Nucleic Acids Research Advance Access published May 21, 2009

Nucleic Acids Research, 2009, 1–9
doi:10.1093/nar/gkp408

In silico selection of RNA aptamers

Yaroslav Chushak[1,*] and Morley O. Stone[2]

[1]Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, MD 21702 and [2]Human Effectiveness Directorate, 711 Human Performance Wing, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

Received February 11, 2009; Revised April 30, 2009; Accepted May 2, 2009

ABSTRACT

***In vitro* selection of RNA aptamers that bind to a specific ligand usually begins with a random pool of RNA sequences. We propose a computational approach for designing a starting pool of RNA sequences for the selection of RNA aptamers for specific analyte binding. Our approach consists of three steps: (i) selection of RNA sequences based on their secondary structure, (ii) generating a library of three-dimensional (3D) structures of RNA molecules and (iii) high-throughput virtual screening of this library to select aptamers with binding affinity to a desired small molecule. We developed a set of criteria that allows one to select a sequence with potential binding affinity from a pool of random sequences and developed a protocol for RNA 3D structure prediction. As verification, we tested the performance of *in silico* selection on a set of six known aptamer–ligand complexes. The structures of the native sequences for the ligands in the testing set were among the top 5% of the selected structures. The proposed approach reduces the RNA sequences search space by four to five orders of magnitude—significantly accelerating the experimental screening and selection of high-affinity aptamers.**

INTRODUCTION

Aptamers are single-stranded DNA or RNA molecules that bind to a specific ligand with high affinity and specificity. They have been extensively explored for sensing and diagnostic applications and for the regulation of gene expression via synthetic riboswitches (1). Aptamers can be selected *in vitro* from a random pool of DNA or RNA molecules (typically 10^{14}–10^{15} different sequences) using an iterative process called systematic evolution of ligands by exponential enrichment (SELEX) (2,3). The SELEX process consists of multiple cycles of selection and amplification: (i) a pool of RNA molecules is screened and aptamers with a binding affinity to a target molecule are separated from non-aptamers and (ii) retained aptamers are amplified by the polymerase chain reaction (PCR) to create a pool of sequences for the next round of enrichment. The entire selection process typically requires up to 15 rounds of selection and can take from a few days to a few months to complete (4).

In recent years, RNA microarrays have emerged as a new approach for high-throughput aptamer selection (5–7). Current technology allows rapid preparation of a large custom microarray with tens of thousands of probes (Agilent Technologies, CombiMatrix). For example, DNA microarrays have been used recently to explore the relationship between the aptamer sequences and binding properties of immunoglobulin E (IgE)-binding aptamers (7,8). The application of high density microarray chips to the aptamer selection process has the potential to speed up the generation of aptamers with high affinity and specificity.

One of the main problems in the application of microarray technology to the selection of RNA aptamers is the design of the initial pool of RNA molecules for screening. This requires one to pre-select 10^4–10^5 RNA sequences for the microarray chip from a set of 10^{13}–10^{14} possible sequences. In this article we propose an *in silico* approach to create a list of RNA sequences with potential binding affinity to a desired small molecule. Our approach consists of three steps:

(1) **Step 1** *Selection of RNA sequences based on their secondary structure*. The analysis of randomly generated RNA pools shows that the majority of sequences have simple stem-loop or slightly branched structures, while the more complex structures are very rare (9–11). Furthermore, high-affinity aptamers are thermodynamically different from the random sequences. It was found that the free energies of the secondary structure formation of GTP aptamers are significantly lower than the same-length random sequences (12). Based on these findings, we developed a set of criteria that limited the presence of sequences with abundant

*To whom correspondence should be addressed. Tel: +1 937 904 9542; Fax: +1 937 255 1474; Email: yaroslav.chushak@wpafb.af.mil

Published by Oxford University Press 2009
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

simple structural motifs and maximized the presence of stable low-energy structures. These criteria selected approximately one RNA sequence from about 2500 random sequences. Only the sequences that passed the selection criteria were forwarded to the next step.

(2) **Step 2** *Generation of 3D structures*. Computational prediction of RNA tertiary structure is a very intensive field of research. For example, during the past year, three different approaches for *ab initio* RNA structure prediction have been proposed (13–15). We used the Rosetta package (13) developed by the Baker group at the University of Washington to predict three-dimensional (3D) structures of selected sequences. The fragment assembly of RNA molecule in Rosetta is based on the simplified energy function that takes into account the backbone conformational and side-chain interaction preferences observed in the experimental RNA structures (13). We developed a protocol that includes minimization of Rosetta-generated structures using the AMBER force field (16) and generalized Born implicit solvent (17). It is widely accepted that ligand binding can drastically alter the receptor's conformation (18). To account for such conformational flexibility, the five lowest energy structures for each sequence were placed into a library of RNA molecules to perform ensemble docking.

(3) **Step 3** *Screening the library of RNA molecules*. Computational docking is a common tool used to identify small-molecule ligands that bind to proteins (19). While most docking methods have been developed for proteins, recent evaluation of AutoDock and DOCK programs has demonstrated their ability to dock compounds to RNA molecules (20). Docking tools are usually used to screen a library of small molecules in order to find a ligand that binds to a specific protein or RNA receptor. In our approach, we screened the library of RNA molecules to find receptors with the highest binding affinity to a desired small molecule. We used a modified version of the DOVIS package (21) for high-throughput virtual screening of the entire RNA library. DOVIS uses AutoDock4 software (22) as the docking engine and runs in parallel on Linux clusters. The original version of DOVIS divides the library of small molecules between multiple processors. We modified the parallelization scheme by dividing the library of receptors (RNA molecules in our case) between the multiple processors for parallel receptor-ligand screening.

In the proposed *in silico* approach, RNA sequences are screened at two levels (Figure 1). At the first level, selection of RNA sequences is based on analysis of secondary structure of the generated sequences. At this screening level, the selected sequences are not target specific. At the second screening level, we used computational docking to identify RNA molecules that bind to a specific target ligand. At this point, the selected RNA molecules are specific to the desired target molecule and they are placed into a pool of sequences for experimental verification and selection of high-affinity aptamers. The developed computational approach allows one to pre-select sequences from the initial pool of RNA molecules and effectively leads to a reduction in sequence search space by four to five orders of magnitude.

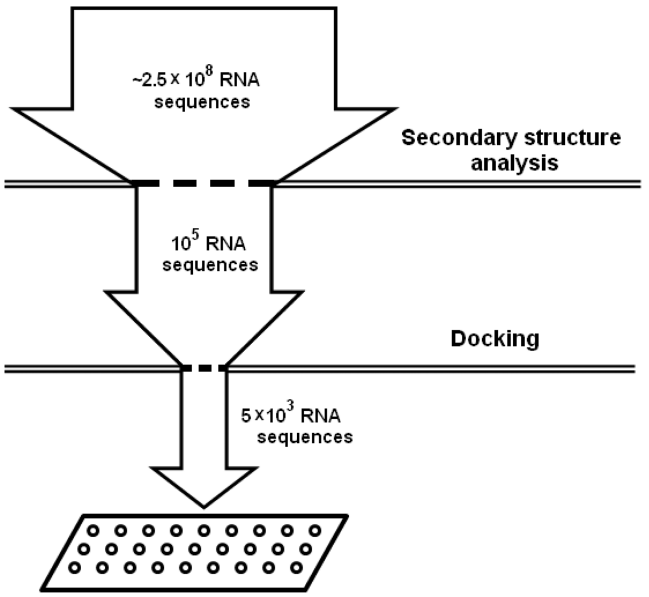

Figure 1. Reduction in size of the RNA sequence space for experimental screening and selection of RNA aptamers by *in silico* approach. The secondary structure of more than 2.5×10^8 RNA sequences was analyzed to select 100 000 sequences for the RNA 3D structure library. The high-throughput virtual screening of the developed library selected 10^3–10^4 sequences suitable for the experimental screening and verification.

MATERIALS AND METHODS

Random RNA sequences of a given length were generated by randomly selecting one of the four RNA bases for every position. The lowest energy secondary structure of a generated sequence was evaluated by the Vienna RNA package (23) using a default set of parameters and prohibiting isolated base pairs by setting noLonelyPairs = 1. The isolated base pairs are usually unstable with respect to thermodynamic perturbations and can lead to significant conformational changes upon ligand binding. Since the folding algorithm implemented in Vienna RNA package cannot predict the formation of structures with pseudoknots, such structures were excluded from the consideration.

For each sequence length, we calculated a mean value of the free energy ($\overline{E}$) and a standard deviation (SD) using a set of 10 000 random sequences. To compare the free energy of an aptamer (E_{apt}) with the mean free energy of the same-length random sequences, we calculated Z-score using a standard equation (24): $Z = (E_{apt} - \overline{E})/SD$.

The pool of 27-mer sequences with a randomized region of 21 nucleotides was generated by applying the following constraints: (i) in the lowest energy conformation bases 1-2-3 form pairs with bases 27-26-25; (ii) the free energy

of secondary structure formation is lower than −5.7 kcal/mol corresponding to *Z*-score equal −1.0; (iii) there are at least 11 bases that do not form Watson–Crick base pairs, i.e., they form a loop or bulge and (iv) the number of the same structural motifs (e.g. stem-loop structure with a stem containing eight base pairs) is limited to 150. More than 2.5×10^8 sequences were screened to generate a pool of 100 000 sequences that satisfied criteria (i–iv). The generation of random RNA sequences and analysis of their secondary structure are the least computationally intensive part of the approach. It took around 4 h on a single Intel P4 3.06 GHz CPU to screen $>10^8$ sequences and to generate a library.

For each of the sequences in the pool, two files were created: a sequence file in FASTA format and a pairing file that contained information about the Watson–Crick base pairing. These two files were used as input files by the Rosetta package (13) for tertiary structure generation. The generated structures were minimized using the AMBER10 package (http://www.ambermd.org).

The automated 3D structure prediction for each sequence in the pool involved the following steps:

(1) Generate 500 decoys using the Rosetta++ *−prna* function with the FASTA-type sequence file and a file with Watson–Crick base pairing as input.
(2) Score generated decoys using Rosetta energy functions and select 100 best decoys.
(3) Create all-atoms files in PDB format for the selected decoys using Rosetta++ *−extract* function.
(4) Prepare AMBER10 input files for each of the PDB files using *tleap* program.
(5) Run energy minimization using the *sander* program and AMBER99 force field for 1000 steps in implicit solvent using generalized Born model (*igb* = 1) with a cut-off value of 16.0 Å.
(6) Rank minimized structures based on the final value of their energy.
(7) Convert five of the best structures into PDB format using *ambpdb* program.
(8) Use the *OpenBabel* program (http://www.openbabel.org) to convert PDB format file into Mol2 file format with the molecule centre at (0, 0, 0) and to assign Gasteiger atomic charges.
(9) Put five Mol2 files for each sequence into a library of RNA structures.

The generation of a library of tertiary structures demands significant computational resources. It takes about 4 h to generate 3D structures for a single sequence or 400 000 CPU hours for a library of 100 000 sequences. We used computational resources at the Air Force Research Laboratory DoD Supercomputing Resource Center. By using 200–300 processors on HP XC Opteron supercomputer, we completed the generation of a library of RNA structures in 4 months.

We performed a validation of our computational approach on a set of six known aptamer–ligand complexes. The 3D structures for these aptamers were generated as described above and placed into a library together with randomly generated 27-mer structures to bring the library size up to 5000 RNA structures. The commercial software SYBYL 7.0 (Tripos Inc., St. Louis, MO) was used to create ligand molecules. For three molecules (theophylline, gentamicin and flavin mononucleotide), coordinates were taken from the experimental PDB files (1EHT, 1BYJ and 1FMN), and Gasteiger charges were added by SYBYL. For the three other molecules (codeine, guanine and isoleucine), coordinates were generated by SYBYL's Concord tool using the SMILES description of the molecule and then optimized by running energy minimization for 1000 steps with the Tripos force field and Gasteiger atomic charges.

The modified version of the DOVIS 2.0 package (25) was used to screen the generated library of RNA structures for RNA–ligand binding. DOVIS 2.0 uses the recently released docking package AutoDock 4.0 (http://www.autodock.scripps.edu) as the docking engine and runs in parallel on Linux clusters. The original version of DOVIS divides the library of small molecules between multiple processors. We modified the parallelization scheme to divide the library of RNA structures between the multiple processors for parallel RNA–ligand screening. The receptor and ligand files were prepared for docking using Python scripts of AutoDockTools (http://www.autodock.scripps.edu/resources/adt/index.html). The non-polar hydrogen atoms on RNA molecules were merged before the docking and Gasteiger charges were assigned by *OpenBabel*. The grid box was centred at (0, 0, 0) with 60 points in each dimension and the default value of 0.375 Å for spacing between the grid points. For each of the RNA–ligand complexes, 20 docking experiments were performed using the Lamarckian genetic algorithm conformational search, with the population size of 150, one million energy evaluations, and a maximum of 27 000 generations per run. The docking results were scored with AutoDock 4.0 scoring function. Thus, screening a library of RNA structures for RNA–ligand binding is not as computationally demanding as generating a library. It required around 8000 CPU hours to screen 500 000 RNA structures to select aptamers to a desired small molecule. Computational docking was performed using 50–100 processors on Linux Networx Evolocity II supercomputer at the US Army Research Laboratory DoD Supercomputing Resource Center.

RESULTS

For our study, we selected 27-mer RNA molecules 5′–GGC–N21–GUC-3′ with a central random region of 21 nucleotides. The constant sequences at the 5′-end and 3′-end correspond to sequences in the well-known theophylline aptamer (26). The 21-nt randomized region potentially contains 4^{21} or about 4×10^{12} different sequences. It is impossible to span the entire space of all the possible sequences; furthermore, the vast majority of sequences do not possess the potential ability for high-affinity binding. Therefore, our aim is to select sequences that have the potential to bind ligands with high affinity and selectivity.

Table 1. The free energy of secondary structure formation for RNA aptamers that bind different ligands comparing with the mean free energy of the same-length random sequences

Ligand (reference)	Aptamer length (bases)	Mean free energy (kcal/mol)	SD (kcal/mol)	Aptamer free energy (kcal/mol)	Aptamer *Z*-score
ATP (31)	40	−6.54	3.23	−17.7	−3.5
Codeine (32)	34	−4.96	2.90	−7.60	−0.9
Flavin (33)	35	−5.27	3.00	−18.0	−4.2
Gentamicin (34)	27	−3.18	2.47	−13.9	−4.3
Guanine (35)	32	−4.39	2.78	−13.7	−3.4
Isoleucine (36)	27	−3.18	2.47	−7.30	−1.7
Neomycin[a] (37)	23	−2.17	2.10	−11.0	−4.2
Neomycin[b] (27)	31	−4.23	2.70	−7.50	−1.2
Theophylline (26)	33	−4.73	2.80	−11.6	−2.5
Tobramycin (28)	26	−2.90	2.35	−8.70	−2.5

[a]Non-functional neomycin aptamer.
[b]Neomycin aptamer that can be integrated into riboswitch to regulate gene expression.

Thermodynamics of RNA secondary structure

Analysis of experimentally selected GTP aptamers shows a significant correlation between the dissociation constant K_d and the free energy of secondary structure formation (12). It was found that the free energy of aptamers is significantly lower than the median same-length random sequence value. We applied a similar analysis to a set of 10 aptamers specific to different small molecules by calculating *Z*-score (Table 1). To calculate the mean value of free energy and SD for the same-length random sequences, we generated 10 000 random sequences. Results presented in Table 1 show that aptamers have *Z*-score values in the range of −0.9 to −4.3. Based on these results, we chose a free-energy cut-off value of −5.7 kcal/mol corresponding to $Z = -1$ of random 27-mer sequences used in our study. Although, in several cases, aptamer sequence free energy has *Z*-score of about −4, we based our choice on the results for neomycin aptamers. The neomycin aptamer (a) with 23 bases (see Table 1) has a free energy of secondary structure of −11 kcal/mol corresponding to $Z = -4.2$. However, no *in vivo* gene regulatory activity was observed using this aptamer (27). The second neomycin aptamer (b) with 31 bases has a free energy of secondary structure of −7.5 kcal/mol and $Z = -1.24$. Interestingly, the binding of ligand changes the conformation of this aptamer and aptamer (b) has a gene regulatory activity in the presence of neomycin (27). We speculate that the possible difference in the regulatory activity of these two aptamers is caused by the secondary structure of aptamer (a), with the extremely low free energy 'locking' the aptamer in the low energy configuration and preventing it from changing conformation upon ligand binding. Since we are interested in functional RNA aptamers, we selected a higher free energy cut-off value.

Watson–Crick base pairs and ligand binding

Detailed analysis of experimental 3D structures for a number of aptamer-ligand complexes provides important information about the molecular recognition and interaction of nucleic acids with ligands (28). We analyzed molecular recognition for several aptamer–ligand complexes (Figure 2). The aptamer nucleotide bases that form a ligand-binding pocket are blue-circled in Figure 2 while the bases that directly participate in ligand recognition by forming hydrogen bonds with the ligand are red-circled. It can be clearly seen that RNA bases involved in molecular recognition do not form Watson–Crick pairs with other bases. A similar conclusion can be drawn from Figure 1 in paper by Carothers *et al.* (29) that shows secondary structures for 11 classes of GTP aptamers. The bases with high informational content, which is important for the high-affinity binding, are always unpaired and located in loops or bulges. There are two possible reasons for this: firstly, unpaired RNA bases are more flexible, so they can easily change their conformation to form a binding pocket and accommodate a ligand, and secondly; unpaired bases have available donor or acceptor atoms for potential formation of hydrogen bonds with the ligand. Therefore, we set a constraint that the secondary structure of our sequences with 27 bases should have at least 11 unpaired bases. This number seems optimal for us, since the higher number of unpaired bases will significantly reduce the presence of sequences with high free energy of the secondary structure while the lower number will increase the presence of sequences with low binding affinity.

Distribution of RNA structural motifs

Initially, we started with $\sim 5.8 \times 10^6$ 27-mer sequences, and these sequences were screened to select 10^5 sequences that met the constraints defined above. These sequences folded into 725 different secondary structures. The frequency distribution of structural motifs is extremely heterogeneous (Figure 3). The stem-loop structure with 8 base pair stem was observed more than 6700 times; 4 structures were present more than 2000 times, while 18 other structural motifs were observed more than 1000 times. On the other hand, almost 140 secondary structures were presented only once or twice.

To increase the diversity of the generated pool of sequences, we imposed an additional constraint by limiting to 150 the number of times each structural motif appeared in the pool. Now, to generate the same

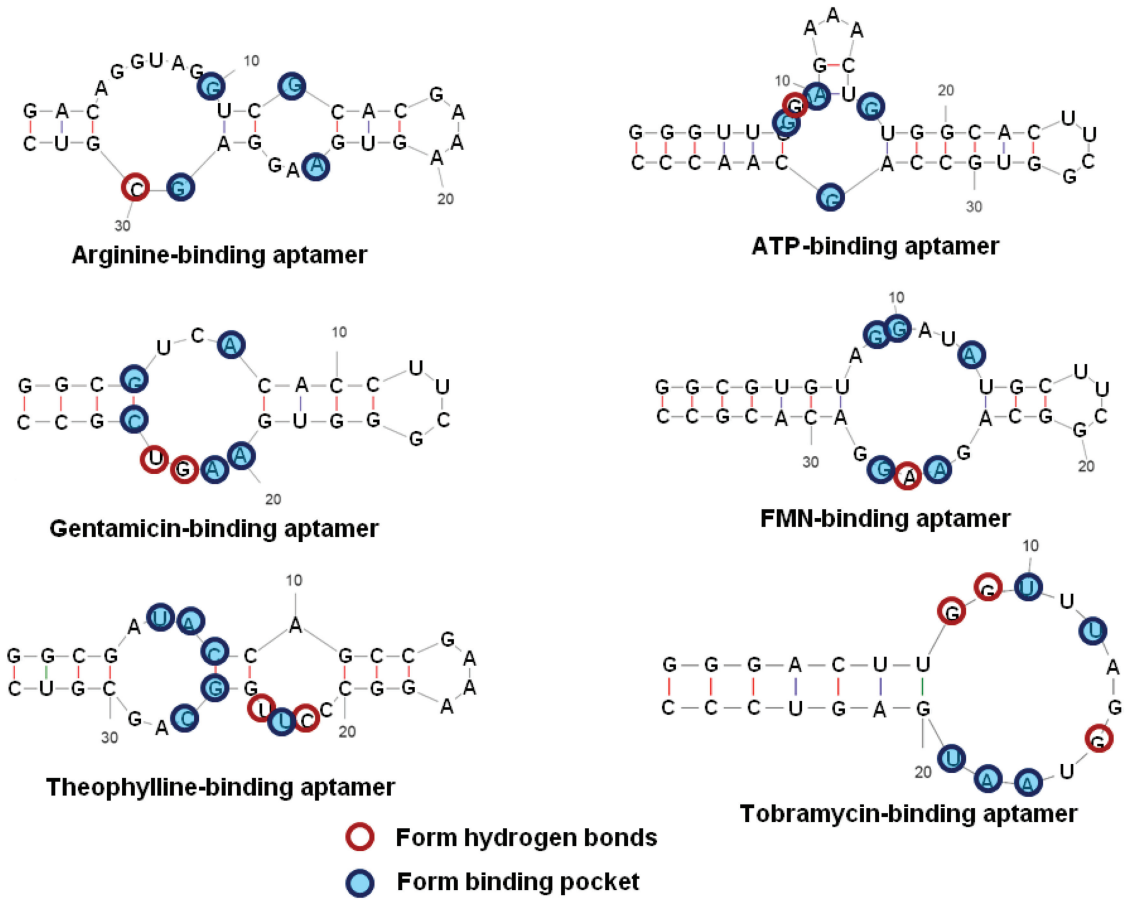

Figure 2. Secondary structure of RNA aptamers for different small-molecule ligands. Blue-circled bases participate in the formation of ligand-binding pocket and red-circled bases form hydrogen bonds with the ligand.

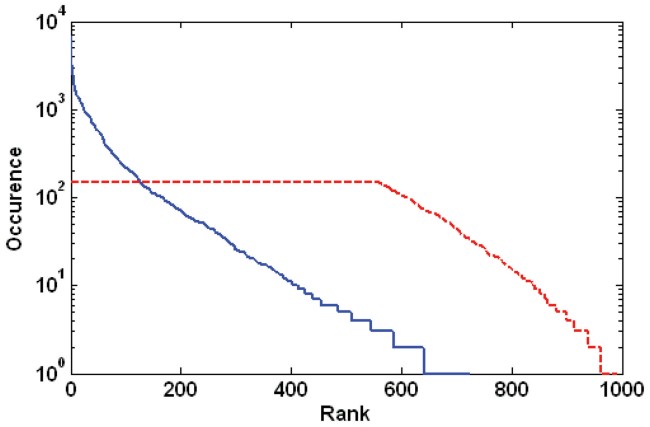

Figure 3. Distribution of structural motifs in 10^5 random RNA sequences. The solid blue line represents sequences with two constraints: the free energy of secondary structure less than −5.7 kcal/mol and the number of unpaired bases is at least 11. The sequences are folded into 725 different secondary structures. The dashed red line shows a distribution when the number of the same structural motifs is limited to 150. In this case, the pool of 100 000 sequences contains 997 different structural motifs.

number of 10^5 sequences, we screened more than 2.5×10^8 random sequences. The newly generated pool contained 997 different structural motifs, corresponding to a more than 35% increase compared with the previous case (Figure 3). In the new pool of sequences, almost 560 motifs appeared 150 times, covering 84% of all sequences. These 100 000 sequences were selected to be included in the library of 3D RNA structures. Setting the number of structural motifs repeats to a lower value, e.g. to 100, will require a significantly higher number of available different structural motifs. On the other hand, the number of possible structural motifs for the sequences with 21 randomized bases and constraints (i–ii) is limited. For that reason, we were not able to generate 100 000 sequences by limiting the number of abundant structures to 100.

RNA 3D structure prediction

The tertiary structure of RNA molecules was predicted using the Rosetta package (13). One of the main problems in the 3D structure prediction is to select a native-like structure from hundreds or thousands of models generated by Rosetta. The Rosetta's simplified energy function takes into account backbone conformational and side-chain interaction preferences, but does not always correctly predict the native-like structure (13). We developed a protocol that includes energy minimization of Rosetta-generated structures using AMBER99 force field (16) and generalized Born implicit solvent (17) (see 'Materials and Methods' for details). To validate the proposed protocol, we compared the predicted 3D structures with a native

Table 2. The average backbone RMSD of the five lowest energy-predicted structures from the native structure

PDB	N_{bases}	Rosetta Score[a] RMSD (Å)	Minimization[b] RMSD (Å)
1BYJ	27	5.2 (1.2)	3.8 (1.3)
1EHT	33	7.4 (0.9)	6.9 (0.6)
1ESY	19	5.6 (0.3)	4.9 (0.6)
1KKA	17	6.9 (1.1)	6.2 (0.7)
1Q9A	27	6.5 (0.4)	6.2 (0.4)
1QWA	21	7.3 (0.7)	6.1 (0.8)
28SP	28	4.2 (0.9)	3.7 (0.6)
2F88	34	6.4 (0.6)	5.8 (0.4)
2TOB	20	5.7 (0.6)	5.1 (0.5)

For NMR models, the first model was designated as the reference structure. Values in parentheses are standard deviations.
[a]Predicted structures were ranked using Rosetta scoring function (13).
[b]Predicted structures were minimized using AMBER force field and ranked based on their final energy.

Table 3. Ranking and binding energy of the native RNA–ligand complex from a pool of 5000 generated 3D structures based on docking procedure

Ligand	Native rank	Predicted ΔG_b (kcal/mol)	Expt. ΔG_b (kcal/mol)	Expt. K_d (μM)
Codeine	216	−9.64	−7.62	2.56
FMN	116	−7.67	−8.59	0.50
Gentamicin	127	−11.4	−10.9	0.01
Guanine	131	−7.87	−7.83	1.80
Isoleucine	57	−4.24	−3.98	1200
Theophylline	102	−4.91	−8.72	0.40

The RNA structures were ranked based on the binding affinity to a small molecule ligand. The native aptamer structure typically is ranked among the top 5% of the best structures.

structure for nine single-chain RNA molecules for which detailed structural information exists (Table 2). For every experimental RNA structure, we generated 500 candidate structures and ranked them using the Rosetta-scoring function. The 100 highest ranking structures were selected for energy minimization, and they were ranked based on the final energy value. The backbone root mean square deviation (RMSD) was calculated for the five lowest energy structures predicted by the Rosetta scoring function and for the five lowest energy structures predicted by the energy minimization technique. For all the tested structures, energy minimization improves the accuracy of the predicted models by reducing RMSD 10–20%. Furthermore, it was found that energy minimization also refined the Rosetta-generated structures by fixing some bond lengths that appeared to be too long. In four cases (1Q9A, 1QWA, 28SP and 2F88), the same model (one out of five) was selected by both Rosetta-scoring function and energy minimization.

The developed protocol that included energy minimization was used to create a library of RNA structures for 100 000 of the 27-mer-generated sequences. The final structured library contained ~500 000 RNA 3D structures with five structures per sequence. By including five 3D structures for every sequence, we were able to perform ensemble docking and, thus, to account for conformational flexibility of the RNA molecules. This library was used for virtual high-throughput screening to select aptamers with binding affinity to a specified small molecule.

Validation via docking known RNA aptamer sequences to their ligands

To test the performance of our computational approach for selection of RNA aptamers, a small pool of 1000 sequences was created. This pool contained six sequences of known aptamers that bind to small molecule ligands in addition to the 994 randomly generated 27-mer sequences. For each of the 1000 sequences, we generated 3D structures using the above protocol and placed five of the lowest-energy structures into the library. In total, our library of RNA molecules contained 5000 generated structures. We used the modified version of DOVIS to screen this library of RNA molecules against each of the ligands in the test set. The results of database docking are summarized in Table 3. Ideally, the structure of the native sequence should be ranked among the structures with the highest binding affinity. It was found that for all the tested 3D RNA–ligand complexes, the native aptamer structure ranked among the top 5% of the best structures. For the five tested ligands, the predicted binding energy for the native structure was within 30% of the experimental value. However, for one small molecule, theophylline, the predicted binding energy is almost two times lower than the experimental value. The possible reason for such disagreement can be the changing of aptamer conformation upon ligand binding. Experimental measurements on the structure of the aptamer–theophylline complex using nuclear magnetic resonance (NMR) spectroscopy revealed significant rearrangements of residues in the internal loop of the aptamer induced by binding of theophylline (26). Furthermore, the high-affinity theophylline binding was observed only in the presence of divalent metal ions (Mn^{2+}, Co^{2+} or Mg^{2+}). In the absence of metals ions, binding affinity of this aptamer was reduced by $\sim 10^4$ (30). In our docking experiments, RNA molecules were considered as rigid and no metal ions were added to the system; these two factors can significantly reduce the binding energy for theophylline binding. However, since the computational database docking is complemented by experimental high-throughput screening, it implies that prioritizing of selected sequences is more important than the accuracy of predicted binding energy.

As an example of the performance of developed *in silico* approach, we presented in Figure 4 some details of aptamer selection for gentamicin molecule. To test the performance of the Autodock4 package, we initially performed docking of gentamicin to the experimental NMR structure of gentamicin-binding aptamer (PDB code 1byj). Figure 4a shows a superposition of the experimental (red) and the best docking ligand pose (yellow). The RMSD between the docked and experimental configuration of ligand was 1.27 Å. The predicted free energy of binding was −9.46 kcal/mol, which is in a good agreement

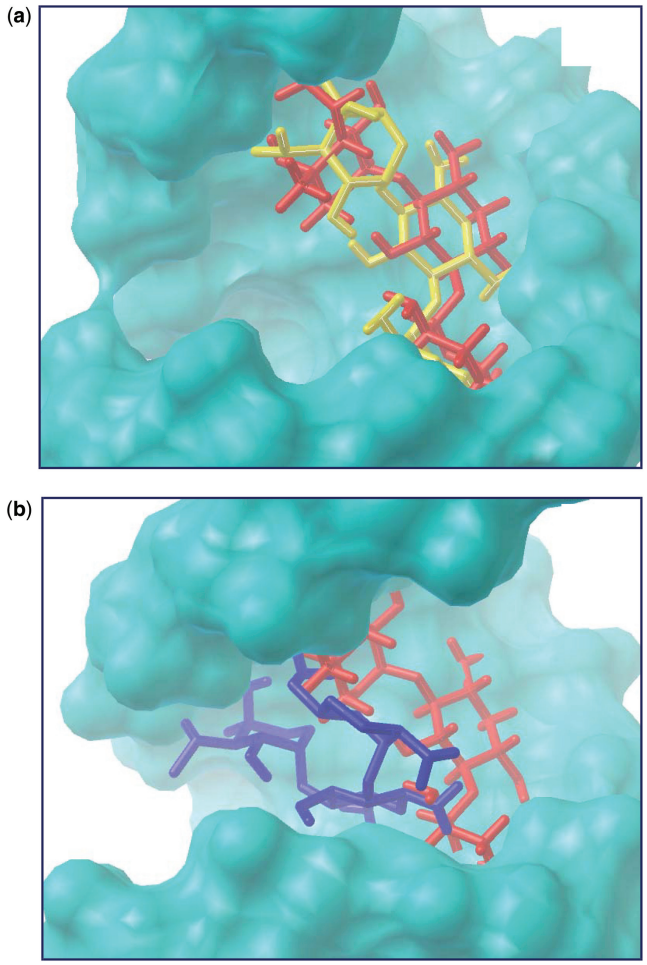

Figure 4. Comparison of predicted docking structures with the experimental results for gentamicin-binding aptamer. (**a**) The predicted docking pose with the highest score of gentamicin (yellow), and the experimental pose (red) inside the experimental NMR structure of RNA aptamer. The first NMR model was taken as the reference structure. (**b**) The predicted docking pose of gentamicin (blue) and the experimental configuration (red) inside the predicted structure of RNA aptamer. The predicted RNA structure was aligned with the experimental using backbone atoms. The experimental pose of gentamicin molecule was fixed in the same position as in the experimental aptamer.

with the experimental value of −10.91 kcal/mol. In the next step, we analyzed the gentamicin-docking conformations from the database-docking experiments described above where the computer-generated structures of aptamers were used. In Figure 4b, we show the best pose of gentamicin molecule (blue) inside the predicted structure of gentamicin-binding aptamer. The predicted 3D structure of RNA aptamer was aligned with the experimental structure using the backbone atoms. We also presented in Figure 4b the experimental configuration of gentamicin (red) inside the experimental-binding pocket. Clearly, the docking pose of ligand is in close proximity to the experimental ligand position and the deviation of the predicted ligand configuration from the experimental was 3.6 Å. The free energy of binding to the predicted RNA structure was −11.43 kcal/mol that is also in a good agreement with the experimental value. We found these results very encouraging taking into account that aptamer's 3D configuration was generated computationally.

CONCLUSIONS

Development of microarray technology for RNA aptamer selection requires one to pre-select 10^4–10^5 RNA sequences for the microarray chip from a set of $\sim 10^{12}$ possible sequences. We have developed a computational approach to reduce the search space of RNA molecules and to create a pool of sequences with potential binding affinity to a desired target ligand. This pool of sequences can be used for experimental screening and selection of high-affinity aptamers using microarrays or other techniques.

The RNA sequences were screened at two levels. In the first level, we analyzed the secondary structure of the generated sequences. On the basis of the secondary structure analysis of known RNA–ligand complexes, we developed a set of criteria that allowed us to select sequences with the potential to bind ligands with high affinity from a pool of random sequences. These criteria were: (i) the free energy of secondary structure formation has *Z*-score lower than −1; (ii) there are at least 11 bases that do not form Watson–Crick pairs and (iii) the number of abundant structures is limited to 150 to increase the diversity of structural motifs in the pool. The applied criteria selected approximately one RNA sequence in 2500 random sequences.

For sequences that passed our selection criteria, we generated 3D structures. We developed a protocol to automate the generation of 3D structures that includes the generation of decoys using the Rosetta package and minimization of low-energy structures using the AMBER force field. The five lowest energy structures for each sequence were placed into a library of RNA molecules, thus allowing us to perform ensemble docking and to account for conformational flexibility in the RNA molecules. At this screening level, the selected sequences were not target specific.

In the second screening level, we used computational docking to identify RNA molecules that bind to a specific target ligand. The high-throughput screening of the developed RNA structure library was performed using the modified version of the DOVIS package. At this point, the selected RNA molecules were specific to the desired target molecule and they were placed into a pool of sequences for experimental screening and selection of high-affinity aptamers.

We validated the proposed computational approach using a set of six known aptamer–ligand complexes. The small library containing 3D structures for six aptamer sequences and 994 randomly generated 27-mer sequences was screened against the ligands from the testing set, and structures were ranked based on their binding affinity. It was found that the structures of the known native aptamer sequences were in the top 5% of the best structures demonstrating the remarkable performance of our method in the selection of potential receptors to small molecule ligands.

The question may be raised concerning the sensitivity of the results of selection on the set of RNA sequences placed into the screening library. Although, we have screened more than 2.5×10^8 RNA sequences to generate a pool of 100 000 of the 27-mer sequences, that number is still four orders of magnitude lower than $\sim 10^{12}$ possible sequences. The set of sequences selected to the pool also depends on the applied selection criteria. Furthermore, since RNA sequences are randomly generated, even for the same set of selection parameter we can get a different set of sequences in the library. However, the experimental screening of aptamers typically uses several round of selection with mutated sequences at each round. Since the RNA sequences placed into a microarray chip are known, it is possible to develop a sequence-fitness landscape and to design aptamers with desired binding affinity (38). Therefore, the high-affinity aptamers can be picked-up during the next rounds of experimental selections even if they were not present in the initial pool of sequences. We developed a computational approach that allows experimentalists to design the initial set of RNA sequences with potential binding affinity to a desired target ligand.

In conclusion, our proposed approach reduces the search space of RNA sequences by four to five orders of magnitude. We anticipate that our approach could be used to create the initial pool of RNA sequences for experimental selection of high-affinity aptamers, greatly accelerating the process of finding the desired aptamer sequence.

ACKNOWLEDGEMENTS

We thank Dr Wanda Lyon from the Air Force Research Laboratory for insights on using microarrays for RNA aptamers selection. We also are thankful to Drs A. Wallqvist, M. Lee and X. Jiang from the Biotechnology HPC Software Applications Institute for useful discussions about the 3D structure prediction and using DOVIS package for RNA–ligand docking.

FUNDING

Air Force Office of Scientific Research (AFOSR), the Defense Advanced Research Projects Agency (DARPA) and the US Department of Defense High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) initiative. Funding for open access charge: US Government contract FA8650-05-2-6518.

Conflict of interest statement. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army, U.S. Air Force, or the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

1. Breaker,R.R. (2004) Natural and engineered nucleic acids as tools to explore biology. *Nature*, **432**, 838–845.
2. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
3. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
4. Osborne,S.E. and Ellington,A.D. (1997) Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.*, **97**, 349–370.
5. Collett,J.R., Cho,E.J., Lee,J.F., Levy,M., Hood,A.J., Wan,C. and Ellington,A.D. (2005) Functional RNA microarrays for high-throughput screening of antiprotein aptamers. *Anal. Biochem.*, **338**, 113–123.
6. Li,Y., Lee,H.L. and Corn,R.M. (2006) Fabrication and characterization of RNA aptamer microarrays for the study of protein-aptamer interactions with SPR imaging. *Nucleic Acids Res.*, **34**, 6416–6424.
7. Katilius,E., Flores,C. and Woodbury,N.W. (2007) Exploring the sequence space of a DNA aptamer using microarrays. *Nucleic Acids Res.*, **35**, 7626–7635.
8. Fischer,N.O., Tok,J.B.-H. and Tarasow,T.M. (2008) Massively parallel interrogation of aptamer sequence, structure and function. *PLoS ONE*, **3**, e2720.
9. Sabeti,P.C., Unrau,P.J. and Bartel,D.P. (1997) Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. *Chem. Biol.*, **4**, 767–774.
10. Gevertz,J., Gan,H.H. and Schlick,T. (2005) In vitro RNA random pools are not structurally diverse: a computational analysis. *RNA*, **11**, 853–863.
11. Stich,M., Briones,C. and Manrubia,S.C. (2008) On the structural repertoire of pools of short, random RNA sequences. *J. Theor. Biol.*, **252**, 750–763.
12. Carothers,J.M., Oestreich,S.C. and Szostak,J.W. (2006) Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J. Am. Chem. Soc.*, **128**, 7929–7937.
13. Das,R. and Baker,D. (2007) Automated *de novo* prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.
14. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 52–55.
15. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
16. Case,D.A., Cheatham,T., Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R. (2005) The AMBER biomolecular simulation programs. *J. Computat. Chem.*, **26**, 1668–1688.
17. Tsui,V. and Case,D.A. (2001) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers (Nucleic Acid Sciences)*, **56**, 275–291.
18. May,A., Sieker,F. and Zacharias,M. (2008) How to efficiently include receptor flexibility during computational docking. *Curr. Compu.-Aided Drug Des.*, **4**, 143–153.
19. Taylor,R.D., Jewsbury,P.J. and Essex,J.W. (2002) A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.*, **16**, 151–166.
20. Detering,C. and Varani,G. (2004) Validation of automated docking programs for docking and database screening against RNA drug targets. *J. Med. Chem.*, **47**, 4188–4201.
21. Zhang,S., Kumar,K., Jing,X., Wallqvist,A. and Reifman,J. (2008) DOVIS: an implementation for high-throughput virtual screening using AutoDock. *BMC Bioinformatics*, **9**, 126.
22. Huey,R., Morris,G.M., Olson,A.J. and Goodsell,D.S. (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, **28**, 1145–1152.
23. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
24. Olson,C.L. (1987) *Essentials of Statistics: Making Sense of Data*. Allyn and Bacon Inc., Newton, MA.
25. Jiang,X., Kumar,K., Hu,X., Wallqvist,A. and Reifman,J. (2008) DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0. *Chem. Cent. J.*, **2**, 18.

26. Zimmermann,G.R., Jenison,R.D., Wick,C.L., Simorre,J.-P. and Pardi,A. (1997) Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. *Nature Struct. Biol.*, **4**, 644–649.
27. Weigand,J.E., Sanchez,M., Gunnesch,E.-B., Seiher,S., Schroeder,R. and Suess,B. (2008) Screening for engineered neomycin riboswitches that control translation initiation. *RNA*, **14**, 89–97.
28. Patel,D. J. and Suri,A. K. (2000) Structure, recognition and discrimination in RNA aptamer complexes with cofactors, amino acids, drugs and aminoglycoside antibiotics. *Rev. Mol. Biotechnol.*, **74**, 39–60.
29. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–5135.
30. Zimmermann,G.R., Wick,C.L., Shields,T.P., Jenison,R.D. and Pardi,A. (2000) Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. *RNA*, **6**, 659–667.
31. Sassanfar,M. and Szostak,J.W. (1993) An RNA motif that binds ATP. *Nature*, **364**, 550–553.
32. Win,M.N., Klein,J.S. and Smolke,C.D. (2006) Codeine-binding RNA aptamers and rapid determination of their binding constants using a direct coupling surface plasmon resonance assay. *Nucleic Acids Res.*, **34**, 5670–5682.
33. Burgstaller,P. and Famulok,M. (1994) Isolation of RNA aptamers for biological cofactors by in vitro selection. *Angew. Chem. Int. Ed. Engl.*, **33**, 1084–1087.
34. Yoshizawa,S., Fourmy,D. and Puglisi,J.D. (1998) Structural origins of gentamicin antibiotic action. *EMBO J.*, **17**, 6437–6448.
35. Kiga,D., Futamura,Y., Sakamoto,K. and Yokoyama,S. (1998) An RNA aptamer to the xanthine/guanine base with a distinctive mode of purine recognition. *Nucleic Acids Res.*, **26**, 1755–1760.
36. Lozupone,C., Changayil,S., Majerfeld,I. and Yarus,M. (2003) Selection of the simplest RNA that binds isoleucine. *RNA*, **9**, 1315–1322.
37. Jiang,L., Majumdar,A., Hu,W., Jaishree,T.J., Xu,W. and Patel,D.J. (1999) Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure*, **7**, 817–827.
38. Knight,C.G, Platt,M., Rowe,W., Wedge,D.C., Khan,F., Day,P.J.R., McShea,A., Knowles,, , and Kell,D.B. (2009) Array-based evolution of DNA aptamers allows modeling of an explicit sequence-fitness landscape. *Nucleic Acids Res.*, **37**, e6.