Structure of a DNA G-quadruplex that modulates SP1 binding sites architecture in HIV-1 promoter

Aurore De Rache, Julien Marquevielle, Serge Bouaziz, Brune Vialet, Marie-Line Andréola, Jean- Louis Mergny, Samir Amrane

▶ **To cite this version:**

Aurore De Rache, Julien Marquevielle, Serge Bouaziz, Brune Vialet, Marie-Line Andréola, et al.. Structure of a DNA G-quadruplex that modulates SP1 binding sites architecture in HIV-1 promoter. 2023. ⟨hal-04275486⟩

HAL Id: hal-04275486

https://hal.science/hal-04275486v1

Preprint submitted on 8 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Structure of a DNA G-quadruplex that modulates SP1 binding sites architecture in HIV-1 promoter

Aurore De Rache[1,2,§,+], Julien Marquevielle[1,2,+], Serge Bouaziz[5], Brune Vialet[1,2], Marie-Line Andréola[1,3], Jean-Louis Mergny[4] and Samir Amrane[1,2*]

[1] Université de Bordeaux, Bordeaux, France.

[2] ARNA laboratory, INSERM U1212, CNRS UMR 5320, IECB, Bordeaux, France.

[3] MFP laboratory, UMR5234, CNRS, Bordeaux, France.

[4] Laboratoire d'Optique & Biosciences, École Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, Palaiseau, France.

[5] UMR 8038 CNRS, Faculté de Pharmacie de Paris, Université Paris Descartes, Sorbonne Paris Cité

[§] Present address: Department of Chemistry, U. Namur, 61 rue de Bruxelles, B5000 Namur, Belgium

[+] These authors contributed equally to this work
* Corresponding author: Samir Amrane; Email: samir.amrane@inserm.fr

Highlight :

- **Context:** Nucleic acid sequences containing guanine tracts can form G-quadruplexes (G4s), impacting gene regulation.
- **Principal results**: HIVpro2 DNA sequence, derived from the HIV-1 promoter, forms a hybrid G4 structure with a single-nucleotide bulge. HIVpro2 G4 can modulate SP1 binding sites architecure, potentially regulating viral transcription and latency.
- **Conceptual advance**: Discovery of a G4 structure in HIVpro2 suggests a novel mechanism for HIV-1 gene regulation.
- **Significance**: Understanding the structural switch between G4s and canonical duplexes may offer new strategies for HIV-1 therapy and latency control.

Manuscript (docx or tex file) Click here to view linked References

Structure of a DNA G-quadruplex that modulates SP1 binding sites architecture in HIV-1 promoter

Aurore De Rache[1,2,§,+], Julien Marquevielle[1,2,+], Serge Bouaziz[5], Brune Vialet[1,2], Marie-Line Andréola[1,3], Jean-Louis Mergny[4] and Samir Amrane[1,2*]

[1] Université de Bordeaux, Bordeaux, France.

[2] ARNA laboratory, INSERM U1212, CNRS UMR 5320, IECB, Bordeaux, France.

[3] MFP laboratory, UMR5234, CNRS, Bordeaux, France.

[4] Laboratoire d'Optique & Biosciences, École Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, Palaiseau, France.

[5] UMR 8038 CNRS, Faculté de Pharmacie de Paris, Université Paris Descartes, Sorbonne Paris Cité

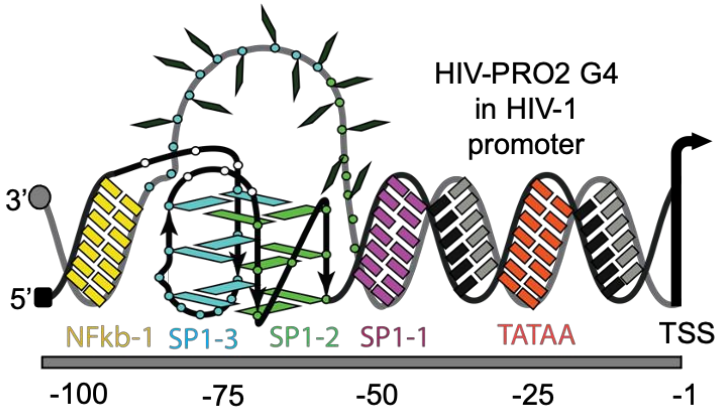
[§] Present address: Department of Chemistry, U. Namur, 61 rue de Bruxelles, B5000 Namur, Belgium

[+] These authors contributed equally to this work

* Corresponding author: Samir Amrane; Email: samir.amrane@inserm.fr

Keywords
DNA structure; SP1 Transcription Factor; Gene expression regulation; HIV-1 promoter, G-Quadruplex

Graphical abstract



ABSTRACT

Nucleic acid sequences containing guanine tracts are able to form non-canonical DNA or RNA structures known as G-quadruplexes (or G4s). These structures, based on the stacking of G-tetrads, are involved in various biological processes such as gene expression regulation. Here, we investigated a G4 forming sequence, HIVpro2, derived from the HIV-1 promoter. This motif is located 60 nucleotides upstream of the proviral Transcription Starting Site (TSS) and overlaps with two SP1 transcription factor binding sites. Using NMR spectroscopy, we determined that HIVpro2 forms a hybrid type G4 structure with a core that is interrupted by a single nucleotide bulge. An additional reverse-Hoogsteen AT base pair is stacking on top of the tetrad. SP1 transcription factor is known to regulate transcription activity of many genes trough the recognition of Guanine-rich duplex motifs. Here, the formation of HIVpro2 G4 may modulate SP1 binding sites architecture by competing with the formation of the canonical duplex structure. Such DNA structural switch potentially participates to the regulation of viral transcription and may also interfere with HIV-1 reactivation or viral latency.

INTRODUCTION

G-quadruplexes (or "G4s") are non-canonical nucleic acid structures formed by guanine-rich DNA or RNA sequences. They are based on the formation of G-tetrads, formed by the arrangement of four guanines connected by eight hydrogen bonds. Potassium or sodium cations, abundantly present in the cellular environment, stabilize these G-tetrads thanks to specific electrostatic interactions with the oxygen of the carbonyl groups of the four guanines. The stacking of 2 or more G-tetrads constitutes the core of the G4 structure which is composed, at each corner, of four guanine pillars or strands. G4s are polymorphic structures that can sample a vast space of conformations (reviewed here (1)). Indeed, in the case of intramolecular G4s, the four strands can be interconnected by three different types of loops (lateral, diagonal or propeller) of variable sizes and sequences. Consequently, each of the four pillars can be oriented in a parallel or antiparallel orientation to its adjacent strands. Thermodynamics and kinetics studies have shown that G4s are often stable far above the physiological temperature (Tm>>37 °C) (2) and can form within a few tens of milliseconds (3). Hence, the cell has certainly evolved mechanisms that use the valuable properties of G4s *in vivo*. Indeed, an increasing number of publications demonstrate their involvement in key biological processes (reviewed here (4)). Several *in cellulo* studies using specific G4 probes (antibodies or ligands) have confirmed the existence of G4s in the genomic DNA and cellular RNA (5-7) and their involvement in telomeres dynamics (8), RNA splicing /translation (9), DNA replication and transcription (10).

In the case of transcription, genome wide bioinformatics studies have identified 1.5 million putative G4s (pG4s) in the Human genome with up to 66% of the human promoters presenting at least one pG4. Interestingly, important transcription factor binding sites, such as Sp1, MAZ, Krox and ZF5 are positioned near or overlaps the pG4s (11) suggesting that G4 formation within promoters plays an important role in transcription regulation (reviewed here (12)). G4 formation at promoters has been directly related to the level of chromatin compaction, i.e. G4s are enriched in relaxed nucleosome-depleted regions (NDRs) located upstream of transcription start sites (TSSs) (11). Furthermore, an increase in the proportion of G4s formed at promoters has been observed when inducing chromatin relaxation by histone deacetylase inhibitors (13). In some cases, once these G4s are formed, they are able to enhance gene transcription by recruiting transcription factors involved in RNA pol II machinery (14). This is supported by the high G4-binding affinity observed in vitro for Sp1 (15), CNBP (16), and LARK (17) transcription factors. G4s can even further stimulate transcription by acting as a hub that enables the simultaneous recruitment of a variety of transcription factors (18).

G4s have also been involved in the replication cycle of a number of viruses such as HIV-1, HCV (19), coronaviruses (20) and many others reviewed here (21). HIV-1 is a RNA retrovirus that infects CD4 cells and induces a deficiency of the immune system causing AIDS disease. HIV-1 viral particle carries two identical RNA genomes. Shortly after infection, the two viral RNAs are reverse-transcribed into double-stranded DNAs. The DNA proviral genome is then integrated in the genetic material of the infected cell. After integration, the provirus uses the transcription machinery of the host cell to transcribe its genetic content. We have recently identified ten evolutionary conserved G4 forming sequences in the HIV-1 genome (22). Most of these G4s contain crucial regulatory elements such as the PPT and cPPT sequences as well as the U3 region. Interestingly, proviral transcription is regulated by a G-rich sequence located on the U3 region of the viral promoter. This sequence contains the three Sp1 transcription factor binding sites. It is located 50 nucleotides upstream of the transcription-starting site, next to the TATA box, on the U3 region

of the 5' LTR (Figure 1A, B). We previously analyzed the central part of this G-rich region spanning the three Sp1 binding sites, from the 5' extremity of Sp1-3 to the 3' extremity of Sp1-1 (HIVpro1) (Figure 1B). It formed a stable two G-tetrad antiparallel G4 with an additional Watson−Crick CG base pair (23). The structure of a second G4 spanning the two first Sp1 binding sites has also been solved (LTR-IV) (Figure 1B). It formed a parallel-stranded G-quadruplex containing a single-nucleotide thymine bulge (24). A more recent study showed the formation of a G4−duplex hybrid structure (LTR-III) (Figure 1B) that spans the 3 Sp1 binding sites (25).

Along this line, we characterize here the structure of a fourth possible G4 conformation adopted by the G-rich fragment of HIV-1 promoter. This 22 nucleotides segment, referred to as HIVpro2, spans the second and third Sp1 transcription factor binding sites (5'-AGGGAGGTGTGGCCTGGGCGGG-3'). It forms a three G-tetrads structure with a hybrid type G4 core that is interrupted by a single nucleotide bulge. An additional reverse Hoogsteen type AT base pair stacks on top of the upper tetrad. Such G4 structure is potentially able to form during transcription; therefore, a conformational interplay between these G4s might intervene in the regulation of promoter activity.

MATERIAL AND METHODS

Oligonucleotides: Site-specific 5% ^{15}N and ^{13}C-labelled oligonucleotides were synthesized relying on the phosphoramidite methodology and using an H8 automated synthesizer (K&A Labs, Germany) at one micro molar scale on 1000 Å primer support (Link Technologies SynBase CPG). Necessary standard phosphoramidites reagents (Bz-dA, dT, iBu-dG and Ac-dC) and solvents were purchased from Glen Research. The labelled Bz-dA, iBu-dG, and dT phosphoramidites ($^{13}C10$, 98%; $^{15}N5$; 98%, CP 95 %) were purchased from Cambridge Isotope Laboratories. The cleavage and deprotection of the oligonucleotides were performed by incubation with ammonium hydroxide at 55 °C for 16h. All oligonucleotide solutions were prepared in 20 mM phosphate buffer pH 6.9 with 70 mM KCl.

Circular dichroism (CD) spectroscopy: CD-spectra were measured in 1 cm path-length quartz cells using a Jasco J-815 equipped with a Peltier temperature controller. Individual spectra correspond to the average of five scans recorded at a speed of 50 nm min^{-1} with a bandwidth of 2 nm and a data integration time of 1s. The sample concentration was 2.5 µM. CD-melting experiment was performed by increasing the temperature by 0.2 °C min^{-1} between 20 and 95 °C and recording a full spectrum (3 accumulations) every 1°C.

UV spectroscopy and melting: UV spectroscopic measurements were performed using a SAFAS UVmc2 double-beam spectrophotometer (Monte Carlo, Monaco). The temperature was controlled with a thermostatable 10-cell holder regulated by a high-performance Peltier temperature controller. Thermal absorption difference spectra (TDS) were obtained by subtraction of the spectrum recorded at 3.5°C from the one obtained at 90.4°C. Melting temperatures were determined from the variation of the absorbance at 295 nm as a function of temperature. Starting from 95 °C, the temperature was decreased at a rate of 0.2°C min^{-1} to 2°C and then heated back up to the initial temperature at the same rate.

NMR experiments: Except otherwise stated, NMR experiments were performed at 20°C either on a 700 MHz or on an 800 MHz Bruker spectrometer equipped with a cryoprobe. Sample concentrations ranged from 100 µM to 2.5 mM. Resonance assignments were based on site specific low-enrichment labelling (26) and through bond correlations at natural abundance (JRHMBC, HMQC and TOCSY) (27,28). They were independently verified using

NOESY experiments. All spectral analyses were performed using the SPARKY software (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco).

Structure calculation: Distance restraints between protons were obtained from {^{1}H–^{1}H} NOESY spectrum (350 ms) at 293 K using SPARKY and CCPN Analysis software. NOESY spectra in H_2O were acquired at 50, 100, 200 and 350 ms mixing times and 150, 250, 350 and 400 ms for D_2O spectra. Distance restraints were derived from the NOESY spectrum recorded at 350 ms mixing time. Interproton distances were calibrated using the average volume of H7–H6 isolated cross-peaks of T8, T10 and T15 corresponding to a distance of 3.0 Å. All restraints files were included in CNS1.2 within ARIA2.3 software for structure calculation. Minimum distance bounds have been set at 1.8 Å˚ and the maximum at 7.0 Å. For each distance a fluctuation of ±25% was allowed in order to prevent errors from peak integration, overlaps or spin diffusion (35). Hoogsteen hydrogen bonds and planarity restraints were set considering the three tetrads and the base-pairing. Dihedral restrains have been set according to *syn* and *anti* conformation geometry determined by H1'/H8 correlations in {^{1}H–^{1}H} NOESY spectrum. Eight iterations were performed from 100 to 750 calculated structures with mixed Cartesian and torsion angle dynamics during the simulated annealing runs. For distances and hydrogen bonds, 50 kcal.mol^{-1}.A^{-2} was applied during the initial stage of dynamics and was increased up to 100 kcal.mol^{-1}.A^{-2} for the remaining steps of the dynamics. Twenty structures were selected and analyzed in order to select the ten lowest energy structures before going through structure refinement using AMBER20 molecular dynamics (see below).

MD simulation: Structure refinement in explicit solvent was performed using AMBER12 molecular dynamics (MD). For each structure selected from ARIA, two K+ cations were inserted within the G-core. The system was then neutralized with potassium cations and solvated with water molecules using a truncated octahedral TIP3P box. The first step was a minimization using harmonic potential position restraints at 25 kcal.mol^{-1}.A^{-2} over 2000 steps of steepest descent minimization. The next step consisted in heating the system from 100 to 293 K for 25 ps with position restraints at 50 kcal.mol^{-1}.A^{-2} in order to perform dynamics with the structure fixed. This was followed by several equilibrations with a progressive reduction of positional constraints: 22, 20, 17, 15, 12 and finally 10 kcal.mol^{-1}.A^{-2}. The last step was the equilibration of the system without positional restraints for 2.5 ns at 293 K. The ten lowest-energy structures were extracted and went through a series of minimizations to eliminate clashes and to correct bonds lengths and torsions.

RESULTS

HIVpro2 forms a stable three tetrad intramolecular G4

In K^+ solution, the proton NMR spectrum of the HIVpro2 sequence presents twelve well-resolved imino peaks located in the 10-12 ppm region (Figure 1C). As each G-quartet involves four imino protons, the presence of twelve peaks indicates the formation of a single G-quadruplex structure composed of three G-quartets. The presence of a lower field peak at 13.7 ppm suggests the presence of an additional base-pair. The thermal absorption difference spectrum (TDS) of HIVpro2 displays a negative minimum at 295 nm and two positive maxima at 280 and 240 nm (Figure 1D), typical of a G-quadruplex structure (29). The melting temperature determined by UV-melting is 59°C (Figure 1E). The reversibility of the melting profile strongly suggests that the structure is monomeric. The CD-spectrum of the HIVpro2 sequence (Figure 1F) presents a negative peak at 237 nm and two positive one at 260

and 295 nm. This signature is typical for 3+1 G-quadruplexes. The melting temperature determined by CD-melting is 63°C at 260 nm and 62°C at 295 nm (Figure 1G), slightly higher than the one measured by UV-melting.

NMR chemical shifts assignments

The unambiguous assignments of the guanine imino (H1) and aromatic (H8) protons were performed by site specific 5%-enrichment ^{15}N and ^{13}C guanine labelling. The ^{15}N filtered experiments (Figure 2A) show that the imino protons from G2, G3, G4, G9, G11, G12, G16, G17, G18, G20, G21 and G22 are involved in the G-tetrads hydrogen bonds network, while those from G6 and G7 are not. All aromatic protons were assigned based on ^{13}C-HSQC experiments on site specific enriched guanines (Figure 2B). ^{13}C-^{1}H-HMBC (Figure 2-C) confirmed the consistency of the assignments of imino and aromatic protons for all guanines. The lower field peak at 13.7 ppm was assigned to the H3 of T15 based on ^{15}N filtered experiments on site specific 5%-enrichment ^{15}N thymine labelling (Figure 2A). The presence of this peak indicates the formation of a base pair involving T15 but not T8. In order to assign the C residues, we proceeded to C to T substitutions at positions 13, 14 or 19. Those mutations were used to assign the cytosine resonances using the disappearance of the specific H5/H6 TOCSY correlation of the mutated C (Figure S1). The intense intra-residue NOE cross-peaks between the H1' of the sugar and the H8 of the aromatic base of G2, G9, G11, G16, G20 (Figure 3C) suggest that these guanines adopt *syn* glycosidic conformations. The presence of 5'-*syn*-*anti*-3' steps between G2-G3, G11-G12, G16-G17 and G20-G21 are evidenced by specific rectangular NOE patterns(30) composed of a double sequential H1'/H8 correlations observed for the 5'-G2-G3-3' and 5'-G12-G13-3' steps (Figure 3C). Based on the described assignments, the classical H8/H6-H1′ NOE sequential connectivity could be traced from A1 through G22 (Figure 3C).

HIVpro2 forms a 3+1 G-quadruplex structure with an additional AT reversed-Hoogsteen base pair

The three G-tetrads G2-G20-G16-G12, G3-G11-G17-G21 and G4-G9-G18-G22 are defined by characteristic H1-H8 proton cyclic NOE connectivity patterns (Figure 3B and S2). The stacking of these G-tetrads results in the formation of a 3+1 G-quadruplex with a bulge in T10 (Figure 3-D). The first two loops are lateral, the A5-G6-G7-C8 loop spans a wide groove and the C13-C14-T15 one spans a narrow groove. The third C19 loop is propeller and spans a medium groove. This topology is confirmed by inter-tetrad H1-H1 correlations (Figure 3A) between the guanines of the two bottom G-tetrads (G3-G4, G9-G11, G17-G18 and G21-G22) and between G11 and G12. HIVpro2 high resolution structure has been computed using 165 NOE restrains obtained in H_2O and 535 derived from D_2O spectra. MD simulation at 2.5 ns has been performed and the 10 lowest energy structures were extracted (Table 1). The structure of the HIVpro2 G-quadruplex is shown in Figures 4A and 4B. On the top of the 3+1 G-quadruplex core, the A1/T15 interaction differs from a classical Watson-Crick base-pair and corresponds to a reversed-Hoogsteen one (Figures 4C, 4D and S3). This base pair involves hydrogen bonds between H3 and O2 of the thymine instead of H3 and O4 in the case of classical Hoogsteen or Watson-Crick base pairs. In order to provide the Hoogsteen face with N7 and H6, A1 adopts a *syn* conformation. The two other residues of the first lateral loop, C14 and C13, point toward the solvent and the groove. Likewise, the residues of the first lateral loop, A5-G6-G7-T8 at the bottom of the structure do not stack on the tetrad and point towards the solvent. ^{15}N filtered experiment did not show any addition hydrogen bond involving G6, G7 and T8 (Figure 2). Indeed, all these residues are pointing outside the structure, toward the solvent (Figure 4C and 4F).

In order to confirm the relative lack of organization of the first lateral loop and the absence of stacking on the tetrad, we assessed the solvent accessibility of the imino protons of the G-quartet core and the T15 residue by performing H_2O to D_2O exchange experiments (Figure S3). After five minutes of D_2O exchange, we observed an almost immediate loss of 6 of the 13 peaks. The rapid disappearance of the imino peaks G20, T15, G4 G9, G18, and G22 implies that these residues are accessible to the solvent, whereas the relative persistence of G2-G16-G12 and G3-G11-G17-G21 suggests that they are protected from exchanging with D_2O molecules. As expected, G3-G11-G17-G21, which imino proton intensities remain unchanged during the first four hours were indeed the most protected because they belong to the central tetrad and are protected by the whole G-quadruplex core. In the same way, G2-G16-G12 are a little less protected from solvent exchange but not G20. This is explained by the structure of HIVpro2, whose A1/T15 base-pair better protects residues G2-G16-G12 (Figure S4) while residue G20 is more accessible to solvent via the medium groove. Finally, the rapid disappearance of G4, G9, G18 and G22 show that they are extensively accessible to solvent molecules. This confirms that A5-G6-G7-T8 lateral loop is poorly structured and does not stack on the G4-G9-G18-G22 tetrad.

Influence of point mutations on HIVpro2 stability

In order to better understand the contribution of the loop residues to the overall stability of the structure, we probed the effects of point mutations of HIVpro2 sequence by using CD spectroscopy, ^{1}H-1D-NMR and UV-melting (Table 2, Figure 5, S5 and S6). The effects of the mutations strongly depended on the positions. At position 1 or 15, A1T or T15A mutations disrupted the A1/T15 base-pair without preventing the formation of the G4 as suggested by the ^{1}H-1D-NMR spectra (Figure 5B) for which the T15 imino peak at 13.7 ppm disappears. In the case of A1T, the CD signal remained identical to the wild-type sequence (Figure 5A), and the melting temperature decreased by 7°C due to a possible disruption of this base-pair. Nevertheless, for T15A, the change in the relative intensity of the CD peaks indicates an important change in the overall conformation. The resulting structure is certainly different from HIVpro2 as also suggested by an increase of the melting temperature by 4°C. A5T, T8A and T10C mutations did not disrupt the Hoogsteen base-pair, thus the global HIVpro2 structure is maintained as monitored by ^{1}H -1D-NMR. In the particular case of A5T, we observed an increase in the UV-melting temperature (+4°C) as well as significant changes in the CD signal. CD signal is usually very sensitive to base stacking, thus, the A to T mutation induced the formation of additional base stacking as compared to HIVpro2 for which no base stacking was observed at this position. In contrary to A5T, the T8A mutation only slightly affected the CD signal but induced a 3°C decrease of the T_m. Thus, the presence of A residues at positions 5 or 8 of the bottom lateral loop hinders base stacking and destabilizes the structure as compared to T residues. Similar results were obtained with G to T mutations of the bottom lateral loop. G6T and G7T also preserved the global structure of HIVpro2 as monitored by the presence of the A1/T15 base-pair on the NMR spectrum. Yet, these substitutions also induced an increase in T_m (+2°C and +6°C) and a significant change in the CD profiles. This may also result from the presence of additional base stacking interactions on the bottom tetrad that may be favored by G to T substitutions. The decrease in steric hindrance when replacing a purine by a smaller pyrimidine residue may favor base stacking. T10C mutation at the bulge position or C13T and C14T mutations at the second lateral loop did not affect the overall structure as monitored by NMR. Interestingly, the T10C mutation at the bulge did not affect the CD signal and no change in base stacking pattern was detected. However, it induced a clear decrease in the melting temperature (-4°C). At this bulge position of HIVpro2 structure the T10 residue is pointing toward the solvent. The destabilization induced by the T10 to C10

replacement may be caused by unfavorable entropic contributions related to a reorganization of the solvent around the structure. For C13T and C14T, there was almost no effect on the melting temperature, yet the CD signal was altered. These C-to-T mutations, which are located in the vicinity of the A1T15 base pair, certainly induced a slight repositioning of the A1T15 base pair leading to a change in the stacking mode on the 3' tetrad as detected by CD.

DISCUSSION

Base-pair stacking and loop organization in hybrid type G4 structures

HIVpro2 adopts a 3+1 hybrid type G4 core with the following single descriptor nomenclature (SDN): 3(+lw+ln+p) *i.e* it is composed of 3 tetrads, the first loop is lateral and spans a wide groove (lw), the second is also lateral but spans a narrow groove (ln) and the third is a propeller type and spans a medium groove (p) (Figure 3D). Of note, the human telomeric sequence TTGGG(TTAGGG)$_3$A (pdb code 2GKU (31)) also adopts a 3+1 hybrid structure, but it is clearly different from the one adopted by HIVpro2. Indeed, it has an opposite SDN (3(−P−Ln−Lw)) ; the G4 core starts with a propeller loop at the first position and finishes with the two lateral loops. HIVpro2 has the exact same G4 core as the one adopted by the human telomeric sequence (TTAGGG)$_4$TT (pdb code 2JPZ (32) and the telomeric sequence of tetrahymena (TTGGGG)$_4$ (pdb code 186D (33). These three sequences have different loop sequences and lengths but it is still useful to compare these three structures. HIVpro2 presents an A1/T15 reverse Hoogsteen base pair that is not present in the two other structures and, to the best of our knowledge, has never been observed in another G4 before. Indeed, at the equivalent positions of A1 and T15, we find A3 and A15 in the case of 2JPZ and T2 and G15 for 2GKU, and no additional base pair was observed in these structures. We also showed in this study that the AGGT lateral loop composed of four nucleotides is poorly structured and does not stack on the bottom tetrad. We also observed that replacing a purine (A or G) by a smaller pyrimidine residue (T) tends to favor base stacking and increases stability. As a comparison, the equivalent of the first AGGT lateral loop of HIVpro2 is the TTA loop in 2JPZ or the GTTG loop in 2GKU. In the case of the shorter TTA loop of 2JPZ, which also has more purines than pyrimidines, a clear stacking of a T and A residues is observed. In the case of the GTTG loop of 2GKU, only the first G of the loop is stacking while the others are not. Therefore, loop organization and base stacking are clearly linked to the length and nucleotide composition of the loop.

The sequences and overall conformations of the loops are therefore very important in order to favor the formation of canonical or non-canonical base pairs that will stack on the G4 tetrads, as observed in the literature. For example, a GGA triad is present in the Giardia telomeric G4 (pdb code 2KOW (34)), a CT base pair is present on the G4 structure of the *C. elegans* telomeric sequence (pdb: 7OQT (8)) while a canonical AT base pair stabilizes the CEB25 G4 structure (pdb: 2LPW (35)). These pairings contribute significantly to the stability of the structure. In the case of HIVpro2, CEB25 or ASC20, the absence of these base pairs induces a decrease in T_m of 4°C to 10°C. Moreover, these pairings are stacked on the tetrads and can interfere with the recognition of the tetrad by protein partners or ligands.

Influence of DNA G4 structures in transcription initiation

Transcription is a highly regulated biological process in all cells. A small imbalance can have serious consequences in humans such as cancers or genetic disorders. Transcription initiation is mainly controlled by the presence of DNA sequences in the promoter ahead of the transcription start site (TSS). These standard sequences such as the

TATA box, CCAAT box or GC box allow the recruitment of an ensemble composed of transcription factors and a TATA binding protein (TBP) to ensure proper functioning of the RNA Polymerase. However, only 60% of eukaryotic gene promoters contain one of these standard sequences, and, in the particular case of the TATA box, it is present in only 15% of promoters (36). Thus, other properties like DNA architecture must intervene in the transcription initiation process. For instance, adenine-rich sequences can increase DNA curvature and influence the accessibility of the promoters (37). Furthermore, non-canonical DNA structures like hairpins and cruciforms are recognized by several regulatory proteins, such as p53 and p73 transcription factors (38,39). Likewise, G4's robustness and folding features also strongly suggest that they are able to play a role during DNA replication or transcription. For instance, assuming its complementary strand is transiently sequestered by a protein during transcription, the G-rich strand is isolated and free to form a G4 structure. The half-life of G4s is relatively long, they can withstand the annealing in the presence of high excess of their complementary strand (up to 50 times) (40). Kinetics studies have also shown that the formation of intramolecular DNA or RNA G4s can be relatively fast, in the order of 60 ms, which is very close to the folding rate of a RNA hairpin (41). The progression rate of DNA replication (20 ms/nt) or transcription (200 ms/nt) would in theory give enough time to the G-rich strand to adopt a G4 structure (42). Indeed, a role of G4s in transcription regulation has already been suggested for several genes (43) and oncogenes (43-45). For instance, G4 formation has been proposed to act as transcription silencer that reduces the expression of the c-myc and c-kit oncogenes (46-48).

Sp1 transcription factor binds to G4s

Sp1 is a transcription factor which regulates gene expression (49). It binds DNA via its zinc finger DNA binding domain. This domain is highly conserved (50) and is known to recognize a specific Guanine-rich duplex motif (5 '(G/T)GGGCGG(G/A)(G/A)(C/T) -3 ') trough major groove recognition of GC base-pairs. These motifs are often localized just upstream of the TSS of promoters that contain either single or multiple Sp1 binding sites (51). More than 12,000 Sp1 binding sites are present in the human genome and they are involved in both activation or repression of many genes. Indeed, in the presence of Sp1, chromatin will either adopt a more permissive structure that will enhance transcription or may condense and inhibit gene transcription (52,53). Like in HIV-1 promoter, Sp1 binding sites often overlap with putative G4 forming sequences suggesting that the formation of G4s also participate to the regulation mechanism of these genes. For instance, the c-Kit promoter presents one Sp1 binding site that can be embedded in three different G-quadruplex structures (15). In this context, it has been shown that Sp1 protein can bind to both double-stranded DNA and G4 DNA structure (15,54). Recently, Piekna-Przybylska et al studied the interaction between Sp1 protein and a fragment of the HIV-1 promoter sequence that fold into a G4. Using pull down experiments, they showed that Sp1 protein can bind the HIV-1 promoter sequence when it adopts a G4 conformation (48).

HIV-1 promoter activity and viral latency

G4 structures formed at the promoter region of HIV-1 provirus can intervene in viral transcription. At least four different G4 structures, including HIVpro2 described here, have been described to form within the three successive SP1 binding sites (Figure 6). Notably, these four topologies are mutually exclusive: the formation of one of them in the promoter will prevent the formation of the three alternative conformations. Furthermore, the formation of the G4s will also exclude the formation of the canonical duplex structure. Depending on the type of structure that is

formed at the promoter, the three SP1 binding sites will be embedded into a G4 that can be hybrid, parallel or antiparallel or into a duplex. These structural modulations of the SP1 binding sites will certainly have an impact on both the recruitment of the SP1 protein and the activity of the promoter. If SP1 activates transcription through the recognition of GC base pairs, then formation of G4s such as HIVpro2 and LTR-IV will play a repressive role by masking the SP1 binding sites. Conversely, the formation of G4s such as HIVpro1 and LTR-III, which exhibit both GC base pairs and guanine tetrads, may be able to recruit SP1 protein and serve as transcription activators. Indeed, Perrone *et al.* analyzed the effect of the G4 structures formed in the HIV promoter (55) using LTR promoters cloned upstream of a firefly luciferase gene. They proposed that LTR-IV formation acts as repressor elements in the transcriptional regulation of HIV-1 while the formation of the other structures might in contrary enhance viral transcription (55). However, other proteins like Nucleolin (56) and FUS (57) could also participate to promoter regulation. These two proteins are known to stabilize G4 structures and silence HIV-1 viral transcription.

Infection with HIV-1 generates viral reservoirs where the virus is latent and silent into the host genome, mostly in long-lived memory T cells. The reactivation mechanism of these silent viruses is still unknown. The G4s/duplexes structural interplay that takes place at viral promoter possibly interferes with virus reactivation from latency. The current therapeutic arsenal does not act on HIV-1 reservoir which represents a major obstacle to completely eradicate the virus. Therefore, an interesting strategy would be to target the G4s located in the promoter with G4 stabilizing ligands (58). Even though this strategy looks attractive, with the current knowledge about G4 functions in promoters, it is difficult to predict whether a given G4 ligand would reactivate HIV-1 reservoirs or in contrary inhibit reactivation.

DATA AVAILABILITY

The NMR chemical shifts have been deposited in the Biological Magnetic Resonance Bank (accession code, 34565) and coordinates have been deposited in the Protein Data Bank (accession code 7ALU).

ACKNOWLEDGEMENTS

We thank Aurore Guédin, Axelle Grelard and Estelle Morvan for helpful advices. This work benefited from the facilities and expertise of the BPCS platform in UMS3033/US001, http://www.iecb.u-bordeaux.fr/index.php/fr/plateformestechnologiques. This work was conducted with the support of the *ANRS Maladies Infectieuses Emergentes – Agence autonome de l'Inserm*, the *Centre National de la Recherche Scientifique* (CNRS), the *Institut National de la Sante´ et de la Recherche Médicale* (Inserm), and the *Université de Bordeaux.*

FUNDING

Agence Nationale de Recherches sur le Sida et les Hepatites Virales (ANRS) [ECTZ35927, ECTZ103899, ANRS AAP1-2015]. Funding for open access charge: INSERM. J.M and A.D.R. benefited from post-doctoral fellowships from the ANRS (ECTZ35927 and ANRS-AAP1-2015). A.D.R. thanks the Belgian Fonds national de la Recherche Scientifique (FNRS) for her Postdoctoral Researcher grant.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Jana, J., Mohr, S., Vianney, Y.M. and Weisz, K. (2021) Structural motifs and intramolecular interactions in non-canonical G-quadruplexes. *RSC Chemical Biology*, **2**, 338-353.
2. Guedin, A., Gros, J., Alberti, P. and Mergny, J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic acids research*, **38**, 7858-7868.
3. Zhang, A.Y. and Balasubramanian, S. (2012) The kinetics and folding pathways of intramolecular G-quadruplex nucleic acids. *Journal of the American Chemical Society*, **134**, 19297-19308.
4. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. and Balasubramanian, S. (2020) The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology*, **21**.
5. Biffi, G., Di Antonio, M., Tannahill, D. and Balasubramanian, S. (2014) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.*, **6**, 75-80.
6. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182-186.
7. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature biotechnology*, **33**, 877-881.
8. Marquevielle, J., De Rache, A., Vialet, B., Morvan, E., Mergny, J.L. and Amrane, S. (2022) G-quadruplex structure of the C. elegans telomeric repeat: a two tetrads basket type conformation stabilized by a non-canonical C-T base-pair. *Nucleic acids research*.
9. Lyu, K., Chow, E.Y., Mou, X., Chan, T.F. and Kwok, C.K. (2021) RNA G-quadruplexes (rG4s): genomics and biological functions. *Nucleic acids research*, **49**, 5426-5450.
10. Robinson, J., Raguseo, F., Nuccio, S.P., Liano, D. and Di Antonio, M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription? *Nucleic acids research*, **49**, 8419-8431.
11. Kumar, P., Yadav, V., Baral, A., Kumar, P., Saha, D. and Chowdhury, S. (2011) Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide. *Nucleic Acids Res.*, **39**, 8005 - 8016.
12. Sengupta, A., Roy, S.S. and Chowdhury, S. (2021) Non-duplex G-Quadruplex DNA Structure: A Developing Story from Predicted Sequences to DNA Structure-Dependent Epigenetics and Beyond. *Acc. Chem. Res.*, **54**, 46-56.
13. Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267-1272.
14. Shen, J., Varshney, D., Simeone, A., Zhang, X., Adhikari, S., Tannahill, D. and Balasubramanian, S. (2021) Promoter G-quadruplex folding precedes transcription and is controlled by chromatin. *Genome Biology*, **22**, 143.
15. Raiber, E.-A., Kranaster, R., Lam, E., Nikan, M. and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res.*, **40**, 1499-1508.
16. Sengupta, P., Bhattacharya, A., Sa, G., Das, T. and Chatterjee, S. (2019) Truncated G-Quadruplex Isomers Cross-Talk with the Transcription Factors To Maintain Homeostatic Equilibria in c-MYC Transcription. *Biochemistry*, **58**, 1975-1991.
17. Niu, K., Xiang, L., Jin, Y., Peng, Y., Wu, F., Tang, W., Zhang, X., Deng, H., Xiang, H., Li, S. *et al.* (2019) Identification of LARK as a novel and conserved G-quadruplex binding protein in invertebrates and vertebrates. *Nucleic Acids Res.*, **47**, 7306-7320.
18. Spiegel, J., Cuesta, S.M., Adhikari, S., Hänsel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biology*, **22**, 117.
19. Jaubert, C., Bedrat, A., Bartolucci, L., Di Primo, C., Ventura, M., Mergny, J.L., Amrane, S. and Andreola, M.L. (2018) RNA synthesis is modulated by G-quadruplex formation in Hepatitis C virus negative RNA strand. *Scientific reports*, **8**, 8120.
20. Kabbara, A., Vialet, B., Marquevielle, J., Bonnafous, P., Mackereth, C.D. and Amrane, S. (2022) RNA G-quadruplex forming regions from SARS-2, SARS-1 and MERS coronoviruses. *Frontiers in chemistry*, **10**, 1014663.
21. Abiri, A., Lavigne, M., Rezaei, M., Nikzad, S., Zare, P., Mergny, J.L. and Rahimi, H.R. (2021) Unlocking G-Quadruplexes as Antiviral Targets. *Pharmacological reviews*, **73**, 897-923.
22. Amrane, S., Jaubert, C., Bedrat, A., Rundstadler, T., Recordon-Pinson, P., Aknin, C., Guedin, A., De Rache, A., Bartolucci, L., Diene, I. *et al.* (2022) Deciphering RNA G-quadruplex function during the early steps of HIV-1 infection. *Nucleic acids research*, **50**, 12328-12343.

23. Amrane, S., Kerkour, A., Bedrat, A., Vialet, B., Andreola, M.-L. and Mergny, J.-L. (2014) Topology of a DNA G-Quadruplex Structure Formed in the HIV-1 Promoter: A Potential Target for Anti-HIV Drug Development. *J. Am. Chem. Soc.*, **136**, 5249-5252.
24. De Nicola, B., Lech, C.J., Heddi, B., Regmi, S., Frasson, I., Perrone, R., Richter, S.N. and Phan, A.T. (2016) Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome. *Nucleic Acids Res.*, **44**, 6442-6451.
25. Butovskaya, E., Heddi, B., Bakalar, B., Richter, S.N. and Phan, A.T. (2018) Major G-Quadruplex Form of HIV-1 LTR Reveals a (3 + 1) Folding Topology Containing a Stem-Loop. *J. Am. Chem. Soc.*, **140**, 13654-13662.
26. Phan, A.T. and Patel, D.J. (2002) A Site-Specific Low-Enrichment 15N,13C Isotope-Labeling Approach to Unambiguous NMR Spectral Assignments in Nucleic Acids. *J. Am. Chem. Soc.*, **124**, 1160-1161.
27. Phan, A.T.n. (2000) Long-range imino proton-13C J-couplings and the through-bond correlation of imino and non-exchangeable protons in unlabeled DNA. *J. Biomol. NMR*, **16**, 175-178.
28. Phan, A.-T., Guéron, M. and Leroy, J.-L. (2002) In Thomas L. James, V. D. and Uli, S. (eds.), *Methods Enzymol.* Academic Press, Vol. Volume 338, pp. 341-371.
29. Mergny, J.-L., Li, J., Lacroix, L., Amrane, S. and Chaires, J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
30. Adrian, M., Heddi, B. and Phan, A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, **57**, 11-24.
31. Luu, K.N., Phan, A.T., Kuryavyi, V., Lacroix, L. and Patel, D.J. (2006) Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold. *Journal of the American Chemical Society*, **128**, 9963-9970.
32. Dai, J., Carver, M., Punchihewa, C., Jones, R.A. and Yang, D. (2007) Structure of the Hybrid-2 type intramolecular human telomeric G-quadruplex in K+ solution: insights into structure polymorphism of the human telomeric sequence. *Nucleic acids research*, **35**, 4927-4940.
33. Wang, Y. and Patel, D.J. (1994) Solution structure of the Tetrahymena telomeric repeat d(T2G4)4 G-tetraplex. *Structure*, **2**, 1141-1156.
34. Hu, L., Lim, K.W., Bouaziz, S. and Phan, A.T. (2009) Giardia telomeric sequence d(TAGGG)4 forms two intramolecular G-quadruplexes in K+ solution: effect of loop length and sequence on the folding topology. *Journal of the American Chemical Society*, **131**, 16824-16831.
35. Amrane, S., Adrian, M., Heddi, B., Serero, A., Nicolas, A., Mergny, J.L. and Phan, A.T. (2012) Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite. *Journal of the American Chemical Society*, **134**, 5807-5816.
36. Brazda, V., Bartas, M. and Bowater, R.P. (2021) Evolution of Diverse Strategies for Promoter Regulation. *Trends in genetics : TIG*, **37**, 730-744.
37. Bacolla, A., Zhu, X., Chen, H., Howells, K., Cooper, D.N. and Vasquez, K.M. (2015) Local DNA dynamics shape mutational patterns of mononucleotide repeats in human genomes. *Nucleic acids research*, **43**, 5065-5080.
38. Fleming, A.M., Zhu, J., Jara-Espejo, M. and Burrows, C.J. (2020) Cruciform DNA Sequences in Gene Promoters Can Impact Transcription upon Oxidative Modification of 2'-Deoxyguanosine. *Biochemistry*, **59**, 2616-2626.
39. Brazda, V., Cechova, J., Battistin, M., Coufal, J., Jagelska, E.B., Raimondi, I. and Inga, A. (2017) The structure formed by inverted repeats in p53 response elements determines the transactivation activity of p53 protein. *Biochemical and biophysical research communications*, **483**, 516-521.
40. Risitano, A. and Fox, K.R. (2003) Stability of Intramolecular DNA Quadruplexes: Comparison with DNA Duplexes. *Biochemistry*, **42**, 6507-6513.
41. Zhang, A.Y.Q. and Balasubramanian, S. (2012) The Kinetics and Folding Pathways of Intramolecular G-Quadruplex Nucleic Acids. *J. Am. Chem. Soc.*, **134**, 19297-19308.
42. Rigo, R., Dean, W.L., Gray, R.D., Chaires, J.B. and Sissi, C. (2017) Conformational profiling of a G-rich sequence within the c-KIT promoter. *Nucleic Acids Res.*, **45**, 13056-13067.
43. Basundra, R., Kumar, A., Amrane, S., Verma, A., Phan, A.T. and Chowdhury, S. (2010) A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *FEBS J.*, **277**, 4254-4264.
44. Rigo, R., Palumbo, M. and Sissi, C. (2017) G-quadruplexes in human promoters: A challenge for therapeutic applications. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1861**, 1399-1413.
45. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. and Balasubramanian, S. (2020) The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology*, **21**, 459-474.
46. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences*, **99**, 11593-11598.

47. Zorzan, E., Elgendy, R., Giantin, M., Dacasto, M. and Sissi, C. (2018) Whole-Transcriptome Profiling of Canine and Human in Vitro Models Exposed to a G-Quadruplex Binding Small Molecule. *Scientific reports*, **8**, 17107.
48. Piekna-Przybylska, D. and Maggirwar, S.B. (2018) CD4+ memory T cells infected with latent HIV-1 are susceptible to drugs targeting telomeres. *Cell Cycle*, **17**, 2187-2203.
49. Mir, R., Sharma, A., Pradhan, S.J. and Galande, S. (2018) Regulation of Transcription Factor SP1 by the β-Catenin Destruction Complex Modulates Wnt Response. *Mol. Cell. Biol.*, **38**, e00188-00118.
50. Oka, S., Shiraishi, Y., Yoshida, T., Ohkubo, T., Sugiura, Y. and Kobayashi, Y. (2004) NMR Structure of Transcription Factor Sp1 DNA Binding Domain. *Biochemistry*, **43**, 16027-16035.
51. Yu, B., Datta, P.K. and Bagchi, S. (2003) Stability of the Sp3–DNA complex is promoter-specific: Sp3 efficiently competes with Sp1 for binding to promoters containing multiple Sp-sites. *Nucleic Acids Res.*, **31**, 5368-5376.
52. Kang, J.E., Kim, M.H., Lee, J.A., Park, H., Min-Nyung, L., Auh, C.K. and Hur, M.W. (2005) Histone Deacetylase-1 Represses Transcription by Interacting with Zinc-Fingers and Interfering with the DNA Binding Activity of Sp1. *Cell. Physiol. Biochem.*, **16**, 23-30.
53. Lee, J.-A., Suh, D.-C., Kang, J.-E., Kim, M.-H., Park, H., Lee, M.-N., Kim, J.-M., Jeon, B.-N., Roh, H.-E., Yu, M.-Y. *et al.* (2005) Transcriptional Activity of Sp1 Is Regulated by Molecular Interactions between the Zinc Finger DNA Binding Domain and the Inhibitory Domain with Corepressors, and This Interaction Is Modulated by MEK. *J. Biol. Chem.*, **280**, 28061-28071.
54. Da Ros, S., Nicoletto, G., Rigo, R., Ceschi, S., Zorzan, E., Dacasto, M., Giantin, M. and Sissi, C. (2021) G-Quadruplex Modulation of SP1 Functional Binding Sites at the KIT Proximal Promoter. *International Journal of Molecular Sciences*, **22**, 329.
55. Perrone, R., Nadai, M., Frasson, I., Poe, J.A., Butovskaya, E., Smithgall, T.E., Palumbo, M., Palù, G. and Richter, S.N. (2013) A Dynamic G-Quadruplex Region Regulates the HIV-1 Long Terminal Repeat Promoter. *J. Med. Chem.*, **56**, 6521-6530.
56. Tosoni, E., Frasson, I., Scalabrin, M., Perrone, R., Butovskaya, E., Nadai, M., Palu, G., Fabris, D. and Richter, S.N. (2015) Nucleolin stabilizes G-quadruplex structures folded by the LTR promoter and silences HIV-1 viral transcription. *Nucleic acids research*, **43**, 8884-8897.
57. Ruggiero, E., Frasson, I., Tosoni, E., Scalabrin, M., Perrone, R., Marusic, M., Plavec, J. and Richter, S.N. (2022) Fused in Liposarcoma Protein, a New Player in the Regulation of HIV-1 Transcription, Binds to Known and Newly Identified LTR G-Quadruplexes. *ACS infectious diseases*, **8**, 958-968.
58. Luo, J., Wei, W., Waldispühl, J. and Moitessier, N. (2019) Challenges and current status of computational methods for docking small molecules to nucleic acids. *European Journal of Medicinal Chemistry*, **168**, 414-425.
59. Amrane, S., Kerkour, A., Bedrat, A., Vialet, B., Andreola, M.L. and Mergny, J.L. (2014) Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. *Journal of the American Chemical Society*, **136**, 5249-5252.
60. Butovskaya, E., Heddi, B., Bakalar, B., Richter, S.N. and Phan, A.T. (2018) Major G-Quadruplex Form of HIV-1 LTR Reveals a (3 + 1) Folding Topology Containing a Stem-Loop. *Journal of the American Chemical Society*, **140**, 13654-13662.
61. De Nicola, B., Lech, C.J., Heddi, B., Regmi, S., Frasson, I., Perrone, R., Richter, S.N. and Phan, A.T. (2016) Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome. *Nucleic acids research*, **44**, 6442-6451.

NMR Restrains	H_2O	D_2O
distance restrains	165	535
intraresidue distance restrains	56	397
sequencial (i, i+1) distance restrains	45	120
long-range (i, ≥ i+2) distance restrains	41	2
short-range non sequencial distance restrains	23	16
dihedral restrains	12	
H bonds restrains	26	
Structural statistics		
structure calculation		
total calculated structures	750	
NOE violations		
number (>0.3 Å)	0.049	
RMSD of violations (Å)	0.16 ± 0.014	
molecular dynamics		
simulation time (ns)	2.5	
extracted structures (lowest energy)	10	
RMSD		
All heavy atoms (Å)	0.59	

Table 1. Statistics of the computed structures of HIVpro2 obtained after structure calculation and refinement by molecular dynamics

Name	Sequence (5' → 3')	T_m (°C)
HIVpro2	AGGGAGGTGTGGCCTGGGCGGG	59
A1T	**T**GGGAGGTGTGGCCTGGGCGGG	52
T15A	AGGGAGGTGTGGCC**A**GGGCGGG	63
A5T	AGGG**A**GGTGTGGCCTGGGCGGG	63
G6T	AGGGA**T**GTGTGGCCTGGGCGGG	61
G7T	AGGGAG**T**TGTGGCCTGGGCGGG	65
T8A	AGGGATG**T**GTGGCCTGGGCGGG	56
T10C	AGGGAGGTG**C**GGCCTGGGCGGG	55
C13T	AGGGAGGTGTGG**T**CTGGGCGGG	59
C14T	AGGGAGGTGTGGC**T**TGGGCGGG	61

Table 2. List of DNA sequences studied here. Base substitutions are underlined and bold. Melting temperatures (T_m) were determined by UV melting experiments recorded in a 20 mM phosphate buffer pH 6.9 supplemented with 70 mM KCl. The standard error of T_m determination is 1°C.

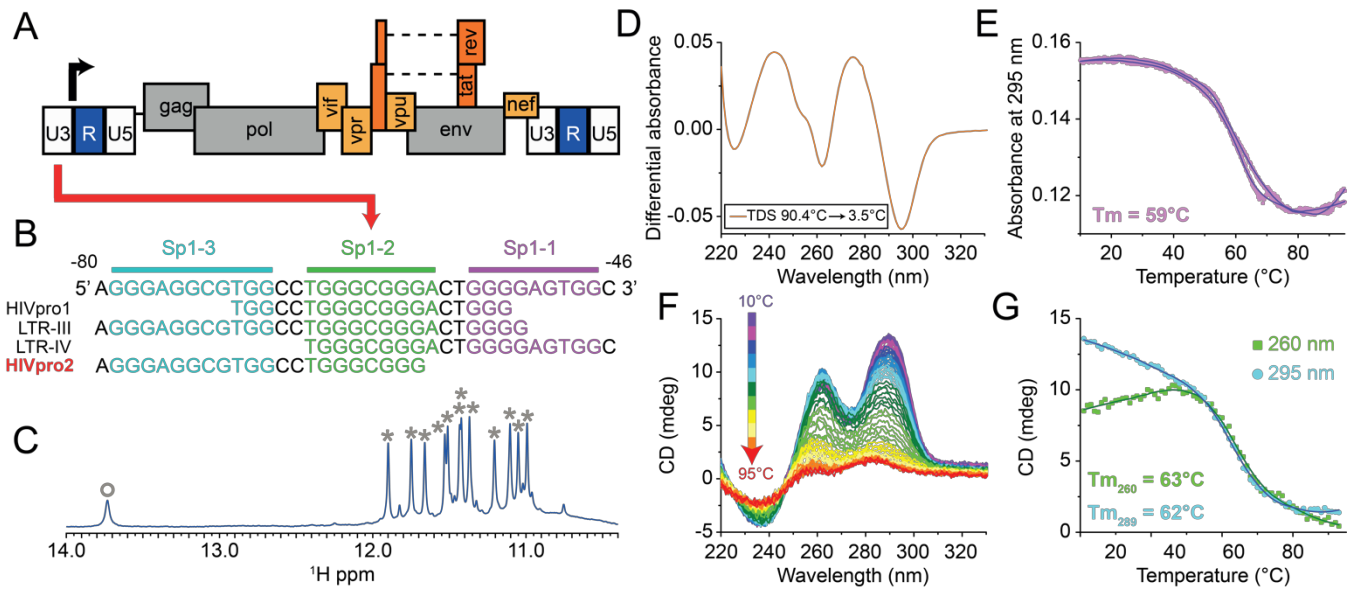


Figure 1. The HIVpro2 sequence and its biophysical properties. (A) and (B) Localization of HIVpro2 sequence in the viral genome. (C). ^{1}H NMR spectrum at 293 K in 90% H_2O /10 % D_2O 20 mM phosphate buffer pH 6.9 supplemented with 70 mM KCl. Guanine imino proton peaks have been identified with (✱) and Thymine imino proton peak with (○). (D), (F), (G) Spectroscopic analysis on a 3 µM HIVpro2 sample (20 mM phosphate buffer pH 6.9, 70 mM KCl) by (D) Thermal differential spectrum (TDS), (E) UV-melting (a signal artefact is observed around 70°C), (F) CD spectra from 20°C to 95°C and (G) CD melting curves extracted from (F).

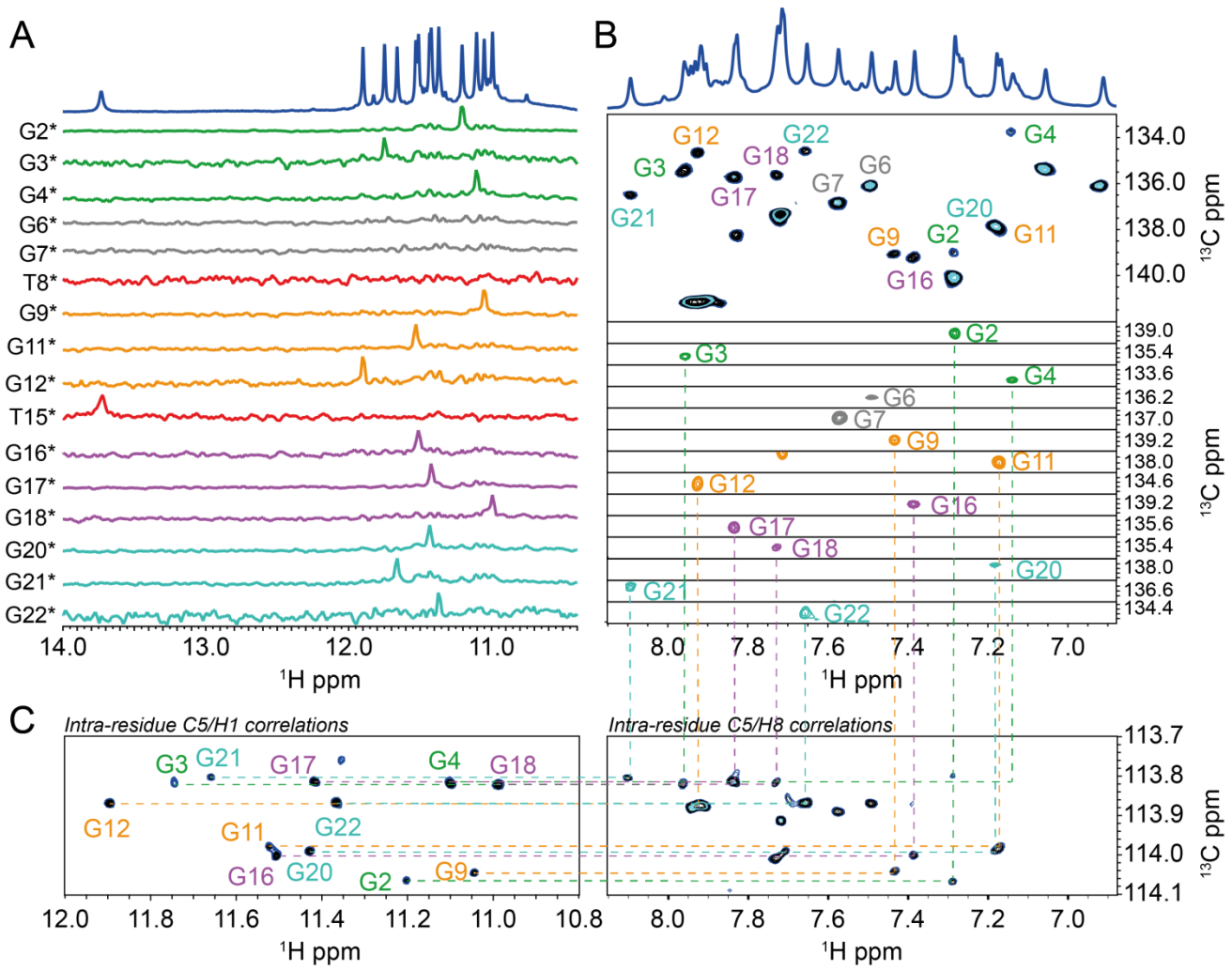


Figure 2. Unambiguous imino protons assignments of HIVpro2 by site specific labelling with 5%-enriched ^{15}N and ^{13}C guanines. **(A)** ^{15}N filtered ^{1}H-1D NMR spectra using 5%-^{15}N-enrichment at the indicated position. **(B)** {^{13}C–^{1}H}-HSQC experiment at natural abundance (Top) and with 5%-^{13}C-enrichment of guanines at the indicated position (bottom). **(C)** {^{13}C–^{1}H}-HMBC at natural abundance showing intra-residue H1/H8 correlations used to identify H8 of guanines implicated in the tetrads. Experiments were performed at 293 K in 90% H_2O /10 % D_2O 20 mM phosphate buffer pH 6.9, 70 mM KCl.

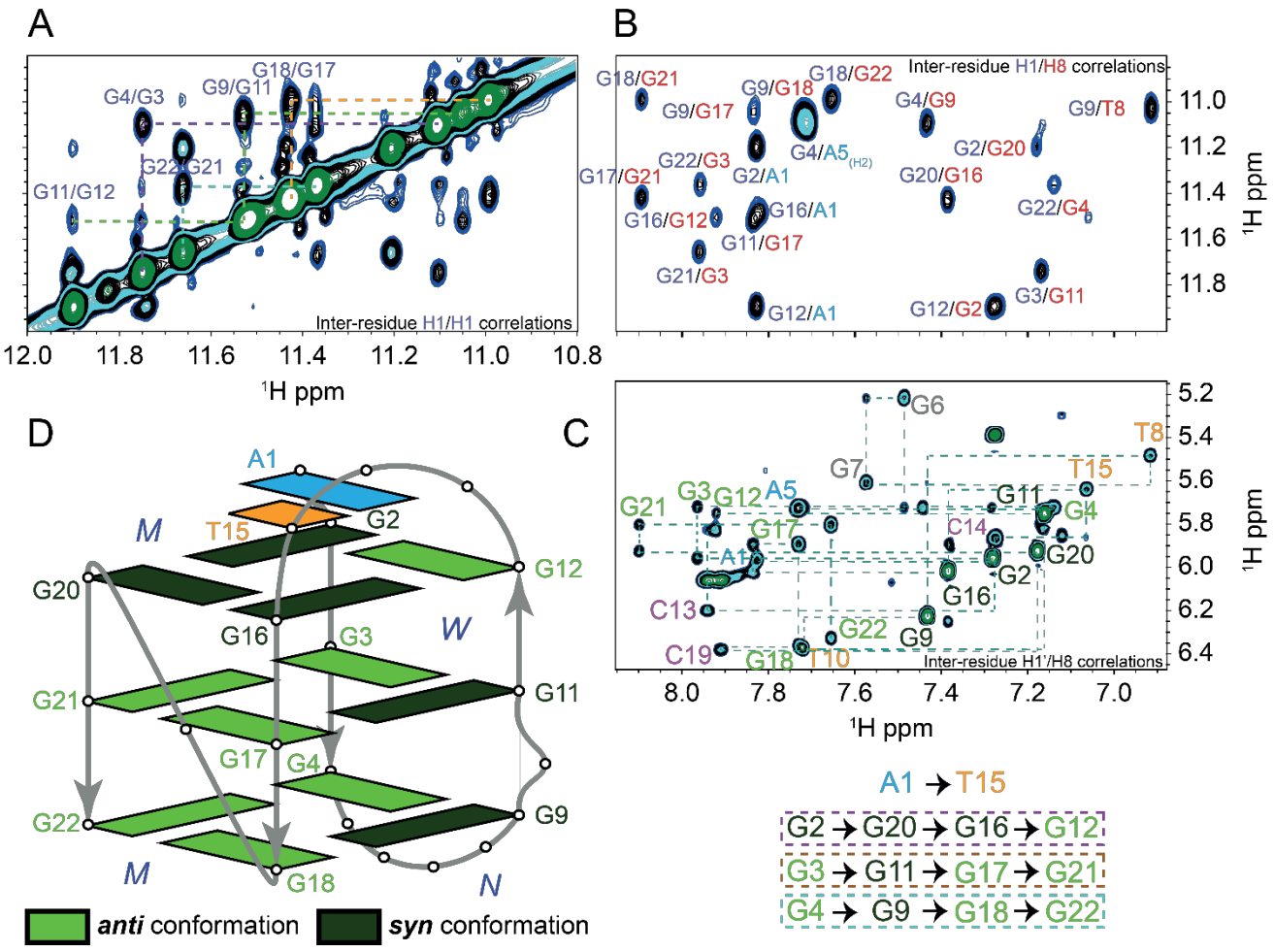


Figure 3. Determination of HIVpro2 topology. **(A)** and **(B)** Inter-residue H1/H1 and H1/H8 correlations used to determine tetrads patterns in a {^{1}H–^{1}H} NOESY experiment with a 350 ms mixing time. **(C)** Inter and intra-residue H1'/H8 correlations ("walk") between H8 from a residue and H1' from the previous residue in a {^{1}H–^{1}H} NOESY experiment with a 350 ms mixing time. **(D)** Topology of the HIVpro2 G-quadruplex with *syn* and *anti* guanines represented in dark and light green, respectively. Grooves have been characterized as W, N and M referring to wide, narrow and medium, respectively. Experiments were performed at 293 K in 90% H2O /10 % D2O 20 mM phosphate buffer pH 6.9, 70 mM KCl.

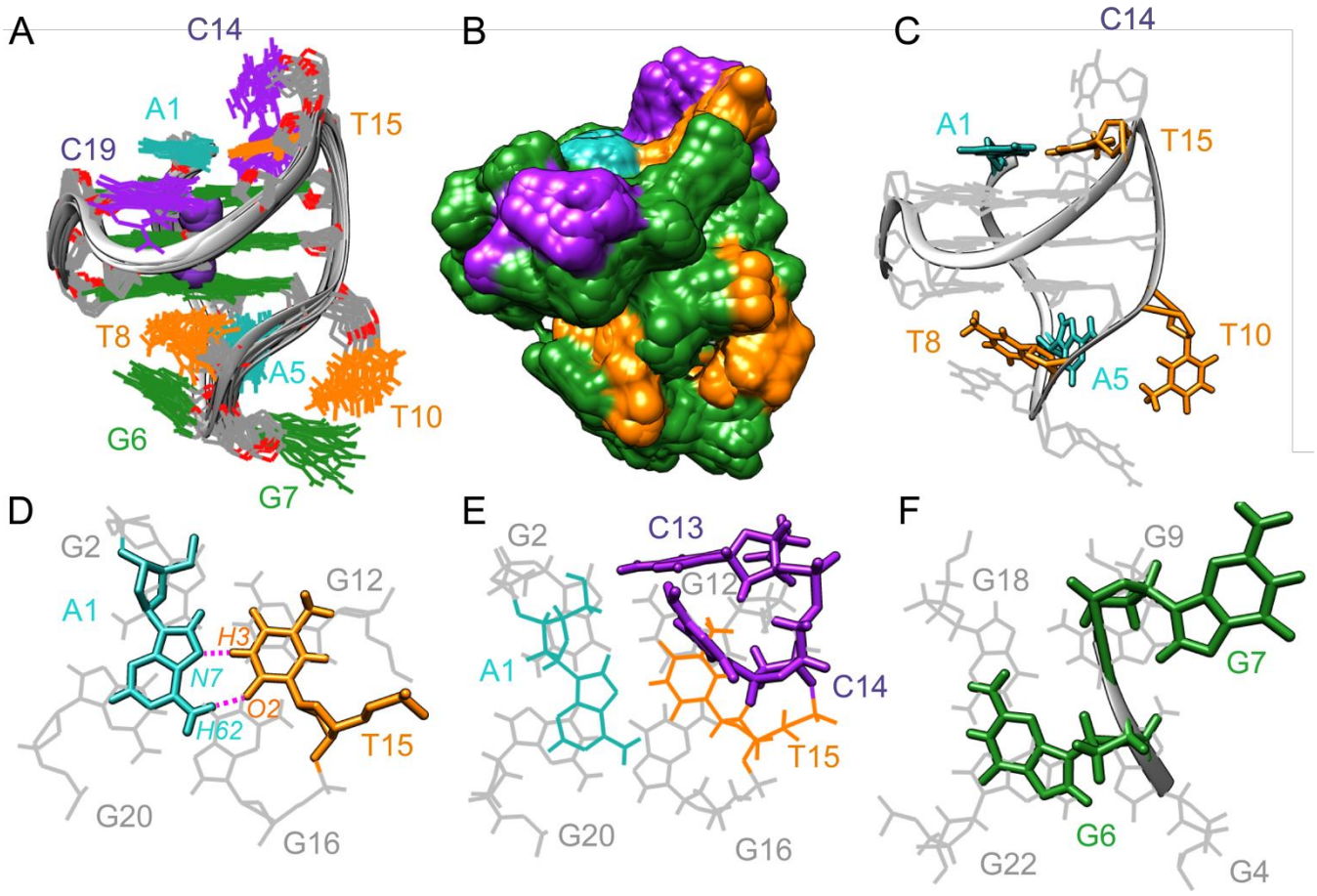


Figure 4. Structure of the HIVpro2 G-quadruplex (PDB 7ALU). (A) Ensemble of the ten best structures obtained after structure calculation and refinement. Guanines are coloured in green, Adenines in cyan, cytosines in purple and thymines in orange. (B) Surface view of the ensemble. Specific features of HIVpro2 structure are highlighted from (C) to (F). (C) The unstructured A5, T8 and T10 residues are highlighted in the first lateral loop. (D) Top view of the reverse Hoogsteen A1-T15 base pair that stacks on the upper tetrad. (E) C13 and C14 (both in purple) plunge in the wide groove behind the reverse Hoogsteen A1-T15 base pair. (F) G6 and G7 with their particular orientation towards the outside of the structure.

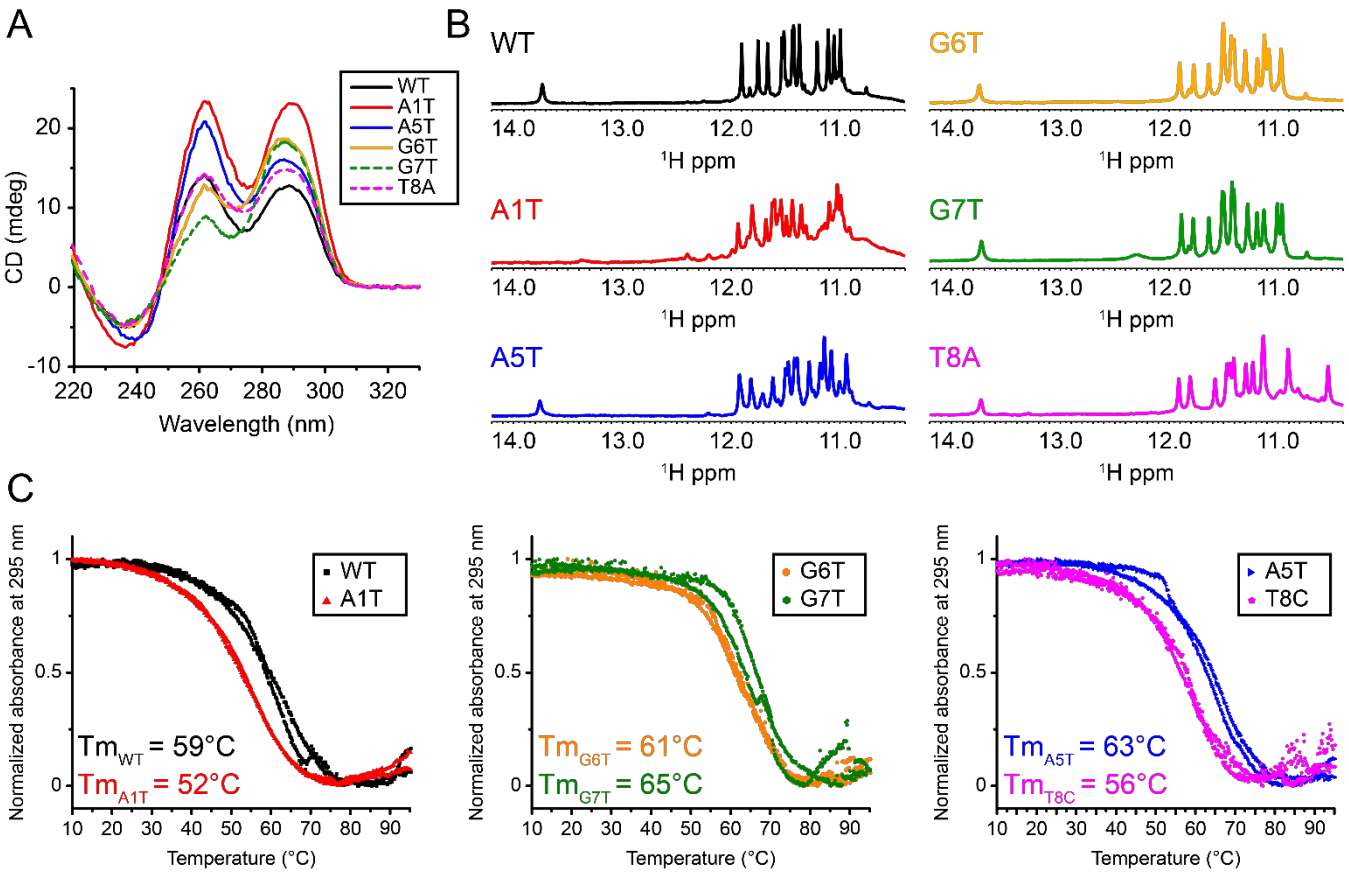


Figure 5. Biophysical studies of HIVpro2 mutants. **(A)** CD spectra at 20°C, **(B)** ^{1}H-1D-NMR imino proton region. All mutants are compared to the wild-type. **(C)** UV-melting experiments showing both heating and cooling curves at 295 nm (Signal artefacts are observed around 70°C for WT and G6T profiles). Experiments were performed in 20 mM phosphate buffer at pH 6.9 supplemented with 70 mM KCl.

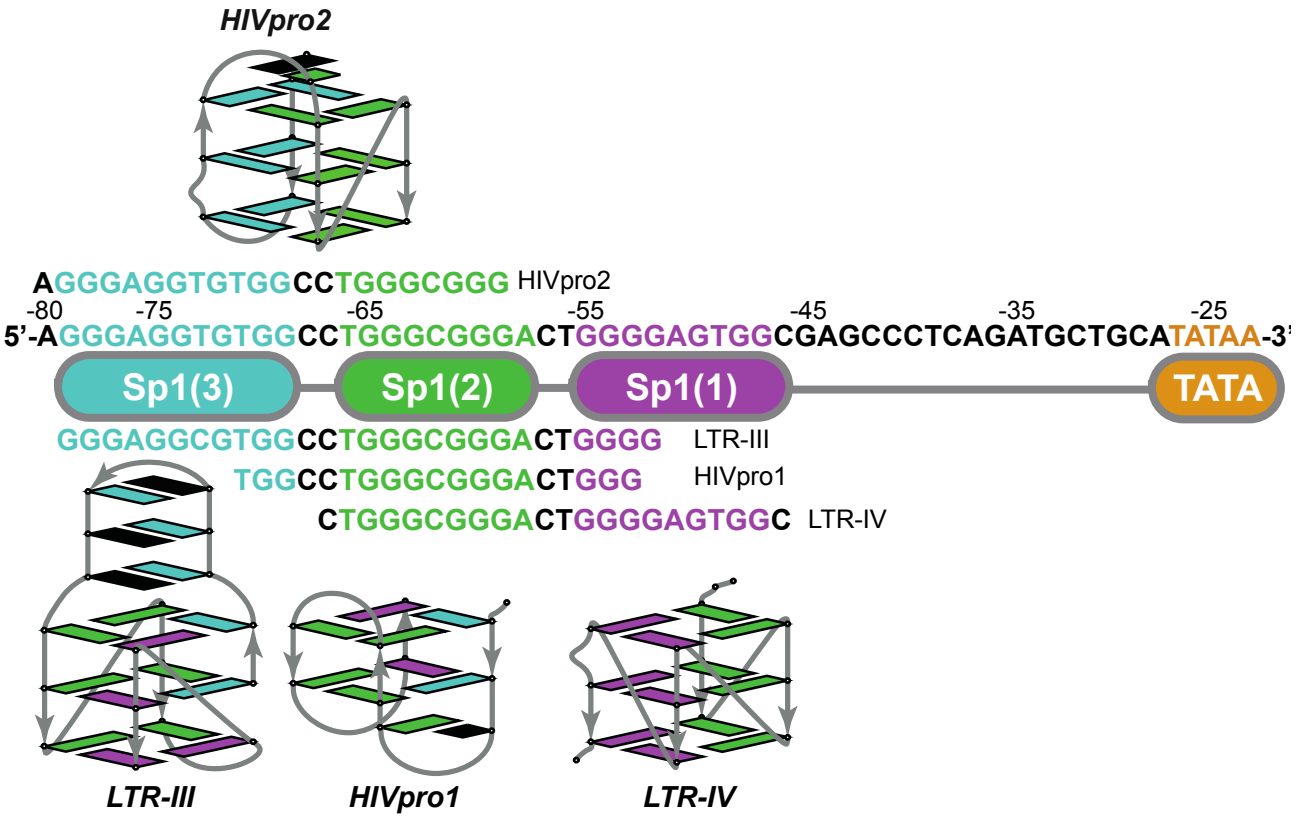


Figure 6: Mutually exclusive G4 structures formed within three adjacent SP1 binding sites in HIV-1 promoter. HIVpro2 (this work, pdb code: 7ALU), HIVpro1 (59), LTR-III (pdb code: 6H1K (60)) and LTR-IV (pdb code: 2N4Y (61)). The distance to the TSS is shown on top of the sequence.

Supplementary Information

Structure of a DNA G-quadruplex that modulates SP1 binding sites architecture in HIV-1 promoter

Aurore De Rache[1,2,§,+], Julien Marquevielle[1,2,+], Serge Bouaziz[5], Brune Vialet[1,2], Marie-Line Andréola[1,3], Jean-Louis Mergny[4] and Samir Amrane[1,2*]

[1] Université de Bordeaux, Bordeaux, France.

[2] ARNA laboratory, INSERM U1212, CNRS UMR 5320, IECB, Bordeaux, France.

[3] MFP laboratory, UMR5234, CNRS, Bordeaux, France.

[4] Laboratoire d'Optique & Biosciences, École Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, Palaiseau, France.

[5] UMR 8038 CNRS, Faculté de Pharmacie de Paris, Université Paris Descartes, Sorbonne Paris Cité

[§] Present address: Department of Chemistry, U. Namur, 61 rue de Bruxelles, B5000 Namur, Belgium

[+] These authors contributed equally to this work

* Corresponding author: Samir Amrane

Email: samir.amrane@inserm.fr

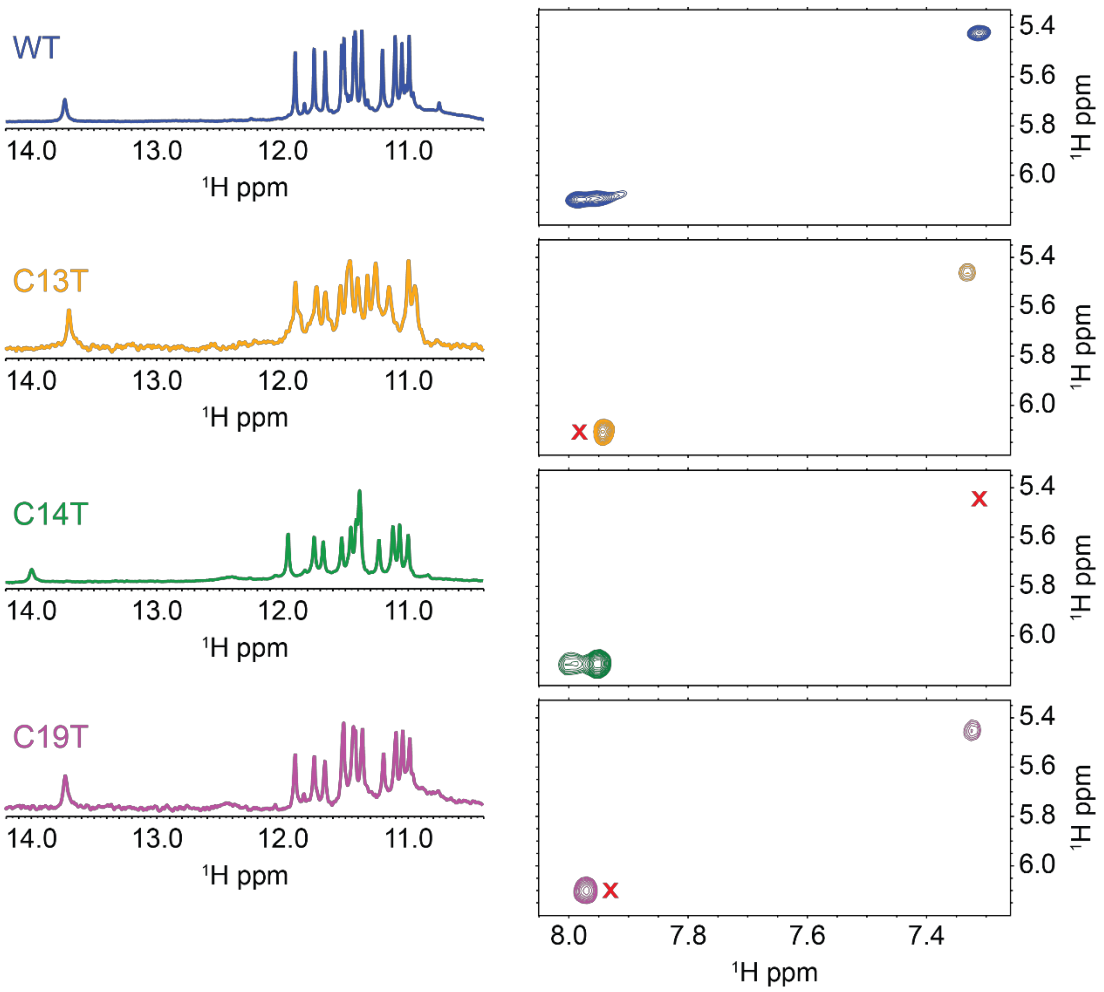


Figure S1. Assignment of cytosine aromatic protons based on C mutants using {1H-1H} TOCSY spectra of wild type and C13, C14 and C19 mutants (C into T) with a 100 ms mixing time. Spectra have been acquired at 20°C (293K) in 90% H2O /10 % D2O 20 mM phosphate buffer pH 6.9 supplemented with 70 mM KCl.

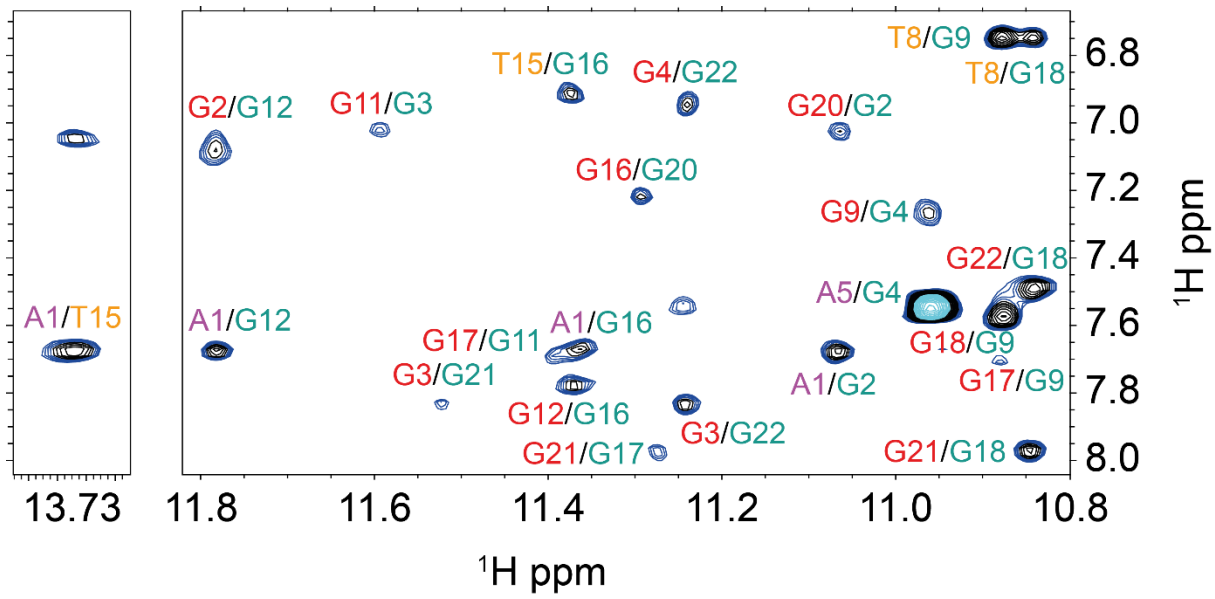


Figure S2. Inter-residue H1/H8 and H3/H8 correlations in {1H-1H} NOESY experiment with a 350 ms mixing time at 5.5°C in 90% H2O /10 % D2O 20 mM phosphate buffer pH 6.9 supplemented with 70 mM KCl.

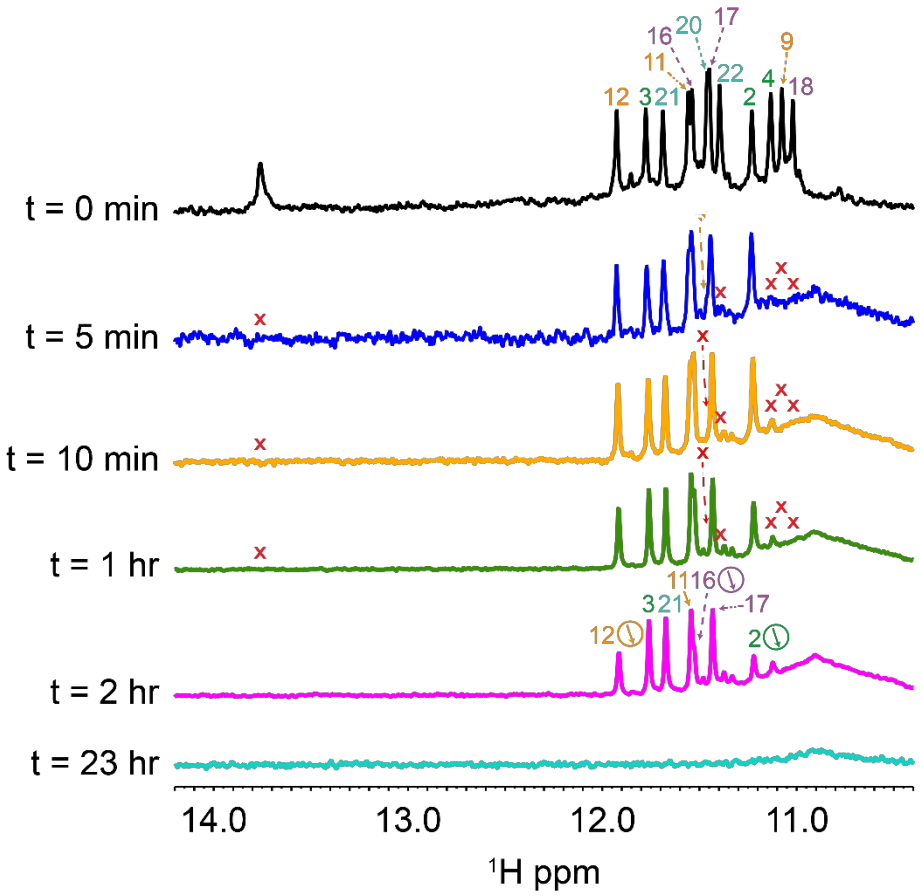


Figure S3. H_2O/D_2O exchange experiments. Spectrum at t = 0 min has been recorded at 20°C in 90% H2O /10 % D2O 20 mM phosphate buffer pH 6.9 supplemented with 70 mM KCl. After lyophilisation, sample has been dissolved in D_2O buffer and spectra have been recorded at different times.

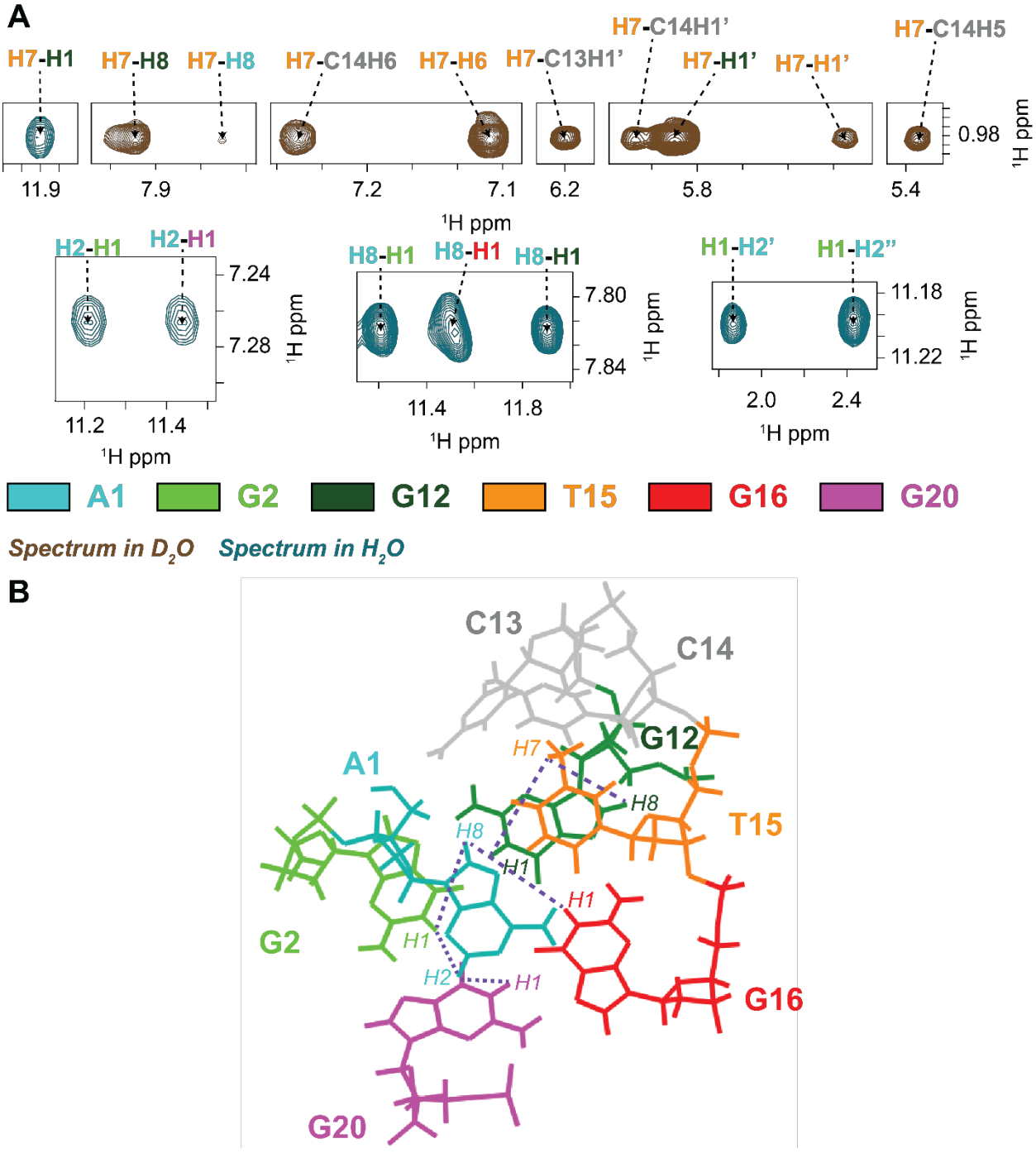


Figure S4. **(A)** Correlations between A1, T15 and the Gs from the top tetrad (G2-G20-G16-G12) identified in {^{1}H–^{1}H} NOESY spectrum recorded at a mixing time of 350 ms. Acquisition was done in 90% H2O/10 % D2O 20 mM potassium phosphate buffer pH 6.9 supplemented with 70 mM KCl at 293 K. **(B)** The corresponding correlations are shown in the structure and used in the computation.

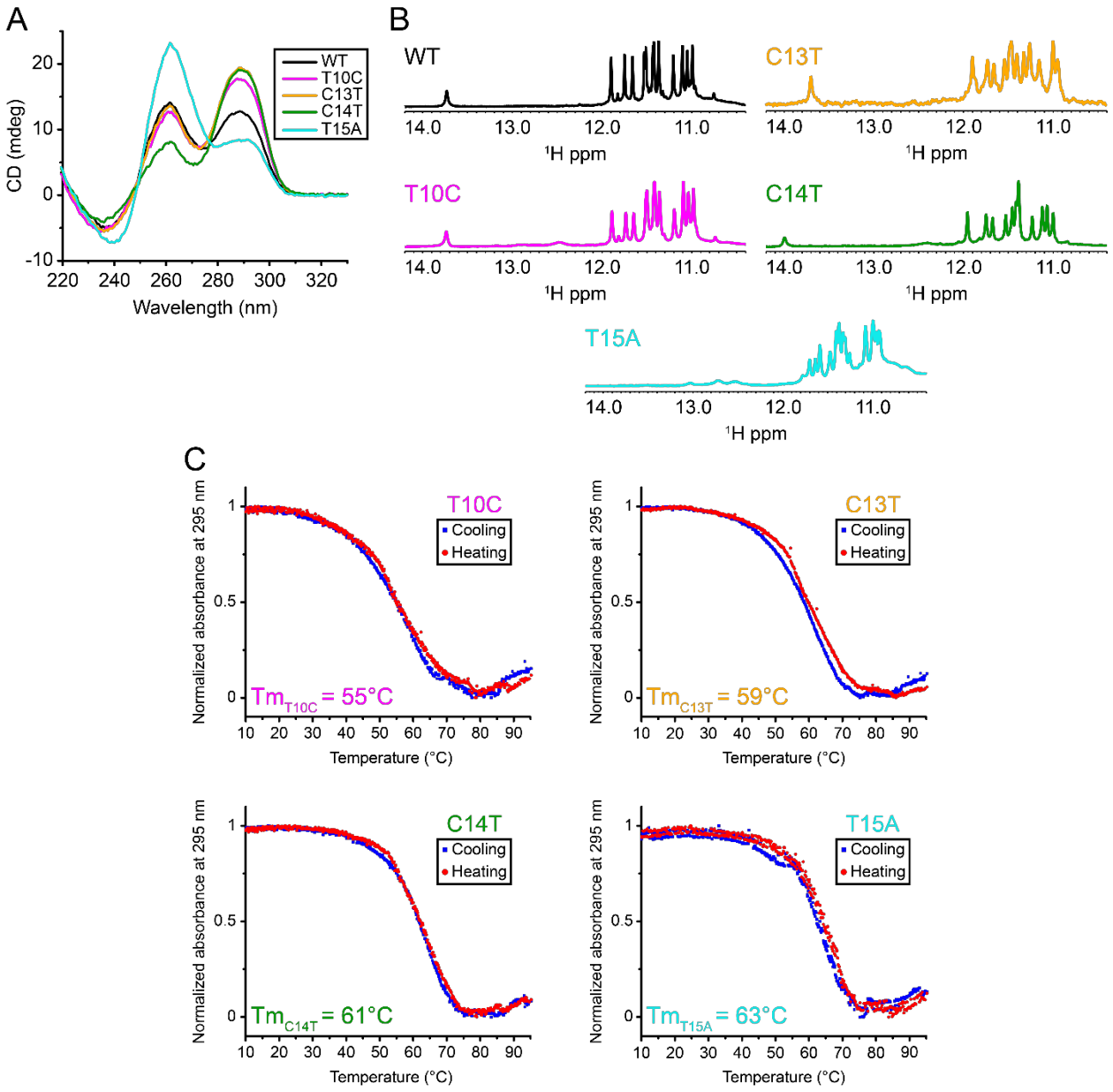


Figure S5. Biophysical studies of HIVpro2 mutants. **(A)** CD spectra at 20°C, in **(B)** ^{1}H NMR imino region. All mutants are compared to the wild-type. **(C)** UV-melting experiments with heating (red) and cooling (blue) curves at 295 nm. Experiments were performed in 20 mM phosphate buffer at pH 6.9 supplemented with 70 mM KCl.

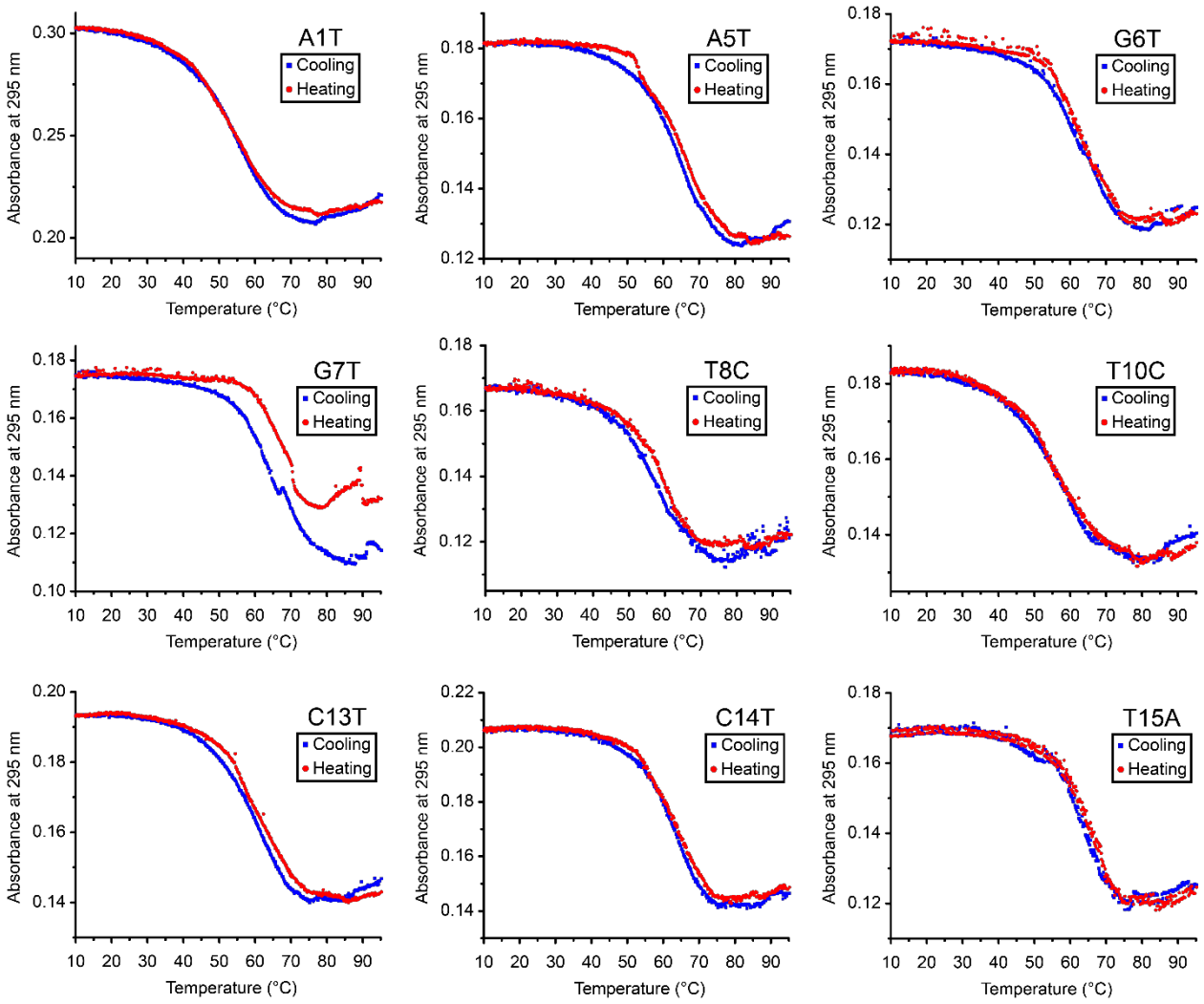


Figure S6. UV-melting experiments raw data with heating (red) and cooling (blue) curves at 295 nm. Experiments were performed in 20 mM phosphate buffer at pH 6.9 supplemented with 70 mM KCl.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: