Supplementary Material for

Exploring the mutational robustness of nucleic acids by searching genotype neighborhoods in sequence space

Qingtong Zhou,[†] Xianbao Sun,[‡] Xiaole Xia,[¶] Zhou Fan,[†] Zhaofeng Luo,[§] Suwen Zhao,[†] Eugene Shakhnovich,[*, ‖] and Haojun Liang[*,‡]

[†] *iHuman Institute, ShanghaiTech University, Pudong New District, Shanghai 201210, China.*

[‡] *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, University of Science and Technology of China, Hefei, Anhui 230026, China*

[¶] *Key Laboratory of Industrial Biotechnology, Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi, Jiangsu 214122, China.*

[§] *Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China.*

[‖] *School of Life Science, University of Science and Technology of China, Hefei, Anhui 230026, China.*

[⊥] *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA*

[*] To whom correspondence should be addressed: Email: shakhnovich@chemistry.harvard.edu, or hjliang@ustc.edu.cn

Supplementary Text S1: Computational Methods

1.1 Secondary Structure Prediction

The secondary structure of nucleic acid was predicted by NUPACK[1] using nearest-neighbor empirical parameter **dna1998** for DNA. The temperature was set to 25℃. As demonstrated by previous research[2,3], the L-Arm binding 1OLD aptamer (5'-GATCGAAACGTAGCGCCTTCGATC-3') have specified secondary structure (shown as target structure in Figure S2). By fully mutated nine bases on non-canonical region of 1OLD aptamer (base 8, 10, 11, 12, 13, 14, 15, 16, 17), we prepared an **aptamer neighborhood sequence library**, which is consisting by 262,144 sequences. For each sequence, minimum free energy (MFE) secondary structure and its corresponding free energy ΔG_{MFE} were predicted. Besides, the free energy of nucleic acid at target motif state was calculated, which was defined as ΔG_{target} (Figure S2). Finally the free-energy gap (FEG) between the lowest secondary structure energy state and the target secondary structure state $\Delta\Delta G_{gap}$ for a current sequence, defined as $\Delta G_{target}-\Delta G_{MFE}$, was obtained.

1.2 Molecular Dynamics Simulation

The parmbsc0 Amber,[4] TIP3P,[5] and Åqvist's force field[6] were used for DNA, water, and ions, respectively. The ligand L-Arm was geometry optimized by Gaussian 09[7] with a level of HF/6-31G*.The L-Arm-DNA complex was placed in a cubic box that the boundary of the box were at least 15 Å to the solute. To neutralize the charge of the system, three Cl^- ions and 24 sodium ions were introduced into the complex. Long-range electrostatic interactions were calculated by particle mesh Ewald method[8]. Van der Waals interactions and real space Coulomb interactions were calculated with a cutoff of 10 Å. The LINCS algorithm was used to constrain bonds involving hydrogen atoms.[9] Neighbor lists were updated every 5 time steps. Firstly, the entire system was energy minimized using five rounds of steepest descent, then it was heated to a final temperature of 300 K with restraints. The restraints were reduced gradually, with a simulation step of 1 fs. Finally, the time step was increase to 2fs and the system was run without restraints in the NPT ensemble at 300 K and 1 bar using a v-rescale thermostat[10] and Berendsen barostat[11]. All molecular dynamic simulation was performed by GROningen MAchine for Chemical Simulations v 4.6.5.[12]

1.3 Binding Free Energy Calculation

The binding affinity of L-Arm to DNA was estimated by binding free energy prediction employing the single trajectory of simulated complex by MM/PBSA combined with entropy change determination.[13,14] The binding free energy change calculated by MM/PBSA and MM/GBSA ($\Delta G_{MMPB(GB)SA}$) can be divided into several individual terms, like follows:

$$\Delta G_{MMPB(GB)SA} = \Delta G_{ele}^{MM} + \Delta G_{vdW}^{MM} + \Delta G_{int}^{MM} + \Delta G_{nonpol}^{sol} + \Delta G_{pol}^{sol}. \qquad (4)$$

where ΔG_{vdW}^{MM} and ΔG_{ele}^{MM} are the van der Waals and electrostatic contributions to the molecular mechanics free energy difference, respectively. ΔG^{sol}_{nonpol} and ΔG^{sol}_{pol} are the polar and nonpolar solvation terms, respectively. The internal and external dielectric constants were set to 1 and 80, respectively. The salt concentration performed was 100 mM and the grid spacing of 0.5 Å was employed for the cubic lattice. The non-polar solvation free energy ΔG^{sol}_{nonpol} was calculated from the solvent-accessible surface area, with a probe radius of 1.4 Å. The surface tension γ and the off-set β were set to 0.00542 kcal/(mol•Å^2) and 0.92 kcal•mol^{-1}, respectively. The changes in configurational entropy upon ligand association ΔS were estimated by an all-atom normal-mode analysis. Prior to normal-mode calculations, the complex, receptor, and ligand were subjected to minimization, with a distance-dependent dielectric constant of $\varepsilon = 4\,r$ and convergence tolerance tighter than a root-mean-squared gradient of drms < 10^{-4} kcal/(mol•Å). MM/PBSA.py[15] in AmberTools12 package was employed.

1.4 Workflow of SELEX in silico

Here in our research, we focus on searching genotype neighborhood in sequence space, thus we limit the number of different bases of mutant DNA is no more than three when compared with wildtype L-Arm-binding 1OLD aptamer. There are 27 single mutants, 324 double mutants, and 2268 triple mutants in our **aptamer neighborhood sequence library**. By considering the formations of secondary structure, we omit the sequences whose $\Delta\Delta G_{gap}$ are greater than 2kcal/mol, thus only 25 single mutants, 293 double mutants, and 2052 triple mutants (2370 sequences in total) are selected for MD-based virtual screening. The binding poses of L-ARM with mutants were generated by performing in silico base mutations on the initial coordinate (PDB code 1O15) by the program mutate_bases of X3DNA[16]. The MD-based virtual screening comprised several rounds. Firstly, 2 ns restraint-free MD simulation was performed on the L-ARM-DNA complex for 2370 mutated sequences. One hundred snapshots taken at 5 ps interval from the last 500 ps MD trajectories were used to calculate the binding free energy by MM/PBSA.py. Then, the 936 DNA sequences with high stability of binding complex or low binding free energy or forming more than three hydrogen bonds with L-Arm were selected for 10 ns MD simulations. 200 snapshots at 10 ps interval and 20 snapshots at 100 ps interval from the last 2 ns of 10 ns MD trajectory were used to calculate the enthalpic and entropic contributions, respectively. In the third round, the MD simulation time increased to 50 ns for the selected 100 sequences from 936 sequences. Extracted from the last 10ns MD trajectory, 500 snapshots and 50 snapshots was prepared for ΔG_{MMGBSA} and -TΔS_{NM} calculation, respectively. To improve the accuracy of binding free energy prediction, twenty replicas of 10 ns MD simulations were performed for each mutant, then the binding affinity averages calculated from the full 20 member ensembles are chose for the final comparison of DNA-L-Arm binding affinities. Considering the heavy computational burden, only 20 sequences are selected for multi-replicas MM/PB(GB)SA binding free energy estimation.

Supplementary Text S2: Experimental validation

DNA oligonucleotides were synthesized by Sangon Biotech Co., Ltd. (Shanghai, China) while L-Argininamide (L-Arm) was obtained from Sigma Chemical Co. Ultrapure water was used in all the experiments. All thermodynamic experiments were performed in a buffer composed of 10 mM sodium phosphate, 25 mM NaCl at pH 6.5.

Circular dichroism measurement[2,17,18] was performed at 25 °C on a Jasco J-810 spectropolarimeter (Jasco, Inc., Easton, MD) equipped with a heating/cooling device with nitrogen purging facilities. The CD spectrum were recorded from 320 to 220 nm with four time scans at a scanning rate of 100 nm•min^{-1}. The concentration of DNA was fixed at 4.5 mM while different concentrations of ligand (0, 5, 10, 25, 50, 75, 125, 250 mM) were titrated to form ligand-DNA complexes.

Model-free ITC protocol was adopt to determine the binding enthalpies of L-argininamide with aptamer because such method could obtain multiple estimates of ΔH^o and avoid any possible fitting bias, as demonstrated by previous research[18]. Typically, 10 μL aliquots of 20 μM aptamer stock solution were added to 1.44 mL of a 3 mM L-argininamide solution at intervals of 300 s with continuous stirring of the solution at 47 rpm over the course of the experiment. The initial delay was set to 300 s and the thermal equilibration was set at 25 °C. To calculate the heats of aptamer dilution, we also injected aptamer solution into the reaction cell loaded with buffer alone. Data were collected as heat released (μcal/mol) in the exothermic reactions versus time (s) and the area under each peak was integrated to give the measure of the heat. For each aptamer, the averaged binding enthalpies (ΔH^o) were obtained after normalization and necessary correction for any heats of dilution.

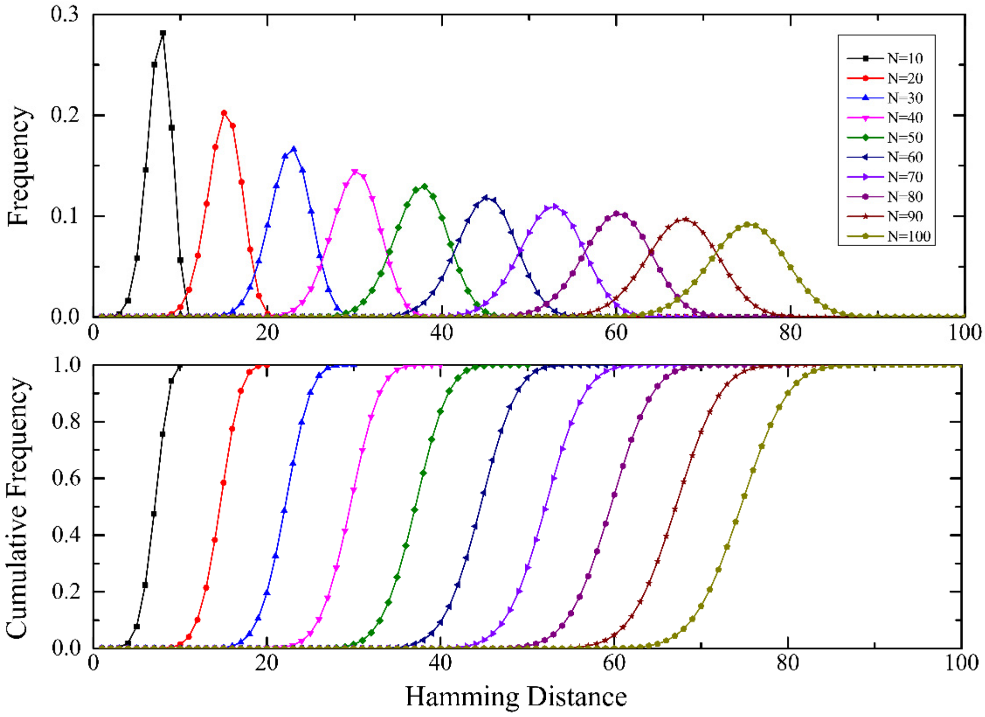

Figure S1. Frequency distribution of the sequences that having identical Hamming distance from a specific sequence in whole nucleic acid (DNA or RNA) sequence space. Hamming distance was defined as the number of positions at which two sequences differ. If the sequence length is $\boldsymbol{L}$, the corresponding sequence space is comprised of $\boldsymbol{\Omega} = \mathbf{4}^{\boldsymbol{L}}$ sequences. The number of sequences at a Hamming distance $\boldsymbol{d}$ from a given sequence in sequence space is $\mathbf{N(d)} = \mathbf{3}^{\boldsymbol{d}} \times \frac{\boldsymbol{L}!}{\boldsymbol{d}!(\boldsymbol{L}-\boldsymbol{d})!}$, which accounts for the proportion $\mathbf{P(d)} = \frac{\mathbf{3}^{\boldsymbol{d}}}{\mathbf{4}^{\boldsymbol{L}}} \times \frac{\boldsymbol{L}!}{\boldsymbol{d}!(\boldsymbol{L}-\boldsymbol{d})!}$ in sequence space.

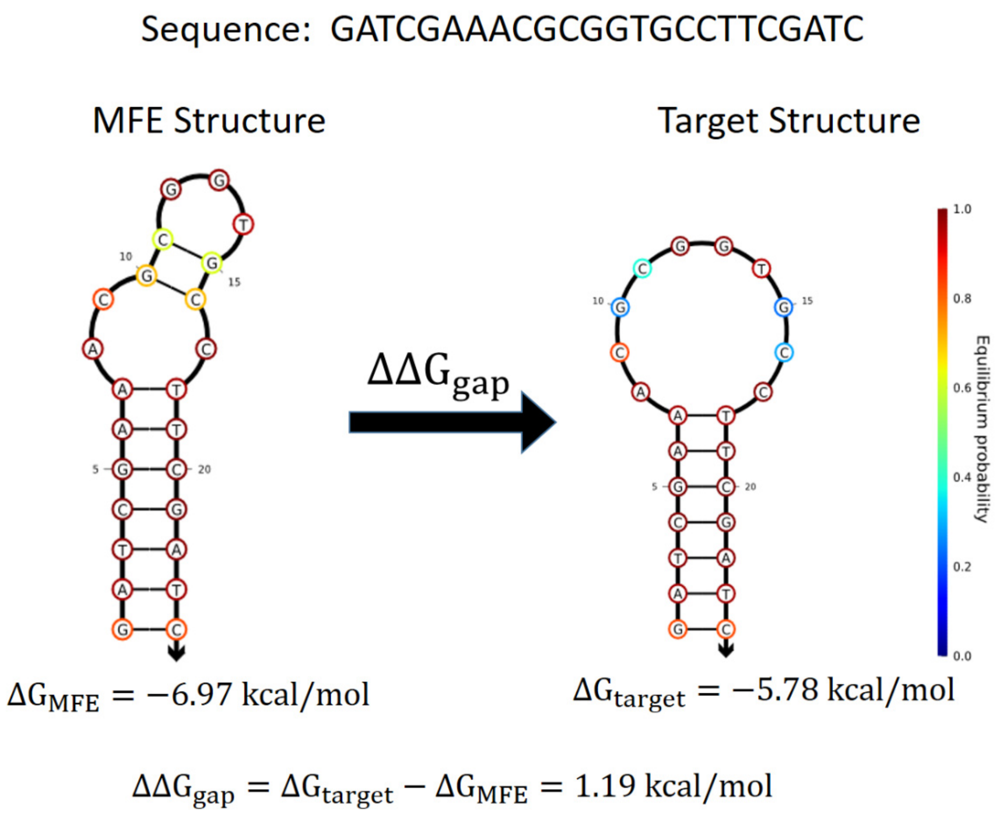

Figure S2. Schematic of the free-energy gap of target motif transformation. For each sequence, free energy in MFE state ΔG_{MFE} and that in target motif ΔG_{target} were calculated by NUPACK at temperature 25℃, and the difference is defined as free-energy gap $\Delta\Delta G_{gap}$. Each base is shaded according to the probability that it adopts the depicted paired or unpaired state at equilibrium.

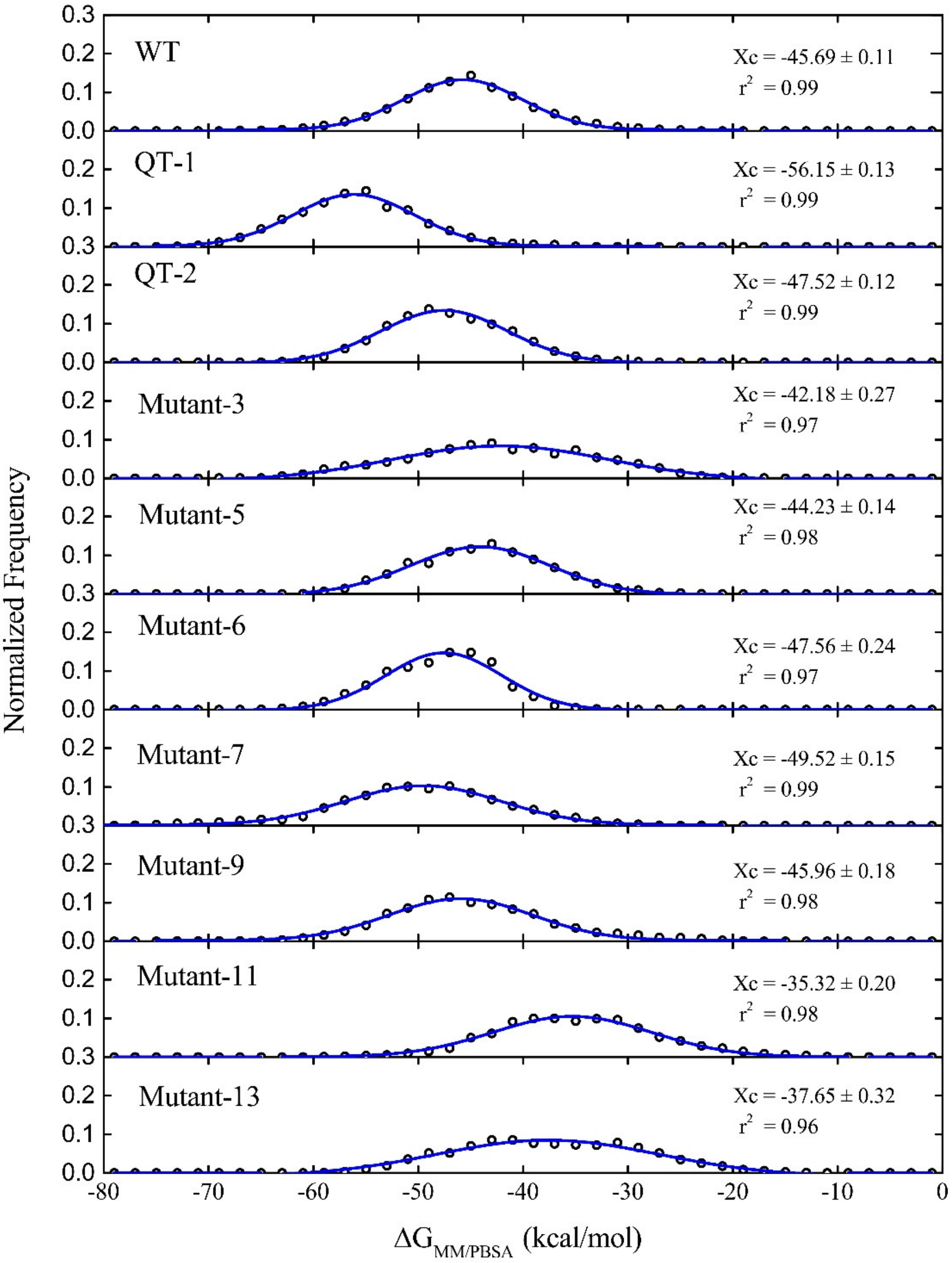

Figure S3. Normalized frequency distribution analysis of the calculated binding free energies $\Delta G_{MM/PBSA}$, which are shown in per snapshot for the WT L-Arm-binding aptamers, its genotype neighborhoods including two in silico selected aptamer QT-1 and QT-2. The data is shown with open circles in black and the expected normal distribution is shown by the blue lines.

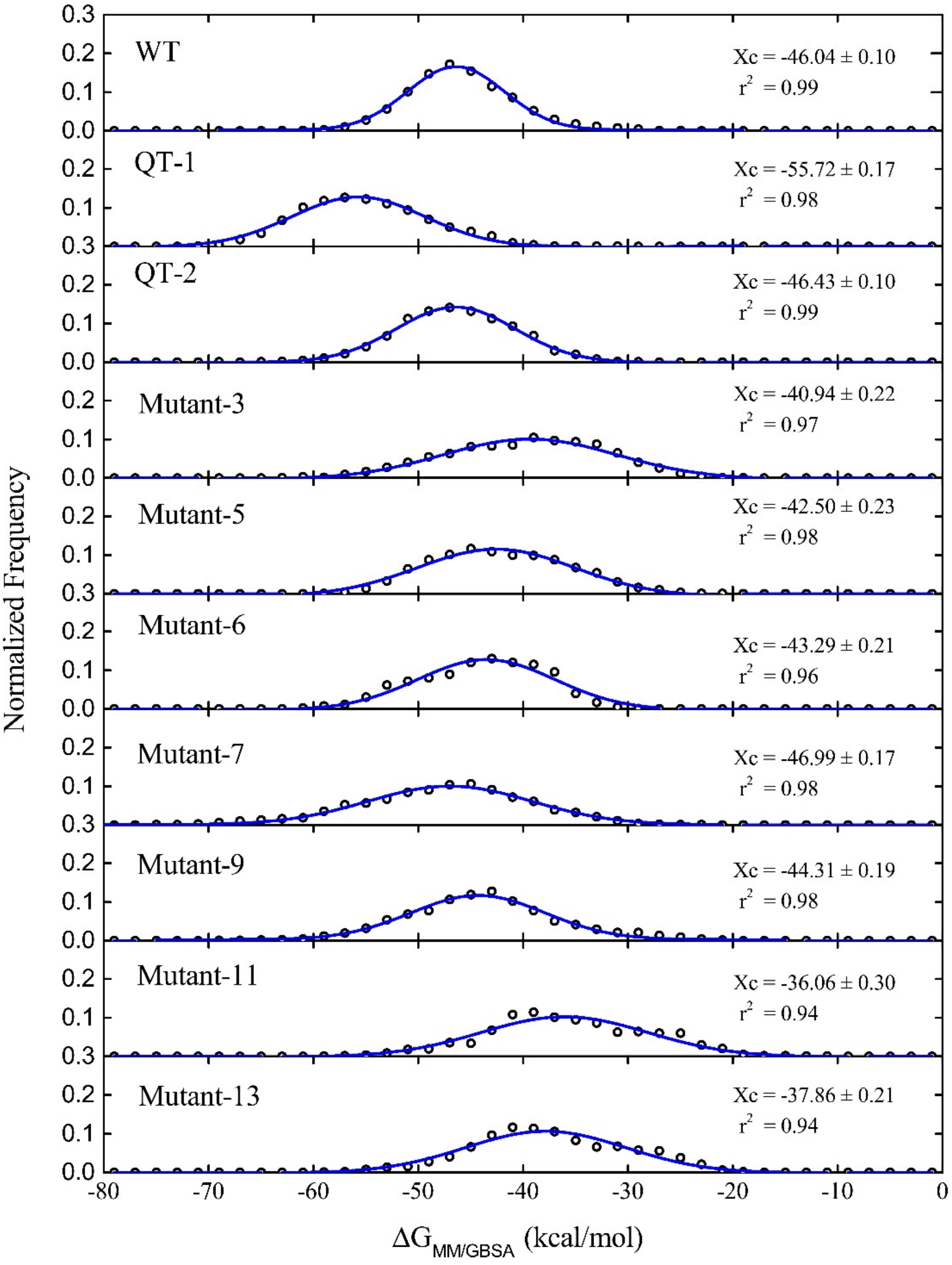

Figure S4. Normalized frequency distribution analysis of the calculated binding free energies $\Delta G_{MM/GBSA}$, which are shown in per snapshot for the WT L-Arm-binding aptamers, its genotype neighborhoods including two in silico selected aptamer QT-1 and QT-2. The data is shown with open circles in black and the expected normal distribution is shown by the blue lines.

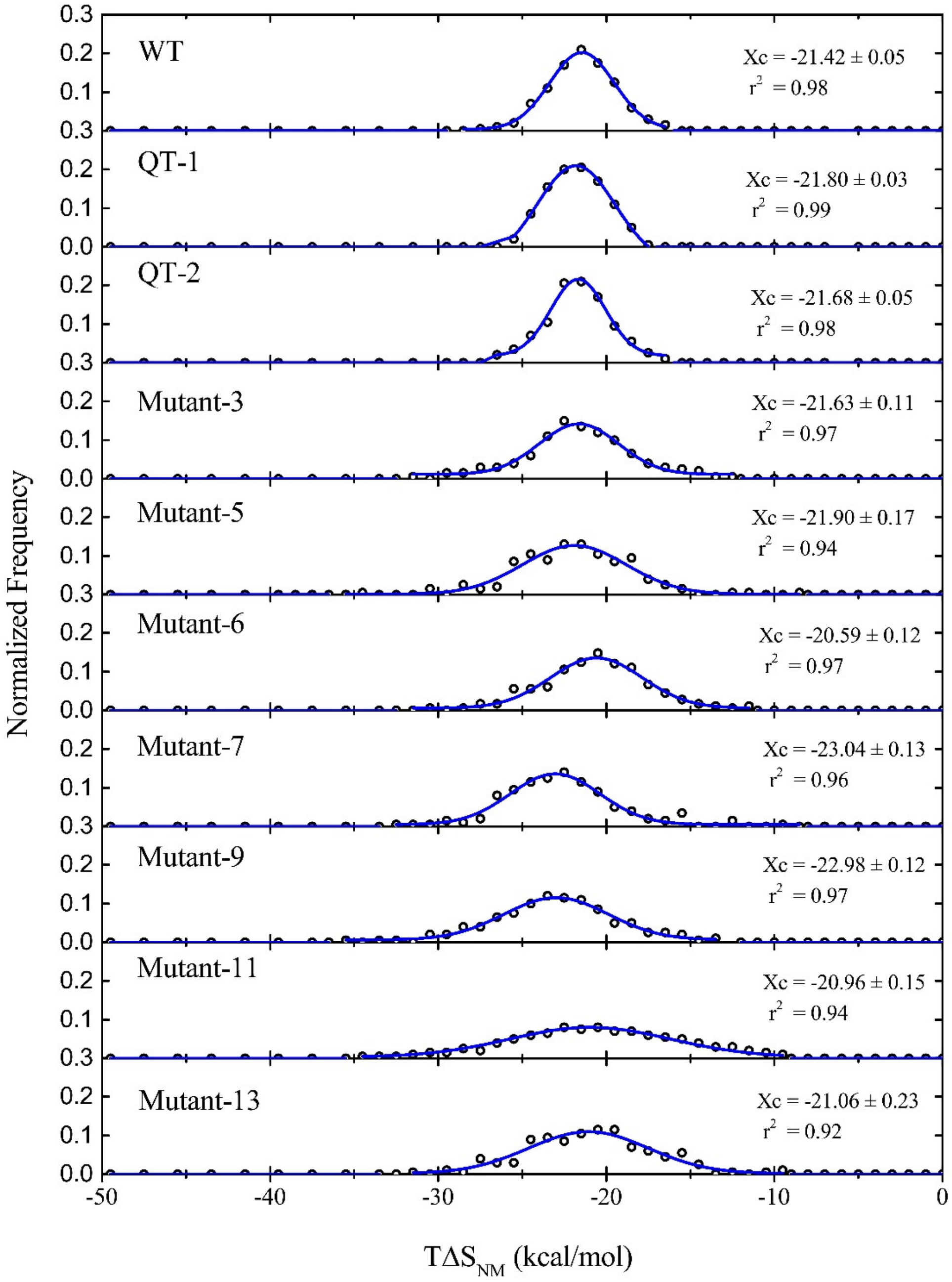

Figure S5. Normalized frequency distribution analysis of the calculated binding free energies $T\Delta S_{NMA}$, which are shown in per snapshot for the WT L-Arm-binding aptamers, its genotype neighborhoods including two in silico selected aptamer QT-1 and QT-2. The data is shown with open circles in black and the expected normal distribution is shown by the blue lines.

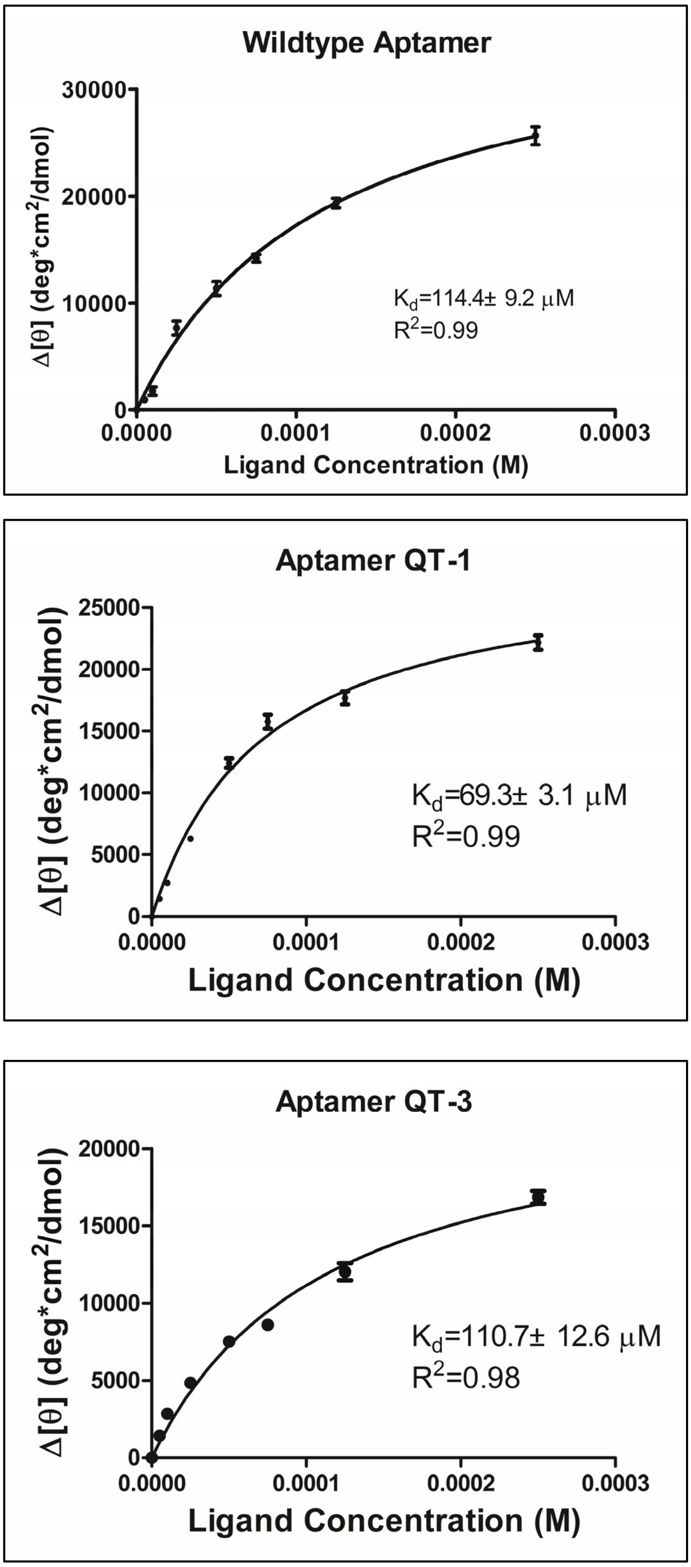

Figure S6. Binding affinity determination for L-Arm-binding DNA aptamers by circular dichroism (CD) spectrum measurement.

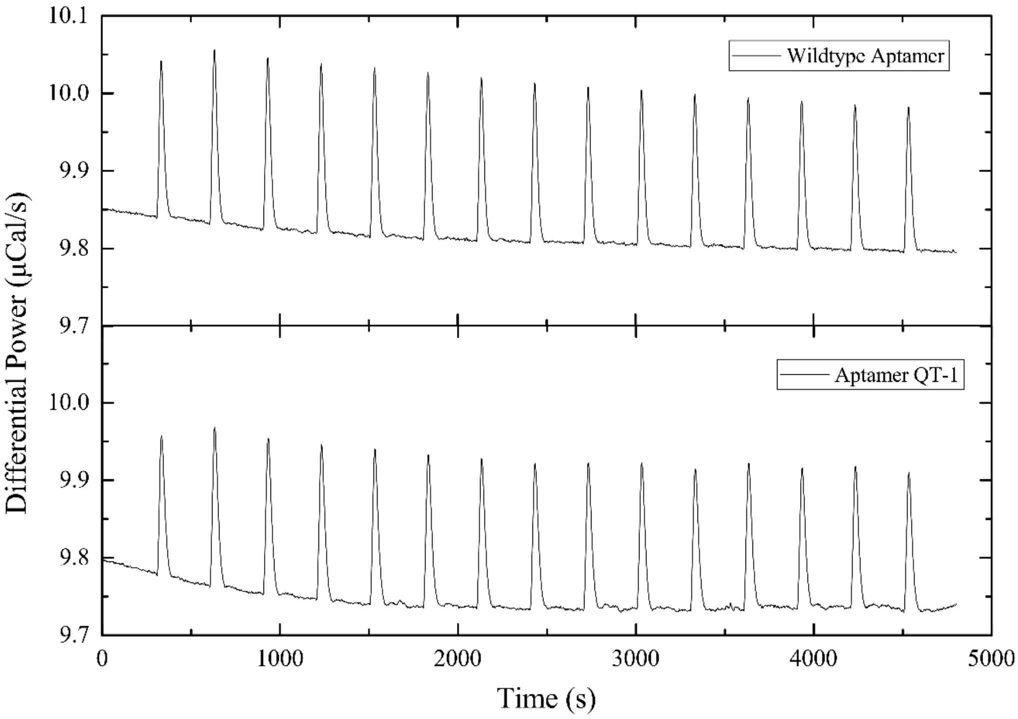

Figure S7. Representative primary data from isothermal titration calorimetry (ITC) experiments for the wildtype aptamer and in silico screened aptamer QT-1. ITC data at 25℃ are shown for the titration addition of 10 μL aliquots of 20 mM aptamer stock solution to 1.44 mL of a 3 mM L-argininamide solution at 5 min intervals.

Table S1. Experimental binding data for L-Arm binding aptamer and its mutants.

	Hairpin mutants of longer aptamer (clone 12-28) [a]	Corresponding mutations on 1OLD aptamer	Required salt concentration to elute (compared with wildtype hairpin) [b]	K_A(/M)	K_d(μM)
Wildtype	-	-	0	8,000	125
EMBO-1995	A10G	A8G	-105	178 [c]	5622[c]
	A10T	A8T	-96	263 [c]	3797 [c]
	C11T	C9T	-62	1018 [c]	982 [c]
	G12A	G10A	-96	263 [c]	3797 [c]
	G12I	G10I	-57	1223 [c]	817 [c]
	T13C	T11C	-48	1689 [c]	592 [c]
	A14G & G15T	A12G & G13T	+14	12053 [c]	83 [c]
	G15T	G13T	+14	12053 [c]	83 [c]
	C16T	C14T	-100	222 [c]	4512 [c]
	G17A	G15A	-48	1689 [c]	592 [c]
	G17I	G15I	-28	3330 [c]	300 [c]
	C18T	C16T	-96	263 [c]	3797 [c]
	C19A	C17A	-115	113 [c]	8870 [c]
	A9T & T20A	A7T & T18A	-96	263 [c]	3797 [c]
	A9G & T20C	A7G & T18C	-96	263 [c]	3797 [c]
QT-1	-	T11C & A12G & C14T	-	14,430[d]	69 [d]
QT-2	-	C14A & G15T	-	9,033[d]	111 [d]
Non-functional genotype neighbors in this study	-	A8C &T11C &C14G	-	NA[e]	-
	-	A8G &A12G &C14T	-	NA[e]	-
	-	A8G &A12G &C17T	-	NA[e]	-
	-	A8T &G13C &C14A	-	NA[e]	-
	-	G10C &A12C &C16A	-	NA[e]	-
	-	T11G &G13C &C17T	-	NA[e]	-
	-	T11G &G13T &C14T &G15A	-	NA[e]	-
	-	A12G &G15C &C16T	-	NA[e]	-

[a] Aptamer clone 12-28 is a 28-mer L-Arm binding aptamer[2]. The difference between clone 12-28 and 1OLD aptamer is that clone-28 has two extra base pairs (G1•C28 and G2•C27) on the stem region.

[b] Demonstrated by Harada, K. et.al[2], mutant DNAs were eluted from an arginine column using an NaCl gradient, while the numbers indicate the difference in salt concentration required to elute each mutant compared with the wild-type hairpin. Negative values indicate weaker binding and positive values indicate stronger binding to arginine.

[c] According to previous research[2,19], the salt elution profiles of the DNAs correlated with relative arginine binding affinities, here the binding parameter K_A was calculated from the linear fitting between logK_A and log[NaCl] salt concentration.

[d] The binding parameter K_A for aptamer QT-1 and QT-2 were determined by circular dichroism studies.

[e] There are no change in CD signal in the presence of L-Arm for these 1OLD random genotype neighbors, thus K_A were referred as NA (not active).

References

(1) Zadeh, J. N.; Steenberg, C. D.; Bois, J. S.; Wolfe, B. R.; Pierce, M. B.; Khan, A. R.; Dirks, R. M.; Pierce, N. A. *Journal of computational chemistry* **2011**, *32*, 170.

(2) Harada, K.; Frankel, A. D. *The EMBO journal* **1995**, *14*, 5798.

(3) Lin, C. H.; Patel, D. J. *Nature structural biology* **1996**, *3*, 1046.

(4) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. *Biophysical journal* **2007**, *92*, 3817.

(5) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926.

(6) Aqvist, J. *J Phys Chem-Us* **1990**, *94*, 8021.

(7) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, Inc.: Wallingford, CT, USA, 2009.

(8) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *The Journal of Chemical Physics* **1995**, *103*, 8577.

(9) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *Journal of computational chemistry* **2005**, *26*, 1701.

(10) Bussi, G.; Donadio, D.; Parrinello, M. *The Journal of chemical physics* **2007**, *126*, 014101.

(11) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684.

(12) 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.

(13) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. *Annual review of biophysics and biomolecular structure* **2001**, *30*, 211.

(14) Stoica, I.; Sadiq, S. K.; Coveney, P. V. *Journal of the American Chemical Society* **2008**, *130*, 2639.

(15) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. *Journal of chemical theory and computation* **2012**, *8*, 3314.

(16) Lu, X. J.; Olson, W. K. *Nature protocols* **2008**, *3*, 1213.

(17) Lin, P. H.; Tong, S. J.; Louis, S. R.; Chang, Y.; Chen, W. Y. *Physical chemistry chemical physics : PCCP* **2009**, *11*, 9744.

(18) Bishop, G. R.; Ren, J.; Polander, B. C.; Jeanfreau, B. D.; Trent, J. O.; Chaires, J. B. *Biophysical chemistry* **2007**, *126*, 165.

(19) Tao, J.; Frankel, A. D. *Biochemistry* **1996**, *35*, 2229.