

## Electronic Supplementary Information

### **Aptamers targeting protein-specific glycosylation in tumor biomarkers: general selection, characterization and structural modeling**

Ana Díaz-Fernández<sup>1,2</sup>, Rebeca Miranda-Castro<sup>1,2</sup>, Natalia Díaz<sup>1</sup>, Dimas Suárez<sup>1</sup>, Noemí de-los-Santos-Álvarez<sup>1,2</sup> and M. Jesús Lobo-Castañón<sup>1,2</sup> \*

<sup>1</sup> Departamento de Química Física y Analítica. Universidad de Oviedo, Av. Julián Clavería 8, 33006 Oviedo –Spain.

<sup>2</sup> Instituto de Investigación Sanitaria del Principado de Asturias, Avenida de Roma, 33011 Oviedo- Spain

\* Corresponding author [mjlc@uniovi.es](mailto:mjlc@uniovi.es)

## Table of contents

1	Materials and Reagents.....	4
2	Protocols.....	6
2.1	Immobilization of proteins on magnetic particles for SELEX.....	6
2.2	Bradford assay.....	6
2.3	SELEX procedure.....	7
2.4	UV-VIS enrichment assays.....	9
2.5	Remelting study.....	10
2.6	SPR binding curves.....	10
2.7	Electrochemical measurements.....	11
2.8	Circular dichroism.....	12
2.9	UV-VIS thermal melting curves.....	12
2.10	Deglycosylation study.....	12
2.11	Statistical analysis.....	13
3	Detailed computational settings.....	13
3.1	Aptamer model building.....	13
3.1.1	<i>Initial structures of the PSAG-1 aptamer.....</i>	<i>14</i>
3.1.2	<i>Enhanced Molecular Dynamics of the PSAG-1 models.....</i>	<i>15</i>
3.1.3	<i>Structural characterization and free energy reweighting of the enhanced MD simulations.....</i>	<i>17</i>
3.1.4	<i>Conventional Molecular Dynamics.....</i>	<i>19</i>
3.1.5	<i>MM-PBSA calculations.....</i>	<i>20</i>

3.2	Rigid docking calculations.....	21
3.3	Molecular dynamics simulation of the hPSA/T-PSAG-1 complex .....	24
4	Tables.....	27
5	Schemes.....	33
6	Figures .....	35
7	References .....	51

## 1 Materials and Reagents

All DNA sequences were purchased from Metabion (Germany) as lyophilized powder with HPLC purification except the library, which was obtained from Sigma. All individual aptamers tested were acquired with a T5 extension followed by a fluorescein tag at 5' terminus. Prostate Specific Antigen (PSA) from human seminal fluid (P117-7) was purchased from BBI Solutions (UK) and used for the selection. PSA Certified Reference Material BCR®-613 was obtained from Sigma-Aldrich (Spain) and used for aptamers characterization. Recombinant Human PSA (rPSA) from *Escherichia coli* (ab126692) and Recombinant Human Lipocalin-2 (NGAL) from HEK 293 cells (ab167728) were obtained from Abcam (UK). Human alpha-fetoprotein (AFP) purified from pooled human cord serum (PRO-406) was purchased from Prospec (Israel). Lectin PhoSL from the mushroom *Pholiota squarrosa* was a kind gift from Prof. Rafael de Llorens. The protein can also be obtained in a recombinant way<sup>1</sup> or by chemical synthesis with the following sequence:

(NH<sub>2</sub>-APVPVTKLVCDGDTYKCTAYLDFGDGRWVAQWDTNVFHTG-OH)<sup>2</sup>,

maintaining the recognition properties. Transferrin human powder was obtained from Sigma Dynabeads™ MyOne™ Streptavidin C1 (1 μm diameter) and M-280 tosylactivated magnetic beads (2.8 μm diameter) were purchased from Thermo Fisher Scientific (Spain). (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, NaH<sub>2</sub>PO<sub>4</sub>, Na<sub>2</sub>HPO<sub>4</sub>, NaCl, KCl, phosphate buffer 10× (PBS pH 7.4), Tris-HCl pH 7.4 buffer 10×, 1 M NaOH solution, 1 M MgCl<sub>2</sub> solution, bovine serum albumin (BSA) and t-RNA were supplied by Sigma-Aldrich (Spain) as molecular biology grade reagents. L-(-)-fucose 99% pure grade, ethanolamine, 3,3',5,5'-tetramethylbenzidine (TMB) liquid substrate, N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride (EDC), N-hydroxysuccinimide (NHS), anti-fluorescein-Peroxidase Fab fragment, Tween-20, 11-mercaptopundecanoic acid and

mercaptohexanol were also purchased from Sigma (Spain).  $K_2HPO_4$  and  $KH_2PO_4$  salts were obtained from Merck (Germany). Immolase DNA polymerase, buffer, dNTPs, and magnesium salt for PCR were obtained from Bioline (UK). The 5× TBE (0.45 M TrisBorate and 0.01 M EDTA, pH 8.3) was purchased from 5 Prime (USA), and the 6× DNA gel loading buffer was acquired from Novagen (USA). ROX and SYBRGreen Reference Dyes were obtained from Invitrogen. All aqueous solutions were prepared using ultra-pure water from a MilliQ system (Millipore, Spain).

The composition of buffers used was the following:

- PBS 1× also named as PBS-Na: 0.01M phosphate pH 7.4+ 0.154 M NaCl
- PBS-K: 0.01M phosphate pH 7.4 + 0.154 M KCl
- Buffers for tosylactivated-MP modification: (BM): 19 mM  $NaH_2PO_4$ , 81 mM  $Na_2HPO_4$  pH 7.4 and (BM2): 3 M  $(NH_4)_2SO_4$  pH 7.4
- Washing buffers for tosylactivated-MP modification: (BLmod): PBS 1× + 0.5 % BSA and (BLmod2): PBS 1× + 0.1 % BSA
- Selection buffer (BS): PBS 1× pH 7.4
- Washing buffer for selection (BSL): BS+ 0.01 % tween-20
- Buffers for streptavidin-MPs modification (BLstrep): 10 mM Tris-HCl, 2 M NaCl + 0.01 % tween-20
- Acetate buffer (NaAc buffer): 10 mM NaAc pH 5.5

The sequences (5' → 3') of the ssDNA used were the following:

- ssDNA library:  
AGGGTTGATAGGTTAAGAGC(N)<sub>40</sub>CGATGTCAACTAGCTGTTGG
- Direct primer: AGGGTTGATAGGTTAAGAGC
- Reverse primer: Biotin-CCAACAGCTAGTTGACATCG

- Unrelated ssDNA sequence:  
AGGGTTGATAGGTTAAGAGCCTAGGCGAAATATAGCTACAACCTGTCT  
GAAGGCACCCAATCGATGTCAACTAGCTGTTGG
- PSAG-1: GAGCGGGGTTGCTGGGATGATAAGGCCCTTTGATGTCTG
- T-PSAG-1: GAGCGGGGTTGCTGGGATGATAAGGCCCTTT
- PSAG-2: GGACGGCTCTGTTATAAGTGACAGATCTGGACGTAACATT
- PSAG-3:  
CAGCTATCACGATGAGCCTTGCACTAGGGTTGATAGGTTAAGAG
- PSA-1: GGACGGTTGCGCTATATTTAACCAAAGTCTGGATTAACA
- Scrambled ssDNA:  
GGTCCTGGGAGACGAACAACCCCGTCACAAGCCTCGTAG

## 2 Protocols

### 2.1 Immobilization of proteins on magnetic particles for SELEX

5 mg of Dynabeads M-280 tosylactivated magnetic particles were washed with 1 mL of BM and resuspended in 100 µg of the protein (hPSA, rPSA or BSA) in BM and 100 µL of BM2 and incubated for 12-18 h at 37 °C at 1300 rpm. Then the supernatant was collected and the particles incubated with 1 mL of BLmod for 1 h at 37 °C at 1300 rpm, washed twice with 1 mL of BLmod2, resuspended in 250 µL of BLmod2 and stored at 4 °C. The amount of protein immobilized onto the surface of the particles was quantified using Bradford assay.

### 2.2 Bradford assay

100 µL of buffer BM/BM2 3:2 (blank), standard BSA solutions from 0.1 µg/ µL to 1.0 µg/ µL prepared in the same buffer and the supernatant from protein-MPs interaction were pipetted into test tubes. Next 5 mL of Bradford Reagent 1× was added to each tube

and incubated for 5 min at RT and the absorbance measured at 595 nm in a Genesys 10 S UV-VIS (Thermo Fisher). The concentration of the protein immobilized into the surface is the difference between the initial concentration of the protein before the incubation (0.4 µg/ µL) and the concentration found in the supernatant.

### 2.3 SELEX procedure

The initial ssDNA library was synthesized by Sigma-Life Science. It consists of a collection of 80-nt oligonucleotides, which was purified by PAGE. Each oligonucleotide contains a central randomized 40-nt sequence, flanked by two fixed regions of 20-nt:

5'-AGGGTTGATAGGTTAAGAGC(N)<sub>40</sub>CGATGTCAACTAGCTGTTGG-3'.

The nucleotide distribution in the library was determined by high throughput sequencing as 27.4% dA, 27.7% dT; 23.7 % dC and 21.2% dG.

This library was subjected to a thermal shock in BS buffer (98 °C, 4 min) and immediately cooled in ice for 4 min before adding BSA-MPs (negative selection) in a 10:1 (protein:DNA) molar ratio. 1.23 µg/mL of BSA, to minimize the unspecific binding, and 1.23 µg/mL of tRNA, as a competitor keeping the molar ratio [tRNA]/[DNA] at 0.1, were also added to this solution. After the interaction at 25 °C and 1300 rpm in a thermomixer (Eppendorf Ibérica) and magnetic separation (DynaMag-2, Thermo Fisher), the supernatant was recovered and incubated first with rPSA-MPs (counter selection) and subsequently with hPSA-MPs (positive selection). Then the supernatant was discarded and the particles washed with BSL buffer. The bound DNA was eluted with 30 µL of hot water (95 °C, 30 min). The eluted DNA was fully amplified in 50 µL vials each containing 2 µL of the eluted ssDNA, 1 µM of both direct (5'-AGGGTTGATAGGTTAAGAGC-3') and biotinylated reverse (5'-Biotin-CCAACAGCTAGTTGACATCG-3') primers, 3 mM MgCl<sub>2</sub>, 0.2 mM dNTPs and 2.8 U

hot-start immolase<sup>TM</sup> DNA polymerase using GeneAmp PCR system 9700 (Applied Biosystems). The PCR protocol included the enzyme activation at 95 °C for 10 min, 15 cycles at 94 °C - 57 °C - 72 °C (45 s each); and a final elongation step at 72 °C for 10 min. To confirm the correct amplification an aliquot of the amplified DNA was run in a 2% agarose gel and then quantified by fluorescence using dsDNA intercalating dyes (Qubit® 2.0 Fluorometer from Invitrogen). 250 pmol of ssDNA were used in each round except in the first one where 1 nmol was used. When the PCR amplification yielded less than 250 pmol of DNA, a new aliquot was amplified.

Before starting a new round, strand separation (250 pmol) was performed using streptavidin MPs. Sufficient number of MPs were washed 3 times with BLstrep buffer. An equal volume of DNA to separate and BLstrep buffer (without tween-20) was added and incubated for 15 min. After three washing steps with BLstrep, 50 µL of NaOH 100 mM was added and incubated for 10 min. The unbound strands released were collected, neutralized with HCl 1 M, and diluted with BS.

To direct the selection through the core-fucose of the PSA two strategies were performed after round 3. In option A, an extra counter selection with hPSA-MPs blocked with the lectin PhoSL was added before the positive selection. The interaction and elution steps were performed as described above. Option B consisted on the elution of the bound DNA sequences with PhoSL lectin in BS for 1 h at 25 °C, through the competition between the ssDNA and the lectin for the core-fucose. Then the eluted DNA was amplified using the described PCR protocol. The negligible influence of the presence of lectin PhoSL in the PCR amplification solution was verified with 2.5 nM of the initial library and 2 µM of PhoSL lectin under the conditions previously described. The 2 % agarose gel showed the expected band size (**Figure S15**). The stringency of the selection was modified by varying the incubation time and number of washing steps as

indicated in **Table 6**. Concentrations of competitor and BSA were modified accordingly to the DNA concentration.

Sequences from round 10B were amplified using direct and reverse primers without tags under the PCR conditions described for the SELEX procedure. The product was purified using MinElute PCR Purification Kit (Qiagen) and a 2% agarose gel was run to check the purity and size of amplicons. The purified product was cloned using TA Cloning® Kit Dual Promoter (Invitrogen) and transformed into One shot® Competent cells (Invitrogen). After plating, 38 colonies were picked and the plasmid was eluted with water at 95 °C for 15 min. Afterward the plasmid was amplified using M13 reverse and forward primer provided with the kit to sequence the insert using an Applied Biosystem-Hitachi 3130xl Genetic Analyzer. The sequences were analyzed using BioEdit software.

#### *2.4 UV-VIS enrichment assays*

25 pmol of ssDNA from each round in BS were heated at 98 °C and cooled in ice for 4 min each and then incubated with 25 pmol of hPSA or rPSA-MPs for 10 min at 25 °C at 1300 rpm. The supernatant was collected (unbound fraction) and the particles were washed twice with BSL. The bound DNA was eluted as in SELEX and collected after magnetic separation. Both the bound and unbound fractions were quantified at 260 nm by UV-VIS spectrophotometry using a Genesys 10 S UV-VIS (Thermo Fisher) equipped with an ultra-micro quartz cuvette (Sigma-Aldrich). The DNA concentration was referred to an external calibration built with an unrelated 80-nt DNA sequence:

5'-AGGGTTGATAGGTTAAGAGCCTAGGCGAAATATAGCTACAACCTGTCT  
GAAGGCACCCAATCGATGTCAACTAGCTGTTGG-3'.

## 2.5 Remelting study

Pool sequences from round 10B were amplified using PCR primers without tags under the SELEX conditions. ROX and SYBRGreen reference dyes were added to a final concentration of 1× each in a total volume of 25 μL. The remelting curves were acquired with a 7900HT Fast Real-Time PCR System (Thermo Fisher Scientific) as follows: 95 °C for 15 s, then 50 °C for 30 s and finally a 20 min progressive increment of temperature until 95 °C that was maintained for 30 s.

## 2.6 SPR binding curves.

The hPSA or rPSA modified SPR gold chip (Xantec Bioanalytics) was prepared as follows: after cleaning with piranha solution (70:30 H<sub>2</sub>SO<sub>4</sub>/H<sub>2</sub>O<sub>2</sub>) for 10 min, it was rinsed with water and ethanol and dried with N<sub>2</sub>. A mixed self-assembled monolayer (SAM) of a 1:3 mixture of 11-mercaptoundecanoic acid and mercaptohexanol prepared from 1 mM solutions in ethanol was constructed overnight at 4 °C in a wet atmosphere. The weakly bound molecules were removed with ethanol and water and the disk was placed onto the prism of an Autolab ESPRIT SPR instrument (Ecochemie) after drying with N<sub>2</sub>. The carboxylic groups of the SAM were activated with three injections of 10 min each of a mixture of EDC (100 mM) and NHS (25 mM) in water and then washed with the running buffer (NaAc buffer). Next, 35 μL of 50 μg/mL of hPSA or rPSA in running buffer were injected in both channels for 20 min and after removing the unbound proteins with the running buffer the unmodified carboxylic groups were capped with a ethanolamine (35 μL 1 M in BS) for 15 min. Serial dilutions of ssDNA from each round were consecutively injected to construct the binding curve. Each interaction consisted on the acquisition of the baseline by triplicate injection of PBS-Na or PBS-K buffer in both cuvettes for 5 min, injection of 35 μL of ssDNA in PBS-Na or PBS-K buffer for 10 min (association step), washing with running buffer and

dissociation in PBS-Na or PBS-K buffer for 10 min. All experiments were performed in automatic mode controlled by Data Acquisition Software 4.4 at  $25 \pm 1$  °C (HaakeD1 thermostat). All signals were background-subtracted.

### *2.7 Electrochemical measurements.*

Screen-printed gold electrodes (SPAuE, DRP-220BT Methrom-DropSens) cleaned with ethanol and water and dried with N<sub>2</sub> were electrochemically polished in 0.5 M H<sub>2</sub>SO<sub>4</sub> through 10 potential cycles between 0 V and 1.3 V at 100 mV/s using a  $\mu$ AUTOLAB type II potentiostat with NOVA 2.1 software (EchoChemie). The counter and reference electrodes were covered with a transparent non-surfactant thin layer, therefore external Ag|AgCl, KCl (3 M) |KNO<sub>3</sub> (3 M) reference and Pt counter electrodes were used.

A mixed SAM of 11-mercaptopundecanoic acid/mercaptohexanol (1:3) in 1 mM NaAc buffer was built overnight at 4 °C under wet conditions to covalently anchor the protein in two steps: activation in 100 mM EDC and 25 mM NHS in water (30 min) and attachment to protein from 50  $\mu$ g/mL solution of hPSA, rPSA, NGAL or AFP in NaAc buffer (30 min). Blocking was achieved with 1 M ethanolamine solution in BS (15 min).

For the binding assays, 10  $\mu$ L of 6-FAM labelled aptamers prepared in BS were incubated onto the modified working electrode for 30 min and labelled with 10  $\mu$ L of 0.5 U/mL Fab-anti-fluorescein-peroxidase in BS-0.5% casein for 30 min. Finally, 35  $\mu$ L of TMB-H<sub>2</sub>O<sub>2</sub> was added on the three electrodes and after 30s of enzymatic reaction, the oxidized product of TMB was measured at -0.2 V for 60 s by chronoamperometry. The analytical signal was the averaged current from the last 10 s. After each step, the electrode was washed with the next buffer used and dried with nitrogen.

## 2.8 *Circular dichroism*

Circular dichroism (CD) spectrum of 5  $\mu\text{M}$  of aptamer PSAG-1 in phosphate buffer with sodium or potassium chloride was acquired using a JASCO J-815 spectropolarimeter (Jasco, Germany) in a 2 mm quartz cuvette at 25  $^{\circ}\text{C}$  controlled with a Peltier pump. CD spectra were recorded from 500 nm to 200 nm at a scanning rate of 200 nm/min and 1 nm resolution. Each spectrum is the average of three successive scans after baseline correction with CD spectrum of the buffer.

## 2.9 *UV-VIS thermal melting curves*

UV-VIS melting was performed in a Carry 60 UV-VIS spectrophotometer equipped with a single cell Peltier accessory and a cuvette temperature controller (Agilent Technologies, USA). 5  $\mu\text{M}$  of aptamer in phosphate buffer with potassium ions or just the buffer (for background correction) was placed into a high precision quartz cell (Hellma Analytics, Germany), heated to 90  $^{\circ}\text{C}$ , slowly cooled to 25  $^{\circ}\text{C}$  and then heated again to 90  $^{\circ}\text{C}$ . Absorbance was monitored between 500 nm to 200 nm at scanning rate of 600 nm/min and measurements were performed every ten degrees interval except between 25  $^{\circ}\text{C}$  and 30  $^{\circ}\text{C}$ . The subtracted spectra were finally obtained at each temperature and buffer.

## 2.10 *Deglycosylation study.*

The interaction with the specific enzymes was performed on SPAuE modified as indicated above in this order: neuraminidase-A (P0722S),  $\beta$ -galactosidase (P0746S), N-acetylhexosaminidase (P0744), mannosidase (P0768S), fucosidase (P0748S) and PNGaseF (P0704S) (New England Biolabs). All enzymes were prepared in their specific buffer and in the concentration specified in the manufacturer's protocols. All the interactions were performed in 10  $\mu\text{L}$  at 37  $^{\circ}\text{C}$  in a wet atmosphere for 2 h except with PNGaseF (24 h). After the interaction with the enzymes, an electrochemical

binding assay was performed with 100 nM of PSAG-1 in PBS-K or 500 nM of PSA-1 in BS as indicated above.

### 2.11 *Statistical analysis*

Data are presented as means  $\pm$  standard deviation for n=3. The difference between means was estimated using the two-tail unpaired Student's t-test taking into account whether the standard deviations are significantly equal by using a two-tailed F-test. In all cases a minimum of 95% confidence level was used. Error bars represent the standard deviation for n=3.

## **3 Detailed computational settings.**

### 3.1 *Aptamer model building*

Current molecular modelling methods are becoming useful tools to determine the secondary and tertiary structure of aptamers as well as to study their interaction with proteins.<sup>3</sup> Of particular interest may be the computational protocol pursued by Jeddi and Saiz<sup>4</sup> for building structural models from the nucleotide sequence of ssDNA aptamers. These authors have shown that the available methods for the structural prediction of single stranded RNAs are also useful to obtain 3D ssDNA models as long as the ssDNA structures are fully relaxed by means of all-atom molecular mechanics calculations involving both energy minimization and molecular dynamics simulations in explicit solvent. Interestingly, Jeddi and Saiz have validated their approach by comparing between the *in silico* and the crystallographic structures for a set of 24 ssDNA hairpin molecules, finding a good agreement between computational predictions and experiment and concluding that their protocol works “exceptionally well” for the hairpin-like structural motif.<sup>4</sup>

Herein, we implemented a similar strategy to that of Jeddi and Saiz albeit with variations in the choice of computational techniques and settings, aimed to increase the reliability of the predicted ssDNA models. Most remarkably, we consider that Molecular Dynamics (MD) simulations on the  $\mu\text{s}$  time scale would be required to characterize the structural and dynamical properties of the PSAG-1 aptamer, which is a relatively large and flexible system. Furthermore, we propose to apply both enhanced MD techniques and conventional MD methods. The workflow of the computational protocol is summarized in Scheme S1.

### *3.1.1 Initial structures of the PSAG-1 aptamer*

To obtain a guess of the secondary structure associated to the PSAG-1 sequence, we employed the mfold algorithm<sup>5</sup> that builds a series of secondary structures combining loop and stack motifs and ranks them in terms of thermodynamic data for canonical base pairing including nearest neighbor effects. Default settings were used for the mfold runs selecting a  $T$  value of 25 °C and an ionic strength of 0.150 M. For the PSAG-1 aptamer, mfold yields three secondary structures that have  $\Delta G$  scorings of -3.16, -3.16 and -2.72 kcal/mol. For the construction of the corresponding 3D models, we resorted to the RNA Composer tool,<sup>6</sup> which decomposes the secondary structure(s) provided as input into different motifs (loop, stems, etc.), selects geometries for the various fragments from the RNA FRABASE structural database<sup>7</sup> allowing base replacement if necessary, employs the machine translation algorithm to generate coordinates for unmatched motifs, and builds a global 3D model by performing superimposition operations and structural relaxation. The coordinates of the resulting structures were saved in PDB format and transformed into the equivalent ssDNA molecules using the Discovery Studio Visualizer software.<sup>8</sup>

### 3.1.2 *Enhanced Molecular Dynamics of the PSAG-1 models*

The coordinates of the three most likely PSAG-1 models (A1-3) generated by *mfold/RNA-composer* were processed with the *tLEaP* program included in the AMBER18 suite of programs<sup>9-11</sup> in order to add the missing H atoms and to assign the molecular mechanics parameters. We described the ssDNA molecules with the *parmbsc1* force field<sup>12</sup> that complements the AMBER charges and atom types for DNA<sup>13</sup> with refined torsion parameters that have been derived from high level QM calculations. To remove bad contacts in the initial ssDNA structures, we made use of the SANDER program included in AMBER18 to carry out 500 steps of conjugate-gradient minimization with a distance-dependent dielectric constant. Using again *tLEaP*, the minimized ssDNA structures were centered in an octahedral box of TIP3P<sup>14</sup> water molecules that extended at least 16 Å from the solute atoms. Na<sup>+</sup>/Cl<sup>-</sup> counterions described by parameters adapted from those of Aqvist<sup>15</sup> were added to the solvent box by *tLEaP* in order to neutralize the negative charge (-39) of the systems and provide 0.150 M ionic strength in agreement with experimental conditions. These settings resulted in simulation boxes containing a total of ~74000 (A1), ~92000 (A2) and ~55000 (A3) atoms.

The solvent molecules and counterions were initially relaxed by means of energy minimizations and 100 ps of molecular dynamics (MD) using SANDER. Then the full systems were minimized and heated gradually to 300 K using 60 ps of constant volume (NVT) MD with a 1 fs time step and using the PMEMD program in AMBER18. Subsequently, the density was adjusted by means of 2.0 ns of constant pressure (NPT) MD with a 2 fs time step and using the Monte Carlo barostat as implemented in PMEMD. Langevin dynamics was employed to control the temperature (300 K) with a collision frequency of 2 ps<sup>-1</sup>. The SHAKE algorithm<sup>16</sup> was selected to constraint all R-H

bonds, and periodic boundary conditions were applied to simulate a continuous system at constant pressure (NTP). A non-bonded cutoff of 9.0 Å was used and the Particle-Mesh-Ewald method<sup>17</sup> was employed to include the contributions of long-range interactions.

To explore multiple free energy basins of the conformational space of the ssDNA models, we employed the Gaussian Accelerated Molecular Dynamics (GMAD) technique<sup>18,19</sup>. A GaMD simulation performs an enhanced and unconstrained sampling of the systems using harmonic boost potentials to smooth out the potential energy surface, accelerating thus transitions between low-energy conformational states. We applied two boost potentials: one to the torsion energetic term and another to the total potential energetic term including waters and counterions. The upper limits of both boost potentials were set to their default values (6.0 kcal/mol). The two numerical parameters that define the boost potentials (a threshold energy  $E$  and harmonic constant  $k$ ) were estimated in preliminary MD and GaMD runs from the average and standard deviation of potential energies. Following the recommended prescriptions in the AMBER18 manual, the number of simulation steps (NTAVE) used to update the potential energy statistics was about 4 times the total number of atoms in each system (*e.g.*, NTAVE=30000 for A1). The preliminary MD run comprised a total of 14xNTAVE simulation steps, yielding an initial guess of the  $E$  and  $k$  parameters. After an equilibration GaMD phase (14xNTAVE steps),  $E$  and  $k$  are refined during a GaMD run with a length of 10x14xNTAVE steps. For example, the preparation of the A1 GaMD simulation requested a total of 50700000 2fs steps (~100 ns). Finally, the production phase of each GaMD NTP simulation was extended up to 2.5  $\mu$ s. Coordinates and the values of the boost potentials were saved for analysis every 10000

simulation steps (2.5 ps). The GPU accelerated version of the PMEMD code<sup>9,20</sup> was employed both in the conventional MD and in the GaMD runs.

### 3.1.3 Structural characterization and free energy reweighting of the enhanced MD simulations

To characterize the shape and conformation of the ssDNA molecule along the production GaMD simulations, two structural indexes were computed. On one hand, we obtained the root-mean-squared-deviation (RMSD) of the heavy atoms (P,C,N,O) in residues 3-38 with respect to the initial structure. On the other hand, we computed the so-called interaction network fidelity index (INF)<sup>21</sup>, which is built from the sets of characteristic intramolecular interactions in a reference structure ( $S_r$ ) and in a given GaMD snapshot ( $S_m$ ). To determine and analyze such interactions in the ssDNA molecules, we employed the DSSR software<sup>22</sup> to identify both base pair interactions and non-pair interactions (*i.e.*, base stacking or base contacts). Hence, the INF value is computed from (i) the set of common interactions between the two structures, which are counted as true positives  $T_p = S_r \cap S_m$ ; (ii) the set of the interactions in the reference structure that are not present in the GaMD snapshot, which are counted as false positives,  $F_p = S_m \setminus S_r$ , and (iii) the number of interactions absent in the GaMD snapshot but present in the reference structure, which are termed false negatives,  $F_N = S_r \setminus S_m$ . It has been proposed<sup>21</sup> then to calculate the INF index as a Mathews correlation coefficient considering true and false positives and negatives:

$$INF = \sqrt{\left(\frac{|T_p|}{|T_p| + |F_p|}\right)\left(\frac{|T_p|}{|T_p| + |F_N|}\right)} \quad (1)$$

If the given GaMD snapshot reproduces all the interactions of the reference structure, then  $|F_P|=|F_N|=0$ , and  $INF=1$ . When the GaMD snapshot does not reproduce any of the interactions of the reference structure, then  $INF=0$ , since  $|T_P|=0$ . It must be also noticed that the values of the RMSD and INF indexes along the GaMD trajectories are largely uncorrelated to each other so that they behave as orthogonal descriptors of the ssDNA configurations.

The enhanced MD methods allow the calculation of a potential of mean force ( $F = -kT \ln p(A)$ ) by estimating the canonical probability distribution  $p(A)$  along a selected coordinate  $A$ . To this end, the probability distribution  $p^*(A)$  derived from the enhanced simulation is reweighted to recover  $p(A)$  as<sup>23</sup>:

$$p(A_j) = p^*(A_j) \frac{\left\langle \exp\left(\frac{\Delta V(\mathbf{r})}{kT}\right) \right\rangle_j}{\sum_j \left\langle \exp\left(\frac{\Delta V(\mathbf{r})}{kT}\right) \right\rangle_j} \quad j = 1, \dots, M \quad (2)$$

where  $\Delta V$  is the boost potential and  $M$  is the selected number of bins along the  $A$  coordinate. In the case of GaMD, the adoption of the harmonic boost potential results in an accurate approximation of the exponential reweighting term through a cumulant expansion to the 2nd order:

$$\left\langle \exp\left(\frac{\Delta V}{kT}\right) \right\rangle \approx \exp\left\{ \frac{\langle \Delta V \rangle}{kT} + \frac{\langle \Delta V^2 \rangle - \langle \Delta V \rangle^2}{2(kT)^2} \right\} \quad (3)$$

This approach is implemented in various *Python* scripts developed by Miao et al.<sup>23</sup>, which were used here to reweight the GaMD simulations of the ssDNA models and estimate their 2D free energy maps in terms of the RMSD/INF coordinates, selecting a

number of bins  $M \approx 15-20$  along each coordinate. In addition, we obtained the unweighted population distribution over the 2D bins defined in terms of the RMSD/INF coordinates (*i.e.*,  $p^*$  estimated as the fraction of GaMD snapshots located within each bin).

The Chimera visualization system<sup>24</sup> was employed to draw the ribbon models of the aptamer while the secondary structures, as determined with DSSR<sup>22</sup>, were annotated and displayed with the VARNA software<sup>25</sup>, representing the base pairing with the Leontis Westhof<sup>26</sup> graphical conventions.

### 3.1.4 Conventional Molecular Dynamics

For each PSAG-1 model (**A1-A3**), we picked up one structure from the subset of GaMD snapshots located within a low energy basin in the free energy map. In particular, we chose the snapshot that had the lower value of the boost potential, which would be presumably closer to those structures that populate the NPT ensemble. We also ensured that the representative GaMD structure is taken from a highly-populated conformational subspace by excluding the boundary regions of the free energy map that were scarcely populated.

The selected solute structures were solvated within a new octahedral box of water and  $\text{Na}^+/\text{Cl}^-$  counterions (0.150 M ionic strength) adjusted to their overall shape. Subsequently, the models were further relaxed by means of a conventional MD simulation at NPT conditions and using the same settings for the preliminary stages (minimization, thermalization and pressurization) as those used in the precedent GaMD calculations. The production phase was extended up to 2.5  $\mu\text{s}$  to allow the systems to explore their equilibrium conformations in aqueous solution. The coordinates of the ssDNA models along the MD trajectories were clustered using the CPPTRAJ program<sup>27</sup>

in AMBER18 with the average-linkage clustering algorithm and a sieve of 250 frames. The distance metric between frames was calculated via best-fit coordinate root mean square deviation (RMSD) using the coordinates of the heavy atoms (C, N, O, P). The clustering was finished when the minimum distance between clusters was greater than 5 Å.

We also computed an MD trajectory (2.5 μs) for a truncated ssDNA aptamer (**T-A1**) that maintains the residues 1-32. The initial structure was taken from the representative of the most populated cluster according to the A1 cMD simulation. After removing the coordinates of the last 8 residues, the truncated system was solvated within an octahedral box of water and Na<sup>+</sup>/Cl<sup>-</sup> counterions. The settings of the T-A1 MD simulation were equivalent to those of the former simulations. The average-linkage clustering algorithm was done with an RMSD threshold of 3 Å.

### 3.1.5 MM-PBSA calculations

MM-PBSA calculations<sup>28,29</sup> were carried out on 1000 equally-spaced snapshots extracted from the last half of the A1-A3 MD trajectories. After having removed all solvent molecules and counterions, the MM-PBSA energy of the solute atoms was computed as:

$$G_{MM-PBSA} = E_{MM} + \Delta G_{solv}^{PB} + \Delta G_{solv}^{non-polar} \quad (4)$$

where  $E_{MM}$  is the molecular mechanics energy including the  $3RT$  contribution due to six translational and rotational degrees of freedom,  $\Delta G_{solv}^{PB}$  is the electrostatic solvation energy obtained from Poisson-Boltzmann calculations<sup>30</sup>, and  $\Delta G_{solv}^{non-polar}$  is the non-polar part of solvation energy due to cavity formation and dispersion interactions between the

solute and the solvent molecules. All these energy components were evaluated with the SANDER and PBSA programs included in the AMBER18 suite.

The  $E_{MM}$  term in eq. (4) was calculated with no cutoff. The  $\Delta G_{solv}^{PB}$  term was obtained by solving the non-linear Poisson-Boltzmann (PB) equation<sup>31</sup> on a cubic lattice with a grid spacing of 0.33 Å and using an iterative finite-difference method. The solute is represented by the atomic charges and radii taken from the AMBER/bsc1 representation and treated as a low dielectric medium ( $\epsilon_{in}$ ). Several values of  $\epsilon_{in}$  (2, 4, 8, 10, 12, 20) were considered to compute the PB solvation energy and the Coulombic contributions to  $E_{MM}$ . The surrounding solvent was treated as a continuum of dielectric constant ( $\epsilon_{out}=80$ ) with a 1:1 electrolyte (0.150 mM) distributed according to a Boltzmann weighted average of the mean electrostatic potential. The solute-solvent dielectric boundary was the contact surface between the radii of the solute atoms and the radius (1.4 Å) of a water probe molecule. The nonpolar solvation term  $\Delta G_{solv}^{non-polar}$  was estimated as the sum of dispersion and cavity terms according to the implicit non-polar solvent model of Tan et al.<sup>32</sup> as implemented in PBSA.

### 3.2 Rigid docking calculations

According to our MD simulations, the structure of the truncated T-PSAG-1 aptamer in aqueous solution is quite compact and stable so that ligand flexibility is expected to play a minor role in the hPSA/aptamer docking calculations. Thus, the coordinates of the most important cluster representative extracted from the T-PSAG-1 MD simulation were taken for the docking analysis.

Initial coordinates for the hPSA protein were obtained from the 3QUM crystal structure (3.2 Å)<sup>33</sup>. In this structure, hPSA is forming a sandwich complex with two monoclonal antibodies that were removed for the subsequent steps. The unit cell contains two

protein molecules (labeled as P and Q in the PDB file), both glycosylated by a sialated triantennary carbohydrate (15 monosaccharides long), *N*-linked to the Asn<sub>61</sub> side chain. Molecule P also displays a short disaccharide O-linked at Thr<sub>125</sub>, but only the glycan bound to the consensus Asn<sub>61</sub> side chain is included in the molecular modeling of the hPSA/aptamer complex. Although the coordinates of the hPSA protein in the crystallographic P and Q molecules present only minor structural differences, the dissimilar carbohydrate linkage between mannose in position 4 and *N*-acetylglucosamine in position 5 (*i.e.*  $\beta$ -(1,4) for P and  $\beta$ -(1,6) for Q), and the crystalline environment result in a different orientation of the hPSA-glycan moieties. However, as noticed by Stura et al.<sup>33</sup>, the overall shape of the branched carbohydrate in the 3QUM structural is modeled without doubt for molecule P, but the interpretation of the sugar branches for molecule Q is less certain. Consequently, only the coordinates of molecule P were taken to carry out the docking calculations.

The protonation state for the titratable residues in hPSA were assigned according to structure-based  $pK_a$  calculations performed with the H++ web server (version 3.1)<sup>34</sup>. For the internal dielectric constant, we used three different values ( $\epsilon_{\text{int}} = 4, 10, \text{ and } 20$ ) to examine the consistency of the  $pK_a$  results. By combining the computed  $pK_a$ s and the visual inspection of the hPSA structure, we finally assigned a positive charge to all lysines and arginines and a negative one to all aspartic and glutamic acids. Histidines were modelled neutral and protonated at the N $\epsilon$  atom, except for His<sub>57</sub> in the catalytic triad that was protonated at N $\delta$  to maintain the catalytically relevant Ser<sub>195</sub>-O $\gamma$ H $\cdots$ N $\epsilon$ -His<sub>57</sub>-N $\delta$ H $\cdots$ O $\delta$ -Asp<sub>102</sub> H-bonds.

To explore different binding modes between the hPSA protein and the truncated aptamer, we performed rigid docking calculations. Previous docking results obtained for protein/aptamer complexes have employed different codes like ZDOCK<sup>35-37</sup>,

PatchDock<sup>38,39</sup> and HADDOCK<sup>39,40</sup> that are usually executed through the available web servers. In the present case, experimental evidences suggest that T-PSAG-1 binds at the hPSA-glycan interface so all the protein, sugar, and DNA units should be described by the docking scoring function on a similar basis in order to get reliable results. For this reason, we decided to use Autodock 4.2<sup>41</sup>, a standalone program initially developed to dock small and flexible ligands to proteins, DNA, RNA, and other macromolecules. Compared to the other docking programs, Autodock is computationally more demanding because it uses a physics-based scoring function that includes specific terms to model van der Waals, electrostatic, and H-bond interactions as well as a desolvation contribution to the binding free energy. The weight of each energy term was originally calibrated using the structures and binding energies experimentally obtained for 188 protein/ligand complexes<sup>42</sup>, but the resulting scoring function provides also accurate results for the docking of small molecules to nucleic acid targets<sup>43</sup>. Therefore, the various atom types and charges (heavy atoms and polar H atoms) that are parametrized within Autodock may provide a balanced description of all interactions among the protein, nucleic acid, and sugar atoms.

Starting from the coordinate files of the hPSA receptor and the T-PSAG-1 ligand, we used the available *Python* scripts in Autodock 4.2 to assign atom types and Gasteiger charges, and to merge non-polar hydrogens. To explore multiple binding options, Autodock requires pre-calculated grid maps for the receptor, in which information on electrostatics, H-bond and steric constraints are stored for each atom type in the ligand. We centered these grid maps on the fucose C1 atom and, due to the large size of the ligand, we defined 256 grid points along each Cartesian axis with a spacing of 0.333 Å. This defined a docking search space centered at the hPSA-sugar interface, but width enough (*i.e.*, a cubic box of 85 Å length) to allow the aptamer to rotate/translate around

hPSA and produce different binding poses. To sample the hPSA/aptamer conformational space, we selected the genetic search algorithm to generate 500 conformations that were finally clustered with an RMSD tolerance of 2.0 Å. The resulting cluster representatives were then relaxed to relieve bad contacts and optimize favorable interactions, introducing thus some flexibility in the docking protocol. To this end, we assigned the AMBER ff14SB force field to represent the hPSA protein<sup>44</sup>, *parmbsc1* for the DNA aptamer<sup>12</sup>, and the GLYCAM\_06j parameters for the carbohydrate moiety<sup>45</sup>. The selected Autodock complexes were edited to add H atoms with tLEaP. Using SANDER, we run 5000 steps of conjugate gradient minimization on each structure and using the Hawkins, Cramer, Truhlar pairwise generalized Born implicit solvent model<sup>46</sup>. After the geometry optimization, the minimized structures corresponding to the cluster representatives were finally ranked using the Autodock scoring function. The docked complexes were visualized with the Chimera program<sup>24</sup>.

### 3.3 Molecular dynamics simulation of the hPSA/T-PSAG-1 complex

After the rigid docking and rescoring calculations, the relevance and stability of the hpSA-aptamer interactions in the most likely docking complex were further assessed by carrying out an unrestrained MD simulation using the conventional MD algorithm and similar settings to those employed in the trajectories of the isolated aptamer molecules. Thus, the top-scoring pose was placed in an octahedral box of TIP3P waters that extended at least 24 Å from the solute atoms, which were described with the AMBER ff14SB, *parmbsc1* and GLYCAM\_06j parameter sets. The tLeAP program neutralized the total charge of the system by adding 94 Na<sup>+</sup> and 60 Cl<sup>-</sup> counterions (0.150 M ionic strength), resulting in a simulation box containing ~109000 atoms.

Using the SANDER/PMEMD programs, the hPSA/T-PSAG-1 system was subject to minimization, thermalization and pressurization applying the same protocols as those

used in the aptamer MD simulations. Subsequently, a conventional NPT MD run was extended up to 2.5  $\mu\text{s}$  and MD frames were saved for analysis every 2.5  $ps$  (solute atoms) or 50  $ps$  (solute+water). During the simulation, we observed that the relative positioning of the sugar antennas changes substantially, evolving on a hundred-ns time scale from their X-ray conformation to more stable ones that favor the formation of stable contacts between the sugar residues and the aptamer. These wide and relatively slow motions were characterized in terms of the time evolution of the RMSD values for selected structural elements with respect to the initial geometry. The molecular surface of protein/glycan/DNA fragments was also computed using the linear combination of pairwise overlaps (LCPO) method<sup>47</sup>. In addition we evaluated the MM-PBSA energy of the hydrated hPSA/T-PSAG-1 complex over 1000 MD snapshots evenly distributed along the trajectory. The MM-PBSA energies ( $\epsilon_r=1.0$ ) were obtained using identical settings and programs to those employed for the aptamer simulations.

The largest structural changes and the decreasing trend in the MM-PBSA energies ends after  $\sim 1.0 \mu\text{s}$  of simulation time so that the first half of the hPSA/T-PSAG-1 trajectory was considered as a “search phase” in which the starting conformation with high energy relaxes to a more stable conformation. Only the second half was then considered when carrying out the rest of structural analyses. Thus, the coordinates of the hPSA/aptamer complex were clustered adopting two RMSD distance metrics (CNOP atoms of T-PSAG-1 or C1-C6 atoms in the sugar residues) and requesting a minimum distance between clusters of 2.5  $\text{\AA}$ . The secondary structure of the hPSA-bound aptamer was annotated/visualized using the DSSR/VARNA programs and the coordinates of the most populated cluster representative. Concerning the hPSA/T-PSAG-1 contacts, all H-bond interactions were characterized on the basis of geometrical criteria (*e.g.*,  $X\cdots Y$  distance  $< 3.5 \text{\AA}$  and  $X-H\cdots Y$  angle  $> 120^\circ$ ). Non-polar interactions were scored in

terms of an empirical dispersion potential<sup>48</sup> evaluated over the pairs of atoms belonging to different hydrophobic groups. The criteria for assessing the occurrence of dispersion interactions were: (a) the total dispersion energy is larger than 0.5 kcal/mol in absolute value; (b) the distance between the centers of mass of the two interacting groups is below 10.0 Å. The CPPTRAJ software was used for all the RMSD, molecular surface and clustering calculations while the H-bond/dispersion contacts were characterized using a software program developed locally. The modeled complexes were visualized with the Chimera program.<sup>24</sup> The computational protocol for the construction of the hPSA/T-PSAG-1 complexes is summarized in Scheme S2.

## 4 Tables

**Table 1** Relative  $G_{MM-PBSA}$  energies of the **A1/A2/A3** structural models of the PSAG-1 aptamer. For each model, its average  $G_{MM-PBSA}$  is computed from 1000 equally-spaced snapshots along the corresponding MD trajectory. The MM-PBSA calculations were repeated using different values for the solute dielectric constant ( $\epsilon_{int}$ ). The conventional (in bold) and the block-averaged (in Italics) standard errors of the mean are also indicated (block size ranged from 1 up to a quarter of the total number of frames).

<b>Model</b>	$\epsilon_{int}=1$	$\epsilon_{int}=2$	$\epsilon_{int}=4$	$\epsilon_{int}=8$	$\epsilon_{int}=10$	$\epsilon_{int}=12$	$\epsilon_{int}=20$
<b>A1</b>	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>A2</b>	37.1	34.9	27.4	25.2	23.8	22.8	19.9
	<b>2.6</b>	<b>1.6</b>	<b>1.6</b>	<b>1.5</b>	<b>1.5</b>	<b>1.5</b>	<b>1.5</b>
	<i>3.8</i>	<i>5.5</i>	<i>5.5</i>	<i>5.5</i>	<i>5.3</i>	<i>5.5</i>	<i>5.3</i>
<b>A3</b>	56.8	47.7	44.3	42.2	39.2	38.3	33.4
	<b>2.6</b>	<b>1.6</b>	<b>1.6</b>	<b>1.6</b>	<b>1.5</b>	<b>1.5</b>	<b>1.4</b>
	<i>4.9</i>	<i>7.5</i>	<i>7.7</i>	<i>8.3</i>	<i>7.5</i>	<i>7.4</i>	<i>7.1</i>

**Table 2**  $\Delta_{\text{bind}}G$  binding energies (kcal/mol) obtained using the Autodock 4.2 scoring function for the relaxed cluster representatives derived from the hPSA/T-PSAG-1 docking calculations. Type of receptor/ligand contacts, *both* for the interaction with the protein and the glycan or *protein* for interaction only with protein atoms, is also indicated.

Structure	$\Delta_{\text{bind}}G$	Contacts	Structure	$\Delta_{\text{bind}}G$	Contacts	Structure	$\Delta_{\text{bind}}G$	Contacts
1	-15.77	<i>both</i>	13	-11.15	<i>both</i>	25	-9.66	<i>protein</i>
2	-13.79	<i>both</i>	14	-11.13	<i>both</i>	26	-9.62	<i>both</i>
3	-13.34	<i>both</i>	15	-11.10	<i>protein</i>	27	-9.51	<i>both</i>
4	-12.97	<i>protein</i>	16	-11.06	<i>both</i>	28	-9.48	<i>protein</i>
5	-12.86	<i>both</i>	17	-10.98	<i>both</i>	29	-9.46	<i>protein</i>
6	-12.53	<i>both</i>	18	-10.96	<i>protein</i>	30	-9.22	<i>Protein</i>
7	-12.28	<i>protein</i>	19	-10.73	<i>protein</i>	31	-9.14	<i>both</i>
8	-11.89	<i>both</i>	20	-10.50	<i>both</i>	32	-8.92	<i>both</i>
9	-11.81	<i>protein</i>	21	-9.89	<i>both</i>	33	-8.07	<i>protein</i>
10	-11.81	<i>both</i>	22	-9.86	<i>both</i>	34	-7.78	<i>both</i>
11	-11.65	<i>both</i>	23	-9.78	<i>both</i>	35	-6.96	<i>both</i>
12	-11.61	<i>both</i>	24	-9.74	<i>both</i>			

**Table 3.** Percentage of occurrence and average distance (Å) computed for the most abundant polar contacts between sugars along the second half of the MD trajectory computed for the PSA/T-PSAG-1 complex. Contacts mediated by water molecules are also included.

<b>Polar Contacts: Glycan···Glycan</b>		
NAG5'@N2H···O2N@NAG5	84%	2.91
NAG5'@O3H···O6@NAG5	56%	2.93
FUC8@O2H···O6@NAG2	52%	2.92
NAG5@O6H···O4@GAL6'	51%	2.90

**Table 4.** Percentage of occurrence and average distance (Å) computed for the most abundant PSA···Aptamer polar contacts along the second half of the MD trajectory computed for the PSA/T-PSAG-1 complex. Contacts mediated by water molecules are also included.

<b>Polar Contacts: Protein···Aptamer</b>		
Lys <sub>95E</sub> @NζH···N3@G5	76%	2.99
Lys <sub>95E</sub> @NζH···O2@C4	75%	2.88
Arg <sub>95I</sub> @Nη1H···OP1@T31	67%	2.95
Arg <sub>95G</sub> @Nη1H···O2@C4	66%	2.87
Lys <sub>95E</sub> @NζH···O4'@G5	59%	3.10
Arg <sub>95I</sub> @Nη2H···OP1@T31	56%	2.94
Met <sub>95A</sub> @O···WAT···OP1@G7	77%	4.21
Lys <sub>95E</sub> @NζH···WAT···N2@G5	70%	3.75
Arg <sub>95G</sub> @Nη1H···WAT···OP1@T32	64%	4.52
Asp <sub>95</sub> @Oδ1···WAT···OP1@T30	57%	4.27
Asn <sub>95F</sub> @NH···WAT···OP1@G6	52%	5.59
<b>Polar Contacts: Glycan···Aptamer</b>		
NAG1@O3H···OP2@G7	100%	2.64
NAG2@O3H···N7@G8	100%	2.85
MAN3@O2H···O6@G7	100%	2.75
MAN3@O4···HN4@C26	99%	2.98
MAN4'@O4H···OP2@G24	97%	2.69
FUC8@O3H···OP2@G6	73%	2.76
MAN4@O6H···OP2@G6	68%	2.71
MAN3@O5···HN4@C26	66%	3.16
MAN4'@O3H···OP2@G24	64%	2.80

**Table 5.** Percentage of occurrence, average distance (Å) and empirical vdW energy (kcal/mol) computed for the most abundant van der Waals PSA···Aptamer contacts along the second half of the MD trajectory computed for the PSA/T-PSAG-1 complex.

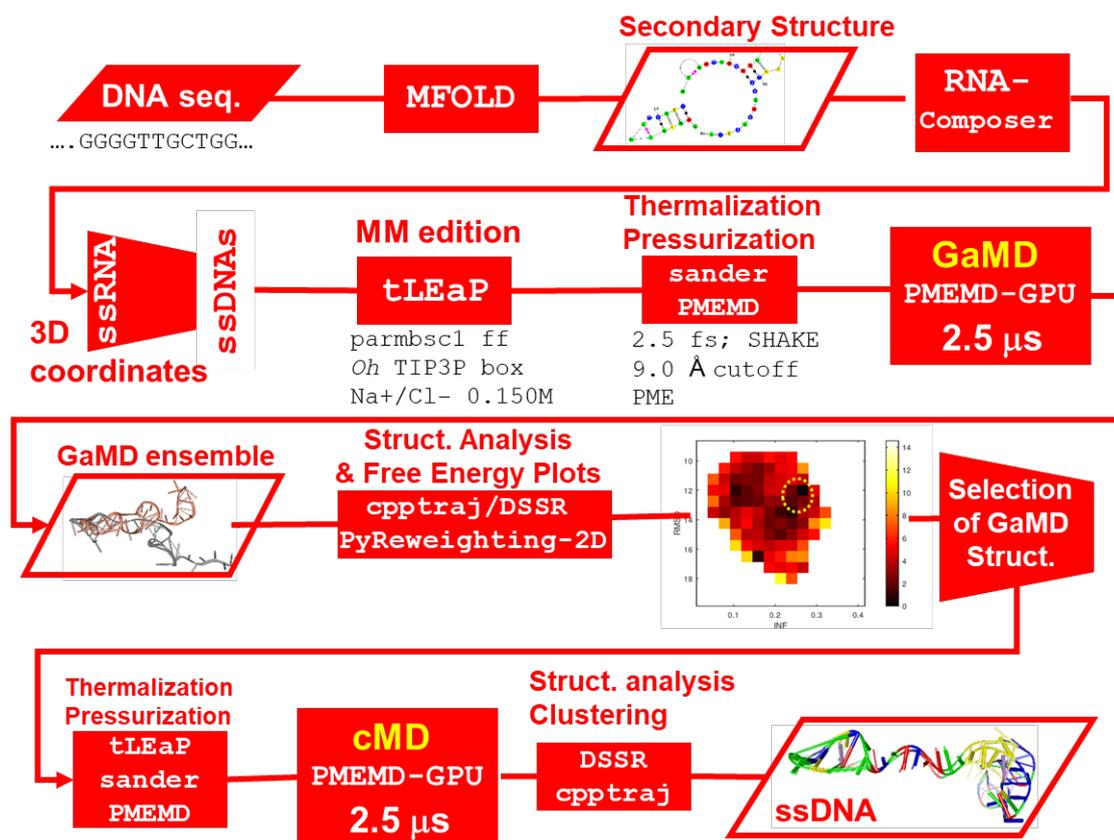
<b>vdW Contacts: Protein···Aptamer</b>			
Met <sub>95A</sub> @sidechain···sugar@G7	96%	5.58	-1.00
Leu <sub>95D</sub> @sidechain···sugar@G6	71%	7.51	-0.17
<b>vdW Contacts Glycam···Aptamer</b>			
NAG2@ring···base_ring@G7	100%	5.14	-1.79
MAN3@ring···base_ring@C26	100%	5.83	-1.07
NAG2@ring···base_ring@G8	100%	6.49	-0.87
MAN3@ring···base_ring@C27	100%	6.69	-0.58
NAG1@ring···sugar_ring@G6	100%	6.49	-0.58
MAN3@ring···base_ring@G8	100%	6.38	-0.57
MAN3@ring···base_ring@G7	100%	6.65	-0.54
FUC8@ring···sugar_ring@G6	100%	6.09	-0.51
MAN4'@ring···base_ring@G24	100%	7.68	-0.41
NAG1@ring···sugar_ring@G7	100%	6.55	-0.34
NAG1@ring···base_ring@G7	100%	8.54	-0.30
NAG2@ring···base_ring@G6	100%	6.78	-0.25
NAG2@ring···sugar_ring@G7	100%	6.73	-0.24
MAN3@ring···base_ring@G6	99%	7.31	-0.23
NAG2@ring···base_ring@T9	99%	7.82	-0.22
NAG2@ring···sugar_ring@G6	98%	6.37	-0.36
MAN4'@ring···sugar_ring@A23	98%	7.16	-0.29
MAN4'@ring···sugar_ring@G24	97%	6.59	-0.44
MAN3@ring···base_ring@G25	95%	6.82	-0.37
MAN4@ring···base_ring@G6	83%	8.11	-0.17
MAN3@ring···base_ring@T9	56%	7.72	-0.12

**Table 6** Variable SELEX conditions to modify the stringency of each round

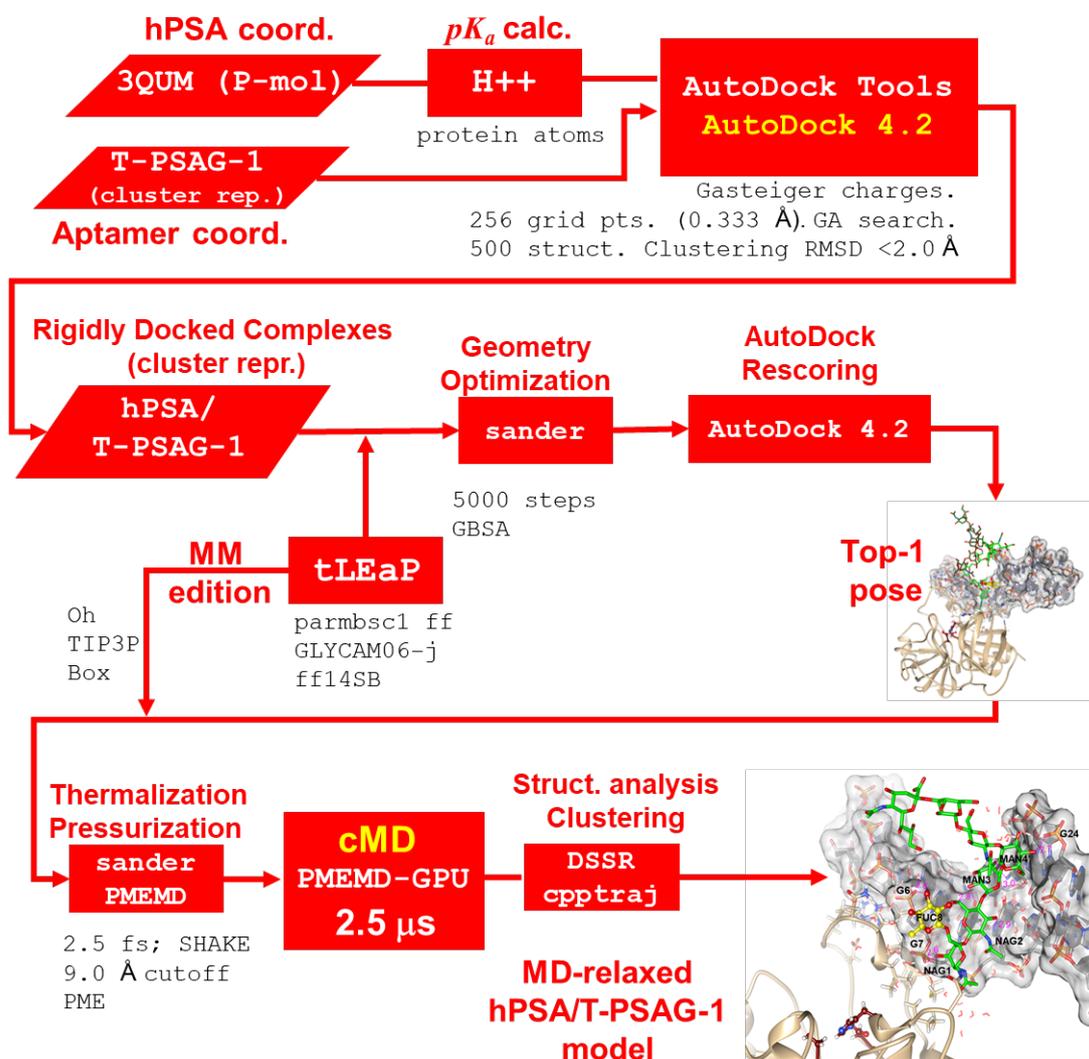
<b>Round</b>	<b>Incubation time (BSA, rPSA, hPSA)</b>	<b>Incubation time hPSA+PhoSL</b>	<b>PhoSL concentration used for elution</b>	<b>Number of washing steps</b>
1	60	-	-	2
2	60	-	-	2
3	<b>30</b>	-	-	2
4A	30	30	-	2
4B	30	-	<b>1 <math>\mu\text{g/mL}</math></b>	2
5A	30	30	-	<b>5</b>
5B	30	-	1 $\mu\text{g/mL}$	5
6B	<b>15</b>	-	1 $\mu\text{g/mL}$	5
7B	15	-	<b>2 <math>\mu\text{g/mL}</math></b>	5
8B	15	-	2 $\mu\text{g/mL}$	<b>10</b>
9B	<b>10</b>	-	2 $\mu\text{g/mL}$	10
10B	10	-	2 $\mu\text{g/mL}$	<b>15</b>

## 5 Schemes

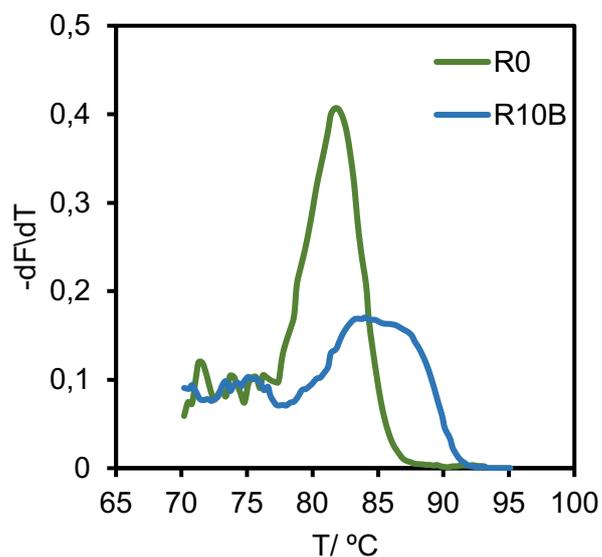
**Scheme S1.** Workflow of the computational protocol for the construction of the aptamer models.



**Scheme S2.** Workflow of the computational protocol for the construction of the hPSA/T-PSAG-1 model.



## 6 Figures



**Figure S1. Remelting curve analysis.** The curves were obtained after 30 s reannealing at 50 °C for the initial library (R0) and the pool after the 10th round (R10B). They indicate a reduction in diversity from the initial library ( $T_m=81.7$  °C) to the pool in R10 ( $T_m=84$  °C), in good agreement with the enrichment found by UV-vis and SPR.

❖ Family I: Sequences with direct primer inserted in the sequence

	10	20	30	40	50	$\Delta G$ (kcal/mol)
1	.... .... .... .... .... .... .... .... .... .... ....	GCAAGCAATGGCGACCCGAACAGTAGGATAGGGTTGATAGGTTAAGAG				-3.54
5	-----GGGA.ACGAGGAGC-CTAG.A.....					-3.86
6	---ACTCAGTC.AT..AGAATTAC.G...G.....CTATATA					-2.30
10	-----TCTATCGAAATTGATTCCCT.G.....					-1.72
13	----GG.GATTTCG..A.TGTTT.CCCT.....					-1.39
14	-----CCTT.GAGTAGTACCC.....					-3.91
16	-----CA.TCCTG.ACCTCCAT.AC.....C-----					-1.71
18	-----TA..-G.CC.....CTATATA					-1.56
19	---AGCTGG.GCA.TCTAAACT.G..TTCG.....C-----					-2.63
22	-----TA.GTG.GCAGTCGATC.....CTATAA-					-0.69
23	-----G..GGGATGAT..G..T.....C-----					+1.42
25	---A.CC.GTG..GTGTTGATCGGCCCT.....C-----					-4.51
29	-----G.TGC..C.....C-----					-0.10
35	---CAGCTATCACGATGAGC-CT.GCAC.....					-3.76
36	-----AG...AG.GGTCC.....					-1.53

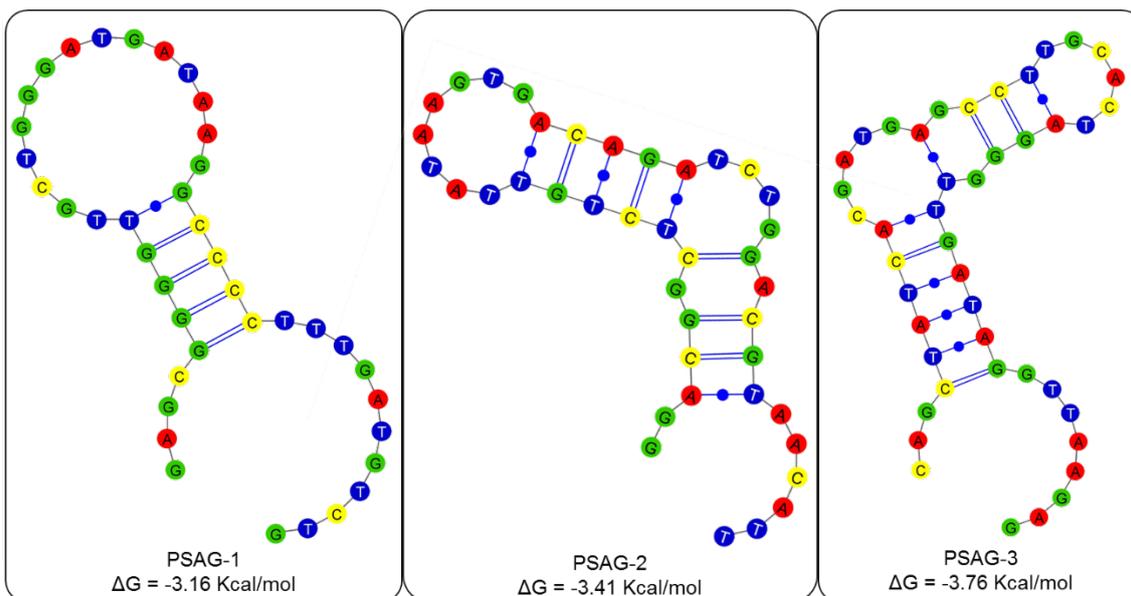
❖ Family II: Sequences similar to PSA-1 aptamer

	10	20	30	40		$\Delta G$ (kcal/mol)
PSA-1	.... .... .... .... .... .... .... .... .... ....	GGACGGTTGCGCTATATTTAACCAAAAGTCTGGAT-TAAC-				-3.07
3	..AG...GGTT...GGGA.G.TAAGGCCCT.T...G.CTG----					-3.16
12	---TC.CATT..GGGAG..GAT-TGC...TA..GGCGAGTTC					-4.12
15	.....C.CT.T...AG.G...-GA.....CG.....TT-					-3.41

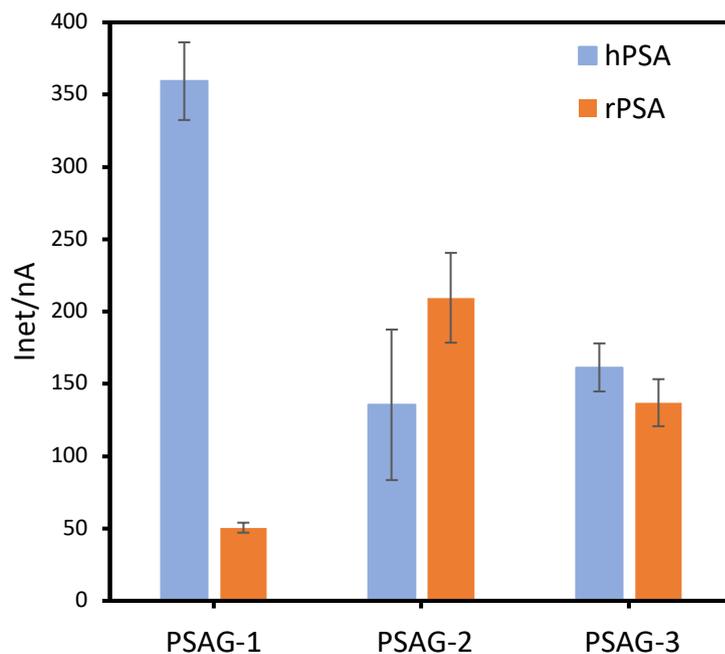
❖ Family III: Unrelated sequences

	10	20	30	40		$\Delta G$ (kcal/mol)
4	.... .... .... .... .... .... .... .... .... ....	GGATCAGAATCTTCCACTCCCTGGCATTATTCATTGCCAA				-4.31
7		GTAAACTACTGAAGGGTCTCCTGGGGTATGTCGAGTATG				-1.43
8		ACAGGCTAGTTC AAGGACAGCAGCCGCCATCAAAATGCCG				-2.55
9		TGCGAAAAGTGGGCTTCGTAAC TACTTGCACTTCGCGACCC				-3.13
11		CGGTATAGGTTAATGAGAGGATAGCGTTATCAGGACTAGT				-1.63
17		GTAGTCTCTAATCAACGTGAGTGTGTAACCGCAGTACGCC				-1.96
20		TTAGCGGATTGGTCAATTCGACACTTGTTGAGCATATGTG				-3.06
21		CTGATACATTACAAGTAGATCCTGGTCCACGAACTGTGCA				-2.78
24		TAGGAAGGACTCCATTTCACGATACGGTGGCACAGGGACG				-3.12
26		CTCGGGTGGGTCAGGTATATGAGTACAGTTTTTCGAGTCT				-0.57
27		CCAGGCAGATTCTCTAGAGCTAAACAATTGAGTAGGCTGT				-1.06
28		GATGTCAACTAGCTGTTGGGGAGATGAGGGATATGTT				-3.05
30		CCATTGGTTGCTAAGGATGACGAATTTCCCAT AAGCGGTA				-0.46
31		TTAAGCGTATCTCTGATGGATGGATACGATACCGAACAT				-5.20
32		GGGTAGCTGGGATACGGCATTTCCTACCGCGCCTGGAT				-3.61
33		GATGTCAACTAGCTGTTGGACTGGGATGTC TTAATGCTC				-3.26
34		TTATGGCAGGGCAGTGGTATCTTAAACCCTGAAGTCAGCTG				-4.87
37		ATATACGATGTCAACTAGCTGTTGGACGGTAGCTATAAAATTT				-3.05

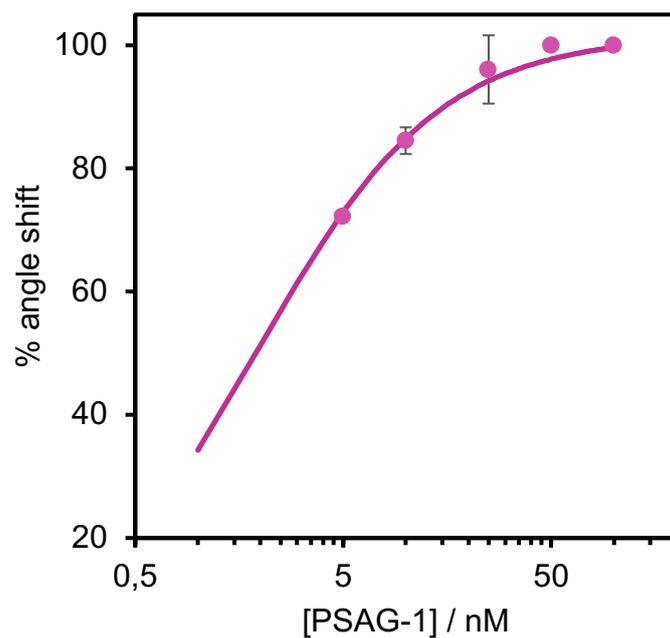
Figure S2. Sequence analysis. Sequences obtained from pool 10 grouped into three families after trimming their primer-binding sequences and their corresponding lowest free energies predicted by using mfold<sup>5</sup>.



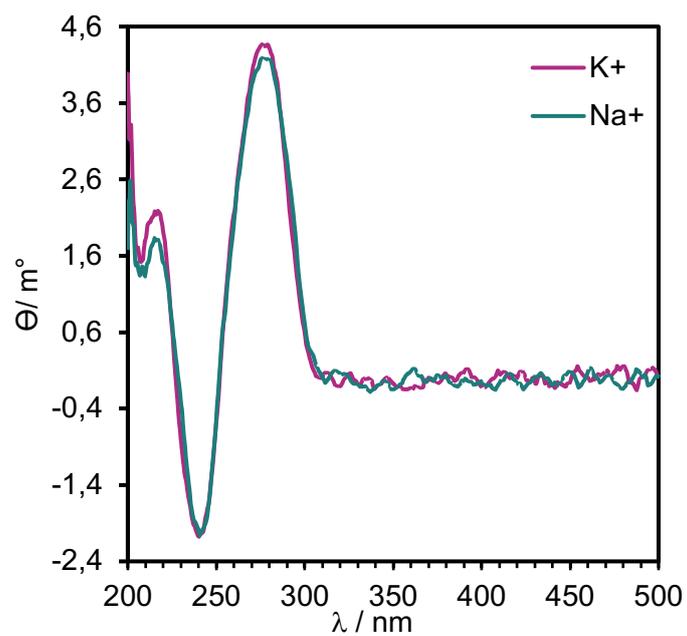
**Figure S3. Secondary structure of the aptamers selected.** The secondary structures of the clones selected from the 10th round of the SELEX drawn using VARNAGui software and its free energy calculated using mfold<sup>5</sup>. PSAG-1 corresponds to clone 3, PSAG-2 is clone 15 and PSAG-3 clone 35.



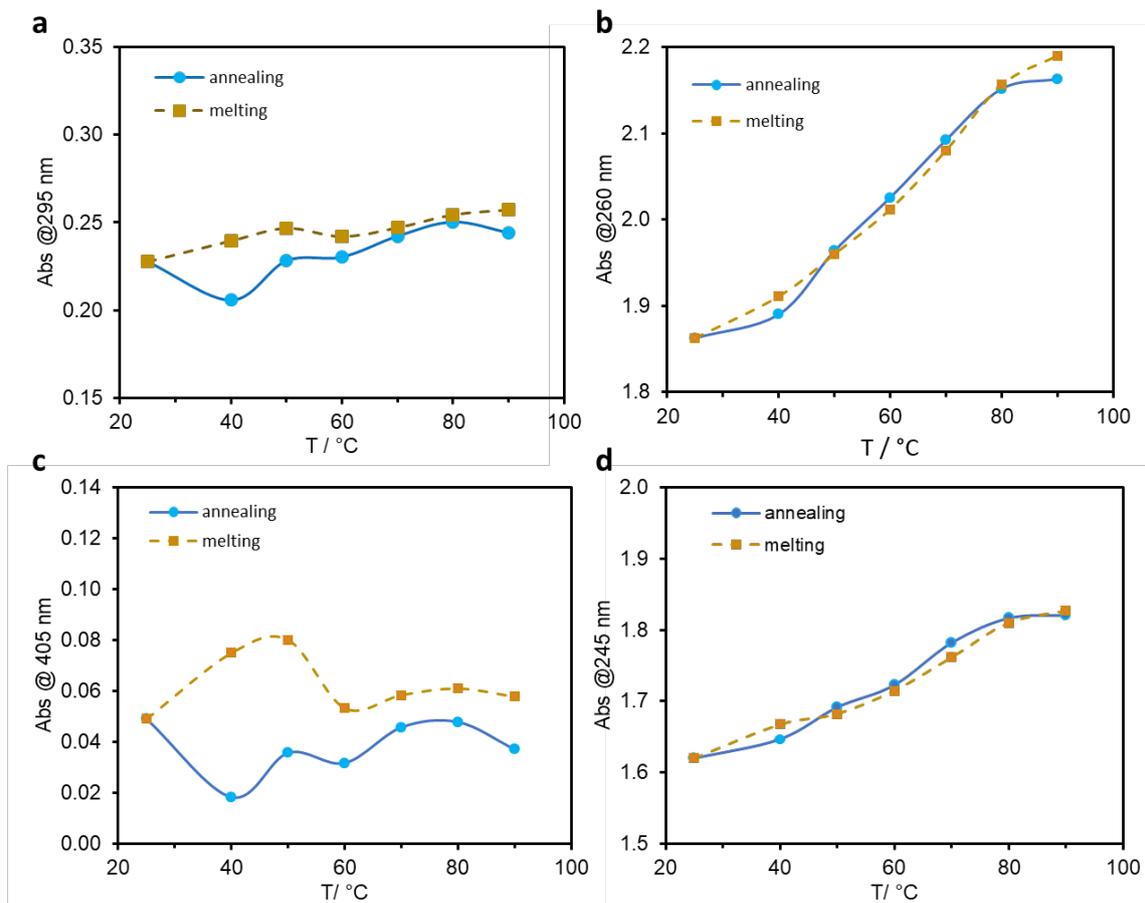
**Figure S4. Selectivity analysis of the selected clones.** Net currents obtained in SPAuE modified with hPSA (blue) or rPSA (orange) through a covalent amide bond to a SAM of mercaptoundecanoic acid and mercaptohexanol mixture. The fluorescein-tagged aptamers were labelled with Fab-antifluorescein-POD conjugate and measurement was carried out by chronoamperometry at -0.2 V after addition of the enzyme substrate (TMB+H<sub>2</sub>O<sub>2</sub>) for 30 s. Only PSAG-1 is able to clearly discriminate glycosylated from unglycosylated PSA forms.



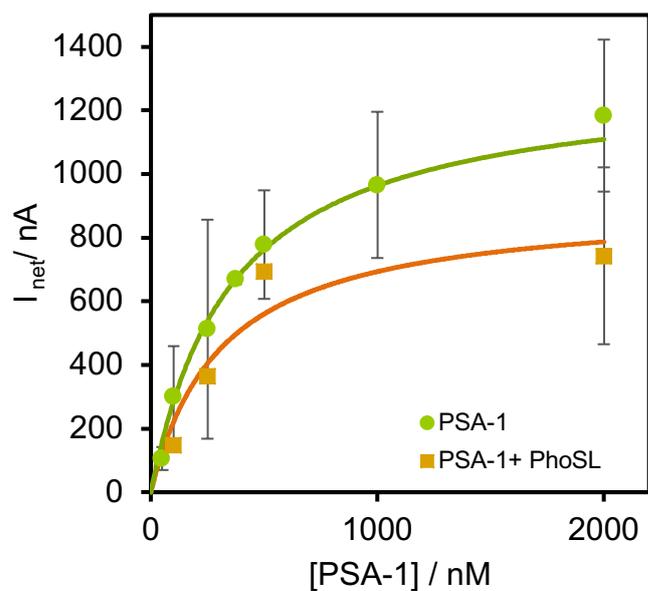
**Figure S5. SPR binding curve.** Data were obtained from the binding of increasing amounts of PSAG-1 to the SPR chip modified with hPSA, in PBS-K. Curve was fitted to the Langmuir model, resulting in a  $K_d$  value of  $1.9 \pm 0.2$  nM.



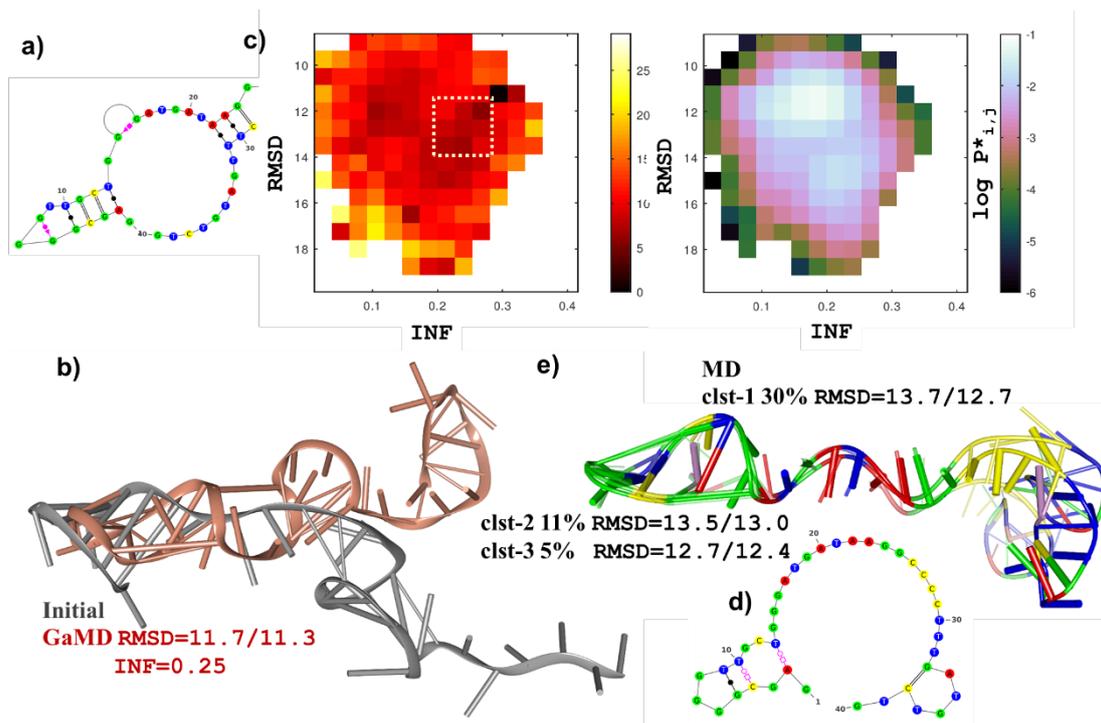
**Figure S6. Circular dichroism spectra.** CD spectra of 5 μM PSAG-1 measured in phosphate buffer pH 7.4 containing 154 mM Na<sup>+</sup> (blue) or K<sup>+</sup> ions (magenta).



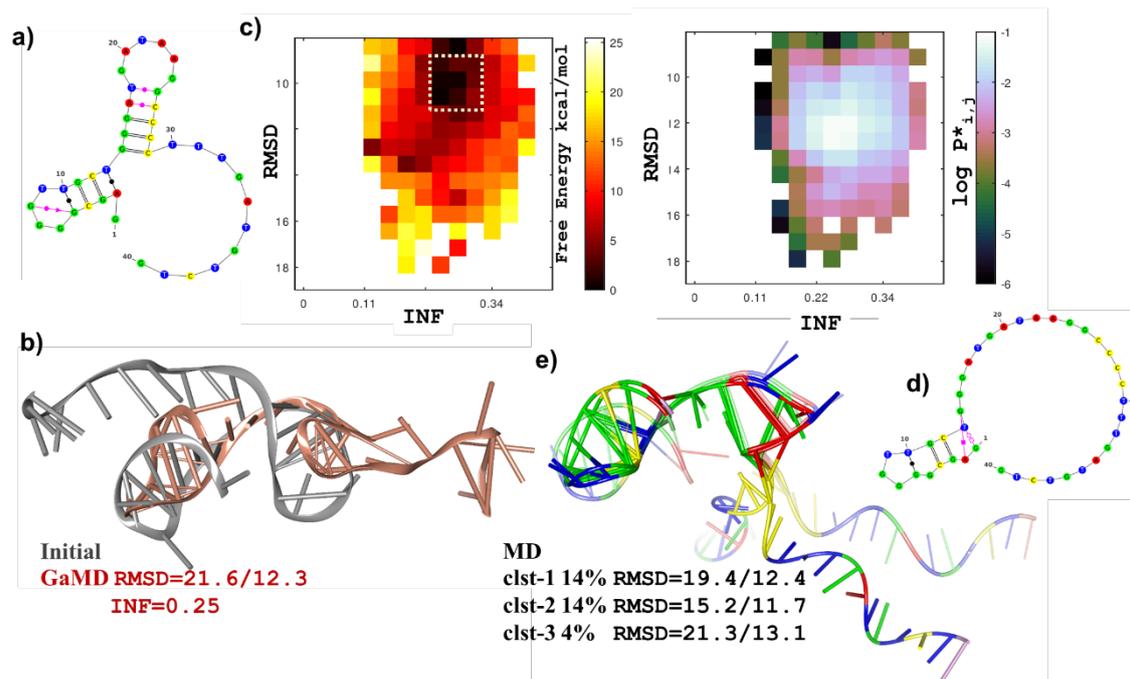
**Figure S7: UV melting curves.** Representative absorbance plots vs temperature at 295 nm (a), 260 nm (b) 405 nm (c) and 245 nm (d) of 5  $\mu$ M PSAG-1 in PBS buffer containing  $K^+$  ions, pH 7.4. The annealing cycle (cooling) is shown in blue circles while the melting cycle (heating) is in brown squares. The absence of an inverted melting curve (negative slope) at 295 nm is a strong evidence that the structure is not a G-quadruplex. At 260 nm the typical melting curve is obtained while at high wavelength (405 nm) the absorbance is mostly unchanged and very low as expected indicating that there are not artifacts affecting the measurement. To further verify the absence of G-quadruplex the plot at 245 nm is also shown.



**Figure S8:** Binding curve obtained for increasing concentrations of PSA-1 to hPSA (green circles) and to PhoSL-blocked hPSA (brown squares) in PBS-Na buffer, both immobilized on SPAuE.

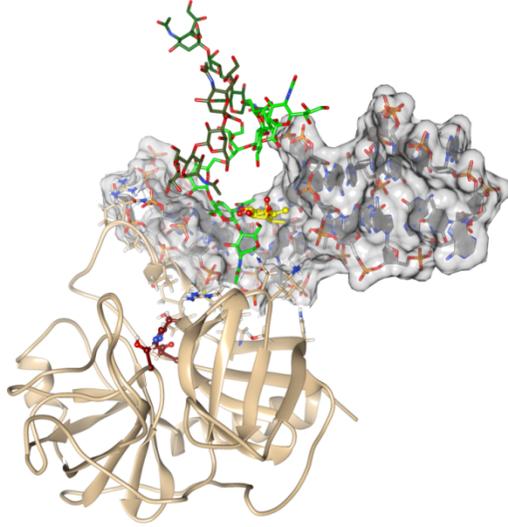


**Figure S9. PSAG-1 model A2** a) Secondary structure of the initial **A2** model for PSAG-1. Base pairing is represented using the Leontis Westhof graphical conventions. b) Ribbons representation of the superposition of the initial **A2** model and the selected GaMD snapshot. RMSD data (in Å; considering CNOPS atoms in residues 3-38/5-30) between both structures are also given. c) Free-energy (left) and logarithm of the GaMD population (right) maps in terms of the RMSD (residues 3-38) and INF structural indexes. The dashed line encloses the free energy basin with significant population from which the GaMD snapshot shown in b) was selected. e) Superposition of the cluster representatives of the three most populated clusters derived from the equilibrium MD simulation. Percentage population and RMSD values (in Å) with respect to the initial **A2** structure are also given.

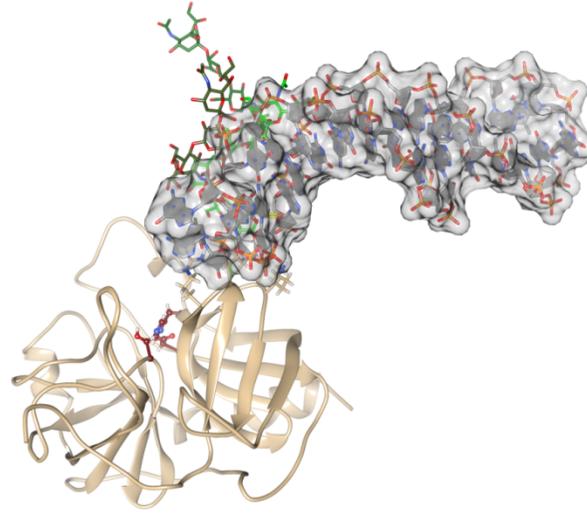


**Figure S10. PSAG-1 model A3** a) Secondary structure of the initial **A3** model for PSAG-1. Base pairing is represented using the Leontis Westhof graphical conventions. b) Ribbons representation of the superposition of the initial **A3** model and the selected GaMD snapshot. RMSD data (in Å; considering CNOPS atoms in residues 3-38/5-30) between both structures are also given. c) Free-energy (left) and logarithm of the GaMD population (right) maps in terms of the RMSD (residues 3-38) and INF structural indexes. The dashed line encloses the free energy basin with significant population from which the GaMD snapshot shown in b) was selected. d) Secondary structure of the fully relaxed **A3** model obtained from the major cluster representative. e) Superposition of the cluster representatives of the three most populated clusters derived from the equilibrium MD simulation. Percentage population and RMSD values (in Å with respect to the initial **A3** structure are also given.

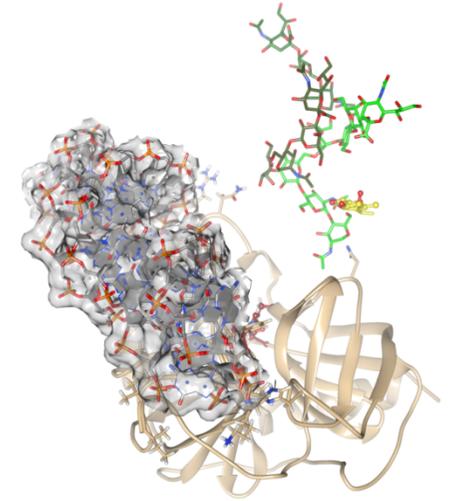
clst-01 -15.77



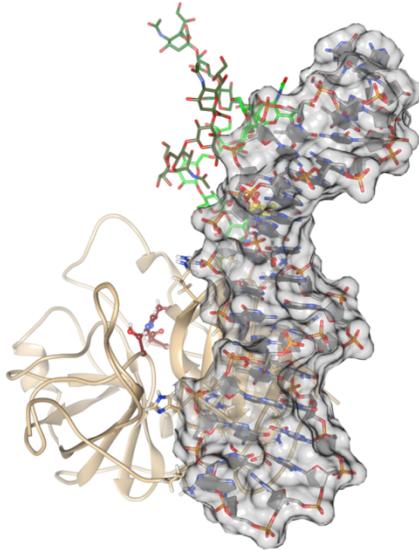
clst-02 -13.79



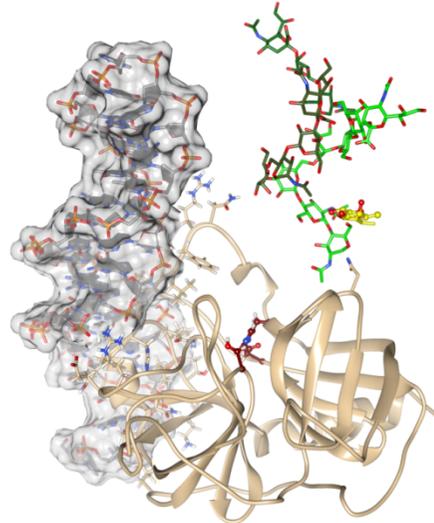
clst-04 -12.97



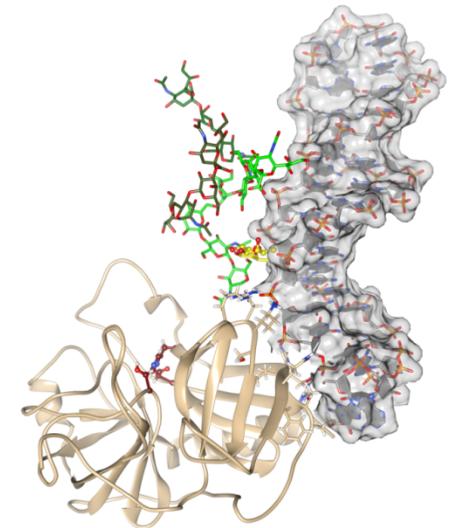
clst-08 -11.89

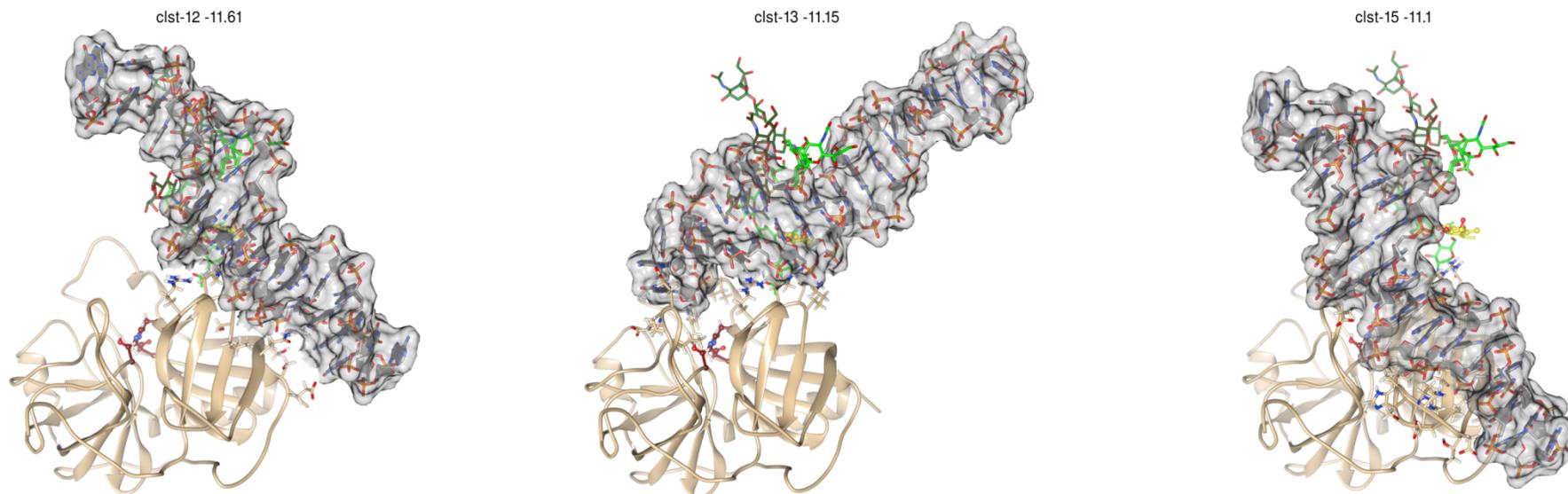


clst-09 -11.81

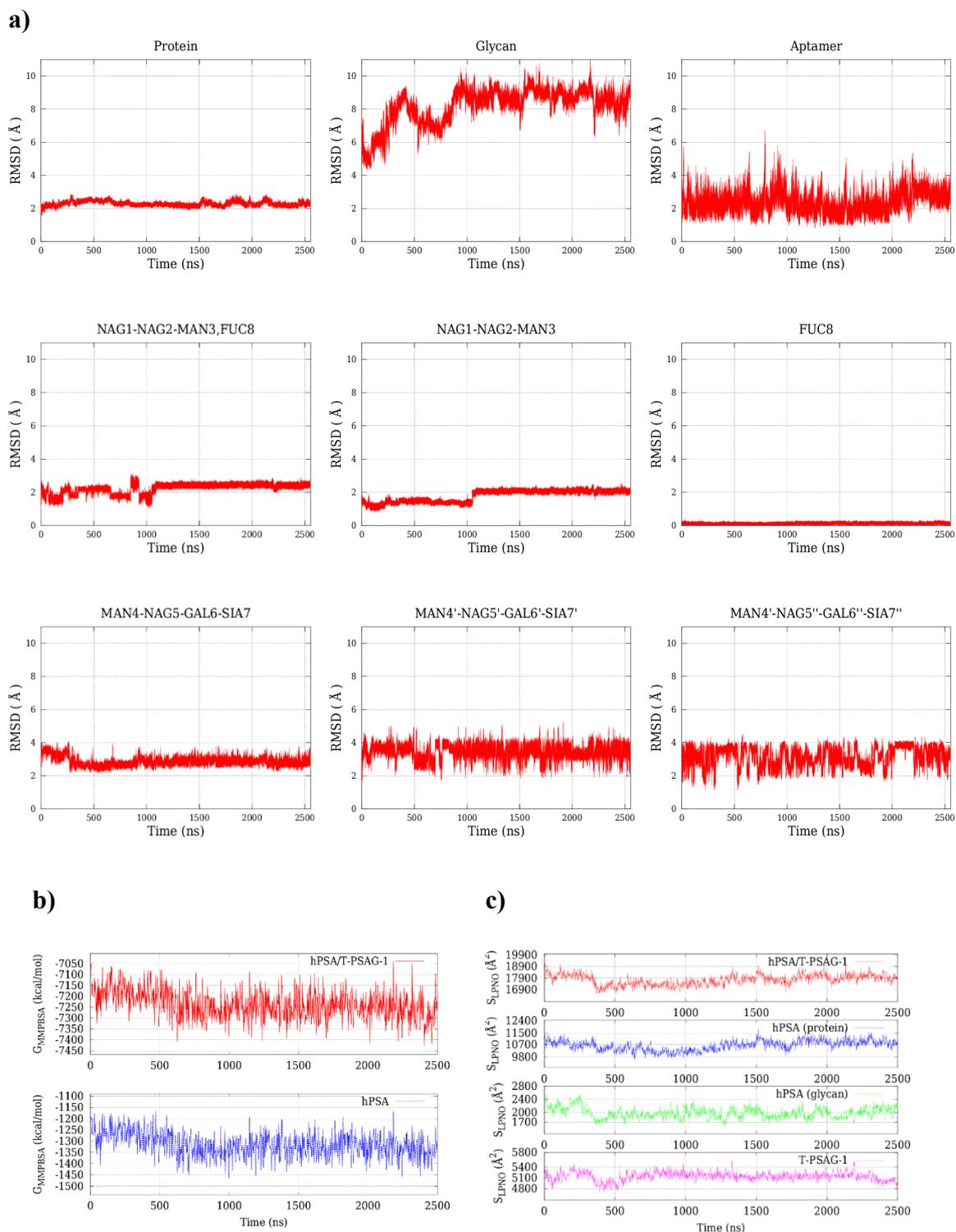


clst-10 -11.81

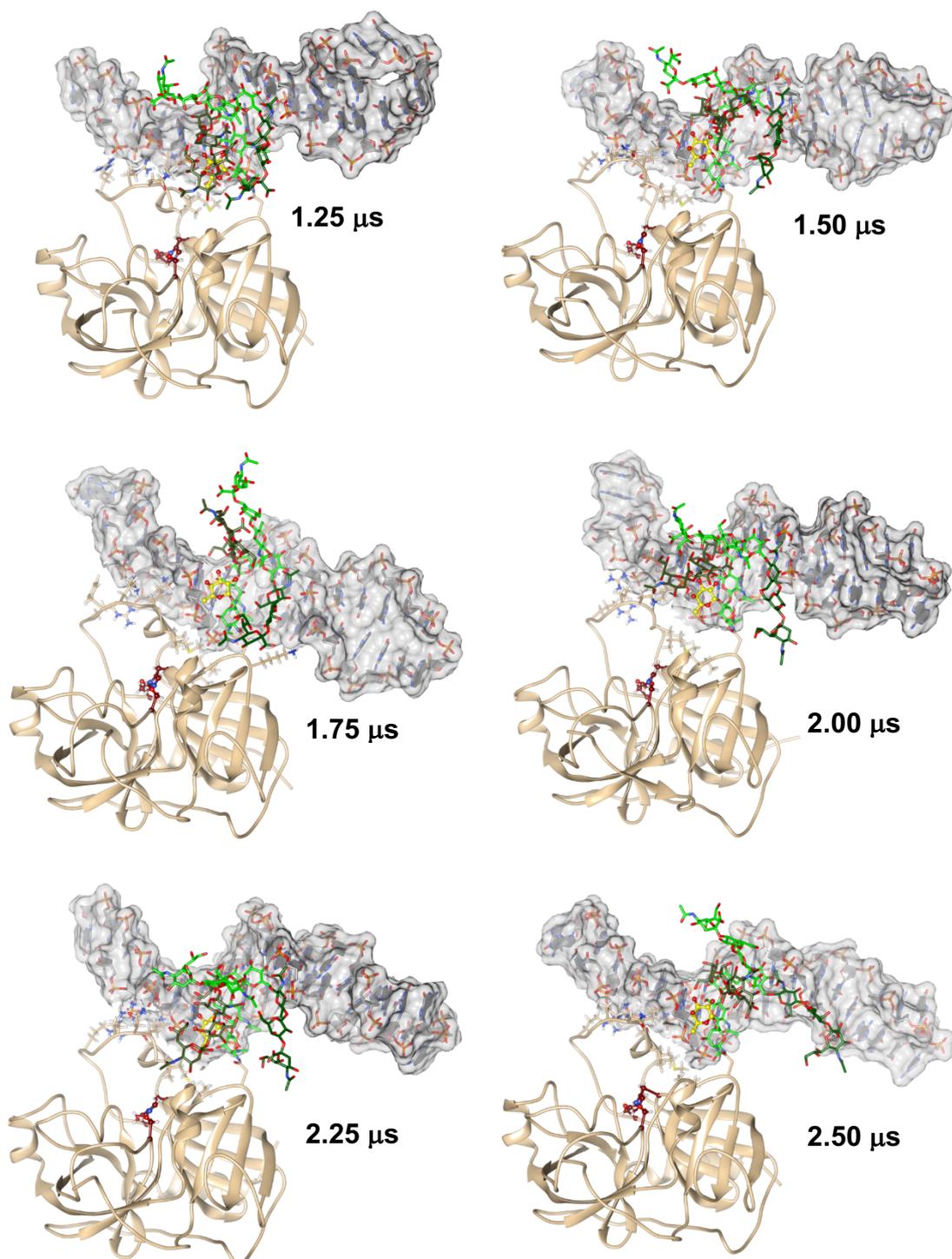




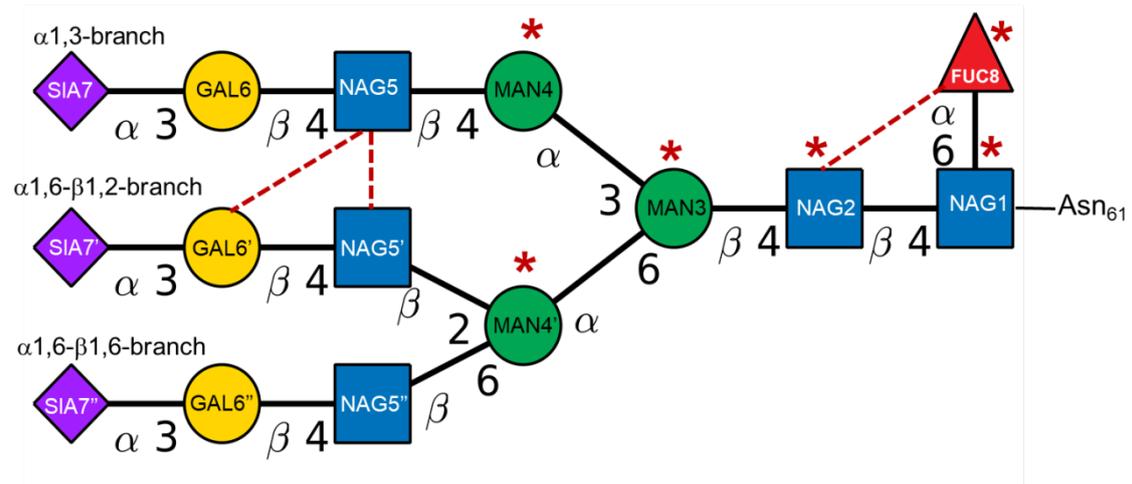
**Figure S11. Ribbon and stick representation of selected docking poses exhibiting diverse binding modes.** The molecular surface of the T-PSAG-1 aptamer is displayed in a translucent mode. The orientation of the hPSA protein is identical in all the structures. The  $\Delta_{\text{bind}}G$  scoring (kcal/mol) is also indicated.



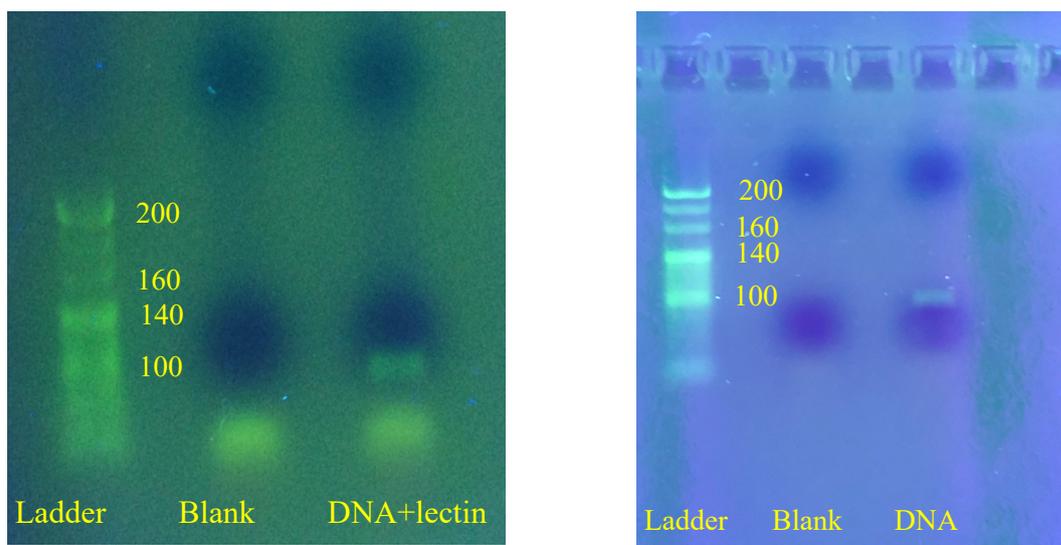
**Figure S12.** Time evolution of structural and energetic descriptors along the 2.5  $\mu$ s MD simulation of the hPSA/T-PSAG-1 complex. **a)** Root mean squared deviation (RMSD, in  $\text{\AA}$ ) plots for various structural elements: the backbone heavy atoms of PSA, the heavy atoms of the glycan moiety, the heavy atoms of the T-PSAG-1 aptamer, and selected sugars of the glycan moiety. The best docking pose obtained for the PSA/T-PSAG-1 complex is the reference structure. **b)** MM-PBSA energy (kcal/mol). **c)** Molecular surface (in  $\text{\AA}^2$ ) computed with the LCPO method.



**Figure S13. Ribbon and stick representation of selected snapshots extracted from the last half of the MD simulation.** The molecular surface of the T-PSAG-1 aptamer is displayed in a translucent mode. The orientation of the hPSA protein is identical in all the structures.



**Figure S14.** Triantennary oligosaccharide linked to the Asn<sub>61</sub> side chain of the hPSA protein in the 3QUM X-ray structure (molecule P). Sugar units marked with an asterisk give close contacts with the T-PSAG-1 aptamer according to the MD simulation. The most abundant sugar-sugar contacts are also indicated.



**Figure S15. Negligible influence of the lectin PhoSL in the PCR amplification.** 2 % agarose gels show the size ladder (left lane), a blank PCR with no library (central lane), and the PCR product of 2.5 nM initial library in the presence of 2  $\mu$ M PhoSL lectin (right lane, left panel) and in the absence of lectin (right lane, right panel).

## 7 References

1. A. Cabanettes, L. Perkams, C. Spies, C. Unverzagt, A. Varrot, *Angew. Chem. Int. Ed.* **2018**, *57*, 10178-10181.
2. Y. Kobayashi et al., *J. Biol. Chem.* **2012**, *287*, 33973-33982.
3. A. Krüger, F. M. Zimbres, T. Kronenberger, C. Wrenger, *Biomolecules* **2018**, *8*, 83.
4. I. Jeddi, L. Saiz, *Sci. Reports* **2017**, *7*, 1178
5. M. Zuker, *Nucleic Acids Res.* **2003**, *31*, 3406-3415.
6. M. Popena, M. Szachniuk, M. Antczak, K.J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, R.W. Adamiak, *Nucleic Acids Res.* **2012**, *40*, e112
7. M. Popena et al., *BMC Bioinformatics* **2010**, *11*, 231.
8. Discovery Studio Modeling Environment, Dassault Systèmes BIOVIA, **2016**.
9. R. Salomon-Ferrer, D.A. Case, R.C. Walker, *WIRE Comput. Mol. Sci.* **2013**, *3*, 198-210.
10. D.A. Case et al., *J. Comput. Chem.* **2005**, *26*, 1668-1688.
11. D.A. Case et al., AMBER 2018, University of California, **2018**.
12. I. Ivani et al., *Nat. Methods* **2016**, *13*, 55-58.
13. T.E. Cheatham III, D.A. Case, *Biopolymers* **2013**, *99*, 969-977.
14. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, *J. Chem. Phys.* **1983**, *79*, 926-935.
15. J. Åqvist, *J. Phys. Chem.* **1990**, *94*, 8021-8024.

16. J-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327-341.
17. U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, L.G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577–8593.
18. Y. Miao, V.A. Feher, J.A. McCammon, *J. Chem. Theory Comput.* **2015**, *11*, 3584-3595.
19. G. Palermo, Y. Miao, R.C. Walker, M. Jinek, J.A. McCammon, *Proc. Natl. Acad. Sci.* **2017**, *114*, 7260-7265.
20. S. Le Grand, A.W. Götz, R.C. Walker, *Comput. Phys. Commun.* **2013**, *184*, 374-380.
21. M. Parisien, J.A. Cruz, E. Westhof, F. Major, *RNA* **2009**, *15*, 1875-1885.
22. X.J. Lu, H.J. Bussemaker, W.K. Olson, *Nucleic Acids Res.*, **2015**, *43*, e142.
23. Y. Miao, W. Sinko, L. Pierce, D. Bucher, R.C. Walker, J.A. McCammon, *J. Chem. Theory Comput.* **2014**, *10*, 2677-2689.
24. E.F. Pettersen et al., *J. Comput. Chem.* **2004**, *25*, 1605-1612.
25. K. Darty, A. Denise, Y. Ponty, *Bioinformatics* **2009**, *25*, 1974-1975.
26. N.B. Leontis, E. Westhof, *RNA* **2001**, *7*, 499-512.
27. D.R. Roe, T.E. Cheatham, *J. Chem. Theory Comput.* **2013**, *9*, 3084-3095.
28. P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, *Acc. Chem. Res.* **2000**, *33*, 889-897.

29. E. Wang et al., *Chem. Rev.* **2019**, *119*, 9478-9508.
30. P. Grochowski, J. Trylska, *Biopolymers* **2008**, *89*, 93-113.
31. Q. Cai, M-J. Hsieh, J. Wang, R. Luo, *J. Chem. Theory Comput.* **2010**, *6*, 203-211.
32. C. Tan, Y-H. Tan, R. Luo, *Phys. Chem. B* **2007**, *111*, 12263-12274.
33. E.A. Stura, B.H. Muller, M. Bossus, S. Michel, C. Jolivet-Reynaud, F. Ducancel, *J. Mol. Biol.* **2011**, *414*, 530-544.
34. R. Anandakrishnan, B. Aguilar, A.V. Onufriev, *Nucleic Acids Res.* **2012**, *40*, W537-W541.
35. W-P. Hu, J.V. Kumar, C-J. Huang, W-Y. Chen, *BioMed Res. Int.* **2015**, *8*.
36. M. Heiat, A. Najafi, R. Ranjbar, A.M. Latifi, M.J. Rasaei, *J. Biotechnol.* **2016**, *230*, 34-39.
37. G. Oliviero et al., *Anal. Chem.* **2016**, *88*, 2327-2334.
38. Y-H. Lao, H-Y. Chiang, D-K. Yang, K. Peck, L-C. Chen L-C., *Chem. Commun.* **2014**, *50*, 8719-8722.
39. R. Ahirwar, S. Nahar, S. Aggarwal, S. Ramachandran, S. Maiti, P. Nahar, *Sci. Rep.* **2016**, *6*.
40. I.A. Baig, J-Y. Moon, S-C. Lee, S-W, Ryoo, M-Y. Yoon, *Biochim. Biophys. Acta* **2015**, *1854*, 1338-1350.
41. G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, *J. Comput. Chem.* **2009**, *16*, 2785-2791.

42. R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, *Comput. Chem.* **2007**, *28*, 1145-1152.
43. P.A. Holt, J.B. Chaires, J.O. Trent, *J. Chem. Inf. Model* **2008**, *48*, 1602-1615.
44. J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696-3713.
45. K.N. Kirschner et al., *J. Comput. Chem.* **2008**, *29*, 622-655.
46. G.D. Hawkins, C.J. Cramer, D.G. Truhlar, *J. Phys. Chem.* **1996**, *100*, 19824-19839.
47. J. Weiser, P.S. Shenkin, W.C. Still, *J. Comput. Chem.* **1999**, *20*, 217-230.
48. M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, E. Kaxiras, *J. Chem. Phys.* **2001**, *114*, 5149-5155.