Efficient HIV-1 inhibition by a 16 nt-long RNA aptamer designed by combining *in vitro* selection and *in silico* optimisation strategies

Francisco J. Sánchez-Luque[a,1], Michael Stich[b,2], Susanna Manrubia[b], Carlos Briones[b,c,*] and Alfredo Berzal-Herranz[a,*]

[a]Department of Molecular Biology, Instituto de Parasitología y Biomedicina "López-Neyra" (IPBLN-CSIC) - PTS Granada, Armilla, Granada, 18016, Spain.
[b]Department of Molecular Evolution, Centro de Astrobiología (CAB-CSIC/INTA), Torrejón de Ardoz, Madrid, 28850, Spain.
[c]Centro de Investigación Biomédica en Red de enfermedades hepáticas y digestivas (CIBERehd), Spain.

* To whom correspondence should be addressed. Alfredo Berzal-Herranz; Tel: +34958181648; Fax: +34958181632; aberzalh@ipb.csic.es. Correspondence may also be addressed to Carlos Briones; Tel: +34915206411; Fax: +34915201074; cbriones@cab.inta-csic.es.

[1]Present Address: Genome Plasticity and Disease Group. Mater Medical Research Institute-University of Queensland. Level 4, TRI Building. 37 Kent st, Woolloongabba (QLD 4102, Australia).

[2]Present Address: Michael Stich, Non-linearity and Complexity Research Group (NCRG), School of Engineering and Applied Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, UK.

Supplementary Information

Supplementary table S1. Repeated sequences found along the *in vitro* selection process, MFE of each of those sequences and group (see main text) to which the sequences can be assigned.

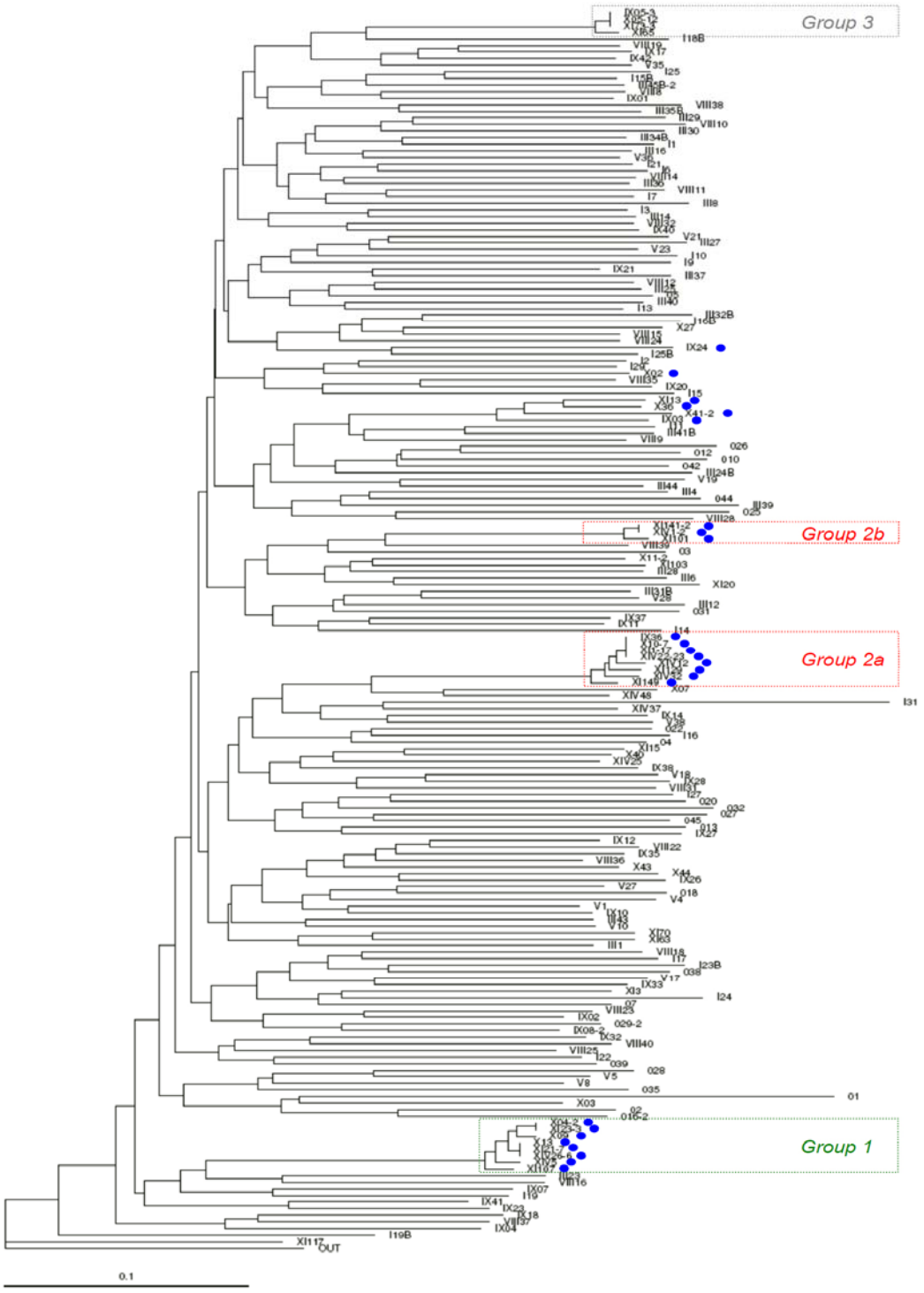
0	I	III	V	VIII	IX	X	XI	XIV	ΔG of the MFE (Kcal/mol)	Group
016-2									-10.70	
029-2									-8.60	
		III45B-2							-5.10	
					IX05-3	X05-12	XI73-3		-4.03	Gr. 3
					IX08-2				-11.10	
					IX36	**X10-7**	**XI1-17**	**XIV22-23**	**-10.80**	**Gr. 2a**
						X04-2	**XI23-3**		**-20.20**	**Gr. 1**
						X11-2			-6.61	
							XI21-7	**XIV26-6**	**-15.70**	**Gr. 1**
							XI141-2	**XIV1-2**	**-11.80**	**Gr. 2b**
						X41-2			**-15.20**	

Notes: Equal sequences are shown on the same row; repeated sequences within each round are noted as 'Seq-R', R being the number of repetitions; sequences including the octanucleotide 5'-GGCARGGA-3' are shown in boldface.

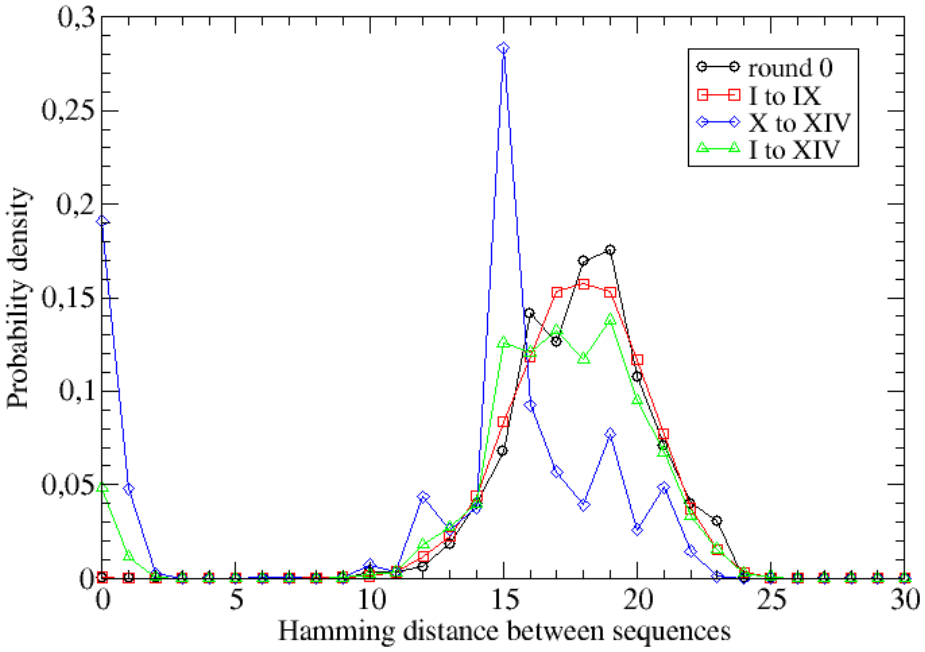
Supplementary table S2. Thermodynamic properties of the aptamer:target binding sites depicted in Fig. S7. Nucleotides of the aptamer not complementary to the binding sequence of the target (both canonical and wobble G-U pairs are allowed) are shown in lowercase letters.

Aptamer molecule	Interaction region (target)	Interacting positions (target)	Interacting positions (aptamer)	Binding sequence (5'-3', target)	Binding sequence (5'-3', aptamer)	Binding energy (Kcal/mol)
XIV22-23	(a)	81-86	24-29	CUUGCC	GGCAAG	-9.43
	(a')	79-84	55-60	AGCUUG	CGAGCU	-7.48
	(b)	138-145	1-8	GAGAUCCC	GGGAaUUC	-5.55
	(c)	249-254	24-29	CUUGCU	GGCAAG	-5.98
	(d)	290-293	11-14	GGUG	CACC	-6.24
XIV1-2	(a)	81-86	23-28	CUUGCC	GGCAAG	-9.43
	(a')	79-84	55-60	AGCUUG	CGAGCU	-7.32
	(b)	138-145	1-8	GAGAUCCC	GGGAaUUC	-5.16
	(c)	249-254	24-29	CUUGCU	GGCAAG	-5.98
	(e)	117-123	10-16	GUUGUGU	ACACAAC	-6.23
XIV26-6	(a)	81-86	15-20	CUUGCC	GGCAAG	-9.73
	(b)	138-145	1-8	GAGAUCCC	GGGAaUUC	-6.67
	(c)	249-255	15-20	CUUGCU	GGCAAG	-6.31
RNApt16	(a)	81-86	5-10	CUUGCC	GGCAAG	-9.73
	(c)	249-254	5-10	CUUGCC	GGCAAG	-6.08
	(f)	151-161	5-15	CCCUUUUAGUC	GGCaAGGAGGG	-4.52

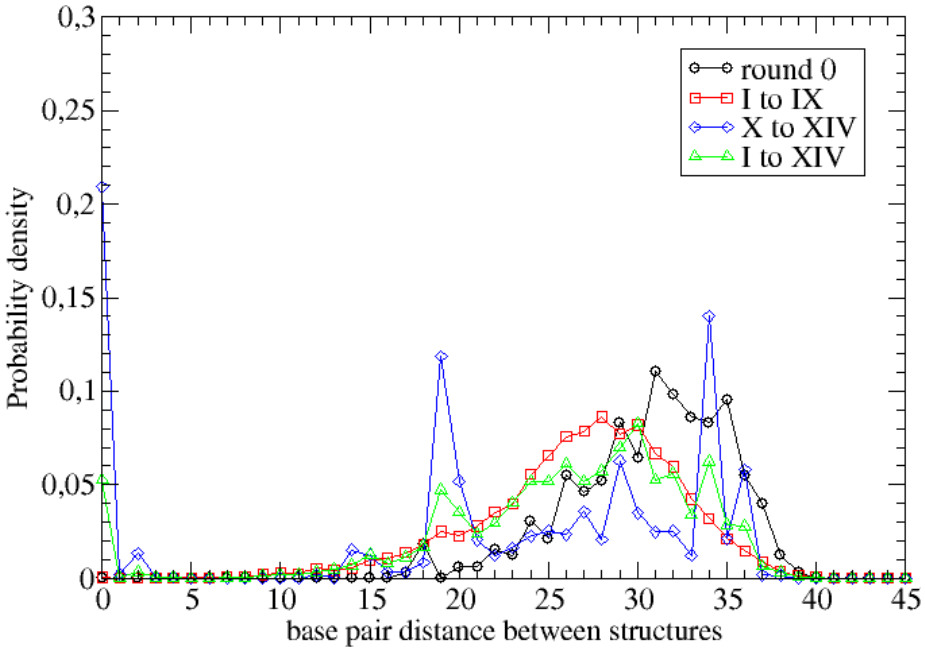
Supplementary figure S1. Clustering analysis of the 188 different, 64 nt-long sequences obtained through the *in vitro* selection process. Groups 1 to 3 of related sequences (with mutual Hamming distances not exceeding 0.03 mutations per nucleotide) are shown with rectangles. Sequences showing the conserved motif 5'-GGCAAGGA-3' are marked with blue dots. The branch lengths are proportional to the Hamming distance between the sequences (scale bar: 0.1 mutations per nucleotide).



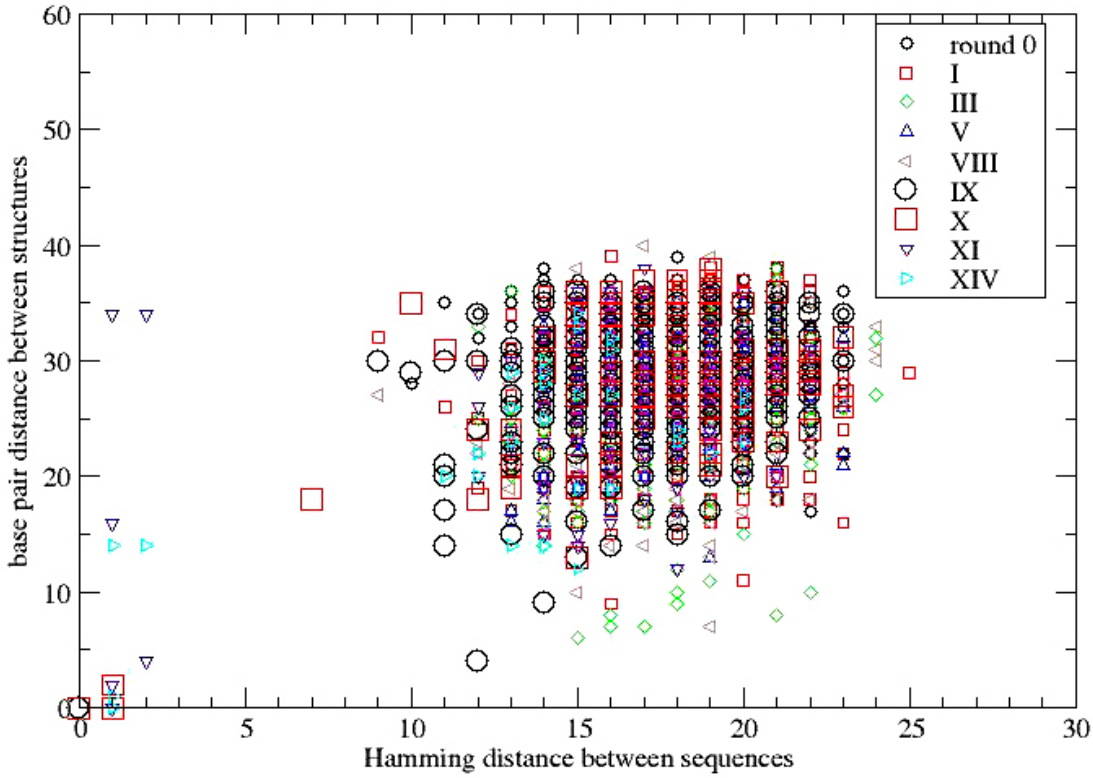
Supplementary figure S2. Normalized distributions of the Hamming distances between sequences of round 0, rounds I to IX (accumulated in one set), and rounds X to XIV (accumulated in one set). The distribution of the Hamming distances between all pairs of sequences selected along the *in vitro* process (rounds I to XIV) is also accumulated in one set. The appearance of peaks in the distribution signals the formation of groups of similar sequences. The height of the peak at zero is proportional to the frequency of identical sequences.



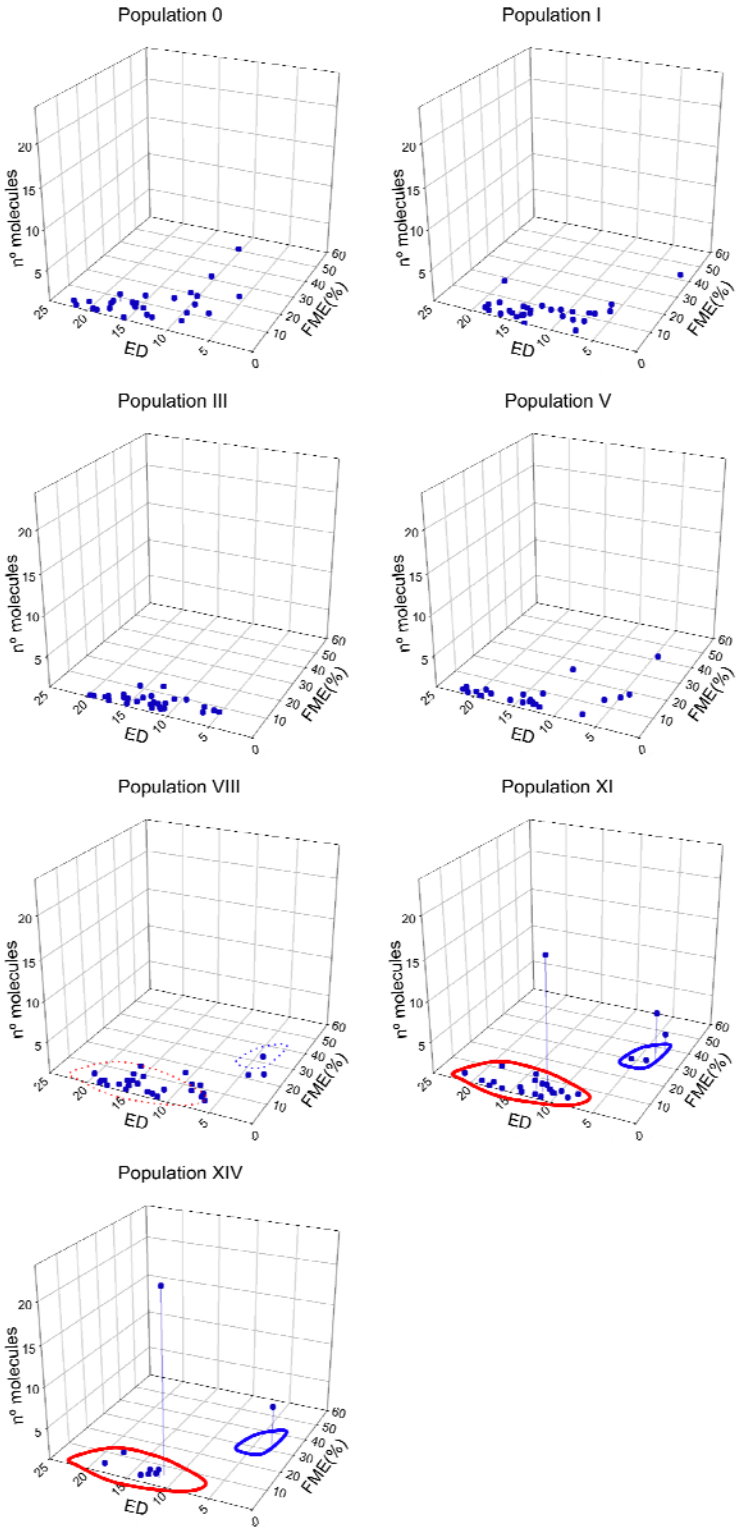
Supplementary figure S3. Normalized distributions of base pair distances between structures of round 0, rounds I to IX (accumulated in one set), and rounds X to XIV (accumulated in one set). The distribution of the base pair distances between all sequences selected along the *in vitro* process (rounds I to XIV) is also accumulated in one set. The appearance of peaks in the distribution signals the formation of groups of similar structures. The height of the peak at zero is proportional to the frequency of identical structures.



Supplementary figure S4. Scatter plot representing the sequence (Hamming) distance *versus* the structure (base pair) distance between pairs of sequences. All the 188 different, 64 nt-long RNA molecules obtained along the *in vitro* selection process are included, yielding a total of 17,578 pairs (points in the plot). The big cloud of points corresponds to poorly related sequences with both large Hamming and base pair distance between each other. More interesting are the few points showing low Hamming distance: i) identical sequences (at a zero Hamming distance, thus at a zero base pair distance) appear in rounds IX and X; ii) sequences at Hamming distance 1 and identical structure are found in rounds X, XI and XIV; and iii) sequences with Hamming distance 1 or 2 and larger differences in base pair distance appear in rounds XI and XIV. Additional interesting examples are found in round IX, where sequences with relatively low Hamming distances (9 to 12) appear. Also, a pair of sequences (X36 and X41-2) in round X presents intermediate Hamming and base pair distances (7 and 18, respectively).



Supplementary figure S5. Evolution of two thermodynamic parameters of the selected sequences along the *in vitro* selection process: ensemble diversity (ED) and frequency of the MFE structure (FME) in the thermodynamic ensemble. The number of repeated sequences is shown in the *z*-axis. Groups of sequences (termed '1' and '2' and encircled with blue and red lines, respectively) appear from population VIII on.



Supplementary figure S6. Clustering analysis of the structures of the 188 different, 64 nt-long sequences obtained through the *in vitro* selection process, based on their base pair distances. Groups 1 to 3 of related sequences (with mutual base pair distances not exceeding 4 base pairs) are shown with rectangles, and their common MFE secondary structure is depicted. Sequences showing the conserved octamer 5'-GGCAAGGA-3' are marked with blue dots. The branch lengths are proportional to the base pair distance among the structures (scale bar: 1 base pair).

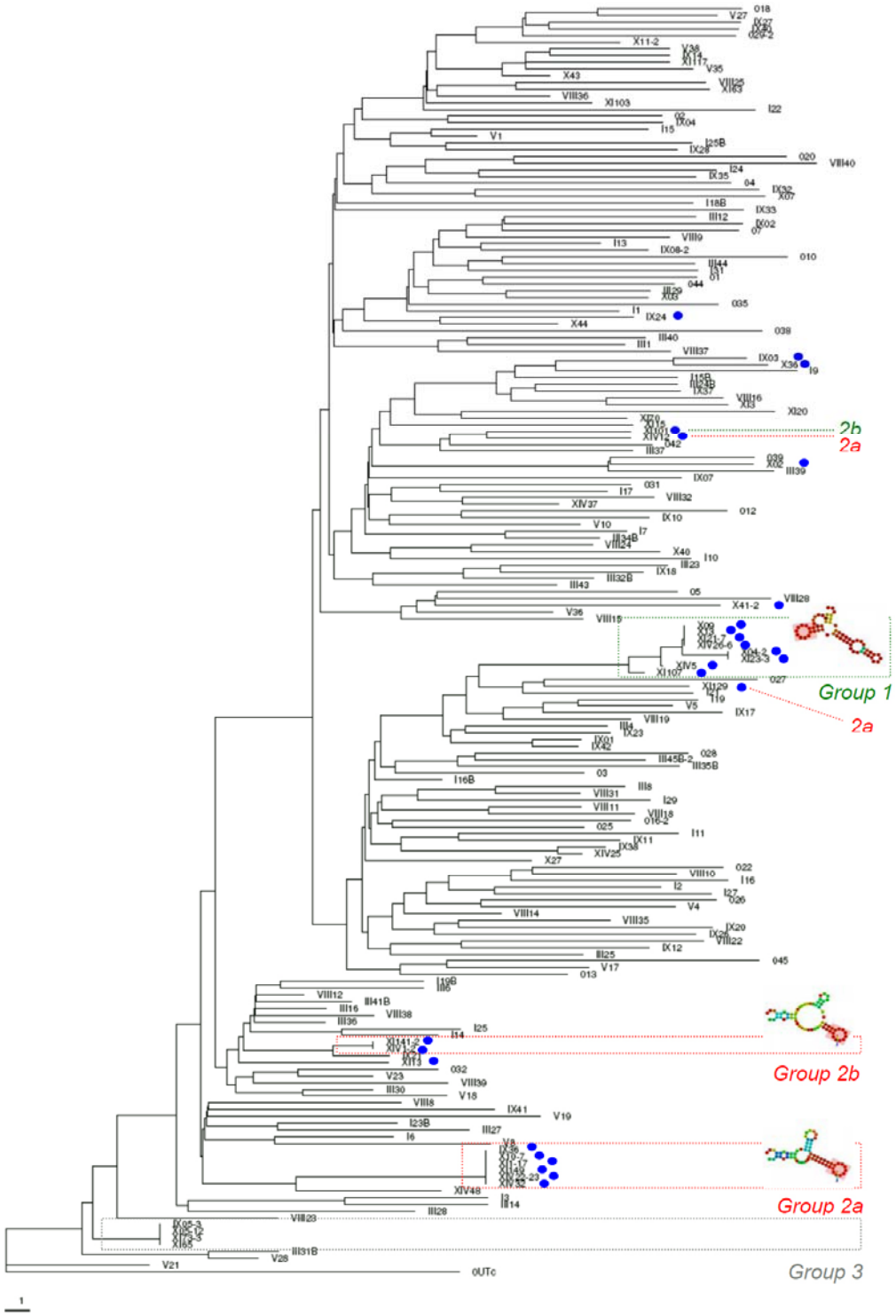
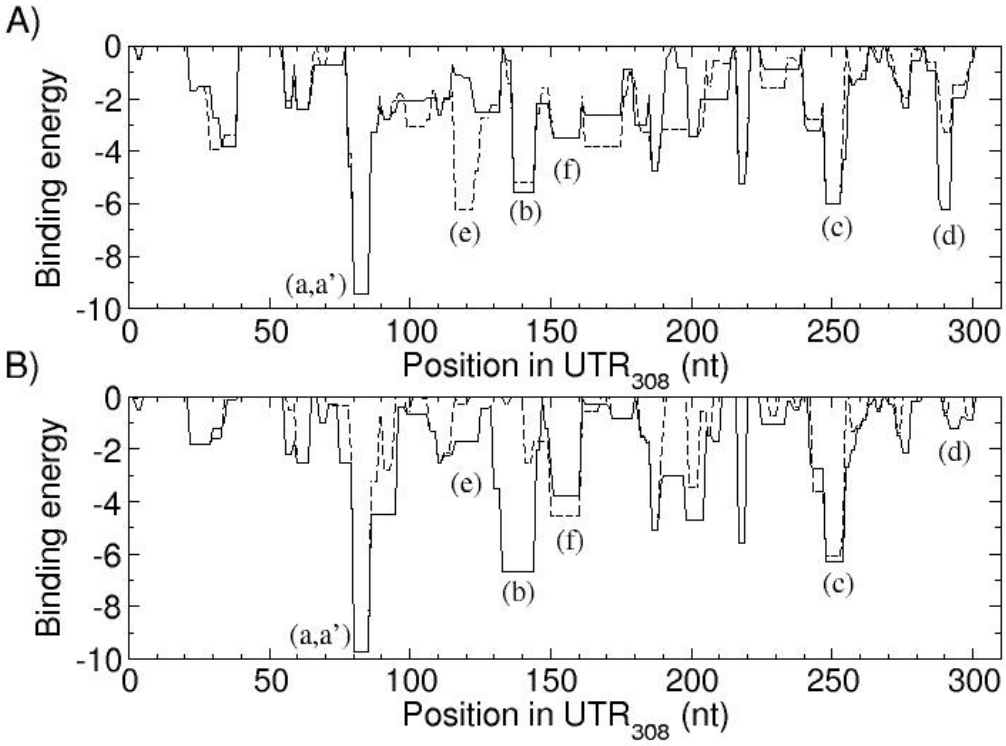
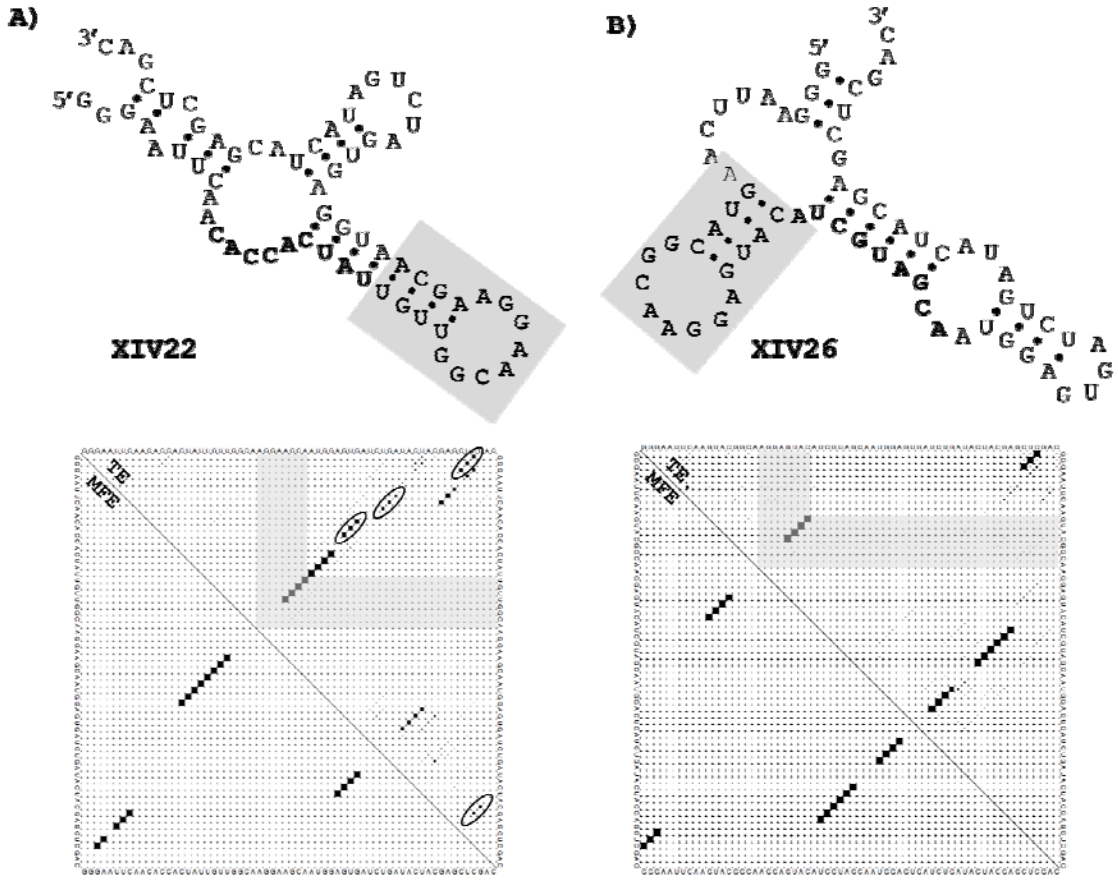


Figure S7. *In silico* prediction of the interaction between the HIV-1 target molecule UTR_{308} and the selected aptamers. Overall local (position wise) binding energy (in Kcal/mol) for the interaction between the target (position of nt on abscissa) and the selected molecules: XIV22-23 (solid line) and XIV1-2 (dashed line) (A), XIV26-6 (solid line) and RNApt16 (dashed line) (B). Regions (a,a') to (f) are the putative binding sites. Useful information can be extracted if we focus (i) on the sites with very high binding energy, and (ii) on the sites where all four curves coincide, in particular, if (iii) the aptamer-target pairing interval is stable enough and does not contain mismatches. Further information can be found on Table S2.



Supplementary figure S8. Thermodynamic Ensemble (TE) *vs* Minimum Free Energy (MFE) structure of aptamers XIV22 **(A)** and XIV26 **(B)**. Dot plots represent possible base pairs of the given sequence. The lower left triangle of the dot plot depicts the MFE structure explicitly represented on top. The upper right triangle depicts the TE, where the intensity of each dot is proportional to the probability of finding such a base pair within the ensemble of structures. In XIV22, some stems relatively abundant in the TE but not present in the MFE structure are encircled. In turn, the similarity between the XIV26 MFE and TE is indicative of the stability of the depicted secondary structure. In both cases, the shadowed region represents the 16 nt-long motif common to most of the aptamers.



Supplementary figure S9. Sequence alignment of a selected region of the 5'UTR of HIV-1 and simian immunodeficiency virus (SIV). The predicted binding site (a) for the aptamers is highly conserved in all HIV-1 groups and subtypes, and even in some SIV strains (red box). Both, the polyadenylation sequence motif 5'-AATAAA-3' and the aptamer binding site (a) within the poly(A) domain are indicated in the figure. Adapted from Kuiken C., *et al.* 2011.