JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Searching the Sequence Space for Potent Aptamers Using SELEX in Silico

Qingtong Zhou,[†,‡,⊥] Xiaole Xia,[§,‡,⊥] Zhaofeng Luo,[∥] Haojun Liang,[*,†] and Eugene Shakhnovich[*,‡]

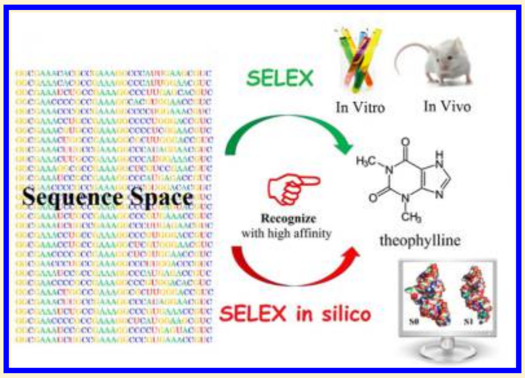
[†]CAS Key Laboratory of Soft Matter Chemistry, Collaborative Innovation Center of Chemistry for Energy Materials, Department of Polymer Science and Engineering, Hefei National Laboratory for Physical Sciences at Microscale, [∥]School of Life Science, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[§]Key Laboratory of Industrial Biotechnology, Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi, Jiangsu 214122, People's Republic of China

S *Supporting Information*

ABSTRACT: To isolate functional nucleic acids that bind to defined targets with high affinity and specificity, which are known as aptamers, the systematic evolution of ligands by exponential enrichment (SELEX) methodology has emerged as the preferred approach. Here, we propose a computational approach, SELEX in silico, that allows the sequence space to be more thoroughly explored regarding binding of a certain target. Our approach consists of two steps: (i) secondary structure-based sequence screening, which aims to collect the sequences that can form a desired RNA motif as an enhanced initial library, followed by (ii) sequence enrichment regarding target binding by molecular dynamics simulation-based virtual screening. Our SELEX in silico method provided a practical computational solution to three key problems in aptamer sequence searching: design of nucleic acid libraries, knowledge of sequence enrichment, and identification of potent aptamers. Six potent theophylline-binding aptamers, which were isolated by SELEX in silico from a sequence space containing 4^{13} sequences, were experimentally verified to bind theophylline with high affinity: K_d ranging from 0.16 to 0.52 μM, compared with the dissociation constant of the original aptamer-theophylline, 0.32 μM. These results demonstrate the significant potential of SELEX in silico as a new method for aptamer discovery and optimization.



1. INTRODUCTION

Understanding the relationship between sequence and function is important for elucidating the functional roles and biomedical applications of nucleic acids. Aptamers are short, single-stranded oligonucleotides that bind potently and selectively to their targets by adopting distinct secondary and tertiary structures. A general approach to searching the sequence space of an aptamer to achieve a desired strong binding to a target is SELEX,[1,2] which begins with an initial library of randomized sequences and then proceeds to repeat successive steps of selection (binding, partition, and elution) and amplification. Finally, after sequencing the enriched library, a few hundred sequences are chosen for detailed analysis to identify the optimal aptamer sequence. The secondary structures of the resulting aptamer[3] or the 3D structures of the bound complex[4] are further investigated to rationalize the selection. Because of the SELEX technique and dozens of variations thereof, aptamer-binding targets have expanded from small molecules and metal ions to proteins, biological cells, and tissues. Although is has been successful for many applications, SELEX is a time-consuming, low-throughput, and labor-intensive approach. Furthermore, SELEX suffers from several limitations, such as the dependence of the outcome on the choice of the initial library and the nonexhaustive search in the sequence space, in which many strong-binding aptamers might be missed.

The initial library of SELEX currently comprises up to 10^{18} sequences, but it represents only a tiny fraction of the total sequence space (10^{36} for a 60 nt random region library), consequently reducing the repertoire of discovered motifs. To fill this gap, the initial libraries obtained in experimental[5,6] and computational[7−9] approaches have been enriched to increase the likelihood of identifying high-affinity aptamers by narrowing the explored sequence space or by increasing the chemical and structural diversity of the nucleotides.[10−12] In an interesting application of the method, genomic SELEX[13] was performed to identify aptamers with adenosine triphosphate (ATP) and guanosine-5′-triphosphate (GTP)-binding motifs encoded in genomic sequences,[14,15] providing an interesting perspective on gene regulation. Nonetheless, the sequence space is far from being fully explored, and the outcome depends on the design of the initial library.

Received: July 25, 2015
Published: October 27, 2015

DOI: 10.1021/acs.jctc.5b00707

While SELEX has been efficient in finding initial hits, subsequent aptamer optimization remains challenging. The difficulty of exploring the fitness landscape around a binding aptamer is likely caused by the diminished selection pressure after SELEX arrives at a viable solution (law of diminishing returns in evolution).[16] Sequencing technologies[17] and parallel characterization methods, such as microarrays,[18] provide an opportunity to understand how SELEX searches sequence space and demonstrate how the enhanced library evolves step by step.[19,20] A comprehensive collection of functional RNA sequences with desired properties provides insights into the RNA sequence−function relationship and the inner workings of SELEX.

The sets of aptamer sequences discovered by SELEX that bind a given target are both highly diverse[21] and locally immutable,[18] suggesting a complicated, highly degenerate sequence−fitness relationship. Specifically, the number and distribution of tightly binding sequences for a given ligand are not known. For example, a triple mutation of an immunoglobulin E-binding aptamer was shown to exhibit a slight increase in binding affinity.[18]

A number of computational tools[22,23] have been employed to facilitate SELEX. Theoretical methods such as the entropic fragment-based approach[24] and RNAiFold[25] have successfully predicted functional RNA aptamers and ribozymes, respectively. RNA tertiary structure prediction,[26−28] structure-based docking,[29,30] and binding free energy calculation[31−34] have shown significant improvements in speed and accuracy. Molecular dynamics (MD) simulations have revealed several crucial mechanistic details related to RNA function.[35−39] For instance, a theophylline-binding RNA aptamer can recognize theophylline with a 10 000-fold greater affinity than that of its closest chemical homologue, caffeine, which differs from theophylline by only a single methyl group at N7. In several instances, MD simulations and binding free energy analysis have provided atomic-level explanations of the binding mechanism.[40,41] Mecozzi's group used MD simulations and free energy calculations to determine the minimal active sequences of RNA aptamers. They successfully reduced the aptamer sequences to 13, 14, and 21 nt for theophylline,[42] flavin mononucleotide,[43] and aminoglycosides,[44] respectively; the sequence lengths of the original aptamers were 33, 35, and 44 nt.

Here, we proposed a computational screening method, SELEX in silico, based on an exhaustive search of the relevant sequence space to identify potent binders. This method combines exhaustive searching in a sequence space for the desired motif with subsequent MD simulation-based virtual screening. SELEX in silico could be used to design an enhanced initial library from a sequence space, to visualize the sequence-enrichment process regarding a desired property, to rationally dissect the contributions of individual bases within the binding mechanism of specific aptamers, and to identify high-binding affinity aptamers. Using this approach, we searched a sequence space containing 4^{13} sequences for new theophylline-binding aptamers. We found six novel aptamers with similar or stronger theophylline-binding affinity than that of the original SELEX-identified aptamer. These results indicated that SELEX in silico is an efficient tool for aptamer discovery and optimization.

2. METHODS

2.1. Principles of SELEX in Silico. SELEX in silico is a two-step approach to search a sequence space for optimized aptamers with a known RNA−ligand complex structure (Figure 1). First, the sequence space is searched exhaustively for sequences that can form (in the apo or holo form) the secondary structure of the target aptamer conformation. For every sequence in the sequence space, the minimum free energy (MFE, ΔG_{MFE}), the free energy in the target motif state ($\Delta G_{\mathrm{target}}$), and the free-energy gap (FEG, $\Delta\Delta G_{\mathrm{gap}}$, defined as $\Delta G_{\mathrm{target}} - \Delta G_{\mathrm{MFE}}$), which represents the energy difference between the lowest energy state and the target state with regard to secondary structure, are calculated (Figure S1). Sequences that adopted the target motif as their MFE structure or with low FEG ($\Delta\Delta G_{\mathrm{gap}} \leq 0.1$ kcal/mol) were selected to populate an enhanced initial library. Second, MD-based virtual screening of the enhanced initial library is performed (Figure 1). The initial ligand-binding conformations of candidate sequences are generated by in silico base mutations of the known aptamer−ligand complex. Then, unrestrained MD simulations, allowing the initially assumed binding conformations to evolve dynamically to account for the molecular recognition processes of the ligand, are performed. The sequences for which the ligand escaped from the binding pocket or the hydrogen bonds between the ligand and RNA were destroyed are discarded. After several rounds of selection, the surviving sequences are expected to bind potently to the ligand. Finally, the sequences with the best predicted binding free energies were verified by experimental assays.

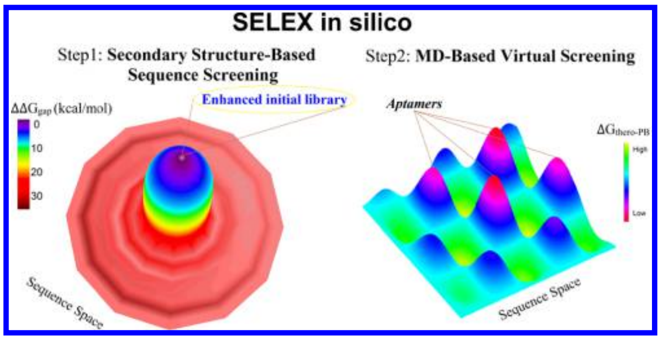

SELEX in silico
Step1: Secondary Structure-Based Sequence Screening
Step2: MD-Based Virtual Screening
Enhanced initial library
ΔΔG_{gap} (kcal/mol)
0
10
20
30
Aptamers
ΔG_{thero-PB}
High
Low
Sequence Space
Sequence Space

Figure 1. Schematic of the SELEX in silico aptamer-screening strategy. The strategy involves two steps: (1) secondary structure-based sequence screening, which aims to collect the sequences that can form a desired RNA motif as an enhanced initial library, and (2) sequence enrichment for theophylline binding by molecular dynamics simulation-based virtual screening.

2.2. Library Design. The original theophylline-binding aptamer consisted of the upper stem region (residues 11−20), the lower stem region (residues 1−4 and 30−33), and the core region (residues 5−10, 21−29, and theophylline).[40] No sequence mutations were considered in the upper and lower stem regions because of their relative distance from the binding pocket. Because the bases of C22 and U24 form hydrogen bonds with theophylline, these two positions were fixed at C22 and U24. The random mutations at sites 5−10, 21, 23, and 25−29 were 5′-GGCGNNNNNNGCCGAAAGGCNCNUNN-NNNCGUC-3′ (N = A, U, C, G). Thus, a sequence space comprising 4^{13} sequences was prepared.

2.3. Computational Procedure of SELEX in Silico. The original theophylline-binding RNA aptamer has a specified secondary structure in its theophylline-bound state, denoted ((((...((.(((....)))....))...)))) in dot-parens-plus notation. For every sequence in the sequence space (Figure S1), ΔG_{MFE} and $\Delta G_{\mathrm{target}}$ were calculated with the same empirical parameter,

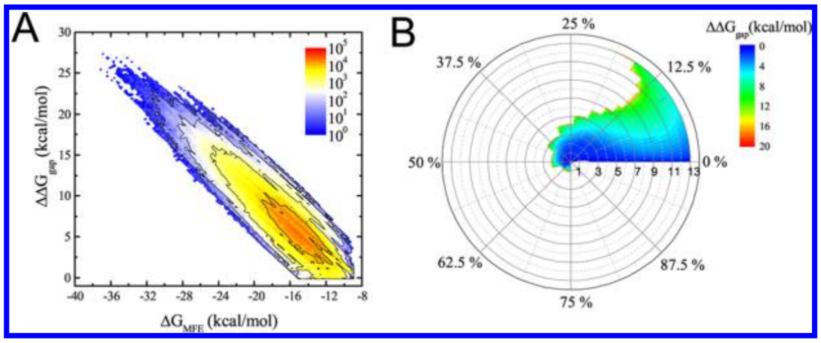


Figure 2. (A) Contour plot of the sequence density surfaces for the 9 437 029 target motif foldable sequences on the basis of ΔG_{MFE} and $\Delta\Delta G_{\text{gap}}$. Each point corresponds to a pair of ΔG_{MFE} and $\Delta\Delta G_{\text{gap}}$, and the color corresponds to the number of sequences that share identical values. Both ΔG_{MFE} (*x*-axis) and $\Delta\Delta G_{\text{gap}}$ (*y*-axis) are plotted in bins with 0.25 kcal/mol intervals. (B) Distribution of the free energy gaps for each subset of sequence space. The center of the polar plot is the original aptamer, the distance from the center indicates the number of different bases compared with the original aptamer, the angle indicates the proportion of target motif foldable sequences in each sequence subspace, and the color represents the corresponding free energy gap, $\Delta\Delta G_{\text{gap}}$. The blank region represents the percentage of sequences that is unable to fold into the target motif.

rna1995, by the NUPACK[45] executables *mfe* and *energy*, respectively. Then, $\Delta\Delta G_{\text{gap}}$ was determined. We selected 61 921 sequences with negligible FEG ($\Delta\Delta G_{\text{gap}} \leq 0.1$ kcal/mol) to populate the enhanced initial library.

Given the difficulty in predicting RNA tertiary structures and the unknown molecular mechanism by which RNA recognizes theophylline,[42] in this study, we focused on expanding the original aptamer[46] in the sequence space while retaining the original binding mode (a sandwich of three base triples). The initial coordinate for theophylline-bound RNA of the original aptamer was determined by Clore et al.[47] (PDB code 1O15). By performing in silico base mutations on the initial coordinate by the program mutate_bases of X3DNA,[48] we generated the binding conformations of RNA−theophylline for mutated sequences. The mutated sequences were explored by MD simulations to determine whether they were stable in the assumed binding mode and retained high binding affinity upon theophylline binding. The binding free energy calculated by the MM/PBSA method was chosen as the main selection criterion. The MD-based virtual screening process comprised several rounds. In the first round, 150 ps of restraint-free MD simulations was performed on the theophylline-binding complexes for 61 921 mutated sequences. In the second round, the 2220 sequences exhibiting high stability of the binding complex (root-mean-square deviation ≤ 0.1 nm) or low binding free energy ($\Delta G_{\text{MM/PBSA}} \leq -33$ kcal/mol or $\Delta G_{\text{MM/GBSA}} \leq -32$ kcal/mol) or forming more than three hydrogen bonds with theophylline were selected from the first round and subjected to 1 ns MD simulations. The MD simulation length in the third round increased to 10 ns for the 270 selected sequences. The fourth round of selection simulated the RNA−theophylline complex over 30 ns for 50 sequences. Finally, 100 ns MD simulations were performed for 24 sequences. The MD simulations were conducted using Gromacs 4.5.3.[49] MM/PBSA.py[50] in the AmberTools12 package was employed to calculate all components of the binding free energy.[51−53] More details on the computational methods and experimental validation are provided in the Supporting Information.

3. RESULTS

3.1. Comprehensive Analysis of RNA Secondary Structure.

The mapping of RNA sequence space into motif space was highly nonuniform, which was consistent with earlier findings for RNA and proteins.[54−56] Regarding the target motif accessibility (Figure S2), 57 671 835 sequences fell into the inaccessible region ($\Delta G_{\text{target}} > 0$) on the sequence density surface, i.e., their folding into the target motif was energetically prohibitive. On the basis of the best-fit least-squares regression on 9 437 029 pairs of ΔG_{MFE} and $\Delta\Delta G_{\text{gap}}$ contributed by target motif foldable sequences, we found the following quantitative relationship:

$$\Delta\Delta G_{\text{gap}} = 0.965 \times [(-9.378) - \Delta G_{\text{MFE}}],\ R^2 = 0.783 \qquad (1)$$

The averaged $\Delta\Delta G_{\text{gap}}$ for these sequences was 6.47 ± 3.36 kcal/mol. Only 59 755 optimal sequences adopted the target motif as their MFE structure, whereas the remaining 9 377 274 sequences had a lower ΔG_{MFE} than ΔG_{target} ($\Delta\Delta G_{\text{gap}} > 0$), indicating that these sequences had to pay a significant free energy penalty to form the target motif (Figure 2A), which weakened the overall binding affinity for theophylline.

To investigate how simultaneous mutations affect the secondary structure, we comprehensively analyzed 13 mutations of the original theophylline-binding aptamer. Here, the sequence subspace was the subset of sequence space defined according to the sequence similarity and composed of mutated sequences with the same Hamming distance to the original aptamer. Thus, it contains $3^n \times C_{13}^n$ sequences, where n is the Hamming distance. As presented in Figure 2B, the percentage of optimal sequences with zero FEG found in the sequence subspace will decrease significantly as n increases, indicating that, from the perspective of secondary structure, the aptamer has little tolerance for few mutations and presents as a sharp peak in a rugged landscape. When seven or more mutations simultaneously occur, the mutated sequences share very low probability (approximately 0.08%) of being an optimal sequence. Interestingly, the number of optimal sequences increased dramatically when the volume of the sequence subspace increased rapidly (Figure S3). These results suggest that the point mutation approach is less efficient in generating optimal RNA sequences and generally fails to improve aptamer binding affinity because the mutated sequences likely lose their secondary structure features.

To improve the efficiency and reduce the computational burden, 61 921 sequences with negligible FEG ($\Delta\Delta G_{\text{gap}} \leq 0.1$ kcal/mol) were collected as the enhanced initial library. As demonstrated by previous research,[57,58] C27 is highly dynamic and not directly involved with theophylline binding, whereas the U27 and G27 mutations were found to bind theophylline with low affinity, probably forming unexpected but stable interactions in the free RNA, such as A7−U27, that block theophylline binding. A7−U27, U7−A27, C7−G27, G7−C27, G7−U27, and U7−G27 were not found in the enhanced initial library, confirming that our secondary structure prediction successfully removed the mutations that prevented base-platform structural motif formation.

3.2. Sequence Enrichment toward Theophylline Binding. The high binding affinity of the original aptamer to theophylline was verified from explicitly solvated MD trajectories with three hydrogen bonds between RNA and theophylline, the structural stability of the binding complex, and the calculated binding free energy. Potent aptamers were expected to have similar features. Thus, the mutated sequences that formed more than three hydrogen bonds with theophylline, that showed high stability of the binding complex, or that had comparable predicted binding free energies were retained for the next round of MD-based virtual screening processes (see Methods). As the simulation time increased, the mutated sequences achieved molecular recognition of theophylline. Similar to SELEX, MD-based virtual screening biased the initial library toward theophylline-binding sequences after several rounds of sequence enrichment and identified potential aptamers.

Sequence logos that indicated the content (relative frequency) of each nucleotide at each position were generated for the surviving sequences after each round of selection. We observed obvious sequence enrichment, especially at positions 6, 7, 8, 23, 26, and 28, accompanied by an ability of the surviving sequences to recognize theophylline with high affinity. As presented in Figure 3A, U at position 6, C at position 8, G at position 26, and A at position 28, exactly the same bases as those in the original aptamer, became the dominant bases and were responsible for the specific binding with theophylline. The base preferences at positions 10, 21, 27, and 29 did not change markedly compared with those in the enhanced initial library, suggesting that these bases contribute to maintaining the specified secondary structure. The bases were randomly distributed across A, U, C, and G for positions 5, 9, and 23.

In addition to the hydrogen-bonding interactions with C22 and U24, theophylline binding was stabilized by two base triples, i.e., U6−U23−A28 and A7−C8−G26. To highlight the importance of base triples for binding, we counted the number of sequences that shared identical base triples and calculated the percentage, especially for the two typical base triples of 6−23−28 and 7−8−26 (Figure 3B). Starting from 2.23% in the enhanced initial library, the percentage of U6−U23−A28 increased to 25% after four rounds of MD-based virtual screening processes. In contrast, the percentage of A7−C8−G26 did not change substantially from 8.25% in the enhanced initial library, increasing to 14.17% among the seven best sequences. U6−G23−A28, which is different only in one base from U6−U23−A28, was found to exhibit the same percentage (25%) as U6−U23−A28 in the fourth round. Consequently, U6 and A28 were relatively conserved. A similar phenomenon was observed for 7−8−26 base triples: the two most dominant base triples were G7−C8−G26 and U7−C8−G26, which showed the same proportion of 37.5% in the 24 best sequences because of the random choice of G and U at position 7. Different enrichment profiles of the base triples provided not only a clear picture of how selection pressure shaped the base triple composition toward theophylline binding but also a qualitative understanding of the inherent relationship among bases.

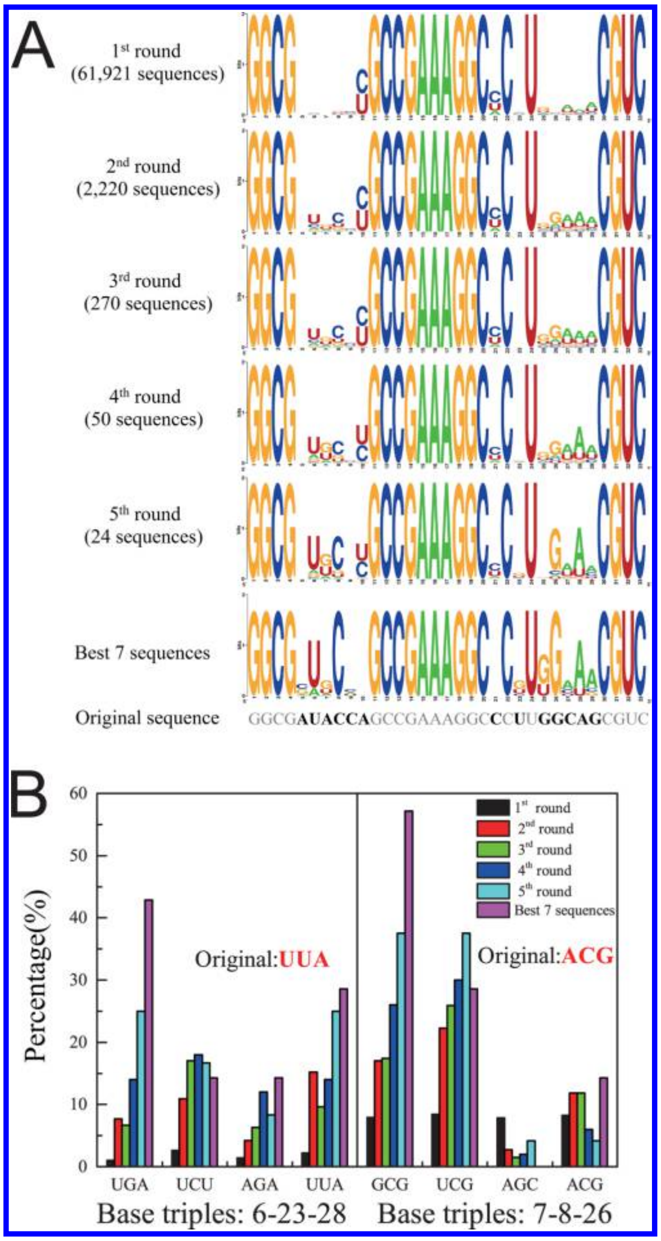


Figure 3. Sequence enrichment regarding theophylline binding using SELEX in silico. (A) Structural RNA logo analysis. The heights of the four letters within the stack of sequence logos generated by Weblogo indicate the content (relative frequency) of each nucleotide at each position in the selected sequences. (B) Evolution of two typical triple base compositions during rounds of SELEX in silico. In addition to the hydrogen-bonding interactions with C22 and U24, theophylline is stabilized by two base triples of the original theophylline-binding aptamer: U6−U23−A28 and A7−C8−G26.

3.3. Theophylline-Binding Aptamers Isolated by SELEX in Silico. As shown in Table 1, the predicted binding free energies of S1 were lower than those of the original sequence S0 for both $\Delta G_{\text{theor-PB}}$ and $\Delta G_{\text{theor-GB}}$. Moreover, the theophylline−S1 binding complex was quite stable, as suggested by a clustering analysis of 2000 frames extracted from the last 20 ns MD simulation trajectory; the most populated cluster accounted for 94.5% of the conformations (Table S1). S2 and S6 formed more than three hydrogen bonds

Table 1. Six Best Theophylline-Binding Aptamers Isolated by SELEX in Silico

sequence ID	sequence (5′ → 3′)	$\Delta G_{\text{theor-PB}}$[a] (kcal/mol)	$\Delta G_{\text{theor-GB}}$[a] (kcal/mol)	T_m (°C)	K_d (μM) theophylline	caffeine
original (S0)	GGCGAUACCAGCCGAAAGGCCCUUGGCAGCGUC	−23.42	−21.58	71.4	0.355 ± 0.074	3522 ± 590
sequence 1 (S1)	GGCGGUGCUCGCCGAAAGGCUCCUGGAUACGUC	−27.83	−24.75	70.0	0.159 ± 0.027	2464 ± 685
sequence 2 (S2)	GGCGGUGCUGGCCGAAAGGCGCUUGGAAACGUC	−26.12	−19.07	70.9	0.391 ± 0.092	4061 ± 1555
sequence 3 (S3)	GGCGGUUCCUGCCGAAAGGCUCGUGGCAACGUC	−26.71	−22.68	68.1	0.519 ± 0.018	5510 ± 1655
sequence 4 (S4)	GGCGCUGCACGCCGAAAGGCCCGUUGAAACGUC	−25.17	−21.73	69.1	0.368 ± 0.081	660 ± 140
sequence 5 (S5)	GGCGCUUCAUGCCGAAAGGCCCGUUGUAACGUC	−24.77	−22.06	63.3	0.322 ± 0.090	2537 ± 824
sequence 6 (S6)	GGCGCAGCCCGCCGAAAGGCACGUGGAACCGUC	−24.83	−23.03	73.7	0.264 ± 0.063	2067 ± 870

[a] The predicted binding free energy, $\Delta G_{\text{theor-PB}}$ or $\Delta G_{\text{theor-GB}}$, was determined for RNA−theophylline complexes.

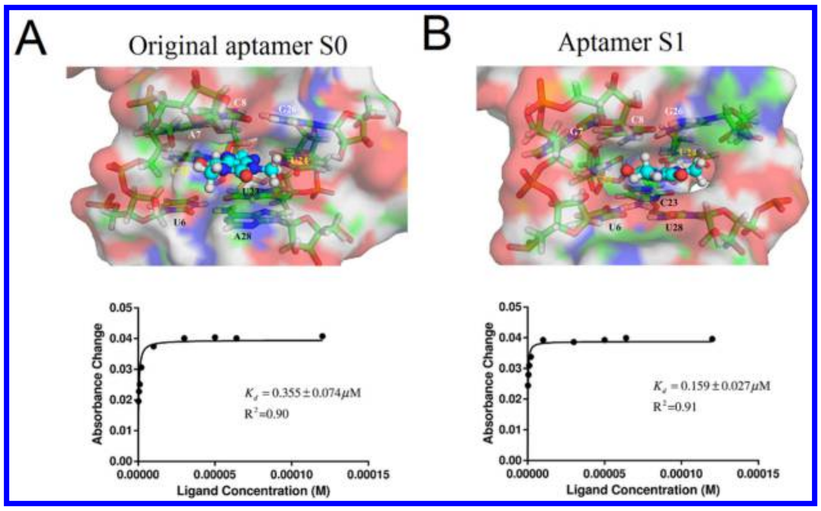


Figure 4. Structural representation of binding pockets and experimental binding affinity measurements of the original theophylline-binding aptamer (A) and the SELEX in silico-identified theophylline-binding RNA sequence, S1 (B).

with theophylline because of the additional hydrogen bond from the 2-hydroxyl of the ribose rings in U6 of S2 or A6 of S6. Additionally, we found that S3− S5 have one identical base triple of U6−G23−A28; S1, S2, S4, and S5 share G7−C8−G26; and S3 and S5 have another identical base triple of U7−C8−G26. Interestingly, S3 and S5 formed the same sandwich of three base triples for theophylline binding and shared similar binding complex structures, probably because the sequence variations between S3 and S5 were located far from the binding pocket.

3.4. Experimental Characterization. In vitro binding assays[42] were conducted for the six best predicted RNA aptamers to evaluate their binding affinity and selectivity between theophylline and caffeine (Supporting Information Text S1). The ultraviolet absorbances of the mixtures were fit to a 1:1 nonlinear binding isotherm (Figure S4). The regression results indicated that all six aptamers have a high binding affinity with theophylline (Table 1). S1 and S6 have a higher K_d than the positive control, S0, which has a binding affinity to theophylline of 0.355 ± 0.074 μM. S1 is an especially strong binder, exhibiting 2-fold higher binding affinity (K_d = 0.159 ± 0.027 μM) (Figure 4). S2, S4, and S5 have binding affinities similar to that of the original aptamer, S0, whereas that of S3 is slightly lower. We found that these six aptamers had 10 000-fold lower binding affinity to caffeine (Figure S5), which differs from theophylline only by one methyl group. The low correlation coefficient (R = 0.29, P < 0.52) between log[K_d(caffeine)/K_d(theophylline)] and log K_d(theophylline) for seven theophylline-binding aptamers (Figure S6) was similar to that found in a previous study on GTP aptamers,[59] which suggested that the increasing specificity upon ligand binding may be achieved by direct selection for specificity rather than as a side effect of an aptamer's high affinity.

Compared with the high accuracy of predicting the binding free energy of the original aptamer upon binding theophylline and its derivatives,[40,41] the variation in the molecular recognition processes of the aptamer−theophylline systems,[60−63] inadequate sampling of the conformational space in limited simulation times, and different entropy change contributions across complexes of the same ligand with different receptors[52,64−66] make it difficult to achieve equivalent accuracy for predicting the binding free energy of different aptamers upon theophylline binding. The correlation between the experimental results and the predicted binding free energy by MM/PBSA method (correlation coefficient R for $\Delta G_{\text{theor-PB}}$ and ΔG_{exp} of 0.49, P < 0.264) was inferior to that produced by the MM/GBSA method (for $\Delta G_{\text{theor-GB}}$ of 0.73, P < 0.065). S2 and S6 had relatively low percentages of finding the most populated cluster in the 2000 extracted snapshots, indicating that 100 ns MD simulations were probably not sufficient for aptamers such as S2 and S6 to achieve the equilibrated binding state. For S2 especially, $\Delta G_{\text{theor-GB}}$ was higher than expected. Without S2, the correlation coefficient R between $\Delta G_{\text{theor-GB}}$ and $\Delta G_{\text{exp.}}$ increased to 0.90 (P < 0.014).

The experimental results verified that the SELEX in silico approach successfully identified theophylline-binding aptamers via an exhaustive search of the sequence space of the aptamer's core region.

DOI: 10.1021/acs.jctc.5b00707

3.5. Fitness Peaks and Fitness Landscape. Experimental SELEX can be viewed as a general approach to mapping the landscape of sequences to their fitness, defined as the binding affinity to a certain target. However, only a small number of isolated aptamer sequences was characterized for the binding affinity at the end, providing only limited insight into the fitness landscape. SELEX in silico scanned a large fraction of the relevant sequence space. Here, the binding affinities for 270 sequences were predicted from 10 ns MD simulations, thereby providing an opportunity to obtain a comprehensive view of the theophylline-binding fitness landscape. Sequences that were ignored or removed during the selection processes of SELEX in silico correspond to zero in the sequence−fitness landscape.

The average intersequence Hamming distance among 270 sequences was 8.24 ± 2.09, whereas the value for S0−S6 was 6.67 ± 1.56 (Figures S7 and S8). By performing sequence alignment, we found that only 41 sequences can be directly mutated to fitness peaks S0−S6 within three mutations and that 31.44% of 264 sequences are within four mutations of S0−S6. In particular, S0 has zero neighbors within three mutations (Figure 5A). These results support the conclusion that the landscape was composed of largely disconnected islands of active sequences, consistent with earlier findings.[19] 264 sequences (except S0−S6) were grouped into fitness peaks S0−S6 by their intersequence Hamming distances (Figure 5B). Interestingly, the fitness peaks have two subtypes: sharp peaks with few neighbors (S0, S3, and S6) and twin peaks surrounded by many neighbors with reduced affinity (S1−S2 and S4−S5). The original aptamer sequence, S0, had zero active neighbors (single, double, or triple mutations), whereas fitness peaks S3 and S6 had three and two active triple neighbors, respectively. In contrast, the S1 and S2 twin peaks have 21 active neighbors, whereas S4 and S5 have 15 neighbors. The fitness landscape near the twin peaks is smooth and contains many mutations with reduced affinity, suggesting that these active sequences probably require step-by-step mutational pathways to connect the peaks (Table S2). Additionally, S4 and S5 could be connected by two different evolutionary pathways. Here, we focus on all 13 mutations around the binding pocket, and 270 sequences are presented. The results imply that the fitness landscape of the full sequence space consists of multiple peaks connected by various evolutionary pathways.

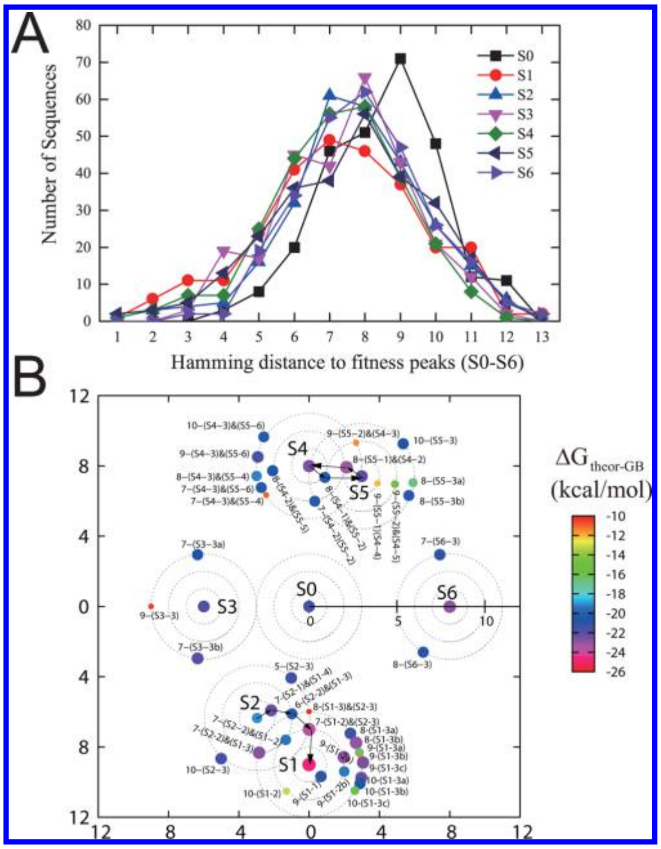


Figure 5. (A) Distribution of Hamming distances for the 270 sequences obtained by SELEX in silico into fitness peaks. For each fitness peak, we calculated the Hamming distance from the other 270 sequences. (B) Network representation of the RNA fitness landscape based on the calculated theophylline-binding affinity. Each sequence is represented by a point, and the binding free energy calculated by the MM/GBSA method determines the point's color and size. S0 is the original theophylline-binding aptamer, and S1−S6 are the top six sequences isolated by SELEX in silico. Related sequences are represented by successive rings of single, double, and triple mutants around the fitness peaks and are named in the graph in the following manner: Hamming distance to S0−(the closest peak-correlative Hamming distance), especially for the sequences whose Hamming distances to the next closest peak are not more than 3, additional arguments & (the next closest peak-correlative Hamming distance).

4. DISCUSSION

In this work, we performed an exhaustive search of the aptamer sequence subspace to identify potential strong binders of theophylline. The search encompassed 4^{13} sequences constituting all possible mutations in 13 positions in the core region. We employed a computational strategy named SELEX in silico, which was conceptually similar to SELEX but was conducted in silico rather than in vitro or in vivo. Our SELEX in silico method provided a practical computational solution to three key problems in aptamer sequence searching: designing the initial library, opening the black box of SELEX by exhaustive sequence enrichment, and identifying fitness peaks in a sequence space.

The initial experimental[5,6] and computational[7−9] nucleic acid library designs demonstrated significant enrichment in high-affinity aptamers versus a totally random library. To rationally design the initial library, we investigated the secondary structures of 4^{13} sequences and found their MFE structure distribution to be highly nonuniform, consistent with earlier findings.[54] From the sequence density surface of the target motif transformation, we found that 85.9% of the 4^{13} sequences were totally unable to form the target motif because of improper bases at certain positions and that only 0.09% of the sequences had the target motif as their MFE structure. We found that for sequences that were foldable into the target motif, the value of $\Delta\Delta G_{\text{gap}}$ between their MFE and free energy in the target motif exhibited a strong linear correlation with ΔG_{MFE}. Therefore, sequences with a low MFE would pay a great free energy penalty to form the target motif that binds theophylline. This cost is paid at the expense of strong binding affinity to the target ligand. We also observed that as the mutated sequences became increasingly different from the original aptamer the number of optimal sequences whose FEG for target secondary structure formation is zero increased significantly and that the percentage of optimal sequences in the corresponding sequence subspace decreased dramatically. Thus, the multimutation strategy in aptamer sequence optimization probably generated a considerable number sequences with correct secondary motifs and increased the probability of discovering a good aptamer sequence. These results suggested that secondary structure analysis in the first step of SELEX in silico effectively biased the initial library design toward the desired RNA motif.

As a direct evolution approach, experimental SELEX is akin to a black box in solving the sequence selection problem. In this study, we conducted several rounds of MD-based virtual screening. While the number of sequences decreased after each round, the length of the MD simulation was increased to provide an accurate estimate of the binding free energies for sequences that survived subsequent rounds of selection. This procedure provided insight into the physical−chemical properties of the subset of sequences selected in the computations under the pressure to bind theophylline with high affinity and stability. This in silico selection pressure enriched the enhanced initial library with sequences having high stability and strong binding affinity. Putative theophylline-binding aptamer sequences were predicted as fitness peaks in the fitness landscape. We observed the enrichment profiles of U6, C8, G26, and A28 and rationalized the dynamic role of C27.[57,58] The asynchronized enrichment of bases can be related to their different roles in the binding. We identified de novo base triples, including U6−G23−A28 and G7−C8−G26, in the theophylline-binding pocket. The results provided a rationale for the selection mechanism to generate a diverse set of predicted functional sequences.

The fitness landscape for theophylline binding was reconstructed based on the predicted binding free energies of 270 sequences, which exhibited a high diversity in their sequence composition. The averaged Hamming distance among the 270 sequences was 8.24, and 223 sequences were more than three bases different from fitness peaks S0−S6; thus, the fitness landscape consisted of largely disconnected, independent fitness peaks.[19] Our comprehensive exploration of the sequence space identified potential step-by-step mutational pathways to connect fitness peaks S4−S5 and S1−S2, which implied that the transformation from one functional activity to another or that a new function[59,67] could be achieved by stepwise mutations among highly connected fitness peaks immersed in an immense sequence space.

As verified by the experimental measurements, the binding affinity of the best sequence determined by SELEX in silico was two times greater than that of the original aptamer, whereas five other sequences had similar binding affinities. These results suggested that the SELEX in silico method successfully enriched the space sequences that can recognize theophylline with high binding affinity. Considering the significant high binding affinity of the original aptamer relative to other aptamers[4] and the relatively high coverage of the sequence space in the SELEX experiment regarding theophylline-binding,[68] the original theophylline-binding aptamer probably arrived at an optimum value,[16] which makes the isolation of aptamers with a nanomolar dissociation constant very challenging. Given the enormous computational demands, we performed MD-based virtual screening for 61 921 selected sequences from among the 332 704 sequences whose $\Delta\Delta G_{\text{gap}}$ are lower than 1.05 kcal/mol (Figure S9). Better aptamers could be achieved by examining more sequences.

The SELEX in silico approach has several limitations. The enhanced initial library containing 61 921 sequences presented a significant computational burden for molecular simulation and is still far from covering the complete set of potential sequences. In this study, to increase the screening efficiency, we assumed that all sequences had a similar structural binding mode to the original theophylline-binding aptamer. This assumption might exclude other potential theophylline-binding conformations[42] and limit the potential to generate a comprehensive view of the fitness landscape. Further investigations in this direction would require using a number of additional tools, such as RNA 3D structure prediction, ligand docking, and other free energy calculation methods, leading to greater computational cost. In this study, we chose a less-demanding approach that started from a known aptamer complex. While computationally less demanding, this approach was limited in its ability to discover novel aptamers with unknown binding complex structures. We cannot exclude the possibility that our selection procedure generated false negatives by discarding sequences that might undergo significant conformational changes from the initial structure, which could not be observed in relatively short simulations. SELEX in silico, similar to any other computational method, generally faced the dilemma of balancing efficiency and accuracy. This method appeared to find an appropriate balance and provided several novel aptamers, thereby extending the solution space for an important target ligand.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00707.

> Computational methods; experimental procedures and characterization data for theophylline-binding aptamers; figures related to detailed binding data, the free energy gap, and the fitness landscape; tables of computed binding free energies and potential evolutionary pathways among fitness peaks (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*(H.L.) E-mail: hjliang@ustc.edu.cn.
*(E.S.) E-mail: shakhnovich@chemistry.harvard.edu.

Author Contributions

[⊥]Q.Z. and X.X. contributed equally to this work.

Funding

We thank the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University, and the Supercomputing Center of USTC for supporting parallel computations. This project is supported by the National Natural Science Foundation of China [21434007 and 91127046 to H.J.], the National Basic Research Program of China [2012CB821500 to H.J.], the National Institutes of Health [GM068670 to E.S.], and the China Scholarship Council.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Ellington, A. D.; Szostak, J. W. *Nature* **1990**, *346*, 818−22.
(2) Tuerk, C.; Gold, L. *Science* **1990**, *249*, 505−10.
(3) Bing, T.; Yang, X.; Mei, H.; Cao, Z.; Shangguan, D. *Bioorg. Med. Chem.* **2010**, *18*, 1798−805.
(4) Hermann, T.; Patel, D. J. *Science* **2000**, *287*, 820−5.
(5) Ruff, K. M.; Snyder, T. M.; Liu, D. R. *J. Am. Chem. Soc.* **2010**, *132*, 9453−64.

(6) Davis, J. H.; Szostak, J. W. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 11616−21.

(7) Kim, N.; Gan, H. H.; Schlick, T. *RNA* **2007**, *13*, 478−92.

(8) Chushak, Y.; Stone, M. O. *Nucleic Acids Res.* **2009**, *37*, e87.

(9) Kim, N.; Izzo, J. A.; Elmetwaly, S.; Gan, H. H.; Schlick, T. *Nucleic Acids Res.* **2010**, *38*, e139.

(10) Pinheiro, V. B.; Taylor, A. I.; Cozens, C.; Abramov, M.; Renders, M.; Zhang, S.; Chaput, J. C.; Wengel, J.; Peak-Chew, S. Y.; McLaughlin, S. H.; Herdewijn, P.; Holliger, P. *Science* **2012**, *336*, 341−4.

(11) Kimoto, M.; Yamashige, R.; Matsunaga, K.; Yokoyama, S.; Hirao, I. *Nat. Biotechnol.* **2013**, *31*, 453−7.

(12) Sefah, K.; Yang, Z.; Bradley, K. M.; Hoshika, S.; Jimenez, E.; Zhang, L.; Zhu, G.; Shanker, S.; Yu, F.; Turek, D.; Tan, W.; Benner, S. A. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 1449−54.

(13) Lorenz, C.; von Pelchrzim, F.; Schroeder, R. *Nat. Protoc.* **2006**, *1*, 2204−12.

(14) Vu, M. M.; Jameson, N. E.; Masuda, S. J.; Lin, D.; Larralde-Ridaura, R.; Luptak, A. *Chem. Biol.* **2012**, *19*, 1247−54.

(15) Curtis, E. A.; Liu, D. R. *Chem. Biol.* **2013**, *20*, 521−32.

(16) Hartl, D. L.; Dykhuizen, D. E.; Dean, A. M. *Genetics* **1985**, *111*, 655−74.

(17) Schutze, T.; Wilhelm, B.; Greiner, N.; Braun, H.; Peter, F.; Morl, M.; Erdmann, V. A.; Lehrach, H.; Konthur, Z.; Menger, M.; Arndt, P. F.; Glokler, J. *PLoS One* **2011**, *6*, e29604.

(18) Katilius, E.; Flores, C.; Woodbury, N. W. *Nucleic Acids Res.* **2007**, *35*, 7626−35.

(19) Jimenez, J. I.; Xulvi-Brunet, R.; Campbell, G. W.; Turk-MacLeod, R.; Chen, I. A. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 14984−9.

(20) Cho, M.; Soo, O. S.; Nie, J.; Stewart, R.; Eisenstein, M.; Chambers, J.; Marth, J. D.; Walker, F.; Thomson, J. A.; Soh, H. T. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 18460−5.

(21) Tran, T.; Disney, M. D. *Nat. Commun.* **2012**, *3*, 1125.

(22) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188−90.

(23) Hoinka, J.; Berezhnoy, A.; Dao, P.; Sauna, Z. E.; Gilboa, E.; Przytycka, T. M. *Nucleic Acids Res.* **2015**, *43*, 5699.

(24) Tseng, C. Y.; Ashrafuzzaman, M.; Mane, J. Y.; Kapty, J.; Mercer, J. R.; Tuszynski, J. A. *Chem. Biol. Drug Des.* **2011**, *78*, 1−13.

(25) Dotu, I.; Garcia-Martin, J. A.; Slinger, B. L.; Mechery, V.; Meyer, M. M.; Clote, P. *Nucleic Acids Res.* **2014**, *42*, 11752−62.

(26) Parisien, M.; Major, F. *Nature* **2008**, *452*, 51−5.

(27) Das, R.; Karanicolas, J.; Baker, D. *Nat. Methods* **2010**, *7*, 291−4.

(28) Fulle, S.; Gohlke, H. *J. Mol. Recognit.* **2009**, *23*, 220−31.

(29) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. *RNA* **2009**, *15*, 1219−30.

(30) Daldrop, P.; Reyes, F. E.; Robinson, D. A.; Hammond, C. M.; Lilley, D. M.; Batey, R. T.; Brenk, R. *Chem. Biol.* **2011**, *18*, 324−35.

(31) Reyes, C. M.; Kollman, P. A. *J. Mol. Biol.* **2000**, *297*, 1145−58.

(32) Xiao, S.; Klein, M. L.; LeBard, D. N.; Levine, B. G.; Liang, H.; MacDermaid, C. M.; Alfonso-Prieto, M. *J. Phys. Chem. B* **2014**, *118*, 873−89.

(33) Furini, S.; Barbini, P.; Domene, C. *Nucleic Acids Res.* **2013**, *41*, 3963−72.

(34) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. *J. Am. Chem. Soc.* **2015**, *137*, 2695−703.

(35) Heldenbrand, H.; Janowski, P. A.; Giambasu, G.; Giese, T. J.; Wedekind, J. E.; York, D. M. *J. Am. Chem. Soc.* **2014**, *136*, 7789−92.

(36) Musiani, F.; Rossetti, G.; Capece, L.; Gerger, T. M.; Micheletti, C.; Varani, G.; Carloni, P. *J. Am. Chem. Soc.* **2014**, *136*, 15631−7.

(37) Matthies, M. C.; Bienert, S.; Torda, A. E. *J. Chem. Theory Comput.* **2012**, *8*, 3663−3670.

(38) Přecechtělová, J.; Munzarová, M. L.; Vaara, J.; Novotný, J.; Dračinský, M.; Sklenář, V. *J. Chem. Theory Comput.* **2013**, *9*, 1641−56.

(39) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham, T. E., III *J. Chem. Theory Comput.* **2014**, *10*, 492−9.

(40) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. *Biopolymers* **2003**, *68*, 16−34.

(41) Freedman, H.; Huynh, L. P.; Le, L.; Cheatham, T. E., III; Tuszynski, J. A.; Truong, T. N. *J. Phys. Chem. B* **2010**, *114*, 2227−37.

(42) Anderson, P. C.; Mecozzi, S. *J. Am. Chem. Soc.* **2005**, *127*, 5290−1.

(43) Anderson, P. C.; Mecozzi, S. *Nucleic Acids Res.* **2005**, *33*, 6992−9.

(44) Anderson, P. C.; Mecozzi, S. *Biopolymers* **2007**, *86*, 95−111.

(45) Dirks, R. M.; Pierce, N. A. *J. Comput. Chem.* **2004**, *25*, 1295−304.

(46) Zimmermann, G. R.; Jenison, R. D.; Wick, C. L.; Simorre, J. P.; Pardi, A. *Nat. Struct. Biol.* **1997**, *4*, 644−9.

(47) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2003**, *125*, 1518−25.

(48) Lu, X. J.; Olson, W. K. *Nat. Protoc.* **2008**, *3*, 1213−27.

(49) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845−54.

(50) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.

(51) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211−43.

(52) Stoica, I.; Sadiq, S. K.; Coveney, P. V. *J. Am. Chem. Soc.* **2008**, *130*, 2639−48.

(53) Ravindranathan, K.; Tirado-Rives, J.; Jorgensen, W. L.; Guimaraes, C. R. *J. Chem. Theory Comput.* **2011**, *7*, 3859−3865.

(54) Schuster, P.; Fontana, W.; Stadler, P. F.; Hofacker, I. L. *Proc. R. Soc. London, Ser. B* **1994**, *255*, 279−84.

(55) *The Aptamer Handbook: Functional Oligonucleotides and Their Applications*; Klussmann, S., Ed.; Wiley-VCH: Weinheim, Germany, 2006.

(56) Bourdeau, V.; Ferbeyre, G.; Pageau, M.; Paquin, B.; Cedergren, R. *Nucleic Acids Res.* **1999**, *27*, 4457−67.

(57) Zimmermann, G. R.; Shields, T. P.; Jenison, R. D.; Wick, C. L.; Pardi, A. *Biochemistry* **1998**, *37*, 9186−92.

(58) Lee, S. W.; Zhao, L.; Pardi, A.; Xia, T. *Biochemistry* **2010**, *49*, 2943−51.

(59) Huang, Z.; Szostak, J. W. *RNA* **2003**, *9*, 1456−63.

(60) Latham, M. P.; Zimmermann, G. R.; Pardi, A. *J. Am. Chem. Soc.* **2009**, *131*, 5052−3.

(61) Lee, S. W.; Zhao, L.; Pardi, A.; Xia, T. *Biochemistry* **2010**, *49*, 2943−51.

(62) Jucker, F. M.; Phillips, R. M.; McCallum, S. A.; Pardi, A. *Biochemistry* **2003**, *42*, 2560−7.

(63) Lin, P. H.; Tsai, C. W.; Wu, J. W.; Ruaan, R. C.; Chen, W. Y. *Biotechnol. J.* **2012**, *7*, 1367−75.

(64) Hou, T.; Yu, R. *J. Med. Chem.* **2007**, *50*, 1177−88.

(65) Xu, L.; Sun, H.; Li, Y.; Wang, J.; Hou, T. *J. Phys. Chem. B* **2013**, *117*, 8408−21.

(66) Sun, H.; Li, Y.; Tian, S.; Xu, L.; Hou, T. *Phys. Chem. Chem. Phys.* **2014**, *16*, 16719−29.

(67) Schultes, E. A.; Bartel, D. P. *Science* **2000**, *289*, 448−52.

(68) Jenison, R. D.; Gill, S. C.; Pardi, A.; Polisky, B. *Science* **1994**, *263*, 1425−9.