Rational Design of Structurally Rich DNA Aptamer Screening Libraries Guided by Empirical Sequence Trends and Stable Loop Motifs

Steven Ochoa, Valeria Tohver Milam

Supporting Information:

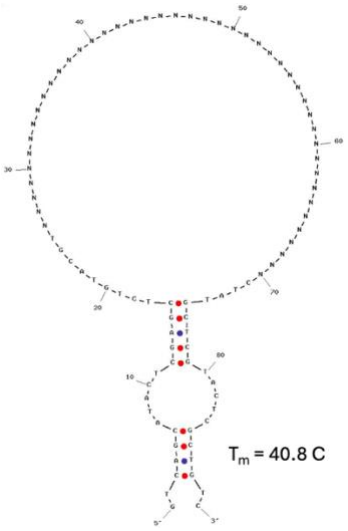
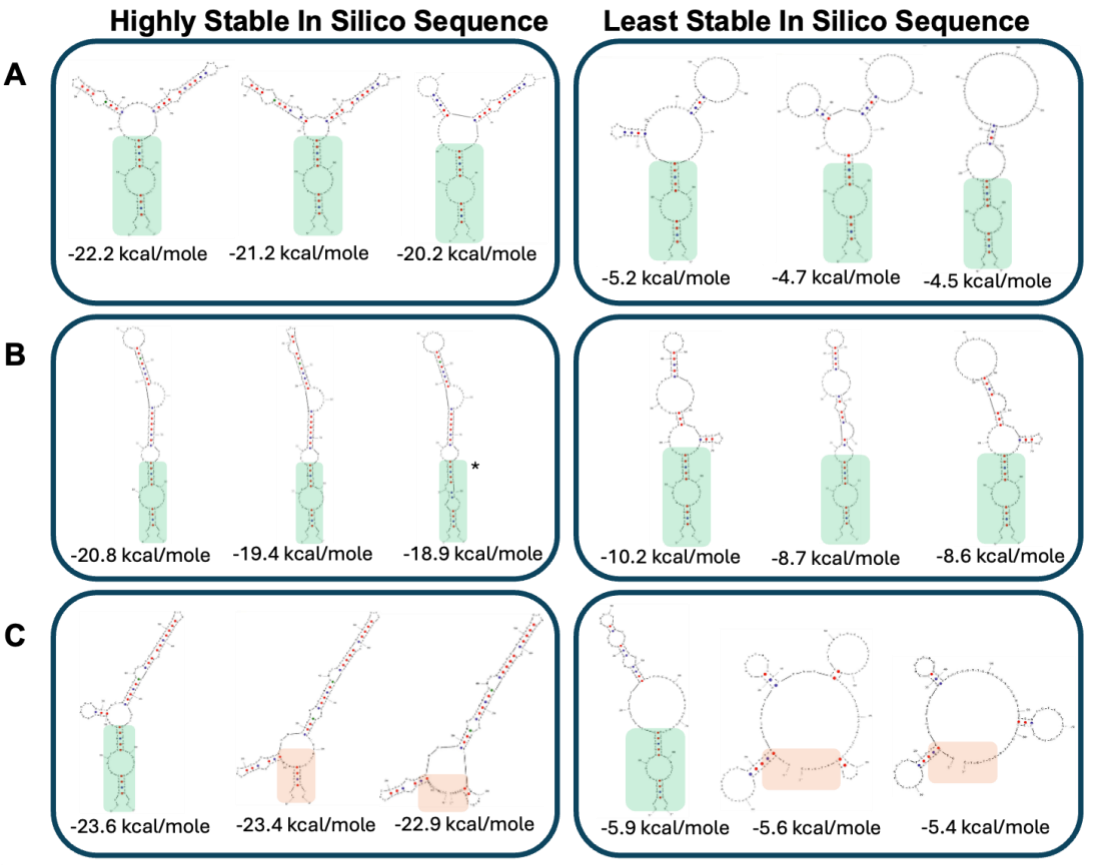


Fig. S1. Secondary structure schematic generated with IDT Oligoanalyzer (https://www.idtdna.com/pages/tools/oligoanalyzer) illustrating the weak stem predicted to form between both fixed-base segments of template sequences for the empirical, tetraloop and standard (shown above) libraries. Multiple suboptimal secondary structures (not shown) predicted similar imperfect stems as shown above.



	Highly Stable In Silico Sequence (5'-3')	Least Stable In Silico Sequence (5'-3')
A	GTCAGCATACTCGAGCTCTGTACGT ATGGGTTTCCCGGCGTAAACCAGCGGGAGCA AATGCTTGCTCGCG CTATGCTCGTACTCGCTGTC	GTCAGCATACTCGAGCTCTGTACGT TTACATACCCCCCAACTATGAACCTCGCCGTC TCCTAAGTCATGT CTATGCTCGTACTCGCTGTC
B	GTCAGCATACTCGAGCTCTGAGTCC CGAGCTACTGCAGACCTGTGGGCGGTAGCGT AATTTTAATCGGGA CTATGCTCGTACTCGCTGTC	GTCAGCATACTCGAGCTCTGAGTCC GGTGGGTTTTGTGTGGTGATATATCCACAGTT TACGGGAGGGGGC CTATGCTCGTACTCGCTGTC
C	GTCAGCATACTCGAGCTCTGCTGGA GAAAGCGCAGGACGATCACGCCAGGGCGTAG CCCTGGGCGTTGTC CTATGCTCGTACTCGCTGTC	GTCAGCATACTCGAGCTCTGCTGGA TATAACCACTGAAGTTCACCGTTGATCCTCCC CTCCGATGTTGTG CTATGCTCGTACTCGCTGTC

Fig. S2. Secondary structure analysis, including suboptimal structures, of representative in silico sequences taken from the extremes of predicted MFE values in Fig. 4 from the A) standard library, B) empirical library and C) tetraloop library. For each library, the left panel shows three mFold predicted secondary structures for a representative sequence among the most thermally stable (i.e., large negative MFE value) sequences; and the right panel shows three predicted secondary structures for a representative sequence among the least stable (small negative MFE value) sequences in the bottom panels. The formation of the intended hairpin structure between

the fixed-base segments highlighted by a green box (intended stem formation) or an orange box (lack of/disrupted stem formation). The asterisk marks one sequence with a predicted stem formation involving different base-pairings from other stems highlighted in green. The primary structures of the six sequences are tabulated in the bottom panel.

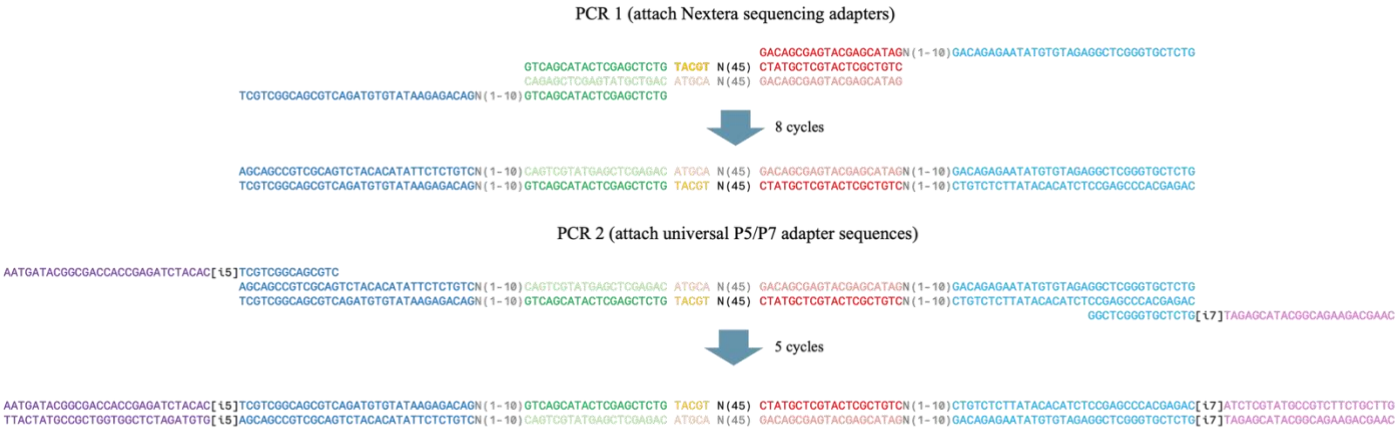


Fig. S3. Schematic of the two-step PCR method used to incorporate heterogeneous sequencing adapters. The raw NGS sequencing data obtained and used in this work is available at https://doi.org/10.5281/zenodo.18719585.

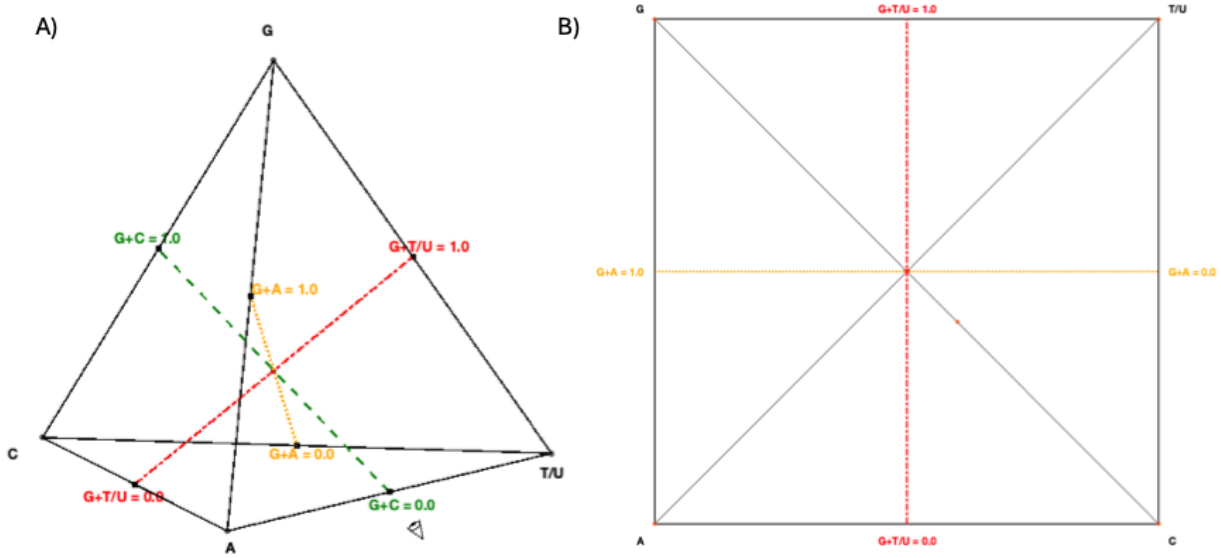


Fig. S4. A) An empty unit simplex tetrahedron with notable axes labeled and shown as colored, dashed lines as an aid for visualizing the 3D geometry and the perspective used to generate the B) 2D projection of the tetrahedron along Chargaff's axis (green dashed line in A).

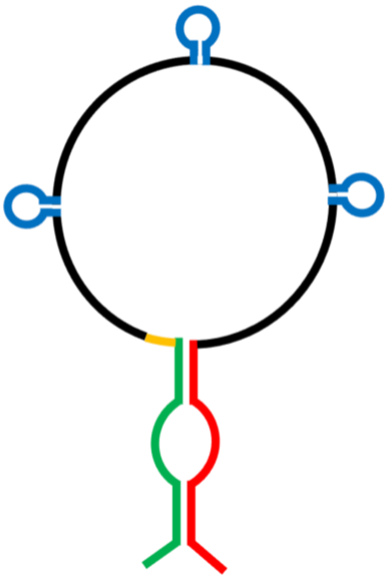


Fig. S5. Schematic of the secondary structure intended with the Tetraloop library design with the tetraloop motifs highlighted in blue.

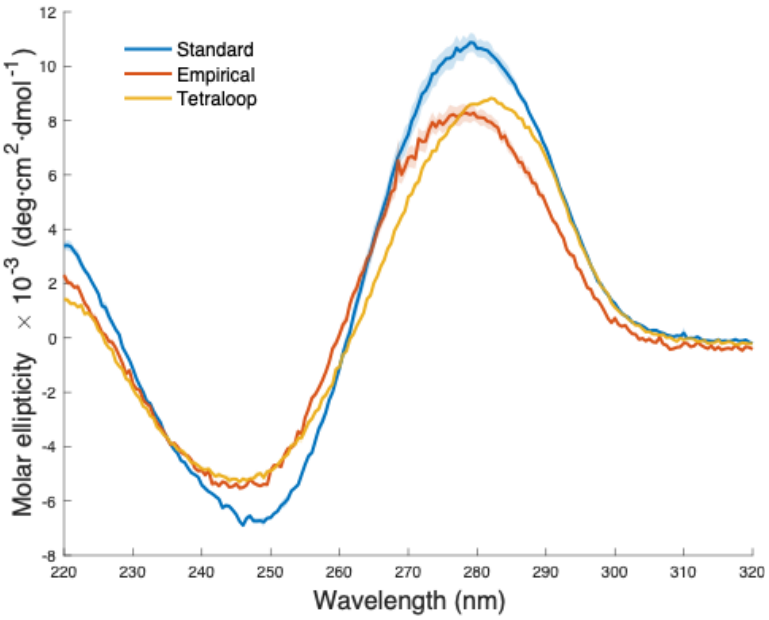


Fig. S6. Circular dichroism spectra measured at 25°C in for all three libraries. Spectra for each library represents the average of $n = 3$ independent scans and the shaded regions indicate the standard error for each dataset.

Table S1. List of Protein Data Bank (PDB) aptamer entries (PDB ID) with experimentally verified G-quadruplex forming structures; aptamer sequence; type of nucleic acid (NA); base length; and the minimum free energy (as change in Gibbs Free Energy) predicted by mFold the most stable, predicted canonical secondary structure.

PDB ID	NA seq 5'-3'	NA type	NA Length	Predicted ΔG (kcal/mol)
2RQJ	GGAGGAGGAGGA	RNA	12	-11.02
7E5P	GGGTGGGTTGGGAGGG	DNA	16	1.84
7XHD	GGGGTGGGTGGTGGGT	DNA	16	2.65
6K84	GGAGGAGGAGGAAGGAGGAGGAGGA	RNA	25	no structure predicted
4DIH	GGTTGGTGTGGTTGG	DNA	15	2.57
6EVV	CGCCTAGGTTGGGTAGGGTGGTGGCG	DNA	26	-3.01
4I7Y	GTCCGTGGTAGGGCAGGTTGGGGTGAC	DNA	27	-0.52
5CMX	GTGACGTAGGTTGGTGTGGTTGGGGCGTCAC	DNA	31	-4.99
8BW5	GGTCAGATGATGGGGATGGGGGGTTGGAGGAATGGATGACC	DNA	41	-1.28