Structural Explorations of NCp7–Nucleic Acid Complexes Give Keys to Decipher the Binding Process

Romain Retureau, Christophe Oguey, Olivier Mauffret, Brigitte Hartmann

Structural explorations of NCp7-nucleic acid complexes give keys to decipher the binding process

Romain Retureau[1], Christophe Oguey[2], Olivier Mauffret[1,*] and Brigitte Hartmann[1,*]

[1] LBPA, UMR 8113, ENS Paris-Saclay - CNRS, 61 avenue du Président Wilson, 94235 Cachan cedex, France

[2] LPTM, CNRS UMR 8089, Université de Cergy-Pontoise, 2 avenue Adolphe Chauvin, 95031 Cergy-Pontoise, France

* corresponding authors; Emails: bhartman@ens-paris-saclay.fr , olivier.mauffret@ens-paris-saclay.fr ; Tel: +33 1 47 40 77 33; Fax: +33 1 47 40 76 71.

ABSTRACT

A comprehensive view of all the structural aspects related to NCp7 is essential to understand how this protein, crucial in many steps of the HIV-1 cycle, binds and anneals nucleic acids (NA) mainly thanks to two zinc fingers, ZF1 and ZF2. Here, we inspected the structural properties of the available experimental models of NCp7 bound to either DNA or RNA molecules, or free of ligand. Our analyses included the characterization of the relative positioning of ZF1 and ZF2, accessibility measurements and the exhaustive, quantitative mapping of the contacts between amino acids and nucleotides by a recent tessellation method, VLDM. This approach unveiled the intimate connection between NA binding process and the conformations explored by the free protein. It also provided new insights into the functional specializations of ZF1 and ZF2. The larger accessibility of ZF2 in free NCp7 and the consistency of the ZF2/NA interface in different models and conditions give ZF2 the lead of the binding process. ZF1 contributes to stabilize the complexes through various organizations of the ZF1/NA interface. This work outcome is a global binding scheme of NCp7 to DNA and RNA, and an example of how protein-NA complexes are stabilized.

INTRODUCTION

The nucleocapsid proteins NCp15, NCp9 and NCp7 from the human immunodeficiency virus type 1 (HIV-1) are the products of the precursor Pr55Gag. They are small basic proteins containing a common NC domain that includes two zinc fingers, ZF1 and ZF2, separated by a short basic linker of 7 residues as earlier shown by a structural study [1]. Owing to their NC domain these proteins form complexes with nucleic acids (NA) and mediate many stages of the HIV-1 cycle, from viral genome recognition to RNA packaging [2–10]. They are able to bind nucleic acids in a non-specific mode as exemplified by the 1000-1500 NCp7 copies covering the RNA genome in the mature viral particle [11,12]. However, the highest NC affinities are measured for particular sites of the viral genome [4,13–15]. Such genomic studies as well as *in vitro* approaches [16–24] showed that NC proteins prefer to interact with accessible, unpaired guanines and, more broadly, with single-strand nucleic acids.

NCp7 is the fully matured form of NC proteins, present in both the host cell and the mature viral particle [2,25–27]. It shows an efficient chaperon activity, reorganizing DNA and RNA molecules and promoting their hybridization, two properties essential for the strand transfers occurring during the reverse transcription process ([28] and references herein). This protein of 55 residues is constituted by the NC domain flanked by unstructured N- and C-terminal regions (Fig. 1).

The earliest structural characterizations of free NCp7 (not bound to nucleic acids) by classical NMR concerned first a 39-residue peptide containing the two zinc fingers and the linker [29], followed by the entire protein [30,31]. However, the disorder in the N- and C-terminal domains and the flexibility of the linker complicated the interpretation of data [31]. This issue was recently overcome by combining small angle solution X-ray scattering experiments and exhaustive NMR investigations on labeled truncated and full length NCp7 [32]. This approach resulted in a series of structural models that, taken together, fit at best the whole experimental datasets. It confirmed that the conformational space sampled by NCp7 covers the so called “closed” and “open” forms which correspond to the presence or absence of ZF1/ZF2 contacts, respectively. These conformations likely originate from the intrinsic dynamics of the linker, also observed by NMR measurements of order parameter S^2 [33].

From a mechanistic point of view, the existence of contacts between ZF1 and the linker [32,33] was interpreted as a factor limiting the accessibility of ZF1 that could thus play a secondary role in the binding process [33]. Indeed, several studies of mutated or designed NCp7 established that the principal function of ZF1 relates to the chaperone activity and RNA packaging [34–38].

3D structures of NCp7 bound to NA were also derived from NMR studies, unveiling the main elements constituting the protein/NA interface. These complexes represent five systems; they contain NCp7 from two different HIV strains associated to RNA [39,40] or DNA [41–43] molecules that are either folded in stem-loop [39–41] or single-strand [42,43]. In particular, these studies elucidated the reason why NCp7 has a preference for guanines and underlined the importance of ZF1 and ZF2 aromatic amino acids in the protein-NA interface. Indeed, a guanine optimizes the contacts with the zinc finger interaction pockets by offering the possibility of 4 hydrogen bonds; in addition, compared to any pyrimidine, a guanine can maximize the stacking with aromatic residues, PHE16 in ZF1 and TRP37 in ZF2, which were consequently usually mentioned as strong elements of hydrophobic clamps. It was also proposed from a qualitative visual inspection of some complexes that the N-ter → C-ter NCp7 orientation and the 5' → 3' course of NAs were parallel in NCp7/DNA complexes while anti-parallel in NCp7-RNA complexes [41,43]. If such a specific polarity remains ascertainable across a larger pool of DNA or RNA sequences, it could reflect the mode of differentiation of DNA and RNA chains by NC proteins.

Understanding the complete molecular mechanism by which NC proteins bind to their RNA and DNA targets ideally requires to collect extensive information about the structural behavior of free NC proteins, free NA and NC-NA complexes. Here, we were interested in carrying out exhaustive analyses of the five available NCp7/RNA [31,40] and NCp7/DNA [41–43] systems to compare them and try to extract information about the binding process. Each of these systems consist of a set of structures – between 10 and 25 – representing in the best possible way NMR data introduced as restraints in refinement protocols. Although the structures in each set are not necessarily representative of a full statistical distribution, the samples are expected to reflect the plasticity of the 3D organization of the considered complexes. In addition, as mentioned above, these different NMR

systems present a diversity of partner composition and conformation. This offers the opportunity to achieve a comprehensive overview of characteristics either common to the organization of very different complexes or, conversely, limited to specific cases.

The first part of this work was devoted to structural explorations of NCp7-NA complexes, in particular regarding the variability of the relative spatial orientations of ZF1 and ZF2, inside each system and across the systems. The most recent models of free NCp7 [32] were also analyzed and compared to bound NCp7 structures. Then, NCp7/NA interfaces were explored using a recent, original tessellation method, called VLDM (Voronoi Laguerre Delauney for Macromolecules) [44,45]. Based on a representation of molecules by a collection of polyhedra filling space without overlaps or gaps, VLDM has the advantage of analyzing interfaces without resorting to any empirical or adjusted parameter. Thus, VLDM deciphers macromolecular interactions but does not evaluate forces or energy directly. Here, VLDM provided an exhaustive inventory of the interacting elements by precisely mapping the NCp7/NA contacts that were further characterized in terms of contact area and nature, specifying the balance between electrostatic and non-electrostatic (hydrophobic, van der Waals) components. Overall, this work provided an in-depth structural overview of the functional elements of NCp7 as well as a detailed description of NCp7/NA interfaces. Importantly, our results lead to a much finer understanding of the reading of NA by NCp7.

RESULTS

Survey of the studied systems

To identify the structural features shared by both NCp7-RNA and NCp7-DNA complexes or, conversely, those that are specific to one or the other family, a total of 84 available NMR-based structures of NCp7-nucleic acid (NA) complexes belonging to five different systems (Table 1) was analyzed. Before reporting the results of our analyses, we need to make a few comments regarding both the proteins and nucleic acids in these systems.

PDB code	NCp7	NA	N_{models}
1A1T	55 residues, full length; NL4-3 strain	RNA of 20 nt from SL3 stem-tetraloop	25
1F6U	55 residues, full length; NL4-3 strain	RNA of 19 nt from SL2 stem-tetraloop	20
2JZW	44 residues: truncation of the first 11 N-terminal residues; M-B strain	DNA of 14 nt from HIV-1 Primer Binding Site (PBS, stem-pentaloop)	19
2L4L	45 residues: truncation of the first 10 N-terminal residues; M-B strain	DNA of 4nt, a single-strand region of the stem-hexaloop mini cTAR	10
1BJ6	42 residues: truncation of the first 11 N-terminal and the last two C-terminal residues; M-B strain	Single-strand DNA of 5 nt	10

Table 1. *Summary of the studied NCp7-NA complexes.*
This Table gives the PDB codes of the NCp7-NA complexes studied here, specifying NCp7 length, eventual truncation and provenance, as well as NA provenance and conformation. The number of models (N_{models}) in each system is given in the last column, summing up to 84. "nt" stands for nucleotide.

Although NCp7 is non-specific in the sense that it binds to any NA accessible region, this protein prefers to interact with single-strand sequences containing at least one guanine [16–19]. A maximal affinity was obtained for NCp7 -not included within Gag- and a RNA sequence derived from SL3 that includes two guanines separated by one base [17,24]. Such double guanine containing motif is present in both NCp7-RNA complexes studied here, GAG in 1A1T [39] and GUG in 1F6U [40]. In the NCp7-DNA systems, the 2JZW [41] and 2L4L [43] complexes are formed with DNA including a TG dinucleotide that is very attractive for NCp7 [16,17]; 1BJ6 [42] contains only one guanine. Most of these different complexes may therefore represent examples of optimal interactions.

Also, the complexes differ in NA length and conformation (Table 1): the 1AIT, 1F6U and 2JZW systems involve NAs folded in stem-loops; 1BJ6 contains a short single-strand DNA fragment; the DNA in 2L4L is the 4 nt single-strand region of a large stem-loop. NCp7 binds to RNA loops in SL2 (1F6U) and SL3 (1AIT) without inducing any stem destabilization while the NCp7-DNA complex 2JZW was considered as a first stage of stem melting because of interactions with both loop

and stem which were suspected to induce a weakening of one C:G base pair [41].

Finally, NCp7 produced from two different HIV-1 strains (Table 1) differ by three amino acids: THR 12, ILE 24 and LYS 26 in 1A1T and 1F6U (NCp7-RNA complexes) are ASN 12, THR 24 and ARG 26 in 2JZW, 1BJ6 and 2L4L (NCp7-DNA complexes). In addition, the full length NCp7, *i.e.* the two zinc fingers ZF1 and ZF2 separated by a short linker and surrounded by C-and N-terminal domains (Fig. 1), is present only in the NCp7-RNA complexes; the three NCp7-DNA complexes contain a N-terminal domain truncated from several residues (Table 1). In the NCp7-RNA complexes, the intact N-terminal domain, folded in a 3_{10} helix, interacts with either the major groove (1A1T, [39]) or the phosphodiester backbone (1F6U, [40]) of the SL3 and SL2 RNA stems, respectively. The importance of such contributions for the complex stability and functions [9] cannot be ignored. However, in our comparative study, we focused on the interactions between NA and the NCp7 part that is common to the five systems, *i.e.* amino acids 12-53.

The diversity of NA and protein compositions and conformations in our pool of data is an advantage for our purpose which is to detect characteristics either common or specific of NCp7/NA organizations. Since the two zinc fingers ZF1 and ZF2 are key components for NA binding, their structures and relative positioning were first scrutinized. The NCp7/NA interfaces were then analyzed and quantified using the VLDM approach, and finally compared to each other.

Structural variability of the zinc fingers ZF1 and ZF2 in NCp7-NA complexes

A previous study of complexes in which NCp7 binds the loop of NA hairpins inspected the conformation of the individual zinc fingers and concluded that their folding was identical regardless of the different NA targets [41]. Here we extended the analysis to our 84 structure dataset by calculating the cross-Root Mean Square Deviations (RMSD) on the backbone heavy atoms of either ZF1 (CYS 15 → CYS 28) or ZF2 (CYS 36 → CYS 49). The very low cross-RMSD values, 0.7 ± 0.25 Å for ZF1 and 0.85 ± 0.6 Å for ZF2 on average, confirms that each type of zinc finger, ZF1 or ZF2, adopts the same 3D conformation across the models, regardless of the systems.

The next step was to compare the spatial positioning of ZF1 and ZF2 with respect to each other. For that we defined the so-called ZF1-ZF2 ensemble in which the residues CYS 15 → CYS 28 and CYS 36 → CYS 49 were considered as a whole. The linker was thus excluded from the analysis. Indeed, the extensive conformational versatility affecting the Φ, Ψ backbone angles all along this short domain submerges the information about the ZF1-ZF2 ensemble by for instance dominating RMSD values. RMSD calculations carried out on the ZF1-ZF2 ensemble (Supplemental Fig. S1-A) show a good conservation of the ZF1 and ZF2 relative positioning within or across the three NCp7-DNA systems. The low cross-RMSD values (average value of 1.4 ± 0.3 Å) associated to 2JZW, 1BJ6 and 2L4L testify of the structural homogeneity of the ZF1-ZF2 ensemble across the NCp7-DNA systems (Supplemental Fig. S1-A). More noticeable cross-RMSD values emerge from the comparison between the two NCp7-RNA systems (Supplemental Fig. S1-A, cross-RMSD values of 3.8 ± 0.2 Å on average for 1A1T *vs* 1F6U) or between NCp7-RNA and NCp7-DNA systems (Supplemental Fig. S1-A, cross-RMSD values of 3.4 ± 0.3 Å on average for 1A1T or 1F6U *vs* 2JZW, 1BJ6 or 2L4L). However, such conformational variability does not imply dramatic rearrangements, as illustrated by the superimposition of typical structures of ZF1-ZF2 ensembles from two RNA- and one NCp7-DNA systems (Supplemental Fig. S1-B).

This global structural conservation is further established by measuring the distance $D_{ZnZF1\text{-}ZnZF2}$ between the two zinc atoms in ZF1 and ZF2 and pseudo-dihedral angles reflecting the relative orientations of ZF1 and ZF2. The values of $D_{ZnZF1\text{-}ZnZF2}$ tend to be slightly shorter in NCp7-RNA than in NCp7-DNA complexes but remain in the same order of magnitude across the systems (Table 2); the standard deviations also indicate limited variations across the models of a given system (Table 2).

	NCp7-RNA		NCp7-DNA		
Distance (Å)	**1A1T**	**1F6U**	**1BJ6**	**2JZW**	**2L4L**
$D_{ZnZF1\text{-}ZnZF2}$	16.4 (0.1)	18.0 (0.5)	19.5 (0.6)	19.0 (0.3)	18.0 (0.2)
Pseudo-dihedral angle (°)	**1A1T**	**1F6U**	**1BJ6**	**2JZW**	**2L4L**
($C\alpha_{CYS15}$- Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{CYS49}$)	17 (4)	78 (11)	49 (7)	40 (5)	63 (3)
($C\alpha_{LYS20}$- Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{GLY43}$)	166 (4)	234 (10)	179 (6)	184 (6)	205 (3)
($C\alpha_{CYS28}$- Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{CYS49}$)	35 (7)	26 (9) or 130 (19)	42 (7)	37 (5)	56 (6)

Table 2*: Structural parameters related to the relative positioning of ZF1 and ZF2 in bound NCp7. The spatial positioning of ZF1 and ZF2 with respect to each other was monitored in each system by calculating $D_{ZnZF1\text{-}ZnZF2}$, the distance between the two zinc atoms in ZF1 and ZF2, and three pseudo-dihedral angles. These angles are all defined around the two zinc atoms in ZF1 and ZF2 but differ by the two Cα atoms that complete the dihedral tetrads: i) CYS 15 and 49, the first and last residues of ZF1 and ZF2, respectively ii) LYS 20 and GLY 43, in ZF1 and ZF2, respectively and iii) CYS 28 and 49, the last and first residues of ZF1 and ZF2, respectively. The data are averaged values calculated on the model set constituting each system. Standard deviations are given in brackets.*

Overall, the form characterized by the proximity of the two zinc fingers (closed form) is preserved, as firmly ascertained by the existence of measurable NMR distances between PHE 16 and ASN 17 on one hand and TRP 37 on the other hand [39,41–43]. The relative orientations of ZF1 and ZF2 was then scrutinized through three Cα-Zn_{ZF1}-Zn_{ZF2}-Cα pseudo-dihedral angles involving the two Zn atoms and two Cα atoms of various residues chosen in three different ways: i) CYS 15 at the beginning of ZF1 and CYS 49, the last residue of ZF2, ii) the ZF1 and ZF2 centers, LYS 20 and GLY 43 and iii) CYS 28 at the end of ZF1 and CYS 49 at the beginning of ZF2. As for $D_{ZnZF1\text{-}ZnZF2}$, the values and standard deviations of these three pseudo-angles are consistent within and across the systems (Table 2). One exception concerns the 1F6U system in which 6 on a total of 19 models show alterations in the ZF1 and ZF2 folding and relative positioning, at least according to the $C\alpha_{CYS28}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{CYS36}$ pseudo-angle (Table 2). However, a clear, dominant scheme characterizes the ZF1-ZF2 ensemble, which can be summarized in terms of couples of vectors. Thus, $Zn_{ZF1}\rightarrow C\alpha_{LYS20}$ and $Zn_{ZF2}\rightarrow C\alpha_{GLY43}$ point towards opposite directions conversely to $Zn_{ZF1}\rightarrow$ $C\alpha_{CYS15}$ and

Zn_{ZF2}→$C\alpha_{CYS49}$ that are directed towards similar directions as well as Zn_{ZF1}→$C\alpha_{CYS28}$ and Zn_{ZF2}→$C\alpha_{CYS49}$ (Fig. 2).

In summary, each zinc finger, ZF1 or ZF2, behaves as a quasi-rigid body in our dataset of NCp7-NA structures. Despite of a residual variability of the relative orientation of ZF1 and ZF2, the organization of the ZF1-ZF2 ensemble is globally remarkably conserved in view of the disparity of the studied systems.

Analysis of free ZF1-ZF2 ensemble and comparison with their bound counterparts

The next question we addressed concerned an eventual resemblance between the free and bound forms of NCp7, in particular regarding the ZF1 and ZF2 relative positioning. For that, we benefited from a recent study that mixed NMR experiments, solution X-ray scattering and simulated annealing to quantitatively depict the conformational space sampled by free NCp7 from the M-B strain [32]. Three clusters of structures were defined, which, taken together, satisfied at best the whole collection of experimental data. At the end of their paper, the authors mentioned that their cluster 1 resembled to the bound NCp7 structure in 1F6U and 2JZW. Here, we focused on the ZF1-ZF2 ensemble and used our own criteria to analyze the 21 models (PDB code 5I1R) that stand for typical free NCp7 structures and to compare them to the 84 models of bound NCp7.

Let us look at the 5I1R models categorized using the $D_{ZnZF1\text{-}ZnZF2}$ distance and $C\alpha$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha$ angles. The existence of four distinct and homogeneous groups (Table 3 and Fig. 3-A) reflects without surprise the variability of the relative positioning of ZF1 and ZF2, which contrasts with the monotony observed in NCp7-NA complexes. The conformational combinations of $C\alpha$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha$ angles that are observed correspond to either short or larger distances between the two Zn atoms, defining closed conformations in groups 2 and 3 and open conformations in groups 1 and 4. The examination of the Φ, Ψ angles along the linker stresses the flexibility of two linker residues, LYS 34 and GLY 35, as previously described [32]. Indeed, various Φ, Ψ combinations of LYS 34 and GLY

35 are associated to each type of ZF1 and ZF2 relative positioning, except from $\Phi_{LYS\ 34}$ in g- in all open models.

Group	Models in 5I1R	$D_{ZnZF1\text{-}ZnZF2}$	($C\alpha_{15}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{49}$)	($C\alpha_{20}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{43}$)	($C\alpha_{28}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{36}$)	Free vs bound NCp7: $RMSD_{av}$
G1	1,4,7	23.0 (0.3)	-56 (8)	100 (7)	-158 (12)	6.4 (0.1)
G2	2,5,8,11,14,17,21	16.6 (0.3)	76 (8)	-91 (5)	68 (5)	4.7 (0.4)
G3	3,6,9,12,15,18,20	15.6 (0.5)	14 (6)	-173 (6)	11 (7)	3.1 (0.2)
G4	10,13,16,19	24.5 (0.7)	77 (47)	-158 (43)	-55 (40)	5.7 (0.3)

Table 3*: Structural parameters related to the relative positioning of ZF1 and ZF2 in free NCp7.*
This table reports the four groups of free NCp7 models resulting from the classification of the ZF1 and ZF2 spatial positioning relative to each other. The 21 models and their numbering are from the 5I1R PDB file. The relative positioning of ZF1 and ZF2 was monitored by the same parameters as in Table 2. The distance $D_{ZF1\text{-}ZF2}$ (Å) and three pseudo-dihedral angles (°) are given in term of averaged values and standard deviations in brackets. The last column shows the average RMSD values (Å) calculated on the heavy backbone atoms of the ZF1-ZF2 ensemble (CYS 15 → CYS 28 and CYS 36 → CYS 49), in free and bound NCp7 (complete sets of 21 and 84 free and bound structures, respectively).

The structures of group 3, characterized by moderate $D_{ZF1\text{-}ZF2}$ values and ($C\alpha_{15}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{49}$) in *g+*, ($C\alpha_{20}$-Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{43}$) in *trans* and ($C\alpha_{28}$- Zn_{ZF1}-Zn_{ZF2}-$C\alpha_{36}$) in *g+* (Table 3) are in quasi perfect agreement with the bound ZF1-ZF2 ensemble (Table 2). Accordingly, the ZF1-ZF2 ensembles of this group remarkably well match their bound equivalents (Supplemental Fig. S2; example in Fig. 3-B), clearly more than those of the other groups (Table 3).

To gain more information about a possible scenario of the NCp7 binding process, VLDM was used to calculate the contact area between each residue in the free NCp7 models and water molecules (for details of model hydration, see Materials and Methods). Irrespective of their location and the conformational group, arginines and lysines largely expose their charged, hydrophilic side chains to the solvent (Fig. 4-A). Less expected because of a poor hydration potential of their large side chains

[46–48], PHE 16 and TRP 37, both essential for NCp7/NA interaction [39–43], also show a consequent accessibility to water molecules. The TRP 37 is a special case because it has a variable accessibility: maximal in the open conformations of groups 1 and 4 (Figs. 4-A and 4-B), this accessibility is reduced in the closed conformations of group 3 and, to a lesser extent, of group 2, due to additional contacts between TRP 37 and linker or ZF1 residues – mainly ASN 17 - (Fig. 4-B). More globally, the residues of ZF2 are more accessible than those of ZF1 in the four conformational groups that we identified (Fig. 4-A).

These analyses support the idea that both open and closed conformations sampled by NCp7 in its free state are exploited during the NA binding process. Indeed, the accessibility measurements reveal that ZF2 and more specifically TRP 37 in free NCp7 open conformations are in the best position to interact with NA targets. However, the fixation of RNA or DNA to NCp7 clearly stabilizes a closed conformation type also explored by the free ZF1-ZF2 ensemble. The next step was to explore the protein/NA interface to understand how the NCp7 closed conformation is maintained in the complexes.

NCp7/NA interface composition and characteristics

The analyses of the NCp7-NA structures with VLDM provided the list of the amino acids and nucleotides that are in contact, and the quantification of the interface in terms of contact areas (CA). Despite the fact that contact areas do not measure energies directly, comparison of surfaces of similar electrostatic/non-electrostatic balance do give indications of the relative strength of interactions. CAs were averaged over the set of models of each system, distinguishing the surfaces interactions involving the different parts from NCp7, ZF1 ($CA_{ZF1/NA}$), the linker ($CA_{linker/NA}$) and ZF2 ($CA_{ZF2/NA}$) (Fig. 5, Supplemental Table S1). The largest CAs are observed with amino acids belonging to one or the other zinc finger, with an additional secondary contribution of residues of N-terminal or linker regions (Fig. 5-A, Supplemental Table S1), which were of course also taken into account in the analyses. The term “extended ZFx/NA interfaces” will designate this extended series of contacts

involving ZFx amino acids and additional relevant residues that, strictly speaking, belong to the N-terminal domain or the linker.

A first result about interfaces relates to the balance between the different component of CAs, electrostatic, hydrophobic or contacts due to the simple proximity (see Materials and Methods). The contribution of these three components is consistent not only across the five considered systems but also across the ZF1/, linker/ and ZF2/NA interfaces (Table 4). Electrostatic CAs correspond to very modest percentages of the total $CA_{ZF1/NA}$, $CA_{ZF2/NA}$ and even $CA_{linker/NA}$ in spite of the linker enrichment in basic residues (Table 4). The electrostatic interactions include hydrogen bonds mostly engaging the guanines interacting with linker and ZF2 residues (Supplemental Fig. S3), as earlier noticed [39]. Supplementing the electrostatic component, hydrophobic contacts represent the main contribution to $CA_{ZF1/NA}$ and a substantial part of $CA_{ZF2/NA}$ (Table 4). This marked hydrophobic character as well as the slightly different balance between the CA types in the ZF1 and ZF2 interfaces resonate with previous physicochemical studies [4,16,49,50]. Hydrophobic contacts primarily implicate aromatic amino acids and aliphatic chains of arginine and lysine on NCp7 side, and base carbon atoms on the NA side. The bases emerge as the major interacting NA elements, accounting for 64 ± 4 % of the total CAs – regardless of the type of contacts (Supplemental Fig. S4). The remaining CAs include sugars (C1', C2', C3', C4', C5', O4' and, when relevant, O2'; 23 ± 5 % of the total CAs) and phosphate groups (P, O1P, O2P, O5' and O3' atoms; 13 ± 4 % of the total CAs) (Supplemental Fig. S4).

Interface	**CA component (%)**	**NCp7-RNA**		**NCp7-DNA**			**AV(SD)**
		1A1T	**1F6U**	**2JZW**	**1BJ6**	**2L4L**	
ZF1/NA	Electrostatic	15	15	18	14	9	14 (3)
	Hydrophobic	44	36	37	46	59	44 (9)
	Proximity	41	49	45	39	32	41 (6)
Linker/NA	Electrostatic	26	18	21	21	30	23 (5)
	Hydrophobic	2	15	15	11	2	9 (6)
	Proximity	72	67	64	68	68	68 (3)
ZF2/NA	Electrostatic	19	18	20	20	14	18 (2)

	Hydrophobic	26	28	29	24	25	26 (2)
	Proximity	55	55	51	56	61	55 (4)

Table 4*: Components of NCp7/NA contact areas.*

The CA contribution of electrostatic, hydrophobic and simple C-O or C-N proximity components is given in percentage of total CA for the five systems studied here. The percentage values were calculated over the whole model set constituting each system. The overall average values (AV) and standard deviations (SD) are given in the last columns.

The interfaces were then examined in detail, starting with the simplest case, namely the ZF2 interface keeping a unique global organization preserved in all models. From the NA point of view, the most important CA contribution arises from a single guanine (Fig. 5-B), extruded from the NA loops in 1A1T, 1F6U and 2JZW and intrinsically accessible in the single-strand DNAs in 2BJ6 and 2L4L. These particular guanines are enclosed in pockets of very similar structure and composition across the systems (Supplemental Fig. S5; Tables 5 and S1). Among their interactions with amino acids, those involving TRP 37 and GLN 45 (Fig. 6-A) prevail, accounting for half or even over half of the total CAs (Table 5). Indeed, the exceptionally large CAs between TRP 37 and these guanines (Table 5) have no equivalent elsewhere in the NCp7/NA interfaces (Supplemental Table S1); they reflect the large overlap of the TRP and guanine aromatic rings (Fig. 6-A). According to quantum mechanical calculations, such stacking is the spatial configuration that corresponds to the best interaction energy for a couple composed of these two elements [51].

system	base	CA(linker/NA)				CA(ZF2/NA)						Sum
		ARG 32	LYS 33	LYS 34	GLY 35	CYS 36	TRP 37	HIS 44	GLN 45	MET 46	LYS 47	
1AIT	G210	no	2 (1)	14 (7)	11 (1)	10 (0)	40 (2)	2 (1)	32 (3)	17 (2)	22 (6)	150 (3)
1F6U	G209	3 (5)	no	2 (3)	9 (2)	4 (1)	37 (3)	no	22 (8)	19 (3)	10 (7)	106 (5)
2JZW	G107	7 (6)	11 (2)	7 (1)	11 (1)	4 (1)	48 (4)	no	29 (3)	14 (2)	16 (5)	147 (3)
1BJ6	G3	18 (5)	12 (2)	7 (2)	10 (1)	8 (1)	56 (2)	no	25 (3)	22 (3)	12 (6)	170 (3)
2L4L	G126	13 (7)	13 (2)	no	6 (2)	7 (1)	45 (2)	3 (1)	30 (3)	21 (5)	1 (1)	139 (3)

***Table 5**. Contact areas associated to key guanines in the linker-ZF2/NA interfaces.*
This table reports, for each system, the detailed contacts of the guanine that engages major contacts with TRP 37 and other amino acids of the linker or ZF2. Each nucleotide/amino acid couple is characterized by the contact area (CA, Å^2) calculated and averaged on the model set of each system; the last column, "Sum", gives the total CA of each considered nucleotide. Standard deviations are given in brackets. "no" stands for not observed.

In addition to interactions with ZF2 amino acids, the residues 32 → 35 of the linker (one ARG, two LYS and one GLY) complete the interface by contacting the sandwiched guanine and sometimes its nearest neighbors (Tables 5 and S1). These contacts are generally more fluctuating than those involving ZF2 residues, apart from the noticeable case of a recurrent hydrogen bond engaging the backbone of GLY 35. Yet, their contribution cannot be underestimated since their substitution by ALA residues affects NCp7 binding properties [9].

Both ZF1/ZF2 and ZF1/NA interfaces show alternative organizations, in contrast with what happens in the uniform ZF2/NA interface. As discussed in the first section, the global conservation of the relative orientation of ZF1 and ZF2 does not totally preclude some variability in backbone courses, which can be amplified by the diversity of side chain conformations. Thus, although ZF1 and ZF2 are always interconnected *via* PHE 16 and TRP 37, these residues do not occupy the same relative position in NCp7-RNA and NCp7-DNA complexes (Fig. 6-B). Despite the visual impression of large structural differences, four of the PHE 16 /TRP 37 arrangements (in 1A1T, 2JZW, 2BJ6 and 2L4L) display similar CAs, 11.7 ± 1.5 Å^2 on average. The remaining interaction in 1F6U has a very weak CA of 1.3 ± 5 Å^2. Besides, it should be noted that this ZF1/ZF2 interface is strengthened in the NCp7-RNA complexes by an ASN 17/ TRP 37 contact, as observed in the conformational group 3 of free NCp7.

Two distinct spatial arrangements are also detected in the ZF1/NA interfaces. These interfaces extend over a variable number of nucleotides (Fig. 5-B). One spectacular case is the amino acid 26 (LYS or ARG in NCp7-RNA, NCp7-DNA complexes, respectively) which covers from two (2L4L) to

nine (2JZW) nucleotides (Supplemental Table S1) and, accordingly, has a variable $CA_{LYS\ or\ ARG\ 26}$ value (Fig. 5-A and Supplemental Table S1). A potential source of variability is the NA folding/unfolding stem-loops offering *a priori* a possibility to multiply the contact points conversely to short single-strand segments. This in fact occurs in two out of the three systems containing contacted stem loops: the NA folding in 1F6U and 2JZW allows proximity interactions that increase the total CAs of the ZF1/NA interfaces ($CA_{av\text{-}ZF1/NA}$ = 313 ± 18 Å^2 for 1F6U, 348 ± 35 Å^2 for 2JZW; 242 ± 10 Å^2 for 1A1T; 161 ± 23 Å^2 for 1BJ6 and 147 ± 20 Å^2 for 2L4L).

Independently of the above considerations, the ZF1/NA interfaces are built around three elements: two amino acids surrounding one nucleotide. One of these elements, PHE 16, is common to all systems whereas the other two, the amino acid 24 and the nucleotide, are either ILE and guanine in NCp7-RNA or THR and pyrimidine in NCp7-DNA complexes. As a first remark, it should be noticed that the CAs associated to guanines are significantly lower with ZF1 (Table 6) than with ZF2 (Table 5).

system	base	N-ter domain		ZF1					Sum
		VAL 13	LYS 14	CYS 15	PHE 16	ILE 24	ALA 25	LYS 26	
1AIT	G212	8 (1)	13 (2)	1 (1)	27 (1)	29 (3)	12 (1)	18 (3)	108 (2)
1F6U	G211	2 (2)	9 (2)	3 (1)	19 (2)	29 (3)	10 (2)	28 (7)	100 (3)
system	base	VAL 13	LYS 14	CYS 15	PHE 16	THR 24	ALA 25	ARG 26	Sum
2JZW	T106	5 (2)	no	no	27 (3)	13 (2)	4 (2)	18 (8)	67 (3)
1BJ6	C2	9 (6)	3 (4)	1 (2)	31 (6)	21 (2)	10 (4)	16 (9)	91 (5)
2L4L	T124	13 (6)	8 (4)	1 (2)	32 (7)	12 (5)	2 (4)	no	68 (4)

Table 6*. Contact areas associated to key nucleotides in extended ZF1/NA interfaces.*
This table concerns specifically each nucleotide of each system which engages major contacts with PHE 16 and other amino acids of the N-terminal domain or ZF1. Each individual nucleotide/amino acid couple is characterized by the contact area (CA_{av}, Å^2) calculated and averaged on the models of

each system; the last column, "Sum", gives the total CA_{av} of each considered nucleotide. Standard deviations are given in brackets. "no" stands for not observed.

Observing that the electrostatic/non-electrostatic CA ratios are comparable (0.25 ± 0.04 on average), we could interpret this difference as a relative weakness of the ZF1/guanine interactions. Focusing on the extended ZF1/NA interfaces (Table 6) reveals that the nature of the sandwiched nucleotide, guanine or pyrimidine, does not systematically affect the interface area (see for instance guanine 211 in 1F6U *versus* cytosine 2 in 1BJ6 in Fig. 5-B and Table 6). Yet, the presence of a pyrimidine clearly disfavors the electrostatic component, the electrostatic/non-electrostatic CA ratio decreases down to 0.06, against 0.25 for a guanine.

In the three NCp7-DNA complexes (Fig. 6-C), and thus irrespective of the DNA folding, a pyrimidine is twisted so that large CAs occur between the attached sugar, in particular the O4', C4' and C5' atoms, and THR 24 (Supplemental Table S1); the same base also interacts with PHE 16 *via* either face or stacked configurations known to be energetically equivalent for all PHE/nucleotide couples [51]. This pyrimidine acts as a shield precluding PHE 16 contact with any other nucleotide (Supplemental Table S1). The NCp7-stem loop RNA complexes are constituted by PHE 16 and ILE 24 surrounding a guanine (Fig. 6-D) in a manner that evokes the ZF2/NA interface (Fig. 6-A) while here, the guanine and PHE 16 adopt a face conformation. The other side of the same guanine engages large contacts with ILE 24 (Fig. 6-D and Table 6), comprising CH/π interactions recurrently observed in protein/RNA interfaces [52,53]. A very similar interaction scheme was described for PHE 16-GUA-ILE 24 in a NCp7- single-strand RNA complex for which corresponding model coordinates are unfortunately unavailable [54], indicating that, as for the NCp7-DNA complexes, the NA folding/unfolding does not change the interface pattern. Another point shared by all the complexes concerns the CAs associated to PHE 16/nucleotide/amino acid 24 which show similar values, 45 ± 6 $Å^2$ and 51 ± 6 $Å^2$ in NCp7-DNA and NCp7-RNA complexes respectively, despite the change of interacting partners and major structural differences. For comparison, the surface of the TRP 37/GUA/GLN 45 interface reaches 73 ± 8 $Å^2$.

A last comment relates to the orientation of NCp7 and NA, parallel or anti-parallel, previously postulated to be a hallmark of NCp7-DNA and NCp7-RNA complexes [41,43]. Here, the CA profiles along the 5' → 3' course of NA sequences (Fig. 5-B) already indicate that the NCp7 N-ter → C-ter directions are not identical in NCp7-DNA and NCp7-RNA complexes. Focusing on the nucleotides contacted by the couple typical of ZF1, PHE 16 and residue 24, much more clearly shows that, in NCp7-DNA or NCp7-RNA structures, they precede (N-ter → C-ter and 5' → 3') or follow (N-ter → C-ter and 3' → 5') the guanine stacked with TRP 37 and GLN 45 of ZF2, respectively (Supplemental Fig. S6). However, nothing in the interface composition – for instance specific contacts with the OH group of RNA sugars - relates the relative NCp7/NA direction to any discrimination mode between RNA and DNA.

Our VLDM approach provides in particular an objective and quantitative description of the interfaces involving both zinc fingers and some residues of the linker. In addition, it emphasizes the contrast between the variability of the extended ZF1/NA interfaces and the robust organization of ZF2/NA contacts, almost perfectly reproduced across the systems. Keeping in mind that the ZF2 sequence and the linker contacted segment are identical in all the complexes, the uniformity of the extended ZF2/NA interface reveals its insensitivity to the other two variables, the NA conformation and sequence. The extended ZF1/NA interfaces testify of an ability to accommodate different i) amino acids at position 24 (ILE vs TRP) and 26 (ARG vs LYS), ii) NA nature (RNA vs DNA) and iii) NA sequences.

DISCUSSION

We analyzed here the available models of NCp7 bound to NA, a collection of five NCp7-NA systems differing by the NA nature (RNA vs DNA), the NA folding (stem-loop vs single strand), the base sequence, and, ultimately, the origin of NCp7. Our investigations also incorporated the free NCp7 models, recently published ([32]. These sets of models were exploited for their ability to provide an experimental-based view of the NCp7 shape variability, keeping in mind that they do not necessary

reflect the relative populations of different 3D organizations when several structural families are explored.

In a first part, we scrutinized the structure of the ensemble constituted by the two zinc fingers ZF1 and ZF2 that have a leading role in NA binding. Then, the NCp7/NA interfaces were described in detail by mapping the contacts between amino acids and nucleotides; their quantification was achieved from a strict topological point of view by measuring contact areas. Our rationale and quantitative approach enabled us to collect information that led in particular to a scenario for the NCp7-NA binding process, unifying and complementing some aspects already proposed.

Introducing the notion of a temporal succession between the bindings of ZF1 and ZF2 is very tempting but premature given our current state of knowledge. However, the larger accessibility of ZF2 compared to ZF1 in free NCp7 (Fig. 4-A) argues for an initial event involving ZF2. During this step ZF2 specifically recognizes one accessible guanine, ignoring available nucleotides of other type. The open conformations sampled by free NCp7, by enhancing the exposition of TRP 37 to the solvent (Fig. 4-B), likely facilitate the detection and the fixation of the targeted guanine by this residue. This guanine in fact concentrates most contacts (Fig. 5-B) that primarily involve TRP 37 and GLN 45 but also engage additional amino acids belonging to both linker and ZF2 (Table 5 and Supplemental Fig. S5). The organization of this interface is remarkably consistent across the NCp7-NA models, implying that the binding step involving ZF2 is insensitive to the guanine neighboring sequence as well as to the NA conformation and nature (Fig. 6-A).

After or simultaneously with the ZF2 binding event, ZF1 interacts with the region adjacent to the guanine contacted by ZF2, according to two modes that are further discussed below. PHE 16 and ARG 26 are the most accessible ZF1 residues in free NCp7, independently of its intrinsic dynamics (Fig. 4). Considering their important implication in the ZF1 interfaces (Fig. 5-A), these amino acids may be decisive at this stage of NA binding. Finally, once assembled to NA, NCp7 adopts a restricted collection of conformations that, globally, corresponds to a unique relative orientation of the two zinc fingers with respect to each other (Table 2; Fig. 2). This 3D arrangement perfectly matches one closed

type of conformation sampled by free NCp7 (Fig. 3-B), whose assembling in complexes should therefore induce minimal energetic cost.

In the light of the former description, NCp7 appears as a very interesting case in which not only one but all the conformations sampled by the protein in its free state have the potential to be exploited during the NA binding process. Thus, NCp7 meets the criteria of the "linkage scheme for binding" previously described as a mix between two mechanisms, the conformational selection and the induced fit ([55]. According to this scheme, a macromolecule exists as multiple conformations capable of interacting with its targets but, after binding, one of the free forms is trapped in the complex, which is stabilized *via* interactions with the target that obviously alter the free energetic landscape.

Concerning the functions of the two zinc fingers, the elements concerning ZF2 in the above considerations (large accessibility in free NCp7, conserved interface across the systems) clearly corroborate the strategic role of this zinc finger in the guanine recognition process. That ZF2 is the principal actor of this function is supported by the dramatic decrease in affinity for RNA observed when ZF2 was deleted in a Gag context, an effect that does not appear with ZF1 deletion [49]. The ZF1/NA interfaces, despite their variable organizations, show characteristics shared by the five studied systems, which at last relate to the ZF1 function. The major contact areas of these interfaces do not clearly depend on a particular type of nucleotide (Table 6). In addition, the nucleotides, guanines as well as pyrimidines, interacting with the key ZF1 amino acids 16 and 24 are never so firmly anchored as the guanines contacted by ZF2 (Table 6 *versus* Table 5). Nevertheless, ZF1 and some neighboring amino acids cover numerous nucleotides, so that the total areas of the corresponding interfaces reach extensive values in complexes containing stem-loops of various sequences (Supplemental Table S1). These observations point towards the ability of ZF1 to lock NCp7-NA complexes without really needing well-defined NA sequence. In addition, ZF1 was previously assumed to be responsible for the NCp7 chaperone activity [34,35]. In this context, 2JZW was presented as an example of the early stage of unfolding a strong secondary structure [41]. Indeed, our analysis shows that a set of five ZF1 residues (GLY 22, HSD 23, THR 24, ARG 26 and ASN 27) engage substantial contacts with the first C_{105}:G_{111} base pair next to the loop (Supplemental Table S1), which gives clear signs of destabilization

[41]. In sum, rather than for distinct roles, our results advocate for two specializations of ZF1 and ZF2: ZF2 would be in charge of the recognition phase itself, while ZF1 would assist the stabilization of the NCp7-NA complexes and, if appropriate, carry out the chaperone function.

We now come back to the extended ZF1/NA interfaces that, regardless of the NA conformation (stem-loop *versus* single strand), are organized in two global patterns (Fig. 6-C) and display either parallel or anti-parallel mutual orientation of NCp7 and NA. At first sight, these organizations are typical of the NA nature (NCp7/RNA *versus* NCp7/DNA). However, the complexes also differ by other features comprising the type of the amino acids at position 24 (ILE or THR) and 26 (ARG or LYS). It would be amazing that changing ARG for LYS or the inverse is the main source of the interface alterations since both residues have equivalent properties – here, they essentially interact thanks to their similar long aliphatic side chains. This argument can no longer be invoked for polar THR and hydrophobic LEU or ILE. Indeed, THR → LEU mutation affects the NCp7 affinity for specific RNA sequences ([38]. Thus, before we are able to conclude about the reasons that preside over the stabilization of either one or the other organization, it would be desirable, if not essential, to determine the effect of the presence of LEU or ILE 24 on the interface with DNA, reminding that these amino acids are present in 95% of HIV-1 NCp7 (http://hivmut.org, ([56]). Whatever the causes, the ZF1/NA arrangements reveal a definite structural plasticity of the NCp7. This is a very important point, given the ability of this protein to assume many different functions.

Finally, our analyses also stress some additional points concerning the protein/NA interfaces. Most ARG and LYS exhibit a large surface accessible to the water molecules in free NCp7 (Fig. 4) ; in the complexes they are associated to important CAs (Fig. 5), in line with their role in NA interaction highlighted by using mutants [9]. Nevertheless, they are generally engaged in contacts that are quite variable in terms of number and location of nucleotides (Supplemental Table S1), suggesting that they have a role in finely adjusting the partners together rather than in the binding events, strictly speaking. Although positively charged, the major contribution of these amino acids to the interfaces consists in making hydrophobic interactions thanks to the aliphatic part of their side chains. With the additional contribution of other amino acids such as aromatic residues, hydrophobic interactions are omnipresent

in the NCp7-NA complexes (Table 4). This should be put in perspective with previous studies that also underline the substantial occurrence of non-electrostatic contacts in protein–DNA structures [45,57]. The case of NCp7 containing complexes reinforces the idea that protein/NA interfaces are stabilized by both hydrophobic and electrostatic interactions.

MATERIALS AND METHODS

NCp7/NA models

We examined available experimental models of NCp7-DNA and NCp7-RNA complexes, all based on NMR data and deposited in the PDB under the codes 1A1T, 1F6U, 2JZW, 1BJ6 and 2L4L. As usual in the case of structures from NMR, each PDB file provided a series of models (Table 1) that were all studied here. Information about the nucleic acids (NA) and protein contained in each complex is supplied in the Table 1. Since two different VIH-1 strains were used to produce NCp7, THR 12, ILE 24 and LYS 26 in NCp7-RNA complexes (1A1T, 1F6U) becomes ASN 12, THR 24 and ARG 26 in NCp7-DNA complexes (2JZW, 1BJ6, 2L4L). Fig. 1 illustrates the composition(s) and numbering of NCp7.

Interface analysis

The NCp7/NA interface was analyzed by VLDM (Voronoi Laguerre Delaunay for Macromolecules), a software originally developed for proteins [44] and recently extended to nucleic acids [45]. VLDM relies on a tessellation method, that is, a partition of space into a collection of polyhedra filling space without overlaps or gaps. The 3D structure entered as input is initially solvated by an 8 Å thick water layer using the Solvate procedure [44,45] to avoid open or distorted polyhedra in the tessellation. The partition of space is carried out on the solute and the solvent atoms considered as a set of sites defined by atomic positions and weights depending on the atom van der Waals radii. Technically, the Delaunay tessellation is first built on all atoms of the whole system; then the Laguerre tessellation is deduced as the geometric dual of the Delaunay diagram. In the Laguerre tessellation,

each polyhedron is convex and most often encloses a single atom. The shape of these polyhedra is variable, but it only depends on the mutual positions of neighboring atoms. In this sense, the Laguerre partition is a faithful representation of the structure, free from adjustable parameters. The contacts are represented as facets shared by two nearest neighbor polyhedral. In the present analysis, only the heavy atoms of the solute or water molecules were considered.

In this approach, a contact occurs whenever two atoms share a common face in the tessellation. The interface between two molecules or molecular groups is a polygonal surface, quantified by its area (CA for contact area). The protein or NA accessibility was represented by the contact areas between water molecules and solute residues. The NCp7/NA interfaces were quantified by the contact areas between protein residues and nucleotides. These interfaces were also analyzed according to the contact nature. Electrostatic contacts involve N and O atoms (N-N, N-O or O-O), excluding repulsive interactions between two donors or two acceptors; hydrogen bonds and salt bridges belong to this category. The other types report either hydrophobic contacts involving carbon atoms exclusively (C-C) or a simple proximity of N-C or O-C atoms. Examination of the distances characterizing the hydrophobic or electrostatic contacts showed maximal distribution peaks at 4 Å (from 3.5 to 6 Å) for C-C contacts and 2.5 Å (from 2.5 to 5 Å) for N-O, N-N and O-O contacts.

Hydrogen bonds between donor (D) and acceptor (A) were calculated with HBPlus [58], using as existence criteria D-A distance < 3.9 Å and D-H-A angle > 120°.

ACKNOWLEGMENTS

The authors thank Dr Philippe Fossé (LBPA, ENS Paris-Saclay) for interesting discussions about the biology of NCp7; they also thank Ahmad Elbahnsi who carried out preliminary investigations.

FUNDINGS

The authors gratefully thank SIDACTION for financial support (Ref 15-2-AEQ-04-01) of their research.

AUTHOR CONTRIBUTIONS

Romain Retureau carried out the analyses and was in charge of figures. Christophe Oguey, the developer of the VLDM software, and Olivier Mauffret participated in discussions and in rereading the manuscript. Brigitte Hartmann wrote the manuscript.

REFERENCES

[1] M.F. Summers, L.E. Henderson, M.R. Chance, J.W. Bess, T.L. South, P.R. Blake, I. Sagi, G. Perez-Alvarado, R.C. Sowder, D.R. Hare, Nucleocapsid zinc fingers detected in retroviruses: EXAFS studies of intact viruses and the solution-state structure of the nucleocapsid protein from HIV-1, Protein Sci. Publ. Protein Soc. 1 (1992) 563–574. doi:10.1002/pro.5560010502.

[2] E.O. Freed, HIV-1 assembly, release and maturation, Nat. Rev. Microbiol. 13 (2015) 484–496. doi:10.1038/nrmicro3490.

[3] E. Mailler, S. Bernacchi, R. Marquet, J.-C. Paillart, V. Vivet-Boudou, R.P. Smyth, The Life-Cycle of the HIV-1 Gag-RNA Complex, Viruses. 8 (2016). doi:10.3390/v8090248.

[4] M. Comas-Garcia, S.R. Davis, A. Rein, On the Selective Packaging of Genomic RNA by HIV-1, Viruses. 8 (2016). doi:10.3390/v8090246.

[5] J.A. Thomas, R.J. Gorelick, Nucleocapsid protein function in early infection processes, Virus Res. 134 (2008) 39–63. doi:10.1016/j.virusres.2007.12.006.

[6] J.-L. Darlix, J.L. Garrido, N. Morellet, Y. Mély, H. de Rocquigny, Properties, functions, and drug targeting of the multifunctional nucleocapsid protein of the human immunodeficiency virus, Adv. Pharmacol. San Diego Calif. 55 (2007) 299–346. doi:10.1016/S1054-3589(07)55009-X.

[7] S.B. Kutluay, P.D. Bieniasz, Analysis of the initiating events in HIV-1 particle assembly and genome packaging, PLoS Pathog. 6 (2010) e1001200. doi:10.1371/journal.ppat.1001200.

[8] J.G. Levin, M. Mitra, A. Mascarenhas, K. Musier-Forsyth, Role of HIV-1 nucleocapsid protein in HIV-1 reverse transcription, RNA Biol. 7 (2010) 754–774.

[9] H. Wu, M. Mitra, M.N. Naufer, M.J. McCauley, R.J. Gorelick, I. Rouzina, K. Musier-Forsyth, M.C. Williams, Differential contribution of basic residues to HIV-1 nucleocapsid protein's nucleic acid chaperone function and retroviral replication, Nucleic Acids Res. 42 (2014) 2525–2537. doi:10.1093/nar/gkt1227.

[10] H. Wu, M. Mitra, M.J. McCauley, J.A. Thomas, I. Rouzina, K. Musier-Forsyth, M.C. Williams, R.J. Gorelick, Aromatic residue mutations reveal direct correlation between HIV-1 nucleocapsid protein's nucleic acid chaperone activity and retroviral replication, Virus Res. 171 (2013) 263–277. doi:10.1016/j.virusres.2012.07.008.

[11] J.A.G. Briggs, M.N. Simon, I. Gross, H.-G. Kräusslich, S.D. Fuller, V.M. Vogt, M.C. Johnson, The stoichiometry of Gag protein in HIV-1, Nat. Struct. Mol. Biol. 11 (2004) 672–675. doi:10.1038/nsmb785.

[12] E. Chertova, O. Chertov, L.V. Coren, J.D. Roser, C.M. Trubey, J.W. Bess, R.C. Sowder, E. Barsov, B.L. Hood, R.J. Fisher, K. Nagashima, T.P. Conrads, T.D. Veenstra, J.D. Lifson, D.E. Ott, Proteomic and biochemical analysis of purified human immunodeficiency virus type 1 produced from infected monocyte-derived macrophages, J. Virol. 80 (2006) 9039–9052. doi:10.1128/JVI.01013-06.

[13] E.W. Abd El-Wahab, R.P. Smyth, E. Mailler, S. Bernacchi, V. Vivet-Boudou, M. Hijnen, F. Jossinet, J. Mak, J.-C. Paillart, R. Marquet, Specific recognition of the HIV-1 genomic RNA by the Gag precursor, Nat. Commun. 5 (2014) 4304. doi:10.1038/ncomms5304.

[14] S.B. Kutluay, T. Zang, D. Blanco-Melo, C. Powell, D. Jannain, M. Errando, P.D. Bieniasz, Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis, Cell. 159 (2014) 1096–1109. doi:10.1016/j.cell.2014.09.057.

[15] K.A. Wilkinson, R.J. Gorelick, S.M. Vasa, N. Guex, A. Rein, D.H. Mathews, M.C. Giddings, K.M. Weeks, High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states, PLoS Biol. 6 (2008) e96. doi:10.1371/journal.pbio.0060096.

[16] R.J. Fisher, A. Rein, M. Fivash, M.A. Urbaneja, J.R. Casas-Finet, M. Medaglia, L.E. Henderson, Sequence-specific binding of human immunodeficiency virus type 1 nucleocapsid protein to short oligonucleotides, J. Virol. 72 (1998) 1902–1909.

[17] C. Vuilleumier, E. Bombarda, N. Morellet, D. Gérard, B.P. Roques, Y. Mély, Nucleic acid sequence discrimination by the HIV-1 nucleocapsid protein NCp7: a fluorescence study, Biochemistry. 38 (1999) 16816–16825.

[18] J.A. Berglund, B. Charpentier, M. Rosbash, A high affinity binding site for the HIV-1 nucleocapsid protein, Nucleic Acids Res. 25 (1997) 1042–1049.

[19] S.J. Kim, M.Y. Kim, J.H. Lee, J.C. You, S. Jeong, Selection and stabilization of the RNA aptamers against the human immunodeficiency virus type-1 nucleocapsid protein, Biochem. Biophys. Res. Commun. 291 (2002) 925–931. doi:10.1006/bbrc.2002.6521.

[20] J.K. Grohman, R.J. Gorelick, C.R. Lickwar, J.D. Lieb, B.D. Bower, B.M. Znosko, K.M. Weeks, A guanosine-centric mechanism for RNA chaperone function, Science. 340 (2013) 190–195. doi:10.1126/science.1230715.

[21] P. Allen, B. Collins, D. Brown, Z. Hostomsky, L. Gold, A Specific RNA Structural Motif Mediates High Affinity Binding by the HIV-1 Nucleocapsid Protein (NCp7), Virology. 225 (1996) 306–315. doi:10.1006/viro.1996.0605.

[22] A. Bazzi, L. Zargarian, F. Chaminade, H. De Rocquigny, B. René, Y. Mély, P. Fossé, O. Mauffret, Intrinsic nucleic acid dynamics modulates HIV-1 nucleocapsid protein binding to its targets, PloS One. 7 (2012) e38905. doi:10.1371/journal.pone.0038905.

[23] L. Pappalardo, D.J. Kerwood, I. Pelczer, P.N. Borer, Three-dimensional folding of an RNA hairpin required for packaging HIV-1, J. Mol. Biol. 282 (1998) 801–818. doi:10.1006/jmbi.1998.2046.

[24] A.C. Paoletti, M.F. Shubsda, B.S. Hudson, P.N. Borer, Affinities of the nucleocapsid protein for variants of SL3 RNA in HIV-1, Biochemistry. 41 (2002) 15423–15428.

[25] G. Mirambeau, S. Lyonnais, R.J. Gorelick, Features, processing states, and heterologous protein interactions in the modulation of the retroviral nucleocapsid protein function, RNA Biol. 7 (2010) 724–734.

[26] M. Cruceanu, M.A. Urbaneja, C.V. Hixson, D.G. Johnson, S.A. Datta, M.J. Fivash, A.G. Stephen, R.J. Fisher, R.J. Gorelick, J.R. Casas-Finet, A. Rein, I. Rouzina, M.C. Williams, Nucleic acid binding and chaperone properties of HIV-1 Gag and nucleocapsid proteins, Nucleic Acids Res. 34 (2006) 593–605. doi:10.1093/nar/gkj458.

[27] T. Wu, R.J. Gorelick, J.G. Levin, Selection of fully processed HIV-1 nucleocapsid protein is required for optimal nucleic acid chaperone activity in reverse transcription, Virus Res. 193 (2014) 52–64. doi:10.1016/j.virusres.2014.06.004.

[28] B. René, O. Mauffret, P. Fossé, Retroviral nucleocapsid proteins and DNA strand transfers, Biochim. Open. 7 (2018) 10–25. doi:10.1016/j.biopen.2018.07.001.

[29] J.G. Omichinski, G.M. Clore, K. Sakaguchi, E. Appella, A.M. Gronenborn, Structural characterization of a 39-residue synthetic peptide containing the two zinc binding domains from the HIV-1 p7 nucleocapsid protein by CD and NMR spectroscopy, FEBS Lett. 292 (1991) 25–30.

[30] N. Morellet, H. de Rocquigny, Y. Mély, N. Jullian, H. Déméné, M. Ottmann, D. Gérard, J.L. Darlix, M.C. Fournie-Zaluski, B.P. Roques, Conformational behaviour of the active and inactive forms of the nucleocapsid NCp7 of HIV-1 studied by 1H NMR, J. Mol. Biol. 235 (1994) 287–301.

[31] B.M. Lee, R.N. De Guzman, B.G. Turner, N. Tjandra, M.F. Summers, Dynamical behavior of the HIV-1 nucleocapsid protein, J. Mol. Biol. 279 (1998) 633–649. doi:10.1006/jmbi.1998.1766.

[32] L. Deshmukh, C.D. Schwieters, A. Grishaev, G.M. Clore, Quantitative Characterization of Configurational Space Sampled by HIV-1 Nucleocapsid Using Solution NMR, X-ray Scattering and Protein Engineering, Chemphyschem Eur. J. Chem. Phys. Phys. Chem. 17 (2016) 1548–1552. doi:10.1002/cphc.201600212.

[33] L. Zargarian, C. Tisné, P. Barraud, X. Xu, N. Morellet, B. René, Y. Mély, P. Fossé, O. Mauffret, Dynamics of linker residues modulate the nucleic acid binding properties of the HIV-1 nucleocapsid protein zinc fingers, PloS One. 9 (2014) e102150. doi:10.1371/journal.pone.0102150.

[34] J. Guo, T. Wu, B.F. Kane, D.G. Johnson, L.E. Henderson, R.J. Gorelick, J.G. Levin, Subtle alterations of the native zinc finger structures have dramatic effects on the nucleic acid chaperone activity of human immunodeficiency virus type 1 nucleocapsid protein, J. Virol. 76 (2002) 4370–4378.

[35] M.J. Heath, S.S. Derebail, R.J. Gorelick, J.J. DeStefano, Differing roles of the N- and C-terminal zinc fingers in human immunodeficiency virus nucleocapsid protein-enhanced nucleic acid annealing, J. Biol. Chem. 278 (2003) 30755–30763. doi:10.1074/jbc.M303819200.

[36] P.-J. Racine, C. Chamontin, H. de Rocquigny, S. Bernacchi, J.-C. Paillart, M. Mougel, Requirements for nucleocapsid-mediated regulation of reverse transcription during the late steps of HIV-1 assembly, Sci. Rep. 6 (2016) 27536. doi:10.1038/srep27536.

[37] R.J. Gorelick, D.J. Chabot, A. Rein, L.E. Henderson, L.O. Arthur, The two zinc fingers in the human immunodeficiency virus type 1 nucleocapsid protein are not functionally equivalent, J. Virol. 67 (1993) 4027–4036.

[38] J. Dannull, A. Surovoy, G. Jung, K. Moelling, Specific binding of HIV-1 nucleocapsid protein to PSI RNA in vitro requires N-terminal zinc finger and flanking basic amino acid residues, EMBO J. 13 (1994) 1525–1533.

[39] R.N. De Guzman, Z.R. Wu, C.C. Stalling, L. Pappalardo, P.N. Borer, M.F. Summers, Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element, Science. 279 (1998) 384–388.

[40] G.K. Amarasinghe, R.N. De Guzman, R.B. Turner, K.J. Chancellor, Z.R. Wu, M.F. Summers, NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the psi-RNA packaging signal. Implications for genome recognition, J. Mol. Biol. 301 (2000) 491–511. doi:10.1006/jmbi.2000.3979.

[41] S. Bourbigot, N. Ramalanjaona, C. Boudier, G.F.J. Salgado, B.P. Roques, Y. Mély, S. Bouaziz, N. Morellet, How the HIV-1 nucleocapsid protein binds and destabilises the (-)primer binding site during reverse transcription, J. Mol. Biol. 383 (2008) 1112–1128. doi:10.1016/j.jmb.2008.08.046.

[42] N. Morellet, H. Déméné, V. Teilleux, T. Huynh-Dinh, H. de Rocquigny, M.C. Fournié-Zaluski, B.P. Roques, Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d(ACGCC), J. Mol. Biol. 283 (1998) 419–434. doi:10.1006/jmbi.1998.2098.

[43] A. Bazzi, L. Zargarian, F. Chaminade, C. Boudier, H. De Rocquigny, B. René, Y. Mély, P. Fossé, O. Mauffret, Structural insights into the cTAR DNA recognition by the HIV-1 nucleocapsid protein: role of sugar deoxyriboses in the binding polarity of NC, Nucleic Acids Res. 39 (2011) 3903–3916. doi:10.1093/nar/gkq1290.

[44] J. Esque, S. Leonard, A.G. de Brevern, C. Oguey, VLDP web server: a powerful geometric tool for analysing protein structures in their environment, Nucleic Acids Res. 41 (2013) W373–W378. doi:10.1093/nar/gkt509.

[45] A. Elbahnsi, R. Retureau, M. Baaden, B. Hartmann, C. Oguey, Holding the Nucleosome Together: A Quantitative Description of the DNA-Histone Interface in Solution, J. Chem. Theory Comput. 14 (2018) 1045–1058. doi:10.1021/acs.jctc.7b00936.

[46] R. Wolfenden, L. Andersson, P.M. Cullis, C.C. Southgate, Affinities of amino acid side chains for solvent water, Biochemistry. 20 (1981) 849–855.

[47] L. Biedermannová, B. Schneider, Structure of the ordered hydration of amino acids in proteins: analysis of crystal structures, Acta Crystallogr. D Biol. Crystallogr. 71 (2015) 2192–2202. doi:10.1107/S1399004715015679.

[48] S. Simm, J. Einloft, O. Mirus, E. Schleiff, 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification, Biol. Res. 49 (2016) 31. doi:10.1186/s40659-016-0092-5.

[49] J.A. Webb, C.P. Jones, L.J. Parent, I. Rouzina, K. Musier-Forsyth, Distinct binding interactions of HIV-1 Gag to Psi and non-Psi RNAs: implications for viral genomic RNA packaging, RNA N. Y. N. 19 (2013) 1078–1088. doi:10.1261/rna.038869.113.

[50] S.S. Athavale, W. Ouyang, M.P. McPike, B.S. Hudson, P.N. Borer, Effects of the nature and concentration of salt on the interaction of the HIV-1 nucleocapsid protein with SL3 RNA, Biochemistry. 49 (2010) 3525–3533. doi:10.1021/bi901279e.

[51] L.R. Rutledge, H.F. Durst, S.D. Wetmore, Evidence for Stabilization of DNA/RNA-Protein Complexes Arising from Nucleobase-Amino Acid Stacking and T-Shaped Interactions, J. Chem. Theory Comput. 5 (2009) 1400–1410. doi:10.1021/ct800567q.

[52] H. Zhang, C. Li, F. Yang, J. Su, J. Tan, X. Zhang, C. Wang, Cation-pi interactions at non-redundant protein--RNA interfaces, Biochem. Biokhimiia. 79 (2014) 643–652. doi:10.1134/S0006297914070062.

[53] S.Z. Borozan, B.P. Dimitrijević, S.Đ. Stojanović, Cation-π interactions in high resolution protein-RNA complex crystal structures, Comput. Biol. Chem. 47 (2013) 105–112. doi:10.1016/j.compbiolchem.2013.08.005.

[54] S. Spriggs, L. Garyu, R. Connor, M.F. Summers, Potential intra- and intermolecular interactions involving the unique-5' region of the HIV-1 5'-UTR, Biochemistry. 47 (2008) 13064–13073. doi:10.1021/bi8014373.

[55] A.D. Vogt, N. Pozzi, Z. Chen, E. Di Cera, Essential role of conformational selection in ligand binding, Biophys. Chem. 186 (2014) 13–21. doi:10.1016/j.bpc.2013.09.003.

[56] N.E. Davey, V.P. Satagopam, S. Santiago-Mozos, C. Villacorta-Martin, T.A.M. Bharat, R. Schneider, J.A.G. Briggs, The HIV mutation browser: a resource for human immunodeficiency virus mutagenesis and polymorphism data, PLoS Comput. Biol. 10 (2014) e1003951. doi:10.1371/journal.pcbi.1003951.

[57] N.M. Luscombe, R.A. Laskowski, J.M. Thornton, Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level, Nucleic Acids Res. 29 (2001) 2860–2874.

[58] I.K. McDonald, J.M. Thornton, Satisfying hydrogen bonding potential in proteins, J. Mol. Biol. 238 (1994) 777–793. doi:10.1006/jmbi.1994.1334.

FIGURES LEGENDS

Fig. 1: Schematic representation of NCp7.

This schema presents the 55 amino acids composing NCp7 from the NL4-3 strain of HIV-1; the residues 12, 24 and 26 in brackets are found in NCp7 from the M-B strain. The two zinc fingers (ZF1 in red and ZF2 in cyan) are separated by a short linker (grey) and surrounded by N- and C-terminal domains (black). In 2JZW, 1BJ6 and 2L4L, 10 or 11 amino acids of N-terminal domain are lacking (see Table 1).

Fig. 2: Characteristics of the relative positioning of ZF1 and ZF2 in bound NCp7

The three panels represent the backbone trace of ZF1 (red) and ZF2 (cyan) separated by the linker (grey). The structure used here is the model 1 of 2L4L. Three couples of vectors (yellow arrows) were chosen to characterize the relative orientation of the two zinc fingers with respect to each other: $Zn_{ZF1} \rightarrow C\alpha_{LYS20}$ / $Zn_{ZF2} \rightarrow C\alpha_{GLY43}$ (a), $Zn_{ZF1} \rightarrow C\alpha_{CYS15}$ / $Zn_{ZF2} \rightarrow C\alpha_{CYS49}$ (b) and $Zn_{ZF1} \rightarrow C\alpha_{CYS28}$ / $Zn_{ZF2} \rightarrow C\alpha_{CYS49}$ (c).

Fig. 3: Typical structures of free ZF1-ZF2 ensembles and comparison with a bound structure.

In these representations showing NCp7 backbone traces, ZF1 is in red, ZF2 in cyan, the linker in grey and the N- and C-terminal domains in black. (a): Typical structures of the four groups defined in Table 3 from the 21 models of free NCp7 in 5I1R; model 1, 2, 3 and 10 are representative of Group 1, 2, 3, and 4, respectively. (b) : Superimposition of the backbone trace of the ZF1-ZF2 ensembles in free (model 9 of Group 3 of 5I1R) and bound (model 16 of 2JZW) NCp7; the corresponding RMSD is 2.2 Å.

Fig. 4: Accessibility to water molecules of amino acids in free NCp7 and the particular case of TRP 37.

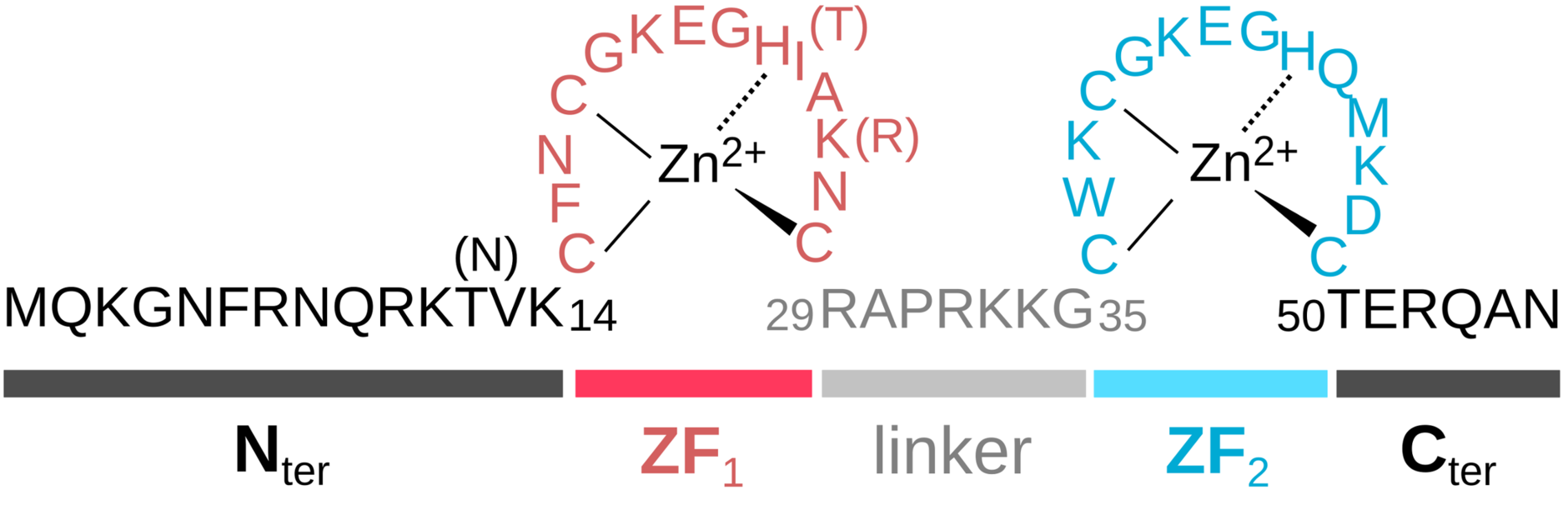
(a): The contact areas (CA_{av}) between amino acids and water molecules are plotted along the free NCp7 sequence for the four structural groups identified in Table 3. The data were averaged over the models that constitute each group; the vertical error bars correspond to standard deviations. (b): These pie charts represent the contact areas ($Å^2$) between TRP 37 and either amino acids or water molecules in the same four structural groups. In both panels, the data associated to ZF1, linker and ZF2 residues are in red, grey and cyan, respectively.

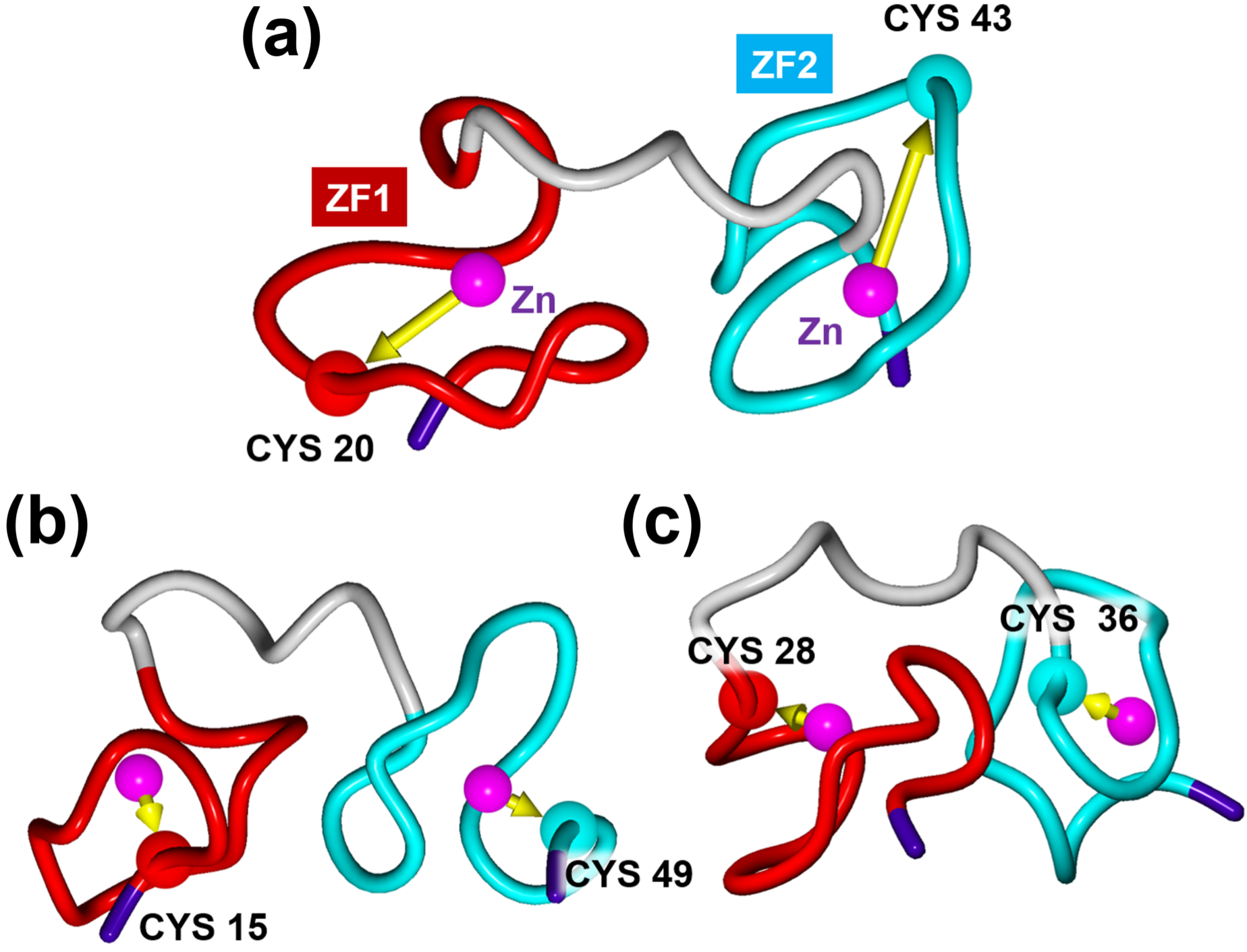
Fig. 5: Amino acids and nucleotides involved in the NCp7/NA interfaces.

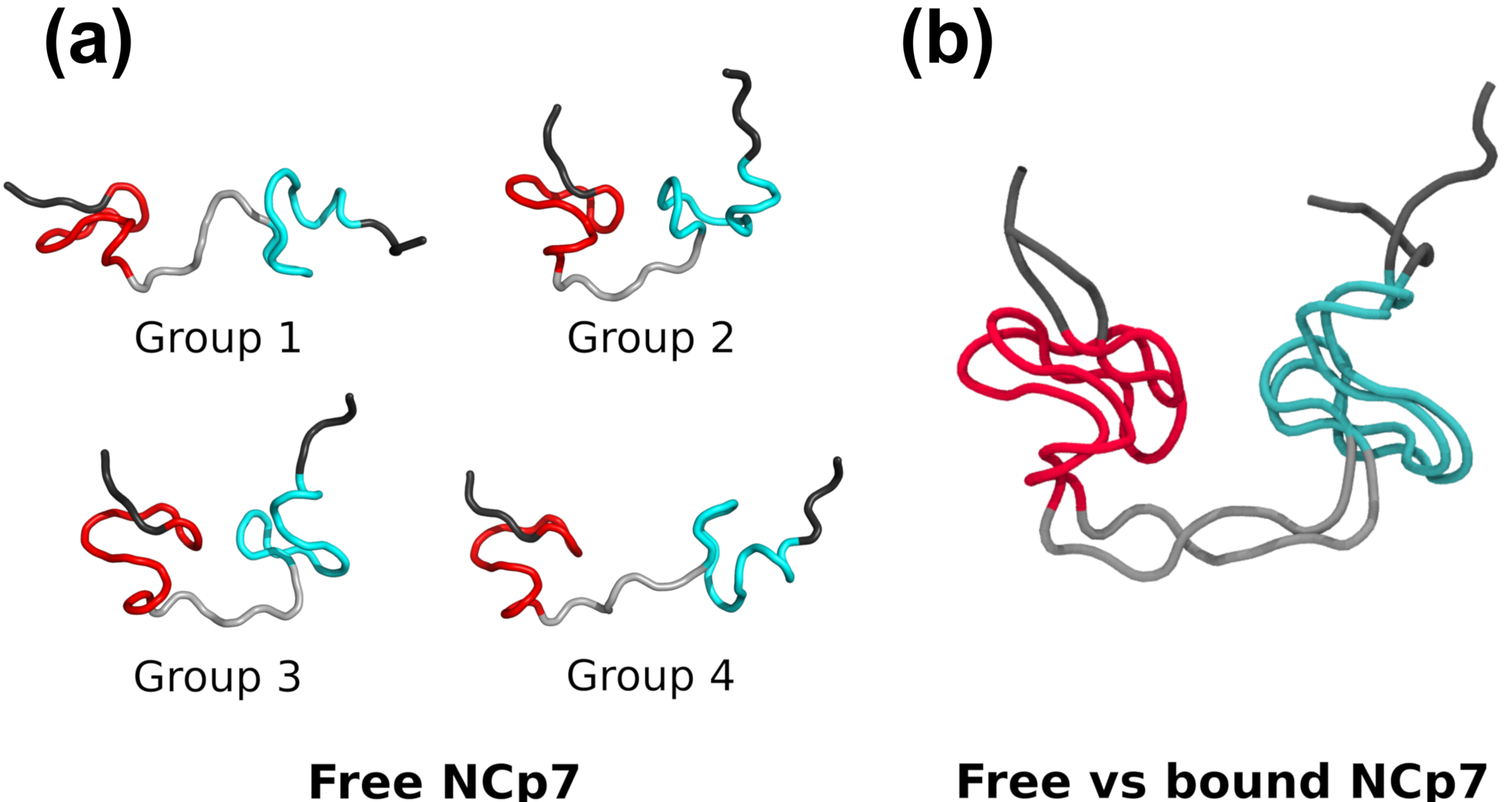
The contact areas (CA_{av}) between amino acids and nucleotides are plotted along the NCp7 sequences (a) or for the contacted nucleotides (b). The NA structures of each studied complex are also schematically represented (c). In (a) and (b) panels, the data associated to N-terminal domain, ZF1, linker and ZF2 residues are in grey, red, grey and cyan, respectively. The data were averaged over the models that constitute each system; the vertical error bars correspond to standard deviations. In the (a) panels, the vertical dashed lines point out the contacts involving PHE 16 and TRP 37, central in the ZF1 and ZF2 interfaces.

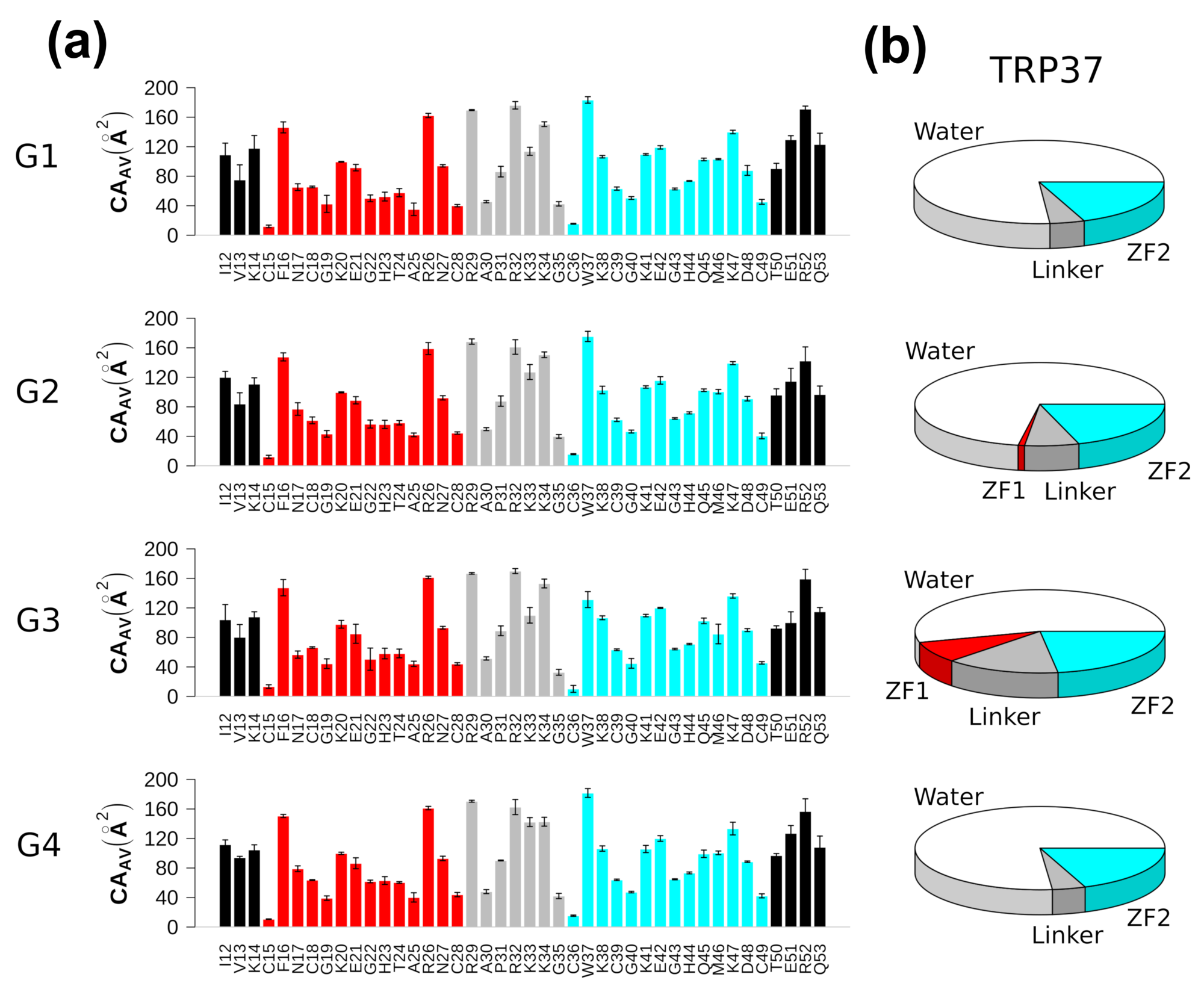
Fig. 6: Typical major interactions in NCp7/NA interfaces.

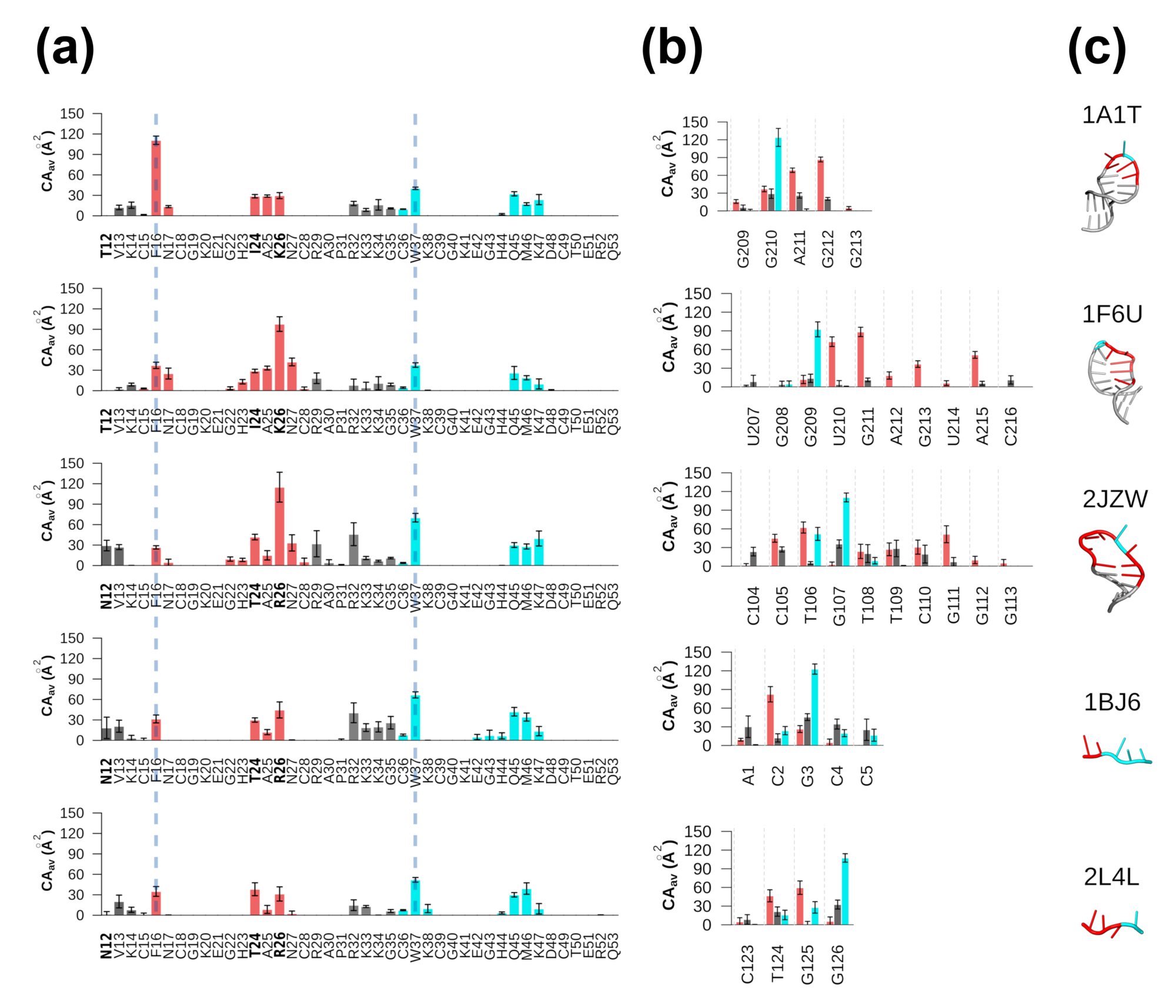
(a): representation of the TRP 37/GUA/GLN 45 interactions observed in all studied NCp7-DNA and NCp7-RNA models; the superimposition was made on the guanines. (b): representation of the TRP 37/PHE 16 interactions observed either in NCp7-DNA or NCp7-RNA models; the superimposition was made on TRP 37. (c): representation of the PHE 16/PYR/THR 24 interactions observed in NCp7-DNA models; the superimposition was made on the pyrimidines. (d): representation of the PHE 16/GUA/ILE24 interactions observed in NCp7-RNA models; the superimposition was made on the guanines.

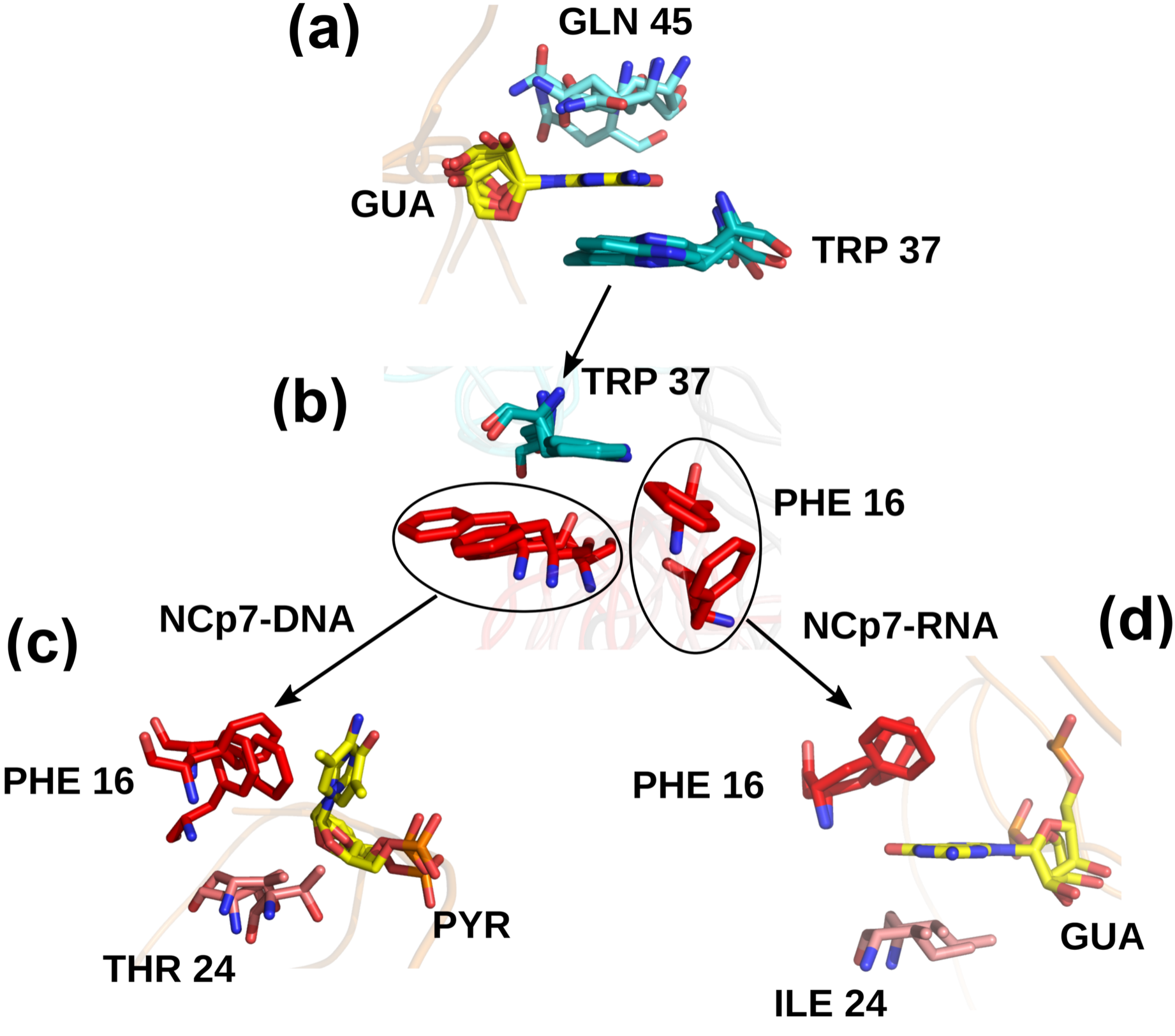

C G K E G H I (T)
N F C A K (R) N C
Zn2+
C G K E G H Q M
K W C D C
Zn2+
(N)
MQKGNFRNQRKTVK14
29RAPRKKG35
50TERQAN
Nter
ZF1
linker
ZF2
Cter

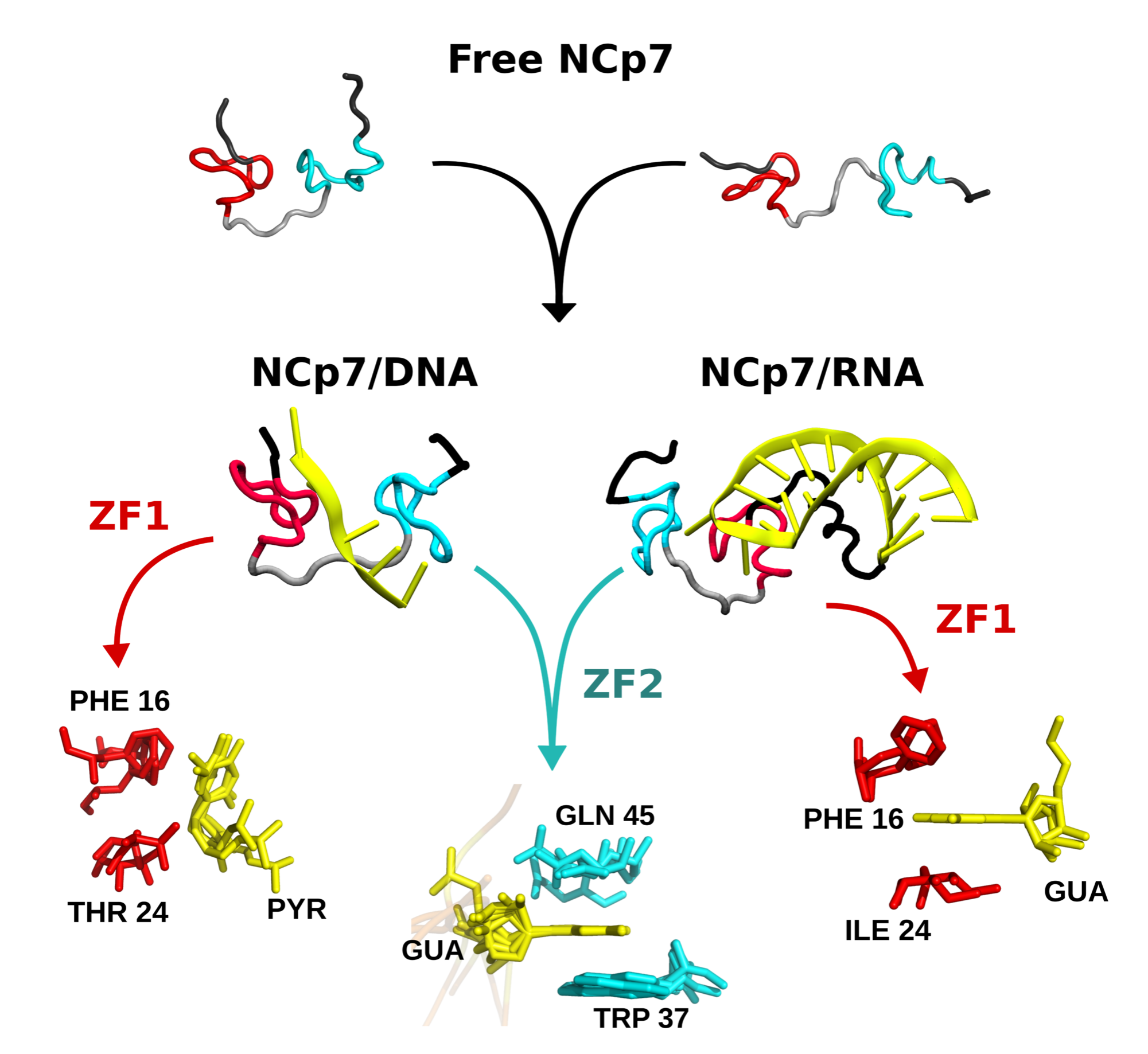

(a)
CYS 43
ZF2
ZF1
Zn
Zn
CYS 20
(b)
CYS 15
CYS 49
(c)
CYS 28
CYS 36


(a)
Group 1
Group 2
Group 3
Group 4
Free NCp7
(b)
Free vs bound NCp7


(a)
(b)
TRP37
G1
G2
G3
G4
$CA_{AV}(Å^2)$
0
40
80
120
160
200
I12 V13 K14 C15 F16 N17 C18 G19 K20 E21 G22 H23 T24 A25 R26 N27 C28 R29 A30 P31 R32 K33 K34 G35 C36 W37 K38 C39 G40 K41 E42 G43 H44 Q45 M46 K47 D48 C49 T50 E51 R52 Q53
Water
ZF1
Linker
ZF2


(a)
(b)
(c)
1A1T
1F6U
2JZW
1BJ6
2L4L
CA_av (Å^2)


(a)
GLN 45
GUA
TRP 37
(b)
TRP 37
PHE 16
NCp7-DNA
NCp7-RNA
(c)
(d)
PHE 16
PYR
THR 24
PHE 16
GUA
ILE 24


Free NCp7
NCp7/DNA
NCp7/RNA
ZF1
ZF2
ZF1
PHE 16
THR 24
PYR
GLN 45
GUA
TRP 37
PHE 16
GUA
ILE 24