	—	−0.58	—
LEAAK	0.40	0.33	—	0.31	—	0.17	—
FEAAK	0.30	0.51	—	0.25	—	0.50	—
VEGGK	−1.00	−0.25	—	−0.52	—	−0.99	—
VEFAK	1.57	1.12	—	1.11	—	1.59	—
VELAK	0.59	0.85	—	0.43	—	0.61	—
AAAAA	−0.10	−0.51	—	−0.11	—	−0.30	—
AAYAA	0.46	0.50	—	0.47	—	0.60	—
AAWAA	0.75	1.11	—	1.11	—	0.72	—
VAWAA	1.43	1.40	—	1.31	—	1.15	—
VAWAK	1.45	1.58	—	1.33	—	1.14	—
VKWAA	1.71	1.43	—	1.68	—	1.74	—
VWAAK	0.04	0.01	—	−0.06	—	0.36	—
VAAWK	0.23	0.13	—	0.26	—	0.93	—
EKWAP	1.30	1.59	—	1.54	—	1.42	—
RKWAP	1.98	1.77	—	1.96	—	1.62	—
VEWVK	1.71	1.91	—	1.89	—	1.67	—
PGFSP	0.90	1.03	—	0.53	—	0.80	—
FSPFR	0.64	0.66	—	1.29	—	0.76	—
RYLPT	0.40	0.46	—	0.08	—	0.13	—
VEWAK	2.73	—	1.79	—	1.48	—	1.47
VKWAP	2.35	—	1.80	—	1.74	—	1.95

[a] Computed from the QSAR equation for the training set compounds. [b] Computed from the QSAR equation for the compounds outside of the training set. [c] —, Not available.

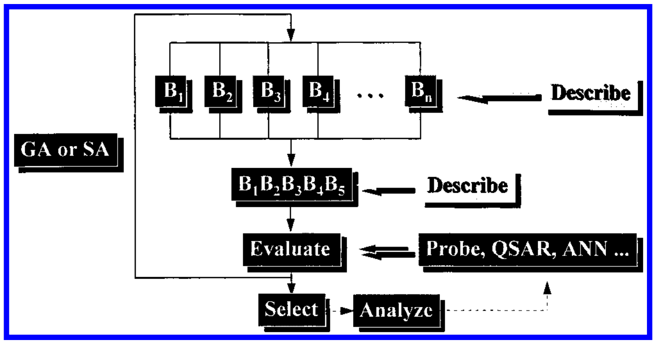


Figure 1. . The schematic diagram of Focus-2D library design.

METHODS

Biological Activity. The log relative activity index (RAI) values of BK-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. Thus, the activities of the training set span a range of three log units, which makes it a good dataset for a QSAR analysis. The detailed description of the assay as well as the calculation of RAI values were described in the original publications.[8,9]

Computational Details. Figure 1 is a schematic diagram of targeted combinatorial library design using Focus-2D, which consists of description, evaluation, and optimization steps. The Molconn-X program[6] was used to generate topological descriptors for pentapeptides (Table 1). These descriptors were developed by Hall and Kier[10] on the basis of chemical graph theory. Programs implemented in Focus-2D as well as GA-partial least squares (GA-PLS) routine for QSAR developed earlier[11] were written in C programming language. The descriptor variables were autoscaled prior to PLS[12,13] and GA-PLS[11] calculations. All calculations were done on the IBM RS6000 workstation (model 340).

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 physicochemical descriptors generated for each amino acid. The topological indices of assembled pentapeptide were calculated using the Molconn-X program.[6] The MOLCONN format,[6] 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 described in advance in this way, and the connection tables of selected amino acids were combined to construct the input file for Molconn-X.[6]

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.*[8] and isotropic surface area (ISA) and electronic charge index (ECI) descriptors reported by Collantes and Dunn[14] (Table 2). In this case, virtual pentapeptides were encoded in the form of a string of descriptor values. Each string consists of 15

Table 2. ISA-ECI[14] and $Z_1-Z_2-Z_3$[8] Descriptors used to Encode Amino Acids

	description method				
	ISA-ECI		$Z_1-Z_2-Z_3$		
amino acid	ISA	ECI	Z_1	Z_2	Z_3
Gly (G)	19.93	0.02	2.23	−5.36	0.30
Ala (A)	62.90	0.05	0.07	−1.73	0.09
Val (V)	120.91	0.07	−2.69	−2.53	−1.29
Leu (L)	154.35	0.10	−4.19	−1.03	−0.98
Ile (I)	149.77	0.09	−4.44	−1.68	−1.03
Phe (F)	189.42	0.14	−4.92	1.30	0.45
Tyr (Y)	132.16	0.72	−1.39	2.32	0.01
Trp (W)	179.16	1.08	−4.75	3.65	0.85
Pro (P)	122.35	0.16	−1.22	0.88	2.23
Asp (D)	18.46	1.25	3.64	1.13	2.36
Asn (N)	17.87	1.31	3.22	1.45	0.84
Glu (E)	30.19	1.31	3.08	0.39	−0.07
Gin (Q)	19.53	1.36	2.18	0.53	−1.14
Ser (S)	19.75	0.56	1.96	−1.63	0.57
Thr (T)	59.44	0.65	0.92	−2.09	−1.40
Cys (C)	78.51	0.15	0.71	−0.97	4.13
Met (M)	132.22	0.34	−2.49	−0.27	−0.41
Lys (K)	102.78	0.53	2.84	1.41	−3.14
Arg (R)	52.98	1.69	2.88	2.52	−3.44
His (H)	87.38	0.56	2.41	1.74	1.11

descriptor values (three descriptors per amino acid) when using Z descriptors, or 10 descriptor values (two descriptors per amino acid) 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 the chemical similarity of a peptide to the biologically active peptide (probe), or by the value of the biological activity of the peptide predicted from a preconstructed QSAR equation (inverse QSAR).[15−17] 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:

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

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 the kth descriptor (see our accompanying paper[5] for more details).

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, PLS[12,13] and cross-validation[13] methods were used to construct QSAR equations (Table 3). For peptides encoded using topological indices, we used a novel QSAR method recently developed in our laboratory,[11] which utilizes GAs and PLS (GA-PLS; see following short description; Table 3).

Library Optimization. To identify potentially active compounds, Focus-2D employs stochastic optimization methods such as SA[17−19] and GA.[20−22] The implementation of SA in Focus-2D was discussed in the accompanying paper.[5] 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 that 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 physicochernical descriptors (i.e., $Z_1-Z_2-Z_3$ 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 preconstructed QSAR equation. Two parent peptides are chosen using the roulette wheel selection method (i.e., high fitting parents are more likely to be selected), and 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.

GA-PLS Method for QSAR. The algorithm of the GA-PLS method[11] is implemented as follows. **Step 1.** The Molconn-X program[6] is applied to generate 462 variables (topological indices) automatically. **Step 2.** All atom ID-dependent descriptors (150 descriptors) and descriptors with zero variance are removed (the atom ID-dependent indices are eliminated because atom IDs are assigned arbitrarily). The resulting number of descriptors initially used for 28 BK-potentiating pentapeptides was 160. **Step 3.** A population of 100 different random combinations of these descriptors is generated. To apply GA methodology, each combination is considered as a parent. Each parent represents a binary string of either one or zero; the length of each string is the same and equal to the total number of descriptors. The value of one implies that the corresponding descriptor is included for the parent, and zero means that the descriptor is excluded. **Step 4.** 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 value of $[1 - (n - 1)(1 - q^2)/(n - c)]$ (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. **Step 5.** Two parents are selected randomly based on the roulette wheel selection method (i.e., high fitting parents are more likely to be selected). **Step 6.** The population is evolved by performing a crossover between two randomly selected parents, which produces two offsprings. **Step 7.** Each offspring is subjected to a random single-point mutation; that is, randomly selected one (or zero) is changed to zero (or one). **Step 8.** The fitness of each offspring is evaluated as already described (cf. Step 4). **Step 9.** If the resulting offsprings give higher values of fitness function, then they replace parents; otherwise, parents are kept. **Step 10.** Steps 5−9 are repeated until a predefined maximum number of crossovers is reached. This algorithm is described in more detail in our previous paper[11] and is available from the QSAR WWW server maintained by the authors at http://mmlinl.pha.unc.edu/∼jin/QSAR.

Table 3. Summary of Statistics

	GA-PLS		GA-PLS		
parameter	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 2000 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] $[1 - (n - 1)(1 - q^2)/(n - c)]$ (see *Computational Detail* section). [h] The residual SD of the X matrix (see the *Computational Detail* section). [i] Standard error of estimate.

The Degree of Fit. The reliability of the prediction by a QSAR equation was evaluated with the modified "degree of fit" method developed originally by Lindberg *et al.*[23] According to the original method, the predicted y values are considered to be reliable if the following condition is met:

$$s^2 < s_a^{\,2}(E_x)F \qquad (2)$$

where s^2 is the residual standard deviation (RSD) of descriptor values generated for a test compound, $s_a^2(E_x)$ is the RSD of the X matrix after dimensions (components) a, and F is the F statistics at the probability level α and $(p - a)/2$ and $(p - a)(n - a - 1)/2$ degrees of freedom. The RSD of descriptor values generated for a test compound is calculated using the following equation:

$$s^2 = \|e\|/(p - a) \qquad (3)$$

where p is the number of x-variables, a is the number of components, and $\|e\|$ is the sum of squared residuals e_i expressed as follows:

$$e_i = x_i - x_iBB' \qquad (4)$$

where x_i is the ith x-variable, and B and B' represent the weight matrix and transposed weight matrix of x variables, respectively. Because the lowest possible value of F is 1.00 at $\alpha = 0.10$ (when both degrees of freedom are equal to infinity), we decided to replace F with the "degree of fit" factor f to simplify the aforementioned condition. Thus, Focus-2D implements the following modified "degree of fit" condition: predicted y values are considered to be reliable if:

$$s^2 < s_a^{\,2}(E_x)f \qquad (5)$$

The "degree of fit" factor was set to 1.0 (Figures 6, 7, and 10), 0.7 (Figure 11a), 0.5 (Figure 11lb), or 0.3 (Figure 11c). This modified "degree of fit" condition is much more stringent than the original one[23] and is especially appropriate for our program.

RESULTS

Generation of QSAR Models. The 28 BK-potentiating pentapeptides were included in the training set to generate QSAR equations using the GA-PLS method.[11] 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 (Table 1). The log RAI values for VEWAK and VKWAP predicted from QSAR equations, constructed with three different types of descriptors, are also listed in Table 1. Although the predicted values for these peptides differ from the actual values, the equations correctly predict them to have higher activities compared with activities of compounds in the training set [the log RAI values of 1.79, 1.48, and 1.47 are obtained for VEWAK using ISA-ECI, $Z_1-Z_2-Z_3$, and topological indices as descriptors (Table 2), respectively, and the log RAI values of 1.80, 1.74, and 1.95 are 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 3. To test the reliability of the prediction using preconstructed QSAR equations with these descriptors, we incorporated the modified "degree of fit" condition (see *Computational Details* section). 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. This condition does not allow the Focus-2D program to over-extrapolate. Because the number of peptides in the training set is very small compared with theoretical number of different pentapeptides (3.2 million), the extrapolation of the 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 3).

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 Focus2-D), as well as after an exhaustive search using ISA-ECI and $Z_1-Z_2-Z_3$ amino acid-based descriptors, are shown in Figures 2−7. The

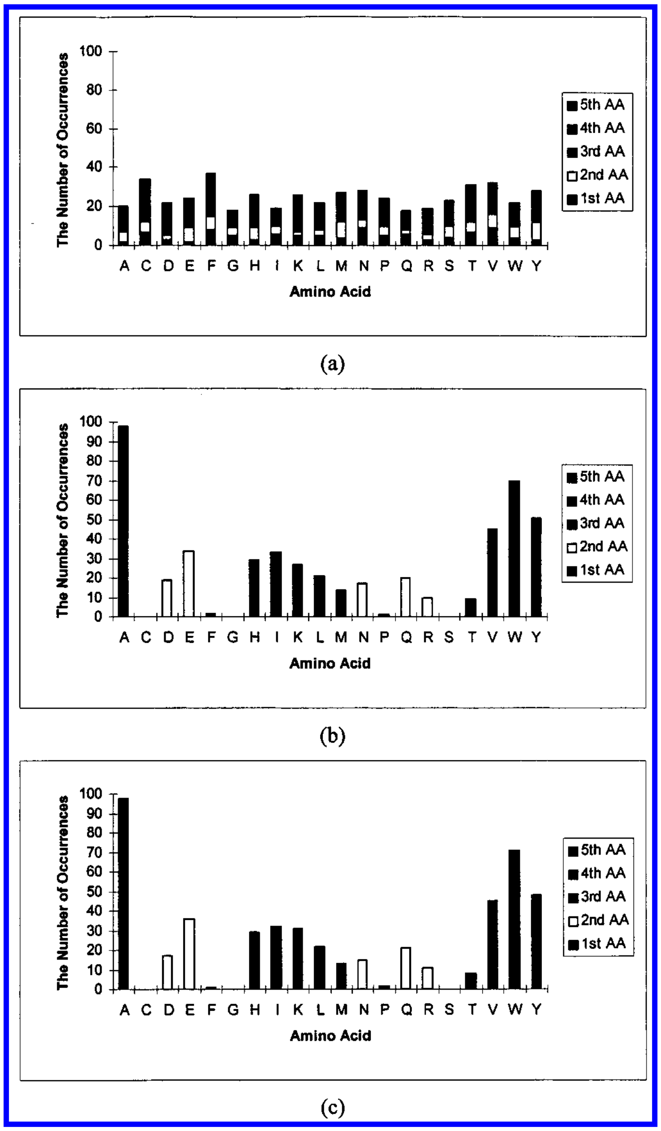


Figure 2. Focus-2D using ISA-ECI description method and VEWAK as the similarity probe: (a) initial population; (b) final population after Focus-2D; and (c) final population after the exhaustive search of 3.2 million peptides.

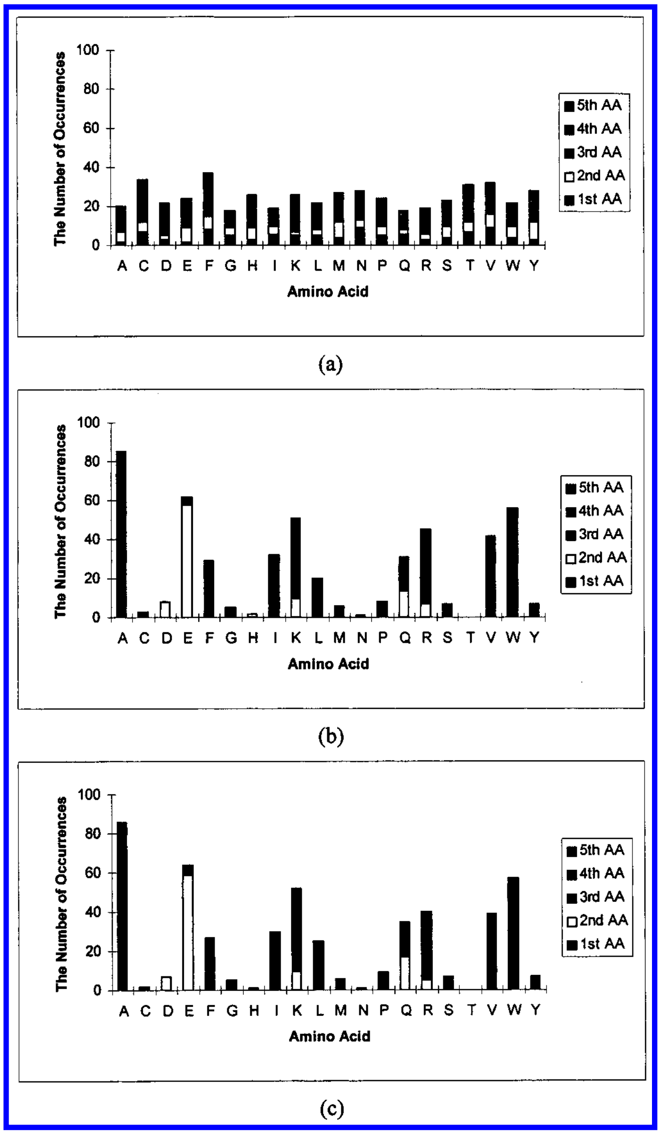
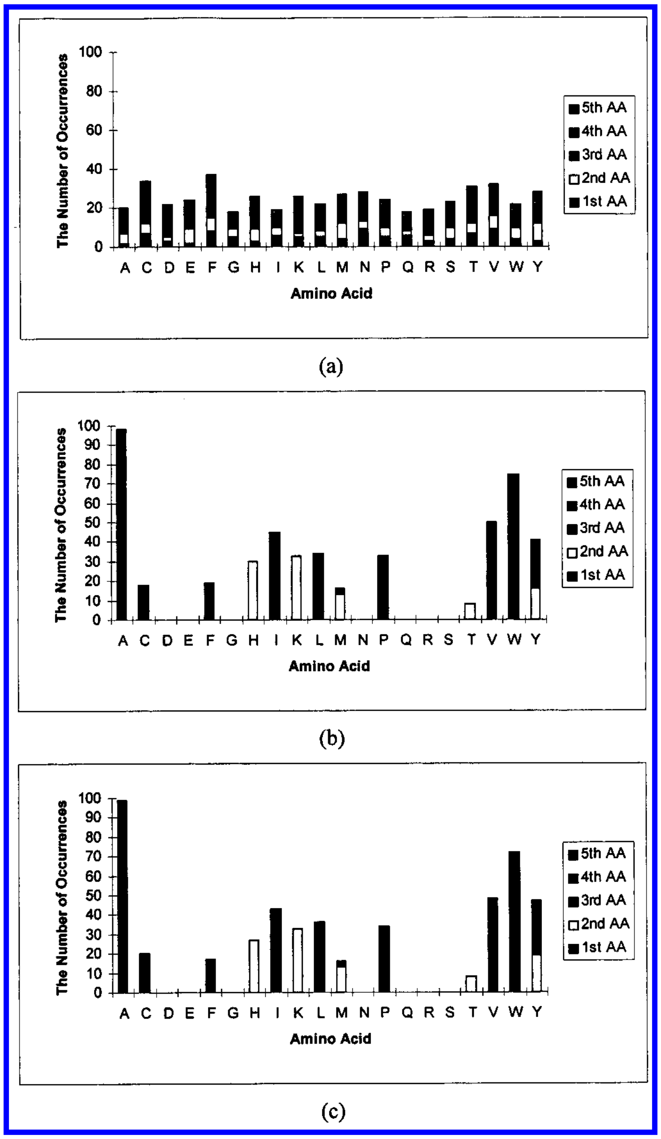


Figure 3. Focus-2D using $Z_1-Z_2-Z_3$ description method and VEWAK as the similarity probe: (a) initial population; (b) final population after Focus-2D; and (c) final population after the exhaustive search.

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. Figures 2a, 3a, 4a, and 5a show the amino acid composition of the initial population before 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. 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 the 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).

In the analogous fashion, the initial random population of 100 pentapeptides was evolved using VKWAP as the similarity probe (Figures 4a and 5a). The amino acid composition of the final population again reflected the dominance of V, K, W, A, and P as well as those amino acids that are similar to them (Figures 4b and 5b). The preferred position of each amino acid is the one found in the probe (V, K W, A, and P prefer the first, second, third, fourth, and fifth positions, respectively). The final population was again compared with amino acid composition derived from the top 100 peptides that are most similar to VKWAP on the basis of exhaustive search (Figures 4c and 5c). For

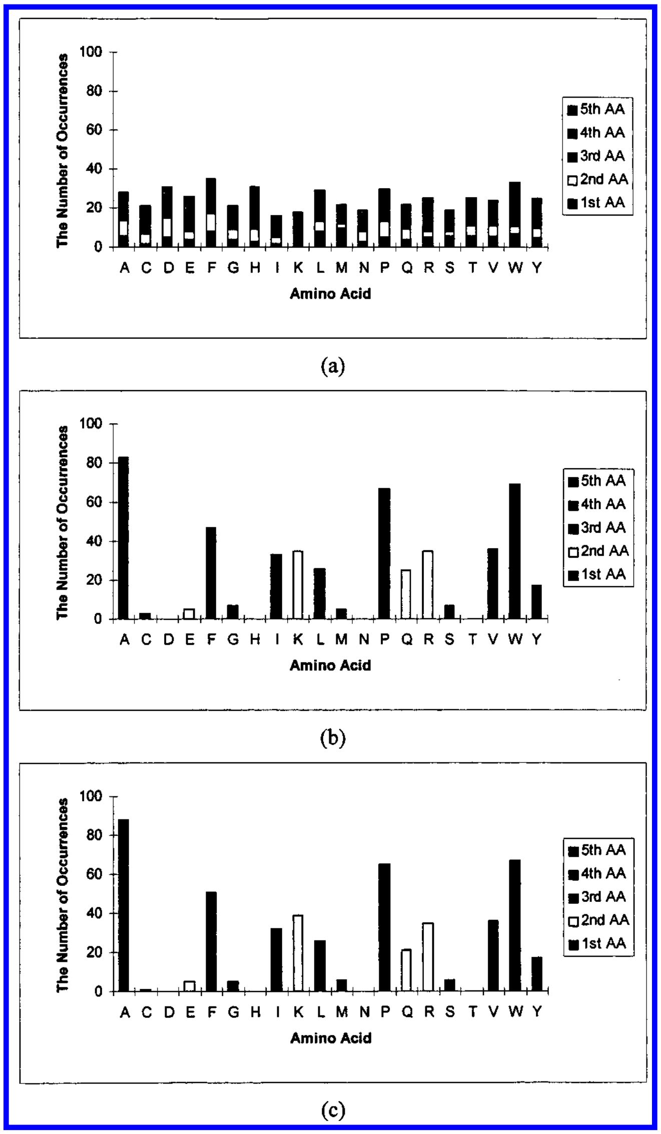


(a)

(b)

(c)

Figure 4. Focus-2D using ISA-ECI description method and VKWAP 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 5. Focus-2D using $Z_1-Z_2-Z_3$ description method and VKWAP as the similarity probe: (a) initial population; (b) final population after Focus-2D; and (c) the final population after the exhaustive search.

both description methods, the final populations after Focus-2D and exhaustive search were almost identical.

Focus-2D using QSAR Equation (Inverse QSAR). The results obtained with Focus-2D using a QSAR based prediction as the evaluation method are shown in Figures 6 and 7 for ISA-ECI and $Z_1-Z_2-Z_3$ descriptors, respectively. Again the populations before (Figures 6a and 7a) and after (Figures 6b and 7b) Focus-2D as well as the population after the exhaustive search (Figures 6c and 7c) are shown. The populations after Focus-2D and the exhaustive search were once again very similar to each other. With ISA-ECI descriptors, Focus-2D analysis selected F, I, L, M, P, R, V, and W to be present in active peptides, and the first, third, and fifth positions were identified as their most likely positions. 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.

Focus-2D using Topological Indices. *Using Similarity Probes VEWAK and VKWAP.* Figures 8 and 9 show the amino acid compositions of the populations before and after Focus-2D using topological indices as descriptors. VEWAK and VKWAP again were used as the similarity probes to evolve the population. Because the topological indices should be calculated for every virtual peptide, the exhaustive search was not performed because of computational limitations. In general, amino acids present in the similarity probe are correctly identified (Figures 8b and 9b). However, the suggested positions for each amino acid are less accurate than those obtained with amino acid-based descriptors (the second and fourth positions are suggested for A when VEWAK is used as the similarity probe; the second, third, and fourth positions are suggested for W when VKWAP is used as the similarity probe; the second, fourth, and fifth positions are suggested for P when VKWAP is used as the similarity probe).

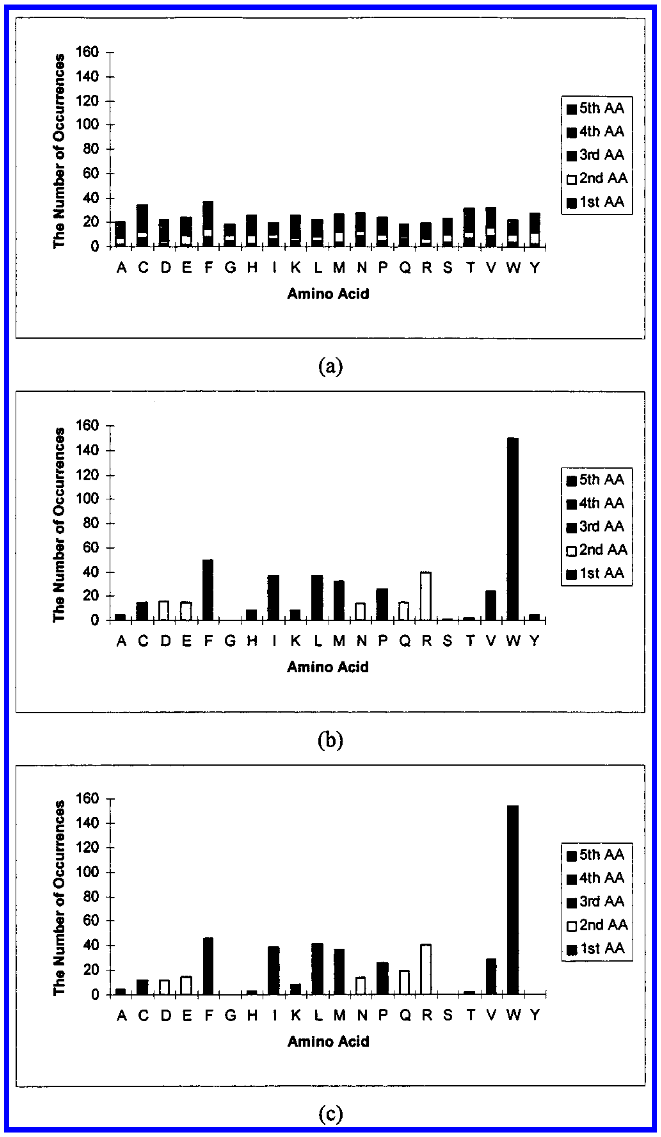


Figure 6. Focus-2D using ISA-ECI description method and the QSAR equation: (a) initial population; (b) final population after Focus-2D; and (c) final population after the exhaustive search.

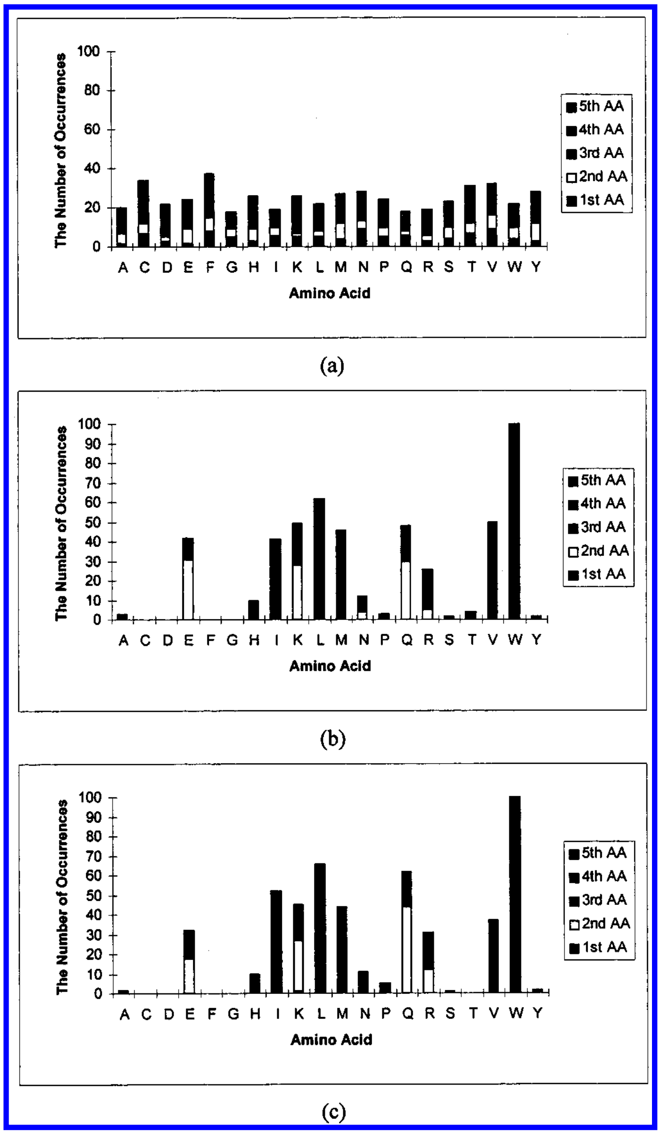


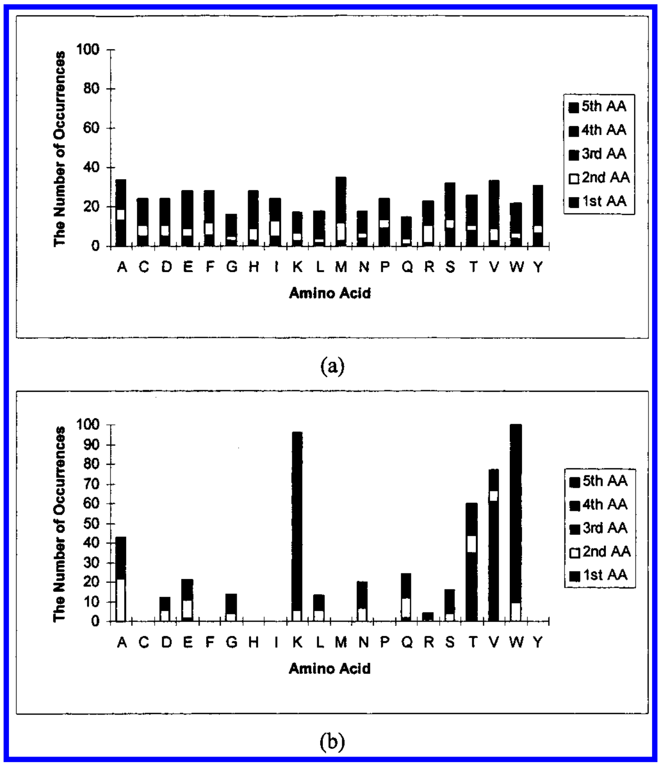
Figure 7. Focus-2D using $Z_1-Z_2-Z_3$ description method and a QSAR equation: (a) initial population; (b) final population after Focus-2D; and (c) final population after the exhaustive search.

Using the QSAR Equation. The results of Focus-2D using topological indices and the preconstructed QSAR equation ($q^2 = 0.845$; ONC = 5), obtained from the GA-PLS method, are shown in Figure 10. The final 100 peptides consisted mainly of F, H, I, R, and W. To examine the effect of the "degree of fit" condition in selecting these amino acids, we introduced a degree of fitting factor (see the *Computational Detail* section for the definition), which was used to control the level of extrapolation. Figure 11a–c shows the results obtained with the degree of fit factors of 0.7, 0.5, and 0.3, respectively (the degree of fit factor of 1 was used for Figures 6, 7, and 10). Interestingly, lowering the degree of fit factor increased the occurrence of amino acid commonly found in the active pentapeptides (i.e., V, E, and K), although multiple preferred positions were suggested for the selected amino acids (Figure 10b and Figure 11a–c).

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 of lead compounds. In this and our accompanying paper,[5] we introduced a novel computational tool, Focus-2D, which was developed to enhance the rational design of targeted chemical libraries. This method utilizes the existing SAR information to identify virtual library compounds with biological activities, and the building blocks frequently found in these virtual libraries are proposed for use in targeted library synthesis. The current implementation of the program includes two different description methods (building block based and whole molecule based) and two different evaluation (similarity probe and QSAR prediction) protocols that are used along with either GA (this paper) or SA (accompanying paper[5]) optimization methods. The key aspects of the algorithm are described in Figure 1.

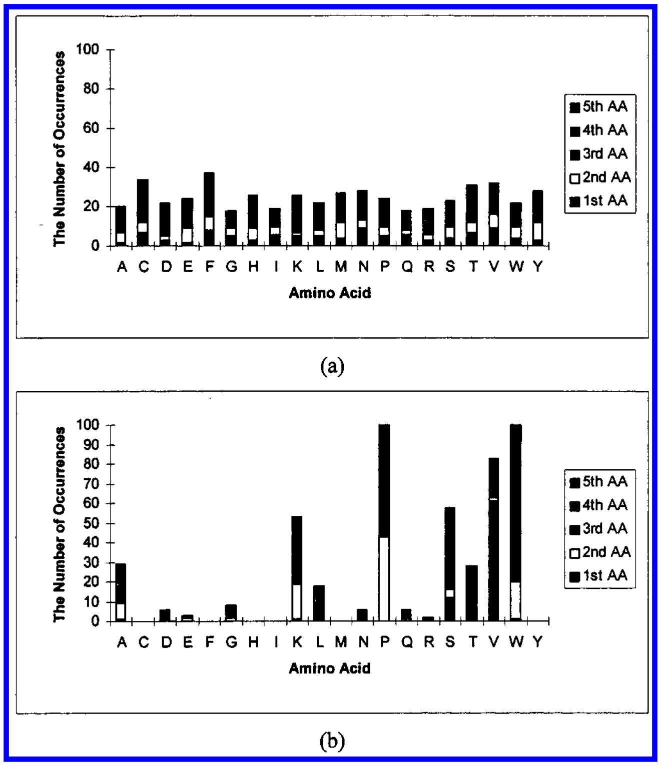
To test this methodology, we selected 30 BK-potentiating pentapeptides as a training set to design a targeted library with BK activity. Selection of a peptide data set was based on the fact that there are almost no published nonpeptide combinatorial chemical libraries that contain SAR information.[24] 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



(a)

(b)

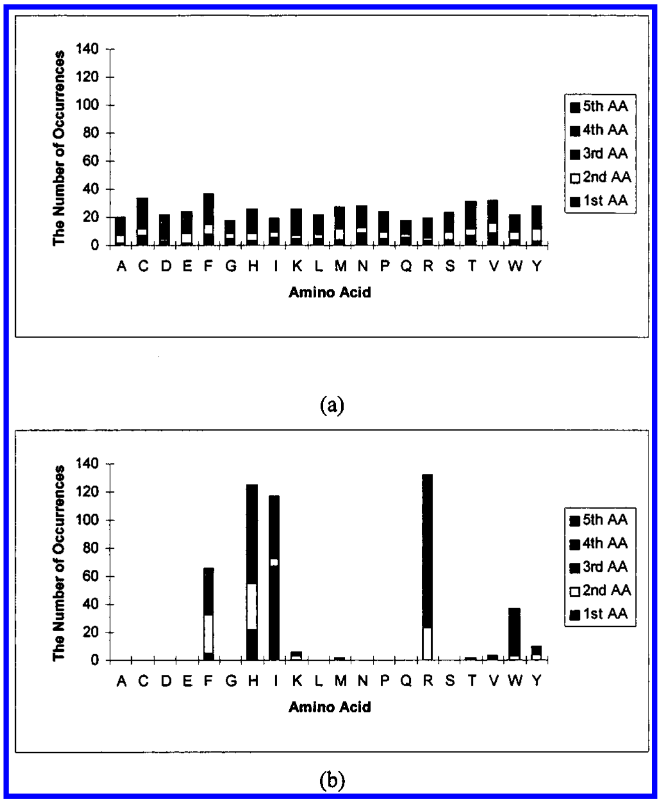
Figure 8. Focus-2D using topological indices as the description method and VEWAK as the similarity probe: (a) initial population and (b) final population after Focus-2D.



(a)

(b)

Figure 9. Focus-2D using topological indices as the description method and VKWAP as the similarity probe: (a) initial population; and (b) final population after Focus-2D.



(a)

(b)

Figure 10. . Focus-2D using topological indices as the description method and a QSAR equation as a fitness function: (a) initial population; and (b) final population after Focus-2D.

there are only 20 naturally occurring amino acids (building blocks).

As one of the ways to guide the GA-based selection process, the similarity of a virtual peptide to one of two active peptides, VEWAK or VKWAP, was measured by its Euclidean distance to the probes. 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 2−5, 8, and 9). However, 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 ISA-ECI and $Z_1-Z_2-Z_3$ descriptors than for the topological descriptors. This result was somewhat expected, because topological indices describe a peptide as a whole and the identity of the amino acid in each position is described implicitly, whereas the amino acid dependent descriptors (i.e., ISA-ECI and $Z_1-Z_2-Z_3$ descriptors) encode the identity explicitly.

The alternative fitting function employed in this paper was based on the use of preconstructed QSAR equations to predict log RAI of virtual peptides (Table 3). The interpretation of the results in terms of the effectiveness of this method to identify amino acids found in the active compounds is more difficult than in the previous case. The reason for this is that selected peptides are active according to the (extrapolated) QSAR equation. Thus, the amino acids found in the final selected peptides are not necessary the same as found in the active peptides (Table 1). The fact that we were trying to extrapolate the existing relationships for 28 peptides to search the structural space defined by 3.2 million peptides further complicated the matters. To address this point, we introduced the modified "degree of fit" condition (see the *Computational Detail* section). This condition made the program intelligent enough not to predict peptides that are structurally too different from peptides in the training set.

As already discussed, we have considered both amino acid- and the whole molecule-based descriptors. One major

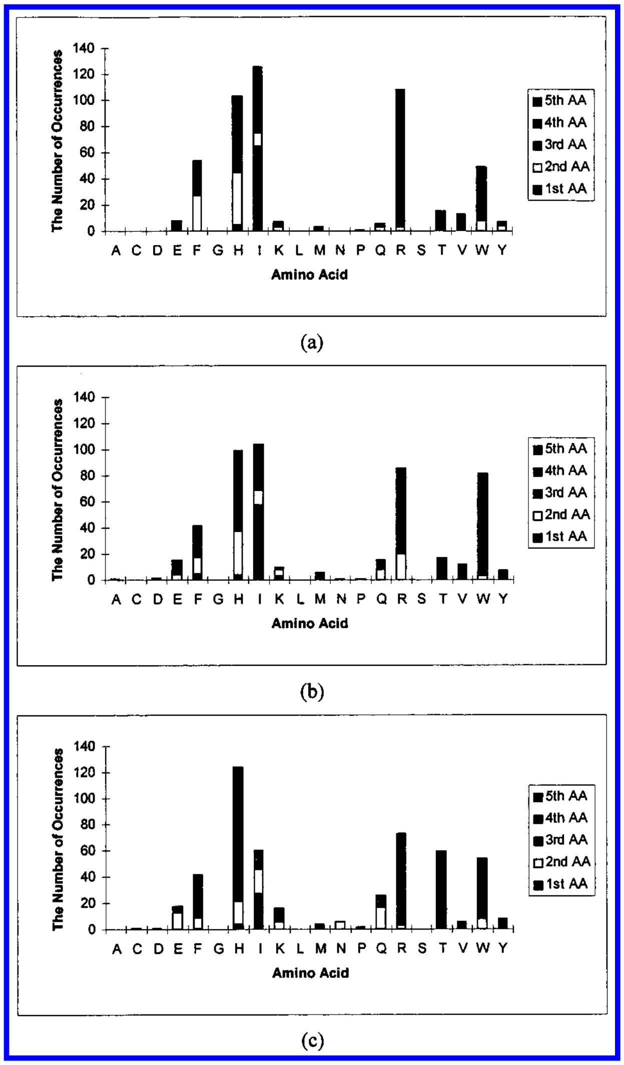


Figure 11. . Focus-2D using topological indices as the description method and a QSAR equation as a fitness function: (a) the degree of fitting factor = 0.7; (b) the degree of fitting factor = 0.5; and (c) the degree of fitting factor = 0.3.

advantage of topological indices, as well as any whole molecule-based descriptors, over amino acid-based descriptors is that topological indices can also describe nonpeptides. This is an important point because peptides similar to a nonpeptide probe or, alternatively, nonpeptides similar to a peptide probe can be identified as illustrated in the accompanying paper.[5] Furthermore, a large number of QSAR studies available in literature can be used to direct combinatorial chemical library design.

An obviously positive result of this work is that it proved the effectiveness of the GA optimization method. In all cases, with either two types of amino acid-based descriptors or the inverse QSAR prediction method, the results of stochastic search were comparable with those obtained after an exhaustive search (cf. Figures 2−7); 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.

Finally, any rational molecular design process should be iterative; that is, closely coupled with experimental design. For instance, the predictive power of a QSAR study in most cases is evaluated by predicting the activities of compounds in the training set and comparing them with their actual activities. After test compounds are synthesized and their activity evaluated, both training and test compounds are combined and used to refine the previous QSAR equation. Similarly, we incorporated this process as a part of Focus-2D, represented by dotted arrows in Figure 1. Together with the modified "degree of fit" condition, the select and analyze steps (shown in Figure 1) represent essential steps in validating our Focus-2D method.

To the best of our knowledge, no experimental targeted library with BK-potentiating activity has been described in the literature. Thus, the present study provides practical suggestions for the rational design of such a library. Our predictions, summarized in Figures 2−7, can be validated by the practical design and evaluation of the BK libraries. This experimental evaluation will also help us to determine the most adequate descriptors among three different types used in this work.

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 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) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

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

(5) Zheng, W.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.

(6) MOLCONN-X version 2.0, Hall Associates Consulting; Quincy, MA.

(7) 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; CINF-068.

(8) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(9) Ufkes, J. G. R.; Visser, B. J.; Heuver, G.; Van Der Meer, C. Structure-Activity Relationships of Bradykinin Potentiating Peptides. *Eur. J. Pharm.* **1978**, *50*, 119−122.

(10) 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.

(11) Cho, S. J.; Cummins, D.; Bentley, J.; Andrews, C. W.; Tropsha, A. An Alternative to 3D QSAR: Application of Genetic Algorithms and Partial Least Squares to Variable Selection of Topological Indices, submitted for publication in *J. Comp. Aided Mol. Design.*

(12) Dunn, W. J., III; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. Multivariate Structure-Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−137.

(13) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(14) Collantes, E. R.; Dunn, W. J., III Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues. *J. Med. Chem.* **1995**, *38*, 2705−2713.

(15) Tetko, I. V.; Luik, A. I.; Poda, G. I. Application of Neural Networks in Structure-Activity Relationships of a Small Number of Molecules. *J. Med. Chem.* **1993**, *36*, 811−814.

(16) Ajay. A Unified Framework for Using Neural Networks To Build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.

(17) So, S. S.; Richards, W. G. Application of Neural Networks: Quantitative Structure-Activity Relationships of the Derivatives of 2,4-Diamino-5-(substituted-benzyl)pyrimidines as DHFR Inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.

(18) Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. Generalized Simulated Annealing for Function Optimization. *Technometrics* **1986**, *28*, 209−217.

(19) Kalivas, J. H.; Sutter, J. M.; Roberts, N. Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet-Visible Spectrophotometry. *Anal. Chem.* **1989**, *61*, 2024−2030.

(20) Goldberg, D. E. *Genetic Algorithm in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(21) Holland, J. H. Genetic Algorithms. *Sci. Am.* **1992**, *267*, 66−72.

(22) Forrest, S. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, *261*, 872−878.

(23) Lindberg, W.; Persson, J.-A.; Wold, S. Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate. *Anal. Chem.* **1983**, *55*, 643−648.

(24) 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.

CI9700945