Deconvolution of Combinatorial Libraries for Drug Discovery: A Model System

Susan M. Freier,*,† Danielle A. M. Konings,‡ Jacqueline R. Wyatt,† and David J. Ecker†

ISIS Pharmaceuticals, 2292 Faraday Avenue, Carlsbad, California 92008, and Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309

Received August 19, 1994®

Iterative synthesis and screening strategies have recently been used to identify unique active molecules from complex synthetic combinatorial libraries. These techniques have many advantages over traditional screening methods, including the potential to screen large numbers of compounds to identify an active molecule while avoiding analytical separations and structural determination of unknown compounds. It is not clear, however, whether these techniques identify the most active molecular species in the mixtures and, if so, how often. Two key factors which may affect success of the selection process are the presence of many active compounds in the library with a range of activities and the chosen order of unrandomization. The importance of these factors has not been previously studied. Moreover, the impact of experimental errors in determination of subset activities or in randomization during library synthesis is not known. We describe here a model system based on oligonucleotide hybridization that addresses these questions using computer simulations. The results suggested that, within achievable experimental and library synthesis error, iterative deconvolution methods generally find either the best molecule or one with activity very close to the best. The presence of many active compounds in a library influenced the profile of subset activities, but did not preclude selection of a molecule with near optimal activity.

Introduction

Chemically synthesized combinatorial libraries provide a new source of compounds for drug discovery.[1,2] The development of automated solid phase synthesis techniques has enabled preparation of chemical libraries with extraordinary diversity and unprecedented numbers of novel compounds. The composition of the libraries can be controlled to have features which make the molecules within the library attractive lead compounds or drug candidates.

When a compound library is determined to have activity in a biological system, a "deconvolution" method must be used to determine which molecule(s) in the library is (are) responsible for the activity. An iterative deconvolution strategy which we refer to as SURF (synthetic unrandomization of randomized fragments)[3,4] can be used to identify a single compound from a mixture (Table 1). Iterative deconvolution strategies have been used to identify peptides which bind tightly to antibodies[5-10] or other protein targets[9,11,12] and oligonucleotides with antiviral activity.[3,4]

One issue is whether the iterative deconvolution method finds the most active molecule in the mixture. For example, among the molecules in complex combinatorial libraries, there may be many that have moderate activity and some with activity nearly equal to that of the most active compound. These molecules with suboptimal activity could make selection of the most active compound difficult. To evaluate the importance of molecules with suboptimal activity on the outcome of SURF deconvolution, we performed calculations using a RNA hybridization model. To our knowledge, oligonucleotide hybridization is the only molecular binding interaction where calculations based on experimentally determined thermodynamic parameters[13,14] can accurately predict association constants of very large numbers of molecules. Thus, it was possible to design a library with several hundred thousand different molecules and calculate the binding affinity for each molecule to the target molecule. The molecule with highest affinity (best binder) was determined and computer simulations of SURF deconvolution experiments were used to ascertain whether molecules that bind with less affinity (suboptimal binders) affect our ability to select the best binder.

We used this model system to ask whether or not SURF deconvolution identified the tightest binder when factors including the order of unrandomization, experi-

Table 1. Deconvolution Strategies for Chemical Libraries of all Tetramers Composed of the Five Monomers A-E

SURF[a]			position scanning[b]		
sequences in subset	molecules per subset	most active subset	sequences in subset	molecules per subset	most active subset
XNNN	125	**A**NNN	**X**NNN	125	**A**NNN
AXNN	25	**AC**NN	N**X**NN	125	N**C**NN
ACXN	5	**ACE**N	NN**X**N	125	NN**E**N
ACEX	1	**ACED**	NNN**X**	125	NNN**D**
selected sequence:		**ACED**	selected sequence:		**ACED**

[a] SURF deconvolution begins with synthesis of a nonoverlapping set of mixtures by incorporating a unique monomer at a common position of each subset. The subsets are tested separately and the one with greatest activity is identified. A second set of compound mixtures is prepared with each subset containing the fixed monomer showing greatest activity from the previous round. In addition, another position is fixed with each of the unique monomers to give another set of subsets. The complexity of the mixture is reduced and the process is repeated until a unique molecule is identified. [b] Position scanning is a noniterative technique which has been used with peptide libraries.[9,17] At each position in the oligomer sequence, a series of mixtures is synthesized with a different monomer in the fixed position. Each of the mixtures is tested separately and the selected molecule is deduced by selecting the monomer from the most active mixture from each position set. In principal, only a single round of screening is required to define the most active molecule.

† ISIS Pharmaceuticals.
‡ University of Colorado.
® Abstract published in *Advance ACS Abstracts,* December 15, 1994.

0022-2623/95/1838-0344$09.00/0

Table 2. Library Molecules with Highest Affinities[a]

9-mer target 5′-GUGUGGGCA-3′		18-mer target 5′-AUGUGUGGGCAACCUAGU-3′		6-mer target 5′-UGGGCA-3′	
sequence	$\Delta G°_{37}$	sequence	$\Delta G°_{37}$	sequence	$\Delta G°_{37}$
GCCCACACA	−17.0	GCCCACACA	−17.4	UAUGCCCAG	−10.5
GCCCACACG	−17.0	GCCCACACG	−17.0	UAUGCCCAA	−10.5
GCCCGCACA	−16.5	GCCCACGCA	−16.9	GAUGCCCAG	−10.5
GCCCACACU	−16.5	GCCCGCACA	−16.9	GAUGCCCAA	−10.5
GCCCGCACG	−16.5	GGUUGCCCA	−16.7	CAUGCCCAG	−10.5
GCCCACGCG	−16.5	GCCCACACU	−16.5	CAUGCCCAA	−10.5
GCCCACGCA	−16.5	GCCCACGCG	−16.5	AUGCCCAGU	−10.5
GCCCACACC	−16.1	GCCCGCACG	−16.5	AUGCCCAGG	−10.5
GCCCGCGCA	−16.0	GCCCGCGCA	−16.4	AUGCCCAGC	−10.5
GCCCGCACU	−16.0	GGUUGCCCG	−16.4	AUGCCCAGA	−10.5
GCCCACGCU	−16.0	GGGUUGCCC	−16.2	AUGCCCAAU	−10.5
GCCCGCGCG	−16.0	UGCCCACAC	−16.2	AUGCCCAAG	−10.5
GCCCGCACC	−15.6	GCCCACACC	−16.1	AUGCCCAAC	−10.5
CGCCCACAC	−15.6	GCCCACGCU	−16.0	AUGCCCAAA	−10.5
GCCCACGCC	−15.6	GCCCGCACU	−16.0	AAUGCCCAG	−10.5
UGCCCACAC	−15.5	GCCCGCGCG	−16.0	AAUGCCCAA	−10.5
GCCCGCGCU	−15.5	GGUUGCCCU	−15.9	UUUGCCCAG	−10.4
GGCCCACAC	−15.5	AGGUUGCCC	−15.9	CGCCCAAUA	−10.4
AGCCCACAC	−15.5	UGCCCACGC	−15.7	GCGCCCAGA	−10.4
GCCCGCGCC	−15.1	UGCCCGCAC	−15.7	AACGCCCAA	−10.4
+4 more at	−15.1	+0 more at	−15.7	+124 more at	−10.4

[a] The library contained all 262 144 possible RNA 9-mers. Hybridization affinities (kcal/mol) to each target were calculated as described in the Experimental Section.

mental error in measurement of association constants, and biased synthesis errors were included. Secondly, we asked how well position scanning (Table 1) compared to iterative SURF. We found that in the presence of suboptimal binders and reasonable experimental errors iterative deconvolution usually selected either the best binder or a molecule that bound nearly as tightly as the best binder. In the presence of experimental error, position scanning was significantly less successful than iterative SURF. These results suggest that SURF deconvolution may be a powerful method for identification of lead compounds even when suboptimal binders are present in the mixtures.

Results

Effect of Suboptimal Binders on Deconvolution Profiles. We used RNA oligonucleotide hybridization to model a library of molecules with a spectrum of affinities. This system was chosen because relatively simple rules are available to calculate association constants (K_A) for any pair of RNA oligomers with reasonable accuracy.[13-15] Three targets of differing length were investigated, 5′-GUGUGGGCA-3′, 5′-AUGUGUGGGCAACCUAGU-3′, and 5′-UGGGCA-3′. The library in each case consisted of all possible 262 144 RNA 9-mers. For all three targets, suboptimal binders included duplexes with mismatches or bulges. When the target length differed from the length of the library oligonucleotides, multiple duplex binding alignments were possible, so there was increased likelihood of several library molecules having similar affinities for the target.

K_As for each molecule were determined as described in the Experimental Section and subset K_As were calculated using eq 1 (see Experimental Section). For the 9-mer target, the tightest binding molecule from the library bound with $\Delta G°_{37} = -17.0$ kcal/mol (Table 2). Two sequences bound with this free energy and, interestingly, they bound 11-fold (1.5 kcal/mol) more tightly than the Watson–Crick complement of the target. This is a result of the stabilizing effect of 3′ dangling ends[16]

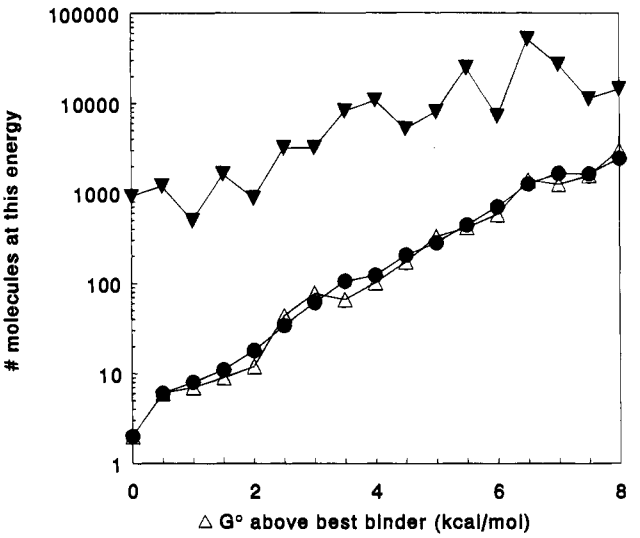

Figure 1. Number of molecules at each energy in a library of 262 144 RNA 9-mers that hybridize to GUGUGGGCA (△), AUGUGUGGGCAACCUAGU (●), or UGGGCA (▼). For each target, energies are plotted relative to the tightest binding 9-mer in the library. For plotting, energies were combined in intervals of 0.5 kcal/mol. Therefore, the number of molecules at 0 kcal/mol above the best includes not only the best binder(s) but also all molecules with $\Delta G°_{37} < 0.5$ kcal/mol above the best.

and was confirmed experimentally with thermal denaturation measurements. The midpoint of the melting transition of the duplex formed by the target and the "best binder" (GCCCACACG) was 7° higher than that of the duplex formed by the target and its Watson–Crick complement (UGCCCACAC). Fits of the melting curves suggested that $\Delta\Delta G°_{37}$ is 1.7 ± 0.2 kcal/mol.

For each target, the 20 library molecules that bound with highest affinity are listed in Table 2; energy distributions at higher energies are plotted in Figure 1. Energy distributions for the 9-mer and 18-mer targets were very similar; less than 15 molecules (0.006%) bound with free energy within 1 kcal of the best binder; less than 3% of the molecules bound with

Table 3. Deconvolution of a Library of 4^9 Oligonucleotide 9-mers Hybridizing to a 9-mer Target[a]

		K_A (M^{-1}) when X =			
round	sequence	A	C	G	U
library	NNNNNNNNN		2.9×10^7 [b]		
1	NNNNXNNNN	**6.5×10^7**	2.2×10^7	2.9×10^7	4.1×10^5
2	NNNNANXNN	**1.8×10^8**	5.2×10^5	8.0×10^7	8.6×10^5
3	NNXNANANN	1.1×10^5	**7.0×10^8**	5.6×10^4	5.6×10^6
4	NNCXANANN	2.3×10^5	**2.8×10^9**	2.7×10^5	4.8×10^7
5	NNCCAXANN	1.6×10^6	**1.1×10^{10}**	1.7×10^6	1.9×10^8
6	NNCCACAXN	6.4×10^8	**4.0×10^{10}**	6.4×10^8	1.9×10^9
7	NXCCACACN	2.0×10^7	**1.6×10^{11}**	1.5×10^7	1.1×10^9
8	NCCCACACX	**2.4×10^{11}**	5.5×10^{10}	**2.4×10^{11}**	1.1×10^{11}
9	XCCCACACA	1.4×10^9	6.4×10^8	**9.5×10^{11}**	5.5×10^8
	selected molecule: GCCCACACA				
	$\Delta G°_{37}$ (kcal/mol)	−17.0			

[a] Association constants for each subset hybridizing to GUGUGGGCA were calculated as described in the text. For each round, K_A of the tightest binding subset is shown in bold. [b] Although K_A for the entire library is usually not measured, it can be calculated from the experimental K_A values for the round 1 subsets: $K_{A,library} = \sum_i K_{A,round\ 1,subset\ i}/4$ where the sum is over the four round 1 subset K_As.

free energy within 8 kcal of the best binder. In contrast, for the 6-mer target, 2414 molecules (0.9%) bound with free energy within 1 kcal of the best binder and 63% of the molecules bound with free energy within 8 kcal of the best binder. For the 6-mer target, the large number of suboptimal binders with affinity near that of the best binder is due to the possibility of multiple alignments and the fact that nonpaired ends in the library oligonucleotide did not significantly affect binding free energy.

Table 3 lists calculated subset K_As for SURF deconvolution of the 9-mer library with the length-matched target (GUGUGGGCA). The selected molecule, GCCCACACA, had a K_A of 9.5×10^{11} M^{-1} (Table 3) and was one of the two tightest binding compounds in the library. If this selected molecule was the only molecule in the library with affinity for the target, K_A of the winning subset would increase by a factor equal to the decrease in subset complexity at each deconvolution step. In our examples with libraries built from four monomers, there were four subsets in each round, so in the absence of suboptimal binders, K_A of the winning subset would increase 4-fold each round and losing subsets would not bind. This hypothetical situation is depicted by the solid symbols in Figure 2a. When suboptimal binders were included in the simulation (open symbols in Figure 2a), calculated affinities increased. The suboptimal binding factor (SBF, see eq 2 of Experimental Section) is a measure of how much more tightly a subset bound than would be expected if the best binder were the only molecule in the subset with affinity for the target. In the example in Table 3 and Figure 2, the SBF of the library was 8.0. Due to the presence of suboptimal binders, the library bound 8.0-fold more tightly than would be expected if the selected molecule was the only molecule with affinity for the target. Although there were thousands of suboptimal binders in the library, a very small fraction of the 262 144 molecules accounted for most of the library activity. The two best binders contributed 25% to the library K_A; the remaining 18 molecules listed in column 1 of Table 2 contributed an additional 50%. The 100 tightest binders (0.04% of the molecules) contributed 90% of the library affinity.

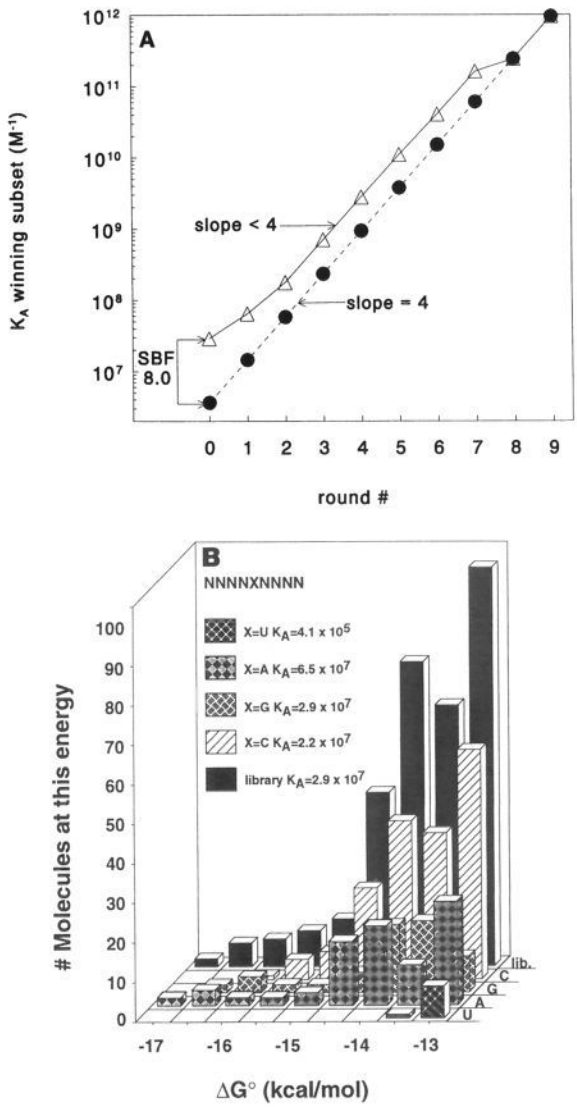

Figure 2. Deconvolution of a library of 4^9 oligonucleotide 9-mers hybridizing to a 9-mer target. Residues were unrandomized in the order 5, 7, 3, 4, 6, 8, 2, 9, 1. (A) Calculated association constant for the tightest binding subset in each round: (△) suboptimal binders were included in calculation of the association constants; (●) association constants were calculated assuming only one molecule in the library, GCCCACACA, binds with $K_A = 9.5 \times 10^{11}$ M^{-1}. (B) The number of molecules at each energy *vs* energy for the entire library (solid bars) and the four subsets (hatched bars) in round 1.

Figure 2b demonstrates division of the 4^9 molecules in the library into four subsets during round 1 of the deconvolution. Examination of the sequences in Table 2 and Figure 2b reveals that the most active compounds were distributed into three subsets (A, C, G) in round 1. No library sequences with U in position 5 bound with energy tighter than −13.1 kcal/mol. As a consequence of distributing the most active compounds into three subsets, each of these three subsets showed substantial activity. The K_A of the winning subset improved only 2.2-fold from the library K_A, instead of 4-fold as would be expected if the best binder were the only active molecule in the library. In subsequent rounds, similar separation of suboptimal binders into losing subsets resulted in observable activity of losing subsets and less than 4-fold improvement in K_A between rounds. In each

Table 4. Deconvolution of a Library of 4^9 Oligonucleotide 9-mers Hybridizing to a 6-mer Target[a]

		K_A (M^{-1}) when X =			
round	sequence	A	C	G	U
library	NNNNNNNNN		1.3×10^5 [b]		
1	NNNNXNNNN	4.7×10^4	**3.6×10^5**	8.6×10^4	1.3×10^4
2	NNNNCNXNN	4.1×10^5	**5.8×10^5**	3.0×10^5	1.5×10^5
3	NNXNCNCNN	4.4×10^5	**8.7×10^5**	5.1×10^5	5.2×10^5
4	NNCXCNCNN	1.1×10^4	1.4×10^6	**2.0×10^6**	2.1×10^4
5	NNCGCXCNN	3.6×10^3	**8.0×10^6**	3.0×10^4	6.2×10^4
6	NNCGCCCXN	**1.8×10^7**	1.1×10^6	1.1×10^7	2.2×10^6
7	NXCGCCCAN	**1.8×10^7**	**1.8×10^7**	**1.8×10^7**	**1.8×10^7**
8	NACGCCCAX	**2.1×10^7**	1.3×10^7	**2.1×10^7**	1.5×10^7
9	XACGCCCAA	**2.1×10^7**	**2.1×10^7**	**2.1×10^7**	**2.1×10^7**
	selected molecule:	AACGCCCAA			
	$\Delta G°_{37}$ (kcal/mol)	−10.4			

[a] Association constants for each subset hybridizing to UGGGCA were calculated as described in the text. For each round, K_A of the tightest binding subset is shown in bold. [b] Although K_A for the entire library is usually not measured, it can be calculated from the experimental K_A values for the round 1 subsets, $K_{A,library} = \sum i\ K_{A,round\ 1,subset\ i}/4$ where the sum is over the four round 1 subset K_As.

round, the subset containing the best binder bound with highest affinity.

Many different orders were tested for unrandomization of the 9-mer library with the 9-mer target. Details of the deconvolution profile depended slightly on the order of unrandomization. For example, K_A of the winning subset in round 1 varied 3-fold depending on the order of unrandomization. The unrandomization order did not, however, affect which molecule was selected. Using 500 different orders of unrandomization, the tightest binding molecule was always selected.

For the previously described library and the 18-mer target (AUGUGUGGGCAACCUAGU), the tightest binding molecule was always selected despite the order of unrandomization. Deconvolution profiles for this target (data not shown) were very similar to those for the 9-mer target. This was likely due to similarities between the energy distributions for these two targets (Figure 1).

A typical deconvolution profile for this 9-mer library hybridizing to the 6-mer target is shown in Table 4. The energy profile for this target (Figure 1) differed from that of the longer targets and results of deconvolution also differed. The average increase in K_A of the best subset between successive rounds was only 2.0 (compared to 3.3 for the 9-mer target). In rounds 9 and 7, when positions 1 and 2 were fixed, no base was preferred in these dangling positions and K_A did not improve at all. The SBF was much larger for the 6-mer target than for the 9-mer target. Suboptimal binders in the library caused K_A to be 1561 times tighter than if only the selected winner (AACGCCCAA) bound. Perhaps the most significant feature of this deconvolution profile was that the selected molecule bound with a free energy of −10.4 kcal/mol. There were 16 molecules in the library that bound slightly more tightly (by 0.1 kcal/mol) than this (Table 2). The origin of this result is revealed by examination of the library sequences in Table 4 and the 16 sequences that bind with $\Delta G°_{37} = -10.5$ kcal/mol (Table 2). In round 2, eight of these 16 molecules were in the A subset and eight in the selected C subset. In round 3, all eight were in the U subset, but that subset was not selected because suboptimal binders caused K_A of the C subset to be greater than that of the U subset. No best binders were in the C subset, so selection of the C subset in round three necessitated selection of a suboptimal binder.

The effect of the unrandomization order was examined for this target. In the case of the 6-mer target, the energy of the selected molecule depended on the unrandomization order. For 500 different orders of randomization, energies of the selected molecule ranged from −10.5 to −9.6 kcal/mol; the average selected energy was −10.34 kcal/mol.

Effect of Suboptimal Binders on Position Scanning. Position scanning[9,17,18] (see Table 1) is a noniterative deconvolution technique where a set of mixtures is synthesized for each position of the oligomer and a single position is fixed in each subset. The sequence of the most active compound is deduced by selecting the monomer from the most active subset at each position. Results of position scanning simulations for this library of 4^9 9-mers hybridizing to each of the three targets, UGGGCA, GUGUGGGCA, or AUGUGUGGGCAACCUAGU, are listed in lines 4−6 of Table 5. With the 9-mer and 18-mer targets, position scanning, like iterative SURF, selected the best binder. In the case of the 6-mer target, however, a molecule which bound 26 times weaker than the best binder was selected. The failure of position scanning to select the best binder with the 6-mer target was a result of multiple binding alignments with similar energies. Position scanning was unable to select a single alignment.

Effect of Assay Experimental Error on Distribution of Selected Molecules. A significant effect of suboptimal binders in the library was that all four subsets bound the target with some affinity. Sometimes the difference in affinities between subsets was quite small. For example, in round 1 with the 9-mer target (Table 3), the G and C subsets bound, respectively, only 2.3 or 2.9 times more weakly than the A subset; the U subset bound 158-fold more weakly. With experimental error in measurement of subset activities, the wrong subset can be selected. If the G or C subset was selected, then the final selected molecule was, respectively, GCCCGCACA (−16.5 kcal/mol) or CGCCCACAC (−15.6 kcal/mol) rather than the tightest binder, GCCCACACA (−17.0 kcal/mol). Thus, if a mistake was made, a suboptimal binder was selected.

To evaluate the global effect of experimental error on SURF deconvolution, Monte Carlo simulations[19] were performed with the assumption that the observed values of log K_A were distributed normally about the true value of log K_A with a standard deviation of log 2. This model reflected a situation where the experimental error was a factor of ±2. For each target, a single order of unrandomization was chosen and 500 simulations were performed with experimental error; energy distributions of selected molecules are plotted in Figure 3 and average selected energies are listed in Table 5. In the absence of experimental error, deconvolution against the 9-mer or 18-mer target always resulted in selection of the tightest binding molecule; however, introduction of experimental error sometimes resulted in selection of a suboptimal binder (Figure 3). Comparison of rows 7−9 to rows 1−3 in Table 5, reveals that experimental error increased the average selected energy slightly (0.24, 0.38, or 0.69 kcal/mol for the 6-mer, 9-mer, or 18-mer

Table 5. Summary of Simulations for Deconvolution of a Library of 4^9 Oligonucleotide 9-mers[a]

row number	target	average selected energy	average selected energy above best[b]	% selected molecules at best energy	% selected molecules within 1 kcal/mol of best	comments
			Iterative SURF			
1	6-mer	−10.34	0.16	21.8	100.0	500 orders of unrandomization tested
2	9-mer	−17.00	0.00	100.0	100.0	
3	18-mer	−17.40	0.00	100.0	100.0	
			Position Scanning			
4	6-mer	−8.50	2.00	0.0	0.0	
5	9-mer	−17.00	0.00	100.0	100.0	
6	18-mer	−17.40	0.00	100.0	100.0	
			Iterative SURF with Experimental Error[c]			
7	6-mer	−10.10	0.40	2.6	97.2	500 Monte Carlo simulations with unrandomization order 5, 7, 3, 4, 6, 8, 2, 9, 1
8	9-mer	−16.62	0.38	55.0	90.3	
9	18-mer	−16.71	0.69	19.8	82.8	
			Position Scanning with Experimental Error[c]			
10	6-mer	−7.74	2.76	0.0	18.5	500 Monte Carlo simulations
11	9-mer	−15.23	1.77	37.0	71.0	
12	18-mer	−13.00	4.40	11.7	31.4	
			Iterative SURF with Nonrandom Synthesis[d]			
13	9-mer, case a	−17.00	0.00	100.0	100.0	unrandomization order 5, 7, 3, 4, 6, 8, 2, 9, 1
14	9-mer, case b	−17.00	0.00	100.0	100.0	
15	9-mer, case c	−15.60	1.40	0.0	0.0	
			Iterative SURF with Experimental Error[c] *and Nonrandom Synthesis*[d]			
16	9-mer, case a	−16.60	0.40	57.2	86.4	250 Monte Carlo simulations with unrandomization order 5, 7, 3, 4, 6, 8, 2, 9, 1
17	9-mer, case b	−16.31	0.69	42.8	72.4	
18	9-mer, case c	−15.82	1.18	26.0	40.8	

[a] Free energies (in kcal/mol) for each selected molecule were calculated as described in the text. Targets were 5′-UGGGCA-3′ (6-mer), 5′-GUGUGGGCA-3′ (9-mer) or 5′-AUGUGUGGGCAACCUAGU-3′ (18-mer). [b] Energies of the best binders are −10.5, −17.0, and −17.4 kcal/mol for the 6-mer, 9-mer, and 18-mer targets, respectively. The energy of the best binder was subtracted from the average selected energy to obtain the average selected energy above the best. [c] Methods for Monte Carlo simulations of unrandomization experiments with experimental errors are described in the text. [d] For the nonrandom syntheses, nucleotide composition at the random positions was as follows: case a, 21.4% C, 26.2% A, 26.2% G, 26.2% U; case b, 18.1% C, 27.3% A, 27.3% G, 27.3% U; case c, 10% C, 30% A, 30% G, 30% U.

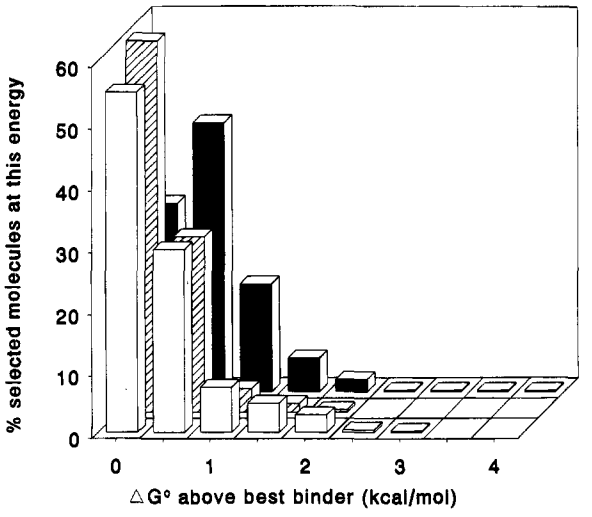

Figure 3. Percent of molecules selected at each energy during Monte Carlo simulations of deconvolution of a 9-mer RNA library hybridizing to GUGUGGGCA (open bars), UGGGCA (hatched bars), or AUGUGUGGGCAACCUAGU (solid bars). Fixed positions were unrandomized in the order 5, 7, 3, 4, 6, 8, 2, 9, 1 and the experimental error was a factor of ±2. For each target, energies are plotted relative to the tightest binding 9-mer in the library. For plotting, energies were combined in intervals of 0.5 kcal/mol. Therefore, the number of molecules at 0 kcal/mol above the best includes not only the best binder(s) but also all molecules with $\Delta G°_{37} < 0.5$ kcal/mol above the best.

targets, respectively) relative to the energy selected in the absence of error.

With the 6-mer target and experimental error, suboptimal molecules were selected >97% of the time. Only 3% of the selected molecules, however, bound less tightly than 1 kcal/mol above the free energy of the best binder (row 7 of Table 5). There was a high likelihood of experimental error leading to selection of a suboptimal binder, but because there were 2398 suboptimal binders within 1 kcal/mol of the best, the suboptimal binder selected was usually within 1 kcal/mol of the best. In contrast, with the 9-mer target and experimental error, a suboptimal molecule was selected only 45% of the time and 10% of the time the selected molecule bound less tightly than 1 kcal/mol above the best (row 8 of Table 5). Although there was less likelihood of selecting a suboptimal binder, there were only 10 suboptimal binders within 1 kcal/mol of the best, so selection of a molecule within 1 kcal/mol of the best was less likely with the 9-mer target than with the 6-mer target.

As discussed above, with the 6-mer target in the absence of experimental error, order of unrandomization affected the selected molecule, but the average selected energy was very close to the best. For example, in the absence of experimental error, unrandomization orders 5, 7, 3, 4, 6, 8, 2, 9, 1 and 2, 6, 4, 8, 1, 9, 7, 3, 5 selected molecules with affinities of −10.4 and −9.6 kcal/mol, respectively. When experimental error was added, these two orders of unrandomization resulted in average selected energies of −10.10 and −10.08, respectively. The fact that an improvement was observed in the presence of experimental error was due to the large number of suboptimal binders for this target.

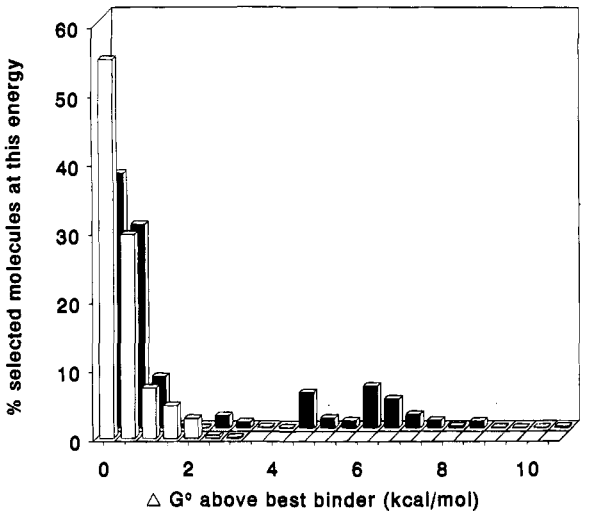

Figure 4. Percent of molecules selected at each energy during Monte Carlo simulations of deconvolution of a 9-mer RNA library hybridizing to GUGUGGGCA using iterative SURF with unrandomization order 5, 7, 3, 4, 6, 8, 2, 9, 1 (open bars) or position scanning (filled bars). Experimental error was a factor of ±2. Energies are plotted relative to the tightest binding 9-mer in the library. For plotting, energies were combined in intervals of 0.5 kcal/mol. Therefore, the number of molecules at 0 kcal/mol above the best includes not only the best binder(s) but also all molecules with $\Delta G°_{37} < 0.5$ kcal/mol above the best.

Effect of Assay Experimental Error on Position Scanning. Figure 4 compares Monte Carlo simulations of position scanning to those for iterative deconvolution. For the 9-mer target, experimental error had a more detrimental effect on position scanning than on iterative SURF. With iterative SURF, for the 9-mer target, all selected molecules bound with free energy within 3.0 kcal of the best binder. In contrast, with position scanning, more than 20% of the selected molecules bound with free energy more than 4.5 kcal/mol above that of the best binder. With the 6-mer and 18-mer targets, differences between iterative SURF and position scanning were even greater (compare rows 10–12 to rows 7–9 in Table 5). Position scanning was least likely to be successful when several different registers of binding were possible.

Effect of Nonrandom Synthesis on Deconvolution Profile. All of the calculations presented above assumed equal concentrations of each molecule in each subset. For oligonucleotide synthesis, however, various strategies for synthesis of random sequences can result in unequal incorporation of the four monomers at each random position. In extreme cases, a single nucleotide may occur as seldom as 10% or as often as 50% at a single "random" position.[20] To examine the effect of nonrandom synthesis on SURF deconvolution, simulations were performed with unequal representation of each nucleotide at each random position. Ratios of C to each of the other nucleotides were 0.9:1.1 (case a), 0.8:1.2 (case b), or 0.5:1.5 (case c). C was chosen to be under-represented because selection of this C-rich oligonucleotide (GCCCACACA) was most affected by under-representation of C. Only in case c was a suboptimal binder selected (lines 13–15 of Table 5). When experimental error was combined with nonrandom synthesis (lines 16–18 of Table 5), the distribution of selected molecules was affected only slightly, except in case c, where only 41% of the selected molecules bound with free energy within 1 kcal/mol of the best binder.

Discussion

Successful selection of the most active compound from a combinatorial library using the SURF deconvolution strategy requires that, in each round, the subset containing the most active compound is selected. This definitely happens if a single compound is active; in each round only one subset will be active and activity will improve as the concentration of active compound increases in subsequent rounds. In reality, however, several compounds in the library are likely to have some activity. If the most active suboptimal binders are together in one subset while the best binder is in another subset, the subset with the suboptimal binders may have the greatest activity and the most active single compound will not be selected at the end of deconvolution.

To evaluate whether suboptimcal binders affect the outcome of SURF deconvolution, calculations were performed using RNA hybridization as the model system. This system was selected because oligonucleotide hybridization is the only molecular binding interaction where simple calculations can predict association constants with reasonable accuracy.[13,21] Moreover, RNA hybridization is not simply a set of independent interactions between base pairs; the free energy of each base pair depends on its context.[22,23] Thus, RNA hybridization is a reasonable model for macromolecular interactions.

Three targets were evaluated and resulted in two types of energy profiles. With the 9-mer and 18-mer targets, there were one or two best binders and less than 15 oligonucleotides bound with free energy near that of the best binder. In contrast, 16 sequences bound to the 6-mer target at the lowest free energy and more than 2400 oligonucleotides (1% of the total library) bound with free energy near that of the best binder (Figure 1). The SBF's for our 9-mer library with the three different targets span the range of SBF's observed experimentally for oligonucleotide and peptide libraries (Table 6).

Affinity distributions for libraries of antigenic determinants binding to immunoglobulins or odorants binding to olfactory receptors have been modeled by Lancet *et al.*[24] Our energy distributions (Figure 1) were qualitatively similar to those for these receptor libraries in that the number of library molecules at each affinity increased rapidly as affinity decreased. Parameters reported by Lancet *et al.*[24] were used to calculate SBF values of 26 and 3 for the immunoglobulin and olfactory receptor library, respectively. These similarities between affinity profiles for the receptors and those for our RNA oligomers suggest that our model of RNA hybridization, with its suboptimal binders, provides useful examples for evaluation of SURF. Thus, our observations on the effects of suboptimal binders may be applicable to many real systems.

In Table 5, we compare the likelihood of selection of the best binder with selection of a "good binder", defined as a molecule that binds with free energy within 1 kcal/mol of the best energy. Association constants of these good binders are within a factor of 5 of that of the best binder. Under some circumstances, suboptimal binders

Table 6. Reported Activities for Deconvolution of Oligomer Libraries

target	composition	no. of monomers	oligomer length	no. of fixed positions	subset complexity	subset activity	winner activity	SBF
			Cellular Screens					
HIV[4]	P=S DNA	4	8	2	4 096	20 μM	0.3 μM	61
HSV[3]	P=S DNA	4	8	1	16 384	70 μM	0.4 μM	94
S. aureus[8]	peptide	18	6	2	104 976	450 μg/mL	3.4 μg/mL	793
S. aureus[9]	peptide	19	6	2	130 321	1730 μg/mL	11 μg/mL	829
			Antibody Binding					
pAB-FMRF[6]	peptide	15	4	0	50 625	1400 μg/mL	0.5 μg/mL	18
pAB-pep3[6]	peptide	16	6	0	16 777 216	6500 μg/mL	0.08 μg/mL	206
mAB-19B10[8]	peptide	18	6	2	104 976	250 μM	0.03 μM	13
mAB-125-10F3[9]	peptide	19	6	2	130 321	20 μM	0.004 μM	26
			Nucleic Acid Binding (Experimental)					
Ha-ras RNA[3]	2′-*O*-methyl RNA	4	9	1	65 536	10 μM	0.01 μM	66
			Protein Binding or Activity					
HIV-protease[11]	peptide	22	4	1	11 132	4400 μM	1.4 μM	3.5
Opioid receptor[9]	peptide	19	6	2	130 321	3.452 μM	0.028 μM	1057
Opioid receptor[29]	peptide	19	6	2	130 321	2.1 μM	0.005 μM	310
			Nucleic Acid Binding (This Work)					
18-mer target	RNA	4	9	0	262 144	0.020 μM	5.5×10^{-7} μM	7.1
9-mer target	RNA	4	9	0	262 144	0.034 μM	1.1×10^{-6} μM	8.0
6-mer target	RNA	4	9	0	262 144	7.9 μM	0.047 μM	1561

[a] The suboptimal binding factors (SBF) were calculated for the deconvolutions of oligomer libraries reported in the recent literature and for the theoretical deconvolutions described in this work. This table lists the assay used in the deconvolution experiment (target) and the description of each library, including the composition, the number of monomers, the length of the oligomer, and the number of fixed positions in the initial round of deconvolution. The number of unique molecules in each initial subset (complexity) is $N^{(L-F)}$ where N is the number of monomers, L is the oligomer length and F is the number of fixed positions. The activity of the most active subset in the initial round (subset activity), the activity of the final compound (winner activity), and the complexity were used to calculate the SBF of the most active initial round subset (see eq 2 of the Experimental Section).

significantly reduced the likelihood of selecting the best binder but had much less effect on the likelihood of selecting a good binder.

Suboptimal binders had very little effect on which molecule was selected by iterative SURF. For the 9-mer and 18-mer targets, the best binder was always selected. For the 6-mer target with its high number of suboptimal binders, some orders of unrandomization resulted in selection of a suboptimal binder. In all cases, however, the selected molecule bound within 1 kcal/mol of the free energy of the best binder. Suboptimal binders resulted in detectable binding of losing subsets, sometimes only 2-fold weaker than the winning subset. Thus suboptimal binders coupled with experimental error reduced the frequency with which the best binder was selected. Good binders, however, were still selected more than 80% of the time.

Position scanning was less successful. When experimental error was considered, position scanning selected a good binder as little as 18% of the time for the 6-mer target (compared to 97% with iterative SURF) and as often as 71% of the time for the 9-mer target (compared to 90% for iterative SURF). Position scanning was particularly unsuccessful when multiple alignments bound with similar energies, as was the case with the 6-mer target. With iterative SURF, selection of a single base in round 1 usually determined alignment of the final selected winner. Position scanning, on the other hand, often selected a different alignment at each position, resulting in a selected sequence with very poor binding.

We were particularly concerned about effects of synthesis errors which may result in unequal incorporation of monomers and nonrandom subsets. The three cases considered roughly reflect the base composition ratios we observed using our best procedures for random synthesis (case a), equimolar amidite mixtures (case b), or *in situ* mixing of amidites by an automated synthesizer (case c).[20] When nonrandom synthesis was included and experimental error considered, good binders were selected more than 70% of the time except with the most asymmetric synthesis. In our worst case, C occurred less than half as often as expected and a good binder was selected less than 50% of the time. Hence, a slight error in synthesis did not destroy the effectiveness of SURF, yet a very asymmetric synthesis had a severe effect.

Although suboptimal binders did not have a large effect on the ability of SURF deconvolution to select a good binder, they did affect the deconvolution profile. Suboptimal binders resulted in activity of more than one subset in each round, in less than 4-fold improvement in activity between rounds, and in values of SBF greater than 1. There were two practical consequences of a SBF value greater than 1. Firstly, activity of the best subset in round 1 was greater than would be expected if only a single molecule was active. This may be an advantage experimentally as it might make it easier to detect active subsets and more complex libraries could be studied. Secondly, between round 1 and the final round, binding did not improve as much as would be expected if only a single molecule was active. This has practical implications in the use SURF for drug discovery, where it must be decided whether it is worth the effort to deconvolute a library after an initial activity is identified. Knowledge of the SBF would allow extrapolation from activity in round 1 to activity of the final selected winner. If the projected activity is unacceptably low for use as a lead compound in drug development, the library can be abandoned and the effort of several rounds of synthesis and screening eliminated.

Unfortunately, the actual value of SBF cannot be known until deconvolution is completed. Table 6, however, provides experimental examples from which

the range of SBF can be estimated. In addition, the deconvolution profile during the early rounds can provide clues for estimation of SBF. As demonstrated above for the example in Table 3 and Figure 2, increased values of SBF were associated with less than 4-fold improvement in K_A between successive rounds and with substantial activity in losing subsets.

Finally, two important factors have been ignored in our modeling. We assumed no interaction between molecules within the library and we assumed that two molecules in the library cannot simultaneously bind to a single target molecule. A consequence of these factors is that the activity of one compound could be affected by other compounds in the library.[25] In principle, our model of nucleic acid hybridization could be extended to include such considerations although the calculations may be prohibitively large. Experiments can also be designed to evaluate the importance of these factors. Despite these potential problems, effective identification of active compounds from mixtures is well-known in the area of natural product screening[26] and is demonstrated by the examples in Table 6. In addition, qualitative similarities between experimental and calculated SBF values (Table 6) suggest that our model may reasonably mimic experimental deconvolution.

In summary, we have modeled SURF deconvolution using two types of energy profiles, one with many fewer suboptimal binders than the other. In both cases, suboptimal binders had only a limited effect on the energy of the molecule selected by iterative SURF and that effect was greatest when errors in synthesis or activity measurements were considered. RNA hybridization provided a useful model for library binding and allowed insight into iterative deconvolution methods of drug discovery.

Experimental Section

Calculation of Free Energies for Library Sequence To Target RNA. The free energy of the intermolecular complex between each possible library sequence and the specific target was calculated using the thermodynamic method MFOLD, described by Zuker and co-workers.[15,27,28] The set of free energy values used were those given by Jaeger *et al.*[15] The standard MFOLD program calculates the free energy for unimolecular folding of RNA. Therefore, library and target sequences were connected and modification of the program allowed us to obtain the free energy of the most stable bimolecular interaction from the intramolecular complex. For example, in the case of the 9-mer target (5′-GUGUGGGCA-3′) and the 9-mer library, 4^9 molecules of the sequence 5′-GUGUGGGCAXXXNNNNNNNNN-3′ were folded where NNNNNNNNN was each of the sequences in the library. The program interpreted XXX as "unknown" bases and did not allow pairing of these residues. This connection between target and library sequences allowed calculation of the most stable interaction between target and library sequences, forcing the XXX into a hairpin loop.

The following modifications were made to the standard MFOLD method.

(1) All hairpin loop contributions were set to +3.4 kcal to represent the helix initiation energy between the target and library sequences.[13] The only hairpin loop structure allowed was one including the connecting XXX sequence in its loop. This structure represents an intermolecular interaction. All other possible hairpin structures represent an intramolecular interaction. If an intramolecular interaction was observed in the optimal folding, the respective molecule was refolded while inhibiting the particular intramolecular interaction.

(2) The standard free energy contributions for the first mismatch in a hairpin loop were modified to allow for calculation of the correct dangling end contribution if one of the two nucleotides of the mismatch was X. If both nucleotides of the mismatch were not X, then they were treated as a terminal mismatch. The association constant for each intermolecular complex was calculated by the formula $K_A = \exp(-\Delta G°_{37})/RT$, were ΔG is the minimal free energy, R is the gas constant (0.001987 kcal/mol/K), and T is temperature (310.15 K).

Apparent association constants for each subset ($K_{A,subset}$) were calculated from the association constants of each molecule in the subset:

$$K_{A,subset} = \sum_i f_i K_{A,i} \quad (1)$$

where the sum is over all the molecules in the subset, f_i is the fraction of that subset that is molecule i, and $K_{A,i}$ is the association constant for molecule i binding to target RNA. If all molecules were equally represented, then $f_i = 1/N$ where N is the number of molecules in this subset. Simultaneous binding of more than one oligonucleotide to the target and interactions between oligonucleotides in the library were not allowed.

SBF (suboptimal binding factor) was defined as the apparent number of winning molecules in a subset:

$$\mathrm{SBF} = \frac{(\text{no. of different molecules in this subset})}{(K_A \text{ of final winner})/(K_A \text{ of this round})} \quad (2)$$

The numerator in eq 2 is equal to the improvement in K_A between this round and the final round, calculated under the assumption that no suboptimal binders exist in the library. The denominator is equal to the improvement in K_A obtained when suboptimal binders are included in the calculation. SBF is, therefore, a measure of how much tighter a subset binds than would be expected if the best binder was the only molecule in the subset with affinity for the target.

Monte Carlo Simulations of SURF. Effects of experimental error were simulated by assuming that observed values of log K_A were distributed normally. Observed K_A values for each subset were generated using standard Monte Carlo techniques. The subset with the largest observed K_A was selected as the "winning subset". Calculations were continued for successive rounds until a single molecule was selected.

Simulations of SURF with Biased Synthesis. SURF deconvolution of nonrandom libraries was performed as described above except that f_i in eq 1 was calculated for each molecule in the library by assuming fixed but nonequal probabilities of each nucleotide at all "random" positions.

Acknowledgment. The authors thank Dr. M. Zuker for assistance in modifying the MFOLD program to calculate bimolecular free energies and Dr. P. Davis for providing base composition data for random oligonucleotide libraries.

References

(1) Gallop, M. A.; Barrett, 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.; Barrett, 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) Ecker, D. J.; Vickers, T. A.; Hanecak, R.; Driver, V.; Anderson, K. Rational screening of oligonucleotide combinatorial libraries for drug discovery. *Nucleic Acids Res.* **1993**, *21*, 1853–1856.

(4) Wyatt, J. R.; Vickers, T. A.; Roberson, J. L.; Buckheit, R. W., Jr.; Klimkait, T.; DeBaets, E.; Davis, P. W.; Rayner, B.; Imbach, J. L.; Ecker, D. J. Combinatorially selected guanosine-quartet structure is a potent inhibitor of human immunodeficiency virus envelope-mediated cell fusion. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 1356–1360.

(5) Geysen, H. M.; Rodda, S. J.; Mason, T. J. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Mol. Immunol.* **1986**, *23*, 709–715.

(6) Blake, J.; Litzi-Davis, L. Evaluation of peptide libraries: An iterative strategy to analyze the reactivity of peptide mixtures with antibodies. *Bioconjugate Chem.* **1992**, *3*, 510–513.
(7) Geysen, H. M.; Rodda, S. J.; Mason, T. J.; Tribbick, G.; Schoofs, P. G. Strategies for epitope analysis using peptide synthesis. *J. Immunol. Methods* **1987**, *102*, 259–274.
(8) Houghten, R. A.; Pinilla, C.; Blondelle, S. E.; Appel, J. R.; Dooley, C. T.; Cuervo, J. H. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **1991**, *354*, 84–86.
(9) Houghten, R. A.; Appel, J. R.; Blondelle, S. E.; Cuervo, J. H.; Dooley, C. T.; Pinilla, C. The use of synthetic peptide combinatorial libraries for the identification of bioactive peptides. *BioTechniques* **1992**, *13*, 412–421.
(10) Edmundson, A. B.; Harris, D. L.; Fan, Z.-C.; Guddat, L. W.; Schley, B. T.; Hanson, B. L.; Tribbick, G.; Geysen, H. M. Principles and pitfalls in designing site-directed peptide ligands. *Proteins* **1993**, *16*, 246–267.
(11) Owens, R. A.; Gesellchen, P. D.; Houchins, B. J.; DiMarchi, R. D. The rapid identification of HIV protease inhibitors through the synthesis and screening of defined peptide mixtures. *Biochem. Biophys. Res. Commun.* **1991**, *181*, 402–408.
(12) Eichler, J.; Houghten, R. A. Identification of substrate-analog trypsin inhibitors through the screening of synthetic peptide combinatorial libraries. *Biochemistry* **1993**, *32*, 11035–11041.
(13) Freier, S. M.; Kierzek, R.; Jaeger, J. A.; Sugimoto, N.; Caruthers, M. H.; Neilson, T.; Turner, D. H. Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A* **1986**, *83*, 9373–9377.
(14) Turner, D. H.; Sugimoto, N.; Jaeger, J. A.; Longfellow, C. E.; Freier, S. M.; Kierzek, R. Improved parameters for prediction of RNA structure. *Cold Spring Harbor Symp. Quant. Biol.* **1987**, *52*, 123–133.
(15) Jaeger, J. A.; Turner, D. H.; Zuker, M. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 7706–7710.
(16) Freier, S. M.; Burger, B. J.; Alkema, D.; Neilson, T.; Turner, D. Effects of 3′ dangling end stacking on the stability of GGCC and CCGG double helices. *Biochemistry* **1983**, *22*, 6198–6206.
(17) Dooley, C. T.; Houghten, R. A. The use of positional scanning synthetic peptide combinatorial libraries for the rapid determination of opioid receptor ligands. *Life Sci.* **1993**, *52*, 1509–1517.
(18) Pinilla, C.; Appel, J. R.; Blanc, P.; Houghten, R. A. Rapid identification of high affinity peptide ligands using positional scanning synthetic peptide combinatorial libraries. *BioTechniques* **1992**, *13*, 901–905.
(19) Bevington, P. R.; Robinson, D. K. Monte carlo techniques. In *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed.; McGraw-Hill: San Francisco, 1992; pp 75–95.
(20) Dr. Peter Davis, personal communication.
(21) Breslauer, K. J.; Frank, R.; Blocker, H.; Marky, L. A. Predicting DNA duplex stability from base sequence. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3746–3750.
(22) Kierzek, R.; Caruthers, M. H.; Longfellow, C. E.; Swinton, D.; Turner, D. H. Polymer-supported RNA synthesis and its application to test the nearest-neighbor model for duplex stability. *Biochemistry* **1986**, *25*, 7840–7846.
(23) He, L.; Kierzek, R.; SantaLucia, J.; Walter, A. E.; Turner, D. H. Nearest neighbor parameters for G-U mismatches: 5′GU3′ is destabilizing in the contexts CGUG, UGUA and AGUU but stabilizing in GGUC. *Biochemistry* **1991**, *30*, 11124–11132.
(24) Lancet, D.; Sandovsky, E.; Seidemann, E. Probability model for molecular recognition in biological receptor repertoires: Significance to the olfactory system. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3715–3719.
(25) Kauffman, S. A. The structure of rugged fitness landscapes: The NK model of rugged fitness landscapes. In *The Origins of Order, Self-Organization and Selection in Evolution*, Oxford University Press: Oxford, UK, 1993; pp 40–45.
(26) Patchett, a. A. Excursions in drug discovery. *J. Med. Chem.* **1993**, *36*, 2051–2058.
(27) Zuker, M. On finding all suboptimal foldings of an RNA molecule. *Science* **1989**, *244*, 48–52.
(28) Jaeger, J. A.; Turner, D. H.; Zuker, M. Predicting optimal and suboptimal secondary structure for RNA. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods in Enzymology, 183;* Doolittle, R. F., Ed.; Academic Press: San Diego, 1990; pp 281–306.
(29) Needels, M. C.; Jones, D. G.; Tate, E. H.; Heinkel, G. L.; Kochersperger, L. M.; Dower, W. J.; Barrett, R. W.; Gallop, M. A. Generation and screening of an oligonucleotide-encoded synthetic peptide library. *PNAS* **1993**, *90*, 10700–10704.

JM9405452