Supplementary Material (ESI) for Integrative Biology

Supplementary Information

Construction of Statistical Model

The statistical model was composed of discrete features describing each thrombin aptamer these included: overall proportion of A,T,C and G, proportion of every combination of dimer and trimer, length of each subsection of the aptamer (L1,L2,L3, length of the complimentary strands and length of each guanine stretch) and the base composition for each of these regions. Describing what constitutes the start of a loop and what constitutes part of a g-quartet when the length of a guanine repeat is greater than 2 is not easy and undoubtedly in reality an ensemble of structures will be formed on the array. In this instance we allow additional guanines to contribute both to the runs of Gs and the proximal loop composition in the model.

In addition to these features the minimum free energy of the sequences and minimum free energy of hybridization of the duplex regions (with interaction with central quadruplex prohibity) were also calculated using hybrid-ss-min from unafold[1].

Model =

0.0875 * length of sequence +
0.0955 * length of first run of G’s +
1.8028 * proportion of A’s within first complementary sequence +
2.5299 * proportion of G’s within first complementary sequence +
1.2166 * proportion of C’s within first complementary sequence +
1.9239 * proportion of T’s within first complementary sequence +
-0.1498 * length of L1 +
-0.124 * length of second run of G’s +
0.1103 * proportion of A’s within L1 +
1.1779 * proportion of G’s within L1 +
-0.1006 * proportion of C’s within L1 +
-0.1822 * length of L2 +
0.1239 * length of third run of G’s +
0.0147 * proportion of A’s within L2 +
0.3539 * proportion of G’s within L2 +
-0.3128 * proportion of C’s within L2 +
-0.197 * length of L3 +
-0.0206 * length of fourth run of G’s +
0.4909 * proportion of A’s within L3 +
0.4536 * proportion of G’s within L3 +
-0.135 * proportion of C’s within L3 +
-0.1365 * proportion of T’s within L3 +
-0.1421 * length of the second complementary region +
0.2183 * proportion of A’s within second complementary sequence +
0.9154 * proportion of G’s within second complementary sequence +
0.2513 * proportion of T’s within second complementary sequence +

0.5016 * spec +
-0.3805 * spec2 +
0.2513 * spec3 +
0.0469 * overall minimum free energy +
-0.0272 * minimum free energy contribution from duplex regions +
1.7933

Sequence – 5' -3'	score^a	percentile
TGGGAGTA**GG**TGT**GG** TGT**GG**T T**GG**GGCTCCCC	76.27	1
GTGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	66.56	2
GGGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GTCTCCCC	40.79	2
GGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC	29.21	5
GGTGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	26.69	5
GGGAGTA**GG**T T**GG**TTGT**GG**T T**GG**GGCTCCTC	21.82	5
GGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GGCTCCCC	14.33	5
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCC	13.18	5
GGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCC	11.55	5
GCGGAGTA**GG**T T**GG**TCGT**GG**T T**GG**GGCTCCCA	11.32	5
GGGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	8.97	10
AGTCGT**GG**TAT**GG** GCA**GG**T T**GG**GTGACT	6.28	10
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCCC	6.10	10
GGGGAGGTAG**GG**T T**GG** TGT**GG**T T**GG**GGCTCCCC*	6.09	10
GGGGAGTA**GG**A T**GG** TGT**GG**T T**GG**GGCTCCCC	5.24	10
TGAGAGTA**GG**C T**GG** TGT**GG**T T**GG**GGCTCCCC	4.68	10
GGGCGAGGTA**GG**T T**GG** TGT**GG**T T**GG**GGCTCCCC*	4.39	10
GGGGAGTA**GG**T T**GG** TGT**GG**ATT**GG**GCTCCGCC	4.25	10
GGGGAGGTA**GG**G T**GG** TGT**GG**T T**GG**GGCTCCCC*	3.98	10
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC	3.94	10
GGGAGAGTA**GG**T T**GG**TGTG**GG**T T**GG**GGCTCCCC	3.92	10
	score^b	percentile
GTGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	1.91	2
GGTGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	1.89	5
GGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GGCTCCCC	1.75	5
GGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC	1.72	5
TGGGAGTA**GG**TGT**GG** TGT**GG**T T**GG**GGCTCCCC	1.63	5
GGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCC	1.47	5
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCCC	1.41	5
GGGAGTA**GG**T T**GG**TTGT**GG**T T**GG**GGCTCCTC	1.40	5
GGGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GTCTCCCC	1.39	5
CGGGGAGTA**GG**T T**GG** TGT**GG**TCT**GG**GGCTCCCC	1.27	10
GGGGAGTA**GG**C T**GG** TGT**GG**T T**GG**GGCTCCCC	1.23	10
AGTCGT**GG**A G**GG**C A**GG**T T**GG**GGTGACT	1.23	10
GGGGAGTA**GG**A T**GG** TGT**GG**T T**GG**GGCTCCCC	1.22	10
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC	1.21	10
GGGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GGCTCCCC	1.20	10
GGGGAGGTAG**GG**T T**GG** TGT**GG**T T**GG**GGCTCCCC*	1.19	10
GGGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC	1.19	10
GGGGAGTA**GG**T T**GG** TGT**GG**ATT**GG**GGCTCCCC	1.19	10
GGGGAGTA**GG**T T**GG**GTAT**GG**T T**GG**GGCTCCCC	1.18	10
GCGGAGTA**GG**T T**GG**TCGT**GG**T T**GG**GGCTCCCA	1.17	10
GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCC	1.15	10
TGAGAGTA**GG**C T**GG** TGT**GG**T T**GG**GGCTCCCC	1.15	10
GGGAGAGTA**GG**T T**GG**TGTG**GG**T T**GG**GGCTCCCC	1.15	10
GGGGAGTA**GG**T T**GG**ATGT**GG**T T**GG**GCTCCCC	1.15	10
GGGGAGTA**GG**T T**GG**TCGT**GG**T T**GG**GTGCTCCCC	1.13	10
GGGGAGTA**GG**T T**GG**TTGT**GG**T T**GG**GGCTCCCC	1.12	10
GGGGAGATA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCCC	1.11	10
GGGGAGTA**GG**T T**GG**ACGT**GG**T T**GG**GCTCCCC	1.10	10
GGGGAGTA**GG**T T**GG**TGTG**GG**T G**GG**GCTCCCC	1.10	10
GGGGAGTA**GG**T T**GG**TCGT**GG**T T**GG**GGCTCCCC	1.10	10
AGTCCGT**GG**A G**GG**C A**GG**T T**GG**GTGACT	1.10	10

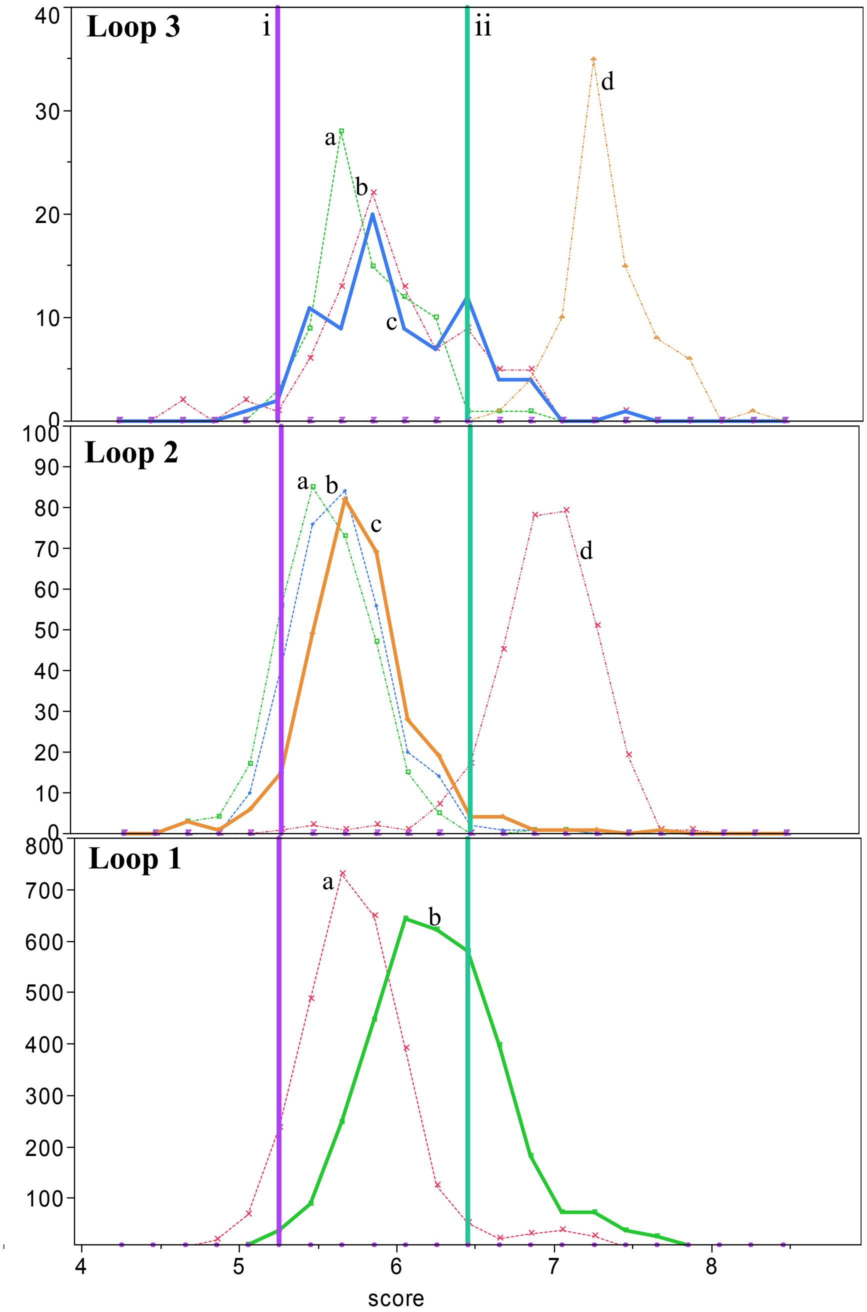
Table 1 Alignments of sequences with top ten 10% of scorers which vary from the GGT(T/A)GGN3GGT(T/A)GG motif. *indicates more than one way of forming the structure. Scores a) th-Bio b) th-cy5.

Correlation between Cy5 and Biotin scores

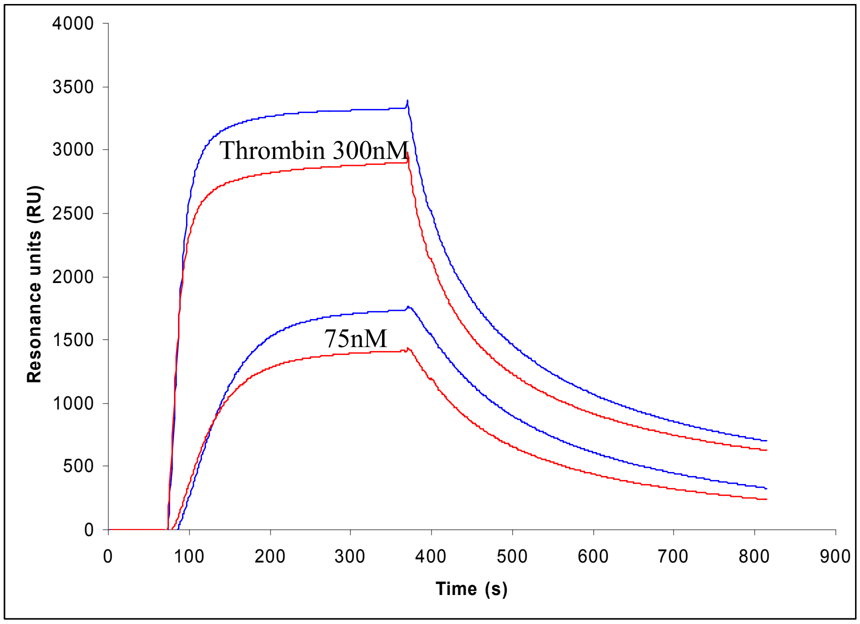
	th-Cy5 chip1	th-Cy5 chip2	th-Bio chip1	th-Bio chip2
th-Cy5 chip1		0.881	0.829	0.784
th-Cy5 chip2	0.881		0.864	0.826
th-Bio chip1	0.829	0.864		0.935
th-Bio chip2	0.784	0.826	0.935	

Table 2. Correlation coefficients for hybridizations performed on THA.

1. R. A. Dimitrov and M. Zuker, *Biophys. J.*, 2004, **87**, 215-226.
2. D. M. Tasset, M. F. Kubik and W. Steiner, *Journal of Molecular Biology*, 1997, **272**, 688-698.



Supplementary Figure 1. (Reprint Figure 4 main text) Relative scores of sequences that represent key aspects detected within the model, solid vertical line *i*, represents the average score of known thrombin binders[29] with Kd's 30, 42, and 126 nm, vertical line *ii* average scores for known thrombin binders with Kd's 0.5, 0.7, 0.9 nm Loop 1) a- dashed line (RED), sequences that do not contain a guanine residues. b- solid line (GREEN) represents sequences that contain at least one G within loop one, Sequence for loop 2 remains fixed. Loop2) a – dot dashed line (GREEN) represents loop 2 sequence TAT, b – dashed line (BLUE) loop 2 sequence of TTG, c – solid line (ORANGE) loop 2 sequence CAG, d – dot dot dashed line (RED) loop 2 sequence TAG, loops 1 and 3 remain fixed at TT. Loop 3) a- dashed line (GREEN) loop 3 sequence GG, b – dot dashed line (RED) loop 3 sequence AA, c – solid line (BLUE) loop 3 sequence TT, d – dot dot dashed line (ORANGE) loop3 sequence TA. N.B. All permutations' of loop sequences were analyzed, shown above is representative subset of data. (color version available in supplementary data)



Supplementary Figure 2. Sensograms from the original aptamer ThB (RED), and the top scoring mutated aptamer (BLUE), as determined by SPR. Thrombin at concentrations of 1µM to 10nM (300, and 75nM shown here) was injected onto the aptamer-immobilized SA chip under running buffer at a flow of 25 µl/min. Measured KD's for ThB and mutated sequence are 26 and 28 nM.