Supporting Methods

Sampling error

Treating the per-position nucleotide frequencies (F_i) as though they were actual probabilities (P_i) results in a sampling error in the information content calculation. Schneider et al.[21] demonstrated a method for calculating the standard error by computing the weighted uncertainties for all possible combinations of bases for a given number of sequences in an alignment. To account for the reduction in uncertainty due to a finite number of sequences, we subtract a correction factor from the information content calculated at each position. The correction factor is the difference between the known maximum uncertainty (2 bits/position for RNA) and the calculated expectation value for the maximum uncertainty in an alignment of a certain size. To compute the values for the expected maximum uncertainty, error correction and variance per position, we implemented a version of CALHNB by Schneider et al.[21] in Perl. We modified their algorithm to adjust the frequency distributions obtained from the sequence alignments to compute the information content required to specify the active structures in completely random RNA sequence space. The values were computed for all of the alignments with prior base probabilities of 0.79, 0.07, 0.07, 0.07. The chart below gives a sense of the magnitude of these values.

Number of sequences	Expectation value (bits)	Error correction (bits)	Variance per position
30	1.74	0.26	0.060
40	1.82	0.18	0.034
50	1.86	0.14	0.020
60	1.89	0.11	0.013
70	1.91	0.09	0.009
80	1.92	0.08	0.006

Normalizing and adjusting base frequencies

We computed the information content of each aptamer using the selection sequence alignments. The sequence pools were synthesized with 21% rate of mutagenesis, where each of the non-wild-type bases was equally likely. Because we know the prior probability of each base type *i* at every position, we can use the frequency distribution of nucleotides from the alignment to compute the information content of the aptamers in relation to random (P_A=P_U=P_C=P_G) sequence space.

As an illustration (see below), consider an aptamer that requires a purine (A or G) in a certain position to be functional. For simplicity imagine that we synthesize a pool of 100 aptamer sequences mutated 21% at that position with A as the original base. In the idealized case we would have 79 sequences with an A, 7 with a U, 7 with a G, and 7 with a C. After selection, only the A- and G-containing sequences remain. The total number of sequences in the pool is reduced to 86, where 79 are A and 7 are G. The relative frequency of A and G after the selection reflects the fact that A and G are equally compatible with a functional molecule and that A was present much more often in the starting pool than G. To normalize each F_i for the skewed starting frequencies we divide each F_i after the selection by the F_i before the selection. Dividing each normalized F_i by the sum of the normalized F_is gives the adjusted frequencies of each base type relative to the others, In the current example, the normalized and adjusted F_A=0.5 and F_G=0.5; the information content of this position is 1 bit, which is the information content needed to specify a two-base varying position in an RNA structure.

Before Selection

base	# seqs	frequency
A	79	0.790
U	7	0.070
G	7	0.070
C	7	0.070
Total	100	1.000

After Selection

base	# seqs	frequency
A	79	0.919
U	0	0.000
G	7	0.081
C	0	0.000
Total	86	1.000

Normalized

base	# seqs	frequency
A	79	1.160
U	0	0.000
G	7	1.160
C	0	0.000
Total	86	2.320

Adjusted

base	# seqs	frequency
A	79	0.500
U	0	0.000
G	7	0.500
C	0	0.000
Total	86	1.000

Supporting Figure 1A-K. GTP aptamer sequence alignments. DNA sequence alignments from each of the selections are shown in order from highest affinity for GTP to lowest. The top line in each alignment contains the minimized sequence (corresponding to the length of the optimized sequence). This is the sequence used as the template for synthesizing the library of mutants. Immediately below that is the secondary structure model of the optimized aptamer shown in bracket form where the 5' -most open bracket "(" base pairs with the 3'-most closed bracket ")". Paired strands are shaded with same color. Point mutations in the loops that were assayed are indicated on the same line as the secondary structure. Blue bases were tested and found not to improve binding activity. Red bases improved the activity of the aptamer relative to the minimized construct and were incorporated into the optimized version.

Mutations in the stem regions of the alignment are shaded to indicate:

Red = Watson-Crick covariation
Orange = new Watson-Crick base pair
Green = new wobble (A-C or G-U) pair
Black = broken pairing

Mutations in the loop are shaded purple

Figure 1A. 9 – 4.

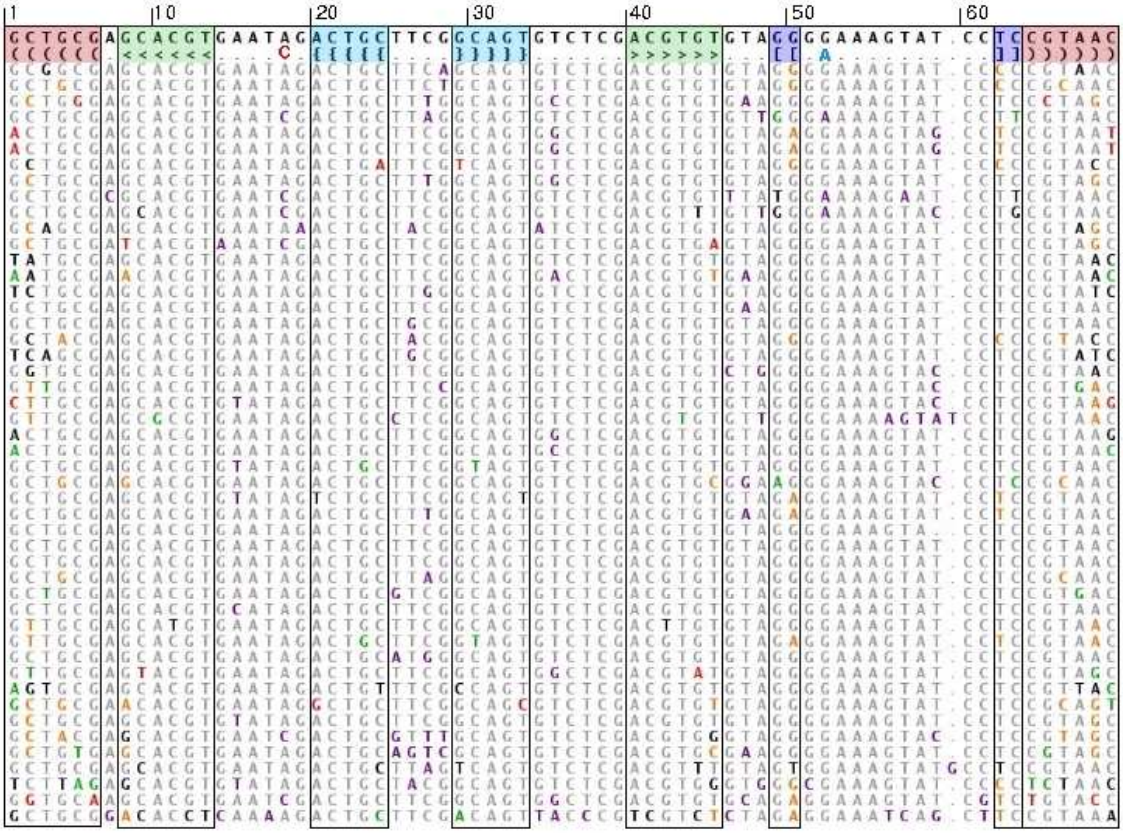


Figure 1B. Class V.

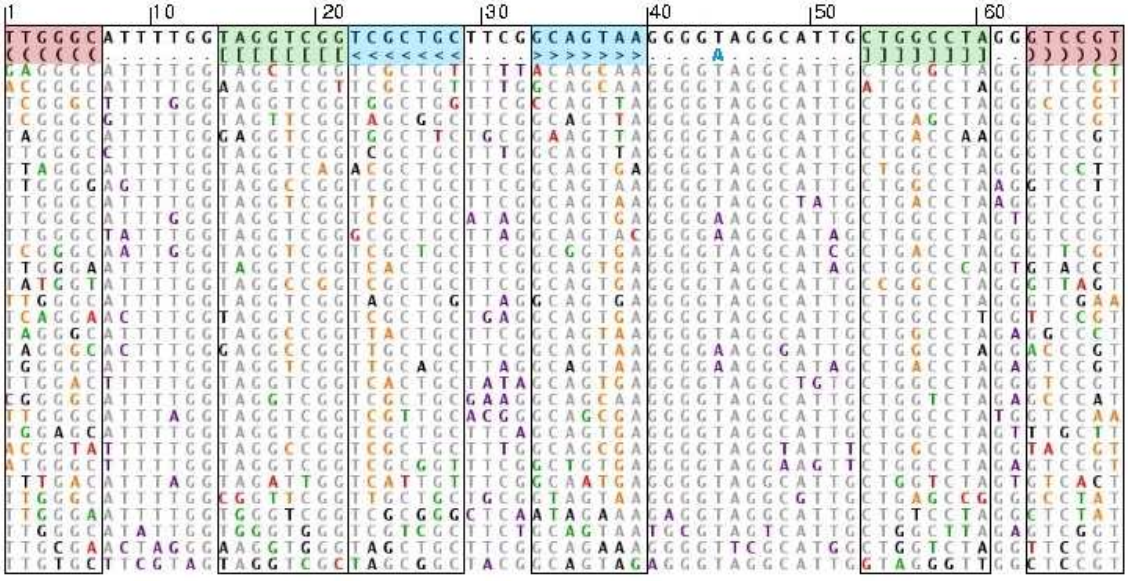


Figure 1C. 10 – 10.

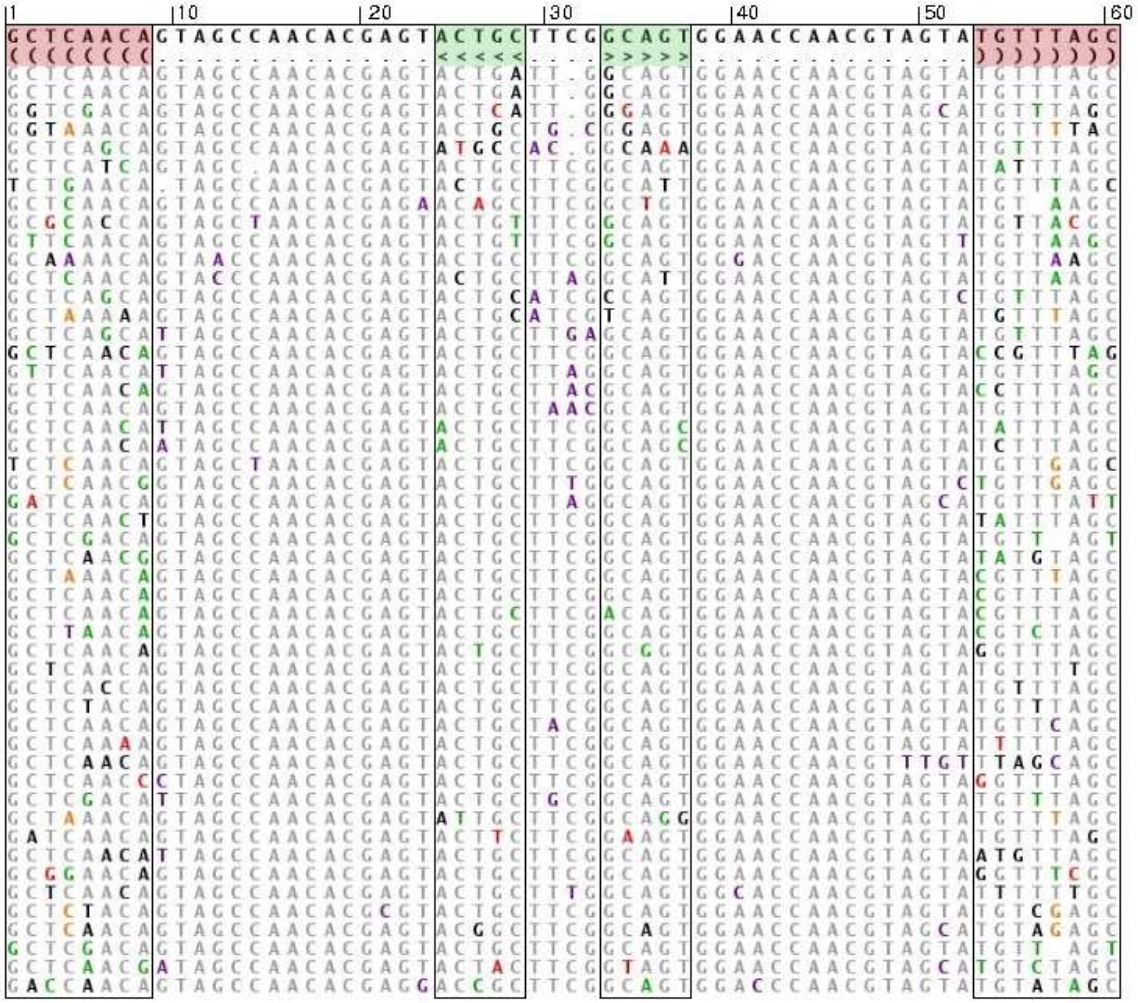


Figure 1D. Class I.

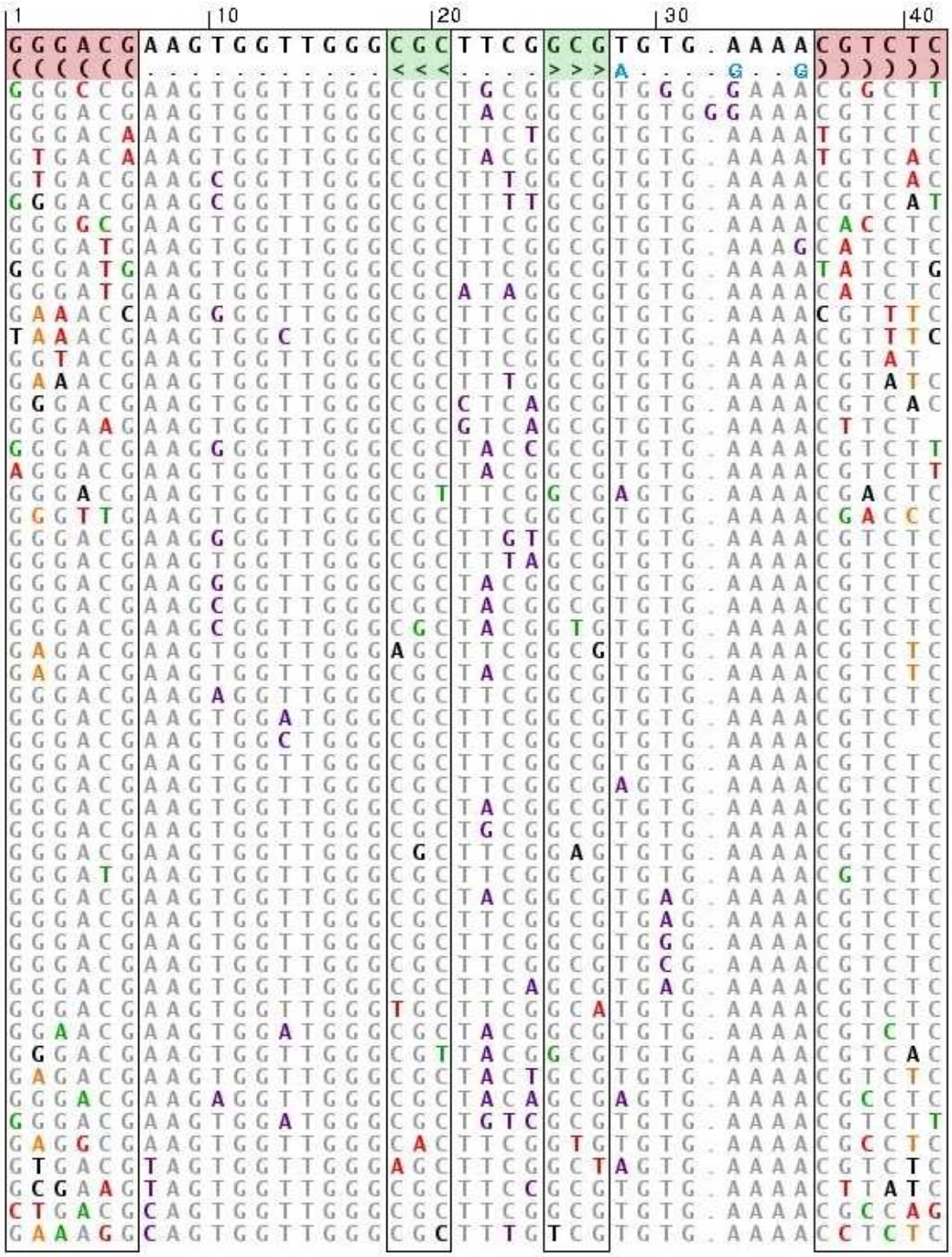


Figure 1E. 10 – 59.

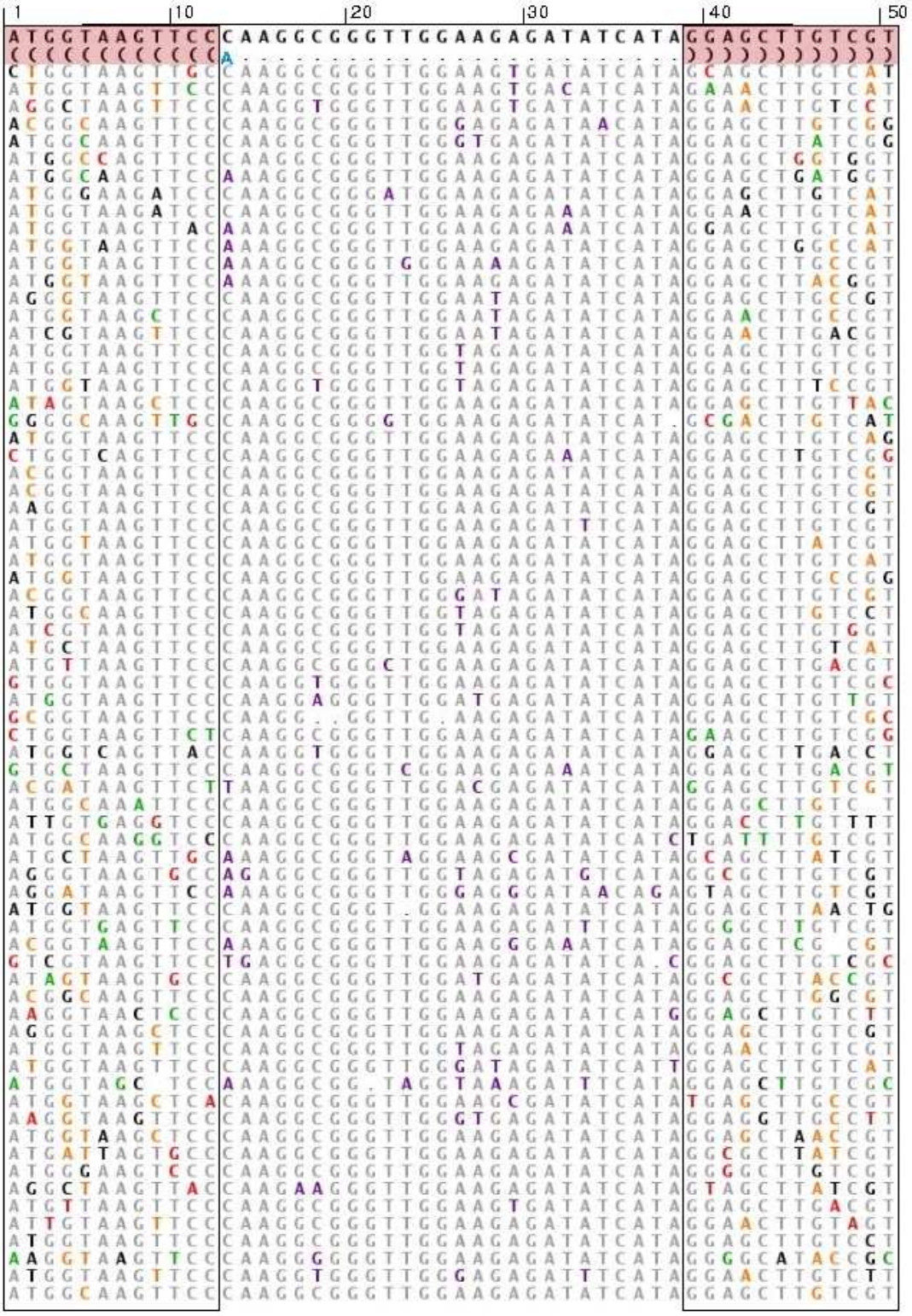


Figure 1F. 10 – 24.

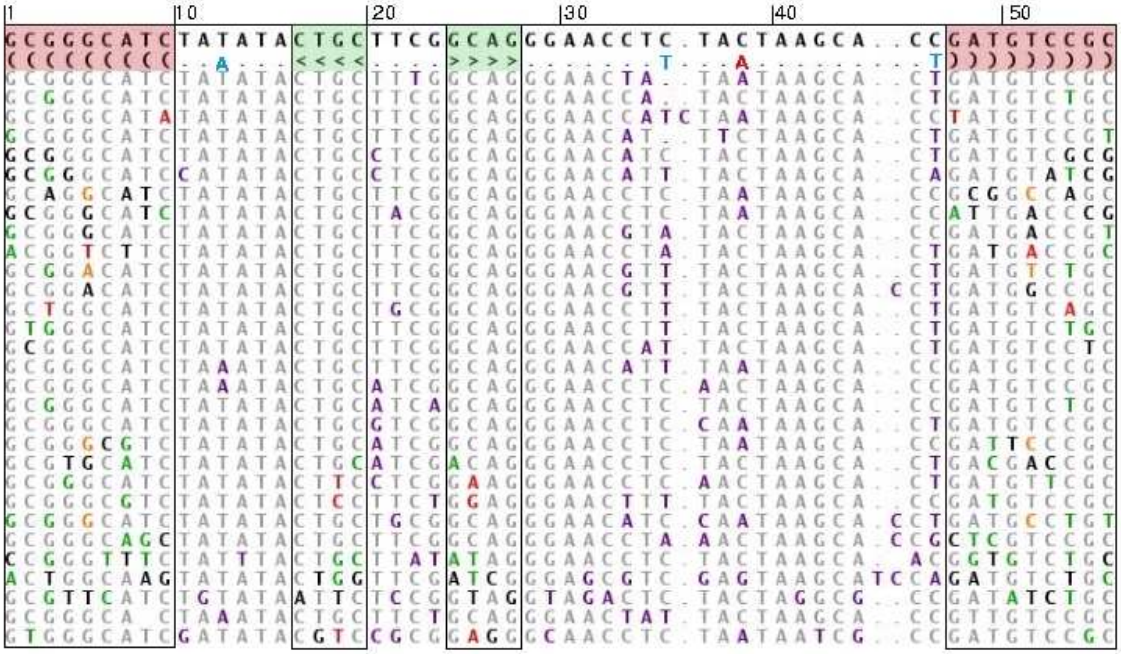


Figure 1G. 9 – 12.

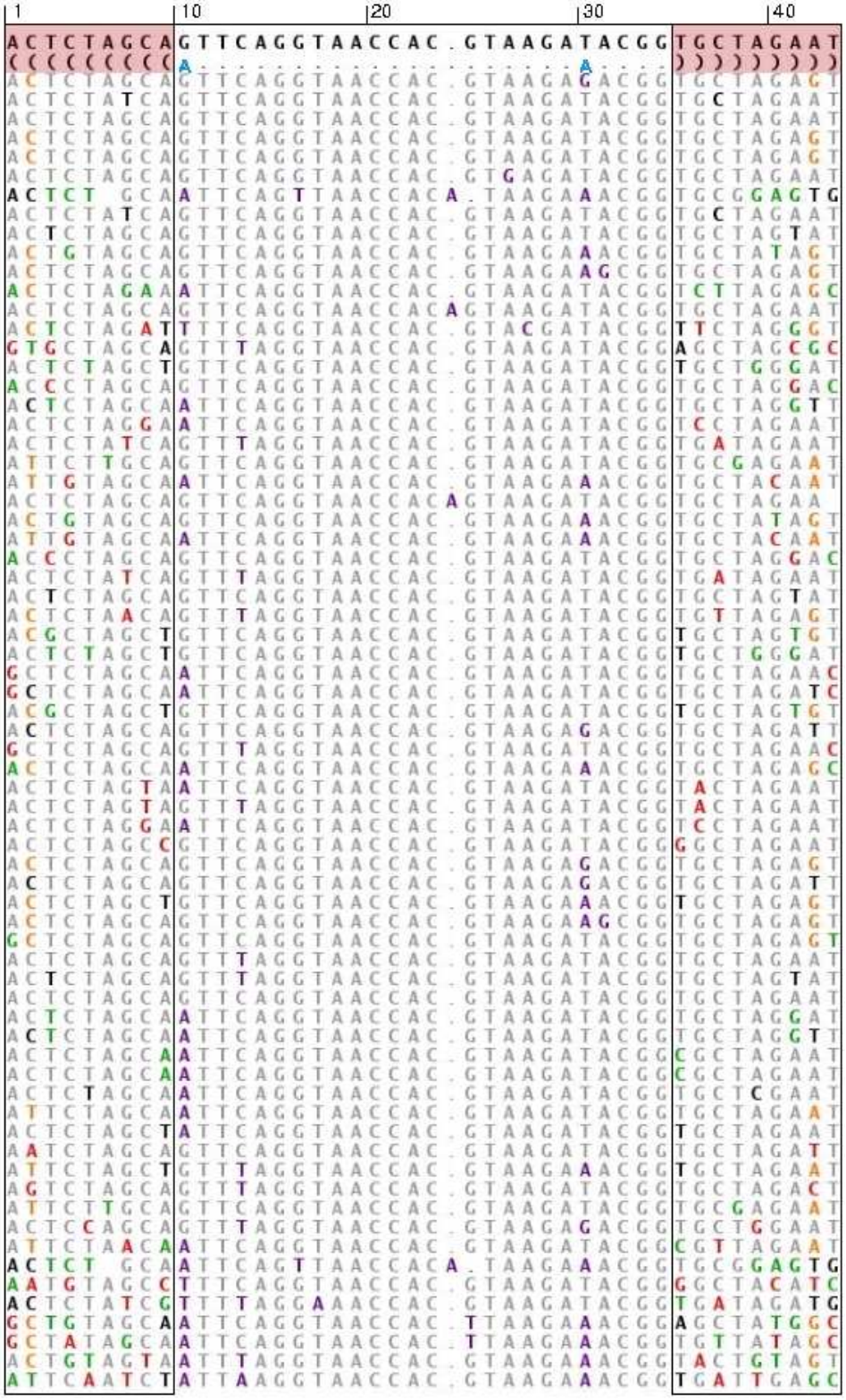


Figure 1H. 10 – 6.

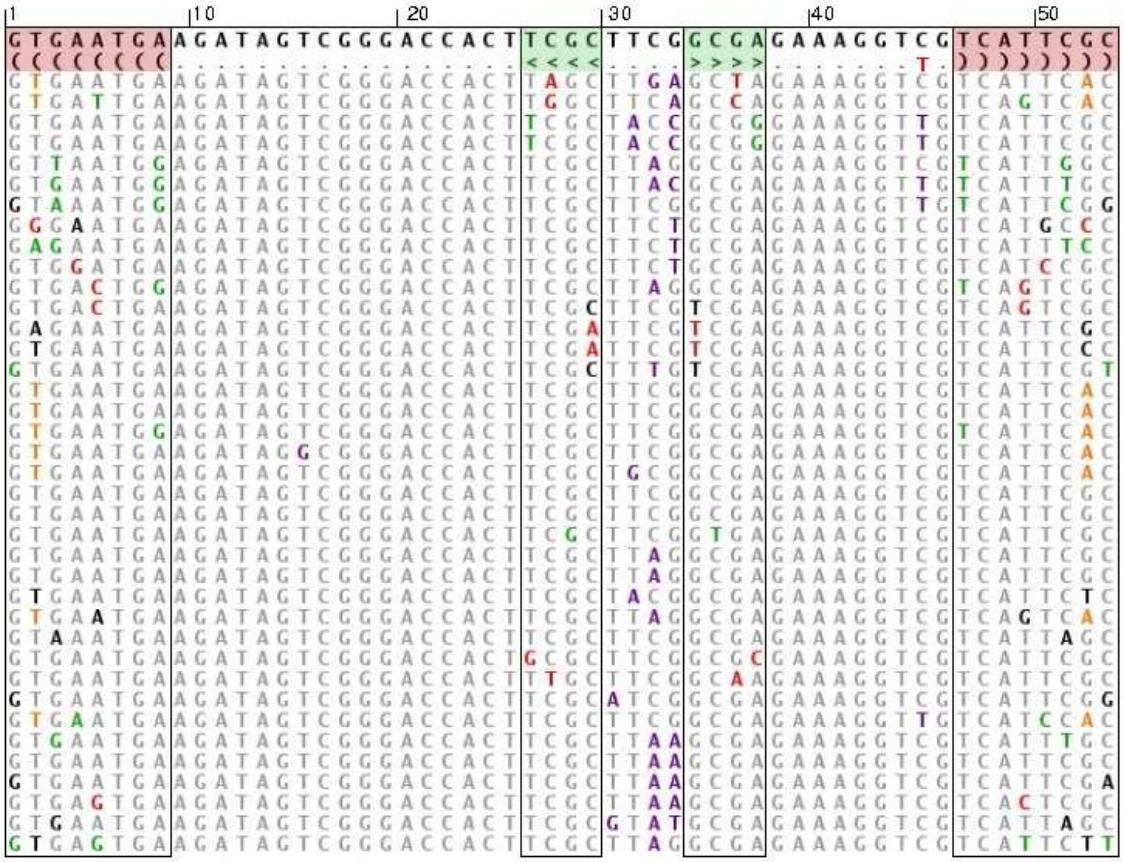


Figure 1I. Class II.

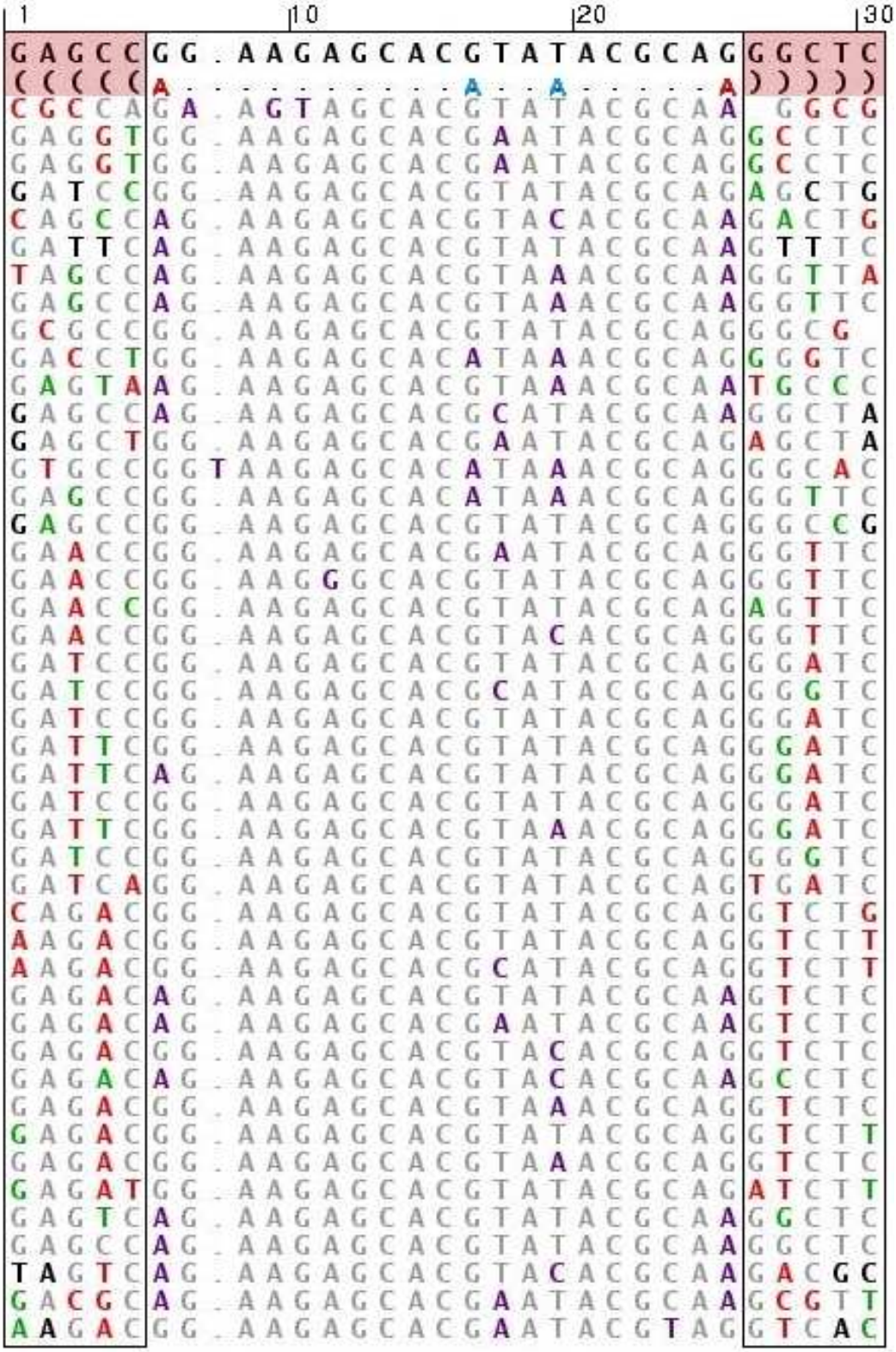


Figure 1J. Class IV.

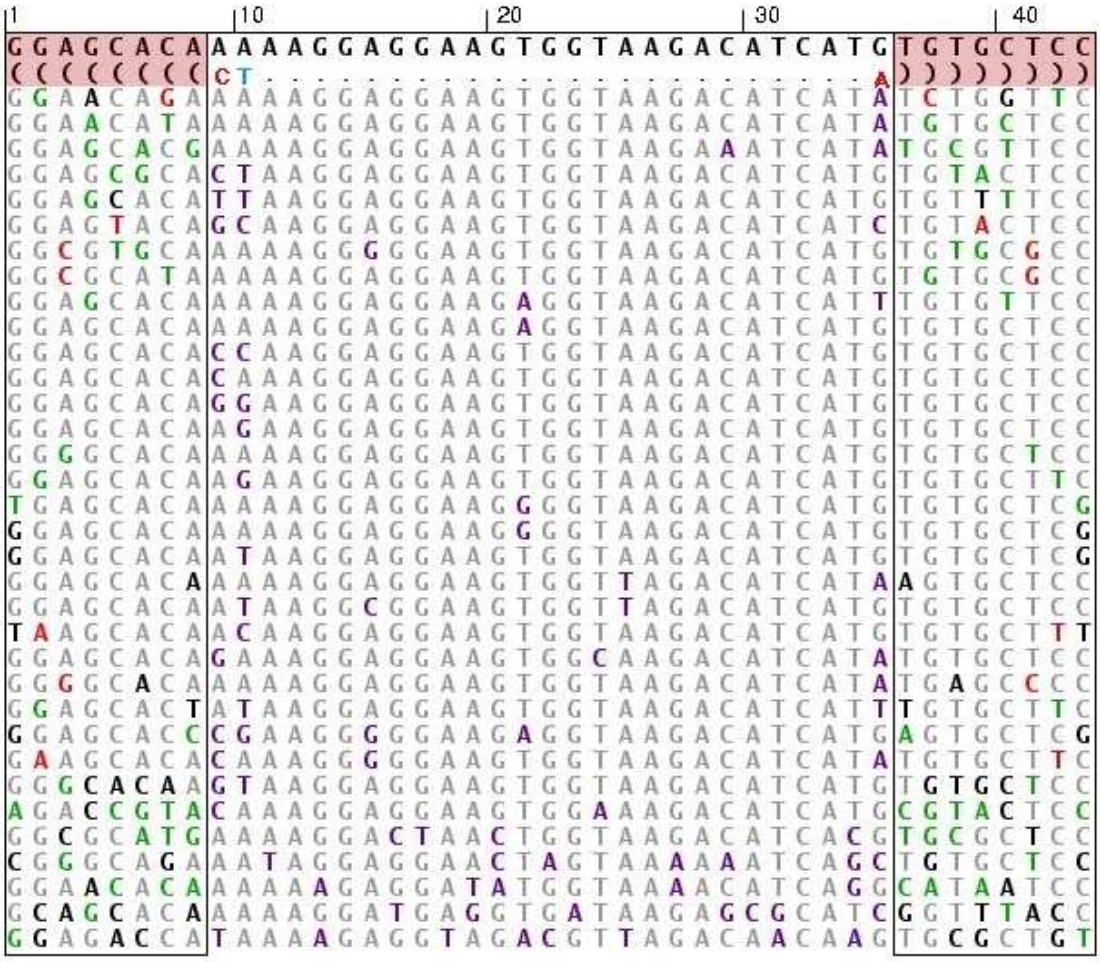
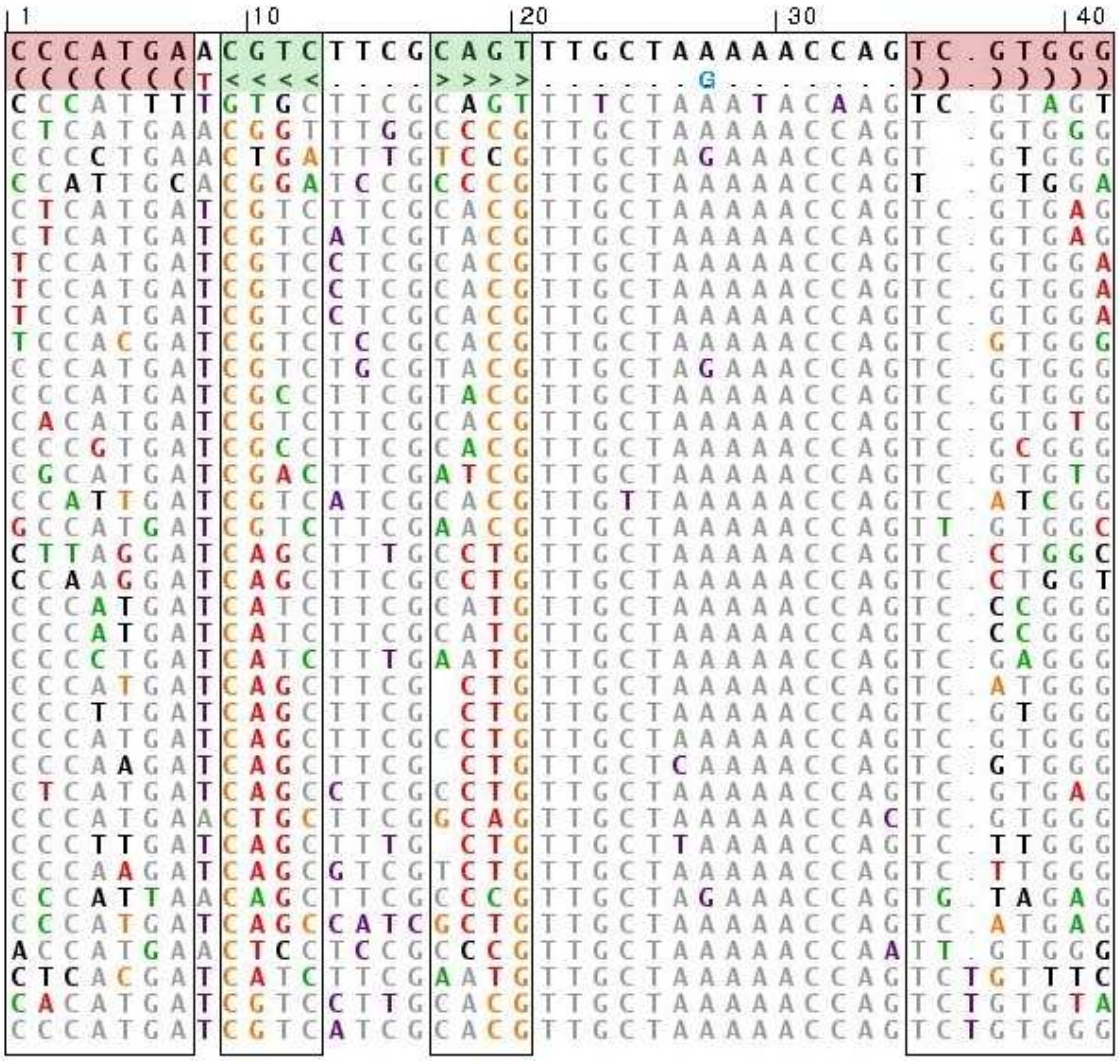


Figure 1K. Class III.



Supporting Chart 1. Original, minimized, and optimized sequences. The chart contains DNA sequences for each aptamer included in our analysis. The top in each group, marked 'ori', is a clone that was identified from the original GTP aptamer selection of Davis and Szostak[19]. Aptamers referred to by number indicate that only one isolate was identified. Aptamers named Class I-V originated multiple times as judged by the presence of diverse flanking sequences - these suggested 5' and 3' boundaries for the functional regions ('ori K_d construct'). For aptamers with multiple isolates (Class I-V), the specific original sequence chosen to derive the optimized versions is displayed. The 'min K_d construct' is the minimized functional sequence determined by deletion end mapping. 'min' is the minimized sequence flanked by constant primer regions and was employed as the template for selecting active sequence variants. The last sequence, 'opt', is the optimized version of each aptamer. K_ds are shown in parentheses. The constant primer-binding sites included for RT-PCR are shown in blue, where applicable. Sequence changes, relative to the original isolate, are shown in red. Occasionally unpaired G's were appended to the 5' ends of the 'min K_d construct' and 'opt' constructs to increase transcription yields; these are colored green.

The optimized sequences will be deposited in GenBank. All sequences, (including selection alignments) are available in FASTA format on our website:

http://genetics.mgh.harvard.edu/szostakweb/publications/extra/informationalcomplexitypaper2003.html

9-4

ori (25 nM)

5'-GGAGGCGCCAACTGAATGAAGTTGCCAGCTGCGAGCACGTGAATAGACTGCTTCGGCAGTGTCTCG
ACGTGTGTAGGGGAAAGTATCCTCCGTAACTAGTCGCGTCAC-3'

min (Kd construct 25 nM)

5'-GGAGGCGCCAACTGAATGAAGTTGCCAGCTGCGAGCACGTGAATAGACTGCTTCGGCAGTGTCTCG
ACGTGTGTAGGGGAAAGTATCCTCCGTAACTAGTCGCGTCAC-3'

min

5'-GGAGCACGAACTCGGTATCCGGAGGCGCCAACTGAATGAAGTTGCCAGCTGCGAGCACGTGAATAG
ACTGCTTCGGCAGTGTCTCGACGTGTGTAGGGGAAAGTATCCTCCGTAACTAGTCGCGTCACGCGAACC
ATTCGGAACATCG-3'

opt (9 nM)

5'-**GGGACG**AGCACGTGAAT**C**GACTGCTTCGGCAGTGTCTCGACGTGTGTAGGGGAAAGTATCC**C**C**CGTCC
C-**3'

Class V

ori (4000 nM)

5'-GGAGGCGCCAACTGAATGAAATTTGGGCATTTTGGTAGGTCGGTCGCTGCTTCGGCAGTAAGGGGT
AGGCATTGCTGGCCTAGGGTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 15000 nM)

5'-**GGG**GGGCATTTTGGTAGGTCGGTCGCTGCTTCGGCAGTAAGGGGTAGGCATTGCTGGCCTAGGGTC
CGTAAC-3'

min

5'-GGTATACGTGCAGAGACGCGTTGGGCATTTTGGTAGGTCGGTCGCTGCTTCGGCAGTAAGGGGTAG
GCATTGCTGGCCTAGGGTCCGTAACGCATGATAGCTGATCGCAGC-3'

opt (17 nM)

5'-**GG**GGGCATTTTGGTAGGTCGGTCGCTGCTTCGGCAGT**G**AGGGGTAGGCATTGCTGGCCTAGGGTCC
CC -3'

10-10

ori (30 nM)

5'-GGAGGCGCCAACTGAATGAATTGCTCAACAGTAGCCAACACGAGTACTGCTTCGGCAGTGGAACCAAC
GTAGTATGTTTAGCATTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 30 nM)

5'-**GG**TTGCTCAACAGTAGCCAACACGAGTACTGCTTCGGCAGTGGAACCAACGTAGTATGTTTAGCAT-
3'

min

5'-GGAGCACGAACTCGGTATCCTTGCTCAACAGTAGCCAACACGAGTACTGCTTCGGCAGTGGAACCAAC
GTAGTATGTTTAGCATGCGAACCATTCGGAACATCG-3'

opt (30 nM)

5'-**GG**GCTCAACAGTAGCCAACACGAGTACTGCTTCGGCAGTGGAACCAACGTAGTATGTT**G**AGC-3'

Class I

ori

5'-GGAGGCGCCAACTGAATGAAUUGCUUCGAGTCTTGAAGTGGTTGGGCTGCTTCGGCAGTGTGAAAATG
AGGCTTTTAAGGGTCCGTAACTAGTCGCGTCAC-3'

ori (K_d construct 160 nM)

5'-GGGAGTCTTGAAGTGGTTGGGCTGCTTCGGCAGTGTGAAAATGAGGCTTTT-3'

min (K_d construct 76 nM)

5'-GGGA**CG**AAGTGGTTGGGCGCTTCGGCGTGTGAAAA**CGTCTC**–3'

min

5'-GGAGCACGAACTCGGTATCCGGGA**CG**AAGTGGTTGGGCGCTTCGGCGTGTGAAAA**CGTCTC**GCGAACC
ATTCGGAACATCG–3'

opt (76 nM)

5'-GGGA**CG**AAGTGGTTGGGCGCTTCGGCGTGTGAAAA**CGTCTC**–3'

10-59

ori (285 nM)

5'-GGAGGCGCCAACTGAATGAACAGGATGGTAAGTTCCCAAGGCGGGTTGGAAGAGATATCATAGGAGCT
TGTCGTTCTGGTCATCCTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 285 nM)

5'-GGATGGTAAGTTCCCAAGGCGGGTTGGAAGAGATATCATAGGAGCTTGTCGT-3'

min

5'-GGAGCACGAACTCGGTATCCATGGTAAGTTCCCAAGGCGGGTTGGAAGAGATATCATAGGAGCTTGTC
GTGCGAACCATTCGGAACATCG-3'

opt (250 nM)

5'-**GGG**ATGGTAAGTTCCCAAGGCGGGTTGGAAGAGATATCATAGGAGCTTGTCGT**CCC**-3'

10-24

ori (500 nM)

5'-GGAGGCGCCAACTGAATGAATGATCCTTTGGAGGGGCATCTATATACTGCTTCGGCAGGGAACTCTAC
TAAGCACCGATGTCACTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 300 nM)

5'-**GC**GGGCATCTATATACTGCTTCGGCAGGGAA**C**CTCTACTAAGCACCGATGTC**CGC**-3'

min

5'-GGGCATGTCCTTTTTCCTAAT**GC**GGGCATCTATATACTGCTTCGGCAGGGAA**C**CTCTACTAAGCACCG
ATGTC**CGC**GTAATAGCAGCATACTCGGA-3'

opt (300 nM)

5'- **GC**GGGCATCTATATACTGCTTCGGCAGGGAA**C**CTCTA**A**TAAGCACCGATGTC**CGC**-3'

9-12

ori (300 nM)

5'-GGAGGCGCCAACTGAATGAACAAAGTTGTTAACACCCACTCTAGCAGTTCAGGTAACCACGTAAGATA
CGGTGCTAGAATGGGATCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 300 nM)

5'-**GGG**ACTCTAGCAGTTCAGGTAACCACGTAAGATACGGTGCTAGAAT-3'

min

5'-GGAGCACGAACTCGGTATCCACTCTAGCAGTTCAGGTAACCACGTAAGATACGGTGCTAGAATGCGAA
CCATTCGGAACATCG-3'

opt (300 nM)

5'- **GGG**ACTCTAGCAGTTCAGGTAACCACGTAAGATACGGTGCTAGA**G**T-3'

10-6

ori (500 nM)

5'-GGAGGCGCCAACTGAATGAAGATAGTCGGGACCACTTCCCTGAGTTCTGCTTCGGCAGATAACTCCAA
GGAGAAAGGTCGTCATTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 500 nM)

5'-**GCG**AATGAAGATAGTCGGGACCACTTCCCTGAGTTCTGCTTCGGCAGATAACTCCAAGGAGAAAGGTC
GTCATT**CGC**—3'

min

5'-GGAGCACGAACTCGGTATCC**GCG**AATGAAGATAGTCGGGACCACTTCCCTGAGTTCTGCTTCGGCAGA
TAACTCCAAGGAGAAAGGTCGTCATT**CGC**GCGAACCATTCGGAACATCG—3'

opt (300 nM)

5'-**GCG**AATGAAGATAGTCGGGACCACTTC**G**CT**TCGGC**GAGAAAGGTCGTCATT**CGC**-3'

Class II

ori

5'-GGAGGCGCCAACTGAATGAAGCTGCTAGCGCGGTAGAAAACCGAGCCGGAAGAGCACGTATACGCAGG
GCTCAACTACATTCCGTAACTAGTCGCGTCAC-3'

ori (K_d construct 4600 nM)

5'-**GGGACG**GCTGCTAGCGCGGTAGAAAACCGAGCCGGAAGAGCACGTATACGCAGGGCTCAACTACA-3'

min (K_d construct 2500 nM)

5'-**GGGACGCTG**AACGAGCCGGAAGAGCACGTATACGCAGGGCTCAACTA**G**-3'

min

5'-GGAGCCGACTAATGATTAAT**CTG**AACGAGCCGGAAGAGCACGTATACGCAGGGCTCAACTA**G**CGAGAG
TCCGCTAACATCGC-3'

opt (400 nM)

5'-**GG**GAGCC**A**GAAGAGCACGTATACGCA**A**GGCTC-3'

Class IV

ori (2400 nM)

5'-GGAGGCGCCAACTGAATGAAATTGACTGTTTGAGCATAAAAAGGAGGAAGTGGTAAGACATCATGTATGTTCACGTAACGGTTGTCCGTAACTAGTCGCGTCAC-3'

min (K_d construct 2400 nM)

5'-**GG**GTT**G**GAGCA**C**AAAAAGGAGGAAGTGGTAAGACATCATGT**G**TG**CTCCC**GT-3'

min

5'-GGACGAGAACTCGCATAAGCGTT**G**GAGCA**C**AAAAAGGAGGAAGTGGTAAGACATCATGT**G**TG**CTCCC**GTGCATGATAGCTGATCGCAGC-3'

opt (900 nM)

5'- GGGAGCACAAAAAGGAGGAAGTGGTAAGACATCATATGTGCTCC-3'

Class III

ori

5'–GGAGGCGCCAACTGAATGAACACACNCCTAAAAGGATACCCATGAACTGCTTCNGCAGTTTGCTAAAAACCACTCGTGGGTACNTTCCGTAACTAAGTCGCGTCAC-3'

ori (K_d construct 112000 nM)

5'-**GGG**ACACACCCTAAAAGGATACCCATGAACTGCTTCGCAGTTTGCTAAAAACCACTCGTGGGTACCTT-3'

min (K_d construct 112000 nM)

5'-**GGG**ACACACCCTAAAAGGATACCCATGAACTGCTTCGCAGTTTGCTAAAAACCACTCGTGGGTACCTT–3'

min

5'-GGAGCCGACTAATGATTAATACACACCCTAAAAGGATACCCATGAACTGCTTCGCAGTTTGCTAAAAACCACTCGTGGGTACCTTCCGTCCTACATCGGGCATTC-3'

opt (8000 nM)

5'- **GGG**ATGA**T**C**GT**CTTCGGA**CG**TTGCTAAAAACCA**G**TC**A**T**CCC**-3'