



**HAL**  
open science

# Interactions acides nucléiques/protéines non spécifiques : le nucléosome et les complexes de la NCp7

Romain Retureau

► **To cite this version:**

Romain Retureau. Interactions acides nucléiques/protéines non spécifiques : le nucléosome et les complexes de la NCp7. Sciences agricoles. Université Paris Saclay (COMUE), 2019. Français. NNT : 2019SACLN057 . tel-02375634

**HAL Id: tel-02375634**

**<https://theses.hal.science/tel-02375634>**

Submitted on 22 Nov 2019

**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.

# Interactions acides nucléiques/protéines non spécifiques : le nucléosome et les complexes de la NCp7

Thèse de doctorat de l'Université Paris-Saclay  
préparée à l'ENS Paris-Saclay

École doctorale n°577 Structure et Dynamique des Systèmes Vivants  
(SDSV)

Spécialité de doctorat : Science de la vie et de la santé

Thèse présentée et soutenue à Cachan, le 24 Octobre 2019, par

**Romain Retureau**

Composition du Jury :

**Thérèse Malliavin**

Dr, Institut Pasteur (UMR3528)

Rapporteuse/Présidente

**Claudine Mayer**

Pr, Université Paris Diderot (UMR 3528)

Rapporteuse

**Véronique Arluison**

Dr, Université Paris Diderot (UMR 12 CEA CNRS)

Examinatrice

**Hugues-Olivier Bertrand**

Sr. Dir. Pre-Sales, Dassault System (BIOVIA)

Examineur

**Luca Monticelli**

Dr, IBCP (UMR 5086)

Examineur

**Brigitte Hartmann**

Dr, ENS Paris-Saclay (UMR 8113)

Directrice de thèse

## Table des matières

|   |    |
|---|----|
| Liste des abréviations .....  | 1  |
| Introduction .....  | 2  |
| Interactions acide nucléique-protéine.....                          | 3  |
| Contexte général.....   | 3  |
| Spécificité et affinité .....                                       | 4  |
| Structures des acides nucléiques .....                              | 5  |
| Principaux motifs protéiques de liaisons aux AN .....               | 8  |
| Types d'interactions .....  | 11 |
| Interaction AN/protéines .....                                      | 14 |
| Le nucléosome.....  | 16 |
| Rôle du nucléosome .....  | 16 |
| Structure du nucléosome .....                                       | 18 |
| Dynamique de la chromatine <i>in vivo</i> .....                     | 26 |
| Assemblage et désassemblage du nucléosome <i>in vitro</i> .....     | 30 |
| Positionnement et spécificité du nucléosome.....                    | 32 |
| Les complexes acides nucléique-NCp7.....                            | 37 |
| Le virus de l'immunodéficience humaine (VIH) .....                  | 37 |
| Rôle de la NCp7 dans le cycle du VIH-1 .....                        | 38 |
| Propriétés structurales de la NCp7 .....                            | 45 |
| Spécificité de la NCp7.....   | 48 |
| Matériel et méthodes .....  | 49 |
| Analyse des interfaces par VLDM .....                               | 49 |
| Fonctionnement mathématique de VLDM : construction des pavages..... | 50 |
| Application de VLDM à l'analyse d'interfaces .....                  | 51 |
| Intérêt de VLDM par rapport à d'autres approches .....              | 52 |
| VLDM : Problématique des macromolécules .....                       | 53 |
| Dynamique moléculaire du nucléosome .....                           | 55 |
| Structures de départ.....   | 55 |
| Protocole de production des dynamiques moléculaires .....           | 56 |
| Analyse des acides nucléiques – Curves+.....                        | 58 |
| Conformation des phosphates .....                                   | 58 |
| Paramètres hélicoïdaux inter-paire de base.....                     | 58 |
| Sillons de l'ADN.....   | 59 |
| Méthode PhAST.....  | 61 |
| Chapitre 1 : Interface ADN/Histones dans le nucléosome.....         | 64 |

|  |     |
|--|-----|
| Sommaire .....   | 64  |
| Article 1 .....  | 66  |
| Annexes article 1 .....  | 80  |
| Conclusion et développements en cours .....                            | 97  |
| Chapitre 2 : Assemblage et désassemblage du nucléosome.....            | 100 |
| Sommaire .....   | 100 |
| Article 2.....   | 102 |
| Annexes article 2.....   | 124 |
| Conclusion et perspectives .....                                       | 128 |
| Chapitre 3 : Dynamique de l'ADN dans le nucléosome .....               | 129 |
| Sommaire .....   | 129 |
| Article 3.....   | 131 |
| Annexes article 3.....   | 154 |
| Conclusion et développements en cours .....                            | 163 |
| Chapitre 4 : Une autre interface AN/Protéine, les complexes NCp7 ..... | 165 |
| Sommaire .....   | 165 |
| Article 4.....   | 167 |
| Annexes article 4.....   | 182 |
| Conclusion.....  | 193 |
| Discussion et conclusion .....   | 194 |
| Le nucléosome.....   | 194 |
| Les complexes AN/NCp7.....   | 196 |
| Conclusion générale .....  | 197 |



# Liste des abréviations

ADN : acide désoxyribonucléique

ADNss : ADN simple brin

AN : acide nucléique

ARN : acide ribonucléique

ARNg : ARN génomique

ARNm : ARN messagers

ARNr : ARN ribosomaux

ARNt : ARN de transfert

DM : dynamique moléculaire

FRET : *Fluorescence Resonance Energy Transfert*

DRN : distance de répétition du nucléosome

MPT : modifications post-traductionnelles

NCP : *Nucleosome Core Particle*

NCp7 : nucléocapside p7

nt : nucléotide

PCH : protéine chaperon des histones

pb : paire de base

PDB : *Protein Data Bank*

RMN : résonance magnétique nucléaire

RRM : *RNA Recognition Motif*

Sida : syndrome d'immunodéficience acquise

SIT : site d'initiation de la transcription

VIH : virus de l'immunodéficience humaine

VLDM : Voronoi Laguerre Delaunay for Macromolecule

# Introduction

---

Les acides nucléiques (AN) et les protéines sont des macromolécules absolument cruciales pour la machinerie cellulaire des organismes vivants. D'une part, les acides nucléiques sont le support de l'information génétique. Ils permettent de déterminer l'identité biologique d'un organisme, de la préserver et de la transmettre. D'autre part, les protéines, présentes dans toutes les cellules vivantes, exécutent et régulent les fonctions vitales de l'organisme. Ces deux unités sont donc à la base du fonctionnement même du vivant. On comprend donc que leurs interactions jouent un rôle majeur dans l'équilibre qui caractérise le fonctionnement des cellules. Dans ma thèse, je me suis proposé de décrypter les mécanismes à l'origine des interactions non spécifiques AN-protéines au travers de deux exemples : le nucléosome et la nucléocapside p7 du virus du VIH-1.

# Interactions acide nucléique-protéine

---

## Contexte général

Classiquement, les interactions AN-protéine sont classifiés selon deux catégories : les interactions spécifiques et les interactions non spécifiques. Les interactions spécifiques AN-protéine sont, de manière générale, mieux caractérisées et impliquent la reconnaissance d'une séquence d'AN par un contact direct avec les bases, dont la succession bien définie le long des AN assure la spécificité (A. Travers 2001; Batabyal et al. 2014). Parmi les protéines impliquées dans ces interactions, on retrouve en particulier les facteurs de transcription (Latchman 1993) qui régulent la transcription de l'ADN double brin en ARN ou encore des protéines de liaison aux ARN qui jouent un rôle majeur dans le contrôle post-transcriptionnel (M.-H. Lee et Schedl 2006). Mais les complexes sur lesquels j'ai travaillé au cours de mes travaux impliquent le deuxième type d'interaction, les interactions non-spécifiques. Dans ce cas, le partenaire protéique est capable de se fixer à une large gamme d'AN sans que l'on puisse identifier de cible AN précise. Ces interactions privilégient les contacts avec le squelette phosphodiester. Dans les deux cas, spécifique comme non-spécifique, la formation d'un complexe demande que soit réalisée la complémentarité structurale entre les deux partenaires, un processus qui fait en général appel à la capacité des macromolécules à se déformer dans le sens requis par le complexe en minimisant l'énergie de déformation.

Bien que l'étude des complexes non spécifiques puisse se faire en utilisant des méthodes expérimentales variées (Iwahara, Schwieters, et Clore 2004; Iwahara, Zweckstetter, et Clore 2006; Jankowsky et Harris 2015), ou des méthodes de sédimentation à l'équilibre remises au goût du jour (Ucci et Cole 2004), les mécanismes associés aux complexes AN-protéines non spécifiques restent un challenge majeur en premier lieu en raison de leur faible affinité et de fort taux de dissociation qui génèrent la présence de plusieurs espèces difficiles à stabiliser ou à séparer et donc compliquées à étudier par les méthodes classiques de cristallographie aux rayons X et RMN (Ucci et Cole 2004). Mais nous verrons que les protéines non spécifiques ont en fait des préférences pour certaines séquences d'ADN et/ou de repliement, qui peuvent être exploitées dans les études expérimentales. Néanmoins, les limitations inhérentes aux complexes non spécifiques font de la modélisation un outil précieux pour les étudier.

## Spécificité et affinité

Le concept de spécificité est communément basé sur la constante de dissociation,  $K_d$ , qui correspond au rapport à l'équilibre entre la vitesse d'association  $k_{on}$  et la vitesse de dissociation  $k_{off}$  du complexe AN/protéine. Mais la spécificité des interactions n'est en fait pas un concept binaire, il existe toute une gamme de  $K_d$  et donc de spécificité (von Hippel et Berg 1986; Davidovich et al. 2015; Jankowsky et Harris 2015). Une spécificité dite absolue serait une protéine capable de reconnaître et de se fixer qu'à une et une seule cible d'acide nucléique. Au contraire, une protéine non spécifique devrait présenter une constante de dissociation identique quel que soit l'AN. Biologiquement, aucun de ces deux extrêmes n'existe. On distingue cependant des protéines avec une spécificité très forte pour une séquence ou un repliement d'AN très conservé et des protéines à faible spécificité, avec une reconnaissance comparable pour une grande variété de séquences d'AN ou de repliements. Parmi les protéines qui reconnaissent une large gamme de cibles, beaucoup ont néanmoins des préférences pour certaines séquences d'AN et/ou certains repliements (Ferré-D'Amaré 2016), comme par exemple le nucléosome et la NCp7 dont je parlerai en détail plus loin.

Le concept de l'affinité ( $K_d$ ) est étroitement relié à celui de la spécificité selon une étude basée sur la méthode SELEX et des travaux de bio-informatiques dédiées aux interactions protéine-protéine (Eaton, Gold, et Zichi 1995; Yan et al. 2013). Classiquement, une protéine qui présente une forte affinité pour une séquence va être spécifique à celle-ci et une reconnaissance non-spécifique présente une affinité moins importante. Cependant, une étude par simulation gros grains portant sur les interactions protéines/protéines montre que si on laisse plus de flexibilité aux protéines, la spécificité augmente mais l'affinité diminue (Chu et Wang 2014). Il serait très intéressant d'appliquer cette approche aux interactions AN/protéines, surtout dans le cas des complexes impliquant des structurations de la protéine.

## Structures des acides nucléiques

Les acides ribonucléiques (ARN) et désoxyribonucléiques (ADN) sont des polymères composés de nucléotides dont le pentose et les groupes phosphates forment la partie répétitive des nucléotides alors que les bases (purine et pyrimidine) présentent des compositions atomiques et donc des propriétés différentes (Figure 1). Ce sont ces bases qui sont au centre de la spécificité dans les interactions AN/protéine.

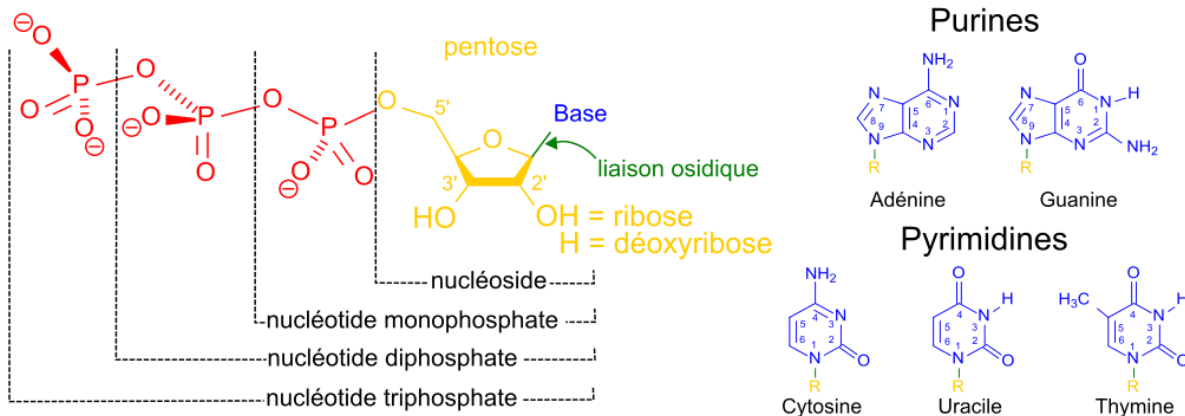
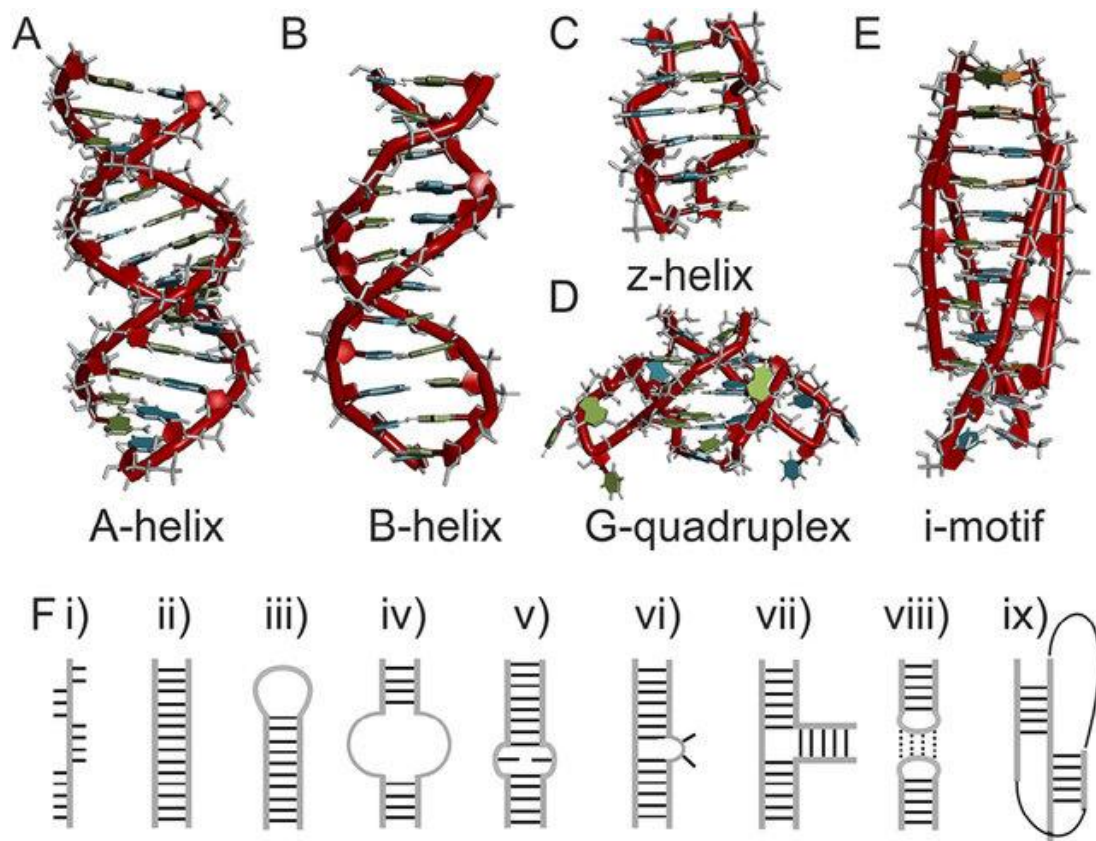


Figure 1 - Nucléotides

### Structure des ADN

La double hélice est la structure la plus courante de l'ADN dans les cellules, caractérisée par l'enroulement de deux brins ADN. Dans cette structure, les bases sont accessibles à la fois du côté des grand et petit sillons et les groupements phosphates sont très exposés au solvant. Il existe plusieurs formes de double hélices (Figure 2A, 2B, 2C), dont la plus répandue est la forme B de type « main droite ». La double hélice A est également de type « main droite » mais les centres des paires de bases sont rejetés loin de l'axe de la double hélice, ce qui modifie complètement les dimensions des sillons ; la forme Z est de type « main gauche ». Des études de diffraction aux rayons X mettent également en évidence d'autres formes de double hélice, dont la présence et les fonctions *in vivo* restent à déterminer, l'ADN C, D et E.

Bien que l'ADN existe majoritairement sous forme de double hélice, on le retrouve également sous la forme simple brin (ADNss). Dans les cellules eucaryotes notamment, des mécanismes de réplication, de recombinaison et de réparation nécessitent la formation d'un intermédiaire simple brin de l'ADN (Shereda et al. 2008) ainsi que chez certains virus, dont le génome est stocké sous forme d'ADN simple brin (Modrow et al. 2013). Sous cette forme, les composants de l'ADN, y compris les bases, sont généralement plus accessibles que l'ADN en double hélice. Cependant, l'ADN simple brin est capable de se replier en structures secondaires et tertiaires comme les épingles à cheveux ou les pseudo-nœuds (Figure 2F – i, iii et ix) qui impactent l'accessibilité des bases.



**Figure 2 – Exemple de structures adoptées par les acides nucléiques. A) double hélice A, B) double hélice B, C) double hélice Z, D) G-quadruplex et E) i-motif. F) Exemples de structures secondaires et tertiaire des AN : i) simple brin, ii) double hélice, iii) épingle à cheveux, iv) boucle interne, v) mésappariement, vi) renflement (de l'anglais bulge), vii) jonction à trois brins, viii) interaction dite « kissing loop » et ix) pseudo-nœud. (source : Börner et al. 2016)**

Plusieurs repliements mettent en jeu plus que deux brins d'ADN. L'intervention des mécanismes cellulaires sur l'ADN implique la formation d'une structure tertiaire impliquant les ADN en double hélice, les jonctions (Figure 2F - vii). La plus connue est la jonction entre quatre brins d'ADN B, dite jonction de Holliday, impliquée dans le mécanisme de recombinaison homologue. Les G-quadruplexes (Figure 2D) sont un repliement des séquences riches en guanines qui forment plusieurs plateaux de base empilés, eux-mêmes constitués par quatre guanines (Parkinson, Lee, et Neidle 2002). On les retrouve par exemple dans les télomères où ils jouent un rôle dans la réplication des extrémités du chromosome (Zahler et al. 1991). Enfin, le *i-motif* est un type d'ADN identifié en 1993 par Gehring, Leroy, et Guéron (Figure 2E), qui a été découvert récemment dans le noyau des cellules, plus précisément dans les régions télomériques et les promoteurs (Zeraati et al. 2018). Ce motif s'apparente à celui des G-quadruplex mais ici, il s'agit d'un empilement de plateaux formés par des cytosines.

## *Structures des ARN*

Les ARN, comme les ADN, sont capables d'adopter un large spectre de conformations. C'est la forme double-hélice A qui est la plus commune dans les ARN double brin (Salazar et al. 1993; Börner et al. 2016), comme par exemple chez certains virus qui stockent leur génome sous forme d'ARN double brin (Wickner 1996). La forme Z existe également (Herbert 2019) alors que la forme B n'est pas adaptée à l'ARN.

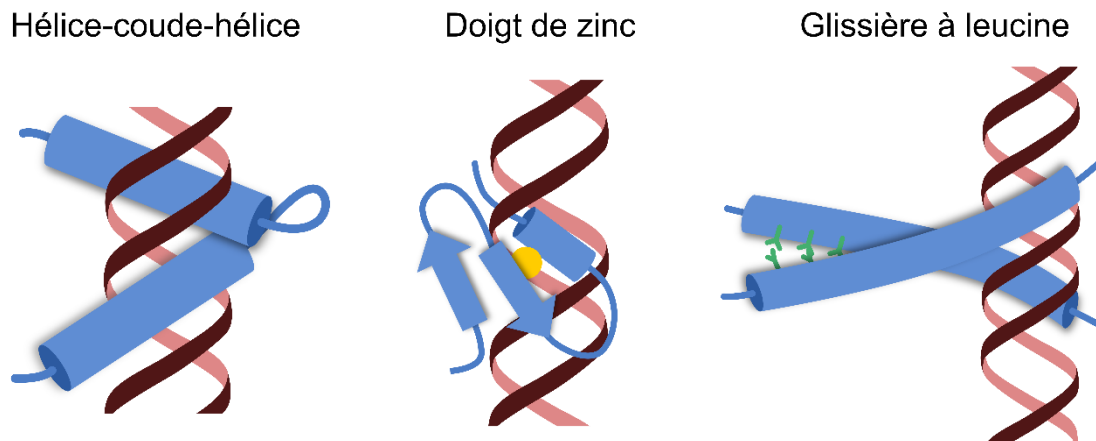
Dans la nature, c'est la forme simple-brin de l'ARN qui prédomine. Chez les eucaryotes, les rôles des ARN simple brin sont très divers et cela se traduit par l'existence de nombreux types d'ARN. Parmi les plus connus, les ARN messagers (ARNm) sont impliqués dans le transport de l'information vers les ribosomes alors que les ARN de transfert (ARNt) et les ARN ribosomiaux (ARNr) permettent la traduction en protéine. Pour jouer ces différents rôles, les ARN sont repliés en structures secondaires, comme les épingles à cheveux (Figure 2F – iii) et les tiges-boucles, et/ou en structures tertiaires (Figure 2F – ix).

Comme pour les ADN, les ARN sont capables de prendre les conformations G-quadruplex et *i-motif*.

## Principaux motifs protéiques de liaisons aux AN

Les domaines de liaison aux AN sont des motifs plus ou moins bien définis qui ont la capacité de se lier avec des AN, sous les différentes formes que ceux-ci adoptent. Ces domaines de liaison qui sont capables de reconnaître une séquence d'AN bien spécifique mais, en général, ont également une bonne affinité pour un spectre plus ou moins large de séquences (Gao et Skolnick 2009). Je décris ici certaines des structures des domaines de liaison aux AN les plus répandus.

### *Domaines de liaison aux ADN*



*Figure 3 – Représentation schématique de trois domaines de liaisons à l'ADN. L'ADN est représenté par une double hélice rouge et noire. Les motifs protéiques sont représentés en bleu, leurs hélices  $\alpha$  sont représentés par des cylindres et leurs feuilletts  $\beta$  par des flèches. Pour le motif doigt de zinc (milieu) l'atome de zinc est représenté en jaune. Pour le motif glissière à leucine (droite), les leucines sont représentées en vert le long de l'hélice.*

### Hélice-coude-hélice

Le motif hélice-coude-hélice est caractérisé par deux hélices  $\alpha$  qui interagissent avec le grand sillon de l'ADN (Figure 3). Bien que ce motif puisse être aussi impliqué dans des interactions non-spécifiques (Percipalle et al. 1995), il reste majoritairement spécifique en raison de l'interaction entre l'hélice C-terminale et les bases du grand sillon de l'ADN (Wintjens et Rooman 1996).

### Doigt de zinc

Il s'agit d'une famille très diverse, caractérisée par un ou plusieurs domaines qui coordonnent un ou plusieurs ions zinc ( $Zn^{2+}$ ), stabilisant leur repliement (Figure 3). C'est notamment le motif le plus courant dans les facteurs de transcription et il est très utilisé en biotechnologie pour développer des reconnaissances très spécifiques. Il existe néanmoins des doigts de zinc non spécifiques (Vandevienne et al. 2013).



### Glissière à leucine

La glissière à leucine est caractérisée par deux longues hélices  $\alpha$  de plusieurs dizaines de résidus, dont les parties C-terminales sont constituées de répétitions de leucines qui permettent la formation d'un homo ou hétéro-dimère (Figure 3); Les parties N-terminales se structurent en hélices  $\alpha$  en interagissant avec le grand sillon de l'ADN. Comme les motifs hélice-tour-hélices, ses interactions avec l'ADN sont majoritairement spécifiques.

### La boîte HMG pour "High Mobility Group box"

La boîte HMG est un domaine protéique caractérisé par trois hélices  $\alpha$  séparées par des boucles. Ces domaines sont retrouvés par paire dans les protéines des groupes HMG1 et HMG2 qui se fixent sur les ADN de manière non-spécifique. La formation d'une interface HMG-box/ADN double hélice impose généralement une déformation importante de l'ADN (J. O. Thomas 2001).

### Domaine « OB-fold »

Le domaine OB-fold (*Oligonucleotide/oligosaccharide-Binding fold*) est caractérisé par un motif de taille variable (entre 70 et 150 acide aminés) peu conservé, composé de tonneaux  $\beta$  formé par 5 feuillettes anti-parallèles et couvert d'une hélice  $\alpha$  sur une extrémité (Murzin 1993). On retrouve beaucoup ce domaine chez les protéines de fixation aux ADN simple brin.

## *Domaines de liaison aux ARN*

### *Motif de reconnaissance de l'ARN*

Les motifs de reconnaissance de l'ARN (« RRM » pour *RNA Recognition Motif*) sont des domaines d'environ 90 acides aminés, composés de quatre brins  $\beta$  antiparallèles et de deux hélices  $\alpha$ , qui se fixent sur les ARN simple brin. Très abondant chez les eucaryotes, ces motifs divers sont les plus communs dans les interactions ARN-protéine (Maris, Dominguez, et Allain 2005) ; ils jouent de nombreux rôles qui impliquent d'interagir avec une grande variété de partenaires ARN (Kenan, Query, et Keene 1991). Les motifs de reconnaissance de l'ARN peuvent être spécifiques ou non-spécifiques (Chaudhury, Chander, et Howe 2010).

### *Motif de liaison à l'ARN double brin*

Les motifs de liaison à l'ARN double brin (« dsRBD » pour *double-stranded RNA Binding Domain*) sont caractérisés par trois feuilletts  $\beta$  antiparallèles encadrés par deux hélices  $\alpha$ . Ce motif joue un rôle critique dans les modifications post-transcriptionnelles, la localisation de l'ARN et l'interférence par ARN (Saunders et Barber 2003; Doyle et Jantsch 2002).

### *Doigts de zinc*

Comme pour les ADN, les doigts de zinc sont capables de reconnaître et de se fixer sur les ARN. Il s'agit donc d'un motif protéique capable d'interagir avec une grande diversité de cibles AN, de manière spécifique ou non.

### *Domaine « OB-fold »*

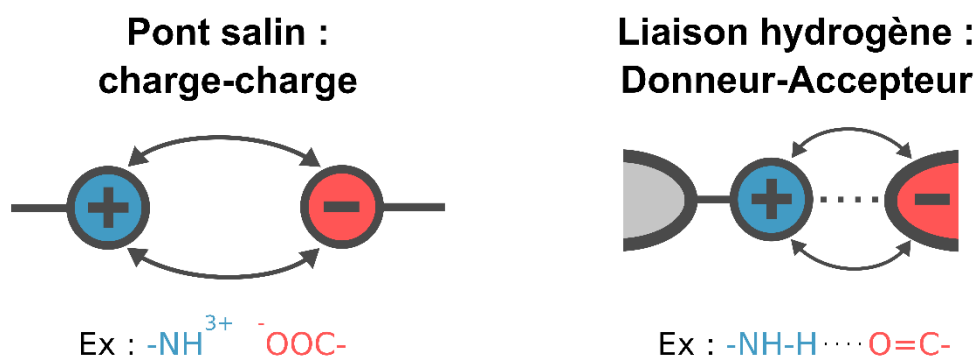
La forte variabilité des domaines *OB-fold* leur permet de se fixer sur les ARN simple brin en plus des ADN (Theobald, Mitton-Fry, et Wuttke 2003).

## Types d'interactions

Les interactions AN/protéine sont très majoritairement des associations transitoires non-covalentes, c'est-à-dire réversibles (Smith et al. 2009). Les protéines interagissent avec les ADN et les ARN avec les mêmes forces physiques. Une analyse statistique de 188 structures de complexes AN/protéines spécifiques et non-spécifiques résolues par cristallographie aux rayons X a permis de mettre en évidence la contribution des différents types d'interactions dans ces interfaces (Lejeune et al. 2005). Pour les complexes contenant ADN et ARN, un simple comptage des interactions montre que les plus fréquentes sont les liaisons hydrogène (~50%), suivi par les interactions de van der Waals et hydrophobes et enfin électrostatiques. Une étude précédente portant sur 129 complexes ADN/protéine soulignait néanmoins que 65% des interactions étaient du type van der Waals (Luscombe, Laskowski, et Thornton 2001). Ces résultats contradictoires tiennent très probablement au fait que les auteurs choisissent différemment les critères qui vont faire que deux atomes interagissent, illustrant ainsi la difficulté qu'il y a à analyser objectivement les interfaces AN/protéines. Je définis ici ces différents types d'interaction.

### Contacts électrostatiques

Les contacts électrostatiques contribuent à de nombreux phénomènes biologiques ; ils sont réputés être particulièrement importants dans les mécanismes de reconnaissance et de complexation de macromolécules. Les contacts électrostatiques peuvent advenir entre charges entières pour donner les liaisons hydrogène et les ponts salins (Figure 4). Plus largement, ils existent aussi entre des charges partielles essentiellement portées par les oxygènes et les azotes.

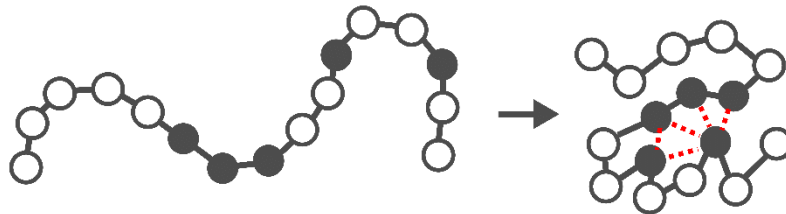


*Figure 4 – Deux types de contacts électrostatiques.*

Pour les nucléotides, ces interactions peuvent intervenir au niveau du groupement phosphate et des bases azotées (Luscombe, Laskowski, et Thornton 2001). Du côté protéique, les liaisons hydrogène peuvent intervenir tout le long du squelette, mais également avec les chaînes latérales des résidus polaires. Les charges portées par les chaînes latérales de certains acides aminés représentent la majorité des contacts électrostatiques entre charges entières.

### *Contacts Hydrophobes*

L'eau est une molécule polaire qui favorise les contacts avec d'autres composés polaires. Ainsi, les composés apolaires se regroupent entre eux (Figure 5) grâce à des contacts appelés hydrophobes. Ces contacts hydrophobes interviennent dans de nombreux phénomènes biologiques (Djikaev et Ruckenstein 2016), notamment le repliement, la structuration et la déstabilisation des protéines (Pace et al. 2011; Bjelić et al. 2008; Silva, Figueiredo, et Cabrita 2014).



**Figure 5 – Représentation schématique des contacts hydrophobes.** Les résidus hydrophiles sont représentés par des cercles blancs et les résidus hydrophobes par des cercles gris. Les contacts hydrophobes sont représentés par des pointillés rouges.

### *Interactions de van der Waals*

Les interactions, ou forces de van der Waals sont des interactions de faible intensité entre deux atomes ou molécules. Elles sont très présentes dans toutes interactions puisqu'elles interviennent dès que deux atomes ou molécules sont proches mais diminuent très vite avec la distance. Elles deviennent également très répulsives lorsque les deux nuages électroniques de atomes s'approchent trop. Ces interactions jouent un rôle important dans la reconnaissance spécifique de forme.

### *Interactions d'empilement aromatique*

Les interactions d'empilement aromatique (ou empilement  $\pi$ - $\pi$ ) se produisent entre les nuages  $\pi$  de deux cycles aromatiques, en particulier entre les paires de bases des AN en double hélice mais elles interviennent aussi dans la stabilité des protéines (Burley et Petsko 1985; Matta, Castillo, et Boyd 2006). Au niveau des complexes AN/protéine, ces interactions peuvent intervenir entre les bases des nucléotides et les acides aminés aromatiques.

## Interaction AN/protéines

### *Interactions ADN/protéines*

Dans le cas de l'ADN double brin, l'étude des domaines de liaison aux ADN présentés précédemment met en évidence que la plupart des interfaces spécifiques se situent au niveau du grand sillon (Domaines de liaison aux ADN, p. 8). Il a en effet été établi il y a de nombreuses années que, dans le grand sillon, les paires de base A:T et G:C présentent un schéma de liaison hydrogène différent, ce qui n'est pas le cas dans le petit sillon (Seeman, Rosenberg, et Rich 1976). Pour le petit sillon, la spécificité ne passe donc pas par la formation de liaison hydrogène. Néanmoins, les dimensions du petit sillon et la séquence des ADN sont liées (Oguey, Foloppe, et Hartmann 2010; X. Xu et al. 2014). A ceci se rajoute que le potentiel électrostatique du petit sillon est à la fois modulé par sa forme et la séquence d'ADN, favorisant en particulier l'insertion de résidus protéiques basiques dans les séquences A:T riches (Rohs et al. 2009). Dans cette logique, l'étude de structures cristallographiques ADN/protéine non-spécifiques indique une prédominance des contacts électrostatiques entre résidus protéiques et phosphate, qui ne portent pas de marque distinctive de la séquence comme les bases le font (von Hippel et Berg 1986; Jones et al. 1999; Nadassy, Wodak, et Janin 1999).

En plus des caractéristiques électrostatiques que je viens de présenter, la déformabilité de l'ADN double brin, c'est-à-dire la capacité de varier autour d'une double hélice idéale d'ADN B, joue un rôle dans la reconnaissance de séquence spécifique (Fujii et al. 2007; Rohs et al. 2010), comme il a été montré au laboratoire sur de nombreux systèmes impliquant des facteurs de transcription très variés (NF- $\kappa$ B (Tisné, Delepierre, et Hartmann 1999), Jun-Fos (Heddi et al. 2008), E2BS (Djuranovic, Oguey, et Hartmann 2004), opérateur du bactériophage 434 (Hartmann, Sullivan, et Harris 2003), Ets1 (Leprivier et al. 2009, 51)). Plus généralement, la formation d'un complexe demande que soit réalisée la complémentarité structurale entre deux partenaires, un processus qui fait appel à la capacité des macromolécules à se déformer dans le sens requis par le complexe en minimisant l'énergie de déformation, et a été illustrée au laboratoire par l'exemple de la DNase I (Heddi, Abi-Ghanem, et al. 2010). La déformabilité de l'ADN s'observe en modélisation au niveau de la paire de base (Lankaš et al. 2004) et au niveau des dinucléotides (Lankas 2004). Expérimentalement, en solution, la flexibilité des dinucléotides concerne aussi bien les conformations des groupement phosphates que les positions relatives de deux plateaux de bases successifs (Djuranovic et Hartmann 2004; Heddi et al. 2008; Heddi, Oguey, et al. 2010; Oguey, Foloppe, et Hartmann 2010; X. Xu et al. 2014). Ainsi, la succession de dinucléotides plus ou moins flexibles est un élément déterminant dans les mécanismes de reconnaissance spécifique de certaines séquences.

## *Interactions ARN/protéine*

La plupart des interactions ARN/protéine connues se font avec de l'ARN simple brin, dont les bases sont très accessibles et offrent de bons sites de fixation (Morozova et al. 2006). La spécificité des complexes ARN/protéine est portée en particulier par des empilements aromatiques favorisés par l'accessibilité accrue des bases ; dans la moitié des interfaces comprenant une phénylalanine, cette dernière est impliquée dans un empilement (Morozova et al. 2006).

Du côté protéine, la plupart des études sur les interactions ARN/protéine se sont concentrées sur le motif de reconnaissance spécifique « RRM ». Il semble que comme pour les ADN, ce soit ces mêmes domaines qui sont à l'origine des interactions non spécifiques. Une étude montre même très clairement, sur l'exemple des protéines C5 de *E. coli*, que l'affinité de ces protéines pour ses cibles est variable, avec une très bonne affinité pour quelques séquences spécifiques et une affinité moyenne pour un grand nombre de séquences ARN (Lin et al. 2016).

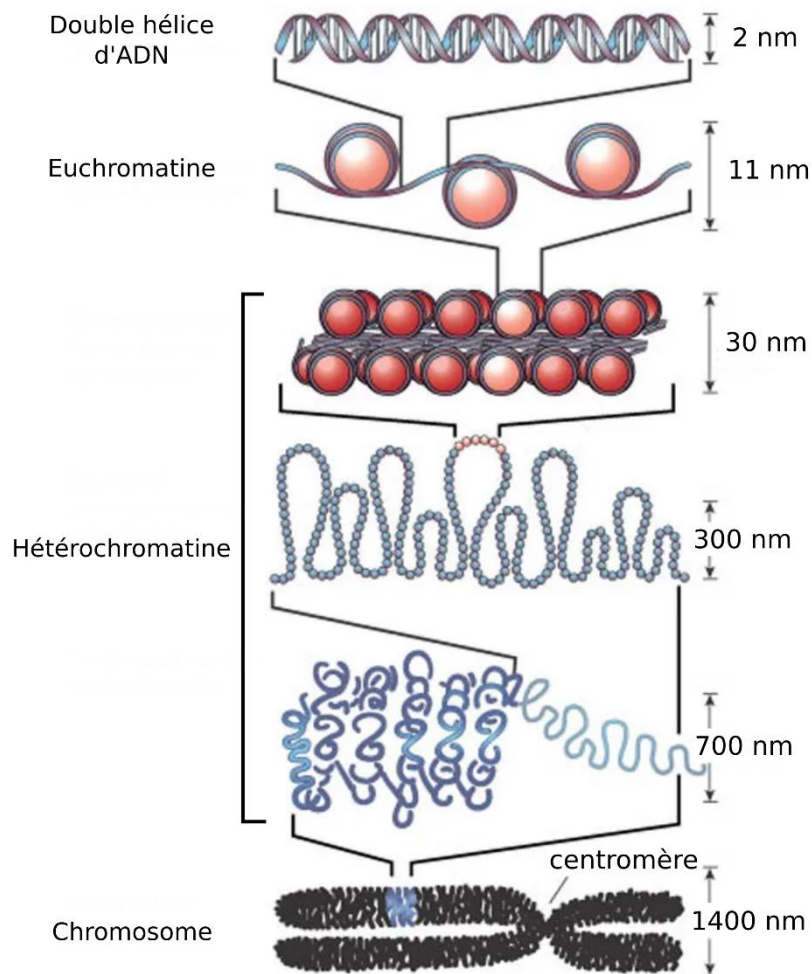
## *Stratégie des protéines dans les interactions AN/protéine*

La spécificité des complexes AN/protéines se situe donc essentiellement au niveau des bases par des liaisons hydrogène et des empilements aromatiques (Jeong et al. 2003; Kalodimos et al. 2004; Allers et Shamoo 2001; Morozova et al. 2006; Yesudhas et al. 2017). Pour obtenir des complexes AN/protéines moins spécifiques, il faut diminuer les contacts avec les bases et favoriser les contacts avec les phosphates. Pour cela, deux stratégies principales vont être adoptées. En fait, les domaines de liaisons spécifiques aux ADN ou aux ARN présentent souvent une affinité moyenne pour de nombreuses autres séquences que leur cible de haute affinité. En raison du large paysage conformationnel que peuvent adopter les AN, il est même difficile pour une protéine d'être complètement non-spécifique (Jankowsky et Harris 2015). Réguler la spécificité s'effectue soit en agrandissant ou en diminuant la taille du site de reconnaissance soit d'autre part en multipliant les sites de liaisons au sein d'une même protéine (Jankowsky et Harris 2015).

# Le nucléosome

## Rôle du nucléosome

Le nucléosome est l'unité de base de la compaction de l'ADN chez les eucaryotes et joue un rôle important dans les autres niveaux de compaction via les interactions nucléosome-nucléosome. Ce rôle structural est au centre de nombreux mécanismes cellulaires puisque l'état de compaction de l'ADN va déterminer son accessibilité et son volume.



**Figure 6 – Schéma des différentes étapes de la compaction de l'ADN chez les cellules eucaryotes. Les nucléosomes sont représentés par des sphères rouges, l'ADN en bleu (d'après Felsenfeld et Groudine 2003).**



L'ADN enroulé autour des histones H3, H4, H2A et H2B forme un collier de perles, l'euchromatine (Figure 6), premier stade de compaction. L'ajout des histones H1 stabilise ensuite les nucléosomes sous la forme d'une fibre de 30 nm, avec l'aide des autres histones qui interagissent entre elles. Les étapes suivantes nécessitent l'intervention de protéines d'échafaudages jusqu'au chromosome mitotique. La structure de la chromatine chez les eucaryotes est majoritairement liée au positionnement des nucléosomes le long des génomes (Woodcock et Ghosh 2010). De nombreux phénomènes impactent le positionnement des nucléosomes qui est dynamique et modulent donc l'accessibilité de l'ADN.

La compaction de l'ADN intervient dans la plupart des mécanismes cellulaires se produisant au sein du noyau, tels que la régulation de l'expression génique. Il existe ainsi de nombreux exemples montrant l'effet de l'hétérochromatine sur l'expression génique, comme la coloration de l'œil des *Drosophila melanogaster* (Zhimulev et al. 1986; Vogel et al. 2009) ou l'évolution de l'expression du gène de la  $\beta$ -globine lors du développement embryonnaire chez les eucaryotes (Kiefer et al. 2008). Ainsi, en s'assemblant et se désassemblant le long du génome, les nucléosomes sont directement associés à l'expression de certaines régions du génome.

## Structure du nucléosome

Dans ce manuscrit, le terme « nucléosome » (ou « NCP » pour l'anglais *Nucleosome Core Particle*) désigne le complexe formé d'un ADN de 145 à 147 paires de bases enroulé autour d'un cœur protéique de huit histones (deux exemplaires des histones H3, H4, H2A et H2B).

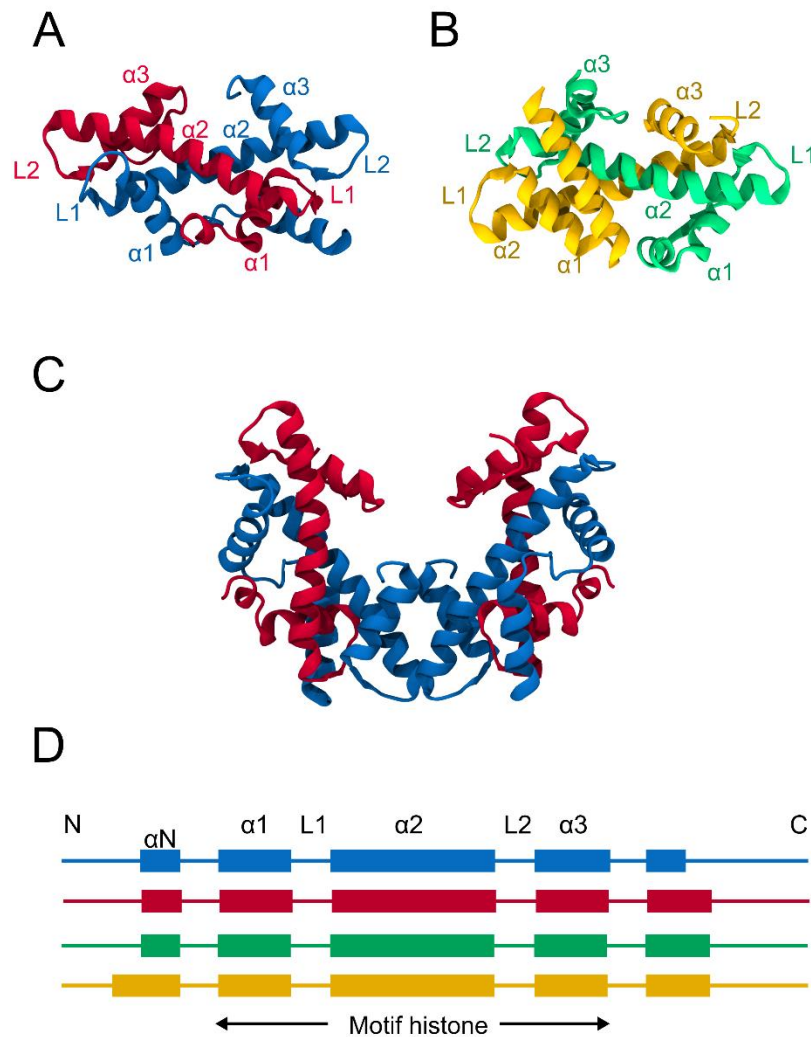
C'est en 1984 que la première structure cristallographique du nucléosome a été résolue à 7Å (T. J. Richmond et al. 1984). Mettant en évidence une structure cylindrique, déjà suggérée par des études de microscopie électronique (Finch et al. 1977), c'était une première étape dans la compréhension de cette organisation. Aujourd'hui, de nombreuses structures du nucléosome à haute résolution (K. Luger et al. 1997; Davey et al. 2002) permettent d'obtenir une description détaillée de son agencement au niveau atomique. Sur la *Protein Data Bank* (PDB), on peut trouver 69 structures du nucléosome avec une résolution inférieure à 3Å sur un total de plus de 120 structures.

Dans ces études, on trouve seulement quatre séquences d'ADN. Les séquences mixtes d'ADN génomique ont été utilisées dans les toutes premières études structurales et ont permis l'obtention de la première structure à basse résolution, à 7Å, du nucléosome (T. J. Richmond et al. 1984). L'utilisation de la séquence ADN codant pour l'ARN ribosomal 5S (ADN 5S) a permis quelques années plus tard d'obtenir une structure à 4.5Å (T. J. Richmond, Searles, et Simpson 1988). C'est avec l'utilisation de la séquence issue de l' $\alpha$ -satellite humain qu'une première structure haute résolution, à 2.8Å (K. Luger et al. 1997) a permis de décrire le nucléosome à l'échelle atomique. Enfin, la séquence 601 (Thåström et al. 1999) est une séquence artificielle identifiée par des expériences SELEX pour sa très forte affinité pour les histones. Cette séquence est la plus populaire pour les études biochimiques *in vivo* et *in vitro* en raison de ses propriétés positionnantes.

### Structure des histones

Le cœur d'histone du nucléosome est composé de huit histones, le tétramère (H3-H4)<sub>2</sub> et deux dimères H2A-H2B, formant deux moitiés symétriques autour d'un axe qualifié de dyadique (K. Luger et al. 1997), que je vais décrire un peu plus loin (p. 21). En plus des histones canoniques, il existe de nombreux variants d'histones essentiels pour répondre aux différentes fonctions de la cellule et aux *stimuli* environnementaux (Talbert et Henikoff 2014). L'ensemble des séquences des histones canoniques et de leur variants ont été regroupées récemment dans la base de donnée HistoneDB 2.0 (Draizen et al. 2016) qui permet de confirmer que malgré les différences, les séquences et les structures des histones H3, H4, H2B, H2A et leur variants restent très conservés (Baxevanis et Landsman 1996; Draizen et al. 2016).

Au niveau de leur structure, les histones sont caractérisées par un motif récurrent, composé de trois hélices  $\alpha$  ( $\alpha 1$ ,  $\alpha 2$  et  $\alpha 3$ ) et de deux boucles (L1 et L2) pour former le motifs  $\alpha 1$ -L1-  $\alpha 2$ -L2-  $\alpha 3$  (Figure 7-D). Ce motif leur permet de s'associer en dimère selon une conformation dite « en poignée de main », maintenue par des interactions hélice  $\alpha 2$  – hélice  $\alpha 2$ , L1-L2 et hélice  $\alpha 1$ - hélice  $\alpha 1$  (Figure 7-AB). Les deux dimères H3-H4 forment un tétramère en interagissant au niveau des hélices  $\alpha 2$  et  $\alpha 3$  des histones H3 (Figure 7-C).

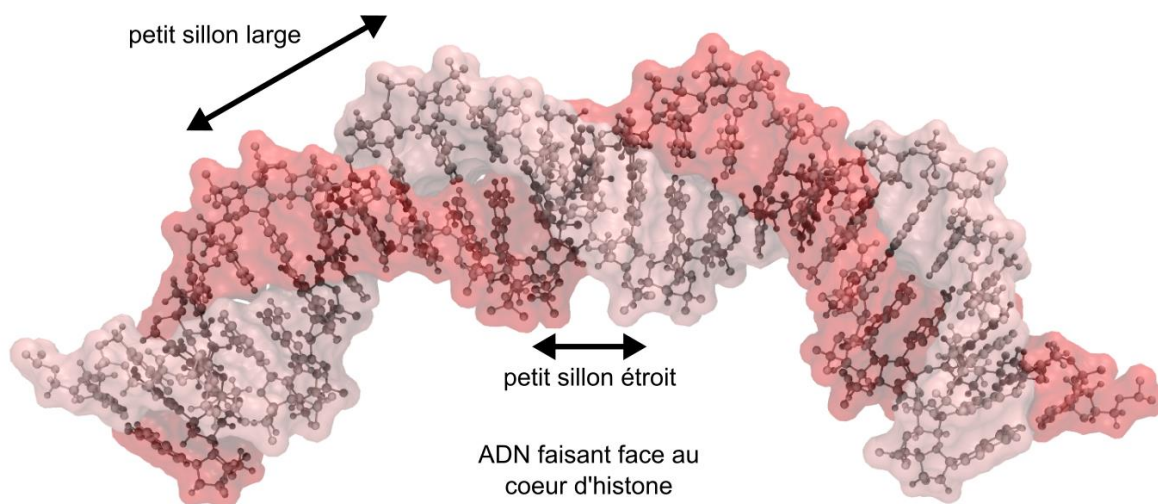


**Figure 7 – Structures secondaires et tertiaires des histones.** A) Structure tridimensionnelle du dimère H3–H4, B) du dimère H2A-H2B et C) du tétramère (H3-H4)<sub>2</sub> formés par les histones H3 (bleu), H4 (rouge), H2A (vert) et H2B (jaune). D) Structures secondaires correspondantes.

Une autre caractéristique des histones réside dans des domaines non structurés, appelés queues. Les queues N-terminales des histones H3, H4 et H2B et la queue C-terminale de l'histone H2A peuvent subir des modifications post-transcriptionnelles importantes (Erler et al. 2014). Elles représentent jusqu'à 30% de la masse des histones mais en raison de leur flexibilité, ne sont pas résolues dans les structures cristallographiques à part dans un système (code PDB 1KX5) où leur conformation est en fait stabilisée grâce à des interactions nucléosomes-nucléosomes intermoléculaires (Zheng et Hayes 2003).

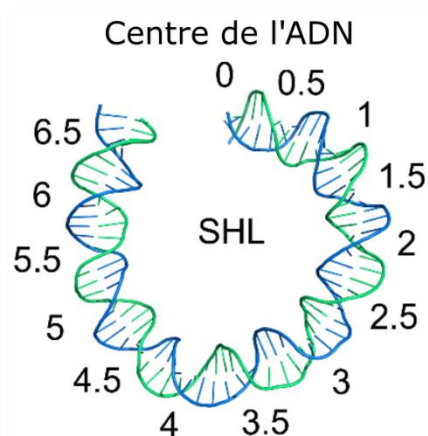
### *Structure de l'ADN nucléosomal*

L'ADN nucléosomal est un ADN double brin de 145 et 147 paires de bases présentant une courbure importante d'environ 30 degrés pour un tour d'hélice (10 pb) (Timothy J. Richmond et Davey 2003). Comme le montrent les études cristallographiques (F. Xu et Olson 2010; Zhurkin, Lysov, et Ivanov 1979) et d'autres analyses structurales (Djuranovic et Hartmann 2004; Heddi et al. 2008; Ben Imeddourene et al. 2015) cette courbure régulière de l'ADN est associée à une variation périodique d'également ~10 pb des paramètres hélicoïdaux décrivant la structure locale (Paramètres hélicoïdaux inter-paire de base, p.58) et à un niveau semi-local, les dimensions des sillons. Dans les régions où le petit sillon fait face aux histones, on observe des *rolls* négatifs, des *slides* positifs et un petit sillon très étroit. Au contraire, lorsque le petit sillon est dirigé vers l'extérieur, les *rolls* sont positifs, les *slides* négatifs et le petit sillon très large (Figure 8). L'annotation SHL (p. 21) permet de localiser une région de l'ADN nucléosomal et elle est liée à cette périodicité.



**Figure 8 – Courbure de l'ADN nucléosomal.** Les brins de l'ADN sont représentés en rouge et en rose.

## L'annotation « Super Helical Location » de l'ADN nucléosomal

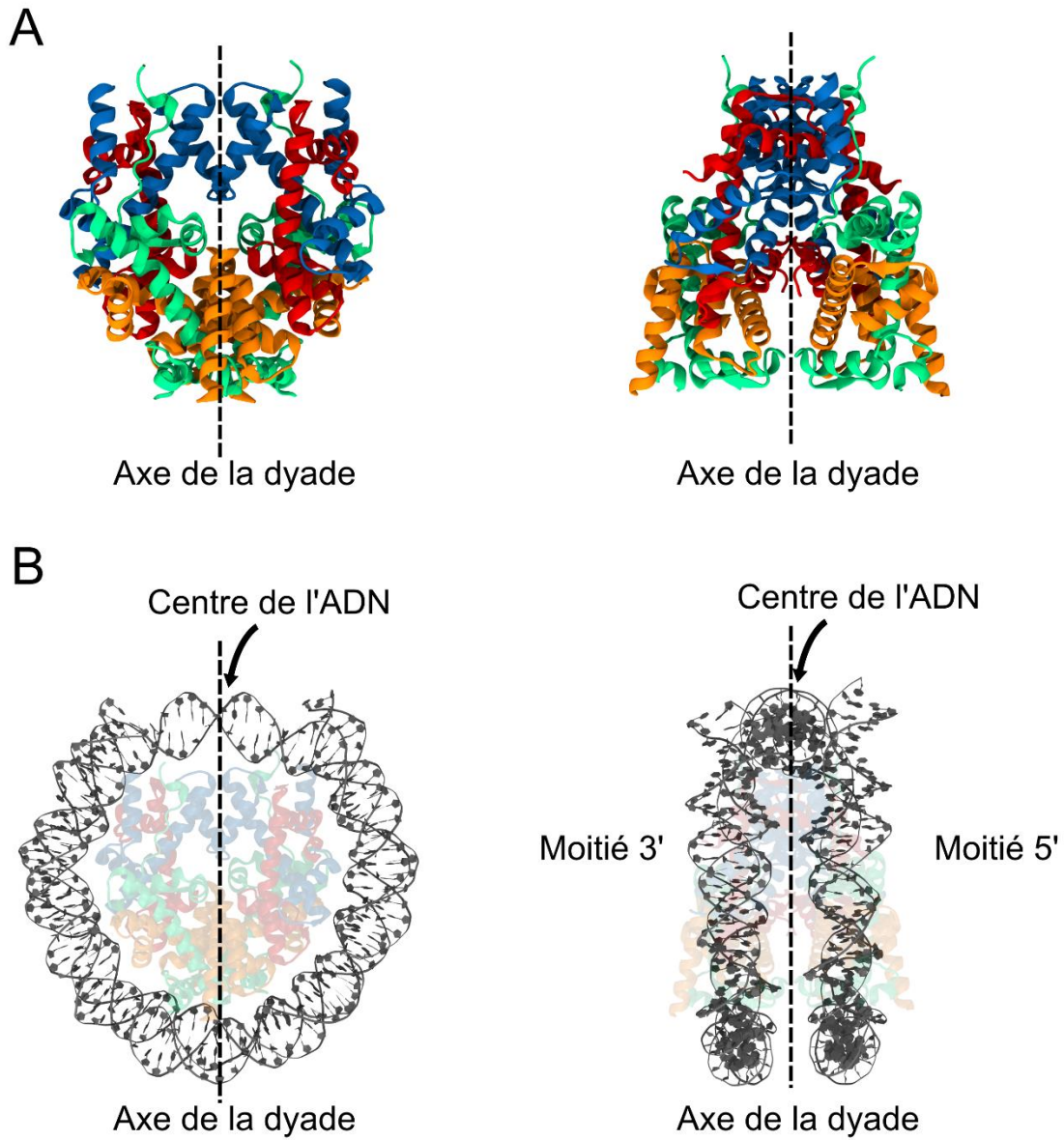


**Figure 9 – Positions SHL sur la moitié 3' de l'ADN nucléosomal.**

Le terme de « SHL » est communément utilisé comme unité de positionnement le long de l'ADN nucléosomal. On note SHL 0 la position centrale de l'ADN nucléosomal. Dans le cadre de la séquence 601 qui comporte 146 pb (comme celle que j'ai utilisé dans mes travaux), SHL 0 correspond au nucléotide 73. On numérote ensuite les positions toutes les 10 pb SHL -1, -2, ... -6 vers la moitié 5' de l'ADN et SHL +1, +2, ..., +6 vers la moitié 3' (Figure 9). Les SHL peuvent également ne pas correspondre à un chiffre entier. Par exemple, SHL +1.5 signifie que l'on parle d'une pb située dans la moitié 3' (le signe +) à 15 nucléotides du center (nt 73).

## La symétrie du nucléosome

Le nucléosome présente une symétrie par rapport à l'axe de la dyade. Cette symétrie est d'abord liée à l'organisation des histones (Figure 10A). Mais elle implique que l'ADN nucléosomal peut être séparé en deux moitiés, que l'on dénommera moitiés 5' et 3'. La moitié 5' correspond aux paires de base 1 à 73, et aux quatre histones H3, H4, H2A et H2B qui sont en interaction avec ces bases alors que la moitié 3' correspond aux bases restantes (73 à 146) en interaction avec les quatre autres copies d'histones (Figure 10B).



**Figure 10 – Symétrie du nucléosome selon l'axe de la dyade.** *A) Structure du cœur d'histone vue de face et de côté, composée des histones H3 (bleu), H4 (rouge), H2A (vert) et H2B (orange). B) Structure du nucléosome vue de face et de côté. L'ADN est représenté en gris et le cœur d'histone en transparence.*

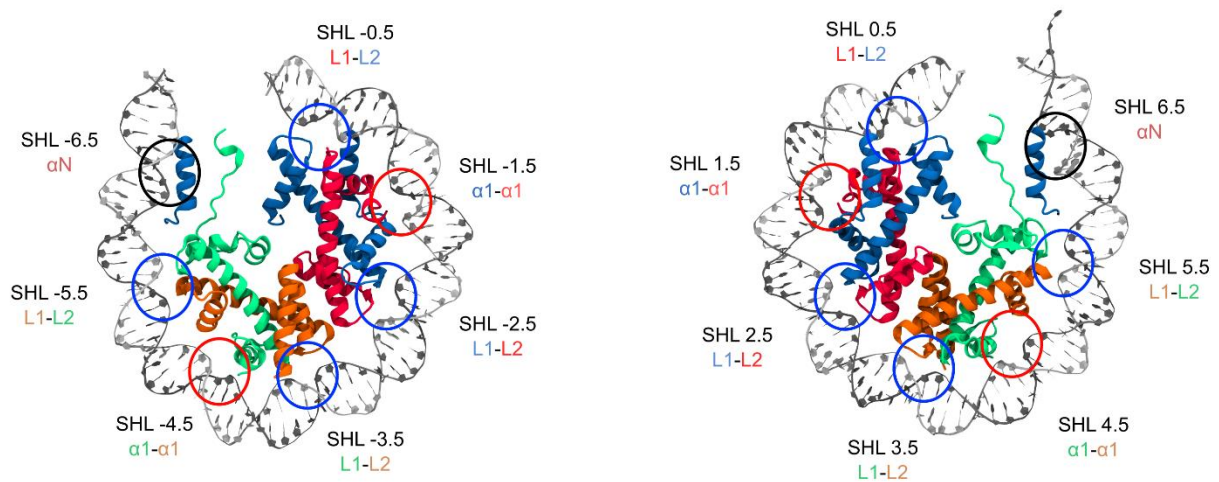


## *Interface ADN/histones*

Nos connaissances du complexe et donc de l'interface ADN/histone proviennent essentiellement de structures cristallographiques de nucléosome. Comme nous l'avons vu, les queues ne sont pas résolues dans ces structures sauf dans la structure dont le code pdb est 1KX5 (Davey et al. 2002). Mais au dire même des auteurs de cette structure, la résolution des queues est malgré tout trop mauvaise pour pouvoir en tirer quelque chose de sûr. Les descriptions de l'interface dont nous disposons concernent donc les interactions entre ADN et domaines structurés des histones.

L'interface a été décrite comme étant caractérisée par de nombreux contacts électrostatiques dus à la présence d'un grand nombre d'acides aminés basiques (ARG et LYS) et polaires qui interagissent avec les groupements phosphate de l'ADN (Mariño-Ramírez et al. 2005). L'analyse de la structure cristallographique à haute résolution (1.9 Å) du nucléosome 1KX5 contenant un ADN dérivé de l' $\alpha$ -satellite humain suggère que de nombreuses liaisons hydrogène additionnelles sont médiées par des molécules d'eau (Davey et al. 2002). Cette même étude propose également que le nombre de liaisons hydrogène serait plus faible pour les régions contactées par les dimères H2A-H2B que pour les régions contactées par le tétramère (H3-H4)<sub>2</sub>.

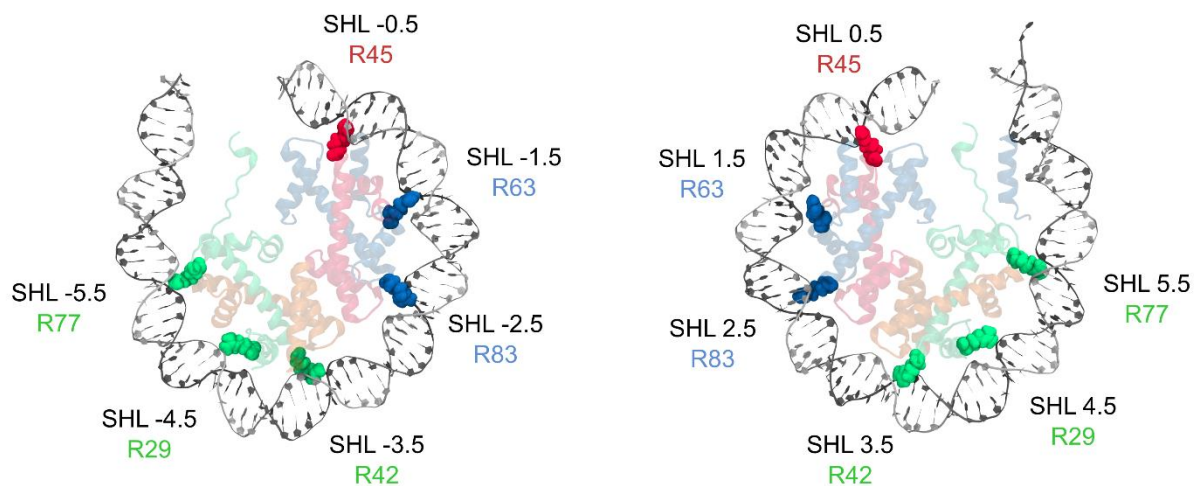
L'interface entre l'ADN et le cœur structuré des histones se caractérise par quatorze sites d'interaction (K. Luger et al. 1997; Davey et al. 2002), situés aux SHL +/- 6.5, 5.5, 4.5, 3.5, 2.5, 1.5 et 0.5. Comme l'organisation des histones est parfaitement symétrique, les sites de fixation impliquant les moitiés 5' (SHL 6.5 → 0.5) et 3' (SHL -6.5 → -0.5) de l'ADN nucléosomal sont symétriques par rapport à l'axe de la dyade. Les sites d'interactions aux SHL +/- 0.5, 1.5, et 2.5 interagissent avec le tétramère (H3-H4)<sub>2</sub> et les sites aux SHL +/- 3.5, 4.5 et 5.5 avec les dimères H2A-H2B (Figure 11). Le site SHL +/- 6.5, proche des extrémités de l'ADN, interagit avec l'hélice  $\alpha$  du côté N-terminal ( $\alpha$ N) de l'histone H3. On distingue trois types d'interfaces ADN/histones, selon que la partie protéique soit une hélice  $\alpha$  ou une boucle : L1-L2 aux SHL +/- 0.5, 2.5, 3.5 et 5.5,  $\alpha$ 1-  $\alpha$ 1 aux SHL +/- 1.5 et 4.5 et  $\alpha$ N aux SHL +/- 6.5 (Figure 11).



**Figure 11 – Sites de liaisons ADN/histone.** Les sites de liaison de type L1-L2 sont entourés en bleu, les sites de liaison  $\alpha 1$ -  $\alpha 1$  en rouge et le site de liaison  $\alpha N$  en noir. L'ADN est représenté en blanc, H3 en bleu, H4 en rouge, H2A en vert et H2B en orange. La moitié 5' du nucléosome est représentée à gauche et la moitié 3' à droite.

Chacun de ces sites est caractérisé par la présence d'une arginine en interaction avec le petit sillon (Davey et al. 2002) (Figure 12). Les sites TTAAA aux SHL +/- 1.5 de la séquence 601 ont plus particulièrement attiré l'attention. Il a été observé à ces sites la formation d'un réseau appelé « sugar clamp » où on met en évidence le rôle important de contacts hydrophobes impliquant des résidus leucine et proline qui viennent compléter les contacts électrostatiques de l'arginine (Wu et al. 2010). Le calcul des champs électrostatiques associés à l'ADN nucléosomal a permis de mettre en évidence qu'un petit sillon à la fois riche en A:T et particulièrement étroit serait très électronégatif et donc favorable à l'insertion de la chaîne latérale de l'arginine (Rohs et al. 2009; West et al. 2010). Cette hypothèse séduisante est malheureusement contredite par mes propres analyses, qui, comme on va le voir, montrent que les arginines SHL +/- 1.5 sont en position apicale, alors que les arginines vraiment insérées interagissent avec des séquences mixtes, A:T et G:C.





**Figure 12 – Arginines à l’interface ADN/histones du nucléosome.** Les arginines sont représentées en sphères opaques. L’ADN est en gris, l’histone H3 en bleu, l’histone H4 en rouge, l’histone H2A en vert et l’histone H2B en jaune. La moitié 5’ du nucléosome est représentée à gauche et la moitié 3’ à droite.

Enfin, une étude sur un ensemble de structures cristallographiques montre que les arginines H4-R45, H3-R83, H2A-R42 et H2A-R77 sont toujours insérées dans le petit sillon, respectivement aux sites SHL +/- 0.5, 2.5, 3.5 et 5.5 (Wang, Ulyanov, et Zhurkin 2010). La mutation de l’arginine H4 R45H, insérée dans le petit sillon au SHL +/- 0.5, entraîne une déstabilisation de la structure du nucléosome (Kruger et al. 1995) qui confirme le rôle clé de cette arginine dans l’interface. La même étude montre que la mutation H3 R116H et H3T118I de résidus proches de l’arginine insérée entraîne également une déstabilisation de la structure, mettant en évidence que l’arginine n’agit pas seule dans la structuration de l’interface. Bien que ces mutations n’existent que pour la site SHL +/- 0.5, les autres sites L1-L2 sont très similaires en structure et une autre étude désigne des résidus équivalents sur le site SHL +/- 2.5 (Karolin Luger 2003).

## Dynamique de la chromatine *in vivo*

La dynamique de la chromatine est liée au positionnement des nucléosomes le long d'un génome, lequel dépend d'un certain nombre de facteurs. Je vais présenter ici quelques-uns de ces facteurs.

### *Modifications post-traductionnelles des histones*

Les extrémités N-terminales des histones H3, H4 et H2B et l'extrémité C-terminale de l'histone H2A subissent de nombreuses modifications post-traductionnelles (MPT) (Zhao et Garcia 2015), une queue pouvant subir plusieurs modifications. L'ensemble de ces modifications forme le « code des histones » (Jenuwein et Allis 2001). De récentes études mettent en évidence le grand nombre de sites potentiels pour ces modifications (Zhao et Garcia 2015), rendant le décryptage du « code des histones » complexe.

Les MPT des histones incluent la méthylation, la phosphorylation, l'acétylation, l'ubiquitination et la sumoylation des acides aminés. Ils affectent le positionnement des nucléosomes et leur organisation de manière directe, en stabilisant ou déstabilisant les interactions nucléosome-nucléosome (Bowman et Poirier 2015) et de manière indirecte en favorisant ou défavorisant le recrutement d'autres facteurs.

### *Protéines chaperons des histones*

Les protéines chaperons des histones (PCH) sont des molécules qui jouent un rôle dans la structuration de la chromatine et la prévention de l'agrégation des protéines histones lors de sa formation (Burgess et Zhang 2013; Valieva, Feofanov, et Studitsky 2016). Aujourd'hui, on dénombre plus de quinze PCH (Tyler 2002; Formosa 2012; Venkatesh et Workman 2015; Mattioli, D'Arcy, et Luger 2015) impliquées dans l'assemblage et le désassemblage du nucléosome, le stockage et le transport des histones, la transcription, la réplication et la réparation de l'ADN (Tableau 1).

Comme on peut le voir, les PCH sont capables d'interagir avec des variants spécifiques des histones (Tableau 1). Il s'agit d'un niveau supplémentaire de régulation de la chromatine (Burgess et Zhang 2013) qui peut même se produire directement sur un nucléosome par le remplacement des histones à l'aide des facteurs de remodelage. Cependant, les interactions PCH/histones sont rarement exclusives et binaires, une PCH peut interagir avec plusieurs histones et variants d'histones, et le contraire est également vrai (Mattioli, D'Arcy, et Luger 2015). En raison de cette complexité, les études structurales sont rares et le fonctionnement de ces protéines, notamment au niveau de l'assemblage du nucléosome, reste incertain. Cependant, elles mettent en évidence le rôle important des variants d'histones dans la dynamique de la chromatine et de nombreux autres processus cellulaires.

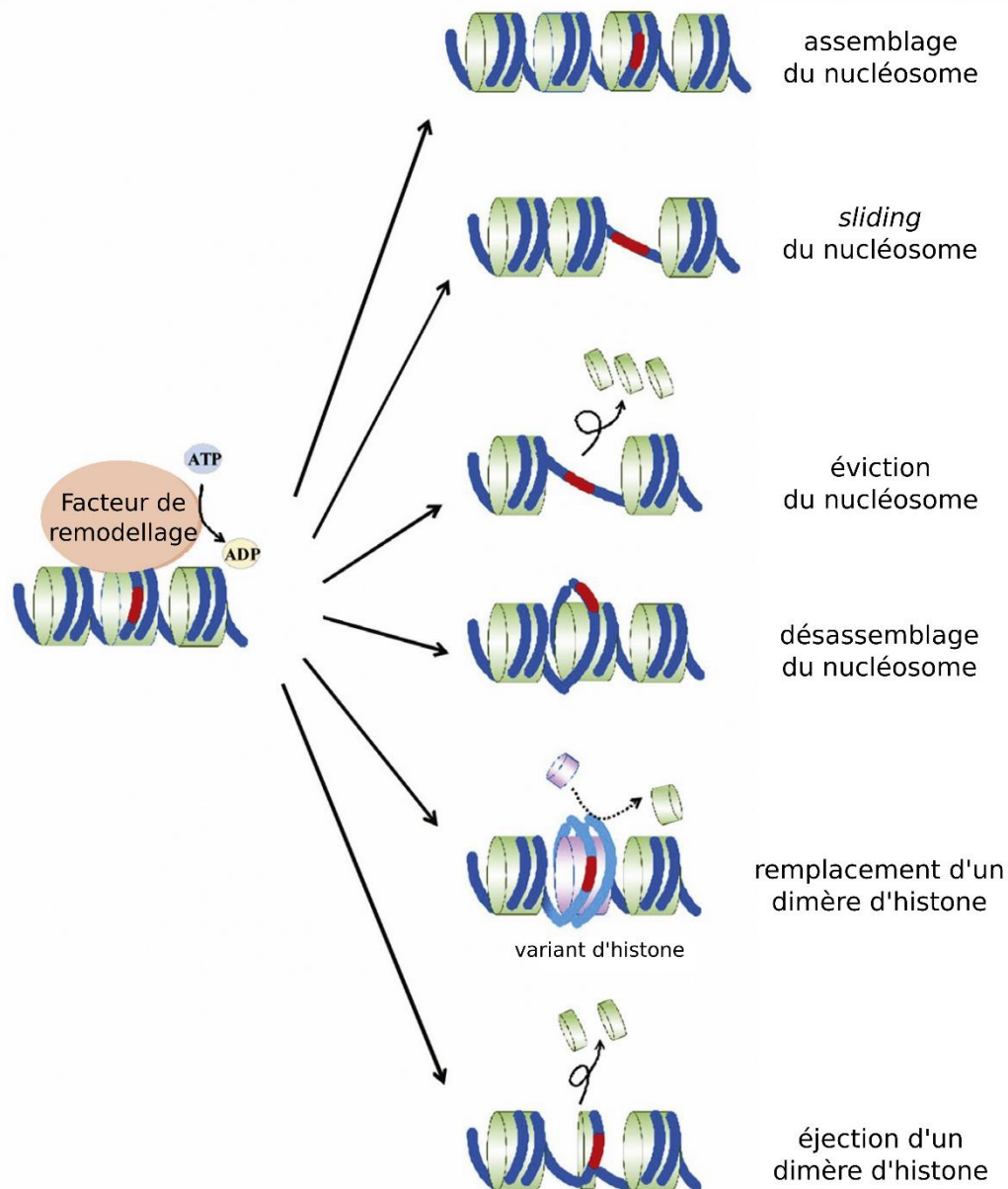
**Tableau 1 – Rôles et partenaires des protéines chaperons d’histone.** (source : Valieva, Feofanov, et Studitsky 2016)

| Histone chaperones | Histones-chaperone partners | Processes involving chaperones                                    |
|--------------------|-----------------------------|---|
| NAP1               | H2A-H2B, H2A.Z-H2B, H3-H4   | Transcription, import of histones into the nucleus from cytoplasm |
| Chz1               | H2A.Z-H2B                   | Transcription   |
| Swr1               | H2A.Z-H2B                   | Transcription   |
| ANP32E             | H2A.Z-H2B                   | Response to DNA damage  |
| FACT               | H2A-H2B, H3-H4              | Replication, transcription, repair                                |
| Spt6               | H3-H4                       | Transcription   |
| Asf1               | H3-H4, H3.3-H4              | Replication, transcription  |
| Rtt106             | H3-H4                       | Replication, transcription  |
| CAF-1              | H3-H4                       | Replication   |
| ANP32E             | H2A.Z-H2B                   | Transcription   |
| DAXX               | H3.3-H4                     | Formation of telomere chromatin                                   |
| Hir                | H3.3-H4                     | Transcription   |
| HIRA               | H3.3-H4                     | Transcription   |
| HJURP              | CenH3 <sup>CENP-A</sup>     | Formation of centromere   |
| Scm                | CenH3 <sup>CSE4</sup>       | Formation of centromere   |
| CAL1               | CenH3 <sup>CID</sup>        | Formation of centromere   |

### *Facteurs de remodelage*

Les facteurs de remodelage sont des enzymes capables de former des complexes multiprotéiques et qui utilisent l’énergie de l’hydrolyse de l’ATP pour induire des changements conformationnels au niveau du nucléosome et des domaines de la chromatine.

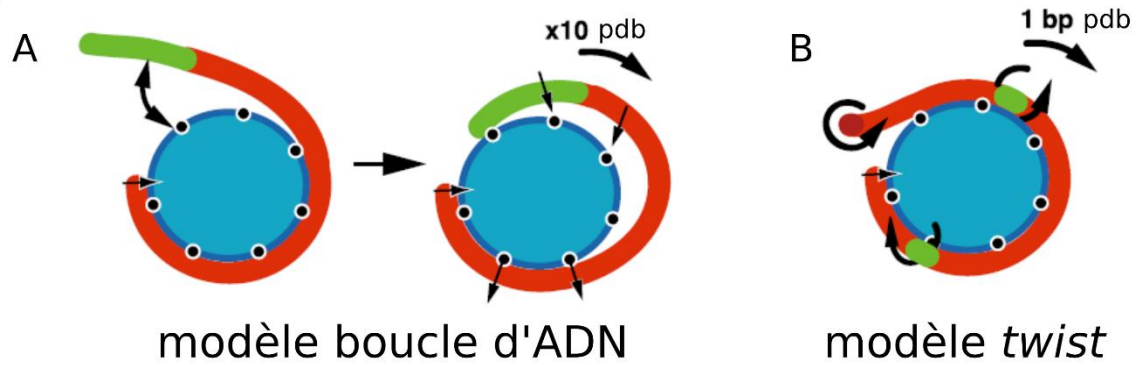
Le remodelage du nucléosome implique des changements dans les interactions ADN/histones afin de perturber ou de déplacer les nucléosomes (Becker et Workman 2013). Historiquement, le premier complexe de remodelage, Swi/Snf a été identifié chez la levure (Winston et Carlson 1992). C’est quelques années plus tard, lorsque le complexe a été purifié (Vignali et al. 2000), que son rôle perturbateur de la structure du nucléosome a pu être mis en évidence. Aujourd’hui, on catégorise les complexes de remodelage selon quatre familles, SNF2, ISWI, CHD et Ino80 (Becker et Workman 2013) aux rôles divers mais partageant plusieurs propriétés : leur affinité est plus forte pour le nucléosome que pour l’ADN libre, elles possèdent toute un domaine de reconnaissance des MTP, un domaine ATPase, un domaine de régulation du domaine ATPase et un domaine d’interaction avec d’autres facteurs de transcription (Volkh et al. 2016).



**Figure 13 – Rôle des facteurs de remodelage sur la conformation de la chromatine.** Les cœurs d'histones sont représentés par des cylindres verts ou rose lorsqu'un dimère histone est remplacé par un variant ; l'ADN est en bleu, avec un site en rouge pour mettre en évidence l'action du facteur de remodelage (d'après W. Chen et al. 2017).

Les facteurs de remodelage peuvent agir de différentes manières sur le nucléosome. Ils peuvent aider à l'assemblage, au désassemblage, à l'éviction des histones et au coulissement du nucléosome le long de l'ADN (Figure 13) (Becker et Workman 2013). Ils peuvent également permettre l'éviction ou le remplacement d'un dimère d'histone au sein d'un nucléosome formé (Figure 13) (Becker et Workman 2013). Les mécanismes d'action de ces facteurs de remodelage ne sont pas bien connus, bien qu'il

soit clair qu'ils agissent en collaboration avec les PCH et sont influencés par les MTP. Deux hypothèses se distinguent pour le repositionnement du nucléosome sans désassemblage : la formation d'une boucle d'ADN d'environ 10 pb qui se déplace le long du nucléosome (Figure 14A) (Becker et Workman 2013; Pasi et al. 2016) et une rotation de l'ADN sur lui-même (*twist*) qui lui permet d'avancer le long de l'interface (Figure 14B)



**Figure 14 – Mécanismes de repositionnement du nucléosome.** L'ADN est schématisé en rouge et en vert, le nucléosome en bleu (d'après Längst et Becker 2004).

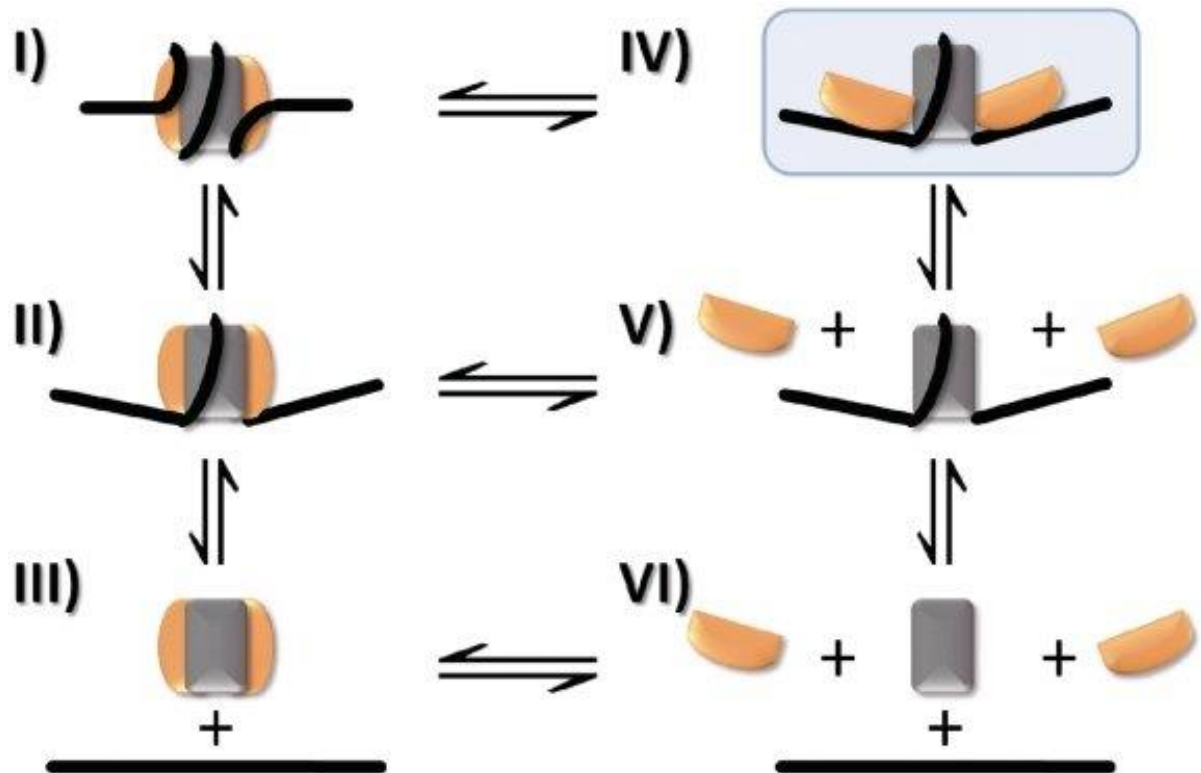
## Assemblage et désassemblage du nucléosome *in vitro*

Comme nous l'avons vu, *in vivo* de nombreux mécanismes permettent la modulation de la chromatine et du nucléosome. Ainsi, il semble clair que l'assemblage et le désassemblage du nucléosome *in vivo* est la résultante de tous ces mécanismes de régulation faisant intervenir les facteurs de remodelage, les protéines chaperons des histones ainsi que les modifications post-traductionnelles des queues d'histones. Cependant, bien que les connaissances sur ces phénomènes aient cru au cours de ces dernières années, en reproduire les effets *in vitro* reste un défi majeur. Par contre, tenter de comprendre les mécanismes physiques (en l'absence d'autres molécules que les ADN et les histones) à l'œuvre lors de l'assemblage et le désassemblage des nucléosomes est possible. Pour éviter que les histones forment des agrégats, l'assemblage des nucléosomes s'effectue en partant d'une haute force ionique et en la diminuant progressivement. La stratégie inverse dissocie le nucléosome. Cette méthode a été souvent employé, par exemple par Park et al. 2005; Hoch, Stratton, et Gloss 2007; Gansen et al. 2009; Böhm et al. 2011; Hatakeyama et al. 2016; Gansen et al. 2018. D'autres études imposent des forces mécaniques à l'ADN pour en étudier le désassemblage (Hall et al. 2009; Gansen et al. 2018).

### Désassemblage du nucléosome

De nombreux travaux cherchent à comprendre le mécanisme de désassemblage du nucléosome. Un des mécanismes proposé est le détachement de l'octamère d'histone en un seul bloc (Brower-Toland et al. 2002) (Figure 15 suivant les étapes I  $\rightarrow$  III  $\rightarrow$  VI). Cependant de nombreuses études FRET (*Fluorescence Resonance Energy Transfert*) (Park et al. 2005; Hoch, Stratton, et Gloss 2007; Gansen et al. 2009; Böhm et al. 2011; Gansen et al. 2018) proposent un modèle séquentiel du désassemblage, maintenant généralement accepté, avec d'abord le relargage des dimères H2A-H2B suivie de celui du tétramère (H3-H4)<sub>2</sub> (Figure 15 suivant les étapes I  $\rightarrow$  IV  $\rightarrow$  V  $\rightarrow$  VI). Ce second modèle est soutenu par des études de microscopie à force atomique à haute vitesse (AFM) (Hagerman et al. 2009; Miyagi, Ando, et Lyubchenko 2011) qui mettent en évidence que les interactions impliquant les extrémités de l'ADN qui interagissent avec les dimères H2A-H2B sont déstabilisées en premier, suivi de celles concernant la région de l'ADN qui interagit avec le tétramère (H3-H4)<sub>2</sub>.

Cependant, de récentes études du nucléosome contenant la séquence 601 par FRET (Gansen et al. 2018) et diffusion RX aux petits angles résolue en temps (*TR-SAXS*) (Y. Chen et al. 2017) mettent en évidence que la libération des deux dimères H2A-H2B est asymétrique et proposent que la libération d'un premier dimère diminue l'affinité du second, qui est donc à son tour libéré très rapidement. Cela rajoute donc une étape aux modèles proposés (Figure 15).



*Figure 15 – Etapes du désassemblage et de l’assemblage du nucléosome. Ce schéma représente différentes étapes possibles du désassemblage et de l’assemblage du nucléosome pour la transition entre le nucléosome complètement formé (I) et les partenaires libres (VI) (source : Böhm et al. 2011)*

### *Assemblage du nucléosome*

Le mécanisme de l’assemblage est bien moins étudié d’après la littérature disponible. Pourtant, très tôt, des études montrent que l’ADN est d’abord contacté par le tétramère (H3-H4)<sub>2</sub> suivis des dimères H2A-H2B (Worcel, Han, et Wong 1978; Crémisi et Yaniv 1980). Récemment, un article du groupe de Malcolm Buckle et Claude Nogues utilisant la méthode PhAST (Méthode PhAST, p. 61) (Hatakeyama et al. 2016) confirme ce schéma séquentiel. En utilisant la séquence 601 non palindromique et un variant de cette séquence, les auteurs montrent également que l’assemblage est sensible à la séquence de l’ADN, pour la fixation de (H3-H4)<sub>2</sub> et, dans une moindre mesure, des dimères H2A-H2B. Ces résultats indiquent donc que le mécanisme de l’assemblage est aussi complexe que celui du désassemblage (Figure 15).



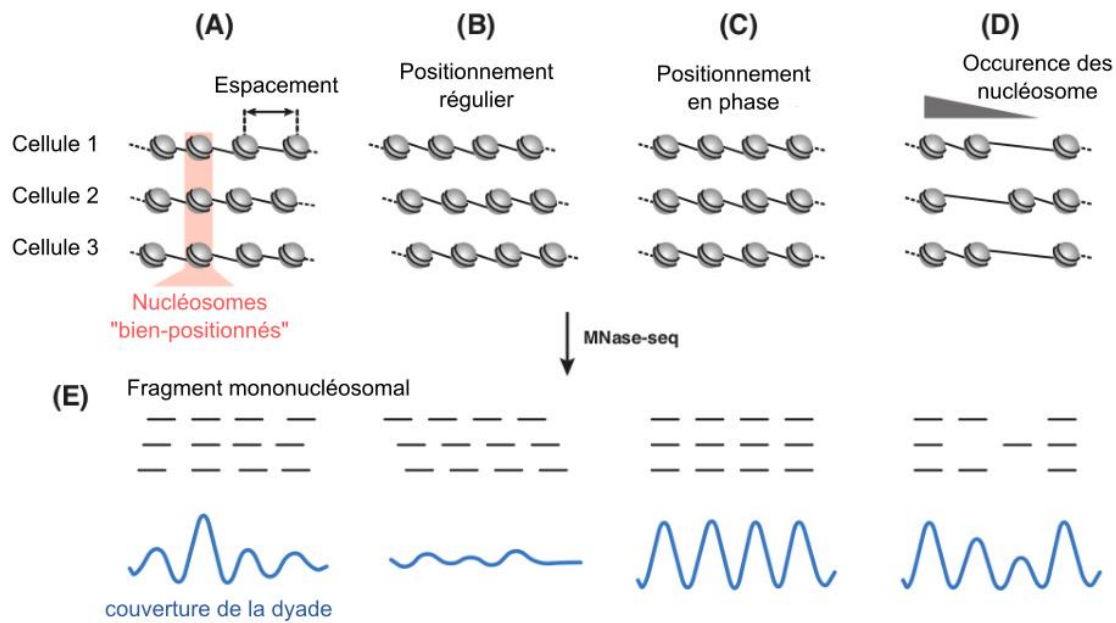
## Positionnement et spécificité du nucléosome

Le positionnement du nucléosome fait référence aux sites occupés par ce complexe le long de l'ADN génomique. Il est notamment caractérisé par l'espacement entre deux nucléosomes successifs, appelée la distance de répétition du nucléosome (DRN ou, en anglais, NRL pour *Nucleosome Repeat Length*). Cette distance varie selon les espèces, voire selon le type cellulaire au sein d'un même organisme (Holde 1989; Valouev et al. 2011). De plus, les DRN dépendent de nombreux facteurs, notamment la séquence d'ADN, la concentration des histones, des facteurs de remodelage et des protéines chaperons des histones. Une étude met également en évidence que les interactions à longue portée entre nucléosomes ont un impact sur leur positionnement (Valouev et al. 2011).

### *Mapping des nucléosomes le long du génome*

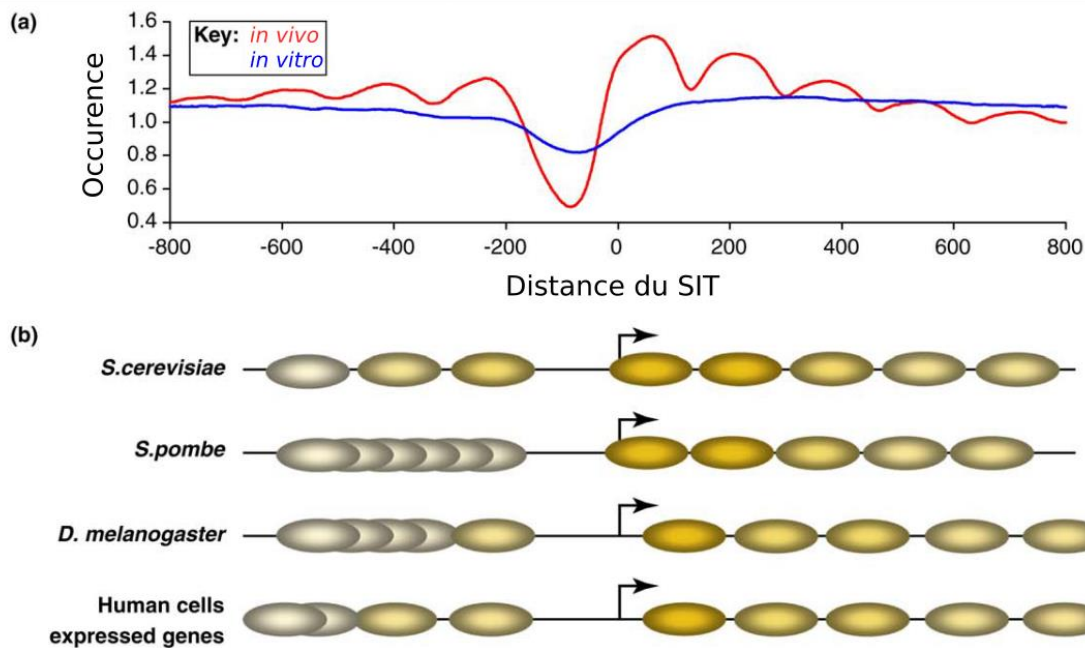
La méthode privilégiée pour cartographier les nucléosomes est l'utilisation de la nucléase micrococcal (MNase) qui est capable de digérer les régions qui relient les nucléosomes (Lantermann et al. 2010; Kent et al. 2011; Valouev et al. 2011; Albert et al. 2007), bien que cette enzyme présente des préférences pour certaines séquences (Dingwall, Lomonosoff, et Laskey 1981) et puisse mieux digérer l'ADN nucléosomal dans des régions très accessibles comme celles proches des sites d'initiation de la transcription (Weiner et al. 2010). Cette méthode, appelée MNase-Seq permet d'obtenir les fragments d'ADN précédemment enroulés autour des histones (Figure 16-E). Ces fragments sont ensuite séquencés et cartographiés sur un génome, afin d'obtenir les positions des nucléosomes le long du génome. Les études MNase-Seq permettent notamment de caractériser les différents types de positionnement du nucléosome. Une revue très récente (Baldi 2019) décrit et définit les différents types de positionnement des nucléosomes dans les populations de cellules. Lorsque les nucléosomes sont positionnés constamment aux mêmes régions génomiques, on les appelle nucléosomes « bien positionnés » (Figure 16-AC). Mais ils peuvent avoir un positionnement régulier avec le même espacement, sans pour autant être en phase dans les différentes populations de cellules (Figure 16-B). On peut mesurer l'occurrence des nucléosomes pour chaque pb, et donc le pourcentage de molécules d'ADN qui sont en interaction avec un nucléosome dans les différentes populations (Figure 16-D).





**Figure 16 – Caractéristiques du nucléosome observées par MNase-seq.** (A) Régions d'ADN génomique garnies de nucléosomes bien positionnés. (B) Positionnement régulier mais faible. (C) Positionnement fort, en phase. (D) Positionnement avec variation de l'occurrence des nucléosomes. (E) Obtention de fragments mononucléosomaux générés par la digestion des MNase et signal correspondant aux dyades des nucléosomes (source : Baldi 2019)

Les régions les plus étudiées du nucléosome par ces méthodes sont les sites d'initiation de la transcription (SIT). Juste en 3' de ces régions, quelle que soit l'espèce, on observe un positionnement en phase très prononcé (Mavrich et al. 2008; Schones et al. 2008; Lantermann et al. 2010; Bai et Morozov 2010). En revanche, la région en 5' des SIT est déplétée en nucléosome (*nucleosome free region*) (Figure 17). Bien que l'effet soit beaucoup plus prononcé *in vivo*, l'analyse des régions SIT *in vitro* (Yong Zhang et al. 2009; Kaplan et al. 2009) met en évidence que l'affinité des ADN pour les histones est plus faible pour cette région en 5' des SIT (Figure 17). Cependant, le positionnement en phase observé en 3' des SIT serait plutôt dû à de nombreux facteurs extérieurs (Yong Zhang et al. 2009).



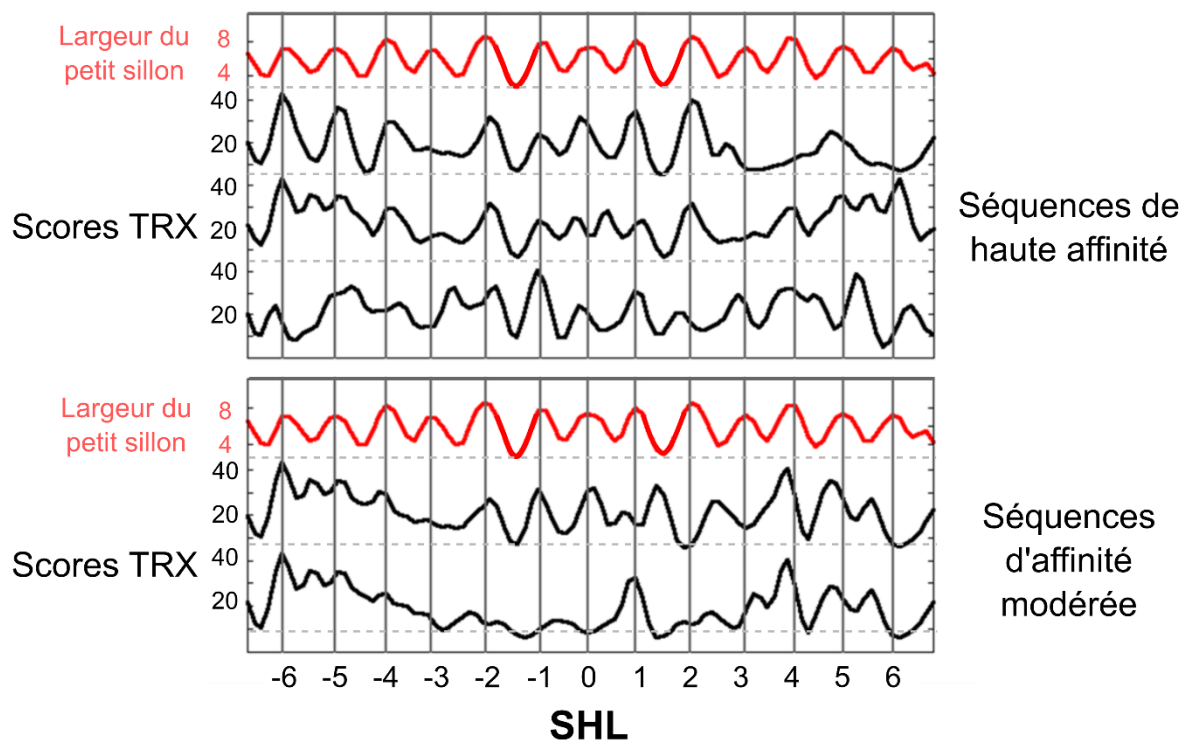
**Figure 17 – Positionnement du nucléosome et sites d'initiation de la transcription.** (a) Occurrence moyenne *in vivo* (rouge) et *in vitro* (bleu) du positionnement des nucléosomes des gènes de *S. cerevisiae* alignés. (b) Représentation schématique du positionnement typique des nucléosomes *in vivo* chez plusieurs espèces (source : Bai et Morozov 2010).

En dehors des SIT, le positionnement en phase est un phénomène plus rare mais toutefois déjà observé, notamment chez les mammifères pour les sites de fixation du répresseur transcriptionnel CTCF et d'autres sites de fixation des facteurs de transcription (Wiechens et al. 2016). Ces résultats montrent que le positionnement en phase est privilégié dans les régions régulatrices de l'expression génique et de la synthèse protéique.

### Rôle de la séquence ADN dans la formation du nucléosome

Comme indiqué précédemment, le positionnement du nucléosome *in vivo* est très clairement impacté par la présence de facteurs externes. Cependant, des études *in vitro* montrent aussi l'impact de la séquence de l'ADN (Kaplan et al. 2009; Bai et Morozov 2010; Baldi 2019). Notamment, la composition des séquences des régions déplétées en nucléosome en 5' des SIT chez la levure sont très riches en A:T (W. Lee et al. 2007) ce qui défavoriserait la formation de nucléosome *in vitro* (Kaplan et al. 2009) et *in vivo* (Raveh-Sadka et al. 2012). Le consensus actuel est que les séquences d'ADN les plus aptes à former des complexes présentent une périodicité lexicographique d'environ dix pb, alternant des motifs riches en A:T dans les régions en interaction avec le cœur d'histone et riches en G:C dans les régions tournées vers l'extérieur. (Drew et Calladine 1987; Jiang et Pugh 2009;

Kaplan et al. 2009; Cui et Zhurkin 2010; Brogaard et al. 2012; Andrew Travers et al. 2010; Struhl et Segal 2013).



**Figure 18 – Scores TRX pour des séquences d'affinité haute et modérée pour le cœur d'histone.** La variation sinusoïdale de la largeur des petits sillons (en Å) le long de l'ADN est représentée en rouge en haut de chaque panel. Les profils des scores TRX de trois séquences de haute affinité (haut) et deux séquences d'affinité modérée (bas) le long de l'ADN sont représentés en noir (d'après Xu et al. 2014).

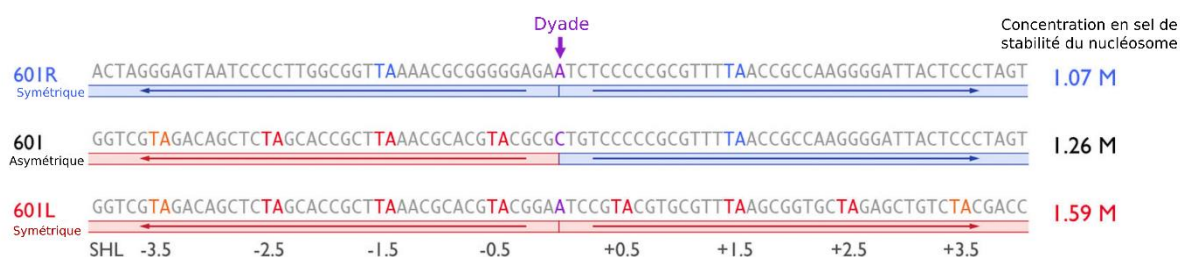
Au laboratoire, sur la base de données RMN « brutes » (en particulier des déplacements chimiques) collectées sur des ADN libres, Brigitte Hartmann et ses collaborateurs ont conçu l'échelle TRX, qui associe à chacun des 10 dinucléotides complémentaires (ApA•TpT, CpG•CpG...) un score qui reflète la dimension de l'espace conformationnel exploré (Heddi et al 2008, 2010, Ben Imeddourene et al 2016). Cette échelle a une valeur prédictive (Xu et al. 2014). D'après cette échelle, les dinucléotides ApA/G•T/CpT, et ApT/C•ApT/C présentent une flexibilité restreinte par rapport aux pas GpG•CpC, CpG•CpG, GpC•CpC et CpA•TpG qui explorent un large espace conformationnel. Ces différences de flexibilité entre motifs riches en A:T et riches en G:C rend parfaitement compte de l'effet de séquence sur l'affinité de l'ADN pour les histones (Heddi et al 2010 ; Xu et al 2014). Pour aller plus loin, un ensemble de séquences nucléosomales artificielles dont les affinités *in vitro* pour le cœur

d'histone sont publiées dans la littérature a ainsi été annoté par TRX. Ces profils TRX ont ensuite été comparés avec les déformations structurales observées dans le nucléosome, illustrées par les variations périodiques sinusoïdales de la largeur du petit sillon, connues pour être conservées indépendamment de la séquence d'ADN. Les meilleurs parallèles entre ces deux types de profils (TRX et dimension du petit sillon) correspondent aux séquences dont l'affinité pour le cœur d'histone était la plus élevée (Figure 18).

Il semble donc clair que les propriétés intrinsèques de l'ADN jouent un rôle important dans l'affinité avec les histones, du moins *in vitro*.

### Le cas de la séquence 601

La séquence 601 (Thåström et al. 1999) est une séquence artificielle identifiée par des expériences SELEX pour sa très forte affinité pour les histones. Son très fort positionnement en fait une séquence de choix pour les études sur la chromatine.



**Figure 19 – Séquence 601 et ses dérivées symétriques.** La symétrisation de la séquence 601 à partir de la duplication de la moitié 3' (601R) diminue l'affinité pour les histones ; la même stratégie appliquée à la moitié 5' (601L) a l'effet inverse (d'après McGinty et Tan 2015).

Cette séquence 601 n'est pas palindromique. L'ADN résultant de la symétrisation de sa moitié 5' (séquence 601L, Figure 19) est plus affine que la séquence d'origine ; par contre la symétrisation de sa moitié 3' (séquence 601R, Figure 19) a un résultat inverse (Chua et al. 2012). En plein accord avec ces résultats, l'étude de l'assemblage du nucléosome montre des étapes intermédiaires traduisant la relative difficulté pour les histones de se fixer solidement à sa moitié 3' (Hatakeyama et al. 2016). Ce serait la périodicité des pas TA toutes les dix paires de bases que l'on trouve dans la moitié 5' qui seraient à l'origine de la préférence des histones (Chua et al. 2012). L'échelle TRX confirme cette idée, en rajoutant que la périodicité de pas flexibles, également observée toutes les dix paires de bases, est aussi importante que celle des pas TA (Xu et al 2014).

# Les complexes acides nucléique-NCp7

## Le virus de l'immunodéficience humaine (VIH)

Le virus de l'immunodéficience humaine (VIH) est un rétrovirus qui agit au niveau des cellules du système immunitaire et nerveux central. Il est responsable du syndrome d'immunodéficience acquise (SIDA), un état affaibli du système immunitaire, rendant vulnérable aux infections opportunistes. La première souche de VIH, le VIH-1 identifiée en 1983 par microscopie électronique (Barré-Sinoussi et al. 1983) est le type le plus commun et le plus pathogène et il présente une grande variété de souches virales que l'on classe selon quatre groupes principaux M, N, O et P (Hemelaar 2012). Le groupe M (*Main*) représente la majorité des infections chez l'homme (plus de 90%) et est divisé en plusieurs sous-types. Les groupes O (*Outlier*), N (*Non-M, Non-O*) et le groupe P (*Pending*) sont des groupes plus rares, classés selon leur localisation.

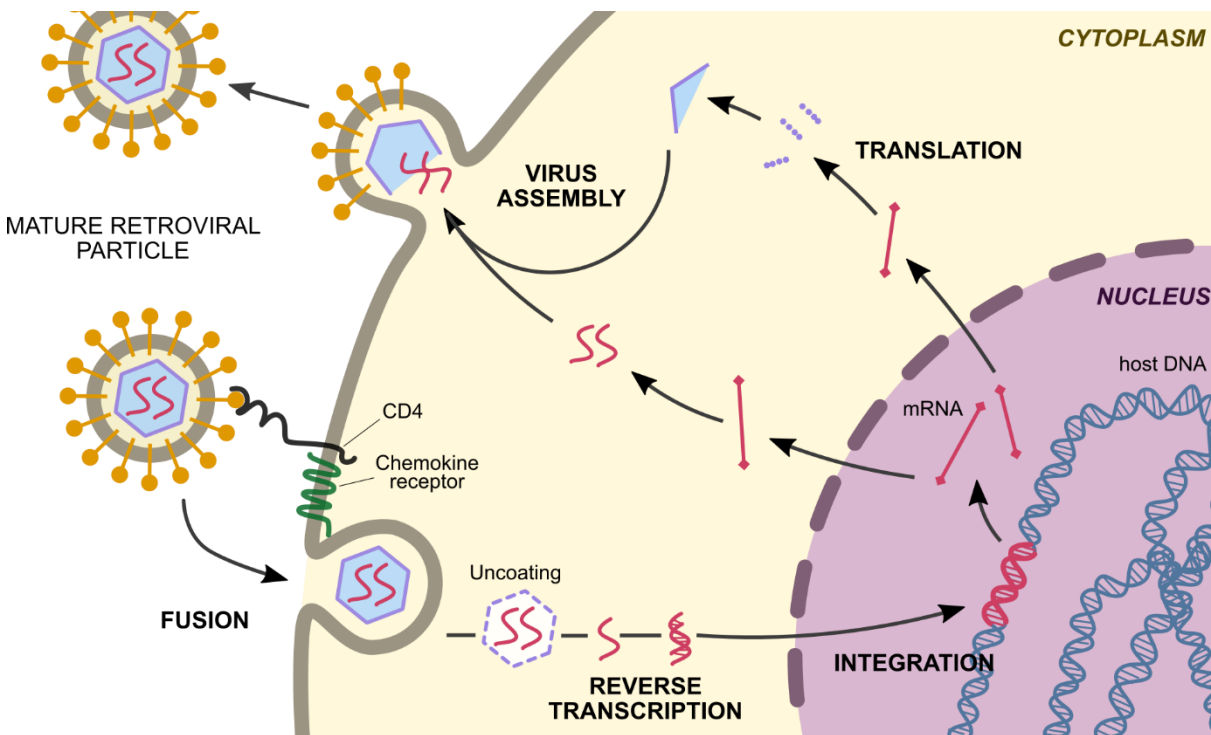
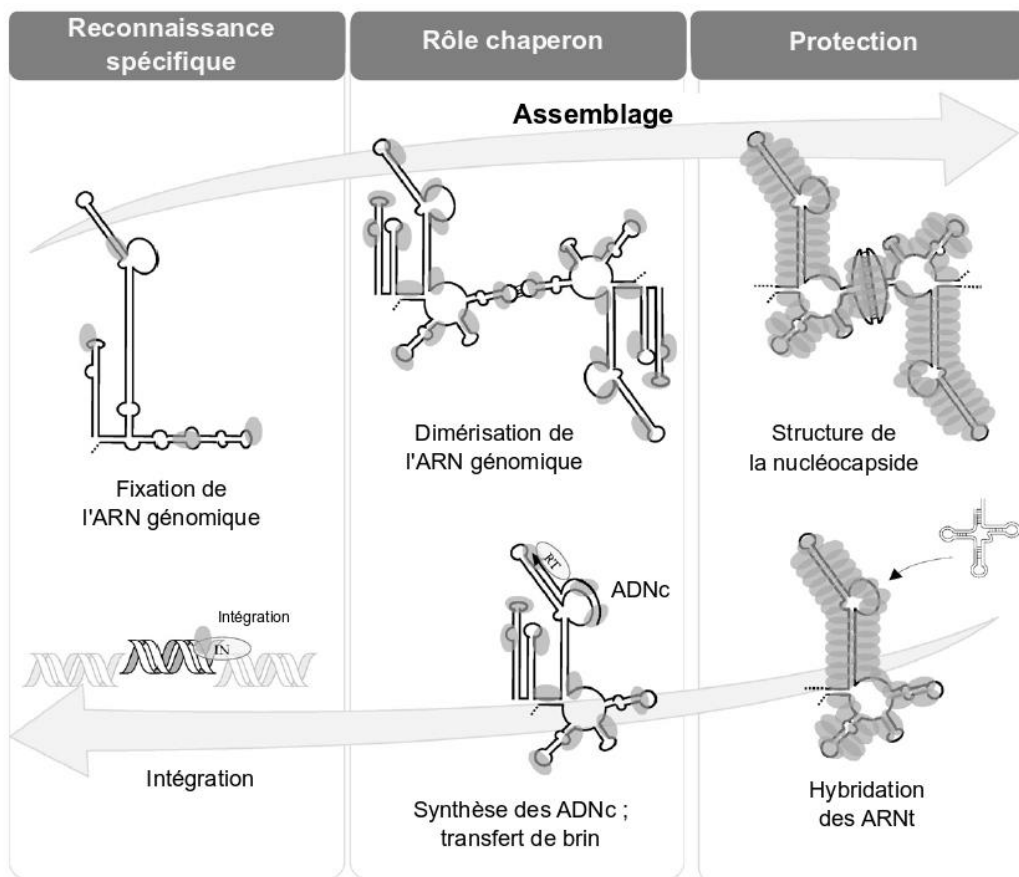


Figure 20 – Représentation schématique du cycle de réplication du VIH-1.

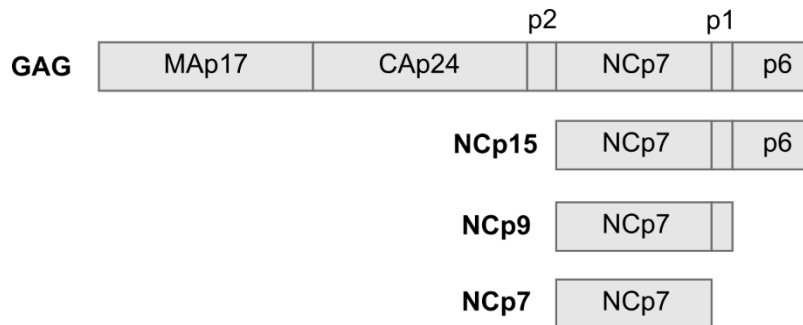
## Rôle de la NCp7 dans le cycle du VIH-1

La protéine 7 de la nucléocapside (NCp7) est une protéine de structure du VIH qui intervient à de nombreuses étapes. Elle participe à la sélection de l'ARN génomique et de l'ARN de transfert, l'hybridation des ARN, la dimérisation et l'encapsidation du génome viral, l'initialisation de la transcription inverse, le transfert de brin et son intégration dans le génome de l'hôte, la formation d'une particule virale et son bourgeonnement et enfin la protection de l'ARN contre les nucléases cellulaires (Figure 21) (J. L. Darlix et al. 1995; Levin et al. 2005; J.-L. Darlix et al. 2007; Levin et al. 2010; J.-L. Darlix et al. 2011; Sleiman et al. 2012; J.-L. Darlix et al. 2014). Elle interagit donc dans de très nombreuses étapes du cycle du VIH (Figure 20) impliquant des acides nucléiques et elle est par conséquent une cible thérapeutique très intéressante (Iraci et al. 2018; Sancineto et al. 2018).



**Figure 21 – Illustration des rôles de la NCp7 dans le cycle de réplication du VIH. (D'après J.-L. Darlix et al. 2011)**

La NCp7 est une forme mature de la polyprotéine Gag (Figure 22). Il s'agit d'une petite protéine de 55 résidus, capable de se fixer aux acides nucléiques (ADN et ARN) avec une préférence pour les AN en simple brin. Elle a en particulier une activité de chaperon des AN, et est ainsi capable de modifier la structure secondaire et tertiaire des ARN et des ADN sur lesquels elle se fixe.



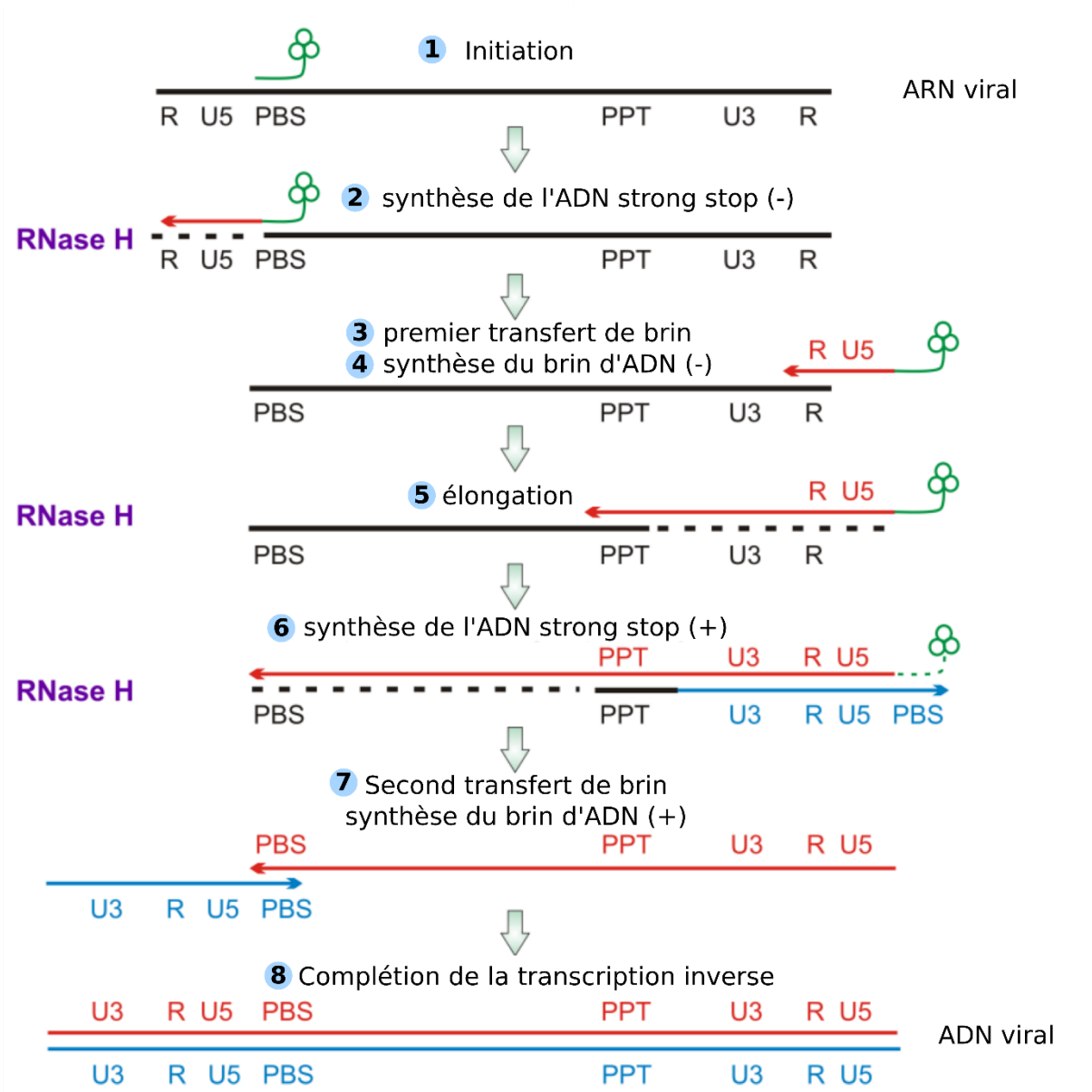
**Figure 22 – Maturation de la NCp7 à partir de la polyprotéine Gag**

### *Protection des acides nucléiques*

Dans la capsid virale, les propriétés d'association et de compaction de la NCp7 permettent la protection de l'ARN génomique (J. L. Darlix et al. 1995; J.-L. Darlix et al. 2011) de manière passive. Des études in vitro ont mis en évidence qu'en condition saturante, la NCp7 se fixe de manière non-spécifique aux ARN (Tanchou et al. 1995; Krishnamoorthy et al. 2003). Ces mêmes études montrent notamment qu'une forte concentration de NCp7 est capable de protéger les ARN contre la dégradation par les nucléases mais de manière limitée.



Le transcription inverse



**Figure 23 – Transcription inverse du VIH-1.** L'ARN est représenté en noir, l'ARN<sup>lys3</sup> vert et les brins d'ADN négatif (ADN(-)) et positif (ADN(+)) en rouge. Les ADN ou ARN dégradés sont représentés par des pointillés. Le détail des étapes est précisé dans le texte (Cf. Transcription Inverse) (d'après Ilina et al. 2012)



C'est essentiellement dans la transcription inverse du VIH que la NCp7 joue son rôle de chaperon des acides nucléiques. La transcription inverse consiste à convertir l'ARN viral génomique simple brin en ADN double brin linéaire. Il s'agit d'un mécanisme complexe, commun à tous les rétrovirus qui se déroule en huit étapes (Ilina et al. 2012; Hu et Hughes 2012) (Figure 23) :

- **1.** la transcription inverse est initiée par l'hybridation de l'ARNt sur le site de liaison d'amorce (PBS pour primer binding site).
- **2.** Synthèse de l'ADN strong-stop (-) à l'extrémité 3' de l'ARN viral et dégradation de la matrice ARN.
- **3.** Premier transfert de brin : l'hybridation des séquences cTAR et TAR permet le transfert de l'ADN strong-stop (-) à l'extrémité 3' de l'ARN.
- **4.** Synthèse de l'ADN négatif (ADN (-)) et dégradation de la matrice ARN.
- **5.** Elongation du brin d'ADN (-).
- **6.** Deuxième transfert de brin
- **7.** Synthèse de l'ADN positif (ADN (+))
- **8.** Complétion de la transcription inverse, l'ARN viral double brin est formé.

Lors de ce processus, la NCp7 intervient à différents niveaux je vais détailler.

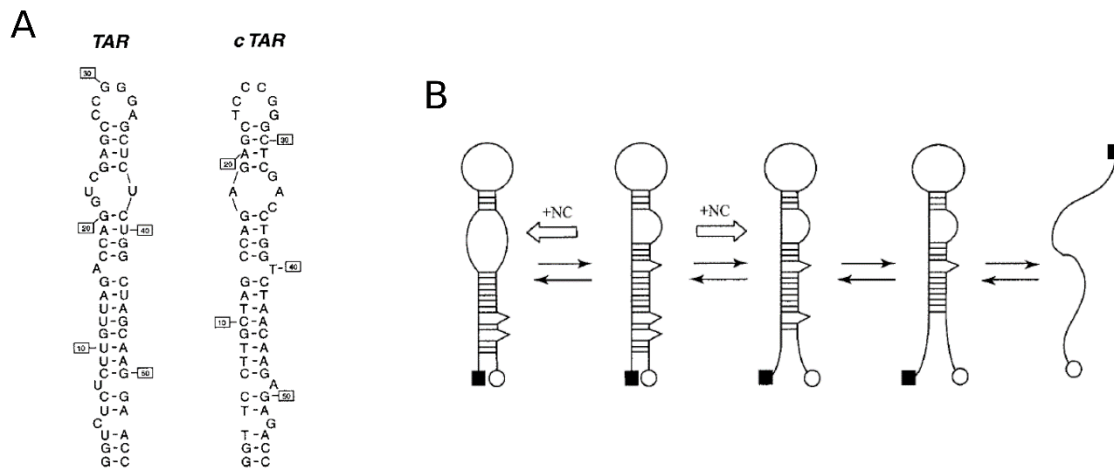
#### Initiation de la synthèse du brin ADN (-)

L'initiation se déroule dans la capsid virale avant la fusion du virus avec la membrane de la cellule cible. Afin de mener à bien cette étape, l'ARN de transfert ARN<sup>tLys3</sup> qui servira d'amorce pour la synthèse du brin ADN(-) (

Figure 23) doit s'hybrider avec le PBS mais il a été montré *in vivo* que les deux partenaires ne s'hybrident pas spontanément (Sleiman et al. 2012). C'est à ce premier niveau qu'intervient la NCp7 qui va stimuler cette réaction, en augmentant le taux d'hybridation d'un facteur  $10^5$  (Hargittai et al. 2004) et en effectuant un réarrangement conformationnel des AN.

#### Premier transfert de brin

Lors de cette étape, l'ADN strong-stop (-) synthétisé à la suite de l'initiation va être transféré vers l'extrémité 3' de l'ARN viral. Cela est rendu possible par la présence de la séquence tige boucle cTAR sur l'ADN strong-stop (-) qui va s'hybrider avec la séquence TAR de l'extrémité 3' de l'ARN (J. L. Darlix et al. 1993). TAR et cTAR sont structurées en tiges boucles mais présentent des mésappariements et des *bulges* (Bernacchi et al. 2002; Beltz et al. 2003) (Figure 24A).



**Figure 24 – Structures secondaires des séquences TAR et cTAR et déstabilisation de cTAR par la NCp7.** A) Structures secondaires des séquences TAR et cTAR. B) Représentation schématique de la déstabilisation de cTAR par la NCp7 (d’après Bernacchi et al. 2002).

L’hybridation spontanée de TAR et cTAR est peu efficace (Beltz et al. 2005). Dans cette étape, la présence de la NCp7 permet une augmentation du taux d’hybridation par un facteur 3000 (Godet et al. 2006). La NCp7 induit la déstabilisation de l’extrémité cTAR (Figure 24B) et son ouverture, facilitant l’hybridation avec son partenaire TAR (Bernacchi et al. 2002; Beltz et al. 2003).

### Second transfert de brin

Le second transfert de brin implique cette fois-ci la séquence PBS positive (PBS (+)), transférée à l’extrémité 5’ du brin d’ADN (-) (

Figure 23). Cette fois-ci, la NCp7 va faciliter l’hybridation entre la séquence PBS (+) et la séquence PBS (-) avec une augmentation d’un facteur 60 du taux d’hybridation (Egelé et al. 2004). La NCp7 ne déstabilise pas la tige boucle de PBS qui est beaucoup plus courte mais induit l’exposition de ses bases nucléotidiques (Bourbigot et al. 2008).

### Transport de l’ADN viral

Le rôle de la NCp7 dans le transport de l’ADN viral vers le noyau est encore mal connu. Des études de fluorescence suggèrent que la NCp7 ne se fixe que de manière non-spécifique sur l’ADN après la transcription inverse et que sa présence dans le complexe pré-intégration et le noyau est très faible (Anton et al. 2015). Elle ne jouerait donc pas un rôle majeur dans le transport de l’ADN viral.

### Intégration de l'ADN viral

L'intégration de l'ADN viral se fait en trois étapes (Figure 25). D'abord, le clivage d'un dinucléotide GT à chaque extrémité 3' de l'ADN viral par l'intégrase, libérant un dinucléotide CA. S'ensuit un transfert de brin avec l'insertion de l'extrémité CA-OH 3' dans le génome cellulaire. Suite à l'insertion partielle de l'ADN viral dans le génome, les enzymes de réparations de la cellule hôte vont rétablir l'intégrité de l'ADN.

Des études *in vitro* montrent que la présence de la NCp7 permet d'accélérer l'étape du transfert de brin (Carteau, Gorelick, et Bushman 1999; Poljak et al. 2003). *In vivo*, une étude de mutation de la NCp7 inhibe l'activité intégrase du virus (J. A. Thomas et al. 2006).

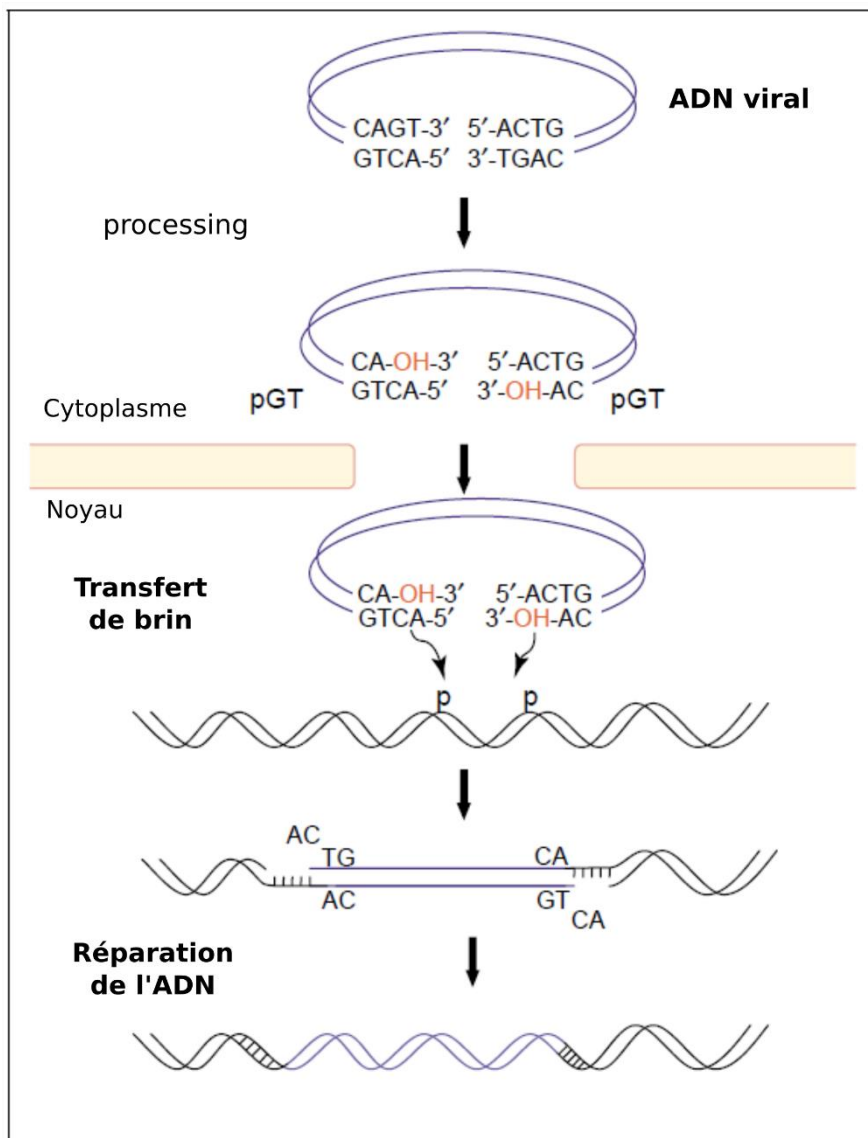


Figure 25 – Schéma du processus d'intégration de l'ADN viral dans l'ADN de la cellule hôte. (d'après Maele et al. 2006)

## Encapsidation, dimérisation et discrimination de l'ARN génomique (ARNg)

Lors de l'assemblage, pour former de nouveaux virions, le virus doit être capable de sélectionner spécifiquement l'ARNg au milieu de nombreux ARN cellulaires et viraux (Abd El-Wahab et al. 2014) afin de procéder à l'encapsidation et la dimérisation de l'ARNg, nécessaire à la reproduction du virion. L'initiation de l'encapsidation et de la dimérisation et la sélection spécifique de l'ARNg se fait par l'intermédiaire du signal d'encapsidation  $\Psi$ , composé des tiges boucles SL1, SL2, SL3 et SL4 (SL pour l'anglais *Stem Loop*) qui se situent dans la région 5' non-codante de l'ARNg (Figure 26).

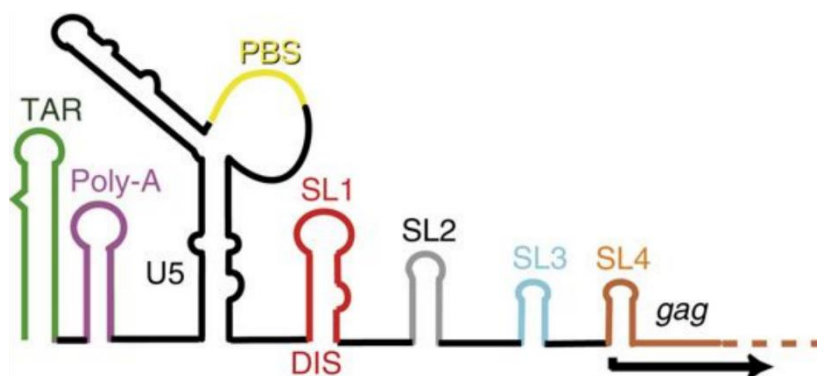


Figure 26 – Représentation schématique de la structure secondaire de la région 5' (500 résidus) non codante de l'ARN génomique du VIH-1. (Source : Abd El-Wahab et al. 2014)

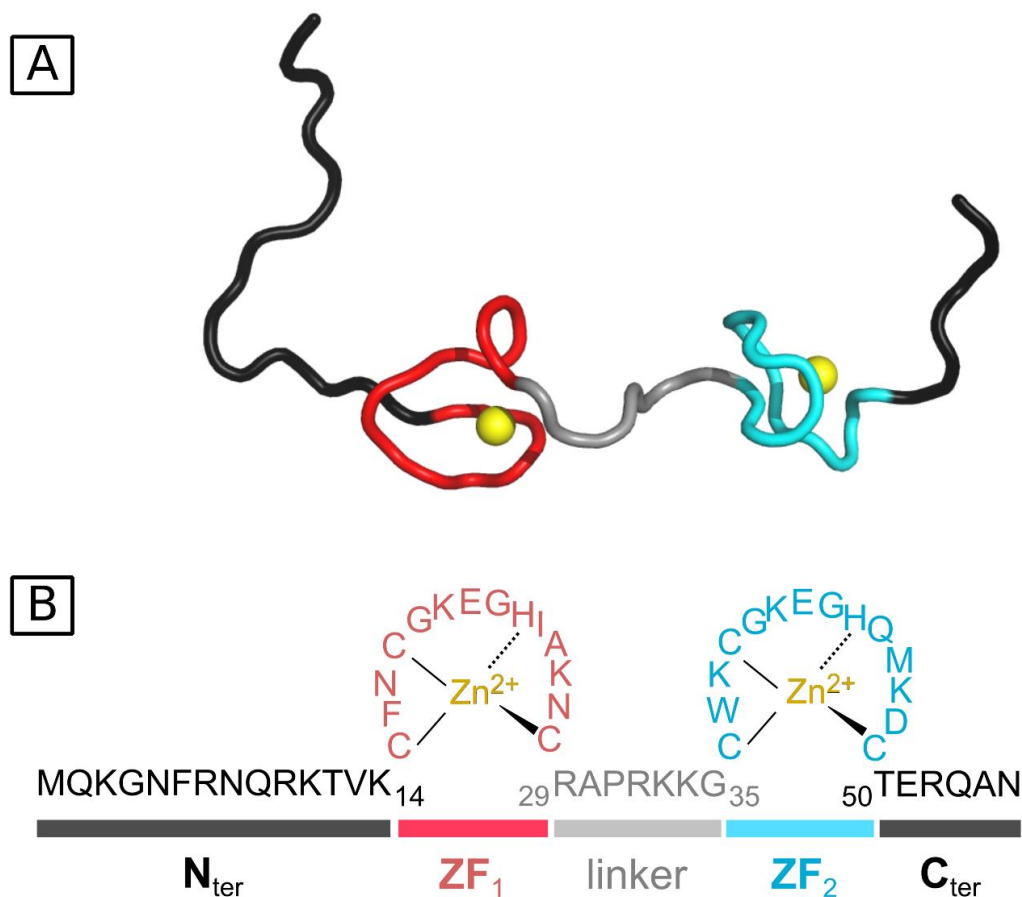
Pendant longtemps, la région SL3 été considérée comme le site initiateur de l'encapsidation le plus important du VIH, mais des études expérimentales montrent un impact mineur de la délétion de SL3 sur l'encapsidation (Harrison et al. 1998; Houzet et al. 2007) alors que la délétion de la région SL1 entraîne une forte réduction de l'infectiosité du virus (Houzet et al. 2007). Le domaine NC de Gag et/ou la NCp7, en se fixant de manière très spécifique sur ces régions, permettraient l'initiation de l'encapsidation (Y. Zhang et Barklis 1995; De Guzman et al. 1998), notamment par la reconnaissance de la séquence  $\Psi$  composée d'une boucle GGAG.

L'initiation de la dimérisation se situe également au niveau de la tige boucle SL1, aussi appelée DIS (pour l'anglais *Dimerization Initiation Site*). Le domaine NC de Gag et/ou la NCp7 interviennent à ce niveaux afin d'aider à la formation d'un duplex entre deux domaines DIS (Muriaux et al. 1996).

# Propriétés structurales de la NCp7

## Structure de la NCp7

Depuis les premières structures obtenues par RMN (Omichinski et al. 1991; Summers et al. 1992; Morellet et al. 1992) on sait que la NCp7 est caractérisée par deux doigts de zinc très structurés de motif X<sub>2</sub>-C-X<sub>4</sub>-H-X<sub>4</sub>-C où X représente des acides aminés variables. Les doigts de zinc sont appelés ZF1 (pour *Zinc Finger 1*) et ZF2 (pour *Zinc Finger 2*) et sont séparés par une région appelée *linker* de sept résidus non structurés. A ses extrémités, la NCp7 possède des queues non-structurées de 14 (en N-terminal) et 6 résidus (en C-terminal) qui sont considérées comme intrinsèquement désordonnées (J.-L. Darlix et al. 2011).

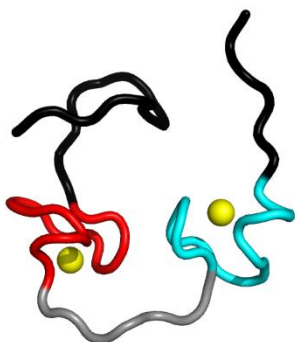


**Figure 27 – Structure et séquence de la NCp7.** A) Une des structures adoptée par la NCp7 libre ; l'image vient des structures publiées par Deshmukh et al. 2016, code PDB 5IIR. Les sphères jaunes représentent les ions Zn<sup>2+</sup> qui coordonnent les doigts de zinc. (B) Représentation de la séquence et de la structure secondaire de la NCp7.

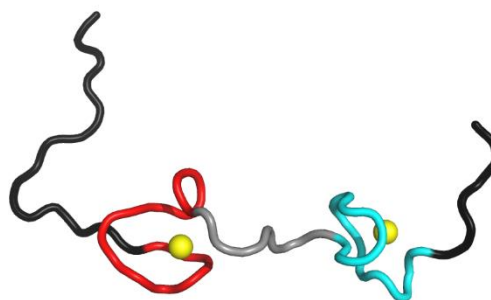
La séquence de la NCp7 est globalement très conservée, surtout sur la partie centrale composée des doigts de zinc et du *linker* (Dorfman et al. 1993; Gorelick et al. 1999). Bien que la queue N-terminale soit plus variable, ses résidus basiques sont strictement conservés par rapport aux autres résidus de la queue (Bourbigot et al. 2008).

Jusque récemment, le comportement du *linker* et le positionnement des doigts de zincs n'étaient pas clairement établis. Certains travaux montraient le rapprochement de ZF1 et ZF2 par le biais du résidu P31 du *linker*, de liaisons hydrogène et d'un plateau hydrophobe (Mély et al. 1994; Morellet et al. 1998). D'autres études proposaient une structure dites « ouverte » de la NCp7, sans interaction entre les doigts de zinc (Summers et al. 1992). C'est une étude récente (Deshmukh et al. 2016) qui combine l'analyse RMN avec de la diffusion des rayons X aux petits angles (SAXS pour l'anglais *Small Angle X-rays Scattering*) qui a permis de décrire le paysage conformationnel de la NCp7 libre et résoudre les contradictions. Les résultats indiquent que la NCp7 est dans un équilibre entre plusieurs conformations, dont des conformations « fermées » qui représentent environ 90% des structures explorées et des conformations « ouvertes » qui représentent les 10% restant (Figure 28). Ces deux types de conformations existent grâce à la dynamique du *linker* (Zargarian et al. 2014).

### Conformation fermée



### Conformation ouverte



**Figure 28 – Conformations fermée et ouverte de la NCp7.** ZF1 est représenté en rouge, ZF2 en cyan, les zincs coordinateurs en jaune, le linker en gris et les extrémités non-structurées en noir. Ces images ont été faites à partir des structures issues de Deshmukh et al. 2016, code PDB 5I1R.

La NCp7 libre présente donc une flexibilité dans le positionnement des doigts de zinc en accord avec les résultats RMN observés sur la dynamique du *linker* (Zargarian et al. 2014), à l'origine de ces mouvements.

### *Structure de la NCp7 complexée avec des acides nucléiques*

Actuellement, cinq structures de la NCp7 avec des acides nucléiques sont résolues au niveau atomique par RMN. Parmi ces structures, deux sont complexées avec un ARN (De Guzman et al. 1998; Amarasinghe et al. 2000) et trois avec un ADN (Bourbigot et al. 2008; Morellet et al. 1998; Bazzi et al. 2011).

Les principales interactions NCp7/AN qui ont été décrites consistent en des interactions électrostatiques non-spécifiques dues aux nombreux résidus basiques de la NCp7 et des empilements entre cycles aromatiques impliquant une base et le TRP37 du ZF2 et la PHE16 du ZF1 (Mély et al. 1993; Morellet et al. 1994; Lam et al. 1994; Bombarda et al. 1999; Avilov et al. 2009). ZF1 et ZF2 forment deux poches hydrophobes et, dans les complexes NCp7/ARN, une guanine s'insère dans chacune des deux poches (De Guzman et al. 1998; Amarasinghe et al. 2000) alors que dans les complexes NCp7/ADN, une seule guanine s'insère dans la poche formée par ZF2 (Bourbigot et al. 2008; Morellet et al. 1994; Bazzi et al. 2011).

Un autre point qui distingue les deux types d'AN est l'orientation de fixation de la NCp7. En effet, dans les complexes ARN, l'orientation de ZF1→ZF2 va du côté 5' de l'ARN vers son côté 3' alors que cette orientation est inversée pour les complexes ADN. Il a été suggéré que les sucres et des interactions hydrophobes associées étaient responsables de ces orientations contraires (Zargarian et al. 2009; Bazzi et al. 2011).

## Spécificité de la NCp7

La protéine NCp7 est capable de se fixer sur de très nombreuses séquences d'AN avec des affinités plus ou moins importantes (Fisher et al. 1998; Urbaneja et al. 1999) cependant, des études mettent en évidence des préférences pour certains sites de fixation (Mély et al. 1993; Bombarda et al. 1999; Wilkinson et al. 2008). La NCp7 est un très bon exemple de l'importance de la spécificité pour les protéines puisque ses différentes fonctions dans le cycle du virus nécessitent différents niveaux de spécificité.

Pour son rôle protecteur des ARN et pour faciliter l'hybridation de l'ARNt dans la transcription inverse, la NCp7 ne requiert pas un fort niveau de spécificité et va essentiellement interagir via ses résidus basiques à la surface des ARN sur environ 6 résidus (Mély et al. 1995).

Dans le cas de son rôle chaperon des AN, la NCp7 va être capable de se fixer, de déstabiliser un duplexe et de se retirer du partenaire AN. Pour jouer ce rôle, elle présente une nette préférence qui s'apparente à une spécificité pour des séquences ARN simple brin riche en A:U suivies d'un duplex dont l'extrémité 5' est une paire de base G:C (Wilkinson et al. 2008).

C'est dans le cadre de son interaction avec les tiges boucles SL1 et SL3 pour l'encapsidation de l'ARNg que la NCp7 présente la plus forte spécificité pour le motif GGXG, X étant un nucléotide quelconque (Wilkinson et al. 2008; J.-L. Darlix et al. 2014).

Mais si l'on s'intéresse à son affinité plus globale, l'ensemble des études structurales sur les complexes AN/NCp7 s'accordent à dire que les plus fortes affinités sont retrouvées pour les motifs d'AN simple brin TG, GXG et TXG, bien que le motif GXG soit plus représenté chez les ARN et le motif TG chez les ADN (J.-L. Darlix et al. 2014).



# Matériel et méthodes

## Analyse des interfaces par VLDM

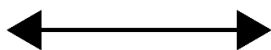
---

VLDM (Voronoi Laguerre Delaunay for Macromolecule) est un outil d'analyse des structures de macromolécules, protéine et/ou acide nucléique, basé sur une méthode géométrique puissante. VLDM est un outil central dans mes travaux de thèse puisque je l'ai utilisé pour l'analyse des interfaces du nucléosome (p. 64) mais également l'interface NCp7/NA (p. 165). VLDM est la version étendue de l'outil VLDP qui était uniquement dédié aux protéines (Voronoi Laguerre Delaunay Protein) (Esque, Oguey, et de Brevern 2010; 2011). VLDM, comme VLDP, est basé sur le pavage géométrique de l'espace euclidien. Il existe déjà de nombreuses applications de plusieurs sortes de pavages géométriques pour l'analyse structurale de molécules biologiques, notamment pour l'assignation des structures secondaires (Dupuis et al. 2005; T. Taylor et al. 2005), la détection de motifs structuraux (Kobayashi et Go 1997; Wako et Yamato 1998; W. R. Taylor 1999; Huan et al. 2004), la mesure volumétrique des cavités (Liang, Edelsbrunner, et Woodward 1998; Le Guilloux, Schmidtke, et Tuffery 2009), la quantification d'interfaces protéine-protéine (Bernauer et al. 2007) ou encore le calcul de l'accessibilité au solvant (Esque, Oguey, et de Brevern 2011). VLDP rassemble plusieurs de ces applications (mesure volumétrique, analyse et quantification des interactions protéine-protéine etc...) en un seul programme. VLDM rajoute la possibilité d'étudier des complexes formés par des protéines et des acides nucléiques.

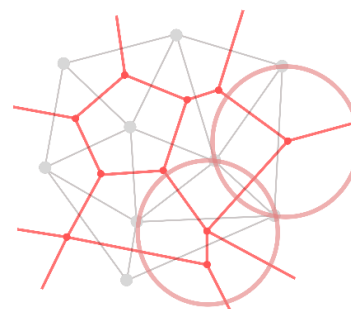
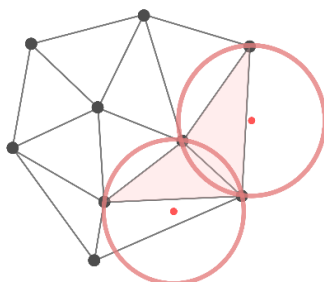
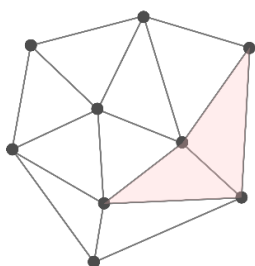
Dans cette section, je présente VLDM et je détaille ses nouvelles fonctionnalités et en quoi elles ont été centrales dans mes recherches ; je parle aussi brièvement du panel d'outils annexes que j'ai développé en particulier pour l'analyse de méga données (« big data ») issues de dynamiques moléculaires (DM) du nucléosome.

## Fonctionnement mathématique de VLDM : construction des pavages

Pavage de **Delauney**



Pavage de **Voronoi**

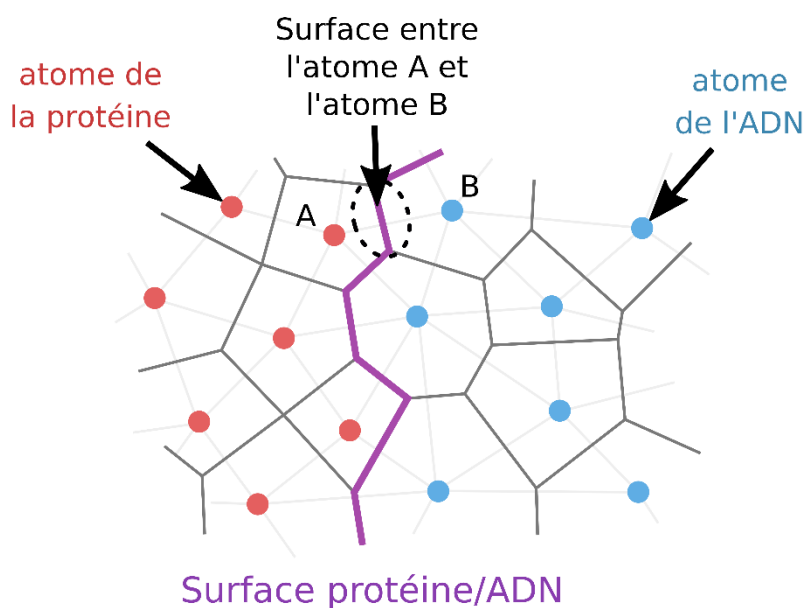


**Figure 29 - Représentation en deux dimensions des pavages de Delaunay et de Voronoï.** A gauche et au milieu, les sommets du pavage de Delaunay sont les positions atomiques du système (points gris). Au milieu, on construit les cercles circonscrits dont les centres (points rouges) sont les sommets des cellules de Voronoï. A droite, les arrêtes rouges du pavage de Voronoï représentent les lignes (surfaces en 3D) de contact entre deux atomes.

L'espace structural de la macromolécule est partitionné selon trois pavages. Le pavage de Delaunay prend pour sommets les centres des atomes (Figure 29). Pour définir les liens et former le pavage tétraédrique (sur la base de 4 atomes), Delaunay définit comme critère que la sphère circonscrite aux quatre atomes doit être vide, c'est à dire qu'elle ne doit pas contenir d'autres atomes. On construit ensuite un deuxième pavage, le pavage de Voronoï, dont les sommets sont les centres des sphères circonscrites de chacun des tétraèdres (Figure 29). Le pavage de Voronoï est composé de polyèdres appelées aussi cellules ; chaque cellule contient un atome et chaque côté définit une frontière entre deux atomes voisins. Enfin, le pavage de Laguerre est une variante pondérée du pavage de Voronoï. Chaque site atomique est alors associé à un poids de telle manière que les propriétés du pavage de Voronoï soient préservées. Il existe ainsi une relation de dualité entre partition Delaunay et partition Laguerre. Dans les applications bio-structurales, une série d'études (Poupon 2004; Esque, Oguey, et de Brevern 2011) ont montré qu'utiliser le carré du rayon de Van der Waals des atomes en tant que paramètre de poids reflète au mieux les données expérimentales de volumes et d'interface.

## Application de VLDM à l'analyse d'interfaces

Au cours de mes travaux, j'ai utilisé VLDM afin de caractériser des interfaces de différente nature entre des partenaires variés tels que les ADN, les ARN, les protéines ou encore le solvant. VLDM quantifie une surface de contact entre deux ensembles d'atomes par la somme de toutes les surfaces entre les atomes du premier ensemble avec ceux du second (Figure 30). Ces ensembles peuvent être des sous-unités protéiques, des chaînes latérales, des nucléotides, des bases, des sucres, etc... De plus, les surfaces se situant au niveau atomique, il est possible de décomposer les contacts par types. Cet outil permet donc une grande flexibilité dans l'analyse des contacts. Ici je détaille les trois types de contact, hydrophobes, électrostatique et proximité, qui nous ont permis de classer les données VLDM.



*Figure 30 – Représentation schématique d'une interface ADN/Protéine.*

### *Contacts Hydrophobes*

Les contacts hydrophobes (Contacts Hydrophobes, p.12) sont définis de manière très simple dans VLDM, il s'agit des contacts entre atomes de Carbone (C) exclusivement (C-C).

### *Contacts électrostatiques*

Dans VLDM, nous définissons les contacts comme électrostatiques (Contacts électrostatiques, p. 11) lorsqu'ils se produisent entre atomes d'azote (N) et d'oxygène (O). Un filtre permet d'éliminer les contacts répulsifs (positif contre positif et négatif contre négatif). Les surfaces issues de ces contacts sont toujours très faibles avec des surfaces inférieures à  $6\text{\AA}^2$  et ils ne représentent que 9 % de tous les contacts entre atomes d'azote (N) et d'oxygène (O).

## *Contacts de proximité*

Cette dernière catégorie de contact correspond à tous les contacts qui ne rentrent pas dans les classes décrites ci-dessus. Ces contacts reflètent les forces de van der Waals (Interactions de van der Waals, p. 12) des atomes qui ne sont engagés ni dans des interactions électrostatiques, ni dans des interactions hydrophobes.

## Intérêt de VLDM par rapport à d'autres approches

Je vais me limiter ici à parler des programmes qui existent pour analyser les contacts entre ADN et protéines. A l'heure actuelle, trois programmes sont proposés pour ce type d'analyse : PDBsum (Laskowski 2001), présenté par défaut dans la « Protein Data Bank » (PDB) où sont déposées les structures de macromolécules, PDIdb (Ferrada and Melo 2009) et DNAproDB (Sagendorf, Berman, and Rohs 2017), deux programmes uniquement spécialisés dans l'analyse d'interfaces NA/Protéine.

PDBsum utilise le programme HBplus (McDonald and Thornton 1994) pour qui deux atomes sont en contact s'ils sont distants de moins de 3.9Å. DNAproDB utilise également HBplus en combinaison avec l'outil SNAP inclus dans 3DNA (Lu and Olson 2008) mais pas encore décrit. Un forum indique que dans SNAP les interactions AN/protéine sont caractérisées grâce à des critères de distances et d'angles qui nécessitent un paramétrage manuel. Enfin, le programme PDIdb établit une définition complexe de l'interface, basée sur des angles entre atomes voisins contenus dans une sphère dite d'interaction. Mais de nouveau, cet outil dépend de deux critères subjectifs, l'angle entre atomes et la taille de la sphère d'interaction.

Selon ces outils, une même interface a donc plusieurs définitions, toutes en partie subjectives. En se basant uniquement sur la disposition des atomes les uns par rapport aux autres, VLDM permet de définir un contact en s'affranchissant de tout ajustement et paramètre manuel. De plus, PDBsum, PDIdb et DNAproDB ne fournissent qu'une liste d'atomes en contact mais pas de quantification d'interfaces, alors que VLDM donne l'aire de ces interfaces. Dans le contexte actuel, VLDM s'inscrit comme un outil performant pour l'analyse, la compréhension et la comparaison d'interface de macromolécules.

Un aspect que VLDM n'aborde pas directement est la quantification de l'énergie des interactions. Actuellement, VLDM ne fait pas de différence énergétique entre une interaction de type électrostatique ou Van der Waals. Cependant, la disposition tridimensionnelle des atomes d'une structure, qu'elle soit issue de la cristallographie, de la RMN ou encore d'une simulation de dynamique moléculaire, est la résultante des forces et des mouvements thermiques. On peut donc considérer que cette disposition prend déjà en compte implicitement ces interactions (une attraction forte va rapprocher les atomes et aura une répercussion sur la géométrie qui sera donc prise en compte par VLDM).

Il serait cependant intéressant dans l'avenir de compléter le programme VLDM avec des mesures d'énergie. Une approche qui semblerait à premier abord intéressante est la méthode MM/PBSA (Genheden et Ryde 2015) qui permet une mesure de l'énergie libre entre deux partenaires avec un bon compromis entre le temps de calcul et la précision de la mesure d'énergie. Cependant, la méthode contient une limitation importante dans le cas des complexes protéines/AN qui vient du fait qu'il faut mesurer l'énergie des partenaires seuls, puis des partenaires complexés. Or, dans le cas du nucléosome par exemple, il n'est pour l'instant pas réaliste d'obtenir la structure d'un ADN libre d'environ 146 pdb et l'analyse MM/PBSA est donc compromise. L'alternative serait de calculer l'énergie potentielle entre chaque paire d'atome en utilisant des paramètres de champ de force. Cette méthode est très intéressante mais comporte une autre limitation. En effet, il a été montré au laboratoire que pour des oligonucléotides issus de la séquence 601, l'image du paysage énergétique exploré par des petits ADN dans des simulations de dynamique moléculaire est très dépendante du champ de force utilisé et qu'il est difficile de savoir quelle paramétrisation est la plus réaliste si on ne dispose pas de données expérimentales (Ben Imeddourene et al. 2015). Mais on pourrait envisager de proposer un choix de champs de force assez large à intégrer à VLDM afin d'associer aux surfaces une nouvelle fonctionnalité de mesure d'énergie potentielle.

## VLDM : Problématique des macromolécules

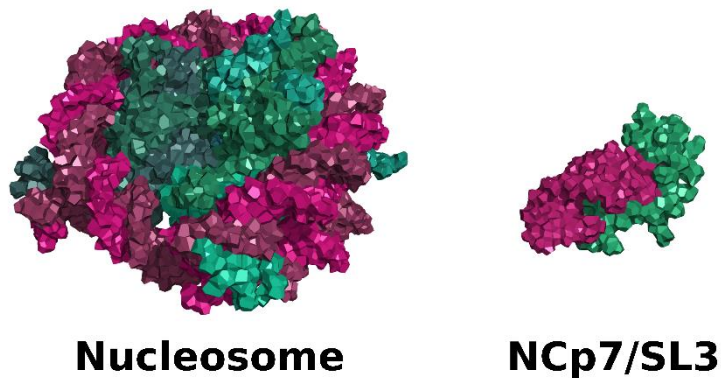
L'étude des macromolécules par VLDM, et dans notre cas d'un gros complexe ADN/protéine, a nécessité des ajustements informatiques.

Dans un premier temps, le programme VLDP a été adapté afin de pouvoir prendre en compte les acides nucléiques. Dans le cas du nucléosome, nous traitons plus d'une microseconde de simulation d'un complexe composé de plusieurs dizaines de milliers d'atomes. Christophe Oguey a donc optimisé la place mémoire de VLDM qui est très sollicitée en raison de la taille du complexe.

J'ai ensuite travaillé avec Christophe Oguey afin de définir des entrées et sorties les mieux adaptées aux macromolécules et faciliter l'étape d'analyse des données. VLDM traite un échantillon de structures instantanées une par une et fournit une liste de tous les contacts entre paires d'atome avec leurs surfaces associées. A la suite d'Ahmad Elbahnsi, j'ai développé un ensemble de script *awk* et *R* pour :

- Isoler les différents ensembles qui caractérisent les interfaces (AN, protéine, solvant, chaînes, etc.)
- Agréger les surfaces de contact calculées pour tous les atomes au niveau des sous-unités, des peptides ou encore des chaînes latérales pour les protéines ou des nucléotides, bases, sucres et phosphates pour les acides nucléiques
- Effectuer ces analyses sur chaque structure extraite d'une DM et générer un fichier de sortie clair et organisé de l'interface souhaitées.

Il est donc possible avec ces scripts de traiter tout type de données VLDM issues de trajectoires de DM ou de structures uniques comme celles issues de la diffraction des rayons X. J'ai d'ailleurs eu l'occasion d'appliquer VLDM sur deux systèmes très différents (Figure 31) : 1) le nucléosome étudié en DM tout atome, c'est-à-dire un très large système et un nombre très conséquent de structures ; 2) complexes NCp7/NA dérivés de données RMN, c'est-à-dire un petit système et un nombre limité de structures.

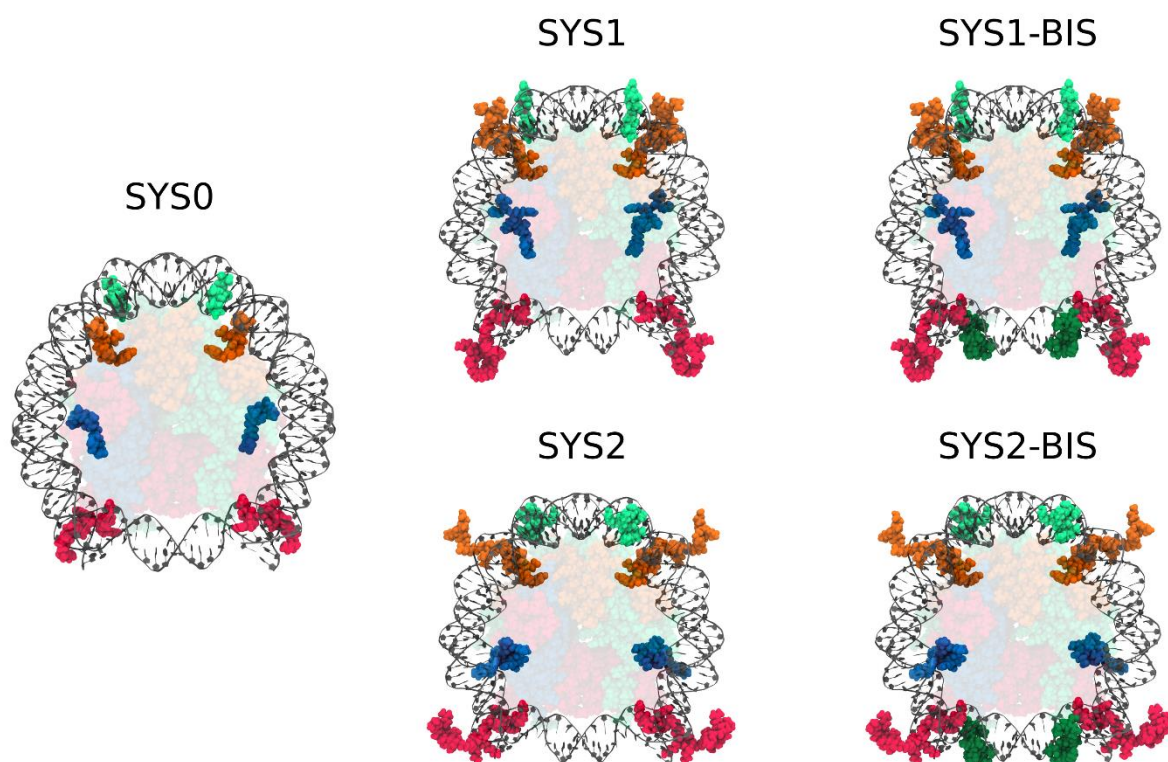


*Figure 31 – Représentation de la structure du nucléosome et du complexe NCp7/SL3 en cellule de Voronoï. A gauche, une structure du nucléosome avec l'ADN en violet et les histones en vert. A droite, une structure de la NCp7 en vert et son partenaire ARN en violet.*

## Dynamique moléculaire du nucléosome

Une partie de mes travaux sur le nucléosome est basée sur l'analyse de cinq DM du nucléosome, réalisées par Ahmad Elbahnsi. Je vais détailler ici les spécificités de chacun des systèmes ainsi que les paramètres de production des DM.

### Structures de départ



**Figure 32 – Structures de départ des dynamiques moléculaires.** L'ADN est représenté par un ruban gris. Les histones sont représentées par des sphères de Van der Waals transparentes pour le cœur structuré et opaques pour les queues. Les histones H3 sont représentées en couleur rouge, H4 en bleu, H2A en vert et H2B en orange. Les queues C-ter de H2A sont représentées en vert foncées et les queues N-ter en vert clair.

Cinq structures de départ, une pour chaque DM, ont été générées à partir de deux structures cristallographiques. Le premier système, SYS0 est composé de la structure cristallographique 3MVD (Makde et al. 2010) uniquement, comprenant essentiellement le cœur structuré des histones issue de *Xenopus laevis* et la séquence 601 composée de 146 paires de bases. Les quatre autres modèles, SYS1, SYS1-bis, SYS2 et SYS2-bis, diffèrent par l'ajout de domaines N-terminal et C-terminal d'histones, de tailles et conformations variées. Ces domaines non structurés, appelés queues, sont issues de la



structure cristallographique 1KX5 (Davey et al. 2002). 1KX5 est la seule structure pour laquelle les queues d’histones ont pu être résolues, même imparfaitement. Dans notre cas, elle n’a pas pu être utilisée en point de départ parce qu’elle est formée avec l’ADN issu du satellite humain  $\alpha$ , qui, en solution, a une affinité médiocre pour les histones. Le tableau (Article 1 – Tableau SD1) et la Figure 32 résument les différences entre ces systèmes.

## Protocole de production des dynamiques moléculaires

En accord avec les résultats expérimentaux de RMN sur les ADN-B libre (Ben Imeddourene et al. 2015), c’est le champ de force CHARMM36 (Best et al. 2012) avec la correction CMAP (Mackerell, Feig, et Brooks 2004) qui a été retenu pour la production des DM. Les calculs ont été fait à l’aide des programmes CHARMM (B. R. Brooks et al. 2009) et NAMD 2.11 (Phillips et al. 2005).

Le soluté est contenu dans une boîte cubique dont la distance minimale entre les bords de la boîte et le soluté est de 10Å. La boîte est remplie de molécules d’eau TIP3P (Jorgensen et Madura 1983). Des ions  $\text{Na}^+$  et  $\text{Cl}^-$  sont ensuite ajoutés pour assurer la neutralité électrique du système et atteindre une concentration de 150 mM, en accord avec les conditions habituelles des expériences *in vitro* sur le nucléosome. Le placement initial des ions se fait selon un algorithme de *Monte Carlo*.

A la fin de l’ajout des molécules d’eau et des ions, les positions atomiques du soluté sont contraintes et deux minimisations, « *steepest descent* » puis « *Adapted Basis Newton-Raphson* », sont effectuées. Ensuite, une première phase d’équilibrage du système est effectuée sous NAMD. Le système est chauffé progressivement par pas de 10K tous les 0.1ps pour atteindre 300K. Pendant cette étape, des contraintes harmoniques sont appliquées sur les positions atomiques du squelette protéique et de l’ADN avec des constantes de force respectives de 1000 et 5000 kcal/mol/Å<sup>2</sup>. Ces contraintes sont ensuite progressivement relâchées au cours d’une seconde phase d’équilibrage d’une nanoseconde dans un ensemble NPT.

**Tableau 2 – Récapitulatif des durées des simulations de dynamique moléculaire pour chacun des systèmes.**

| Système    | SYS0 | SYS1 | SYS1-BIS | SYS2 | SYS2-BIS |
|------------|------|------|----------|------|----------|
| Durée (ns) | 200  | 200  | 300      | 200  | 300      |



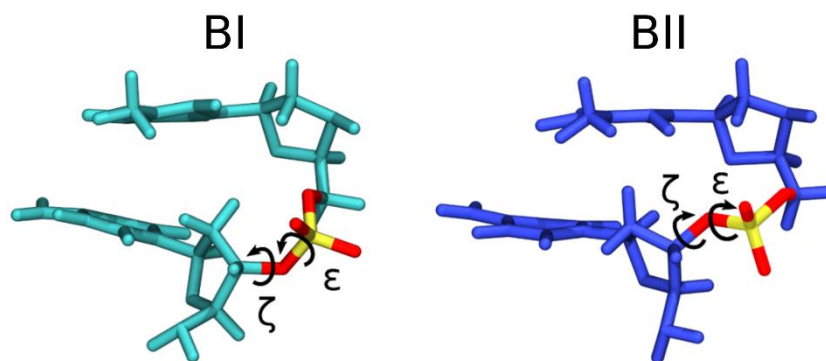
La phase de production est également effectuée dans le même ensemble NPT, à une température de 300K et sous une pression de 1bar. Le thermostat suit l'équation de Langevin (C. L. Brooks et Karplus 1983; Feller et al. 1995) ainsi des conditions périodiques sont appliquées aux limites. Les interactions Coulombiennes à l'infini sont traitées par la sommation d'Ewald (Darden, York, et Pedersen 1993). Pour les interactions de Van der Waals, un algorithme de type « *switch* » est appliqué au-delà de 10 à 12Å. Le pas d'intégration est de 2 fs ; l'algorithme SHAKE contraint les longueurs de liaisons impliquant les atomes d'hydrogène.

## Analyse des acides nucléiques – Curves+

Lors de mon travail, j'ai été amené à étudier de manière approfondie les acides nucléiques et plus particulièrement l'ADN B. Nous avons notamment mis en évidence des caractéristiques propres à l'ADN nucléosomal concernant les conformations des phosphates, les paramètres hélicoïdaux inter-paire de base de l'ADN et les dimensions des sillons (p. 129). Ces analyses structurales ont été réalisées avec le programme Curves+ (Blanchet et al. 2011).

### Conformation des phosphates

La conformation des phosphates est définie ici par les deux angles dièdres  $\epsilon$  et  $\zeta$  dont il a été montré très tôt qu'ils présentent une distribution bimodale dans les structures cristallographiques d'ADN-B

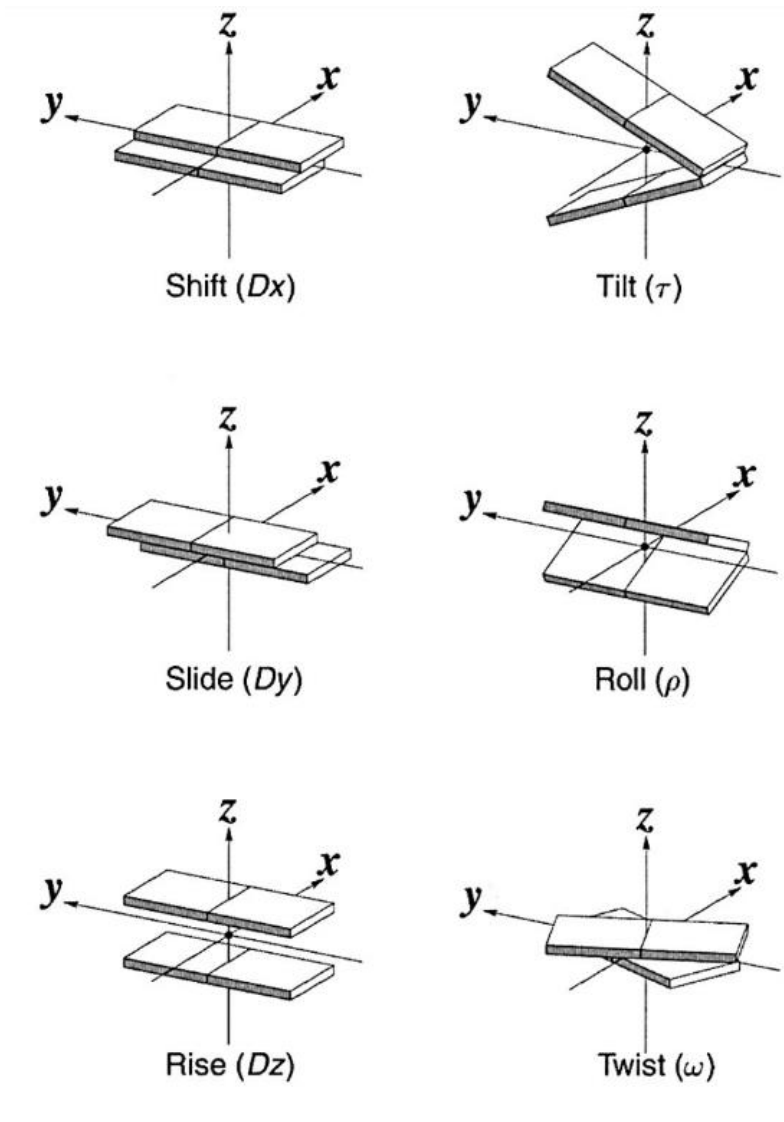


**Figure 33 – Conformation BI et BII des phosphates de l'ADN B.** Représentation de deux dinucléotides où les groupements phosphates sont en BI (à gauche) ou en BII (à droite). L'atome de phosphore du groupement phosphate est représenté en jaune et les oxygènes du groupement phosphate en rouge.

(Schneider, Neidle, et Berman 1997; Berman 1997; Djuranovic et Hartmann 2003; 2004). On nomme BI la conformation du phosphate où  $\epsilon$  et  $\zeta$  sont respectivement en *trans* et *g*- et BII la conformation où  $\epsilon$  et  $\zeta$  sont respectivement *g*- et *trans* (Fratini et al. 1982). La Figure 33 présente des groupements phosphates en conformation BI et BII.

### Paramètres hélicoïdaux inter-paire de base

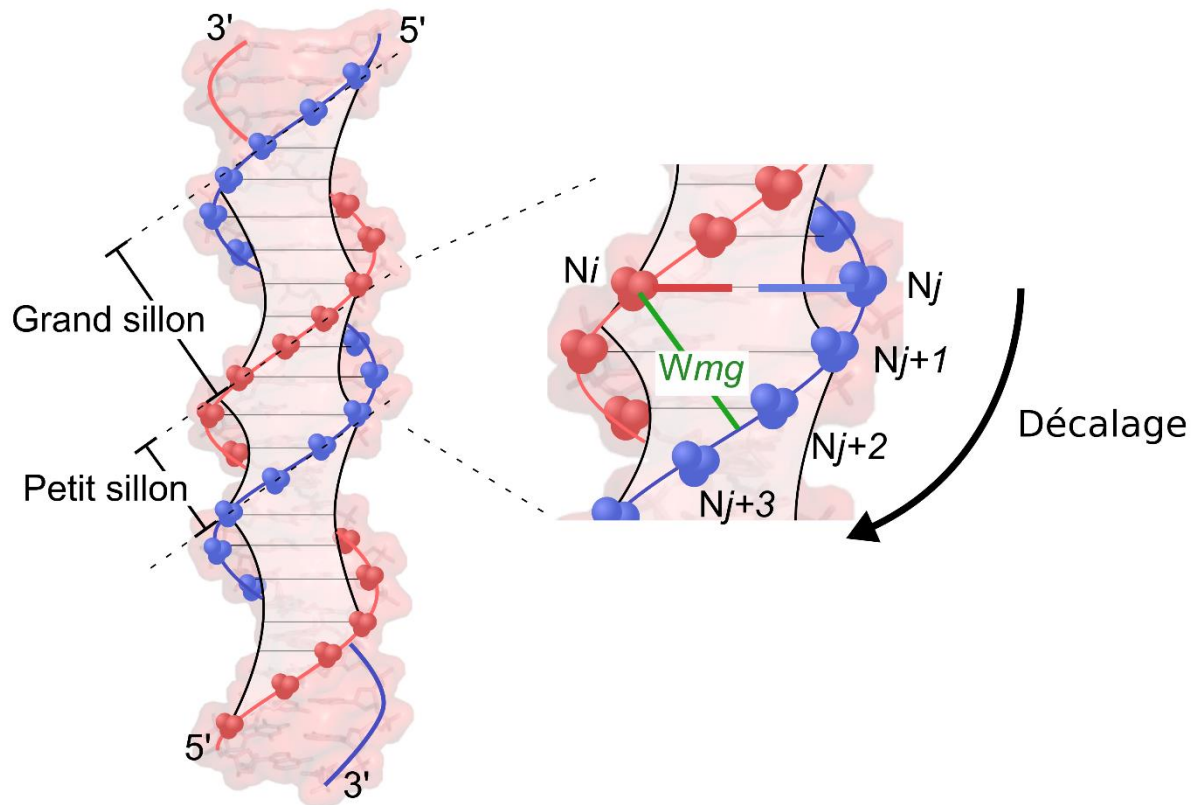
Les paramètres hélicoïdaux inter-paire de base décrivent les 3 translations et les 3 rotations entre deux paires de base successives.



*Figure 34 – Paramètres hélicoïdaux inter-paire de base de l'ADN (Lu et Olson, 2008)*

## Sillons de l'ADN

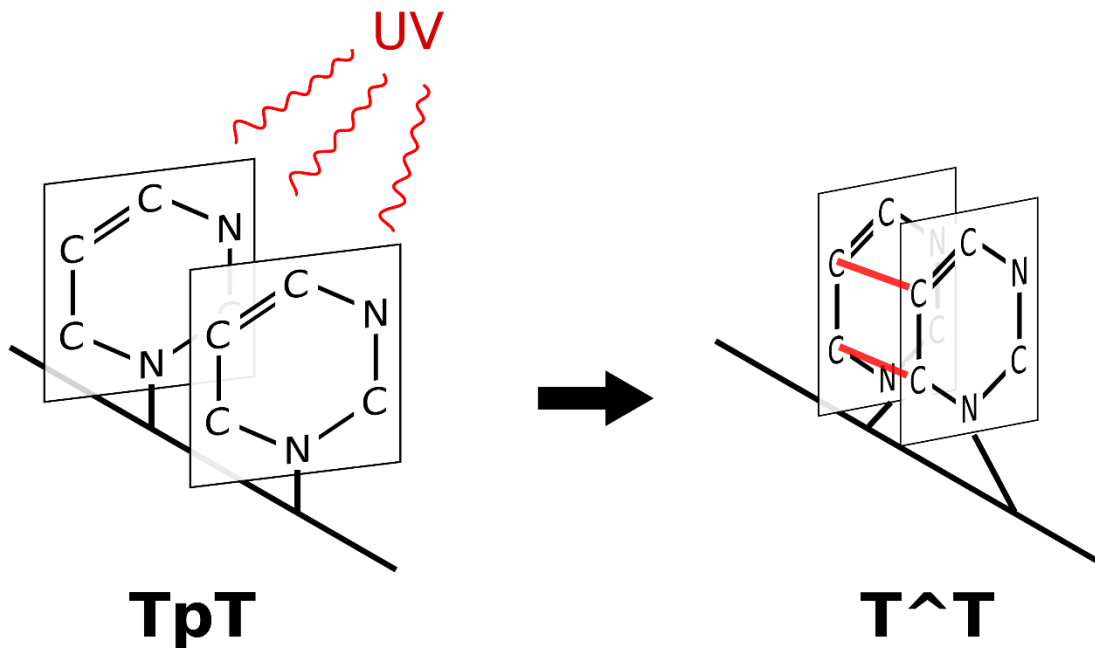
L'ADN B est caractérisé par sa conformation en double hélice et ses deux sillons inégaux, le grand et le petit sillon (Figure 35). Nous allons voir que la largeur des petits sillons joue un rôle important dans les interactions avec les histones (Chapitre 3 : Dynamique de l'ADN dans le nucléosome, p. 129). La mesure de la largeur du petit sillon s'effectue dans Curves+ entre deux points des brins I et J qui ne correspondent pas à une paire de base, mais à deux bases décalées en moyenne de 2,5 nucléotides (Figure 35).



**Figure 35 – Représentation schématique du grand et du petit sillon de l'ADN.**  $N_i$  est le nucléotide à la position  $i$  sur le brin I et  $N_j$  le nucléotide du brin J qui lui est apparié.  $W_{mg}$  correspond à la largeur du petit sillon.

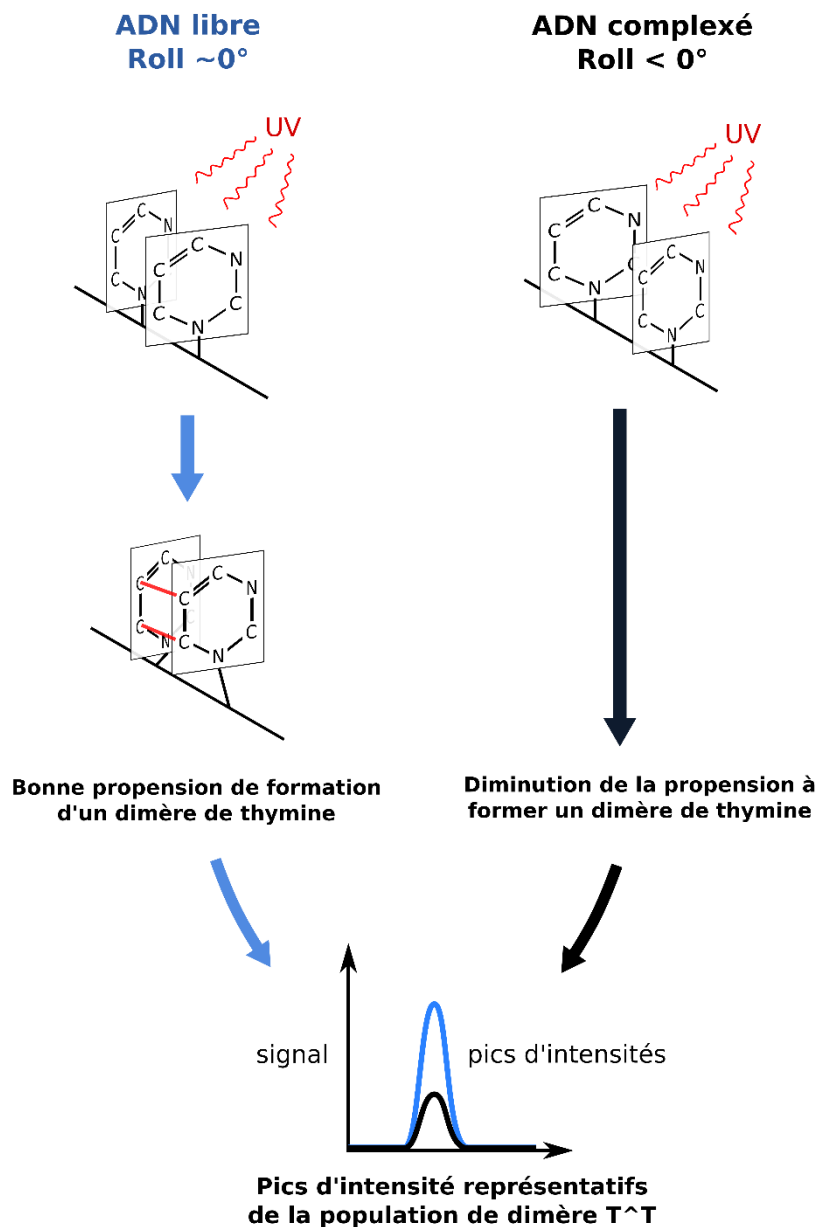
## Méthode PhAST

La méthode PhAST (*Photochemical Analysis of Structural Transitions*) a été initialement mise au point par Malcolm Buckle et son groupe (Hatakeyama et al. 2016). C'est une approche expérimentale non-invasive et qui ne requiert aucune modification chimique des macromolécules que l'on veut étudier, contrairement à des techniques de fluorescence par exemple. Elle nous a permis de faire une analyse précise de l'assemblage et du désassemblage du nucléosome dans les mêmes conditions.



**Figure 36 – Représentation schématique de la formation d'un dimère de thymine  $T^T$ .** A gauche, deux thymines qui se suivent (TpT) sur un brin d'ADN sont soumises à un rayonnement UV et se dimérisent (droite) selon un rendement quantique qui reste faible.

PhAST consiste à mesurer la probabilité de formation d'un dimère de pyrimidine entre deux pyrimidines adjacentes (YpY) localisées sur le même brin d'ADN après irradiation UV (Figure 36). La probabilité de former un dimère est modulée par la nature chimique des pyrimidines (les cytosines ont un plus faible rendement que les thymines) et par la structure locale de l'ADN. Un dimère YpY étant formé grâce à deux liaisons C5-C5 et C6-C6, tout élément de structure rapprochant ces deux atomes favorise l'apparition d'un dimère. Un roll positif, associé à un petit twist augmente ainsi la probabilité d'obtenir un dimère (Hatakeyama et al. 2016).



**Figure 37 – Exemple d’une mesure PhAST sur un pas TpT.** La complexation de l’ADN avec une protéine entraîne un changement dans la structure locale de l’ADN (panneau du haut). Dans cet exemple, les positions relatives des deux thymines successives sont plus favorables à la formation de dimère dans l’ADN libre (roll ~0) que dans l’ADN lié (roll < 0) ; la fixation de la protéine s’accompagne d’une diminution du pic d’intensité représentatif de la population T^T.

Afin d'analyser le comportement de l'ADN lors de l'assemblage et du désassemblage du nucléosome, deux expériences sont faites en parallèle, une sur l'ADN complexé et une sur l'ADN libre (Figure 37). Une fois l'irradiation UV effectuée, on associe à chaque pas YpY un signal dont l'intensité reflète la population du dimère Y<sup>^</sup>Y. Si l'intensité des pics pour l'ADN libre et pour l'ADN complexé est identique, alors la structure de l'ADN complexé est comparable à celle de l'ADN libre. Des intensités différentes témoignent d'un changement de structure locale (par exemple, un *roll* plus faible) que, dans le cas du nucléosome nous avons relié aux interactions avec une histone.

Dans la pratique, on calcule le  $\log_2$  du ratio de l'intensité du pic de l'ADN complexé sur l'intensité du pic de l'ADN libre. Ainsi, lorsque la propension à former un dimère de pyrimidine diminue dans l'ADN complexé, on observe un  $\log_2(IR)$  négatif. Au contraire, un  $\log_2(IR)$  positif témoigne d'une augmentation de la propension à former un dimère de pyrimidine dans l'ADN complexé. Lorsque la structure de l'ADN complexé se rapproche de celle de l'ADN libre, le  $\log_2(IR)$  tend vers 0.

# Chapitre 1 : Interface ADN/Histones dans le nucléosome

## Sommaire <sup>1</sup>

Afin de jouer son rôle de compaction de l'ADN dans les cellules eucaryotes, le complexe nucléosomal doit être capable de se former avec des séquences d'ADN de composition variable. Mais dans tous les cas, il impose une courbure importante à la double hélice. Cette courbure régulière, d'environ 30° pour dix paires de bases, n'est pas observée dans l'ADN libre, quelle que soit la séquence. Les histones doivent donc induire puis maintenir l'ADN dans un état courbé et donc stressé.

L'organisation globale de cette interface a été résolue avec les premières structures cristallographiques du nucléosome. L'interface est décrite par 14 sites de fixation, un toutes les 10 paires de base le long de l'ADN (Article 1 – Figure 1A, 1B). Chacun de ces points de contact est caractérisé par une arginine interagissant avec le petit sillon de l'ADN et l'interface est donc considérée comme ayant un fort caractère électrostatique. Mais ces interactions paraissent peu de chose au regard des distorsions de l'ADN qui doivent être préservées. C'est pourquoi il a aussi été proposé que de nombreuses liaisons hydrogène médiées par des molécules d'eau renforçaient les contacts directs entre ADN et histone. Cependant, ces descriptions de l'interface proviennent d'analyses limitées aux liaisons hydrogène, utilisant en particulier HBplus, et de par la nature cristallographique des complexes, ne peuvent pas donner d'informations sur la dynamique et la stabilité des contacts. Enfin, elles ne prennent en compte que les domaines structurés des histones, les queues étant très mal voire pas du tout résolues en cristallographie.

Le but de ce premier travail était de décrire une interface ADN-histone en solution, qui pourrait servir de référence. Pour cela, nous avons produit 1.2µs de simulations de dynamique moléculaire du nucléosome en solution explicite, c'est-à-dire dans un solvant à 0.15M NaCl, avec les modèles et selon le protocole décrit dans la section dynamique moléculaire du nucléosome (p. 55). Ce nucléosome est composé d'histones issues de *Xenopus laevis* ainsi que de la séquence d'ADN dite 601. Cette séquence a été choisie à cause de sa forte affinité pour les histones qui suggère une interface particulièrement solide. De plus, son caractère fortement positionnant la rend très pratique à utiliser lors d'expériences, comme celles menées par M. Buckle et C. Nogues que je décrirai plus tard.

---

<sup>1</sup> L'ensembles des références de cette section peuvent-être trouvées dans l'article associé au chapitre (p. 72) et dans l'introduction de ce manuscrit (Le nucléosome, p. 17)



Nous avons analysé les simulations avec le programme VLDM, qui permet de caractériser chaque interaction ADN/protéine en termes de surface et d'occurrence de contact. Dans la première partie de l'article, nous avons répertorié l'ensemble des acides aminés et des nucléotides impliqués dans l'interface. Nous relevons que les contacts sont très stables au cours de la dynamique à l'exception de ceux impliquant les parties N-terminales des queues d'histones et les extrémités de l'ADN. En regroupant les contacts par SHL on retrouve une interface très symétrique par rapport au centre de l'ADN, que ce soit pour le cœur (Article 1 – Figure 4) ou les racines des queues d'histone (Article 1 – Figure 6). Puisque les deux moitiés de la séquence 601 ne sont pas symétriques (non palindromiques), l'effet de la séquence de l'ADN semble donc négligeable sur l'interface. Nous soulignons l'importance des contacts hydrophobes, souvent négligés dans la littérature mais dont la contribution en termes de surface est équivalente voire supérieure aux contacts électrostatiques.

Nous avons montré que contrairement à ce que montraient les études cristallographiques, aucune molécule d'eau n'est piégée de manière systématique et durable au niveau de l'interface ADN/Histone. Les liaisons hydrogène entre l'eau et le nucléosome sont donc transitoires et n'ont donc pas un rôle clé dans le maintien du complexe.

Nous avons ensuite analysé les densités de cations ( $\text{Na}^+$ ) à l'interface ADN/histones mais également autour du nucléosome (Article 1 – Figure 8). Bien que certaines positions de l'interface ADN/histones semblent favoriser la présence d'un ion, nous ne retrouvons pas d'ions piégés systématiquement dans toutes les dynamiques. En revanche, nous avons mis en évidence la présence d'un nuage de cations  $\text{Na}^+$  entre les deux doubles hélices de l'ADN qui sont superposées dans le nucléosome ; ce nuage permet de faire bouclier entre les groupements phosphates très proches et donc hautement répulsifs.

# Holding the Nucleosome Together: A Quantitative Description of the DNA–Histone Interface in Solution

Ahmad Elbahnsi,<sup>†,‡,⊥</sup> Romain Retureau,<sup>†,⊥</sup> Marc Baaden,<sup>§</sup> Brigitte Hartmann,<sup>\*,†</sup> and Christophe Oguey<sup>\*,‡</sup>

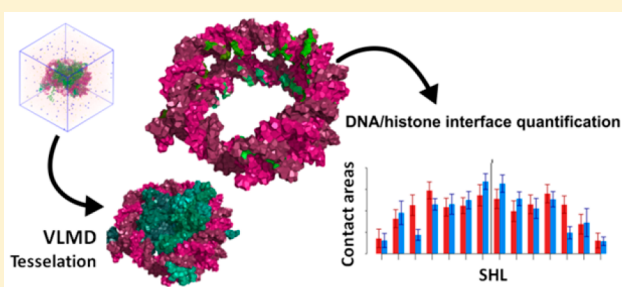
<sup>†</sup>LBPA, UMR 8113, ENS Paris-Saclay - CNRS, 61 avenue du Président Wilson, 94235 cedex Cachan, France

<sup>‡</sup>LPTM, UMR 8089, CNRS, Université de Cergy-Pontoise, 2 avenue Adolphe Chauvin, 95302 Cergy-Pontoise, France

<sup>§</sup>Laboratoire de Biochimie Théorique, CNRS, UPR9080, Université Paris Diderot, Sorbonne Paris Cité, 13 rue Pierre et Marie Curie, 75005 Paris, France

## Supporting Information

**ABSTRACT:** The nucleosome is the fundamental unit of eukaryotic genome packaging in the chromatin. In this complex, the DNA wraps around eight histone proteins to form a superhelical double helix. The resulting bending, stronger than anything observed in free DNA, raises the question of how such a distortion is stabilized by the proteic and solvent environments. In this work, the DNA–histone interface in solution was exhaustively analyzed from nucleosome structures generated by molecular dynamics. An original Voronoi tessellation technique, measuring the topology of interacting elements without any empirical or subjective adjustment, was used to characterize the interface in terms of contact area and occurrence. Our results revealed an interface more robust than previously known, combining extensive, long-lived nonelectrostatic and electrostatic interactions between DNA and both structured and unstructured histone regions. Cation accumulation makes the proximity of juxtaposed DNA gyres in the superhelix possible by shielding the strong electrostatic repulsion of the charged phosphate groups. Overall, this study provides new insights on the nucleosome cohesion, explaining how DNA distortions can be maintained in a nucleoprotein complex.



## INTRODUCTION

The nucleosome is the fundamental building block of packaged DNA in eukaryotic cells. It consists of an octameric histone core, including two copies of histones H3, H4, H2A, and H2B, around which 145–147 base-pairs (bp) of DNA are wrapped making 1.65 turns of a left-handed superhelix.<sup>1–3</sup> The term “gyre” is commonly used to denote a superhelix turn around the histone core, to distinguish it from turns of the DNA double-helix around its own axis. As molecular length, a gyre represents approximately 7 double helix turns. Each histone comprises a folded domain (structured core) composed of 3  $\alpha$ -helices connected by short loops and an unstructured, flexible N-terminal domain called a tail.<sup>4,5</sup> Moreover, H2A has a C-terminal tail. The two copies of each histone type are arranged symmetrically with respect to the dyad axis, which passes through the center of nucleosomal DNA.

*In vivo*, a plethora of interacting factors<sup>6</sup> such as chaperones,<sup>7,8</sup> remodeling complexes,<sup>9,10</sup> histone variants,<sup>11,12</sup> epigenetic modifications,<sup>13,14</sup> and intrinsic, sequence dependent DNA properties<sup>6,15–27</sup> orchestrate the nucleosome positioning. Indeed, eukaryotic DNA transactions are intimately associated with the spatial and temporal distribution of nucleosomes within a specific genome, which regulates the DNA

accessibility. Thus, assembly and disassembly of nucleosomes continuously occur in the nucleus.

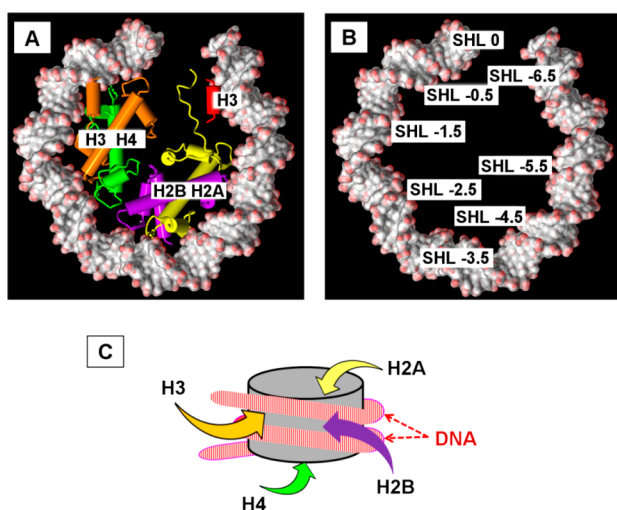
In an attempt to better understand the physical and structural events underlying the disassembly mechanism, earlier studies examined the nucleosome behavior in response to an increased ionic strength, which concluded that the H2A–H2B dimers were released before the H3–H4 tetramer.<sup>28,29</sup> This global dissociation scheme, which may reproduce what occurs in the nucleus,<sup>30–32</sup> was confirmed by *in vitro* experiments using short nucleosome arrays reconstituted with various positioning DNA sequences, the Widom’s 601 sequence,<sup>31–36</sup> and sequences from sea urchin 5S ribosomal gene (sequence 5S)<sup>32,37–40</sup> or GUB.<sup>41</sup> Time-lapse atomic force microscopy imaging,<sup>36</sup> time-resolved small-angle X-ray scattering,<sup>32,33</sup> single-pair fluorescence resonance energy transfer,<sup>31,34,41</sup> or optical tweezers<sup>40,42</sup> results are compatible with a disassembly first stage facilitated by transient unwrapping of one or both peripheral DNA regions; also, a DNA region in the vicinity of the dyad is the major barrier in the course of the nucleosome dissociation, but its length remains debated. Two studies of unzipping 601 and arbitrary DNA sequences in single

Received: September 5, 2017

Published: December 20, 2017

nucleosomes<sup>34,42</sup> described a multistep process and postulated that, in addition to the DNA center, two regions resisting the dissociation are symmetrically located around 40–50 bp from the dyad.

Elucidating the subtleties of nucleosome dissociation may be a pointless exercise without a precise knowledge of the DNA–histone interactions. The global organization of this interface emerged from the first high resolution X-ray structure of a nucleosome containing a DNA from the human  $\alpha$ -satellite.<sup>1–3</sup> This analysis revealed that the DNA–histone interface was subdivided into 14 main sites (seven sites per half particle) spaced by  $\sim 10$  bp from center to center along the DNA (Figure 1A,B).



**Figure 1.** Overviews of the nucleosome. Panel A is an upper view of the structure of the nucleosomal DNA 5' half (in gray, with phosphate groups in red) bound to the histone structured cores of H3 (in orange), H4 (in green), H2A (in yellow), and H2B (in purple). The isolated helix in red belongs to the second H3 copy that contacts the DNA extremity. In panel B, the SHL numbering of DNA regions interacting with the histone structured cores is shown on the same DNA structure as in panel A; SHL specifies the number of double-helix turns between a given region and the DNA center (SHL0). The structures in panels A and B were derived from the model used in the SYS1 simulation presented here, removing the histone tails for clarity. Panel C provides a schematic view of the nucleosome displaying the location of one copy of each histone N-tail, either on the upper or lower face of the nucleosome (H4 and H2A) or passing through the interstice between the two superimposed DNA gyres (H3 and H2B).

The interface is punctuated by arginines interacting at each DNA contact point, which contact the minor groove. These arginines are considered to be determinant for stabilizing the nucleosome, especially when their side chains penetrate the minor groove, making electrostatic and hydrophobic interactions with sugar atoms.<sup>43</sup> Owing to well ordered DNA, histones, and solvent components, another human  $\alpha$ -satellite containing structure (1KX5) solved to 1.9 Å resolution, led to the suggestion that the limited collection of direct hydrogen bonds between DNA phosphate groups and arginines or lysines was supplemented by water molecules making interactions simultaneously with histone and DNA.<sup>44</sup> In the same structure, specific ion-binding sites in histones and DNA were also identified, but they mainly contribute to nucleosome–nucleosome interactions and not to the internal interface. A network of interactions, known as “sugar clamp”, emerged from

a structure of a 601 containing nucleosome.<sup>45</sup> This network surrounds two TTAAA elements, symmetrically located at  $\pm 1.5$  turn of the DNA center and combines electrostatic and hydrophobic contacts between DNA phosphate groups and sugars on one hand and H3-arginine, H3-leucine, and H4-proline on the other hand. Described as unique in the nucleosome,<sup>45</sup> the sugar clamp was suspected to promote the positioning power of these TTAAA elements.<sup>46</sup> Apart from the minor groove inserted arginines and the sugar clamp, it is commonly accepted that electrostatic interactions, comprising those mediated by water molecules, are the major elements responsible for maintaining the DNA wrapped around the histones.<sup>43,47–49</sup> However, a previous exhaustive, careful analysis of protein–DNA X-ray structures highlighted the central importance of nonelectrostatic contacts for complex formation and stability.<sup>50</sup> In the analyses of the DNA–histone interface, such interactions, only sparsely described, clearly deserve to be better investigated and quantified.

Also, despite their essential contribution, X-ray structure analyses cannot answer all questions. The well-known biases due to crystal growth conditions or lattice contacts, which are the major driving force for the formation of nucleosome crystals,<sup>24,45,51</sup> are in fact not the only issue. First, 1KX5<sup>44</sup> is the only crystal to determine the structure of the H3 and H2B N-tails that pass between the two DNA double helices juxtaposed one above the other in the nucleosome as well as of the H4 and H2A tails that enter or exit on the upper or lower sides of the nucleosome (Figure 1C). But even in this case, the tails were globally too poorly ordered to provide a precise map of their contacts with the DNA.<sup>44</sup> Yet, modeling studies mentioned a substantial contribution of certain tail parts to the interface,<sup>52,53</sup> although another study<sup>54</sup> suggested that the interactions with histone tails cannot explain the barriers described from experimental disassembly results. Second, the DNA superhelical wrapping brings phosphate groups of juxtaposed DNA gyres very close to each other. Because of electrostatic repulsion, such a proximity would not be possible without efficient shielding. H3 and H2B tails undoubtedly reduce this repulsion in surrounding DNA regions. Still, what happens in regions where the juxtaposed DNA gyres are not separated by histone tails has not been elucidated.

Thus, our understanding of how the DNA is maintained by the histones remains incomplete. We therefore decided to study the nucleosome from all-atom molecular dynamics simulations with explicit water molecules and ions, focusing more on the DNA–histone interface in solution than on the nucleosome dynamics. To get a reference of a strong interface, we studied a nucleosome formed with the artificial sequence 601, known to generate a very stable complex.<sup>17,55</sup> In addition, our nucleosome models contain a large part of the histone tails. For the interface analysis, we used a tessellation method, called VLDM (Voronoi Laguerre Delaunay for Macromolecules) initially developed for proteins<sup>56–58</sup> and adapted here to nucleic acids. Based on a representation of molecules by a collection of polyhedra filling space without overlaps or gaps, VLDM has the advantage of providing interface information by taking all the constituents into account—comprising the solvent—without resorting to any empirical or adjusted parameter. Here, VLDM gave an exhaustive inventory of the interacting elements and quantitative account of the interactions in terms of contact area and time occurrence.

The DNA–histone contacts were characterized in this way at each DNA binding region. The respective contributions of the

histone structured cores and tails to the interface were ascertained. Special attention was paid to the balance between electrostatic and hydrophobic contacts. Water molecules were analyzed in the interface context to identify their role in possibly supplementing the direct contacts between amino acids and nucleotides. The repulsions between superimposed, close DNA gyres were examined in relation to the presence of cations. In sum, the exhaustive description of the elements involved in the DNA–histone interface in solution provides new elements to better understand how the DNA is stabilized around the nucleosome.

## MATERIALS AND METHODS

**Nucleosome Starting Models.** We built five models, all based on the folded domains of *Xenopus laevis* histones and the 601 sequence of 146 base-pairs (bp) from the nucleosome X-ray structure 3MVD.<sup>59</sup> The first model, SYS0, corresponds to the 3MVD structure without any change. The four other models, SYS1, SYS1-bis, SYS2, and SYS2-bis, differ by the length and conformations of the histone N- and C-terminal domains (called tails), hereafter described.

In addition to a structured, folded domain, each histone comprises a disordered N-terminal domain,<sup>4,5</sup> the histone tail; H2A also contains a disordered C-terminal domain. Only short fragments of 5 to 8 amino acids corresponding to minimal N-tail roots were resolved in 3MVD. Full-length tails were observed in only one nucleosome X-ray structure, 1KX5,<sup>44</sup> which is formed of a DNA sequence issued from the human  $\alpha$ -satellite. In our models, the tails from 1KX5 were added to 3MVD after the 3MVD and 1KX5 histone folded parts, identical in both structures, were superimposed. Nevertheless, the 1KX5 N-tails were partially truncated to prevent their collapse and wrapping around the DNA as observed in modeling studies of 1KX5 involving full-length tails.<sup>53,60,61</sup> Indeed, such folding disagrees with experimental results<sup>5,62</sup> and with the idea that the tails may be easily accessible to proteins impacting nucleosome positioning<sup>6,63</sup> or extended when they interact with neighboring particles.<sup>64</sup> Thus, the N-tail amino acids more than 3.5 Å distant from DNA in 1KX5 were removed so that the remaining regions coincide with the inaccessible parts as delimited by trypsin and clostripain digestions.<sup>62,65</sup>

In addition, the tail conformations of the two copies of each histone in 1KX5 are different. For instance the RMSD (Root Mean Square Deviation) between the N-tail backbone heavy atoms of the two H3 copies is 12.2 Å. We took advantage of this heterogeneity to generate different models all including a symmetrically replicated tail conformation: SYS1 and SYS1-bis were built with truncated tails from 1KX5 chains A, B, C, and D, while SYS2 and SYS2-bis contained truncated tails from 1KX5 chains E, F, G and H.

Finally, the H2A C-tail from 1KX5 was integrated in two models, SYS1-bis and SYS2-bis. The composition of tails in our models is given in Table 1; more details in particular about the tail sequences are presented in Table S1.

In our models, the  $pK_a$  values of histidines were calculated using an empirical equation expressing  $pK_a$  as a function of electrostatic potential, hydrogen bonds, and accessible surface area.<sup>66</sup> The  $pK_a$  values of six histidines (H3-H39, H4-H75, H2A-H32, H2A-H82, H2B-H46, and H2B-H79) were between 7.7 and 9.6. These histidines were therefore protonated. The role of this protonation was demonstrated through an additional simulation carried out with unprotonated histidines,

**Table 1. Histone Tail Composition of the Models<sup>a</sup>**

|            | SYS0      | SYS1 and SYS1-bis    | SYS2 and SYS2-bis      |
|------------|-----------|----------------------|------------------------|
| H3 N tail  | 37 → 44   | 24 → 44 from chain A | 24 → 44 from chain E   |
| H4 N tail  | 20 → 24   | 17 → 24 from chain B | 17 → 24 from chain F   |
| H2A N tail | 12 → 16   | 9 → 16 from chain C  | 9 → 16 from chain G    |
| H2B N tail | 28 → 34   | 19 → 34 from chain D | 19 → 34 from chain H   |
|            | SYS0      | SYS1 and SYS2        | SYS1-bis and SYS2-bis  |
| H2A C tail | no C-tail | no C-tail            | 119 → 128 from chain C |

<sup>a</sup>The composition of the N-terminal tails is given for the four histone types in the five models studied here: SYS0 is the 3MVD structure, and SYS1, SYS1-bis, SYS2, and SYS2-bis each contain added amino acids taken from the specified chains of 1KX5. The H2A C-tail from 1KX5 was integrated in SYS1-bis and SYS2-bis. The tail sequences are detailed in Table S1.

in which irreversible breaking of DNA pairing was observed in several places. Such loss of base pairing, not retrieved either in nucleosome X-ray structures or during the other simulations, points out the importance of histidine protonation for maintaining the double helix integrity in the nucleosome.

**Molecular Dynamics Simulation Set Up.** Molecular dynamics simulations were carried out with the CHARMM36 force field<sup>67</sup> with the CMAP correction<sup>68</sup> using the CHARMM<sup>69</sup> and NAMD 2.11 programs.<sup>70</sup> This force field was chosen for its ability to match with experimental NMR data collected on free B-DNAs.<sup>71</sup>

Each structure was immersed in a cubic box (side length of ~130 Å) filled with TIP3P water molecules<sup>72</sup> preserving at least 10 Å of separation between the solute and the edges of the box. Electric neutrality was achieved by the addition of Na<sup>+</sup><sup>73</sup> and Cl<sup>-</sup><sup>74</sup> reaching a final concentration of 150 mM as commonly used for *in vitro* nucleosome experimental studies. The location of ions was optimized using a Monte Carlo approach. The whole systems contain from ~220 000 to 270 000 atoms.

The water molecule and ion positions were first minimized by steepest descent and adapted basis Newton–Raphson<sup>75</sup> methods, keeping the solute fixed, in the CHARMM program. Then, in NAMD, an equilibration simulation progressively raised the temperature from 0 to 300 K, by steps of 10 K every 0.1 ps. During this heating process, harmonic restraints were applied to the atomic position RMSD of the protein backbones and DNA, using force constants of 1000 and 5000 kcal/mol/Å<sup>2</sup>, respectively. These restraints were then relaxed stepwise during a total of 1 ns equilibration using the NPT ensemble.

The production phase was also carried out in the NPT ensemble at a temperature of 300 K and a pressure of 1 bar. Five simulations were performed starting from the models in Table 1 using the Langevin scheme.<sup>76,77</sup> Periodic boundary conditions were used to avoid artifacts at the box boundaries. Infinite range Coulomb interactions were treated with the Particle Mesh Ewald approach.<sup>78</sup> For the van der Waals interactions, a switching function was applied at 10 Å and the cutoff was set to 12 Å. SHAKE was applied to constrain the bond lengths involving hydrogen atoms. The integration time steps were 2 fs, and coordinates were saved every 1000 steps (2 ps). The duration of the simulations are 200 ns for SYS0, SYS1, and SYS2 and 300 ns for SYS1-bis and SYS2-bis.

**Interface Analysis.** The interface between DNA and histones was analyzed by VLDM (Voronoi Laguerre Delaunay for Macromolecules), a software originally developed for proteins<sup>56–58</sup> and extended here to nucleic acids.



Table 2. Flexible Extremities and Stiff Roots of Histone Tails<sup>a</sup>

|         |     |           |           |            |           |           |           |            |           |           |           |           |           |           |           |           |           |          |          |          |          |          |
|---------|-----|-----------|-----------|------------|-----------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|----------|----------|----------|----------|----------|
| N-tails | H3  | 24        | 25        | 26         | 27        | 28        | 29        | 30         | 31        | 32        | 33        | 34        | 35        | 36        | 37*       | 38        | 39        | 40       | 41       | 42       | 43       | 44       |
|         |     | <i>A</i>  | <i>A</i>  | <i>R</i>   | <i>K</i>  | <i>S</i>  | <i>A</i>  | <i>P</i>   | <i>A</i>  | <i>T</i>  | <i>G</i>  | <i>G</i>  | <i>V</i>  | <i>K</i>  | <b>K*</b> | <b>P</b>  | <b>H</b>  | <b>R</b> | <b>Y</b> | <b>R</b> | <b>P</b> | <b>G</b> |
|         | H2A | <i>9</i>  | <i>10</i> | <i>11*</i> | <i>12</i> | <i>13</i> | <i>14</i> | <i>15</i>  | <i>16</i> |           |           |           |           |           |           |           |           |          |          |          |          |          |
|         |     | <i>K</i>  | <i>T</i>  | <b>R*</b>  | <b>A</b>  | <b>K</b>  | <b>A</b>  | <b>K</b>   | <b>T</b>  |           |           |           |           |           |           |           |           |          |          |          |          |          |
|         | H2B | <i>19</i> | <i>20</i> | <i>21</i>  | <i>22</i> | <i>23</i> | <i>24</i> | <i>25*</i> | <i>26</i> | <i>27</i> | <i>28</i> | <i>29</i> | <i>30</i> | <i>31</i> | <i>32</i> | <i>33</i> | <i>34</i> |          |          |          |          |          |
|         |     | <i>Q</i>  | <i>K</i>  | <i>K</i>   | <i>D</i>  | <i>G</i>  | <i>K</i>  | <b>K*</b>  | <b>R</b>  | <b>R</b>  | <b>K</b>  | <b>T</b>  | <b>R</b>  | <b>K</b>  | <b>E</b>  | <b>S</b>  | <b>Y</b>  |          |          |          |          |          |

<sup>a</sup>In SYS1, SYS2, SYS1-bis, and SYS2-bis simulations, the N-tails of H3, H2A, and H2B showed flexible extremities sampling various conformations (amino acids in italic) and stiff roots covering a much more limited conformation landscape (amino acids in bold). The stars indicate the limit between those two regions according to a NMR study.<sup>5</sup>

VLDM relies on a tessellation method, that is, a partition of space into a collection of polyhedra filling space without overlaps or gaps. The program builds the Delaunay tessellation and its Laguerre dual from a set of atomic data, each atom being characterized by its position in space and a weight depending on its van der Waals radius. For consistency with the simulations, the van der Waals radius values are those of CHARMM36. In the present analysis, only the heavy atoms of the solute and solvent 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. The interface can be analyzed according to the nature of contacts. Hydrophobic contacts correspond to carbon atoms exclusively (C–C). 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. A third type reports a proximity of N–C or O–C atoms.

VLDM analyses of simulated nucleosomes were carried out on snapshots extracted every 250 ps from the trajectories, discarding the first 50 ns. To avoid open or distorted polyhedra in the Laguerre tessellation, an 8 Å thick water layer around the solute was taken together with the solute as input to VLDM. 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.

**Additional Trajectory Analyses.** Root Mean Square Deviations (RMSDs) were computed with cpptraj<sup>79</sup> between the snapshots and the initial configuration derived from the X-ray structures before the heating and equilibration stages. RMSDs were calculated on protein backbone heavy atoms or on all DNA heavy atoms.

Protein secondary structures were analyzed with DSSP.<sup>80</sup> The DNA base-pairing was examined with HBPLUS<sup>81</sup> using distance and angle cutoffs of 3.9 Å and 130°, respectively. The electrostatic potential was calculated using APBS<sup>82</sup> in PyMOL,<sup>83</sup> with the ion charges of CHARMM36 and ionic strength of 150 mM. The cation occupancy was calculated with VolMap tools in VMD, using a grid of 1 Å. VMD<sup>84</sup> and PyMOL<sup>83</sup> were used for visualization.

## RESULTS

**Overview of the Simulations.** Five nucleosome systems were built based on the folded domains of *Xenopus laevis* histones and the 601 sequence of 146 base-pairs (bp) from the nucleosome X-ray structure 3MVD.<sup>59</sup> These initial systems differ by the length and conformations of the histone N- and C-terminal domains (tails) (Tables 1 and S1).

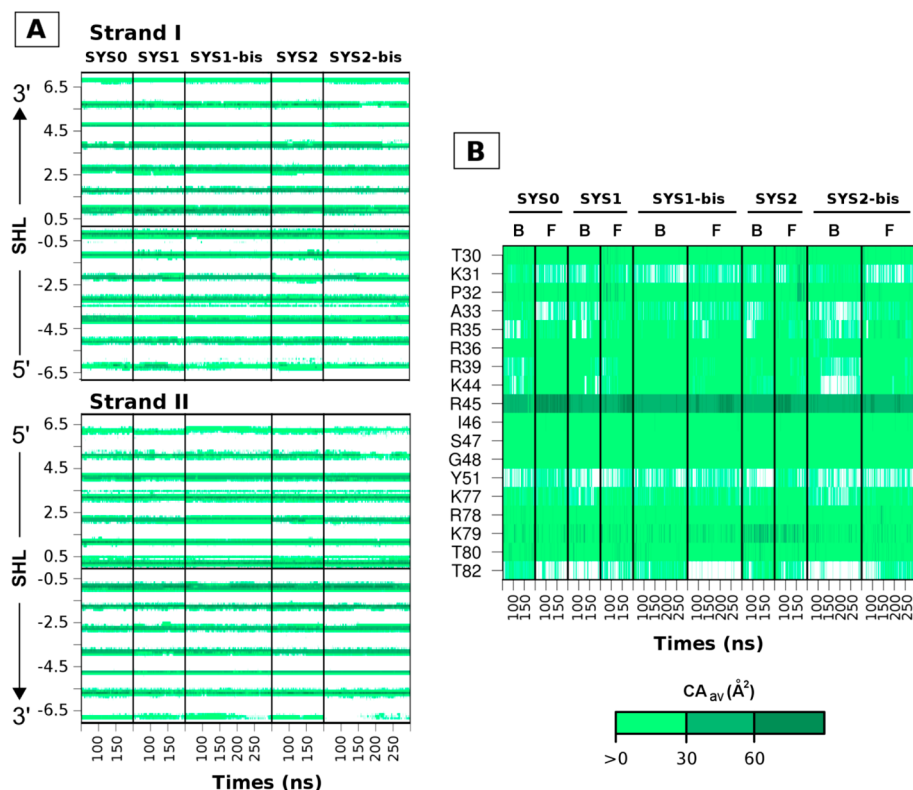
The first model, SYS0, corresponds to the 3MVD structure without any change. Histone N- and C-tail regions from the nucleosome X-ray structure 1KX5<sup>44</sup> were added to SYS0 in the four other models, SYS1, SYS1-bis, SYS2, and SYS2-bis according to the procedure described in the Materials and Methods section. The five systems were simulated with the CHARMM36 force field<sup>67</sup> for a total duration of 1.2 μs. Classical analyses were first performed mainly to ensure that the simulated nucleosomes were not subject to artifactual distortions but also to gain a first estimation of the consistency of the various simulations.

In all simulations, the stability of the radius of gyration ( $R_g$ ) of the whole nucleosomes and DNA double helix (Figure S1) indicated that no histone–histone or DNA–histone disassembly occurred in the trajectories. The system with minimal tails, SYS0, expectedly showed slightly lower  $R_g$  values ( $R_g = 36.2 \pm 0.3$  Å, Figure S1) than the other systems containing longer tails ( $R_g = 37.2 \pm 0.2$  Å on average, Figure S1). The large  $R_g$  values, around 45 Å (see examples in Figure S1), obtained for the DNA in all simulations reflect its location at the periphery of the histone structured cores.

The overall stability of the histones was then assessed through the root-mean-square deviations (RMSD) calculated between the initial configuration and the simulated snapshots (see Materials and Methods). The octameric histone structured cores, first considered, remain globally close to their X-ray counterparts (RMSD  $\approx 2$  Å, Figure S2) in all the systems, with well-preserved  $\alpha$ -helices and loops (example of H3 in Figure S3).

Unlike the structured cores, the tails included in SYS1, SYS1-bis, SYS2, and SYS2-bis deviate from those in the starting structure, which were extracted from 1KX5 tails (Figure S4). This is expected since the tail conformations in 1KX5 are largely shaped by the crystal contacts between neighboring nucleosomes.<sup>44</sup> The highest RMSD values (>10 Å) are reached by H3 and H2B N-tails, which are longer than any other tail (Tables 1 and S1). Although each initial model was built in such a way that the conformations of both copies of each histone type were strictly identical (see Materials and Methods), this pairwise symmetry is lost during the last part of the equilibration stage. The asymmetry is reflected by the RMSDs, which are rarely identical for both copies of a given model (Figure S4). Also, the RMSDs of tails of the same species differ across different systems (Figure S4).

A finer understanding of the tail behavior emerges from additional analyses, discarding the first 50 ns to limit the effect due to the earliest tail rearrangements. Focusing first on H3, H2A, and H2B N-tails, we found that the atomic fluctuations, large and maximal at the extremities (atomic fluctuations up to 10 Å), decrease until reaching a plateau in the regions located before the structured cores and close to the DNA (atomic



**Figure 2.** Contact area maps of the DNA–histone interface. Two examples of maps of DNA–histone contact areas (CAs) plotted as a function of time for the five simulations exploited here. The green color shades code the area value according to the scale given on the right: (A) CA associated with each nucleotide along the two complementary strands (strands I and II) of sequence 601 expressed in terms of super helix location (SHL). (B) CA associated with amino acids of the H4 structured core interacting with the DNA, considering the histone copies separately (chains B and F).

fluctuations below 5 Å) (Figure S5), suggesting flexible extremities and stiff tail roots. However, low atomic fluctuations do not say anything about possible conformation variations across simulations. Systematic cross-RMSDs were thus calculated to better define which tail parts explore similar conformational landscapes. A satisfactory compromise was found for the amino acids 37–44 of H3, 12 or 11–16 of H2A, or 26–34 of H2B since the mean cross-RMSDs between their structures are  $2.6 \pm 1.1$ ,  $3.1 \pm 1.4$ , and  $2.9 \pm 1.1$  Å, respectively. By comparison, the same analyses applied to the remaining N-tail parts led to mean cross-RMSDs of  $14.6 \pm 6.0$  Å for H3,  $7.9 \pm 3.9$  Å for H2A, and  $12.3 \pm 6.6$  Å for H2B. By visual inspection, we found that these high values correspond to a large variety of conformations. Besides, in the latter regions, we observed the formation of a short  $\alpha$ -helix involving amino acids 23–28 of H2B in SYS1 and SYS1-bis, as previously detected in an isolated H2A–H2B heterodimer by NMR.<sup>85</sup>

Globally, the behavior of simulated H3, H2A, and H2B tails is in full agreement with a NMR study of nucleosome in solution, which demonstrated that the tail structural disorder disappears beyond amino acid 37 of H3, 11 of H2A, and 25 of H2B<sup>5</sup> (Table 2). Therefore, although the full range of tail polymorphism can only be partially sampled in the accessible simulation times, the molecular dynamics simulations adequately reproduce the contrast between flexible extremities and stiff roots in H3, H2A, and H2B N-tails. The term “tail root” refers to the amino acids defined as such in Table 2.

Unlike the N-tails of H3, H2A, and H2B, the eight amino acids composing the simulated H4 N-tails are all stiff (atomic fluctuations below 4 Å; Figure S5). The RMSD analyses indicate that these tails adopt two main conformations, one in

SYS1 and SYS1-bis (mean cross-RMSDs of  $3.0 \pm 1.1$  Å) and another, more variable, in SYS2 and SYS2-bis (mean cross-RMSDs of  $6.5 \pm 3.3$  Å). The DNA–histone interface analysis will provide more details on how the H4 tails are stabilized in two distinct conformations owing to differently located interactions with DNA. Concerning the H2A C-tail, most of its amino acids are especially flexible (Figure S5) and the whole tail is highly dynamic (mean cross-RMSDs of  $13.2 \pm 7.2$  Å); both features were observed by NMR.<sup>5</sup> Indeed, we will see that this tail does not engage sizable contacts with DNA.

The last nucleosome component to be scrutinized is DNA. The RMSDs involving all the DNA heavy atoms stabilize around  $2.6 \pm 0.5$  Å on average, considering together the five simulations (Figure S2). This coherency across the systems indicates that the DNA overall structure is affected by neither the large tail truncations in SYS0 nor the presence of H2A C-terminal tails in SYS1-bis and SYS2-bis. Apart from 2 or 3 terminal base pairs, mainly unpaired, the base pair hydrogen bonds are well-conserved in all simulations (90% of simulation time, examples in Figure S6). By comparison, base pairing is maintained 99% of simulation time in the internal part of free DNA dodecamers related to the 601 sequence.<sup>71</sup> This slight difference between free and bound DNAs may reflect the stress induced by the DNA wrapping around the histone core.

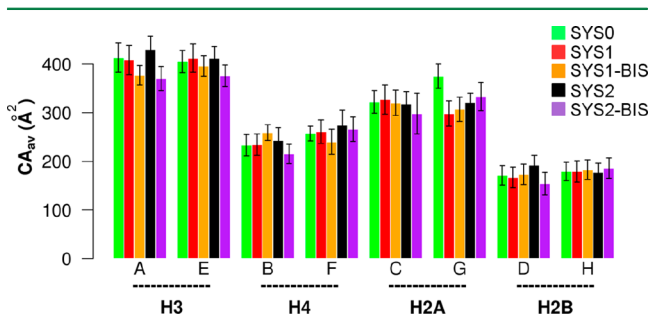
In sum, these preliminary analyses showed that all the simulated nucleosomes preserve the DNA–histone assembly without generating anomalous distortions. In addition, the behavior of histone structured cores and DNA is well preserved across the different simulations. Concerning the tails, an important point is that the proximity to the DNA clearly limits the structural variability of H3, H2A, and H2B N-tail roots,

which is consistent, at least for H3 and H2B, with their position sandwiched in a narrow passage between the two superimposed DNA gyres (Figure 1). In the next sections, devoted to interface analyses, the first 50 ns were systematically discarded to guarantee a correct equilibration of the solvent<sup>86</sup> but also to discard early tail rearrangements from consideration.

**DNA–Histone Interface.** In this section, we present the direct contacts between DNA and either the histone structured cores or tails. Contact maps from VLDM provide an identification of the nucleotides (Figure 2A) and amino acids (example in Figure 2B) participating in the interface as well as the evolution of the contacts over the trajectory time.

The interface was thus monitored in terms of both contact area (CA) and time occurrence. The nucleotide positions are specified in terms of Super Helix Location (SHL), that is, the number of helical turns separating a given base pair from the central base pair, SHL0. SHL is negative or positive for the 5' (Figure 1B) or 3' DNA half sequence, respectively.

**DNA–Histone Structured Core Contacts.** The contact area (CA) associated with DNA–histone structured core is  $2332 \pm 111 \text{ \AA}^2$  on average. Calculated by histone type,  $CA_{av}$  values range from  $391 \pm 37$  for H3 to  $168 \pm 28 \text{ \AA}^2$  for H2B, in the following decreasing order:  $CA_{av}(\text{H3}) > CA_{av}(\text{H2A}) > CA_{av}(\text{H4}) > CA_{av}(\text{H2B})$  (Figure 3).



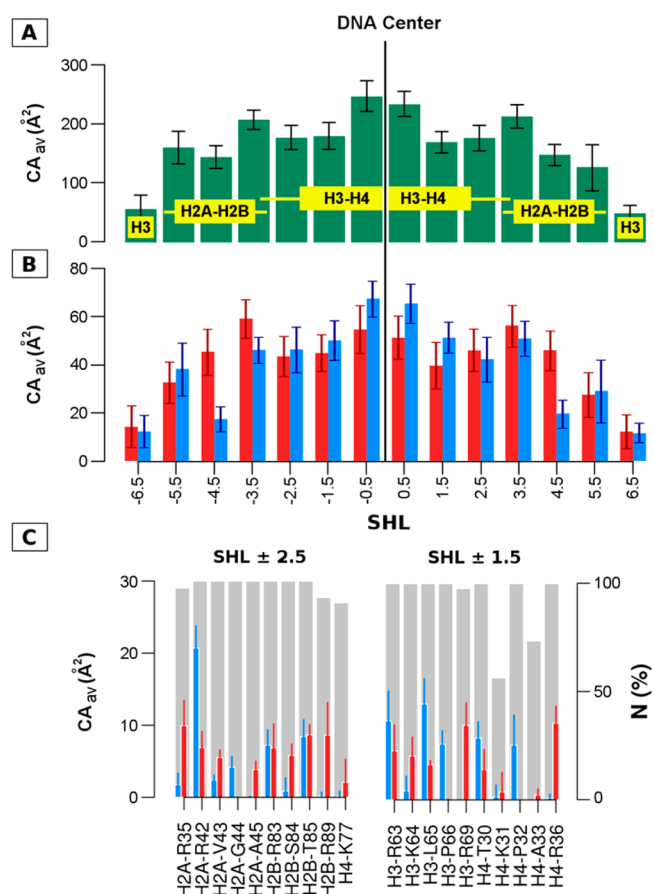
**Figure 3.** Interface between the histone structured cores and DNA across simulations. Contact areas ( $CA_{av}$ ) of the histone structured cores with DNA were extracted from each simulation and averaged over time, considering the histone copies (chains A, E, B, F, etc.) separately. The data associated with the different simulations are colored according to the code given on the top right. The thin vertical error bars correspond to standard deviations.

The correlation coefficients of  $CA_{av}$  calculated at the amino acid level for equivalent copies across different systems (for instance, chain A in  $SYSn$  and  $SYSm$ ) are equal to or higher than 0.9, confirming the consistency of the simulations, visible in Figure 3. Moreover, the interfaces associated with pairs of copies in the same system (for instance, chain A versus chain E in  $SYSn$ ) also compare well (correlation coefficients  $>0.9$ ; see also Figure 3).

Concerning DNA, the contacts are clustered into separate blocks pertaining to nucleotides of both DNA strands (Figure 2A). Successive blocks are separated by 8 to 12 base pairs (bp). The shift between the blocks in strands I and II indicates that the contacts occur across the minor grooves. The simulated interface reproduces the pattern known from nucleosome X-ray structures.<sup>2,44</sup> The comparison of the five data sets of  $CA_{av}$  extracted from simulations and calculated at the nucleotide level led to very high values ( $\geq 0.95$ ) of correlation coefficients, attesting that, as for the histone structured cores, both location and area of contacts are robust across the simulations. In

particular, as evidenced by the similar  $CA_{av}$  values, shortening the tails ( $SYS0$  vs the four other systems) does not affect the DNA–histone structured core interface. As a practical consequence of such a coherence, the five molecular dynamics simulations were analyzed collectively, considering in addition both DNA strands together instead of each strand separately.

The  $CA_{av}$  of the 14 DNA contact regions are symmetric with respect to the center, as shown in Figure 4A.



**Figure 4.** Interface between DNA and the histone structured cores. Average contact areas ( $CA_{av}$ ) of DNA regions, labeled by SHL, with the histone structured cores were calculated considering either all types of contacts (A) or specific hydrophobic (blue) or electrostatic (red) components (B). The yellow flyers indicate the regions where (H3–H4)<sub>2</sub>, the H2A–H2B dimers, or H3 interact. Panel C details the hydrophobic (blue) and electrostatic (red) contributions of amino acids involved in the interface at SHLs  $\pm 2.5$  and  $\pm 1.5$ ; the contact occurrences (N%) are represented by shaded gray area. The data were averaged over the five simulations,  $SYS0$ ,  $SYS1$ ,  $SYS2$ ,  $SYS1$ -bis, and  $SYS2$ -bis. The vertical thin error bars associated with  $CA_{av}$  are standard deviations.

These contacts were analyzed to examine the contribution of the different CA components. Roughly half of  $CA_{av}$  values can be classified as electrostatic and hydrophobic contacts (Figure 4B); the remaining contacts correspond to simple DNA/histone proximity. As expected, the phosphate groups, defined here by O5', O3', P, O1P, and O2P atoms, are involved in 95% of the electrostatic contacts. The hydrophobic contacts implicate sugar carbon atoms and various amino acids, but the largest  $CA_{av}$  values occur with the aliphatic chains of arginines that contact the minor grooves (both copies of H4-R45, H3-R49, H3-R63, H3-R83, H2A-R77, and H2A-R42;



examples of H3-R63 and H2A-R42 in Figure 4C; see also Figure S7). The complete information about the amino acids that participate in electrostatic or hydrophobic contacts at each SHL is furnished in Figure S7, which further illustrates the pairwise similarity of the interface spots with respect to the DNA center. Importantly, a large amount of contacts are present for a large part of the simulation time regardless the  $CA_{av}$  values (examples in Figure 4C; Figure S7). More precisely, 75% of contacts are observed in more than 90% of the snapshots. Such high occurrence rates in time reflect the remarkable stability of the interface between the DNA and histone structured cores (see also Figure 2).

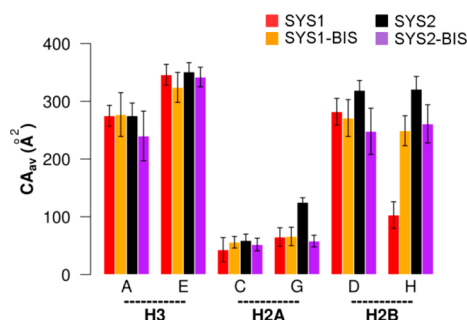
A striking point of this analysis concerns the balance observed between electrostatic and hydrophobic  $CA_{av}$ . Most SHLs are associated with extensive hydrophobic  $CA_{av}$  values, which are even predominant at SHLs  $\pm 5.5$ ,  $-2.5$ ,  $\pm 1.5$ , and  $\pm 0.5$  (Figure 4B; detailed examples in Figure 4C). So the mix of hydrophobic and electrostatic interactions previously described around SHLs  $\pm 1.5$ <sup>45</sup> is a feature reproduced along the whole DNA, SHLs  $\pm 4.5$  being the only locations where electrostatic  $CA_{av}$  prevail (Figure 4B). At a global level, the area of hydrophobic ( $563 \pm 39 \text{ \AA}^2$ ) and electrostatic ( $578 \pm 36 \text{ \AA}^2$ ) contacts are equivalent. However, the most extensive hydrophobic components are associated with the  $(H3-H4)_2$  structured core (Figures 4A and S8). This may enlighten the early step of *in vitro* nucleosome reconstitution that occurs at high ionic strength. At 1.5 M,  $(H3-H4)_2$  robustly binds the DNA exactly as in the complete nucleosome<sup>17</sup> although the electrostatic contacts and consequently the DNA–histone interactions are expected to be weakened. Our results suggest that the substantial hydrophobic component of the  $(H3/H4)_2$  interface complements the weakened electrostatic action so that this histone tetramer is at once appropriately anchored to the DNA.

**DNA–Histone Tail Interface.** Here, we analyze SYS1, SYS2, SYS1-bis, and SYS2-bis that contain elongated tails with respect to SYS0 (Table 1). The total  $CA_{av}$  associated with the tails reaches  $1887 \pm 425 \text{ \AA}^2$ , a substantial quantity compared to the CA involving the structured cores ( $2332 \pm 111 \text{ \AA}^2$ ). However, a large part of the DNA–tail interface is due to the N-tail roots of H3, H2A, and H2B, as explained now.

The extremities of H3, H2A, and H2B N-tails (defined in Table 2), the H4 N-tail, and the H2A C-tail are involved in the interface with the DNA (Figures S9 and S10), but their conformational flexibility described in the Overview section causes significant  $CA_{av}$  variations across the simulations and copies (Figure S9). In addition, their interaction with DNA generally corresponds to rather weak, unstable contacts (Figures S9 and S10). The H2A C-tail is an extreme case since its contribution to the interface is especially weak in terms of both area and time occurrence (Figure S10). In the H4 N-tail, a single arginine, located at the beginning of the simulated sequence (R17), interacts with DNA (Figure S10), binding the minor groove edge at either SHLs  $\pm 2.5$  in SYS1 and SYS1-bis or SHLs  $\pm 1.5$  in SYS2 and SYS2-bis. Finally, the first two or three N-tail amino acids of H3, H2A, and H2B also contact DNA at best 60% of the time (Figure S10). Overall, these results are in full accordance with previous simulations specifically dedicated to the tail behavior.<sup>52,53</sup> They are also strongly supported by a NMR study demonstrating the enhanced flexibility of these tail regions,<sup>5</sup> which is incompatible with any durable, stabilizing interaction with DNA. We thus decided not to further describe the interactions involving these

tail extremities, also because complete sampling would require longer simulation times.

From now on, our analysis focuses on the roots of H3, H2B, and H2A N-tails (defined in Table 2). The roots represent 65% of the total  $CA_{av}$  between DNA and tails, and as seen before, they are characterized by low atomic fluctuations and restricted conformational variabilities. Except for H2B chain H N-tail root in SYS1, whose arginine R30 escaped from the DNA minor groove causing perturbations in this interface region, pairs of  $CA_{av}$  across either simulations or copies are comparable (Figure 5, correlation coefficients calculated at the amino acid level from 0.76 to 0.85). SYS1, SYS2, SYS1-bis, and SYS2-bis simulations were consequently analyzed together.

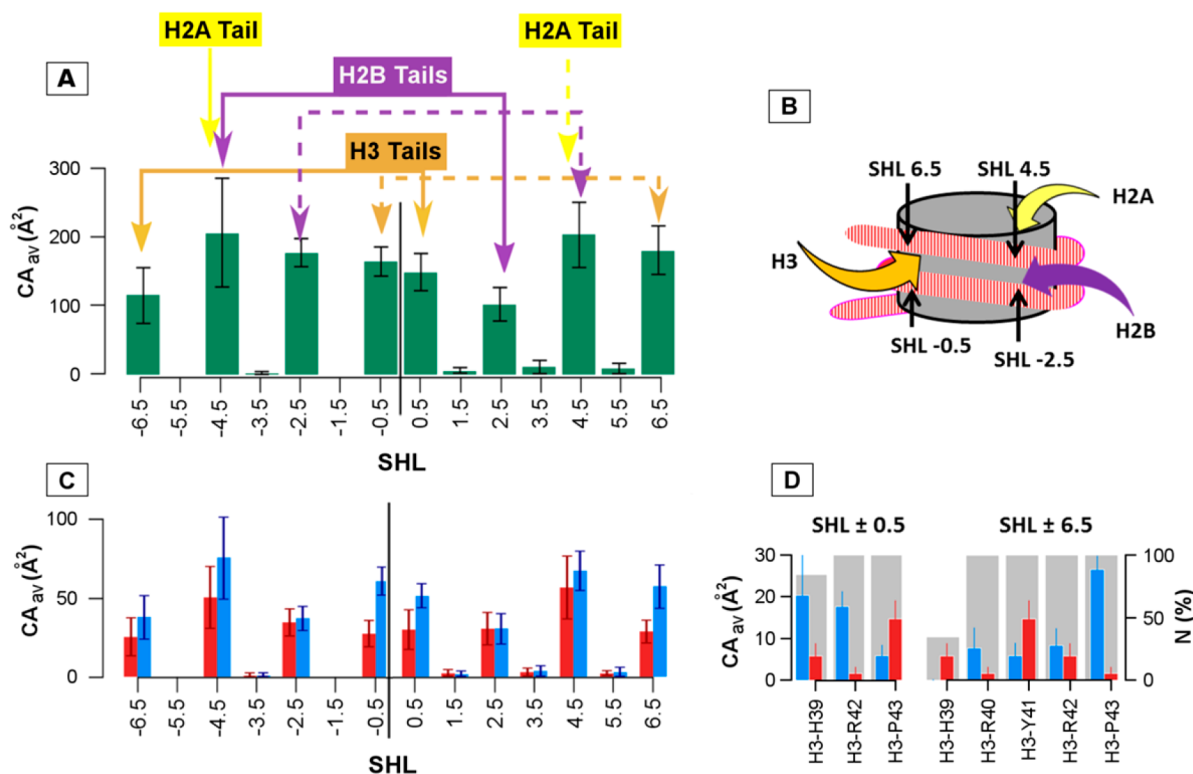


**Figure 5.** Interfaces involving the H3, H2A, and H2B tail roots and DNA across simulations. Contact areas ( $CA_{av}$ ) of the histone tail roots of H3, H2A, and H2B (defined in Table 2) with DNA were extracted from each simulation and averaged over time, considering the histone copies (chains A E, C G, and D H) separately. The data associated with the different simulations are colored according to the color code given on the top. The vertical thin error bars correspond to standard deviations.

Eight SHL spots are contacted by H3, H2B and H2A N-tail roots (Figure 6A). Each H2B and H3 tail root contains 8 amino acids and binds to two juxtaposed double helices (Figure 6A,B) while the H2A tail root is shorter (4 amino acids) and interacts with only one double helix on the upper or lower side of the nucleosome (Figure 6A,B). Accordingly, the H2A N-tail root contributes much less to the interface than the H3 and H2B ones (Figure 5). Nevertheless it maintains contacts with DNA during the whole simulation time (Figure S10), unlike the flexible H4 N-tail and the H2A C-tail (Figure S10). Each of the eight DNA binding spots combines electrostatic and hydrophobic contacts (Figures 6C,D and S11), most of them being very stable during the trajectories (Figures 6D and S11). Overall, in the DNA–tail root interface considered here, the hydrophobic  $CA_{av}$  is almost twice as large as the electrostatic  $CA_{av}$ :  $390 \pm 45 \text{ \AA}^2$  and  $213 \pm 39 \text{ \AA}^2$ , respectively. This particular interface shows a more marked hydrophobic character than the DNA–structured core interface in which electrostatic and hydrophobic  $CA_{av}$  are equivalent.

To conclude on this section, our analyses reveal the importance of hydrophobic contact areas that substantially complement the electrostatic interactions in the DNA–histone interface. Thus, the combination of very stable, direct electrostatic and hydrophobic contacts is the norm for each DNA point anchoring the histone structured core or H3, H2A, and H2B N-tail roots. A noticeable contribution to the cohesion of the DNA superhelical wrapping is provided by the double contact made by each H3 and H2B N-tail root connecting two juxtaposed DNA gyres.





**Figure 6.** Interface between DNA and the tail roots of H3, H2A, and H2B. (A) Average total contact areas ( $CA_{av}$ ) calculated at each SHL region involved in the DNA interface with the N-tail roots of H3, H2A, and H2B; the plain and dashed connectors specify the type of tail copies that contact the DNA. (B) Schematic representation of the nucleosome illustrating the location of one copy of each of the three N-tail species. (C) Hydrophobic (blue) or electrostatic (red) components of the  $CA_{av}$ , the total of which is represented in panel A. (D) Details of the hydrophobic (blue) and electrostatic (red) contribution of the major amino acids involved in the interface at SHLs  $\pm 0.5$  and  $\pm 6.5$ ; the total contact occurrences (N%) are represented in gray. The vertical thin bars in panels A, C, and D are standard deviations associated with  $CA_{av}$ . The data were extracted and averaged from the four simulations, SYS1, SYS2, SYS1-bis, and SYS2-bis.

**Interfacial Water Molecules.** The water molecules were then scrutinized in the context of the DNA–histone interface. This investigation was motivated to a large extent by a previous analysis of hydrogen bonds in the nucleosome crystallographic structure 1KX5, that contained a 147 bp DNA derived from the  $\alpha$  human satellite.<sup>44</sup> In this structure, the only one that allowed the unambiguous identification of a very large number of water molecules, 121 water molecules were found to bind simultaneously to DNA and histone structured cores. These water molecules either sustain direct DNA–histone hydrogen bonds or connect groups that are too distant or improperly oriented to allow direct interactions. The authors of the crystallographic study concluded that direct and water mediated hydrogen bonds equally contribute to the nucleosome stability. The molecular dynamics runs gave us a good opportunity to probe, at least *in silico*, the existence of long-lived water mediated hydrogen bonds in solution.

Equivalent analyses were carried out on the simulated structures, using the same histone parts and criteria ( $D_{\text{Donor-Acceptor}} \leq 3.5 \text{ \AA}$ ; Donor–Hydrogen–Acceptor angle  $>90^\circ$ ) as Davey and collaborators,<sup>44</sup> keeping in mind that our models differ from 1KX5 by the DNA length and sequence. In our simulations, we found a total of 51 amino acids involved in direct hydrogen bonds with DNA compared to 37 in 1KX5 (Table S2). However, this difference is largely canceled by the fact that the water mediated connections between 12 amino acids and DNA in 1KX5 turn into direct hydrogen bonds in the simulated structures (Table S2). This observation illustrates that, compared to the solid state, DNA and histones in solution

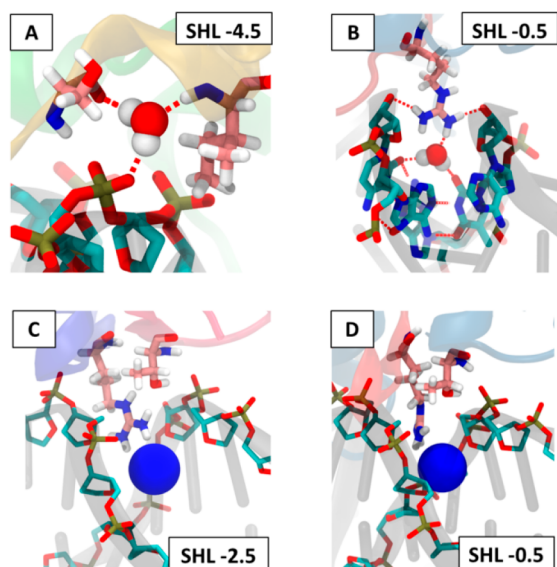
have the potential to undergo subtle local rearrangements promoting direct electrostatic interactions. Overall, most amino acids involved in hydrogen bonds are retrieved in both experiment and simulation (Table S2).

Actually, according to VLDM, very few water molecules are trapped in the simulated interface. At SHLs  $\pm 4.5$ , a water molecule links a phosphate group to S33 and I36 of H2B (Figure 7A) in 74% of snapshots.

Another water molecule interacts with the O2 atom of thymine or cytosine in the minor groove floor at SHL  $-0.5$  and R45 in H4 (Figure 7B) in 40% of snapshots. Both DNA–water–I36 and DNA–water–R45 interactions reinforce direct contacts (Figure S7).

Thus, an overwhelming majority of water molecules circulate all around the DNA and histones, without being confined at precise interface locations as observed in 1KX5. This mobility does not mean that water molecules do not intervene in the interface. By filling gaps between DNA and histone surfaces, reducing the repulsions between electronegative atoms, or making transient hydrogen bonds, water molecules play an essential role in the nucleosome as well as in any DNA–protein complex.<sup>87</sup> Nevertheless, the simulations provide a strong indication that stable, long-lived water molecules bridging DNA and histones are the exception rather than the rule in liquid solution.

**Cations at DNA–Histone and DNA–DNA Interfaces.** In a way similar to water molecules, ions could behave as interface mediators. Visual inspection revealed that some  $\text{Na}^+$  spent a large part of the simulation time at the DNA–histone interface,

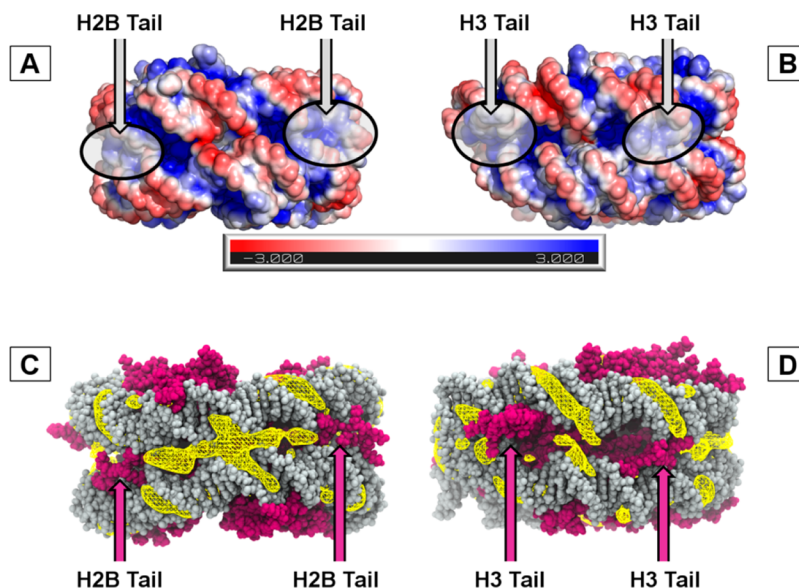


**Figure 7.** Water molecules and  $\text{Na}^+$  cations trapped in the DNA–histone structured core interface. Top panels show examples of water molecules mediating DNA–histone interactions: (A) phosphate group located above the major groove at SHL  $-4.5$  interacts with S33 and I36 of H2B via a water molecule; (B) another water molecule joins the minor groove floor at SHL  $-0.5$  and R45 of H4. These structures were extracted from SYS2. The bottom panels show  $\text{Na}^+$  cations inserted in the DNA minor groove in the vicinity of R83 of H3 and T80 of H4 at SHL  $-2.5$  (C) or R45 of H4 and T118 at SHL  $-0.5$  (D).

contrary to  $\text{Cl}^-$  that are, as expected, always far from the DNA. The existence of long-lived cation binding sites is also attested by  $\text{Na}^+$  with low atomic fluctuations. By systematically calculating the distances between  $\text{Na}^+$  and DNA ( $D_{\text{Na}^+-\text{DNA}}$ ) or histones ( $D_{\text{Na}^+-\text{histone}}$ ) for each ion in each snapshot, as well as distance distributions, we found that seven  $\text{Na}^+$  reside close to both DNA and histones ( $D_{\text{Na}^+-\text{DNA}} \leq 4$  and  $D_{\text{Na}^+-\text{histone}} \leq 5.5$

Å, see examples of distance distributions in Figure S12) between 10% and 100% of the simulation time, but their location and time occurrence often vary across the simulations (Table S3). This disparity is not a reminiscence of the simulation starting points, since none of the observed interfacing ions is already in place at the beginning of the production phase; it may rather reflect an insufficient sampling or local structural differences modulating the electrostatic attractiveness for  $\text{Na}^+$ . X-ray structures of nucleosome cannot help here, because they commonly contain divalent cations instead of  $\text{Na}^+$ , which in any case cannot be easily identified even in high resolution structures. Further specific investigations would be consequently required to determine which parameters influence the presence of ions at the interface. At the current stage, one can yet state that  $\text{Na}^+$  are able to stay as much as 250 ns at a precise interface location (Table S3). Such situations are observed in particular at SHLs  $\pm 4.5$ ,  $\pm 2.5$ , and  $-0.5$ , where one  $\text{Na}^+$  can be enclosed in the DNA minor groove, close to an inserted arginine/threonine couple (examples in Figure 7C,D).

$\text{Na}^+$  cations are also implicated in the stabilization of the DNA superhelical path, which strictly speaking does not relate to the interface. Indeed, the wrapping of the DNA around the histones brings phosphate groups belonging to juxtaposed DNA gyres close to each other. The minimal distance that separates pairs of phosphate atoms ( $D_{\text{p-p}}$ ) is especially short when two minor grooves face each other, dropping to 7 or 8 Å for certain pairs of superposed SHLs. Thus, the severe electrostatic repulsion induced by the proximity of negatively charged phosphate groups requires efficient shielding. Shielding is ensured by direct phosphate group–amino acid interactions in the regions where H3 or H2B tail roots pass between the DNA gyres. To better understand what happens when close phosphate groups are not separated by histone tails, we carried out calculations of electrostatic potential and  $\text{Na}^+$  density, expressed in terms of occupancy (see Materials and Methods).



**Figure 8.** Electrostatic potentials and  $\text{Na}^+$  occupancy. These nucleosome back side views are centered on the regions where the minor grooves at either SHLs  $-4$  and  $+4$  (A and C) or SHLs  $-2$  and  $+6$  (B and D) are juxtaposed. The histone tails passing between the DNA gyres are specified in flyers. (A, B) Electrostatic potentials are represented on the nucleosome solvent accessible surface according to the color scale given underneath, ranging from  $-3.0$  to  $+3.0$  kT/e. (C, D)  $\text{Na}^+$  occupancy is represented by a yellow isosurface corresponding to occupancy = 3% of the simulation time. DNA is in gray, and the histones are in purple.

Globally, there is a good correspondence between electronegative patches (Figure 8A,B) and  $\text{Na}^+$  occupancies (Figure 8C,D). Cation (here  $\text{Na}^+$ ) accumulation in the electronegative parts of minor grooves exposed to the solvent (Figure 8C,D) was previously described for free DNAs<sup>88–90</sup> and is thus not surprising. More relevant for our purpose,  $\text{Na}^+$  ions are also observed in interstices between DNA gyres devoid of histone tails, in particular in the gap extending between the two H2B tail contact points (Figure 8C). This region is centered on an extremely narrow interstice between the juxtaposed minor grooves at SHLs  $-4$  and  $+4$  ( $D_{p-p} = 7.3 \pm 1.0$  Å), around the dyad axis, which generates a marked electronegative potential (Figure 8A). Indeed, a substantial  $\text{Na}^+$  occurrence is observed all along the gap, spreading from the first to the second H2B tail copy (Figure 8C).

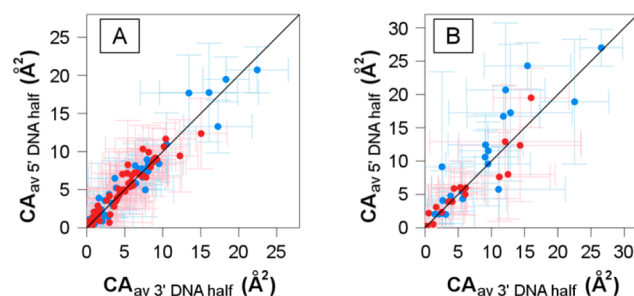
$\text{Na}^+$  ions are not as prevalent in the gyre interstices delimited by the crossing points of H3 and H2B tails (Figure 8D). In this region, the juxtaposed minor grooves are more distant ( $D_{p-p}$  of  $11 \pm 1.6$  and  $14.6 \pm 1.1$  Å at SHLs  $-2/+6$  and  $+2/-6$ , respectively) than for the facing SHLs  $-4$  and  $+4$  and, accordingly, produce a less strong electronegative potential (Figure 8B). However, even in this case,  $\text{Na}^+$  ions overflow from the minor groove and cover the closest phosphate groups at the top of the facing minor grooves (Figure 8D).

The first part of the results presented above shows that  $\text{Na}^+$  can penetrate the interface at minor grooves where arginines and threonines are inserted. The second part reveals that a cloud of  $\text{Na}^+$  intercalates in the narrow gaps between DNA gyres, weakening the repulsive electrostatic forces between close phosphate groups and complementing the shielding action of H3 and H2B tails. So, cations play an important role in stabilizing and preserving the DNA superhelix path.

**DNA–Histone Contacts and DNA Sequence.** The last point relates to a possible sequence effect on the interface, which could play a role in the differential ability of the 5' and 3' halves of the 601 sequence to be maintained around the histone cores.<sup>17,42,46</sup> The 601 sequence is not palindromic; most local contact patterns at SHL positions symmetric with respect to the DNA center (SHL  $-n$  in the DNA 5' half and SHL  $+n$  in the 3' half,  $n = 0.5, 1.5, 2.5$ , etc.) are composed of different sequences (given in Table S4), except for SHLs  $\pm 1.5$  where the TTAAA motif is found on both sides in a way satisfying symmetry. Now each SHL  $-n$ /SHL  $+n$  pair faces identical regions of histone copies and contacts the same amino acids. This particular feature offers the opportunity to finely probe to what extent the DNA sequence modulates the interface by comparing the SHL  $-n$ /histone and SHL  $+n$ /histone contacts.

The similarity within pairs of CAs measured from the point of view of either the histone copies (Figures 3 and 5) or the SHL  $-n$ /SHL  $+n$  pairs (Figures 4, 6, S7, and S11) is especially striking in the case of the histone structured cores (Figures 3, 4, and S7). The very symmetric contact profiles clearly argue against any sequence effect on the interface. The symmetry is confirmed by the excellent correlation existing between contacts measured in the 3' and 5' DNA halves, in terms of both electrostatic and hydrophobic CA with the protein core (Figure 9A). Although slightly weaker, the correlation for the contacts with the tail roots is also convincing, taking into account the standard deviations (Figure 9B).

Thus, our simulations involving sequence 601 nucleosome do not support the idea that the DNA sequence significantly affects the contacts with the histones.



**Figure 9.** Comparison between the contacts involving the 5' and 3' halves of the 601 sequence and the histones. Hydrophobic (blue) and electrostatic (red) components of contact areas ( $CA_{av}$ ) between the histone amino acids and the nucleotides of the 5' DNA half are compared to their equivalents involving the 3' DNA half. The contacts represented are between DNA and either the histone structured cores (A; correlation coefficients of 0.98 for both hydrophobic and electrostatic CAs) or the tail roots (B; correlation coefficients are 0.91 for both hydrophobic and electrostatic CAs). The vertical and horizontal error bars are the standard deviations calculated across the simulations. The diagonal black line represents  $y = x$ .

## DISCUSSION AND CONCLUSION

A nucleosome containing the high affinity 601 sequence was simulated by molecular dynamics in explicit solvent to study the DNA–histone interface in solution. The interface was analyzed by VLDM, a Voronoi based method, in terms of contact area and time occurrence.

The quantitative measurements of the interface at each contacted SHL clarify global features of DNA–histone interactions. That similar large and very stable contacts are observed for eight binding sites, from SHL  $-3.5$  to SHL  $3.5$ , indicates that all of these central SHLs are equally crucial for anchoring the H3 and H4 structured cores. At each of these SHL interfaces, one arginine, intimately interacting with the minor groove, represents nearly 20% of the contacts, and 7 to 9 other amino acids make the complement, the ensemble engaging long-lived contacts with the DNA. The situation is completely different near the DNA entry/exit, especially at SHLs  $\pm 6.5$  where the contact areas with the histone structured domains are four times smaller than those at the central SHLs, and the contact occurrences are relatively low.

Concerning the tails, without surprise, the extremities of H3, H2A, and H2B N-tails, as well as the whole H2A C-tail, are flexible and mobile as previously observed;<sup>5,52</sup> they have a minimal contribution to the DNA–histone interface. In contrast, the conformational variability of the roots of H3, H2B, and H2A N-tails is considerably lower owing to intimate, long-lived interactions with the DNA. Indeed, these proximal histone regions contribute to 35% of the total DNA–histone contact area.

Considering together the structured domain and root tail contributions gives a global view of the contact distribution along both central and peripheral DNA regions. As expected, the central DNA region engages the more extensive and stable contacts, cumulating interactions with H3–H4 structured cores and H3 tail roots. However, another noticeable region emerges around SHLs  $\pm 4.5$ , where large contacts with the H2A and H2B root tails substantially reinforce the interface involving the structured cores of the same histones. This observation remarkably parallels optical tweezer<sup>40,42</sup> and FRET<sup>34</sup> experiments in which a strong resistance to nucleosome disassembly was observed at 40–50 bp from the dyad, conjecturally related



to a local strengthening of the DNA–histone interactions. Finally, the defective interactions are limited to the DNA extremities, in agreement with unwrapping or breathing studies.<sup>15</sup> It is also interesting to note that post-translational modifications that affect amino acids contacting the DNA at SHLs  $\pm 6.5$  (such as T45, R52, R53, or K56 of H3) enhance the DNA unwrapping,<sup>91,92</sup> very likely by further weakening this particular interface region.

Returning to the analysis of interactions, most interfaces at the SHLs contacting the histone structured cores and eventually tail roots combine both types of interactions, electrostatic and hydrophobic (Figures 4 and 6). In several cases, in particular in the DNA–histone tail root interface, the hydrophobic interactions cover larger areas than the electrostatic contacts. Overall, the  $CA_{\text{Electrostatic}}/CA_{\text{Total}}$  and  $CA_{\text{Hydrophobic}}/CA_{\text{Total}}$  ratios are 0.23 and 0.27, respectively. These values stress the central importance of hydrophobic contacts in the nucleosome, which are too often relegated to a secondary role. Actually, the DNA–histone interface resembles those of other typical DNA–protein complexes, which the interfaces comprise two-thirds of nonelectrostatic contacts on average.<sup>50</sup>

With very few exceptions, simulated water molecules fill space between DNA and histones in a fluid manner, without forming durable bridges between DNA and histones. The water mediated hydrogen bonds observed in 1KXS<sup>44</sup> either disappear or, more often, turn into direct hydrogen bonds owing to local adjustments of DNA and amino acids. In the context of a potential structuring role of water inferred from X-ray structures of DNA–protein complexes,<sup>44,50,93,94</sup> our simulations show that water molecules in interstices formed at the DNA–histone interface do not get durably trapped at any specific position. The various types of interaction, for example, electrostatic shielding, polarization, or hydrogen bonds, affect individual solvent molecules, including ions, mostly in a transitory way. The net effect is visible at the statistics level, through decreased fluctuation and diffusion near the DNA or protein surface.

Ions in solution play a decisive role in neutralizing electrostatic repulsions between solute charges. A typical case is the juxtaposed gyres of DNA wrapped around the nucleosome core: in particular in regions where two minor grooves face each other, some phosphate groups are very close to each other. The positively charged residues in H3 and H2B tail roots crossing the DNA belt between the gyres are undoubtedly essential for the superhelix formation and stability. Outside these contact points, the narrowest interstices between DNA gyres are the scene of  $\text{Na}^+$  accumulation. At these locations, cations appear here as fundamental elements in the nucleosome cohesion, explaining why the DNA superhelical path exists notwithstanding phosphate groups in close proximity.

As a final point, the comparison of interfaces involving pairs of histone copies of the same kind bound to different DNA fragments did not supply evidence of any DNA sequence effect. At this stage, we cannot exclude that the DNA–histone interface could be affected by DNA sequences clearly unfavorable to nucleosome formation. However, our result echoes the findings of a recent experimental study that enabled detection of DNA distortions during the sequential steps of histone binding.<sup>17</sup> During the reconstitution process, the DNA sequence clearly influences the early events of nucleosome formation. So, the  $(\text{H3}/\text{H4})_2$  tetramer binds more efficiently

the 5' than the 3' side of the central part of the 601 sequence. Introducing mutations in the 5' side of the 601 sequence also decreases the nucleosome reconstitution efficiency. However, no difference persists once the complexes are stabilized.

As provided by our simulations in explicit solvent and VLDM, the detailed DNA–histone interface in solution can serve as a reference, very comparable to the other nucleoprotein complexes.<sup>50</sup> From a practical point of view, our results can be used as biophysical background to interpret experiments on nucleosome assembly or disassembly, as well as to anticipate the structural consequences of epigenetic modifications or histone mutations. The intricate and cohesive interface depicted here sheds light on how the DNA superhelical wrapping and induced distortions are maintained around the histones in the nucleosome.

On the other hand, the nature of this interface raises the issue of completely dissociating DNA from histones, a process that implies breaking extensive and robust contacts and may thus be a hard, energy consuming task. This concern is supported, in particular, by the rarity of spontaneous partial unwrapping of DNA peripheral regions,<sup>41</sup> still supposed to be the initial step of nucleosome unfolding. It was proposed that, in cell, remodeling factors are required to favor the occurrence of such breathing and to achieve the nucleosome disassembly.<sup>36</sup> The strong DNA–histone interface described here could further explain why the action of remodeling factors is essential in the dynamical positioning of nucleosomes along eukaryotic genomes.

## ■ ASSOCIATED CONTENT

### 📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00936.

Radius of gyration of simulated NCPs and DNA in NCPs, RMSDs of the histone structured core and the DNA, H3 structured core secondary structures, distribution of the RMSD values of the histone tails, atomic fluctuations of histone tails, Watson–Crick base pairing in simulated DNA, interface between the DNA and the histone structured cores, hydrophobic and electrostatic contact areas between the DNA and either the structured cores of the  $(\text{H3}-\text{H4})_2$  tetramer or the H2A–H2B dimers, comparison of the simulated interfaces involving the extremities of H3, H2A, and H2B tails, H4 tail, or H2A C-tail and the DNA, interface between the histone tails and the DNA, interface between the DNA and the histone tail roots, density plots of  $\text{Na}^+$ –DNA and  $\text{Na}^+$ –histone distances, sequences of histone tails, hydrogen bonds between DNA and histone structured cores in 1KXS and molecular dynamics simulations, time occurrence of  $\text{Na}^+$  cations at the DNA–histone interface, and DNA sequences at contacted SHLs (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

\*Brigitte Hartmann. Tel: (+33) 1 47 40 74 21. Fax: (+33) 1 47 40 76 71. E-mail: [bhartman@ens-paris-saclay.fr](mailto:bhartman@ens-paris-saclay.fr).

\*Christophe Oguey. Tel: (+33) 1 3425 7518. Fax: (+33) 1 3425 7500. E-mail: [oguey@u-cergy.fr](mailto:oguey@u-cergy.fr).

### ORCID

Brigitte Hartmann: 0000-0002-0611-8806

### Author Contributions

<sup>†</sup>A.E. and R.R. have equally contributed to the work. A.E. performed the simulations and contributed to the development of VLDM with C.O. A.E. initiated the analyses that were completed and supplemented by R.R., also in charge of most figures. M.B. participated in discussions and in rereading the manuscript. C.O. and B.H., the project leaders, wrote the manuscript.

### Funding

This work was funded by the PEPS (Projets Exploratoires Pluridisciplinaires) program of CNRS. The simulations were carried out on the GENCI-CEA platform. This study was further supported by the “Initiative d’Excellence” program from the French State (Grant “DYNAMO”, ANR-11-LABX-0011-01 and ANR-11-EQPX-0008).

### Notes

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

The authors thank Liliane Mouawad (Institut Curie, Centre universitaire d’Orsay, Orsay) for helpful advice.

### REFERENCES

- (1) Luger, K. Nucleosomes: Structure and Function. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd: Chichester, 2001.
- (2) Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution. *Nature* **1997**, *389* (6648), 251–260.
- (3) Luger, K.; Richmond, T. J. DNA Binding within the Nucleosome Core. *Curr. Opin. Struct. Biol.* **1998**, *8* (1), 33–40.
- (4) Gao, M.; Nadaud, P. S.; Bernier, M. W.; North, J. A.; Hammel, P. C.; Poirier, M. G.; Jaroniec, C. P. Histone H3 and H4 N-Terminal Tails in Nucleosome Arrays at Cellular Concentrations Probed by Magic Angle Spinning NMR Spectroscopy. *J. Am. Chem. Soc.* **2013**, *135* (41), 15278–15281.
- (5) Zhou, B.-R.; Feng, H.; Ghirlando, R.; Kato, H.; Gruschus, J.; Bai, Y. Histone H4 K16Q Mutation, an Acetylation Mimic, Causes Structural Disorder of Its N-Terminal Basic Patch in the Nucleosome. *J. Mol. Biol.* **2012**, *421* (1), 30–37.
- (6) Hughes, A. L.; Rando, O. J. Mechanisms Underlying Nucleosome Positioning in Vivo. *Annu. Rev. Biophys.* **2014**, *43*, 41–63.
- (7) Gurard-Levin, Z. A.; Quivy, J.-P.; Almouzni, G. Histone Chaperones: Assisting Histone Traffic and Nucleosome Dynamics. *Annu. Rev. Biochem.* **2014**, *83*, 487–517.
- (8) Park, Y.-J.; Luger, K. Histone Chaperones in Nucleosome Eviction and Histone Exchange. *Curr. Opin. Struct. Biol.* **2008**, *18* (3), 282–289.
- (9) Clapier, C. R.; Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.* **2009**, *78*, 273–304.
- (10) Längst, G.; Manelyte, L. Chromatin Remodelers: From Function to Dysfunction. *Genes* **2015**, *6* (2), 299–324.
- (11) Kamakaka, R. T.; Biggins, S. Histone Variants: Deviants? *Genes Dev.* **2005**, *19* (3), 295–316.
- (12) Talbert, P. B.; Henikoff, S. Histone Variants on the Move: Substrates for Chromatin Dynamics. *Nat. Rev. Mol. Cell Biol.* **2017**, *18* (2), 115–126.
- (13) Tessarz, P.; Kouzarides, T. Histone Core Modifications Regulating Nucleosome Structure and Dynamics. *Nat. Rev. Mol. Cell Biol.* **2014**, *15* (11), 703–708.
- (14) Zentner, G. E.; Henikoff, S. Regulation of Nucleosome Dynamics by Histone Modifications. *Nat. Struct. Mol. Biol.* **2013**, *20* (3), 259–266.
- (15) Eslami-Mossallam, B.; Schiessel, H.; van Noort, J. Nucleosome Dynamics: Sequence Matters. *Adv. Colloid Interface Sci.* **2016**, *232*, 101–113.

(16) Fernandez, A. G.; Anderson, J. N. Nucleosome Positioning Determinants. *J. Mol. Biol.* **2007**, *371* (3), 649–668.

(17) Hatakeyama, A.; Hartmann, B.; Travers, A.; Nogues, C.; Buckle, M. High-Resolution Biophysical Analysis of the Dynamics of Nucleosome Formation. *Sci. Rep.* **2016**, *6*, 27337.

(18) Heddi, B.; Oguey, C.; Lavelle, C.; Foloppe, N.; Hartmann, B. Intrinsic Flexibility of B-DNA: The Experimental TRX Scale. *Nucleic Acids Res.* **2010**, *38* (3), 1034–1047.

(19) Kaplan, N.; Moore, I. K.; Fondufe-Mittendorf, Y.; Gossett, A. J.; Tillo, D.; Field, Y.; LeProust, E. M.; Hughes, T. R.; Lieb, J. D.; Widom, J.; Segal, E. The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome. *Nature* **2009**, *458* (7236), 362–366.

(20) Korber, P. Active Nucleosome Positioning beyond Intrinsic Biophysics Is Revealed by in Vitro Reconstitution. *Biochem. Soc. Trans.* **2012**, *40* (2), 377–382.

(21) Liu, M.-J.; Seddon, A. E.; Tsai, Z. T.-Y.; Major, I. T.; Floer, M.; Howe, G. A.; Shiu, S.-H. Determinants of Nucleosome Positioning and Their Influence on Plant Gene Expression. *Genome Res.* **2015**, *25* (8), 1182–1195.

(22) Richmond, T. J.; Davey, C. A. The Structure of DNA in the Nucleosome Core. *Nature* **2003**, *423* (6936), 145–150.

(23) Struhl, K.; Segal, E. Determinants of Nucleosome Positioning. *Nat. Struct. Mol. Biol.* **2013**, *20* (3), 267–273.

(24) Travers, A.; Hiriart, E.; Churcher, M.; Caserta, M.; Di Mauro, E. The DNA Sequence-Dependence of Nucleosome Positioning in Vivo and in Vitro. *J. Biomol. Struct. Dyn.* **2010**, *27* (6), 713–724.

(25) Widom, J. Role of DNA Sequence in Nucleosome Stability and Dynamics. *Q. Rev. Biophys.* **2001**, *34* (3), 269–324.

(26) Xu, X.; Ben Imeddourene, A.; Zargarian, L.; Foloppe, N.; Mauffret, O.; Hartmann, B. NMR Studies of DNA Support the Role of Pre-Existing Minor Groove Variations in Nucleosome Indirect Readout. *Biochemistry* **2014**, *53* (35), 5601–5612.

(27) Yuan, G.-C.; Liu, Y.-J.; Dion, M. F.; Slack, M. D.; Wu, L. F.; Altschuler, S. J.; Rando, O. J. Genome-Scale Identification of Nucleosome Positions in *S. Cerevisiae*. *Science* **2005**, *309* (5734), 626–630.

(28) Burton, D. R.; Butler, M. J.; Hyde, J. E.; Phillips, D.; Skidmore, C. J.; Walker, I. O. The Interaction of Core Histones with DNA: Equilibrium Binding Studies. *Nucleic Acids Res.* **1978**, *5* (10), 3643–3663.

(29) Oohara, I.; Wada, A. Spectroscopic Studies on Histone-DNA Interactions. II. Three Transitions in Nucleosomes Resolved by Salt-Titration. *J. Mol. Biol.* **1987**, *196* (2), 399–411.

(30) Arimura, Y.; Tachiwana, H.; Oda, T.; Sato, M.; Kurumizaka, H. Structural Analysis of the Hexasome, Lacking One Histone H2A/H2B Dimer from the Conventional Nucleosome. *Biochemistry* **2012**, *51* (15), 3302–3309.

(31) Böhm, V.; Hieb, A. R.; Andrews, A. J.; Gansen, A.; Rocker, A.; Tóth, K.; Luger, K.; Langowski, J. Nucleosome Accessibility Governed by the Dimer/Tetramer Interface. *Nucleic Acids Res.* **2011**, *39* (8), 3093–3102.

(32) Chen, Y.; Tokuda, J. M.; Topping, T.; Sutton, J. L.; Meisburger, S. P.; Pabit, S. A.; Gloss, L. M.; Pollack, L. Revealing Transient Structures of Nucleosomes as DNA Unwinds. *Nucleic Acids Res.* **2014**, *42* (13), 8767–8776.

(33) Chen, Y.; Tokuda, J. M.; Topping, T.; Meisburger, S. P.; Pabit, S. A.; Gloss, L. M.; Pollack, L. Asymmetric Unwrapping of Nucleosomal DNA Propagates Asymmetric Opening and Dissociation of the Histone Core. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (2), 334–339.

(34) Gansen, A.; Valeri, A.; Hauger, F.; Felekyan, S.; Kalinin, S.; Tóth, K.; Langowski, J.; Seidel, C. A. M. Nucleosome Disassembly Intermediates Characterized by Single-Molecule FRET. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (36), 15308–15313.

(35) Hoch, D. A.; Stratton, J. J.; Gloss, L. M. Protein-Protein Förster Resonance Energy Transfer Analysis of Nucleosome Core Particles Containing H2A and H2A.Z. *J. Mol. Biol.* **2007**, *371* (4), 971–988.

- (36) Miyagi, A.; Ando, T.; Lyubchenko, Y. L. Dynamics of Nucleosomes Assessed with Time-Lapse High-Speed Atomic Force Microscopy. *Biochemistry* **2011**, *50* (37), 7901–7908.
- (37) Hagerman, T. A.; Fu, Q.; Molinié, B.; Denvir, J.; Lindsay, S.; Georgel, P. T. Chromatin Stability at Low Concentration Depends on Histone Octamer Saturation Levels. *Biophys. J.* **2009**, *96* (5), 1944–1951.
- (38) Lyubchenko, Y. L. Nanoscale Nucleosome Dynamics Assessed with Time-Lapse AFM. *Biophys. Rev.* **2014**, *6* (2), 181–190.
- (39) Park, Y.-J.; Dyer, P. N.; Tremethick, D. J.; Luger, K. A New Fluorescence Resonance Energy Transfer Approach Demonstrates That the Histone Variant H2AZ Stabilizes the Histone Octamer within the Nucleosome. *J. Biol. Chem.* **2004**, *279* (23), 24274–24282.
- (40) Brower-Toland, B. D.; Smith, C. L.; Yeh, R. C.; Lis, J. T.; Peterson, C. L.; Wang, M. D. Mechanical Disruption of Individual Nucleosomes Reveals a Reversible Multistage Release of DNA. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (4), 1960–1965.
- (41) Tomschik, M.; van Holde, K.; Zlatanova, J. Nucleosome Dynamics as Studied by Single-Pair Fluorescence Resonance Energy Transfer: A Reevaluation. *J. Fluoresc.* **2009**, *19* (1), 53–62.
- (42) Hall, M. A.; Shundrovsky, A.; Bai, L.; Fulbright, R. M.; Lis, J. T.; Wang, M. D. High-Resolution Dynamic Mapping of Histone-DNA Interactions in a Nucleosome. *Nat. Struct. Mol. Biol.* **2009**, *16* (2), 124–129.
- (43) Wang, D.; Ulyanov, N. B.; Zhurkin, V. B. Sequence-Dependent Kink-and-Slide Deformations of Nucleosomal DNA Facilitated by Histone Arginines Bound in the Minor Groove. *J. Biomol. Struct. Dyn.* **2010**, *27* (6), 843–859.
- (44) Davey, C. A.; Sargent, D. F.; Luger, K.; Maeder, A. W.; Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 Å Resolution. *J. Mol. Biol.* **2002**, *319* (5), 1097–1113.
- (45) Wu, B.; Mohideen, K.; Vasudevan, D.; Davey, C. A. Structural Insight into the Sequence Dependence of Nucleosome Positioning. *Structure* **2010**, *18* (4), 528–536.
- (46) Chua, E. Y. D.; Vasudevan, D.; Davey, G. E.; Wu, B.; Davey, C. A. The Mechanics behind DNA Sequence-Dependent Properties of the Nucleosome. *Nucleic Acids Res.* **2012**, *40* (13), 6338–6352.
- (47) García-Pérez, M.; Pinto, M.; Subirana, J. A. Nonsequence-Specific Arginine Interactions in the Nucleosome Core Particle. *Biopolymers* **2003**, *69* (4), 432–439.
- (48) West, S. M.; Rohs, R.; Mann, R. S.; Honig, B. Electrostatic Interactions between Arginines and the Minor Groove in the Nucleosome. *J. Biomol. Struct. Dyn.* **2010**, *27* (6), 861–866.
- (49) Yusufaly, T. I.; Li, Y.; Singh, G.; Olson, W. K. Arginine-Phosphate Salt Bridges between Histones and DNA: Intermolecular Actuators That Control Nucleosome Architecture. *J. Chem. Phys.* **2014**, *141* (16), 165102.
- (50) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino Acid-Base Interactions: A Three-Dimensional Analysis of Protein-DNA Interactions at an Atomic Level. *Nucleic Acids Res.* **2001**, *29* (13), 2860–2874.
- (51) Harp, J. M.; Hanson, B. L.; Timm, D. E.; Bunick, G. J. Asymmetries in the Nucleosome Core Particle at 2.5 Å Resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56* (12), 1513–1534.
- (52) Erler, J.; Zhang, R.; Petridis, L.; Cheng, X.; Smith, J. C.; Langowski, J. The Role of Histone Tails in the Nucleosome: A Computational Study. *Biophys. J.* **2014**, *107* (12), 2911–2922.
- (53) Shaytan, A. K.; Armeev, G. A.; Goncarencu, A.; Zhurkin, V. B.; Landsman, D.; Panchenko, A. R. Coupling between Histone Conformations and DNA Geometry in Nucleosomes on a Microsecond Timescale: Atomistic Insights into Nucleosome Functions. *J. Mol. Biol.* **2016**, *428* (1), 221–237.
- (54) Ettig, R.; Kepper, N.; Stehr, R.; Wedemann, G.; Rippe, K. Dissecting DNA-Histone Interactions in the Nucleosome by Molecular Dynamics Simulations of DNA Unwrapping. *Biophys. J.* **2011**, *101* (8), 1999–2008.
- (55) Thåström, A.; Bingham, L. M.; Widom, J. Nucleosomal Locations of Dominant DNA Sequence Motifs for Histone-DNA Interactions and Nucleosome Positioning. *J. Mol. Biol.* **2004**, *338* (4), 695–709.
- (56) Esque, J.; Leonard, S.; de Brevern, A. G.; Oguey, C. VLDP Web Server: A Powerful Geometric Tool for Analysing Protein Structures in Their Environment. *Nucleic Acids Res.* **2013**, *41* (W1), W373–W378.
- (57) Esque, J.; Oguey, C.; de Brevern, A. G. Comparative Analysis of Threshold and Tessellation Methods for Determining Protein Contacts. *J. Chem. Inf. Model.* **2011**, *51* (2), 493–507.
- (58) Esque, J.; Oguey, C.; de Brevern, A. G. A Novel Evaluation of Residue and Protein Volumes by Means of Laguerre Tessellation. *J. Chem. Inf. Model.* **2010**, *50* (5), 947–960.
- (59) Makde, R. D.; England, J. R.; Yennawar, H. P.; Tan, S. Structure of RCC1 Chromatin Factor Bound to the Nucleosome Core Particle. *Nature* **2010**, *467* (7315), 562–566.
- (60) Biswas, M.; Voltz, K.; Smith, J. C.; Langowski, J. Role of Histone Tails in Structural Stability of the Nucleosome. *PLoS Comput. Biol.* **2011**, *7* (12), e1002279.
- (61) Roccatano, D.; Barthel, A.; Zacharias, M. Structural Flexibility of the Nucleosome Core Particle at Atomic Resolution Studied by Molecular Dynamics Simulation. *Biopolymers* **2007**, *85* (5–6), 407–421.
- (62) Dumuis-Kervabon, A.; Encontre, I.; Etienne, G.; Jauregui-Adell, J.; Méry, J.; Mesnier, D.; Parello, J. A Chromatin Core Particle Obtained by Selective Cleavage of Histones by Clostripain. *EMBO J.* **1986**, *5* (7), 1735–1742.
- (63) Speranzini, V.; Pilotto, S.; Sixma, T. K.; Mattevi, A. Touch, Act and Go: Landing and Operating on Nucleosomes. *EMBO J.* **2016**, *35* (4), 376–388.
- (64) Pepenella, S.; Murphy, K. J.; Hayes, J. J. Intra- and Inter-Nucleosome Interactions of the Core Histone Tail Domains in Higher-Order Chromatin Structure. *Chromosoma* **2014**, *123* (1–2), 3–13.
- (65) Morales, V.; Richard-Foy, H. Role of Histone N-Terminal Tails and Their Acetylation in Nucleosome Dynamics. *Mol. Cell. Biol.* **2000**, *20* (19), 7230–7237.
- (66) Krieger, E.; Nielsen, J. E.; Spronk, C. A. E. M.; Vriend, G. Fast Empirical PKa Prediction by Ewald Summation. *J. Mol. Graphics Modell.* **2006**, *25* (4), 481–486.
- (67) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; Mackerell, A. D. Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *J. Chem. Theory Comput.* **2012**, *8* (1), 348–362.
- (68) Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604.
- (69) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Cafilisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
- (70) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
- (71) Ben Imeddourene, A.; Elbahnsi, A.; Guéroult, M.; Oguey, C.; Foloppe, N.; Hartmann, B. Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLoS Comput. Biol.* **2015**, *11* (12), e1004631.
- (72) Jorgensen, W. L.; et al. *J. Chem. Phys.* **1983**, *79*, 926.
- (73) Lamoureux, G.; Roux, B. Absolute Hydration Free Energy Scale for Alkali and Halide Ions Established from Simulations with a Polarizable Force Field. *J. Phys. Chem. B* **2006**, *110* (7), 3308–3322.
- (74) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations. *J. Chem. Phys.* **1994**, *100* (12), 9050–9063.



(75) Brooks, B.; Brucoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.

(76) Brooks, C.; Karplus, M. Deformable Stochastic Boundaries in Molecular-Dynamics. *J. Chem. Phys.* **1983**, *79* (12), 6312–6325.

(77) Feller, S.; Zhang, Y.; Pastor, R.; Brooks, B. Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621.

(78) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.Log(n) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(79) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

(80) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637.

(81) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793.

(82) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (18), 10037–10041.

(83) Schrödinger, LLC. *The PyMOL Molecular Graphics System*, version 1.8, 2015.

(84) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33–38.

(85) Moriwaki, Y.; Yamane, T.; Ohtomo, H.; Ikeguchi, M.; Kurita, J.; Sato, M.; Nagadoi, A.; Shimojo, H.; Nishimura, Y. Solution Structure of the Isolated Histone H2A-H2B Heterodimer. *Sci. Rep.* **2016**, *6*, 24999.

(86) Materese, C. K.; Savelyev, A.; Papoian, G. A. Counterion Atmosphere and Hydration Patterns near a Nucleosome Core Particle. *J. Am. Chem. Soc.* **2009**, *131* (41), 15005–15013.

(87) Reddy, C. K.; Das, A.; Jayaram, B. Do Water Molecules Mediate Protein-DNA Recognition? *J. Mol. Biol.* **2001**, *314* (3), 619–632.

(88) Auffinger, P.; D'Ascenzo, L.; Ennifar, E. Sodium and Potassium Interactions with Nucleic Acids. *Met. Ions Life Sci.* **2016**, *16*, 167–201.

(89) Savelyev, A.; MacKerell, A. D. Differential Deformability of the DNA Minor Groove and Altered BI/BII Backbone Conformational Equilibrium by the Monovalent Ions Li(+), Na(+), K(+), and Rb(+). *J. Chem. Theory Comput.* **2015**, *11* (9), 4473–4485.

(90) Sushko, M. L.; Thomas, D. G.; Pabit, S. A.; Pollack, L.; Onufriev, A. V.; Baker, N. A. The Role of Correlation and Solvation in Ion Interactions with B-DNA. *Biophys. J.* **2016**, *110* (2), 315–326.

(91) Bowman, G. D.; Poirier, M. G. Post-Translational Modifications of Histones That Influence Nucleosome Dynamics. *Chem. Rev.* **2015**, *115* (6), 2274–2295.

(92) Lawrence, M.; Daujat, S.; Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet.* **2016**, *32* (1), 42–56.

(93) Schneider, B.; Cerný, J.; Svozil, D.; Cech, P.; Gelly, J.-C.; de Brevern, A. G. Bioinformatic Analysis of the Protein/DNA Interface. *Nucleic Acids Res.* **2014**, *42* (5), 3381–3394.

(94) Sonavane, S.; Chakrabarti, P. Cavities in Protein-DNA and Protein-RNA Interfaces. *Nucleic Acids Res.* **2009**, *37* (14), 4613–4620.

# **Holding the nucleosome together: A quantitative description of the DNA-histone interface in solution**

Ahmad Elbahnsi, Romain Retureau, Marc Baaden, Brigitte Hartmann and Christophe Oguey

## **Supplementary Figures**

**Figure S1:** Radius of gyration of simulated nucleosomes and nucleosomal DNA.

**Figure S2:** RMSDs of the histone structured core and the DNA.

**Figure S3:** H3 structured core secondary structures.

**Figure S4:** Distribution of the RMSD values of the histone tails.

**Figure S5:** Atomic fluctuations of histone tails.

**Figure S6:** Watson-Crick base pairing in simulated DNA.

**Figure S7:** Interface between the DNA and the histone structured cores.

**Figure S8:** Hydrophobic and electrostatic contact areas between the DNA and either the structured cores of the (H3-H4)<sub>2</sub> tetramer or the H2A-H2B dimers.

**Figure S9:** Comparison of the simulated interfaces involving the extremities of H3, H2A and H2B tails, H4 tail or H2A C-tail and the DNA.

**Figure S10:** Interface between the histone tails and the DNA.

**Figure S11:** Interface between the DNA and the histone tail roots.

**Figure S12:** Density plots of Na<sup>+</sup> - DNA and Na<sup>+</sup> - histones distances

## **Supplementary Tables**

**Table S1:** Sequences of histone tails.

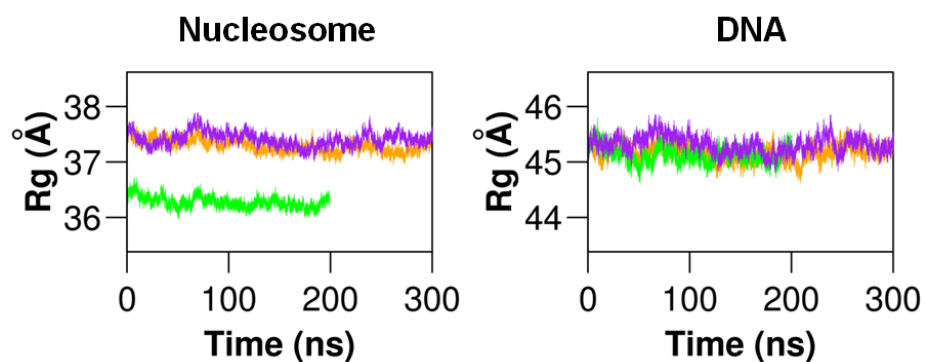
**Tables S2-1 and S2-2:** Hydrogen bonds between DNA and histone structured cores in 1KX5 and MDs.

**Table S3:** Time occurrence of Na<sup>+</sup> cations at the DNA histone interface.

**Table S4 :** DNA sequences at contacted SHLs.

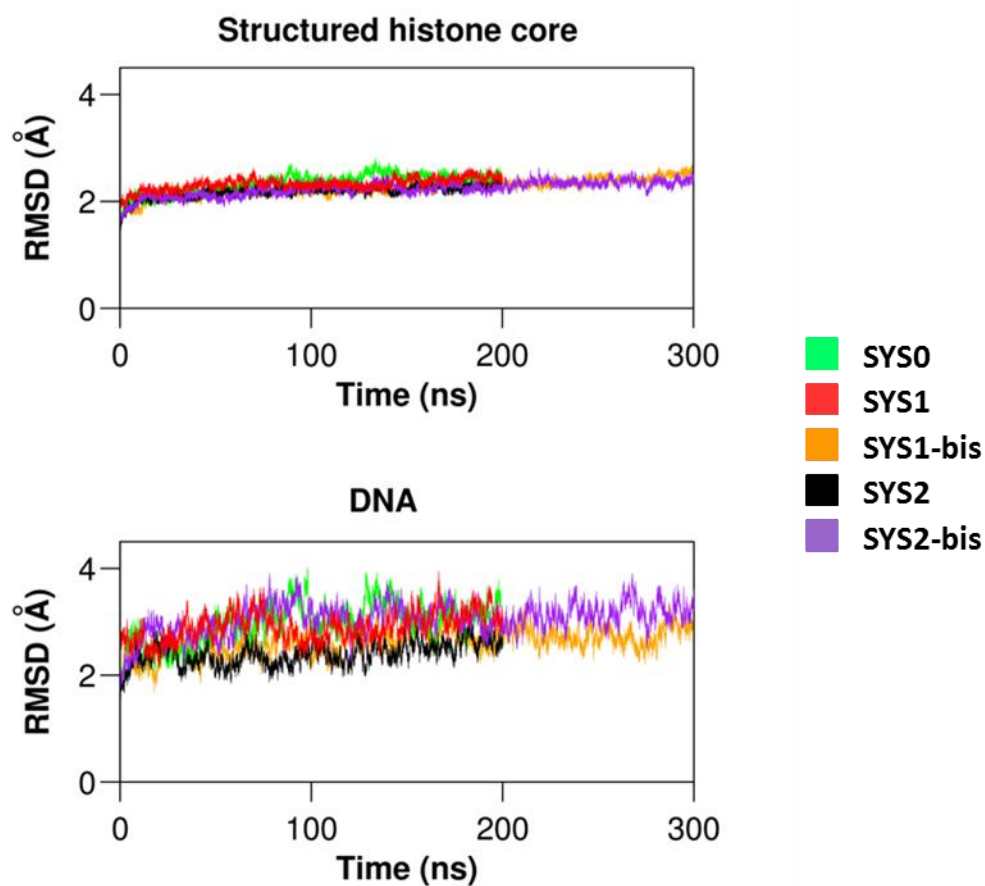


**Figure S1:** Radius of gyration of simulated nucleosomes and nucleosomal DNA.



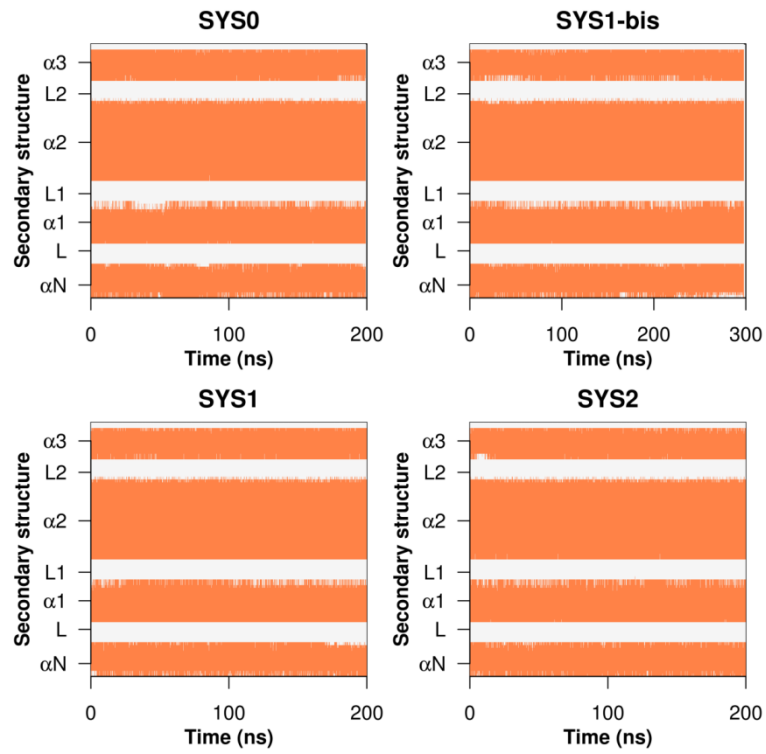
The radius of gyration ( $R_g$ ) is plotted as a function of time for representative systems, SYS0 (green), SYS1-bis (orange), SYS2-bis (purple), considering either the whole nucleosomes (left panel) or only the DNA (right panel). SYS1 and SYS2 behave as SYS1-bis and SYS2-bis.

**Figure S2:** RMSDs of the histone structured core and the DNA.



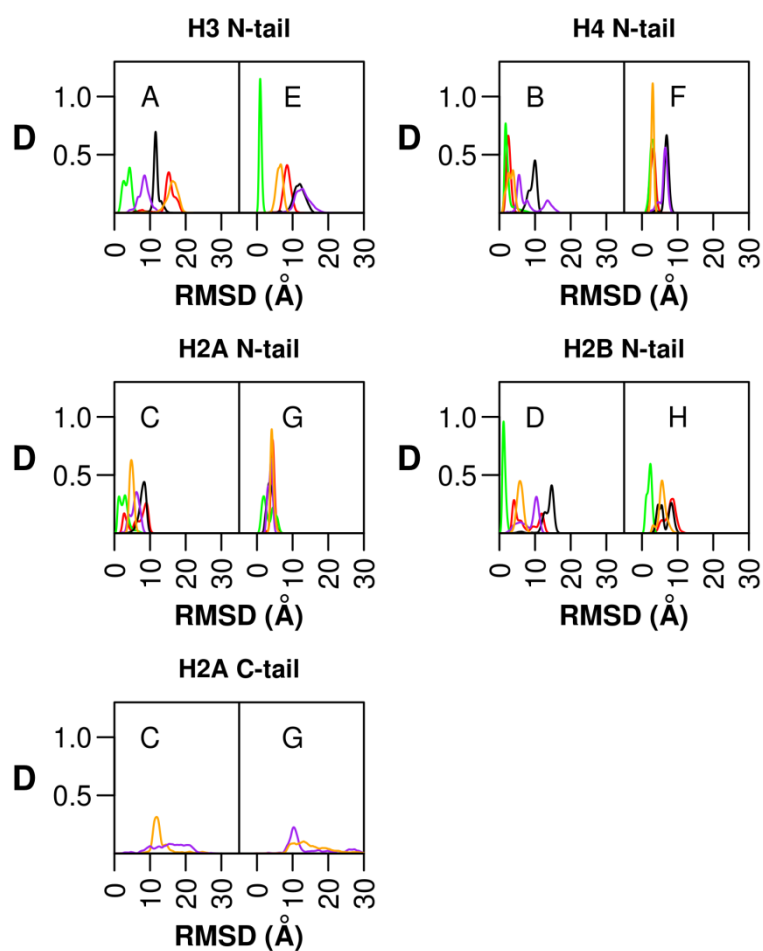
The RMSDs of the structured octameric histone core (top panel) and of the DNA (bottom panel) are extracted from each simulation and plotted as a function of time. The data associated with the different simulations are colored according to the code given on the right. The RMSDs were calculated between the initial model derived from X-ray structures and the simulation snapshots, using the backbone heavy atoms for the histones or all the heavy atoms for DNA. The first and last two base-pairs of DNA, unpaired, were discarded from the analysis.

**Figure S3:** H3 structured core secondary structures.



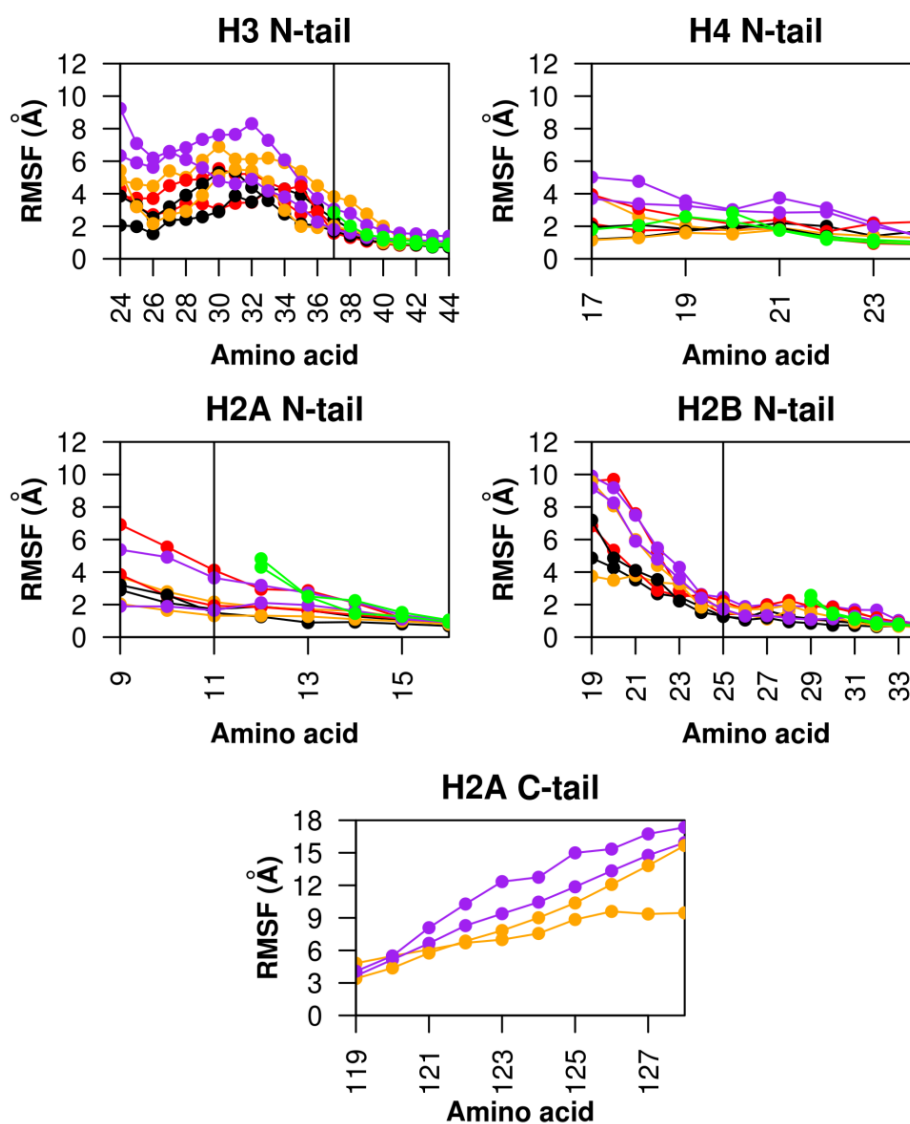
The evolution of the  $\alpha$ -helices (orange) and loops (white) composing the structured core of H3 is shown along the simulations SYS0, SYS1-bis, SYS1 and SYS2. Identical results were obtained with SYS2-bis.

**Figure S4:** Distribution of the RMSD values of the histone tails.



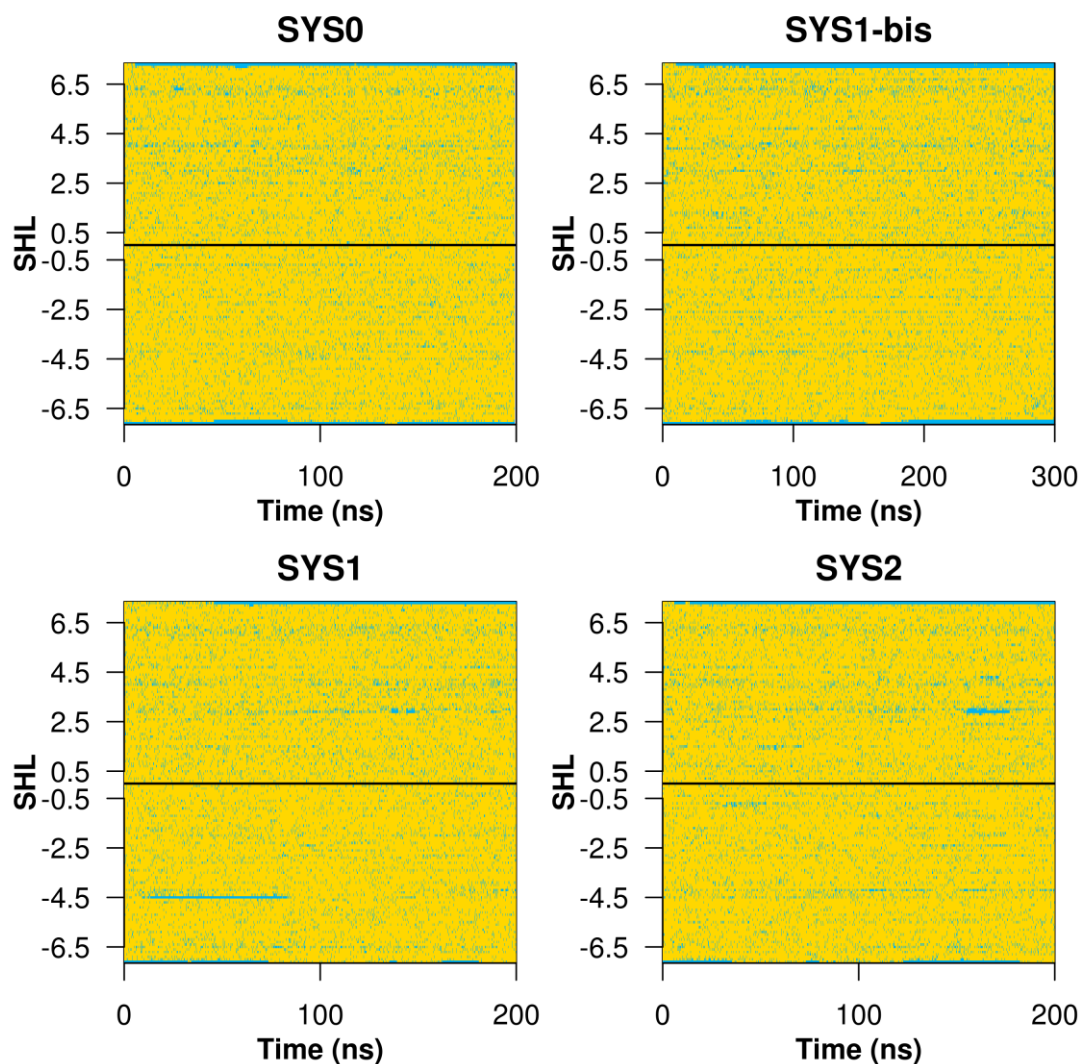
The occurrence distribution (or histogram) (D) of RMSDs is plotted for both copies (left and right of each panel) of the tails, with SYS0 in green, SYS1 in red, SYS1-bis in orange, SYS2 in black and SYS2-bis in purple. For each simulation and each tail of a given chain, RMSDs were calculated between the initial model and the snapshots after the corresponding histone structured core configurations had been superimposed.

**Figure S5:** Atomic fluctuations of histone tails.



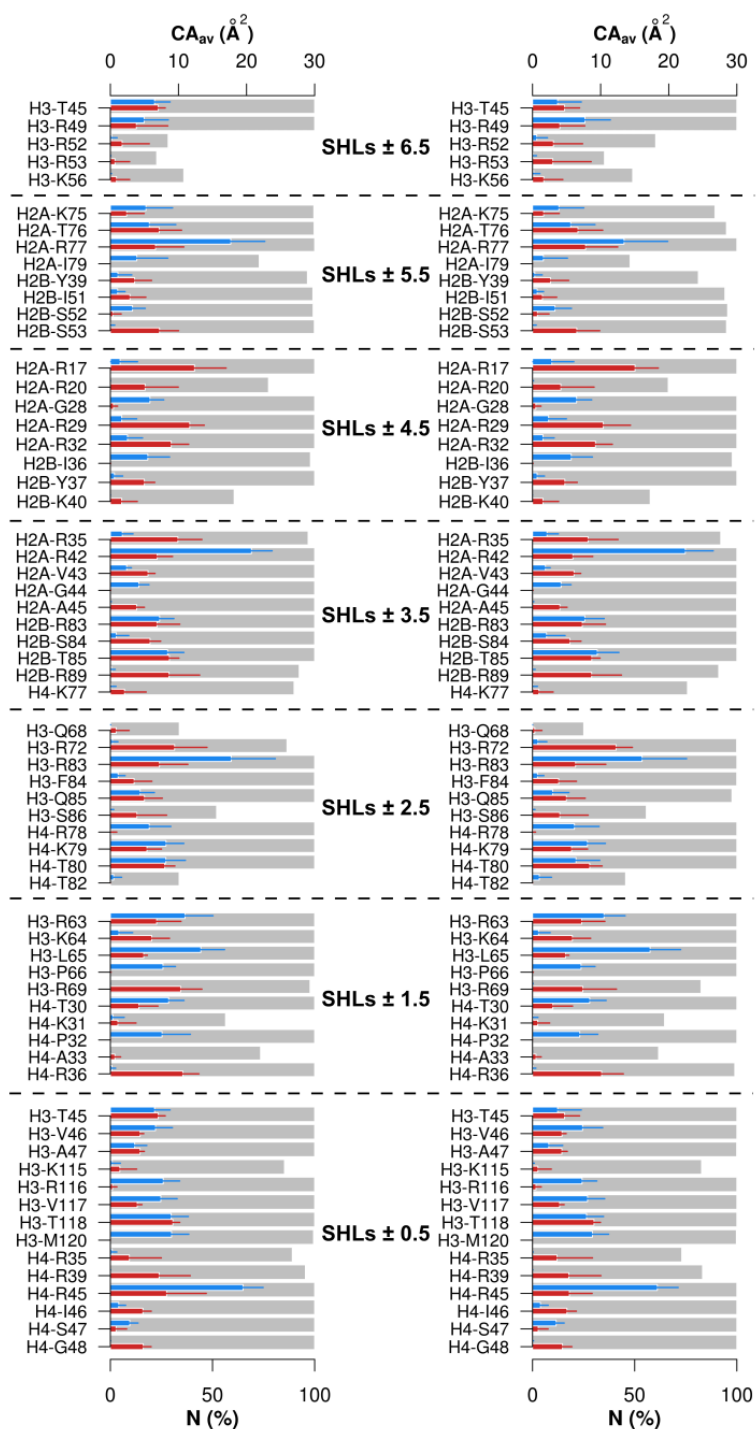
The mean atomic fluctuations (RMSF) per residue were calculated for both copies of each histone tail in the five simulations, SYS0 (green), SYS1 (red), SYS1-bis (orange), SYS2 (black) and SYS2-bis (purple). Note that the y axis of the H2A C-tail panel covers a larger range of values than in the other panels. The vertical black lines correspond to the limit between flexible and stiff regions according to a previous NMR study (1).

**Figure S6:** Watson-Crick base pairing in simulated DNA.



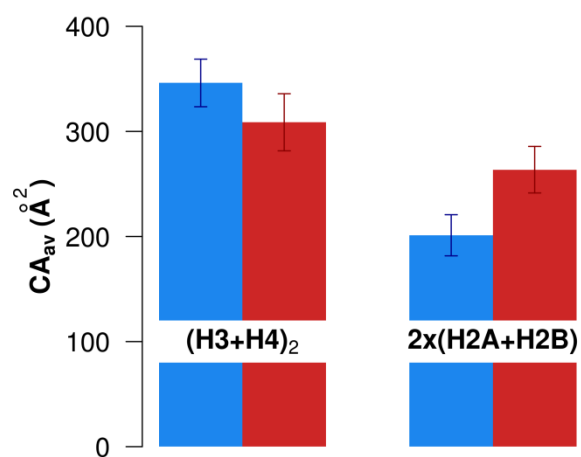
The DNA base pairing along sequence 601 is represented as a function of time for SYS0, SYS1-bis, SYS1 and SYS2. The presence of full base-pairing (two and three hydrogen bonds for A:T or G:C base-pairs, respectively) corresponds to yellow bars; the loss of at least one hydrogen bond is represented by cyan bars. Position along the DNA molecule is expressed in terms of Super Helix Location (SHL), that is, number of double-helix turns from the origin, at the DNA center (SHL0).

**Figure S7:** Interface between DNA and the histone structured cores.



These plots present the amino acids of the histone structured cores involved in each SHL interface; left and right plots correspond to negative and positive SHLs, respectively. Each amino acid is characterized by its hydrophobic (blue) and electrostatic (red) average contact area ( $CA_{av}$ , horizontal thin bars for standard deviations) and its occurrence (N%), represented by shaded gray area. The data were averaged on the five simulations, SYS0, SYS1, SYS2, SYS1-bis and SYS2-bis.

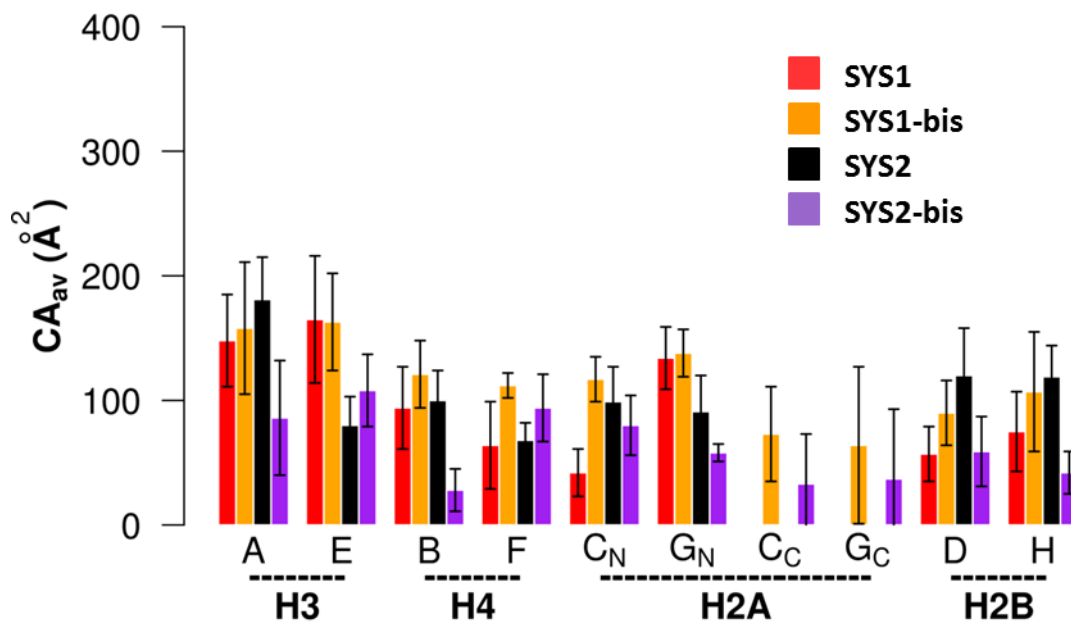
**Figure S8:** Hydrophobic and electrostatic contact areas between DNA and the structured cores of either the  $(\text{H3-H4})_2$  tetramer or the H2A-H2B dimers.



The average contact areas ( $CA_{av}$ ) of the DNA interface with the structured cores of  $(\text{H3-H4})_2$  (left) or both H2A-H2B dimers (right) are shown separately in hydrophobic (blue) and electrostatic (red) components. The thin vertical bars are the standard deviations calculated over the five simulations.

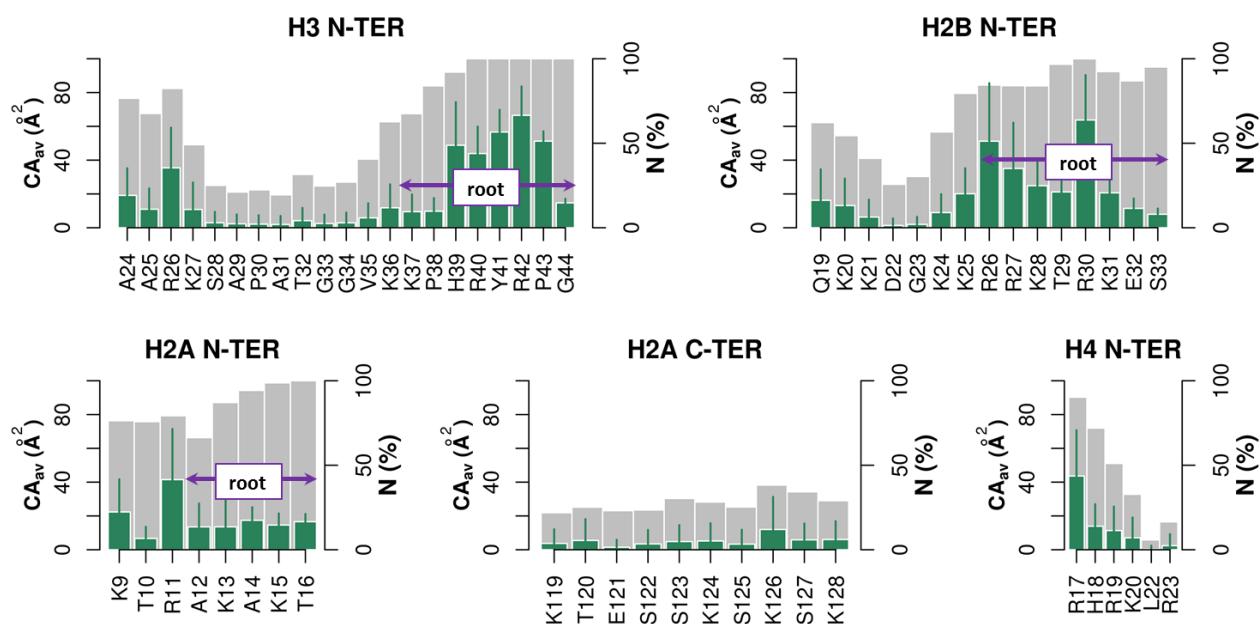


**Figure S9:** Comparison of the simulated interfaces involving the flexible parts of histone tails.



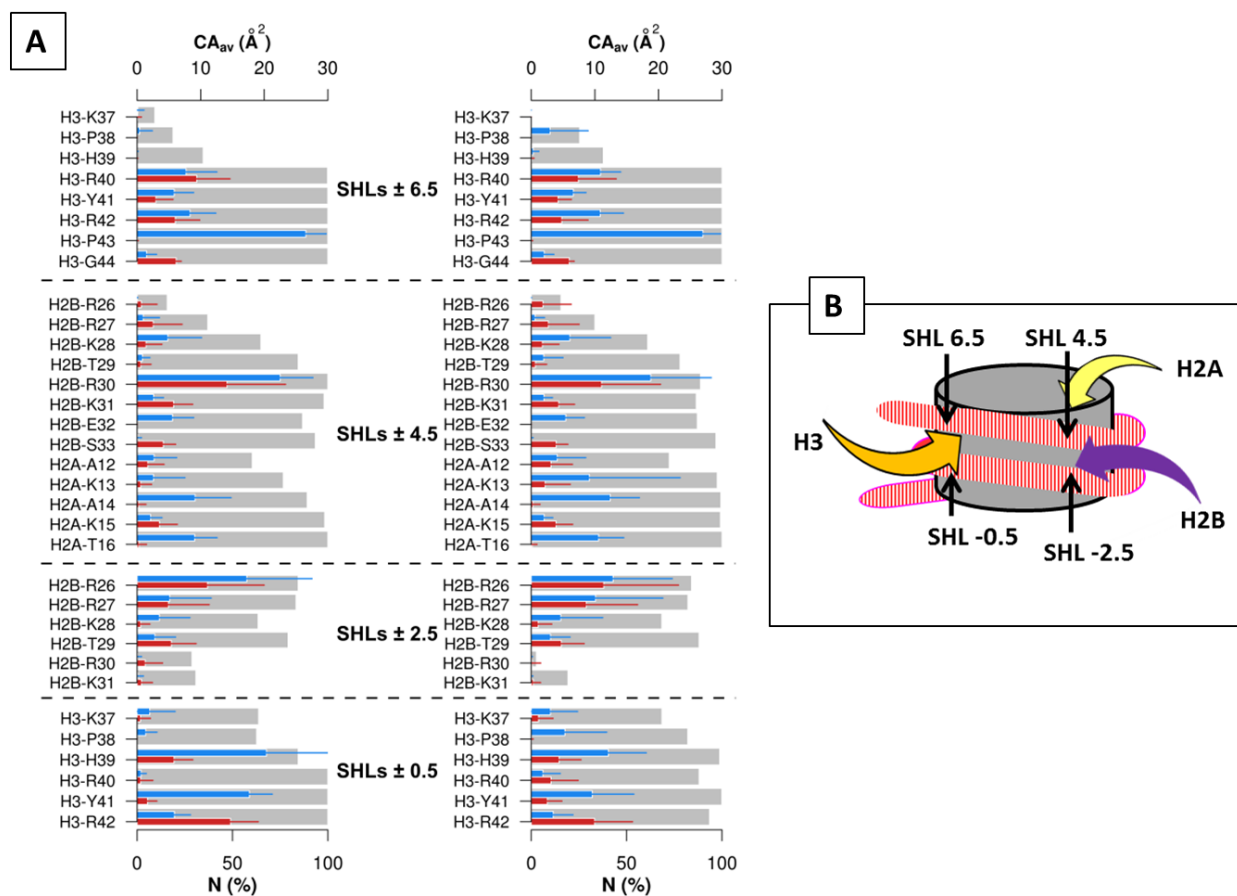
The contact areas ( $CA_{av}$ ) with DNA of the flexible tail extremities of H3, H2A H2B (defined in Table 2), of the H4 N-tails and of the H2A C-tail were extracted from SYS1, SYS1-bis, SYS2 and SYS2-bis simulations and averaged over time, considering all histone copies separately. The data from different simulations are colored according to the code given on top. The vertical thin bars are standard deviations. The N and C subscripts of H2A chains label the N- and C-tails, respectively.

**Figure S10:** Detailed interface between histone tails and DNA.



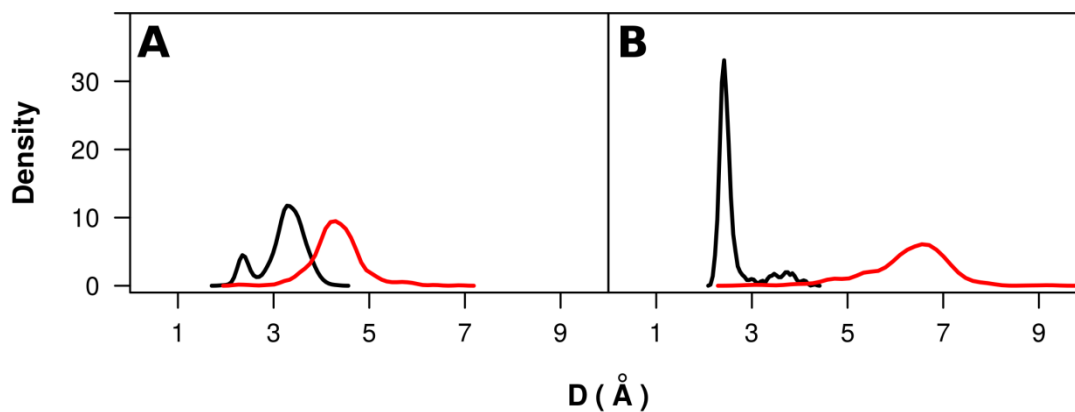
These plots present the histone tail amino acids of H3, H4, H2A and H2B that are involved in the DNA-histone interface. Each contacted amino acid of each histone is characterized by its total average area ( $CA_{av}$ , in green, vertical thin bars for standard deviations) and its occurrence (N%), represented by shaded gray area. The tags and arrows delimit the H3, H2A and H2B tail roots. The data of the N-terminal tails were averaged on the four simulations, SYS1, SYS2, SYS1-bis and SYS2-bis, and the pair of equivalent copies in each system. The contacts of the H2A C-terminal tail were averaged on SYS1-bis and SYS2-bis, the two simulations where these tails are present.

**Figure S11:** Interface between DNA and histone tail roots.



A: Amino acids of the tail roots of H3, H2B and H2A (defined in Table 2) involved at eight SHLs located symmetrically with respect to the DNA center; left and right plots correspond to negative and positive SHLs, respectively. Each amino acid is characterized by its hydrophobic (blue) and electrostatic (red) average area ( $CA_{av}$ , horizontal thin bars for standard deviations) and its occurrence (N%), represented by shaded gray area. The data were calculated and averaged on snapshots extracted from the four simulations, SYS1, SYS2, SYS1-bis and SYS2-bis. B: Schematic representation of the nucleosome indicating the SHLs contacted by one copy of the H3, H2A and H2B tail roots.

**Figure S12:** Density plots of  $\text{Na}^+$  - DNA and  $\text{Na}^+$  - histones distances.



The distances between  $\text{Na}^+$  and DNA (black) or histones (red) were systematically calculated for each ion along the simulations. This figure presents two examples of density plots of these distances for ions characterized by low atomic fluctuations. A: at SHL -1.5, one  $\text{Na}^+$  stays close to both DNA and histone. B: at SHL -2.5, another  $\text{Na}^+$  interacts with the DNA but remains distant from the histones. The data were extracted from SYS1-bis and SYS2-bis for A and B, respectively.

**Table S1:** Sequences of histone tails.

| <b>H3 N tail</b>  |     |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
|-------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1                 | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| A                 | R   | T   | K   | Q   | T   | A   | R   | K   | S   | T  | G  | G  | K  | A  | P  | R  | K  | Q  | L  | A  | T  |
| 23                | 24  | 25  | 26  | 27  | 28  | 29  | 30  | 31  | 32  | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 |
| K                 | A   | A   | R   | K   | S   | A   | P   | A   | T   | G  | G  | V  | K  | K  | P  | H  | R  | Y  | R  | P  | G  |
| <b>H4 N tail</b>  |     |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| 1                 | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| S                 | G   | R   | G   | K   | G   | G   | K   | G   | L   | G  | K  | G  | G  | A  | K  | R  | H  | R  | K  | V  | L  |
| 23                | 24  |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| R                 | D   |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| <b>H2A N tail</b> |     |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| 1                 | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11 | 12 | 13 | 14 | 15 | 16 |    |    |    |    |    |    |
| S                 | G   | R   | G   | K   | Q   | G   | G   | K   | T   | R  | A  | K  | A  | K  | T  |    |    |    |    |    |    |
| <b>H2B N tail</b> |     |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| 1                 | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| A                 | K   | S   | A   | P   | A   | P   | K   | K   | G   | S  | K  | K  | A  | V  | T  | K  | T  | Q  | K  | K  | D  |
| 23                | 24  | 25  | 26  | 27  | 28  | 29  | 30  | 31  | 32  | 33 | 34 |    |    |    |    |    |    |    |    |    |    |
| G                 | K   | K   | R   | R   | K   | T   | R   | K   | E   | S  | Y  |    |    |    |    |    |    |    |    |    |    |
| <b>H2A C tail</b> |     |     |     |     |     |     |     |     |     |    |    |    |    |    |    |    |    |    |    |    |    |
| 119               | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 |    |    |    |    |    |    |    |    |    |    |    |    |
| K                 | T   | E   | S   | S   | K   | S   | K   | S   | K   |    |    |    |    |    |    |    |    |    |    |    |    |

The composition of the full length N- and C-terminal tails is spelled out for the four histone types. In SYS0 (which corresponds to 3MVD), the N-tails are limited to the short regions indicated here by a blue background. N-tail amino acids on yellow background were taken from 1KX5 and added to 3MVD in SYS1, SYS1-bis, SYS2 and SYS2-bis. The H2A C-tail from 1KX5 was integrated in SYS1-bis and SYS2-bis (in green).

**Table S2-1:** Hydrogen bonds between DNA and histone structured cores in 1KX5 and MDs

|               |     |             | 1KX5      |                   | MDs       |
|---------------|-----|-------------|-----------|-------------------|-----------|
|               |     |             | Direct hb | Water mediated hb | Direct hb |
| SHL $\pm$ 5.5 | H2A | <b>T 76</b> | +         | -                 | +         |
|               |     | <b>R 77</b> | +         | -                 | +         |
|               |     | K 75        | -         | -                 | +         |
|               | H2B | <b>Y 39</b> | +         | -                 | +         |
|               |     | S 52        | -         | -                 | +         |
|               |     | <b>I 51</b> | -         | +                 | +         |
| SHL $\pm$ 4.5 | H2A | <b>R 17</b> | +         | -                 | +         |
|               |     | R 20        | -         | -                 | +         |
|               |     | <b>G 28</b> | -         | +                 | +         |
|               |     | <b>R 29</b> | +         | -                 | +         |
|               |     | <b>R 32</b> | +         | -                 | +         |
|               | H2B | R 30        | +         | -                 | -         |
|               |     | K 31        | +         | -                 | -         |
|               |     | S 33        | +         | -                 | -         |
|               |     | <b>I 36</b> | -         | +                 | +         |
|               |     | <b>Y 37</b> | -         | +                 | +         |
| SHL $\pm$ 3.5 | H2A | <b>R 35</b> | +         | -                 | +         |
|               |     | <b>R 42</b> | +         | -                 | +         |
|               |     | <b>V 43</b> | +         | -                 | +         |
|               |     | <b>A 45</b> | +         | -                 | +         |
|               | H2B | <b>R 83</b> | +         | -                 | +         |
|               |     | <b>S 84</b> | +         | -                 | +         |
|               |     | <b>T 85</b> | +         | -                 | +         |
|               |     | R 89        | -         | -                 | +         |
|               | H4  | K 77        | -         | -                 | +         |

This table summarizes the amino acids that form hydrogen bonds with DNA in 1KX5 and in MDs. The data, sorted by SHL and histone type, come from the Supplementary Materials of Davey and coll. (2) for 1KX5 or from our analyses for the simulations. Direct or water mediated hydrogen bonds are reported for 1KX5. However, water mediated hydrogen bonds are listed only if they are retrieved in the form of direct hydrogen bonds in the simulations (gray background). Because there is no stable water mediated hydrogen bond in MDs, only direct interactions are informed here. The amino acids in bold make hydrogen bonds in both 1KX5 and MD snapshots. The signs “+” and “-” indicate the presence or the absence of hydrogen bonds, respectively.

**Table S2-2:** Hydrogen bonds between DNA and histone structured cores in 1KX5 and MDs*Table continued from previous page*

|               |    |              | 1KX5      |                   | MDs       |
|---------------|----|--------------|-----------|-------------------|-----------|
|               |    |              | Direct hb | Water mediated hb | Direct hb |
| SHL $\pm$ 2.5 | H3 | <b>R 72</b>  | +         | -                 | +         |
|               |    | <b>R 83</b>  | +         | -                 | +         |
|               |    | <b>F 84</b>  | +         | -                 | +         |
|               |    | <b>Q 85</b>  | -         | +                 | +         |
|               |    | <b>S 86</b>  | +         | -                 | +         |
|               | H4 | <b>R 78</b>  | -         | +                 | +         |
|               |    | <b>K 79</b>  | +         | -                 | +         |
|               |    | <b>T 80</b>  | +         | -                 | +         |
| SHL $\pm$ 1.5 | H3 | <b>R 63</b>  | +         | -                 | +         |
|               |    | <b>K 64</b>  | +         | -                 | +         |
|               |    | <b>L 65</b>  | +         | -                 | +         |
|               |    | <b>R 69</b>  | +         | -                 | +         |
|               | H4 | <b>T 30</b>  | -         | +                 | +         |
|               |    | <b>K 31</b>  | -         | +                 | +         |
|               |    | <b>R 36</b>  | +         | -                 | +         |
| SHL $\pm$ 0.5 | H3 | R 40         | +         | -                 | -         |
|               |    | Y 41         | +         | -                 | -         |
|               |    | G 44         | +         | -                 | -         |
|               |    | <b>T 45</b>  | -         | +                 | +         |
|               |    | V 46         | -         | -                 | +         |
|               |    | <b>A 47</b>  | +         | -                 | +         |
|               |    | <b>K 115</b> | -         | +                 | +         |
|               |    | <b>V 117</b> | +         | -                 | +         |
|               |    | <b>T 118</b> | +         | -                 | +         |
|               | H4 | <b>R 35</b>  | +         | -                 | +         |
|               |    | <b>R 39</b>  | -         | +                 | +         |
|               |    | <b>R 45</b>  | -         | +                 | +         |
|               |    | <b>I 46</b>  | +         | -                 | +         |
|               |    | <b>S 47</b>  | -         | +                 | +         |
| <b>G 48</b>   | +  | -            | +         |                   |           |

**Table S3:** Time occurrence of Na<sup>+</sup> cations at the DNA histone interface.

|          | SHL  |     |      |     |      |     |      |     |      |     |      |     |      |     |
|----------|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|
|          | -6.5 | 6.5 | -5.5 | 5.5 | -4.5 | 4.5 | -3.5 | 3.5 | -2.5 | 2.5 | -1.5 | 1.5 | -0.5 | 0.5 |
| SYS0     | 0    | 0   | 6    | 47  | 12   | 84  | 55   | 28  | 3    | 93  | 12   | 4   | 99   | 3   |
| SYS1     | 0    | 0   | 16   | 27  | 100  | 87  | 36   | 20  | 80   | 96  | 0    | 0   | 98   | 0   |
| SYS1-bis | 0    | 0   | 1    | 0   | 84   | 61  | 0    | 10  | 99   | 100 | 0    | 0   | 92   | 0   |
| SYS2     | 0    | 0   | 13   | 60  | 8    | 0   | 0    | 2   | 0    | 74  | 0    | 1   | 96   | 13  |
| SYS2-bis | 0    | 0   | 2    | 46  | 0    | 100 | 89   | 6   | 41   | 3   | 32   | 34  | 100  | 0   |

This table reports the percentage of simulation time during which one ion stays at each specified SHL, close to both DNA ( $D_{\text{Na}^+-\text{DNA}} \leq 4\text{\AA}$ ) and histone ( $D_{\text{Na}^+-\text{histone}} \leq 6\text{\AA}$ ). The data are presented for each of the five simulations analyzed here.

**Table S4:** DNA sequences at contacted SHLs.

|          |                    |                      |                    |                  |                    |                    |
|----------|--------------------|----------------------|--------------------|------------------|--------------------|--------------------|
| SHL      | -0.5               | -1.5                 | -2.5               | -3.5             | -4.5               | -5.5               |
| Sequence | CGTACG<br>GCATGC   | GCTTAAA<br>CGAATTT   | TCTAGCA<br>AGATCGT | TCGTAG<br>AGCATC | GCTCAA<br>CGAGTT   | TGCCGAG<br>ACGGCTC |
| SHL      | 0.5                | 1.5                  | 2.5                | 3.5              | 4.5                | 5.5                |
| Sequence | TCCCCCG<br>AGGGGGC | TTTTAACC<br>AAAATTGG | CAAGGGG<br>GTTCCCC | TCCCTA<br>AGGGAT | CCAGGCA<br>GGTCCGT | TGTCAGA<br>ACAGTCT |

This table reports the nucleotide composition of the SHL sites that are contacted by each copy of histone structured cores in the two complementary DNA strands (5'→3' for the first line and 3'→5' for the second line).

### References in Figure S5 and Table S2

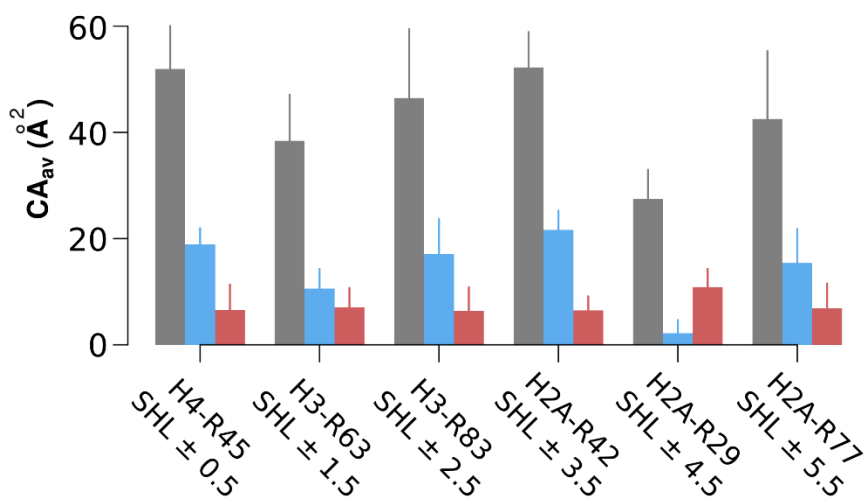
1. Zhou B-R, Feng H, Ghirlando R, Kato H, Gruschus J, Bai Y. Histone H4 K16Q mutation, an acetylation mimic, causes structural disorder of its N-terminal basic patch in the nucleosome. *J Mol Biol.* 2012 Aug 3;421(1):30–7.
2. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol.* 2002 Jun 21;319(5):1097–113.



## Conclusion et développements en cours

Dans ce premier article, grâce à VLDM, j'ai pu décrire sur des bases solides et objectives l'interface ADN/histone en termes de surface et d'occurrence de contact. J'y mets en évidence un réseau d'interaction très dense permettant le maintien d'un ADN courbé. La stricte symétrie de l'interface ainsi définie reste un résultat relativement inattendu puisque la séquence 601 n'est pas palindromique. Ce résultat suggère qu'une fois l'interface formée, la composition nucléotidique de l'ADN intervient peu dans le complexe. Or, il est clair que la séquence impacte le nucléosome, au moins lors de l'assemblage, puisque toutes les séquences d'ADN ne forment pas de nucléosome avec la même affinité. Afin de comprendre ce phénomène, j'ai travaillé sur des données expérimentales d'assemblage et de désassemblage du nucléosome *in vitro*, présenté dans la section suivante (Article 2).

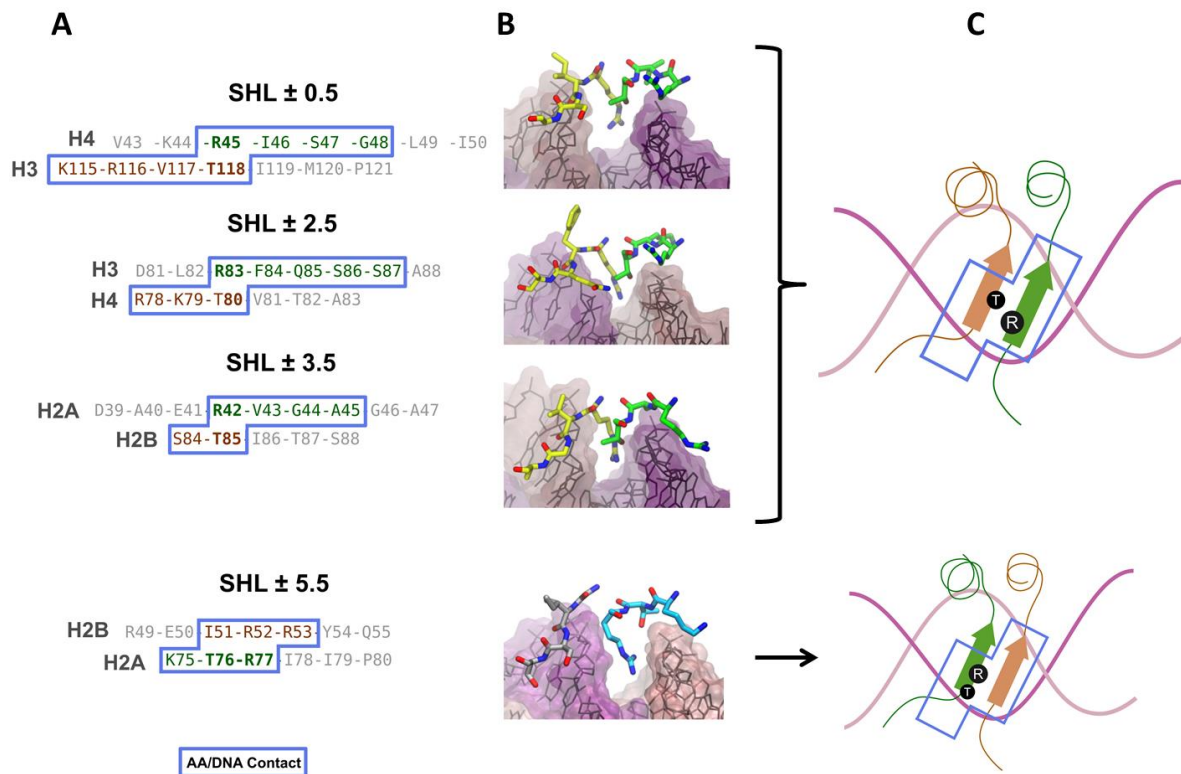
Pour compléter la compréhension de l'interface, je travaille actuellement sur l'analyse plus fine des sites de fixations impliquant des arginines qui interagissent avec les petits sillons de l'ADN aux SHLs +/- 5.5, 4.5, 3.5, 2.5, 1.5 et 0.5 (Figure 12, p. 25).



**Figure 38 – Interfaces arginines/petits sillons de l'ADN.** Les surfaces de contacts moyenne ( $CA_{AV}$ ) sont calculées sur les quatre simulations des systèmes SYS1, SYS1-bis, SYS2 et SYS2-bis (p. 55) pour tous les contacts (gris), les contacts hydrophobes (bleue) et les contacts électrostatique (rouge).

Avec VLDM, j'ai effectué la mesure précise des contacts entre les arginines et l'ADN et j'observe des interactions très stables, présentes tout le long des trajectoires. On note également une prédominance des contacts hydrophobes, à l'exception de l'arginine au site SHL +/- 4.5 (Figure 38).

Les résultats des analyses montrent que les sites contactés par les boucles L1-L2 (Interface ADN/histones, p. 23) (SHLs +/- 5.5, 3.5, 2.5 et 0.5) favorisent l'insertion des arginines dans le petit sillon, cette insertion se produisant de 72% à 100% du temps le long des dynamique. Au contraire, les arginines des hélices  $\alpha 1-\alpha 1$  (SHLs +/- 3.5, 2.5) semblent moins faciles à s'insérer, avec respectivement 36 et 2% d'insertion dans le petit sillon.



**Figure 39 – Structure des pinces d'histones.** A) Séquences protéiques alignées selon les résidus qui se font face dans le petit sillon. B) structure correspondante en interaction avec le petit sillon de l'ADN dans le nucléosome. C) Représentation schématique de l'interaction pour les sites SHLs +/- 0.5, 2.5 et 3.5 (panneau du haut) et les sites aux SHL +/- 5.5 (panneau du bas).

En m'intéressant plus particulièrement aux 8 arginines qui s'insèrent dans le petit sillon, j'ai observé qu'elles étaient caractérisées par un réseau, impliquant plusieurs autres résidus protéiques qui forment une pince protéique autour du petit sillon (Figure 39). Dans 6 cas, ce réseau est constitué par deux fragments de différentes histones, l'un contenant l'arginine et l'autre une thréonine. Il présente une variante pour les sites SHLs +/- 5.5 où l'arginine et la thréonine sont situées sur la même histone H2A. Malgré cette différence, l'arginine associée à ces sites présente un très bon taux d'insertion (88 %).

# Chapitre 2 : Assemblage et désassemblage du nucléosome

## Sommaire <sup>2</sup>

L'assemblage et le désassemblage du nucléosome sont au centre de la régulation de l'expression génique. Pour permettre la réplication, la transcription ou encore la recombinaison, la chromatine doit être très dynamique dans le sens où les protéines doivent pouvoir accéder à l'ADN. C'est l'assemblage et le désassemblage du nucléosome qui vont ainsi permettre la régulation de l'accessibilité de l'ADN.

*In vivo*, l'assemblage et le désassemblage du nucléosome ont lieu en présence de toute une série de facteurs, comme par exemple les célèbres facteurs de remodelage. Cependant, pour mieux comprendre la biophysique de ces mécanismes, nous nous sommes intéressés à l'assemblage et au désassemblage du nucléosome *in vitro*, en l'absence de facteurs protéiques.

La méthode la plus utilisée pour étudier l'assemblage et le désassemblage du nucléosome *in vitro* est de faire varier la concentration en NaCl. Une diminution de la force ionique de ~2M à 0.15M NaCl accompagne l'assemblage, alors que l'augmentation de la concentration en sel va correspondre au désassemblage. A 2M, l'ADN est libre d'histone ; à 0.15M, le nucléosome est complètement formé, du moins dans le cas de séquences positionnantes, de type séquence 601. Faire varier la force ionique par paliers permet d'isoler d'éventuels états intermédiaires.

Des études très précoces avaient montré que le tétramère (H3-H4)<sub>2</sub> se fixait sur l'ADN avant les dimères H2A-H2B. Différentes études combinant la méthode décrite précédemment avec des expériences telles que le FRET ou encore la microscopie à force atomique à haute vitesse ont établi que le désassemblage était séquentiel avec d'abord la dissociation des dimères H2A-H2B suivi de la dissociation du tétramère (H3-H4)<sub>2</sub>. Akiko Hatakeyama, Claude Nogues et Malcolm Buckle ont utilisé une méthode qu'ils ont mis au point au laboratoire (Hatakeyama et al. 2016), la méthode PhAST (p. 61) pour étudier l'assemblage du nucléosome en détail pour la première fois. Dans le travail que je présente ici, la même équipe a voulu comparer l'assemblage et le désassemblage dans des conditions similaires, sur un nucléosome constitué avec la séquence 601 non palindromique qui a servi à mon analyse de l'interface dans mon premier travail. Comme expliqué précédemment (p.

---

<sup>2</sup> L'ensemble des références de cette section peuvent être trouvées dans l'article associé au chapitre (p. 107) et dans l'introduction de ce manuscrit (Le nucléosome, p. 17)

61), PhAST donne des informations sur la structure locale d'un ADN, à la résolution du dinucléotide NpN, sur la base de signaux reflétant la probabilité de formation d'un dimère YpY sous irradiation UV. La variation entre les signaux obtenus sur l'ADN libre et ADN plus ou moins fermement complexé avec des histones a été suivi à différentes concentrations (1.5M, 1.0M, 0.5M et 0.1M) de NaCl (Article 2 – Figure S1) lors d'expériences consacrées soit à l'assemblage soit au désassemblage (Article 2 – Figure 3). Comme nous l'avons déjà vu, le signal PhAST n'est pas directement impacté par le contact ADN/histone mais par le fait que la complexation des deux partenaires induit des modifications structurales locales, notamment du *roll*.

Mon rôle dans cette étude était d'interpréter ces variations expérimentales issues de PhAST grâce à mes résultats sur l'interface (Article 2 – Figure 2). Contrairement à l'interface que nous avons décrite symétrique dans l'article 1, les signaux liés à l'assemblage et au désassemblage présentent des asymétries entre les moitiés 5' et 3' de la séquence 601 (Article 2 – Figure 4). L'analyse révèle un effet de la séquence sur l'interaction avec les histones, la moitié 5' de l'ADN nucléosomal étant plus favorable que la moitié 3' à interagir avec les histones lors de l'assemblage et du désassemblage. Ces résultats sont en accord avec des travaux menés au laboratoire qui avaient permis de caractériser les propriétés de structure et de flexibilité des ADN libres qui jouent en faveur de la formation du nucléosome. Ces travaux avaient ainsi proposé que l'alternance de régions A:T et G:C riches qui est extrêmement marquée le long de la moitié 5' de la séquence 601 rendait cette séquence particulièrement attractive.

# Nucleosome assembly and disassembly pathways.

---

Akiko Hatakeyama, Romain Retureau, Marco Pasi, Brigitte Hartmann, Claude Nogues and Malcolm Buckle.

LBPA, IDA, ENS Paris-Saclay, CNRS, Université Paris-Saclay, F-94235, Cachan, France

## Abstract

Nucleosome assembly and disassembly play a central role in the regulation of gene expression. Here we used PhAST (Photochemical Analysis of Structural Transitions) to understand the pathways of both nucleosome formation and dissociation by monitoring at a base pair level structural changes induced along the DNA upon histone binding or release at various ionic strengths. In addition to confirming the sequential binding/release of (H3-H4)<sub>2</sub> tetramer and H2A-H2B heterodimers, PhAST reveals multiple intermediate states, emerging during assembly or disassembly. These states are all characterised by a subtle asymmetry in histone/DNA interactions around the pseudo dyad axis; this concurs with the sequence dependent properties of free DNA governing its propensity to form nucleosomes. This study offers the first detailed view of both assembly and disassembly pathways of nucleosomes *in vitro* and, most importantly, describes how the DNA sequence influences the structural organisation of nucleosome intermediate states.

## Introduction

The fundamental repeating unit of chromatin is the nucleosome; 145-147 base pairs (bp) of DNA wrapped around an octamer of histone proteins (two H2A/H2B heterodimers and one (H3-H4)<sub>2</sub> tetramer), efficiently compacts genomes into cell nuclei and regulates many DNA functions. The spatial and temporal distribution of nucleosomes as a result of nucleosome assembly and disassembly is involved in all DNA transactions. *In vivo*, a plethora of interplaying factors such as chaperones, remodelling complexes, histone variants, epigenetic modifications and intrinsic, sequence dependent, DNA properties orchestrate the dynamics of nucleosome positioning. Understanding how this complex nuclear machinery operates requires extensive knowledge of the structural pathways adopted by nucleosomes during assembly and disassembly events.

In recent years a large amount of interest has been paid to the disassembly mechanism. Studies were carried out *in vitro* by recording the response of preformed nucleosomes to a gradual increase in ionic strength<sup>1-10</sup> or, less frequently, to an external mechanical force applied to the nucleosome DNA<sup>11,12, 13</sup>. The examination of spontaneous disassembly focused on unwrapping of DNA in the peripheral regions (so-called “DNA breathing”)<sup>14-18</sup>. Fluorescence-based techniques were used to dissect the stepwise disassembly of nucleosomes, each experiment providing data about specific histone-histone, DNA-histone or, eventually, DNA-DNA proximities. For example, the use of three or four pairs of fluorophores gave insights into the behaviour of the DNA’s extremities<sup>1</sup> or interior DNA regions<sup>5, 6, 9</sup> with respect to H4 and H2B, as well as providing information about histone-histone proximities<sup>2</sup>.

In theory, two global pathways for nucleosome disassembly may be envisaged, either dissociation of the octamer as a single entity from the DNA or sequential release of histones or groups of histones; the second hypothesis is now generally accepted. A series of studies of salt-induced dissociation based on FRET (Forster Resonance Energy Transfer) approaches<sup>1, 2, 4, 5, 6, 9, 10</sup> or TR-SAXS (Time-Resolved Small Angle X-ray Scattering)<sup>8</sup> provided coherent arguments in favour of a pathway with two major successive phases: a first step leading to the release of H2A-H2B dimers and a second, distinct step corresponding to (H3-H4)<sub>2</sub>-DNA dissociation. Such a global two-step scheme is likely to be general since it was observed in nucleosomes studied under identical conditions but containing different DNA sequences, *i.e.* 601- and 5S-nucleosomes<sup>7, 4</sup> or 601-, 5S- and MMTV-nucleosomes<sup>3, 6</sup>.

Although technically not straightforward, the existence of very early states of disassembly, before the removal of H2A-H2B, was also examined. With the 5S-nucleosome, the two H2A-H2B dimers were observed to dissociate from the (H3-H4)<sub>2</sub>-DNA complex in a single transition without observable stable intermediates<sup>1, 7</sup>. FRET experiments with 601-nucleosomes proposed that disruption or weakening of the interface between H2A-H2B dimers and (H3-H4)<sub>2</sub> tetramer interface (the so-called “butterfly” state), helped to rupture the DNA / H2A-H2B interfaces<sup>5, 6, 10</sup>. According to models inferred from experiments using SAXS<sup>7, 8</sup>, FRET<sup>12, 8, 10</sup> and single molecule unwrapping associated with FRET<sup>12</sup>, the release of the two H2A-H2B dimers in 601-nucleosomes is asymmetric, starting from one

unwrapped DNA end (the so-called “J”-shaped state). A recent cryo-EM study captured structures related to first events of spontaneous disassembly of 601-nucleosomes<sup>18</sup>, which are hard to observe in solution because nucleosome open states are marginally populated<sup>14, 15</sup>. The cryo-EM structures showed again an asymmetric loss of contacts between H2A-H2B and DNA arising from the spontaneous breathing of one extremity of the DNA fragment; this first nucleosome opening gradually propagates in association with subtle histone rearrangements; the intermediate states, in which one H2A–H2B dimer is no longer visible, resemble those hexasomes (DNA bound to the (H3–H4)<sub>2</sub> tetramer and one H2A–H2B dimer) obtained from SAXS<sup>19</sup> or X-ray<sup>20</sup> data. In all the studies presented above, the asymmetric opening was considered symptomatic of a DNA sequence effect since the strict symmetry of the histone structured domains with respect to the dyad axis<sup>21, 22</sup> cannot account for such phenomena. On the basis of salt titrations<sup>21</sup> and single molecule experiments<sup>11</sup>, it was proposed that the 601 sequence is constituted by “strong” left and “weak” right halves<sup>8</sup>; strong and weak sides may relate to differences in DNA intrinsic flexibility<sup>23, 24, 12</sup>.

A general DNA sequence effect on disassembly was further attested by the fact that 601-nucleosomes better resist chaotropic destabilisation than nucleosomes formed with other sequences<sup>7, 4, 3, 6</sup>. In contrast, nucleosome stability, characterised by the ionic strength at the midpoints of the successive transitions during disassembly, is not affected by the histone origin<sup>6</sup> likely because of the very high degree of conservation of histone sequence and folding<sup>25</sup>.

The large number of studies presented above provided information exclusively on nucleosome disassembly; states preceding H2A–H2B dimer release are better characterised than later events leading to the complete dissociation of the complex. By comparison, assembly remained almost completely ignored apart from initial studies during the 1980’s<sup>26, 27</sup>, a situation that prevented any global approach to understanding both nucleosome association and dissociation. In view of this, we embarked on studies to capture nucleosome intermediate states during NaCl-induced assembly and disassembly (Schema 1) of the 601-nucleosome and to compare the two pathways. We used the PhAST (Photochemical Analysis of Structural Transitions) technique developed in our group<sup>28</sup>.

PhAST possesses the useful property of being able to follow both processes from the DNA point of view and to provide structural information at base pair resolution, independent of conditions. This non-invasive technique can thus be applied to freely diffusing macromolecules in solution, without introducing structural artefacts caused for instance by bulky hydrophobic fluorophores<sup>2</sup>. It is based on the measurement of the probability to form, on the same DNA strand, UV photo-induced cyclobutane dimers between adjacent pyrimidines (YpY dimer, linked by C5-C5 and C6-C6 bonds)<sup>28-30</sup>. In addition to the quantum yields specific to each type of step, the probability of forming YpY dimers depends on the local YpY structure, more precisely on the inter base-pair parameters of roll and twist that are coupled in B-DNA in solution<sup>31</sup>. Thus, low twist and positive roll shorten the YpY C5-C5 and C6-C6 distances and thus favour dimer formation whereas large twist and negative roll has the inverse effect



<sup>28</sup>. The YpY dimer probabilities along a DNA sequence therefore reflect its local average structure. Comparison between probabilities collected on free and bound DNA simultaneously reveals DNA structural changes induced by the presence of proteins.

PhAST proved to be remarkably efficient in following structural changes in DNA as 601- and derived 601-nucleosomes were formed under decreasing ionic strength conditions, *i.e.* during nucleosome assembly <sup>28</sup>. We demonstrated first, the intuitive idea that histone binding induces noticeable structural changes in the local parameters of roll and twist all along the 601 sequence. Then we observed that nucleosome formation starts with the binding of (H3/H4)<sub>2</sub> and ends with the recruitment of H2A-H2B dimers, in agreement with earlier studies <sup>26, 27</sup>. An important contribution of PhAST was to detect and describe for the first time additional steps occurring during nucleosome formation. The structural organisation of nucleosome intermediate states reflected the existence of marked DNA sequence effects that could be unambiguously assigned to specific DNA regions. For example, we highlighted the more robust interaction of (H3/H4)<sub>2</sub> with the 5' side of the 601 sequence than with its 3' side counterpart, a result explained by the experimental asymmetries found in the dynamic properties of the free sequence <sup>23, 24</sup>. That such subtle events could be captured by PhAST encouraged us to further describe how the 601-nucleosome disassembles and thus be able to compare assembly and disassembly processes.

We first examined the effect of potential induced structural perturbations on free DNA of high salt concentrations <sup>32</sup> by considering data collected on the naked 601 sequence at various ionic strengths and comparing the YpY reactivities. We then catalogued differences in YpY reactivities between free and bound DNA thus obtaining a description of DNA structural changes during increased or decreased ionic strength. To interpret the structural changes in the local DNA configuration in terms of strengthening or weakening (ultimately, binding or release) of DNA/histone interactions, we used a very fine mapping of contacts between DNA and both structured and unstructured histone regions obtained from exhaustive simulations in explicit solvent <sup>33</sup>. A parallel between DNA structural changes and DNA/histone interactions led to the identification and characterisation of a series of predominant nucleosome intermediate states. The pathways of the two processes of assembly and disassembly were thus interpreted in terms of structural organisation of nucleosome intermediate states associated with identical or different ionic strengths.

## Results

PhAST generates YpY dimers in DNA using laser photo-radiation; the dimer detection technique produces peaks representing the probabilities of dimer formation; the quantification consists of measuring the peak amplitudes which we will call intensities ( $I$ ) (for calculation of these see Materials and Methods). The intensity ( $I$ ) reflects the DNA local structure, as reported in the Introduction. By comparing data collected on bound and free DNA, PhAST therefore reveals the structural effects of protein on DNA. In the two possible end situations, intensities ( $I$ ) correspond to i) free DNA at high salt concentrations ( $I_f$ ) or ii) DNA fully engaged in a nucleoprotein complex ( $I_b$ ), typically at 0.1 M NaCl. Intermediate values characterise nucleosome intermediate states.

A first series of PhAST experiments were carried out starting from a mixture of free DNA and histones and then gradually decreasing the ionic strength (Schema 1). This nucleosome assembly study has been described in a previous publication<sup>28</sup> and the corresponding intensities ( $I$ ) are reused here. In a new, second series, preformed nucleosomes were submitted to a progressive increase of the ionic strength (nucleosome disassembly, Schema 1). A second approach consisted of transferring the preformed nucleosomes directly by a single step to different NaCl concentrations: the two approaches gave very similar results (Figure S1), and the data presented here are from the two types of experiments considered together. In all cases, free DNA was in parallel photo-irradiated in the absence of histones.



*Schema 1: Outline of the PhAST experiments monitoring successive DNA structural changes induced during nucleosome assembly and disassembly by using stepwise decrease and increase of NaCl concentration.*

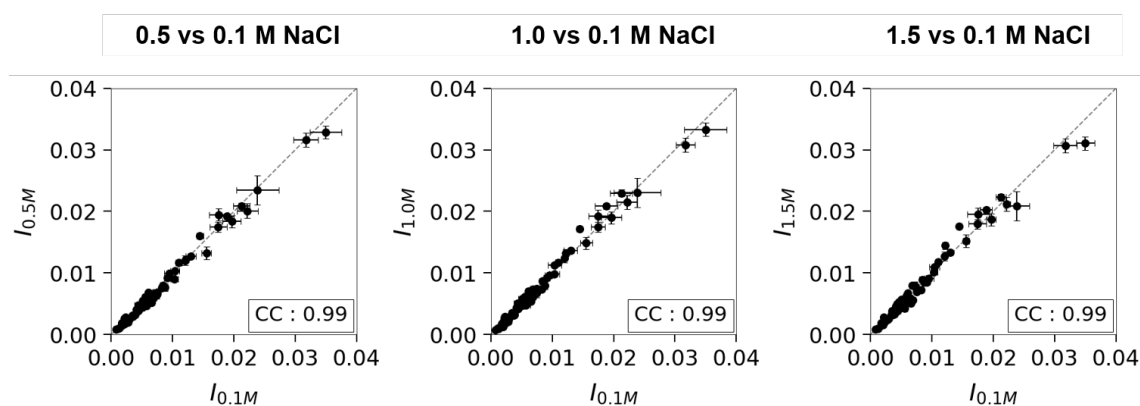
Before presenting our analysis of differences in PhAST signals associated with variations in salt concentration and their interpretation in terms of changes in DNA/histone interactions, we focus on the free DNA to ensure that ionic strength variations in the range used here do not perturb its structure.

### PhAST profiles of naked DNA at various NaCl concentrations

Our studies of nucleosome assembly and disassembly used salt concentrations that classically range from 0.1 to 1.5 M NaCl. Variations within this range modify the DNA melting temperature<sup>34</sup> without strongly perturbing the gyration radius or persistence length of the B-form of DNA<sup>35, 36, 37, 38</sup>. However, to our knowledge, there is no report on a possible effect of such moderate ionic strength variations on the local structure of the B-DNA double helix at room temperature. More specific to the PhAST

approach, the impact of the salt concentration on the quantum yield of YpY dimer formation is also unknown. So, we compared the PhAST signals collected on the 601 sequence at 0.5, 1.0 or 1.5 M NaCl to those data obtained at 0.1 M NaCl (Figure 1), which is considered here as a relevant reference for a physiological solution.

Before commenting on the extremely good correlation between the intensities of dimer formation as a function of NaCl concentration ( $CC = 0.99$ , Figure 1) we focus on two steps that were extremely sensitive to laser photo-radiation regardless of the ionic strength (Figure 1, intensity ( $I$ ) > 0.03). These “hotspots” correspond to TpT dinucleotides immediately 5' to TpA in the two TTTAA segments present in the 601 sequence. According to an NMR study <sup>24</sup>, the free TTTAA block is associated with low twists and positive rolls that both favour YpY dimer formation. In addition, the two adenines facing the photo-reactive TpT step show uncommon behaviour that includes the resonance broadenings of their H2 and H8 protons as well as an exceptional sensitivity of their <sup>31</sup>P chemical shifts to temperature changes. The enhanced photo-reactivities associated with these steps could be a further example of the ability of PhAST to detect unusual structural features in DNA.



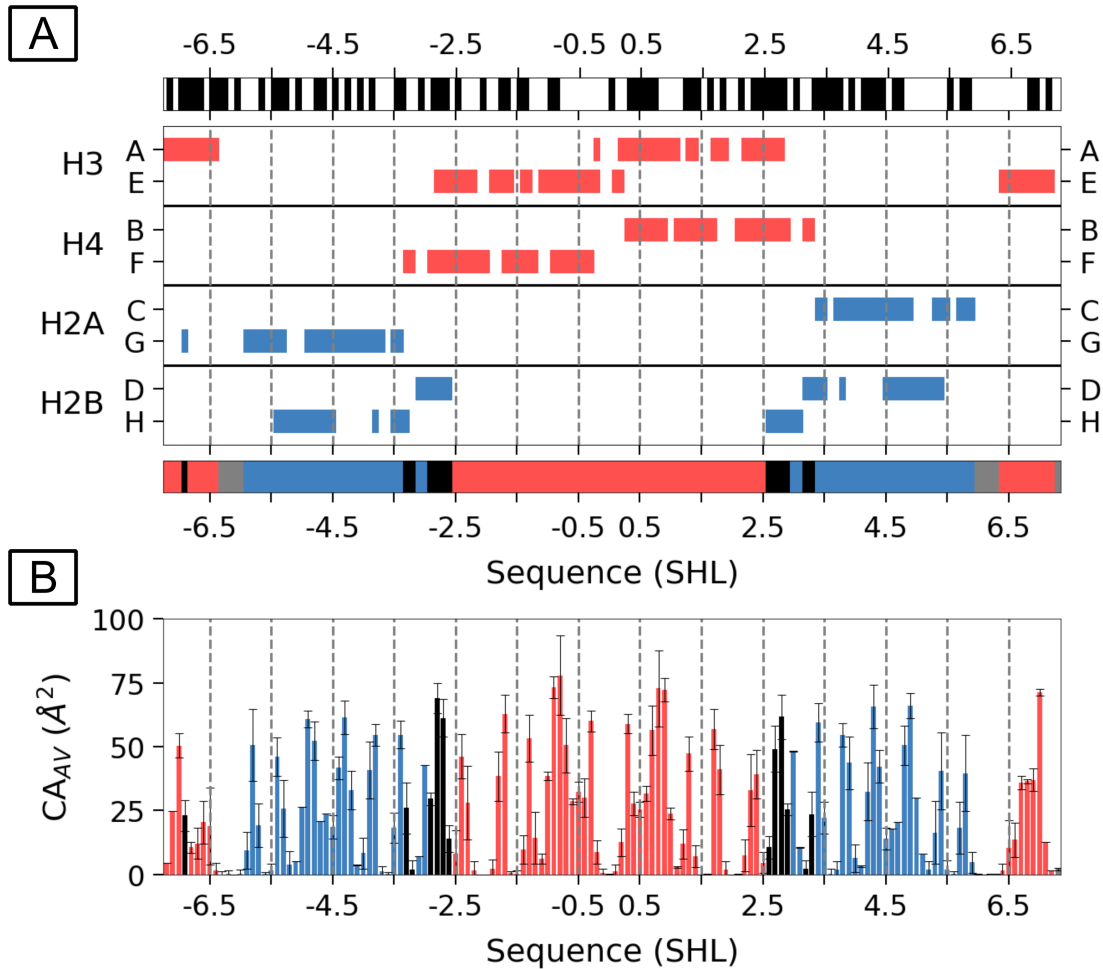
**Figure 1: Comparison of PhAST signals at various ionic strengths.** The intensity ( $I$ ) is expressed in terms of normalised peak height associated with each YpY step of the naked 601 sequence. The data collected at 0.1 M NaCl were systematically compared to those at 0.5, 1.0 and 1.5 M. Each point represents the value of  $I$ , averaged from 6 experiments related to assembly and disassembly studies; the vertical and horizontal bars are standard errors. The correlation coefficients ( $CC$ ) are given in boxes in each panel.

However, as stated above, the PhAST datasets at four NaCl concentrations were highly correlated and aligned according to  $x=y$  (Figure 1). This clear result showed that over the considered range, the salt concentration had a marginal effect on the quantum yield of YpY dimer formation and on the average local DNA structure. This conclusion that dismisses an eventual DNA perturbation is reassuring, given the large number of studies of nucleosome dissociation that implicitly presupposed a negligible salt effect on the DNA structure. In the context of the present work, differences between

PhAST signals collected at various ionic strengths can be confidently interpreted in terms of modifications of DNA/histone interactions. The next requirement is to have a precise description of the DNA/histone interface in solution.

### DNA/histone interface

A 1  $\mu$ s trajectory of molecular dynamics in explicit solvent was recently obtained on a nucleosome containing the 601 sequence and *Xenopus laevis* histones, including large parts of histone tails (see Materials and Methods). The snapshots, analysed with the VLDM program (see Materials and Methods), provided precise knowledge of DNA/histone interactions in solution<sup>33</sup>. Monitoring of the DNA-protein interface identified those DNA regions in which nucleotides of one or the other strand of the double helix interact with the histones H3, H4, H2A or H2B (Figure 2-A). The contribution of each base pair to the interface was also expressed in terms of contact areas involving the DNA and both structured and unstructured histone domains (Figure 2-B). Overall, the contact area associated with (H3-H4)<sub>2</sub> tetramer represent 54% of the total contact area, thus slightly larger than for H2A-H2B heterodimers. The DNA/histone interactions are remarkably symmetric with respect to the pseudodyad axis (Figure 2-B) in perfect resonance with the strict symmetry of the histone structured domains; this implies that the DNA sequence, which is not palindromic, has a marginal effect on the interface once the DNA is fully wrapped around the histone and a complete nucleosome is formed.



**Figure 2: The DNA/histone interface.** **A:** Schematic representation of DNA regions involved in the DNA/histone interface. The nucleotides interacting with the histone structured and unstructured domains are positioned along the 601 sequence, specifying the contacts with the different chains (A, B, C, etc...) of each histone type (H3 and H4 in red, H2A and H2B in blue). The lower banner summarises those regions contacted by H3-H4, H2A-H2B, both H3-H4 and H2A-H2B (black) or none of them (grey). The DNA sequence is labeled by SHL (Super Helical Location, defined in Material and Methods). The dotted lines correspond to intervals of 1 SHL. In the upper banner, the location of YY steps along the two strands of the 601 sequence is pictured by black bars. **B:** Quantification of the DNA/histone interactions. The average contact areas ( $CA_{av}$ ) quantified the DNA/histone interactions at the base pair level of DNA regions, and uses the same color code as in panel A. The vertical error bars are standard deviations over the simulation trajectory. The data presented in panels A and B were extracted from the VLDM analysis of structures from a  $1\mu s$  simulation of a 601-nucleosome.

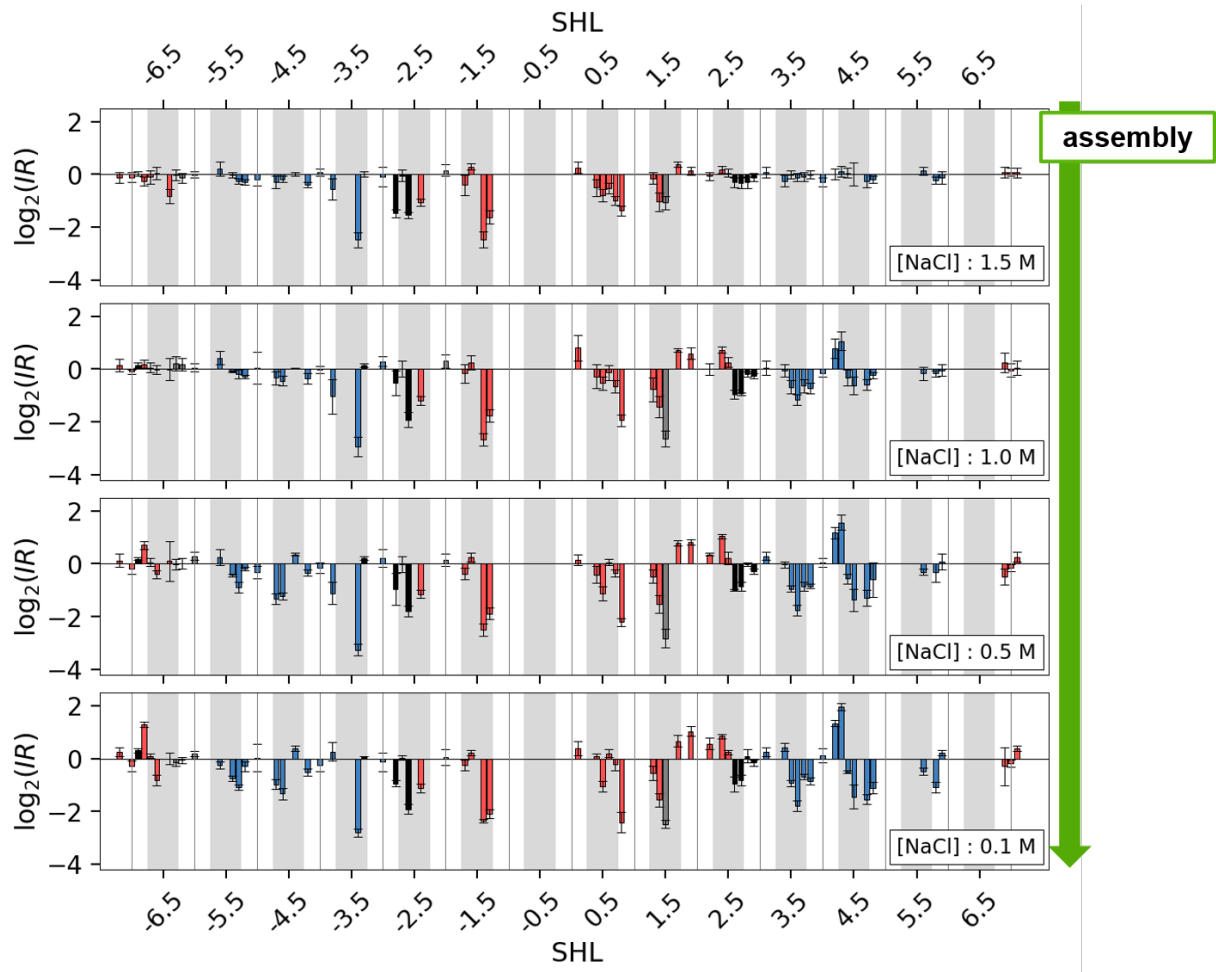
On the one hand, there are no DNA regions that escape histone contact apart from a short fragment of 4 bp around  $SHL \pm 6.25$ . On the other hand, tracts of  $Y_n$  (n from 2 to 7) forming dimers

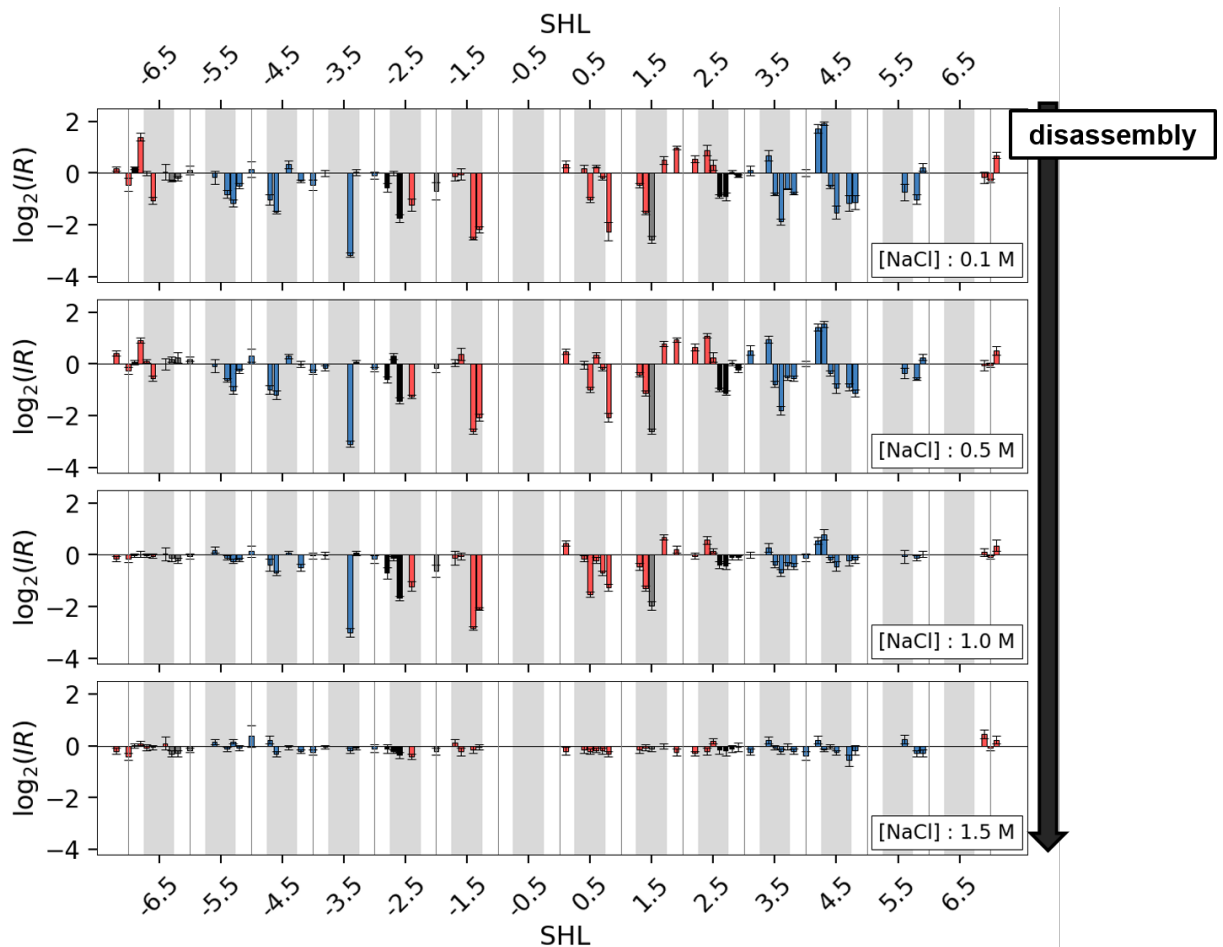
under photo-irradiation are regularly distributed all along the DNA sequence. This situation is clearly propitious to the detection of DNA/histone interface modifications using PhAST. Furthermore, the differentiation between the effects induced by the binding of (H3-H4)<sub>2</sub> tetramer or H2A-H2B dimers is facilitated by the fact that DNA regions where interactions with H3-H4 and H2A-H2B overlap are extremely limited (Figure 2).

### Assembly and disassembly processes

PhAST was applied to naked DNA and nucleosomes at four ionic strengths according to Schema 1. The PhAST signals from assembly experiments were obtained previously<sup>28</sup> and reused here to be compared to those of the disassembly experiments. For comparative purposes they are represented in Figure 3. Changes between PhAST signals of bound and free DNA are quantified by attributing to each YpY position a quantity,  $\log_2$  of Intensity Ratios, defined as  $\log_2(IR) = \log_2(\text{normalized intensity of a given peak in bound DNA} / \text{normalized intensity of the same peak in naked DNA})$ , see also Materials and Methods<sup>28, 30</sup>. It should therefore be borne in mind that when the  $\log_2(IR)$  value is negative this means that the photoreactive signal of the bound DNA ( $(I)_{\text{bound DNA}}$ ) was less than that for the free DNA ( $(I)_{\text{free DNA}}$ ).

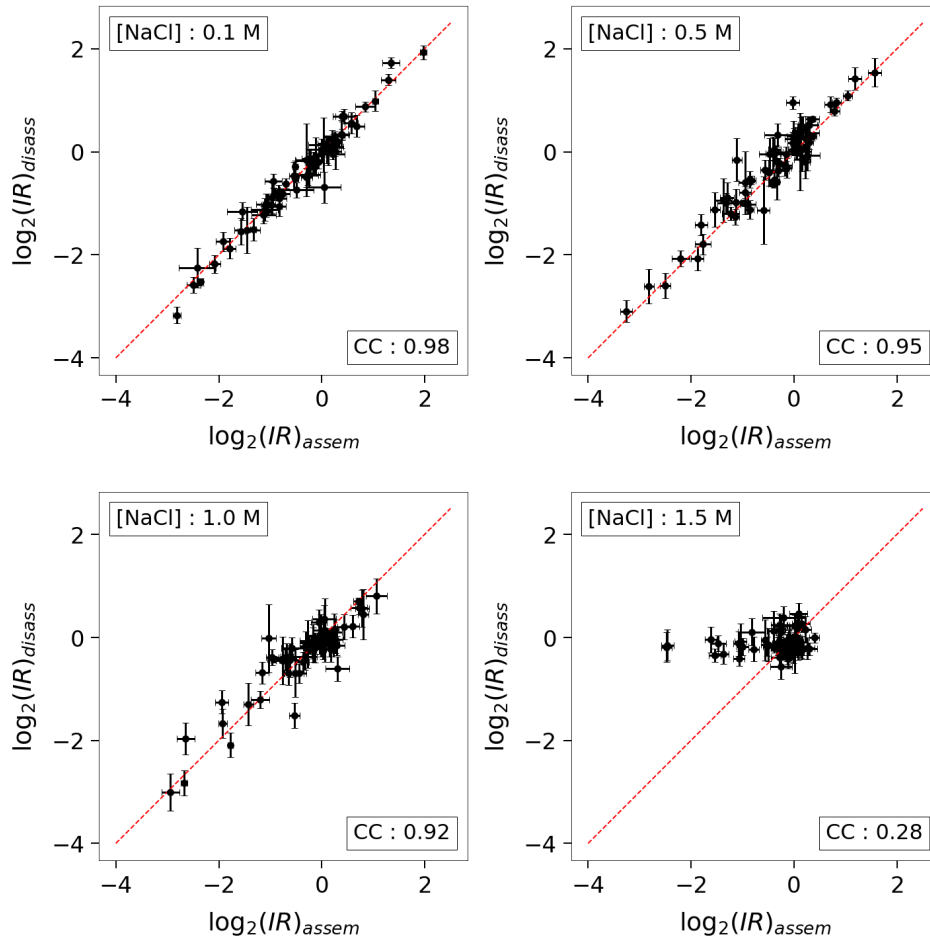
During both assembly and disassembly experiments, the differences in the  $\log_2(IR)$  profiles (Figure 3) advocate for the presence of nucleosome intermediate states that are specific of each of the four salt concentrations used. The comparison between  $\log_2(IR)$  collected at identical ionic strengths during assembly and disassembly experiments is shown in Figure 4.





**Figure 3. Changes in the probability of YpY dimer formation in DNA during nucleosome assembly and disassembly experiments.** PhAST signals are presented in terms of  $\log_2$  of the intensity ratios (IR) along the 601 sequence expressed in SHLs; they are given for decreasing (top panel) or increasing (bottom panel) ionic strengths, as indicated by the green and black arrow respectively. The IR quantities are the ratios calculated between the normalised peak heights of nucleosomal and naked DNA at each YpY position (see the text and Materials and Methods); they represent changes in the probability of YpY dimer formation. Red and blue bars correspond to DNA residues involved in the interface with H3-H4 and H2A-H2B, respectively. The black bars correspond to dinucleotides contacted by both H3-H4 and H2A-H2B. Minor-groove inward facing regions observed in the nucleosome structures are represented by grey boxes; they approximately correspond to the SHL centres. Error bars are standard errors (see Material and Methods for details).

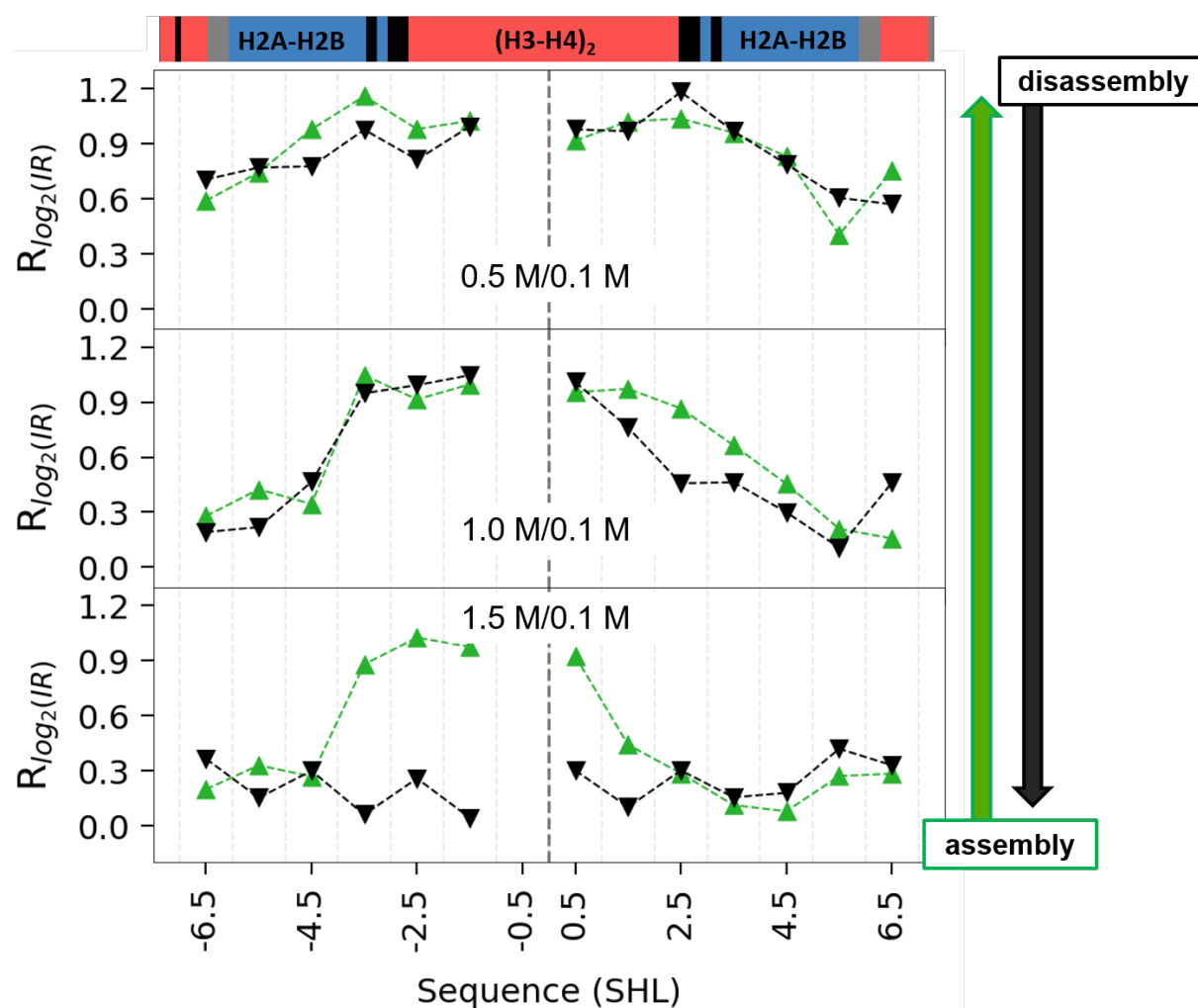




**Figure 4: Comparison of PhAST signal changes from assembly and disassembly experiments.** Comparison of the  $\log_2(IR)$  collected from the assembly ( $\log_2(IR)_{assem}$ ) and disassembly ( $\log_2(IR)_{disass}$ ) experiments at each used ionic strength, i.e. 0.1, 0.5, 1.0 and 1.5 M NaCl. The NaCl concentration and the correlation coefficients (CC) are given in boxes in each panel. The red lines represent  $x=y$ . Error bars are standard errors (see Material and Methods for details).

At 0.1 M NaCl, the nucleosome was completely formed and stable<sup>28</sup>. Most YpY steps distributed along the whole DNA length showed significant values of  $\log_2(IR)$  (Figure 3), almost identical in both assembly and disassembly experiments (Figure 3 and Figure 4, top left). The PhAST signal profile at 0.1 M NaCl represents a signature of the structural effect of the histone octamer on the DNA. As previously shown<sup>28</sup>, a decrease in the probability of formation of a given YpY dimer ( $(I)_{bound\ DNA} < (I)_{free\ DNA}$ ,  $\log_2(IR) < 0$ ) is produced when, from free to bound, the value of roll changes i) from a positive value to a less positive or negative value or ii) from a negative value to a more negative value. The interpretation of an increase in probability of YpY dimer formation arises from the same scheme, substituting negative by positive and positive by negative. So, the sign of  $\log_2(IR)$  finally relates to the structural nature of histone-induced changes.

The similarity between assembly and disassembly experiments at 0.1 M NaCl globally persisted at 0.5 M but began to break down at 1.0 M and finally disappeared at 1.5 M (Figure 4). To further characterize the events occurring during assembly and disassembly experiments at ionic strengths other than 0.1 M, we calculated  $R_{\log_2(IR)}$ , defined as the ratio between  $|\log_2(IR)|$  at a given ionic strength and  $|\log_2(IR)|$  at 0.1 M, the PhAST signal associated with an intact nucleosome being therefore taken as a reference. Note that the  $R_{\log_2(IR)}$  quantity does not take into account the positive or negative sign of PhAST signals that only have a structural significance as mentioned previously. Low  $R_{\log_2(IR)}$  values relate to weak DNA/histone interactions with the extreme, and ideal case of  $R_{\log_2(IR)} = 0$  for a DNA totally free of histones; high  $R_{\log_2(IR)}$  values reflect robust DNA/histone contacts, a DNA fully wrapped around the histone octamer ideally corresponding to  $R_{\log_2(IR)} = 1$ .  $R_{\log_2(IR)}$  values were calculated by summing the values obtained on 10-bp fragments centred on SHLs  $\pm 0.5$ ,  $\pm 1.5$ , etc... according to the DNA/histone interaction pattern (Figure 2). The result of this analysis is summarized in Figure 5 and illustrated by cartoons of molecular models in Figures S2-1 and S2-2.



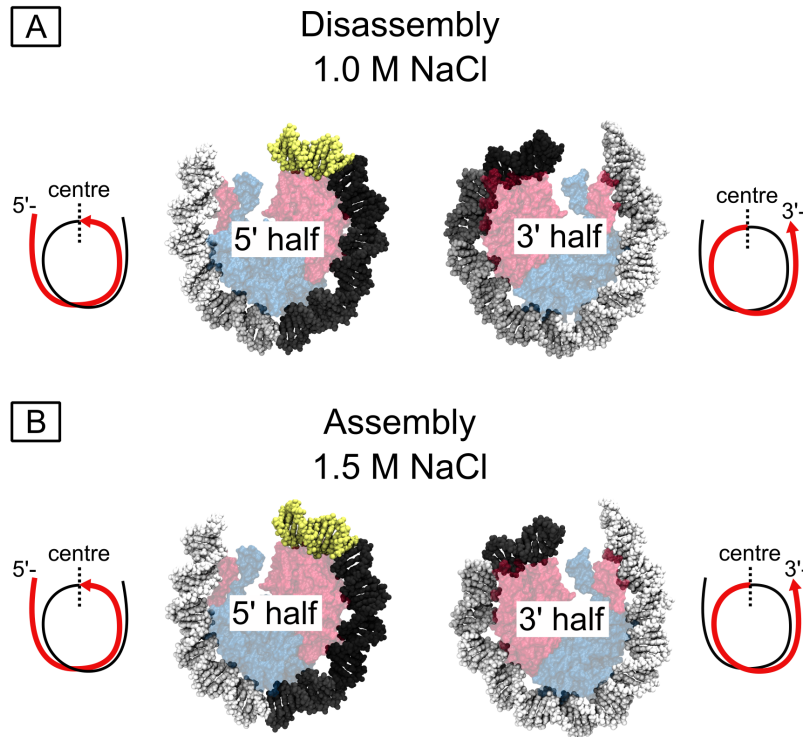
**Figure 5: Changes in PhAST signals by SHLs, in assembly and disassembly experiments. The changes**

of PhAST signals between 0.1 M and the other salt concentration are expressed by the ratios ( $R_{\log_2(IR)}$ ) between the absolute values of  $\log_2(IR)$  at a given ionic strength and  $\log_2(IR)$  at 0.1 M, calculated by SHL. High and low  $R_{\log_2(IR)}$  values indicate increasing and decreasing effects of DNA/histone interactions on DNA shape. The data were sorted and averaged by SHL. The assembly (green triangles) and disassembly (black inverted-triangles) data are compared at each ionic strength. The vertical dashed line indicates  $SHL=0$ . The boxes at the top specify those histones that interact: H3 and H4 in red, H2A and H2B in blue, both H3-H4 and H2A-H2B in black and where none are involved in grey.

In both assembly and disassembly experiments, the resemblances between PhAST signals at 0.5 M and 0.1 M (Figure 5) were compatible with the dominant presence of stable DNA/(H3-H4)<sub>2</sub> interactions around the DNA pseudodyad. The tendency of  $R_{\log_2(IR)}$  values to decrease at the DNA extremities at 0.5 M (Figure 5) indicated that H2A-H2B and especially H2A(C)-H2B(D) around SHL +5.5 were more loosely bound to the DNA than (H3-H4)<sub>2</sub>. This instability also concerns the H3 tails that contact the DNA at SHLs  $\pm 6.5$ .

The comparison of PhAST signals at 1 M and 0.1 M revealed a more complex situation. The low  $R_{\log_2(IR)}$  values associated with the two peripheral DNA regions (Figure 5) signalled a severe weakening of the interface between DNA and the two (H2A-H2B) dimers in both assembly and disassembly experiments. In the DNA centre, the two sides around the pseudodyad responded differently to photo-irradiation, generating a clear asymmetry (Figure 5). On the 3' side, PhAST signal changes expanded from the 3' end to near the centre (Figure 5), reflecting alterations in interactions with H3 and H4, enhanced in the disassembly experiments (illustrated by the cartoons in Figure 6-A). On the 5' side, the high  $R_{\log_2(IR)}$  values observed from SHL 0 to SHL -3.5 (Figure 5) implied that the ~35 bp region 5' of the pseudodyad remained tightly bound to H3 and H4 in both association and dissociation experiments (Figure S2). Within this same region, the abrupt drop of  $R_{\log_2(IR)}$  values between SHL -3.5 and SHL -4.5 (Figure 5) also indicated that the integrity of the H3-H4 contacts was not affected by the serious deterioration of DNA/H2A-H2B dimer interactions present at SHLs -3.5 and -2.5, where the four histones meet (Figure 2).

According to PhAST, assembly and disassembly intermediate states at 1.5 M only, shared the critical deterioration, (Figures 4 and 5) or even the absence, of interactions of DNA with the two H2A-H2B dimers (Figure 5). 1.5 M corresponds to the first step of nucleosome assembly. In this case, there is again a strong asymmetry with respect to the DNA centre (Figures 4, 5 and S2) that matches with an octamer (H3-H4)<sub>2</sub> still firmly attached to half of its DNA site (Figure 6-B). In contrast, at the same salt concentration, the last step of the disassembly process is described by  $\log_2(IR)$  values approaching zero (Figure 3) and low  $R_{\log_2(IR)}$  values (Figure 5) all along the DNA. Clearly here, the asymmetry is totally lost. The close resemblance with what happens with the naked DNA is consistent with a massive histone release (Figures 5 and 6-B), in agreement with experiments using high precision FRET approaches<sup>10</sup>.



**Figure 6: Representation of asymmetries in the DNA/histone interactions on nucleosome models according to PhAST experiments.**

In the panels A and B, the DNA double helix was coloured with a gradient of grey, from light to dark grey for low and high  $R_{\log_2(IR)}$  values (see Figure 6). Low and high  $R_{\log_2(IR)}$  values were interpreted as weak and strong DNA/histone interactions. The double helix in yellow corresponds to the DNA region devoid of YpY steps. The nucleosome DNA was split into two parts containing either the 5' or 3' half of the 601 sequence, as specified in the schemas at the left and right of each molecular representation. Histones are represented by semi-transparent models (H3-H4 in red and H2A-H2B in blue) to reflect their location in the intact nucleosome. For clarity, the histone tails are not represented. Comparison of the PhAST derived data  $R_{\log_2(IR)}$  along the 5' (left) and 3' (right) DNA halves **A**: at 1.0 M NaCl in the experiments of nucleosome dissociation and **B**: at 1.5 M NaCl in the experiments of nucleosome association.

## Discussion

The main aim of this study was to provide a precise analysis of nucleosome formation and dissociation pathways under the same conditions. We used the PhAST approach that is non-invasive and does not require the use of chemical modifications. The unambiguous characterisation of nucleosome intermediate states in terms of DNA base pair and histone contributions is one of the major advantages of PhAST over existing techniques used to scrutinise nucleosome behaviour.

The PhAST results obtained during association and dissociation experiments allowed the identification of multiple intermediate states that formed sequentially during both processes. We infer that binding takes place from the DNA centre and progresses towards the extremities whereas dissociation develops in the direction from the extremities to the centre. These directionally inverse progressions of histone interactions merit a more detailed discussion, in particular regarding the net DNA sequence effects detected by PhAST.

At 0.5 M NaCl, the PhAST signals in the two DNA extremities matched with DNA breathing, perhaps enhanced at the 3' extremity as previously suggested<sup>21, 11</sup>. When the ionic strength exceeded 0.5 M, the DNA/H2A-H2B interfaces were increasingly weakened so that the sequence effect became undetectable. As previously explained<sup>28, 33</sup>, hydrophobic interactions play a non-negligible role at high ionic strength, and the DNA/H2A-H2B interfaces are disadvantaged because of their very poor hydrophobic character. Concerning specifically disassembly, the DNA/histone interface analysis established that the DNA/H2A-H2B interface was less robust than the DNA/H3-H4 interface, an additional factor to favour the release of H2A-H2B at 1 and 1.5 M NaCl.

The real novelty of our work involved the 70-bp region around the pseudodyad, more precisely those segments covering SHLs -3.5 → 0.5 (5' half) and SHLs 0.5 → 3.5 (3' half). PhAST revealed their asymmetric behaviour at 1.5 M NaCl for assembly or at 1 M for disassembly (Figure 6): in both cases the 5' side was clearly more favourable to H3-H4 anchoring than the 3' side.

Concerning assembly, previous results from our TRX annotation provided a quantitative characterisation of the intrinsic structural variability in B-DNA, at the dinucleotide level<sup>23</sup>. In terms of what is predicted from the TRX analysis on free DNA, the 601 sequence would consist of asymmetric halves<sup>23 24</sup>. From the 5' extremity to SHL 2, the alternation of stiff and flexible 5-bp segments was found to perfectly coincide with the periodic, sinusoidal variations of the structural descriptors of the DNA wrapped in a nucleosome. This correspondence was strongly reduced or even disappeared in regions from SHL 2.5 to the 3' extremities, as in other sequences of lower affinity for the histones. It was concluded from these studies that the ability of the 601 sequence to form nucleosome originates from the intrinsic structural and dynamic properties of the whole 5' half extended to the segment just 3' of the dyad (SHLs -7 to 2.5), which limit the cost of DNA wrapping. This idea also provided a convincing explanation for the asymmetry that we observed here during assembly<sup>28</sup>.

This 5' vs 3' side asymmetry involving the central 70 bp is also present during disassembly. A first hypothesis is that this asymmetry signals variations of the DNA-histone interface strength along the DNA in the intact, complete nucleosome, as occurs with H2A-H2B. However, this is clearly not the case since analysis of a stable nucleosome in solution showed that the concerned segments, symmetrically located with respect to the pseudodyad, interact equally with (H3-H4)<sub>2</sub> (Figure 2)<sup>33</sup>. We suggest that the disassembly pathway exploits the most stressed regions in a manner analogous to the relaxation of a stiff spring upon release of constraints. Indeed, our work indicates that those DNA regions that are particularly refractive to nucleosome formation are also the first to break their interactions with the histones.

In summary, the PhAST approach provides a fine picture of the propensities of DNA and histones to interact that, in turn, allows the decryption of the main nucleosome intermediate states during assembly and disassembly. These nucleosome intermediate states are more complex than the relatively simple ensembles of octasomes, hexasomes or tetrasomes described hitherto in the literature. Although under the conditions used here, assembly and disassembly pathways do not correspond to strictly invertible schemes both are extremely sensitive to the DNA sequence. While it is of course difficult to extrapolate these results to the situation *in vivo*, it seems reasonable to expect that such DNA sequence effects modulate chromatin remodelling in conjunction with numerous trans factors such as ATP-dependent proteins that regulate binding, release and sliding of nucleosomes.

**Acknowledgements.** Andrew Travers is warmly thanked for having initiated and guided this study and whose ideas are a continuous thread throughout. The authors also thank Christophe Oguey for helpful advice. This manuscript is dedicated to the memory of Jorge Langowski whose animated and instructive discussions with MB strongly challenged and influenced the ideas expressed in this manuscript.

**Authors contributions.**

C.N. and M.B. conceptualized the study. A.H. performed PhAST experiments. R.R. and M.P. performed the analyses. B.H. and M.B. collected the bibliography and wrote the paper. All authors discussed the results and commented on the paper.

## Methods

### Nucleosome reconstitution

DNA fragments containing 601 sequence were prepared as described previously<sup>28</sup>. Nucleosomes were reconstituted with a salt dilution method according to manufacturer's instructions (New England BioLabs) with a slight modification. Human recombinant histone H2A/H2B dimer (1.5 µg, 54 pmol) and histone (H3/H4)<sub>2</sub> tetramer (1.5 µg, 27 pmol) (New England BioLabs) were mixed with the linearised 601 DNA fragments (6 µg, 16 pmol) in 11 µl of 2M salt buffer (18 mM Tris-HCl, 2 M NaCl, 0.9 mM DTT, 0.9 mM EDTA). The mixture was incubated at room temperature (RT) for 30 min before the salt concentration was lowered to 0.1 M by adding dilution buffer (10 mM Tris-Cl, pH 7.5, 1 mM EDTA, 0.05% NP-40, 5 mM 2-mercaptoethanol, 0.1 mM PMSF) five times every 20 min (from 1.5 M to 1.0 M, 0.5 M, 0.25M, and 0.1M NaCl).

### Dissociation process

Preformed nucleosome complexes were dissociated using two different regimes of increasing salt concentration; either nucleosomes or free DNA were transferred stepwise from 0.1 M NaCl to 0.5 M, 1.0 M and 1.5M. After 15 minutes incubation at each of the salt concentrations, samples were photo-irradiated with the laser as described below. Alternatively, nucleosomes or free DNA at 0.1 M NaCl were transferred directly by one step to either 0.5 M, 1.0 M or 1.5 M NaCl and again irradiated with the laser.

#### One step salt increase:

Twenty microlitre (20 µl) of preformed nucleosome complexes prepared as described above (30 ng/µl DNA) was mixed with 20 µl of high salt buffer (10 mM Tris-Cl, pH 7.5, 1 mM EDTA, 0.05% NP-40, 5 mM 2-mercaptoethanol, 0.1 mM PMSF at one of the following NaCl concentrations; 1M, 2M, and 3M). The final NaCl concentration was 0.5 M, 1 M, or 1.5 M. Final sample volume was 40 µl (DNA concentration was 15 ng/µl) for all conditions. The mixture was incubated for 20 min at RT before irradiation with the laser.

#### Stepwise salt increase:

One hundred microlitres (100 µl) of preformed nucleosome complexes in 0.1 M NaCl buffer prepared as described above (30 ng/µl DNA) was mixed with 11 µl of 4M salt buffer (10 mM Tris-Cl, pH 7.5, 4M NaCl, 1 mM EDTA, 0.05% NP-40, 5 mM 2-mercaptoethanol, 0.1 mM PMSF) to increase the salt concentration to 0.5 M and the mixture was incubated for 20 min. 4M NaCl salt buffer was added in two more steps to increase the NaCl concentration from 0.5M to 1.5M. The mixture was incubated for 20 min after every addition. At each salt concentration, 30 µl of the mixture was taken for Phast analysis



and the DNA concentration adjusted to 15 ng/μl by adding buffer with NaCl before irradiated with the laser.

## Phast analysis

Samples of DNA alone or of reconstituted nucleosomes were irradiated with the UV Laser, followed by primer extension of fluorescent-end-labelled oligonucleotides and separation of the ensuing fragments using capillary electrophoresis (CE, for details on the experimental procedure see <sup>28</sup>). The resulting electrophoretograms (see Figure 1 of our previous publication <sup>28</sup> for an example) were analysed to determine the size (in base pairs) and relative abundance of the fragments present in each sample, using the following procedure. First the electrophoretogram was calibrated by converting migration times to fragment sizes (in units of bases) through piecewise linear fits to the internal size standard (600LIZ) which was run together with each sample. The initial part of each electrophoretogram (up to sizes corresponding to 20 nucleotides) consistently displayed extremely noisy behaviour and was systematically discarded from further analyses. Note that since the primers used for primer extension were larger than 20 b (see above), the retained portion of electrophoretograms also contained the unelongated primers. To facilitate comparison among independent CE runs, electrophoretograms were then normalised using their integral. Peaks were identified based on the analysis of the numerical first derivative of the electrophoretograms, and their maximum height was taken as an estimation of the relative abundance of each fragment. Knowing that a fragment of length  $x$  indicated the formation of a pyrimidine dimer between bases  $x + 1$  and  $x + 2$ , to obtain the relative propensity of forming a pyrimidine dimer at each position along our DNA, we needed to assign to each peak an integer fragment size in units of bases. The starting point for this assignment was the calibrated migration time corresponding to the peak maximum, which was, by the very nature of the calibration process, a fractional quantity in units of bases. Instead of simply taking the closest integer size by rounding, which often leads to artefacts such as assigning two peaks to the same fragment size, we developed an optimization procedure that minimised artefacts by allowing small corrections (on average of about 0.24 bases in either direction), rigorously without changing the order of peaks. In particular, we took advantage of the fact that the size of a fragment implies its sequence, and made the reasonable assumption that the most likely cause for the polymerase to stop is the presence of a photo-induced pyrimidine dimer. Our procedure maximizes, within the aforementioned constraints, the likelihood of the fragment length assignment given its sequence. The resulting sets of peak intensity as a function of DNA sequence for  $n$  independent replicates at the same salt concentration were averaged ( $n = 3$  for the association data,  $n = 6$  for the dissociation data), and the standard error was calculated.

To quantify changes in the likelihood of pyrimidine dimer formation between free and bound DNA at a given salt concentration, we calculated the ratio between the peak intensity in bound DNA over free DNA for each pyrimidine dinucleotide on both strands. These ratios indicate how much more

or less likely it is for a given pyrimidine dimer to form in the presence of histones. As described previously<sup>28</sup>, the comparison is best presented using the  $\log_2$  of these intensity ratios ( $\log_2 IR$ , see e.g. Figure 3). The standard error  $\sigma$  for the  $\log_2 IR$  values was estimated using the following formula:

$$(\ln 2)^2 \sigma^2 = \sigma_B^2 \left[ \frac{1}{\mu_B} \right]^2 + \sigma_F^2 \left[ \frac{1}{\mu_F} \right]^2$$

which propagates the standard errors calculated for each condition ( $\sigma_B$  and  $\sigma_F$  for bound and free DNA, respectively) using a first order expansion around their means ( $\mu_B$  and  $\mu_F$ ).

At a very few positions, PhAST signals corresponded to weak reactivities that were so variable across the experiments that  $\log_2(IR)$  values were lower than the associated standard error; these suspicious points were discarded from the analysis illustrated in Figures 5 and 6.

## Nucleosome models, simulations and DNA/histone interface analysis

The construction of the nucleosome models, the set-up of simulations and the DNA/histone interface analysis were detailed in an original article recently published<sup>33</sup>. We present a brief summary to place the analyses presented here in their initial context.

The four studied nucleosome models contain *Xenopus laevis* histones and the 601 sequence. All models included the histone tails that were partially truncated to only retain those regions not digested by trypsin and clostripain<sup>39, 40</sup>. These nucleosomes were simulated by molecular dynamics in explicit solvent (water molecules and 150 mM NaCl) using the CHARMM36 force field<sup>41</sup> for a total duration of 1  $\mu$ s.

The interface between DNA and histones was analysed by VLDM (Voronoi Laguerre Delaunay for Macromolecules), a software used in particular to depict the DNA/protein interfaces without resorting to any empirical or adjusted parameters<sup>33, 42</sup>. The interface between two structural elements is a polygonal surface, quantified by its area and by its occurrence. In this paper, the contacts occurring less than 20% of the simulation time were not considered.

Note that the DNA sequence is expressed in terms of Super Helical Location (SHL) that is, the number of helical turns separating a given base pair from the central base pair, SHL0; we assume that, on average, one turn corresponds to 10 bp.

- 1 Park, Y. J., Dyer, P. N., Tremethick, D. J. & Luger, K. A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *J Biol Chem* **279**, 24274-24282, doi:10.1074/jbc.M313152200 (2004).
- 2 Hoch, D., A; J,J; Stratton, and Gloss, L.M. Protein-Protein Förster Resonance Energy Transfer Analysis of Nucleosome Core Particles Containing H2A and H2A.Z. *JMB* **371**, 971-988 (2007).
- 3 Kelbauskas, L. C., N; Bash, R; Yidh, J; Woodbury, N. Lohr,D. Sequence-Dependent Nucleosome Structure and Stability Variations Detected by Förster Resonance Energy Transfer. *Biochemistry* **46**, 2239-2248 (2007).

- 4 Gansen, A. *et al.* Nucleosome disassembly intermediates characterized by single-molecule FRET. *Proc Natl Acad Sci U S A* **106**, 15308-15313, doi:10.1073/pnas.0903005106 (2009).
- 5 Bohm, V. *et al.* Nucleosome accessibility governed by the dimer/tetramer interface. *Nucleic Acids Res* **39**, 3093-3102, doi:10.1093/nar/gkq1279 (2011).
- 6 Toth, K. B., V; Sellmann, C; Danner, M; Hanne, J; Berg, M; & Barz, I. G., A; Langowski, J. Histone- and DNA Sequence-Dependent Stability of Nucleosomes Studied by Single-Pair FRET. *Cytometry Part A* **83A**, 839-846, doi:10.1002/cyto.22320 (2013).
- 7 Chen, Y. *et al.* Revealing transient structures of nucleosomes as DNA unwinds. *Nucleic Acids Res* **42**, 8767-8776, doi:10.1093/nar/gku562 (2014).
- 8 Chen, Y. *et al.* Asymmetric unwrapping of nucleosomal DNA propagates asymmetric opening and dissociation of the histone core. *Proc Natl Acad Sci U S A* **114**, 334-339, doi:10.1073/pnas.1611118114 (2017).
- 9 Lee, J. & Lee, T. H. Single-Molecule Investigations on Histone H2A-H2B Dynamics in the Nucleosome. *Biochemistry* **56**, 977-985, doi:10.1021/acs.biochem.6b01252 (2017).
- 10 Gansen, A. *et al.* High precision FRET studies reveal reversible transitions in nucleosomes between microseconds and minutes. *Nat Commun* **9**, 4628, doi:10.1038/s41467-018-06758-1 (2018).
- 11 Hall, M. A. *et al.* High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat Struct Mol Biol* **16**, 124-129, doi:nsmb.1526 [pii] 10.1038/nsmb.1526 (2009).
- 12 Thuy T.M. Ngo, Q. Z., Ruobo Zhou, Jaya G. Yodh, and Taekjip Ha. Asymmetric Unwrapping of Nucleosomes under Tension Directed by DNA Local Flexibility. *Cell* **160**, 1135–1144, doi:10.1016/j.cell.2015.02.001 (2015).
- 13 Sheinin, M. Y., Li, M., Soltani, M., Luger, K. & Wang, M. D. Torque modulates nucleosome stability and facilitates H2A/H2B dimer loss. *Nat Commun* **4**, 2579, doi:10.1038/ncomms3579 (2013).
- 14 Tomschik, M., Zheng, H., van Holde, K., Zlatanova, J. & Leuba, S. H. Fast, long-range, reversible conformational fluctuations in nucleosomes revealed by single-pair fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A* **102**, 3278-3283, doi:10.1073/pnas.0500189102 (2005).
- 15 Koopmans, W. J., Buning, R., Schmidt, T. & van Noort, J. spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes. *Biophys J* **97**, 195-204, doi:10.1016/j.bpj.2009.04.030 (2009).
- 16 Miyagi, A., Ando, T. & Lyubchenko, Y. L. Dynamics of nucleosomes assessed with time-lapse high-speed atomic force microscopy. *Biochemistry* **50**, 7901-7908, doi:10.1021/bi200946z (2011).
- 17 Lyubchenko, Y. L. Nanoscale Nucleosome Dynamics Assessed with Time-lapse AFM. *Biophys Rev* **6**, 181-190, doi:10.1007/s12551-013-0121-3 (2014).
- 18 Bilokapic, S., Strauss, M. & Halic, M. Histone octamer rearranges to adapt to DNA unwrapping. *Nat Struct Mol Biol* **25**, 101-108, doi:10.1038/s41594-017-0005-5 (2018).
- 19 Arimura, Y., Tachiwana, H., Oda, T., Sato, M. & Kurumizaka, H. Structural analysis of the hexasome, lacking one histone H2A/H2B dimer from the conventional nucleosome. *Biochemistry* **51**, 3302-3309, doi:10.1021/bi300129b (2012).
- 20 Kato, D. O., A. Arimura, Y. Mizukami, Y. , Horikoshi, N. S., K. Akashi, S. Nishimura, Y. , Park, S.-Y. N., J. Maehara, K. Ohkawa, Y. & Matsumoto, A. K., H. Inoue, R. Sugiyama, H, Kurumizaka, H. Crystal structure of the overlapping dinucleosome composed of hexasome and octasome. *Science* **356**, 205-208 (2017).
- 21 Chua, E. Y., Vasudevan, D., Davey, G. E., Wu, B. & Davey, C. A. The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Res* **40**, 6338-6352, doi:10.1093/nar/gks261 (2012).
- 22 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

- 23 Heddi, B., Oguey, C., Lavelle, C., Foloppe, N. & Hartmann, B. Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res* **38**, 1034-1047, doi:gkp962 [pii] 10.1093/nar/gkp962 (2010).
- 24 Xu, X. *et al.* NMR studies of DNA support the role of pre-existing minor groove variations in nucleosome indirect readout. *Biochemistry* **53**, 5601-5612, doi:10.1021/bi500504y (2014).
- 25 Sullivan, S. A. a. L., D. Characterization of Sequence Variability in Nucleosome Core Histone Folds. *PROTEINS* **52**, 454-465 (2003).
- 26 Worcel, A., Han, S. & Wong, M. L. Assembly of newly replicated chromatin. *Cell* **15**, 969-977 (1978).
- 27 Cremisi, C. & Yaniv, M. Sequential assembly of newly synthesized histones on replicating SV40 DNA. *Biochem Biophys Res Commun* **92**, 1117-1123 (1980).
- 28 Hatakeyama, A., Hartmann, B., Travers, A., Nogues, C. & Buckle, M. High-resolution biophysical analysis of the dynamics of nucleosome formation. *Scientific Reports* **6**, 27337, doi:10.1038/srep27337 (2016).
- 29 Buckle, M., Geiselmann, J., Kolb, A. & Buc, H. Protein-DNA cross-linking at the lac promoter. *Nucleic Acids Research* **19**, 833-840 (1991).
- 30 Brach, K. *et al.* Photochemical analysis of structural transitions in DNA liquid crystals reveals differences in spatial structure of DNA molecules organized in liquid crystalline form. *Sci Rep* **8**, 4528, doi:10.1038/s41598-018-22863-z (2018).
- 31 Imeddourene, A. B. *et al.* The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res*, doi:10.1093/nar/gkw084 (2016).
- 32 Andrews, A. J. & Luger, K. Nucleosome Structure(s) and Stability: Variations on a Theme. *Annu Rev Biophys* **40**, 99-117, doi:10.1146/annurev-biophys-042910-155329 (2011).
- 33 Elbahnsi, A., Rétureau, R., Baaden, M., Hartmann, B. & Oguey, C. Holding the Nucleosome Together: A Quantitative Description of the DNA-Histone Interface in Solution. *J Chem Theory Comput* **14**, 1045-1058, doi:10.1021/acs.jctc.7b00936 (2018).
- 34 Schildkraut, C. & Lifson, S. Dependence of Melting Temperature of DNA on Salt Concentration. *Biopolymers* **3**, 195-+, doi:DOI 10.1002/bip.360030207 (1965).
- 35 Borochoy, N., Eisenberg, H. & Kam, Z. Dependence of DNA Conformation on the Concentration of Salt. *Biopolymers* **20**, 231-235, doi:DOI 10.1002/bip.1981.360200116 (1981).
- 36 Baumann, C. G. S., S B. Bloomfeld, V A. Bustamante, C. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. USA* **94**, 6185-6190 (1997).
- 37 Pan, S., At Nguyen, D., Sridhar, T., Sunthar, P. & Ravi Prakash, J. Universal solvent quality crossover of the zero shear rate viscosity of semidilute DNA solutions. *Journal of Rheology* **58**, 339-368, doi:10.1122/1.4861072 (2014).
- 38 Kriegel, F. *et al.* Probing the salt dependence of the torsional stiffness of DNA by multiplexed magnetic torque tweezers. *Nucleic Acids Res* **45**, 5920-5929, doi:10.1093/nar/gkx280 (2017).
- 39 Dumuiskervabon, A. *et al.* A Chromatin Core Particle Obtained by Selective Cleavage of Histones by Clostripain. *Embo Journal* **5**, 1735-1742, doi:DOI 10.1002/j.1460-2075.1986.tb04418.x (1986).
- 40 Morales, V. & Richard-Foy, H. Role of histone N-terminal tails and their acetylation in nucleosome dynamics. *Molecular and Cellular Biology* **20**, 7230-7237, doi:Doi 10.1128/Mcb.20.19.7230-7237.2000 (2000).
- 41 Hart, K. *et al.* Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. *Journal of Chemical Theory and Computation* **8**, 348-362, doi:10.1021/ct200723y (2012).
- 42 Rétureau, R., Oguey, C., Mauffret, O. & Hartmann, B. Structural Explorations of NCp7-Nucleic Acid Complexes Give Keys to Decipher the Binding Process. *J Mol Biol* **431**, 1966-1980, doi:10.1016/j.jmb.2019.03.002 (2019).

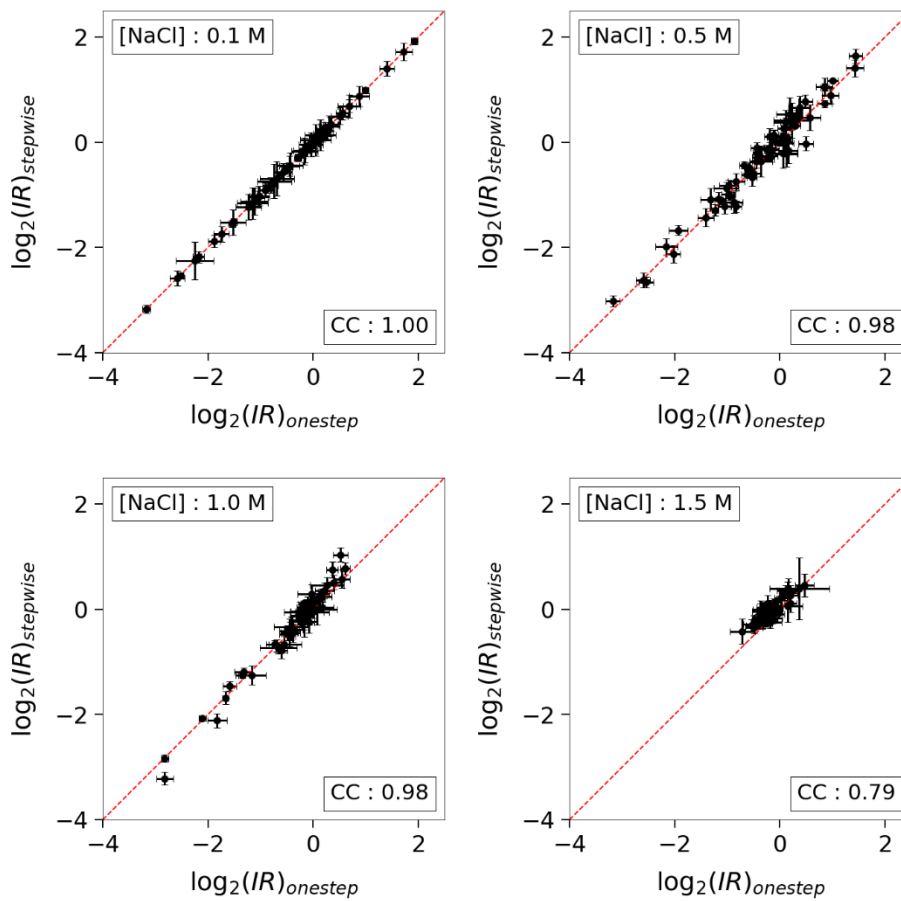
## **Supplementary Data**

Nucleosome assembly and disassembly pathways.

Akiko Hatakeyama, Romain Retureau, Marco Pasi, Brigitte Hartmann, Claude Nogues and Malcolm Buckle.

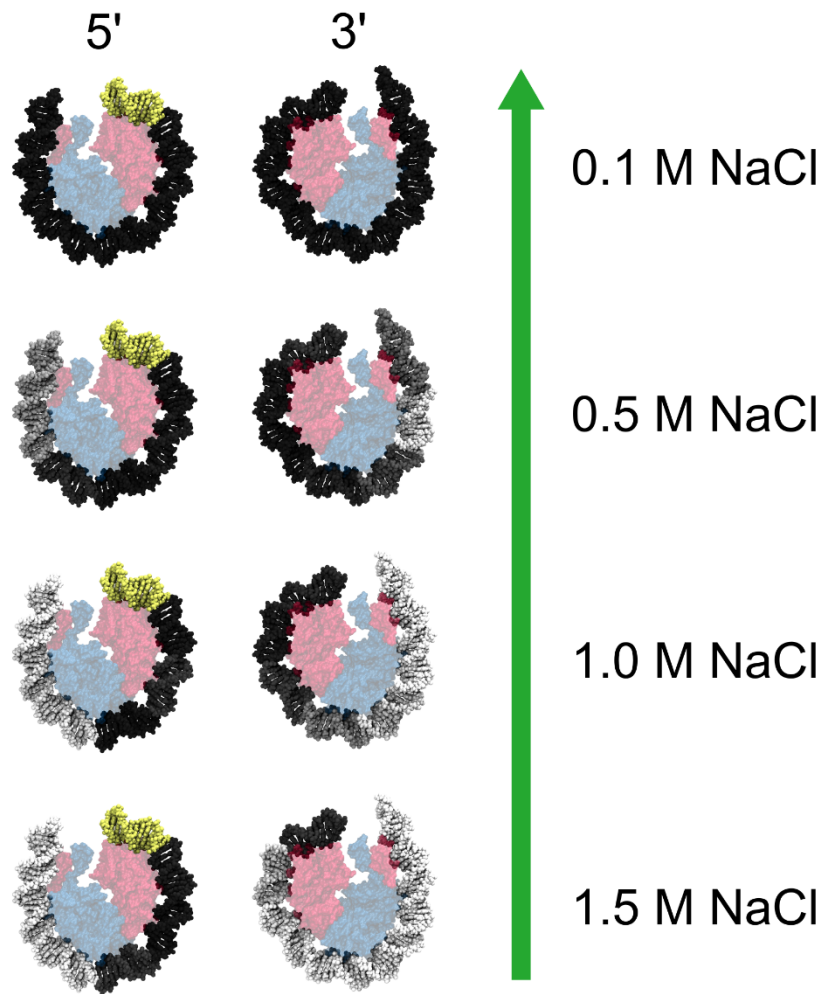
Figure S1: Comparison of PhAST signals from two types of disassembly experiments.

Figures S2-1 and S2-2: Representation of the DNA-histone interactions on nucleosome models according to PhAST experiments of nucleosome assembly (S1-1) and disassembly (S1-2).



**Figure S1: Comparison of PhAST signals from two types of disassembly experiments.**

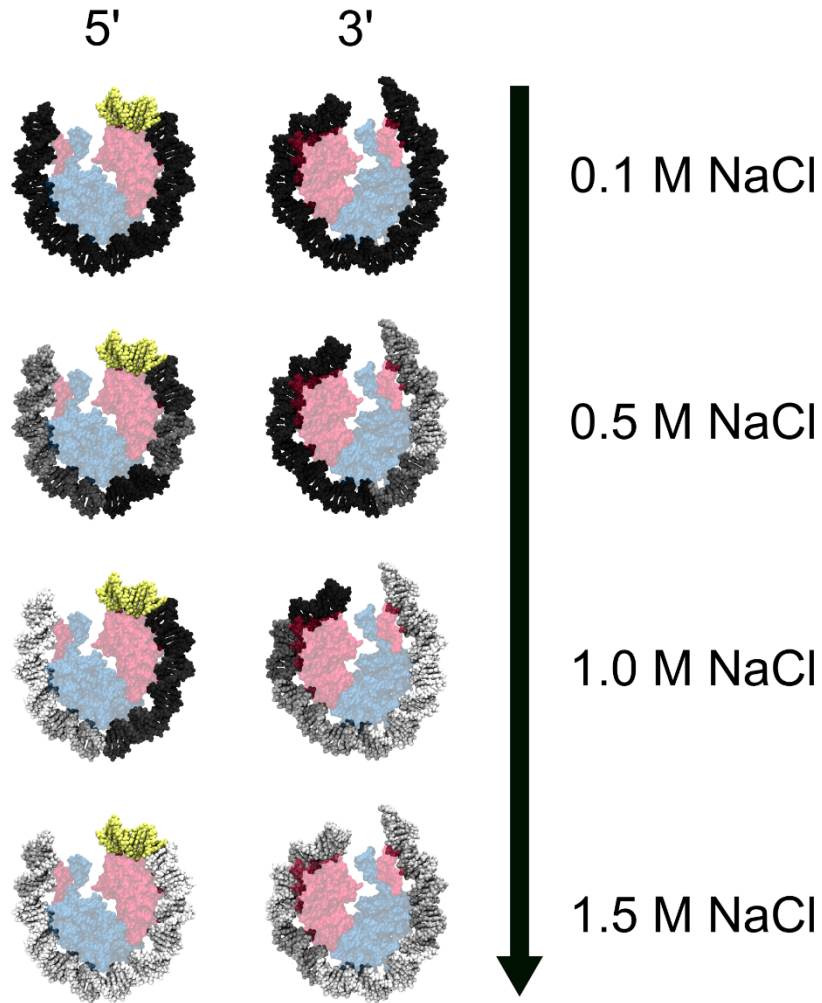
Comparison of the log<sub>2</sub>(IR) collected from the disassembly experiments carried out either by gradually decreasing the ionic strength (stepwise salt increase) or by transferring the preformed nucleosomes directly by a single step to different NaCl concentrations (one step salt increase). The NaCl concentration and the correlation coefficients (CC) are given in boxes in each panel. The red lines represent x=y. Error bars are standard errors (see Material and Methods for details).



**Figure S2-1: Representation of the DNA-histone interactions on nucleosome models according to PhAST experiments of nucleosome assembly.**

The DNA double helices were coloured with a gradient of grey, from light to dark grey for low and high  $R_{\log_2(IR)}$  values (see Figure 5). The data were extracted from PhAST experiments of nucleosome assembly carried out from 1.5 to 0.1 M NaCl. Low and high  $R_{\log_2(IR)}$  values were interpreted as weak and strong DNA-histone interactions. The double helix in yellow corresponds to the DNA region devoid of YpY steps. The nucleosomal DNA was split into two parts containing either the 5' or 3' half of 601 sequence. Histones are represented by semi-transparent models (H3-H4 in red and H2A-H2B in blue) to recall their location in the intact nucleosome.





**Figure S2-2: Representation of the DNA-histone interactions on nucleosome models according to PhAST experiments of nucleosome disassembly.**

The DNA double helices were coloured with a gradient of grey, from light to dark grey for low and high  $R_{\log_2(IR)}$  values (see Figure 5). The data were extracted from PhAST experiments of nucleosome assembly carried out from 0.1 to 1.5 M NaCl. Low and high  $R_{\log_2(IR)}$  values were interpreted as weak and strong DNA-histone interactions. The double helix in yellow corresponds to the DNA region devoid of YpY steps. The nucleosomal DNA was split in two parts containing either the 5' or 3' half of 601 sequence. Histones are represented by semi-transparent models (H3-H4 in red and H2A-H2B in blue) to recall their location in the intact nucleosome.

## Conclusion et perspectives

Dans cette étude, nous avons pu décrypter et comparer directement les étapes clés de l'assemblage et du désassemblage du nucléosome. Les voies de l'assemblage et du désassemblage ne sont pas strictement identiques, mais dans les deux cas, elles sont sensibles à la séquence d'ADN. Il semble donc que la capacité de l'ADN à absorber le stress produit par l'enroulement autour des histones ait surtout un impact lors des étapes transitoires menant vers l'assemblage ou le désassemblage mais, d'après nos simulations, qu'il impacte moins le nucléosome complètement formé.

Cette étude ouvre de nouvelles questions sur le nucléosome et la méthode PhAST offre la possibilité d'y répondre. D'une part, nous observons une légère différence dans les étapes d'assemblage et de désassemblage. Afin de mieux cerner ce phénomène, par exemple de voir si les états transitoires apparaissent progressivement, de manière continue ou bien s'il y a des décrochages à une force ionique donnée, il serait intéressant de faire varier plus finement les concentrations en NaCl (par pas de 0.2M au lieu de 0.5M).

Une autre perspective permise par la méthode PhAST est de pouvoir observer l'assemblage et le désassemblage du nucléosome avec d'autres séquences d'ADN. Notamment, il existe des versions palindromiques de la séquence 601 (p. 36) dont la moitié 5' ou la moitié 3' est dupliquée. Mais on pourrait également proposer une symétrisation de la séquence 601 en termes de propriétés intrinsèques des dinucléotides selon l'échelle TRX (Rôle de la séquence ADN dans la formation du nucléosome, p. 34). Ces variants d'ADN permettraient d'affiner et de rationaliser complètement l'effet de séquence sur l'assemblage et le désassemblage, question encore ouverte aujourd'hui.

# Chapitre 3 : Dynamique de l'ADN dans le nucléosome

## Sommaire <sup>3</sup>

Il est aujourd'hui communément admis que les propriétés intrinsèques de l'ADN libre jouent un rôle dans sa complexation avec les protéines. La flexibilité de l'ADN au niveau des dinucléotides (paramètres hélicoïdaux comme le *roll*, ou le *twist*) et du tour d'hélice (dimension des sillons, courbure) influe directement sur l'affinité avec les protéines. Des analyses de données RMN et de structures provenant de dynamiques moléculaires ont permis d'établir sur des ADN libres les couplages entre l'équilibre BI↔BII du squelette phosphodiester, les paramètres hélicoïdaux locaux de l'ADN et les dimensions du sillon. Ces couplages ont permis de quantifier la flexibilité locale de l'ADN, qui est dictée par la séquence dinucléotidique. Dans cette étude, nous analysons par dynamique moléculaire l'ADN du nucléosome afin de comprendre l'impact de l'interface ADN/histones très dense sur ses propriétés intrinsèques.

Nous avons donc analysé le comportement de l'ADN nucléosomal en utilisant les mêmes 1.2μs de simulations du nucléosome que pour le premier article (p. 66). Nous avons comparé les résultats avec ceux obtenus à partir de quatre dynamiques de dodécamères d'ADN libre de 1μs chacune. Les séquences de ces ADN libres correspondent à 39 paires de base de la séquence 601 (Article 3 – Figure S1).

Dans un premier temps, nous montrons que les fluctuations atomiques au niveau de la paire de base sont peu variables le long de la séquence (Article 3 – Figure 1C) parce que quasiment tout l'ADN est contacté sur un brin ou l'autre à l'exception d'une ou deux paires de base tous les 10 nucléotides.

Malgré ces contacts et la forte courbure induite par la complexation avec les histones, l'équilibre BI↔BII observé dans l'ADN libre persiste dans l'ADN nucléosomal (Article 3 – Figure 3). Cependant, la distribution des conformères le long de la séquence est périodique dans l'ADN nucléosomal avec des pics de fort pourcentage de BII (Article 3 – Figure 3B) correspondant aux zones où le petit sillon de l'ADN fait face aux histones.

L'analyse des paramètres inter-paire de base de l'ADN nucléosomal permet de mettre en évidence les corrélations impliquant le *roll*, le *slide*, et le *twist* et les combinaisons de conformation de

---

<sup>3</sup> L'ensembles des références de cette section peuvent-être trouvées dans l'article associé au chapitre (p. 136) et dans l'introduction de ce manuscrit (Le nucléosome, p. 17)

phosphate qui se font face sur le brin I et le brin J (Article 3 – Figure 7), observées dans l'ADN libre, mais plus marquées dans le nucléosome.

Concernant l'effet de séquence, nous montrons que les dinucléotides les plus réfractaires à prendre la conformation BII dans l'ADN libre présentent un pourcentage moins élevé de BII dans l'ADN nucléosomal (Article 3 – Figure 7) en comparaisons des autres dinucléotides.

# A dynamic view of DNA structure within the nucleosome: biological implications

Romain Retureau<sup>1</sup>, Nicolas Foloppe<sup>3</sup>, Ahmad Elbahnsi<sup>1,2</sup>, Christophe Oguey<sup>2</sup> and Brigitte Hartmann<sup>1,\*</sup>

<sup>1</sup> LBPA, CNRS, ENS Cachan, Université Paris-Saclay, 61 avenue du Président Wilson, 94235 Cachan cedex, France

<sup>2</sup> LPTM, UMR8089, CNRS, Université de Cergy-Pontoise, 2 avenue Adolphe Chauvin, 95302 Cergy-Pontoise, France

<sup>3</sup> 51 Natal Road, Cambridge CB1 3NY, UK

\* Corresponding author: bhartman@ens-cachan.fr (BH)

## ABSTRACT

Most of eukaryotic cellular DNA is packed in nucleosome core particles (NCPs). DNA in NCPs (DNA<sub>NCP</sub>) is wrapped around histones and the effect of this organization on the DNA structural properties is largely unknown. Here this was addressed using MD simulations of a nucleosome containing the 601 DNA sequence and four related free dodecamers. Despite a dense DNA-histone interface, DNA<sub>NCP</sub> motions occur at the dinucleotide level, comprising BI↔BII backbone transitions. To a large proportion, the variability of helical parameters in DNA<sub>NCP</sub> parallels that observed in free DNA, while being slightly attenuated. BI/BII states contribute to the DNA<sub>NCP</sub> super-helical pathway *via* strong couplings with slide, roll and twist, as initially revealed in free DNA. Thus, the DNA wrapping signature consists of well-defined ~10bp periodicities of BI/BII populations as well as slide and roll values. Importantly, BI- and BII-rich regions in DNA<sub>NCP</sub> mainly mirror the sequence dependence of BII propensities in free DNA. In sum, this study examines how DNA behaves inside a large DNA-protein complex. It also provides a rational structural basis about the DNA sequence preferences for a non-specific protein core, strengthening the idea that the DNA intrinsic properties regulate the formation of nucleoprotein complexes.

Keywords: nucleosome; nucleosomal DNA; DNA structure and dynamics; BI ↔ BII equilibrium; modeling; CHARMM36.

## INTRODUCTION

The nucleosome core particle (NCP) is one of the most notorious DNA-protein complexes, as the fundamental building block of packaged DNA in eukaryotic cells. X-ray structures showed that the DNA in NCP ( $\text{DNA}_{\text{NCP}}$ ) wraps  $\sim 1.7$  times around eight histone proteins (two copies of H2A, H2B, H3, and H4) to form a super-helical double helix, as described in several reviews (1–7).  $\text{DNA}_{\text{NCP}}$  of different sequences and lengths from 145 to 147 bp adapt to constant histone binding motifs, regularly positioned at the surface of the histone structured domains. Along the DNA sequences, the values of several inter base-pair parameters and the groove dimensions follow sinusoidal profiles, with oscillation period close to 10 base-pairs (bp) (8–12). Two inter base-pair parameters, roll and slide, are the main actors of the  $\text{DNA}_{\text{NCP}}$  wrapping (9, 13) but their values along  $\text{DNA}_{\text{NCP}}$  are not exactly reproducible across the NCP X-ray structures (12). The  $\text{DNA}_{\text{NCP}}$  anchoring is also assisted by twist and rise adjustments that generate either expanded/overwound (stretched state) or compressed/underwound (unstretched state) short segments (2, 14–17).

The NCP regulates access to the DNA in the nucleus and therefore plays a central role in a large variety of processes (recent review (4)). Yet, the  $\text{DNA}_{\text{NCP}}$  is not totally masked since interactions with ligands, comprising proteins, remain possible. The DNA readability is tightly related to the *in vivo* nucleosome positioning that results, in addition to intrinsic DNA properties, from the action of a series of trans-acting factors (reviews: (18–20), including chromatin remodelers or methyltransferases that operate on  $\text{DNA}_{\text{NCP}}$  (4, 20–22). Various enzymes or transcription factors also bind  $\text{DNA}_{\text{NCP}}$ . The DNase I enzyme is an interesting example, since it has long been exploited in nucleosome studies (23, 24) and still is (25). Essentially, the catalytic site of DNase I fills the DNA minor groove and cleaves the phosphodiester linkage (26, 27). Applied on NCP, this enzyme targets the wide minor grooves pointing outwards, opposite to the histone octamer, and produces a typical oscillatory cleavage profile ((25) and references herein). This property was cleverly exploited to detect transcription factors that bind to  $\text{DNA}_{\text{NCP}}$  and consequently disrupt the DNase I periodic cleavage pattern (25).

It is now widely accepted that the properties of free DNA (intrinsic DNA properties) act on the assembly of nucleoprotein complexes. One of the factors influencing the affinity between a DNA target and its partner is the DNA sequence dependent flexibility at a local (for instance roll or twist) or semi-local (groove dimensions) level, as illustrated on various systems (28–36). As typical examples, the NF- $\kappa$ B DNA targets need two regions whose dynamics transiently expose the specific base atom pattern recognized by the protein (31, 37); also, the DNase I cleavage efficiency is increased by a malleable minor groove that favors the enzyme anchoring (33).

In free DNA, the DNA constituents undergo coordinated motion. Thus, motions of the phosphodiester linkages are concomitant with variations of the relative positions of two successive bases. After earlier analyses of molecular mechanics models of free DNA (38) or X-ray (39, 40) approaches, NMR studies definitively established the strong relation between the BI  $\leftrightarrow$  BII equilibrium involving the  $\epsilon$  and  $\zeta$  backbone angles, and the inter-base parameters of slide, roll, twist (41–43). To a lesser extent, rise, tilt and shift are also coupled to the backbone states (44). Finally, the groove dimensions are associated to the BI or BII density per 4-5 bp segments (45, 46). In sum, there is much overlap between the notions of intrinsic dynamics and mechanical

couplings.

In addition to the structural couplings described just above, the sensitivity of backbone state populations to the DNA sequence, firmly established by NMR in solution (41, 46), offered the opportunity to quantify the malleability of dinucleotides and complementary dinucleotides. Indeed, the backbone state populations of each dinucleotide reflect the conformational landscape that it is able to explore. According to this approach, in free DNA ApA/G•T/CpT, and ApT/C•ApT/C have only access to a restricted conformational region while GpG•CpC, CpG•CpG, GpC•CpC and CpA•TpG oscillate over a larger range of conformations (41, 46).

Then, an important and unresolved question is whether the intrinsic structural and dynamical properties, revealed in free DNA, still operate in the case of a bound, and therefore constrained, DNA. From this point of view, DNA<sub>NCP</sub> is a paradigm: its curvature, more accentuated than in free DNA, is maintained by a dense interaction network involving both the histone structured domains and tail roots along the two DNA strands, as recently observed on a nucleosome simulated in solution (47). Actually, the behavior of such stressed DNA<sub>NCP</sub> remains rather poorly documented as now presented.

Concerning potential mechanical couplings, several studies which examined various datasets of NCP X-ray structures reported qualitative parallel DNA<sub>NCP</sub> profiles for i) roll and twist (48, 49), ii) roll, twist and phosphate group conformers (41), iii) roll, twist, and slide (8, 50) or iv) roll, twist and groove width (12). These analyses suggest that the free DNA properties are at least in part preserved in DNA<sub>NCP</sub>, as observed from large datasets of DNA bound to proteins (39, 45). To our knowledge, only two studies (51, 52) mentioned the dispersion of DNA<sub>NCP</sub> helical parameter values, which is a direct way to characterize the DNA<sub>NCP</sub> malleability. A comparison between X-ray structures of free DNA oligomers and NCP suggested that the dispersion of roll and tilt values was enhanced in DNA<sub>NCP</sub> (51). This result was apparently retrieved in a short (15ns) molecular dynamics (MD) simulation of the best resolved X-ray structure (PDB code: 1KX5) in explicit solvent (52). Finally, the relative variability of the slide and roll values along DNA<sub>NCP</sub> (12) as well as the twist and rise adjustments in response to DNA length or the presence of ligands (15–17) could argue in favor of a residual DNA<sub>NCP</sub> malleability. Overall, those rather scarce insights were mostly gleaned from static DNA<sub>NCP</sub> X-ray structures, potentially influenced by the biases of intermolecular contacts and crystal packing (2, 53, 54).

Thus, we decided to exploit all atoms MD simulations of a nucleosome formed with the 601 sequence (also called Widom sequence) to gain a more direct view of the behavior of DNA<sub>NCP</sub> in solution. This 601 sequence, discovered from synthetic random sequences, remains of special interest because of its high affinity for the histone octamer (55). The simulations, performed in explicit solvent using the CHARMM36 force field (56), were previously used to better understand how DNA<sub>NCP</sub> could be maintained around the histones (47). The exhaustive characterization of the interface – contact list, contact area and contact occurrence - was carried out with VLDM, a Voronoi tessellation-based method analyzing the topology of interacting elements without any empirical or subjective adjustment (47, 57, 58). That study found a DNA-protein interface more robust than previously known, with dense, long-lived non-electrostatic and electrostatic interactions engaging DNA<sub>NCP</sub> and both structured and unstructured histone domains. Another new insight on the nucleosome cohesion was the notable reinforcement of interfaces by the stiff tail roots (the histone tail regions adjacent to the structured



domains) and the stabilization of the close juxtaposed DNA<sub>NCP</sub> gyres by cation accumulation.

Here, the same simulations of a total duration of 1  $\mu$ s were revisited to characterize additional features of the DNA<sub>NCP</sub> structure and dynamics. Moreover, we exploited further MDs of four free DNA dodecamers that together cover 39 base pairs of the 5' half of sequence 601 (59). These MDs, of 1  $\mu$ s each, initially designed to test the CHARMM36 force field against NMR data collected on these dodecamers asserted the satisfactory DNA representation, in particular regarding the BI and BII population balance and its modulation at the dinucleotide sequence level. Since the MDs of free and NCP-bound DNA were performed with consistent protocols and force-fields, they allow a comparison between free and histone bound DNA.

The present study investigates aspects of DNA<sub>NCP</sub> dynamics including atomic fluctuations, base pairing stability, motions of backbone, variations of inter base pair parameters and couplings involving helical descriptors. Bound and free DNA are systematically compared to better emphasize the specific characteristics of DNA<sub>NCP</sub>. Thanks to the exhaustive dataset collected on the DNA-histone interface, we were able to examine the relation between the DNA<sub>NCP</sub> dynamics and the contacts with the histones. Finally, the role of the different dinucleotide intrinsic properties on the propensity of a DNA sequence to assemble into a nucleosome was considered in the light of the properties of DNA<sub>NCP</sub>.

## MATERIALS AND METHODS

Two previous papers presented in details the setup of our Molecular Dynamics (MD) simulations carried out on the nucleosome (47) and the free DNA dodecamers (59). We therefore only summarize here the main aspects of the protocols.

### Nucleosome models

The nucleosome is constituted by two copies of four histones (H3: chains A and E; H4: chains B and F; H2A: chains C and G; H2B: chains D and H) and a double strand DNA. We built four nucleosome models containing the same DNA sequence and the same histone folded domains but differing by the conformations of the histone N- and C-terminal domains – called N- and C-tails (Figure S1-A). The built systems, named here SYS1, SYS1-bis, SYS2 and SYS2-bis, were all based on the folded domains of *Xenopus laevis* histones, and the 601 sequence of 146 base-pairs (bp) from the nucleosome X-ray structure 3MVD (60); the tails were grafted on the folded domains, after transfer from the X-ray structure 1KX5 (61), the only source of resolved tails. We exploited the fact that the two copies of each type of histone tail in 1KX5 adopt different conformations to generate various starting points by symmetrizing the N-tail structures (Figure S1-A and Table S1). These tails were partially truncated (Table S1) so that the kept regions coincide with the inaccessible parts, as delimited by trypsin and clostripain digestions (62, 63). The residues 119-128 of the H2A C-tail from 1KX5 were integrated in SYS1-bis and SYS2-bis. According to  $pK_a$  calculations (64), six histidines (H3-H39, H4-H75, H2A-H32, H2A-H82, H2B-H46 and H2B-H79) have  $pK_a$  values between 7.7 and 9.6 and were therefore protonated.

## DNA sequences

The 601 sequence of 146 bp (Table S2) was initially selected for its very high-affinity for association with the histone octamer (55). In the nucleosome structures, the DNA center corresponds to the pseudo two-fold axis of symmetry, the dyad. According to conventions suggested for the description of the first X-ray structure of NCP (1), the rotational orientation of the DNA is defined relative to the DNA center (Super Helix Location zero, or SHL0). Assuming that one DNA turn corresponds to  $\sim 10$  bp, the 3' half DNA is numbered from SHL -7 to SHL 0, and the 5' half from SHL 0 to SHL +7.

The sequences of the four free (unbound) dodecamers studied by NMR are reported in Table S2; they overlap by three bases to splice the four oligomers, excluding the terminal base pairs, subject to end effects. For instance, the first (Oligo 1) and the second (Oligo 2) oligomer end and begin with the same motif, GCT. Taking into account the overlaps and excluding the extremities, a total of 39 non-redundant base pairs are covered by the four free dodecamers, corresponding to the segment from SHL -3.8 to SHL -0.2 of DNA<sub>NCP</sub> (Figure S1-B). In the present study, these dodecamers will be called “601-dodecamers”.

## Molecular Dynamics Simulations Set-up

Molecular dynamics simulations were carried out with the CHARMM (65) and NAMD 2.11 programs (66), using the CHARMM36 force field (56) with the CMAP correction (67). This force field was chosen for modeling the nucleosome because it gave reasonable agreement between the simulated free 601-dodecamers and their counterpart experimental NMR data (59), while being also reliable for proteins.

The macromolecule solute, nucleosome or dodecamer, was immersed in a box filled with TIP3P water molecules (Jorgensen et al., 1983) and at least 10 Å of separation between the solute and the edges of the box. Electric neutrality was achieved by adding either Na<sup>+</sup> (62) and Cl<sup>-</sup> (68) ions, or only Na<sup>+</sup> ions (minimal salt condition), for the simulations of nucleosome and 601-dodecamers, respectively. After equilibration steps detailed in the previous articles (47, 59), the production phases were carried out in the NPT ensemble at a temperature of 300K and a pressure of 1bar, using periodic boundary conditions, Particle Mesh Ewald treatment (69) and SHAKE (70). For the van der Waals interactions, a switching function was applied at 10 Å and the cutoff was set to 12 Å. The integration time step was 2 fs and coordinates were saved every 1000 steps (2 ps). The duration of the simulations were 200 ns for SYS1 and SYS2, 300 ns for SYS1-bis and SYS2-bis and 1  $\mu$ s for each 601-dodecamer. For every simulation, the first 50ns were discarded from analyses to account for early tail relaxation and equilibration of the solvent.

## DNA-Histone interface

The interface between DNA and histones was previously analyzed by VLDM (Voronoi Laguerre Delaunay for Macromolecules), a software originally developed for proteins (57, 62, 62) and recently extended to nucleic acids (47, 58). VLDM relies on a partition of space into a collection of polyhedra filling space without overlaps or gaps. The Delaunay tessellation and its Laguerre dual were built from a set of atomic data, each atom being characterized by its position in space and a weight depending on its van der Waals radius. For consistency with the simulations, the van der Waals radii were set to the default values of CHARMM36. To avoid open or

distorted polyhedra in the Laguerre tessellation, an 8 Å thick water layer around the solute was taken together with the solute as input to VLDM. VLDM analyses of simulated nucleosomes were performed on snapshots extracted every 250ps from the trajectories, discarding the first 50ns; only the heavy atoms of the solute and solvent were considered. In this approach, the interface between two molecular groups is a polygonal surface, quantified by its area.

### **Additional trajectory analyzes**

Root Mean Square Deviations (RMSDs) and Root Mean Square Fluctuations (RMSFs) were computed on solute heavy atoms with GROMACS (71). The DNA base-pairing was examined with HBPLUS (72) using cutoffs of 3.9 Å and 90° for the donor-acceptor distance and the donor-hydrogen-acceptor angle, respectively. Helicoidal parameters of DNA were analyzed using Curves+ (73) and 3DNA (74). Because of end melting singularities, the first and last five base pairs of DNA<sub>NCP</sub> were excluded from the analyses; thus the statistics were limited to 136 base pairs and 135 dinucleotides (NpN).

## **RESULTS AND DISCUSSION**

### **Overview of the studied trajectories**

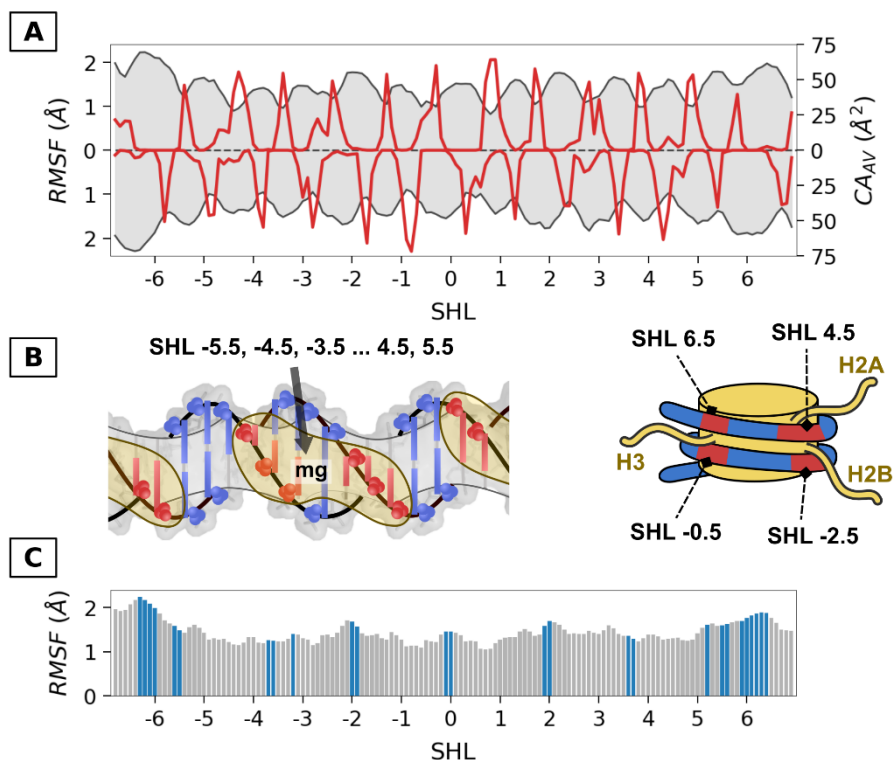
Four nucleosome systems (SYS1, SYS1-bis, SYS2 and SYS2-bis) were considered, which differed initially by the conformations of the histone unstructured domains, called tails (see Figure S1-A, Table S1 and Materials and Methods). As explained in Introduction, the four corresponding simulations were recently used to provide a detailed and quantitative description of the DNA-histone interface in solution (47). In this previous work, the integrity and stability of the simulated nucleosome was checked following standard practice in the analysis of molecular dynamics (MD) simulations. Briefly, the folding of histones was clearly preserved; the tail conformations had no effect on the behavior of the other nucleosome components; no DNA<sub>NCP</sub>-histone disassembly occurred in the trajectories. Calculations of Root Mean Square Deviations (RMSD) specifically for DNA<sub>NCP</sub>, indicated that its overall shape was not severely affected during the trajectories. This is illustrated in Figure S2, showing that simulated DNA<sub>NCP</sub> settles around  $1.5 \pm 0.5$  Å of the average structure, for the four nucleosome systems. The remarkable consistency across the four MDs was in fact also found in the previous analyses, so that the four simulations were considered as a whole (47). We are opting again for the same method in the present work, while providing in Supplementary Data the analyses obtained separately for individual simulations, for key points.

The present work also re-examines simulations previously carried out on four free “601-dodecamers” (59) which, once depleted from their extremities, cover a quarter of the 601 sequence, from SHL -3.8 to SHL -0.2 (Table S2 and Figure S1-B). This extends prior analyses of free DNA by comparisons with DNA<sub>NCP</sub>.

### **Atomic fluctuations of nucleotides in DNA<sub>NCP</sub>**

Root Mean Square Fluctuations (RMSFs) are commonly monitored to estimate the deviation of the position of any structural element with respect to its time-averaged position. They were computed on each nucleotide along DNA<sub>NCP</sub> over all the simulations considered together (Figure 1), taking advantage of the coherency across the

four trajectories (Figure S3).



**Figure 1: Relationship between RMSFs in DNA<sub>NCP</sub> and DNA-histone contacts.**

**A:** Comparison of Root Mean Square Fluctuation (RMSF, grey area) and DNA-histone contact area ( $CA_{av}$ , red lines) profiles along the strands I (y axes oriented upwards) and J (y axes downwards) of DNA<sub>NCP</sub>; the data were calculated for each nucleotide along DNA<sub>NCP</sub>; in the case of CA, the values were time averaged. **B:** Schematic representations of the DNA segments interacting either with the histone structured domains (beige clouds) across the minor groove (mg) (left), or with histone tail roots (right). Contacted and non-contacted nucleotides are colored in red and blue, respectively; the centers of the interacting regions are specified in terms of Super Helix Locations (SHLs). **C:** RMSF of the complementary nucleotides along DNA<sub>NCP</sub>; the grey and blue bars correspond to complementary nucleotides with contact areas larger or smaller than 5 Å<sup>2</sup>, respectively.

The RMSF values of nucleotides follow an oscillating profile along the DNA<sub>NCP</sub> strands but the variations have limited amplitudes, reaching at most  $\sim 0.5$  Å in the innermost region, from SHLs -5 to 5 (Figure 1-A). To examine the histone contact effect on the DNA motions, these RMSF values were compared (Figure 1-A) to the direct DNA-histone interactions previously quantified by VLDM in terms of contact areas (47) (see also Material and Methods). The contacts involving the histone structured domains are clustered into blocks along each DNA strand I and J (Figure 1-A); these blocks group themselves pairwise into larger blocks as the same amino acids contact both strands I and J across the minor groove (Figure 1-B); the successive blocks are spaced by  $\sim 10$  nucleotides and disposed periodically so that their centers are symmetrically located at SHLs  $\pm 0.5, 1.5, 2.5, 3.5, 4.5, 5.5$  and  $6.5$ . Additional interactions centered at SHLs  $\pm 0.5, 2.5, 4.5$  and  $6.5$  involve the

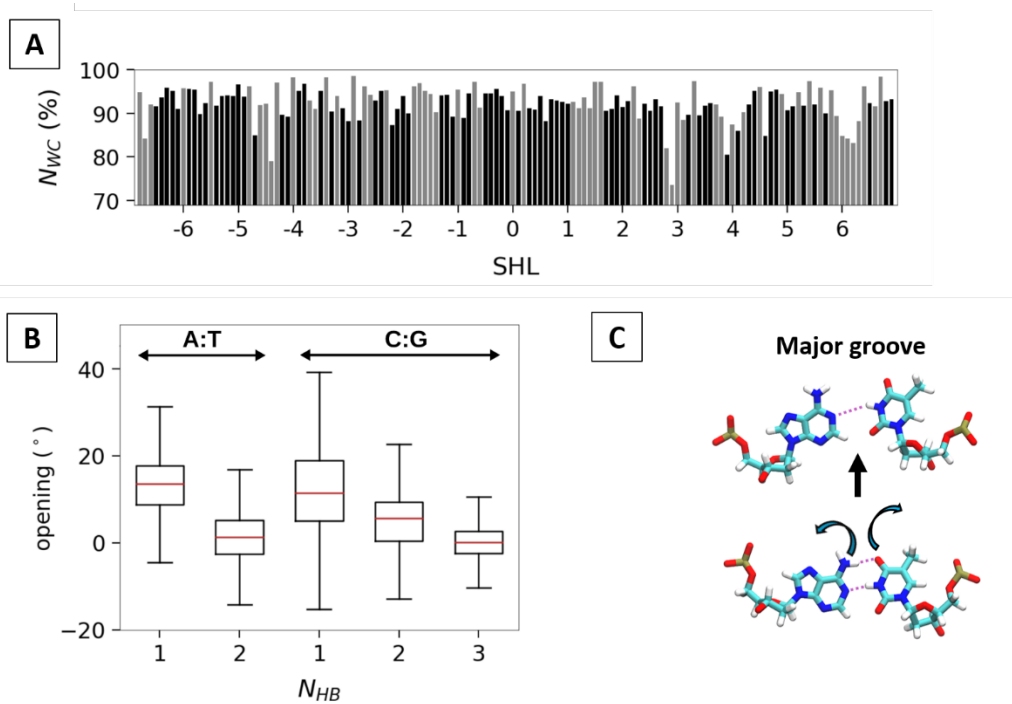
histone tail roots, in particular those of H3 and H2B that pass between the two DNA super helical turns juxtaposed one above the other (Figure 1-B). Overall, the DNA<sub>NCP</sub>-histone contact areas are periodic along strands I and J and have comparable values at each contacted SHL (Figure 1-A).

Strand by strand, the variations of RMSFs and histone contact areas (Figure 1-A) are anti-phased, supporting the physical idea that motions are reduced at the DNA-histone interface, because of interactions. However, this pattern becomes much less marked when one considers base pairs instead of single nucleotides (Figure 1-C). This occurs because most base pairs are contacted only on one nucleotide, the other remaining non-contacted as illustrated in Figure 1-B; therefore, only rare, isolated DNA<sub>NCP</sub> spots escape from interaction with histones (Figure 1-C). One can thus anticipate that the variability of the double helix descriptors (roll, twist ...) is homogeneously affected by the DNA-histone interface all along DNA<sub>NCP</sub>. But before further analyzing in detail the behavior of DNA<sub>NCP</sub> structural elements, we examined the base pairing which is a fundamental aspect of DNA integrity.

### **Stability of base pairs in DNA<sub>NCP</sub>**

The stress associated to the severe bending of DNA<sub>NCP</sub> could affect the base pairing by weakening the hydrogen bonds between the base partners, motivating their inspection. Numerically, the hydrogen bonds were defined by two criteria, a donor-acceptor distance lower than 3.9 Å and a donor-hydrogen-acceptor angle of more than 90°. The DNA starting structure for the simulations contained many distorted base pairs that were regularized in the very early stages of simulations, not considered in analyses. After this equilibrium setup stage, the Watson-Crick scheme is maintained during 92 and 93% of the MD trajectory for A:T and G:C base pairs, respectively (Figure 2-A). The same values were obtained from the simulations of the free 601-dodecamers.

In DNA<sub>NCP</sub>, the disrupted base pairs contain at least one intact hydrogen bond; the complete loss of hydrogen bonds is an extremely rare event, affecting less than 0.1% of base pairs extracted from the combined MDs. Among the six parameters (shear, stretch, stagger, buckle, propeller and opening) describing the base pairs (75), only the opening accounts for the percentage of incorrectly paired structures (correlation coefficient of 0.65). Since the opening values are mainly positive and do not exceed 40° (Figure 2-B), the base pairing disruptions in DNA<sub>NCP</sub> should be perceived as semi-open states corresponding to a partial flip of the bases towards the major groove, as illustrated for an A:T base pair in Figure 2-C. Such trend to open towards the major groove rather than the minor groove was also observed on free oligomers (76), comprising the free 601-dodecamers.



**Figure 2: Watson-Crick pairing in  $DNA_{NCP}$ .**

**A:** The percentage of MD snapshots ( $N_{WC}$  %) in which the Watson-Crick pairing is strictly respected is plotted for each base pair along the 601 sequence. The bars in black and grey correspond to G:C and A:T base pairs, respectively. The standard deviations calculated for each base pair are low, from 0.3 to 4%, apart from one case at 10%. **B:** The boxplots represent statistics on the opening values associated to A:T base pairs with one or two hydrogen bonds and G:C base pairs with one, two or three hydrogen bonds. The few G:C base pairs with only one hydrogen bond contribute to the large opening variability. In these plots, the rectangle represents the second and third quartiles, the red line inside indicates the median value and the lower and upper quartiles are shown as points either side of the rectangle. **C:** Molecular representations of A:T base pairs either perfectly paired (bottom model, opening of  $2^\circ$ ) or partially flipped towards the major groove (top model, opening of  $44^\circ$ ).

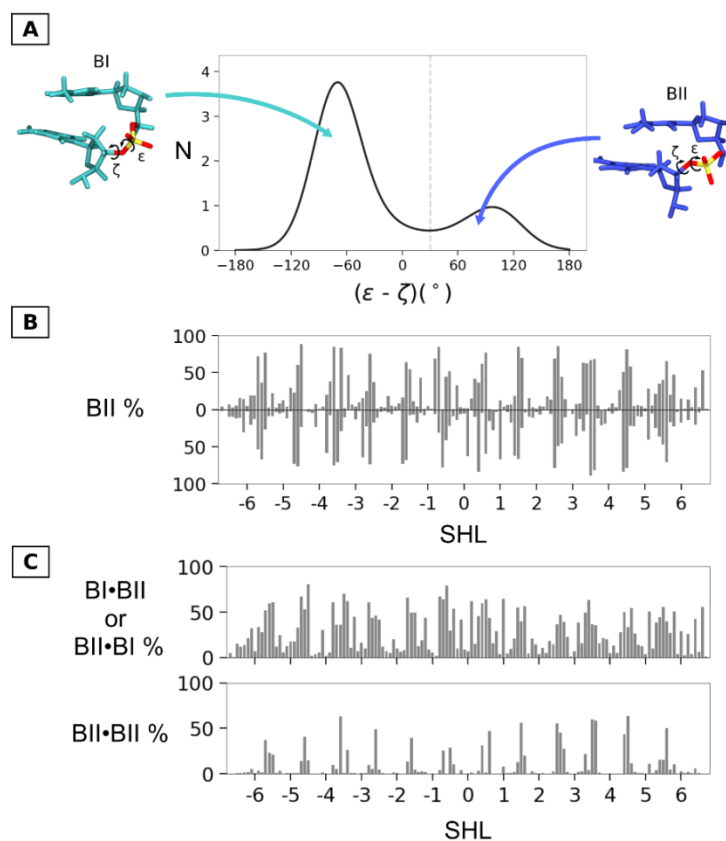
The three regions most susceptible to transitory escape from strict base pairing are composed of four (SHLs -4.5/-4 and 3) or seven (SHL 6.5) successive A:T base pairs (Figure 2-A). However, such A:T accumulation on its own cannot explain the pairing weakness since the two TTAAA segments symmetrically located at SHLs  $\pm 1.5$  remain well paired during our simulations (Figure 2-A). A point common to the three DNA regions at SHLs -4.5/-4, 3, and 6.5 – but absent at SHLs  $\pm 1.5$  – is the multiplicity of contacts occurring with both histone cores and tail roots (47). Although no specific, direct interaction stabilizes the partially open base pairs, the instability of the base pairing in these three regions may reflect the stress caused by an outstanding concentration of contacts, and the effect may be facilitated by the weaker strength of A:T pairing. So, acting as mechanical fuse, A:T pairs at these positions may actually be more favorable to the complex than more robust G:C pairs and may thus be a factor contributing to the binding affinity of the 601 sequence.

The existence of semi-open states, even infrequent, could be exploited in nuclear processes such as the

methylation of the cytosine C5 atom, which is initiated *via* the major groove (77) and can occur when the DNA is organized in nucleosomes (78). Nevertheless, apart from a few specific segments, the Watson-Crick hydrogen-bonding of base pairs is overwhelmingly stable along the DNA<sub>NCP</sub>, remaining central to the DNA double-helical organization.

### BI ↔ BII equilibrium in DNA<sub>NCP</sub> backbone

BI and BII states correspond to correlated crankshaft motions of  $\epsilon$  and  $\zeta$  backbone angles (Figure 3). They are characterized by the pseudo angle  $(\epsilon - \zeta)$ , as indicated by the X-axis of Figure 3A. Indeed, the distribution of the  $(\epsilon - \zeta)$  values in DNA<sub>NCP</sub> shows a bimodal regime in which the peaks of the BI and BII populations are separated by a minimum located at  $(\epsilon - \zeta) = 30^\circ$  (Figure 3-A), in excellent accordance with analyses of X-ray structures of free DNA (39), and simulations of free 601-dodecamers (59). Hence, this minimum value was and is used to define BI ( $(\epsilon - \zeta) < 30^\circ$ ) and BII ( $(\epsilon - \zeta) > 30^\circ$ ) states.



**Figure 3: Distributions of BI and BII conformations in DNA<sub>NCP</sub>.**

**A:** Distribution of  $(\epsilon - \zeta)$  values in DNA<sub>NCP</sub> and typical structural representations of each of the two phosphate group conformations, BI (cyan) and BII (blue) conformations. **B:** BII percentages (BII %) of each phosphate group  $p_i$  and  $p_j$ , along the strands I (y axis oriented upwards) and J (y axis downwards) of DNA<sub>NCP</sub>; the standard deviations are on average 5.5 %. **C:** Percentages of the two combinations of facing phosphate groups  $p_i \cdot p_j$  containing one (BI•BII or BII•BI) or two (BII•BII) BII conformers along DNA<sub>NCP</sub>. The standard deviations are 6 and 3 % for BI•BII and BII•BII%, respectively.



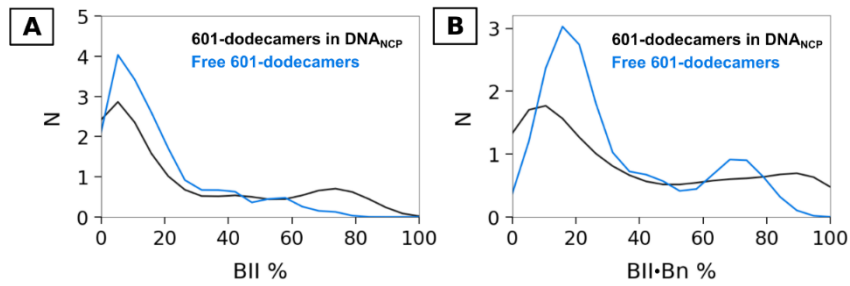
The global percentage of BII conformers in DNA<sub>NCP</sub> is consistent across the simulations, between 22 and 23 % on average in each trajectory (Figure S4). Also, the comparison of the BII percentages associated to each phosphate group (p in NpN, N standing for any nucleotide), extracted from each trajectory, leads to cross-correlation coefficients exceeding 0.85, another strong justification for merging of the simulations.

The global percentage of BII conformers is slightly higher in DNA<sub>NCP</sub> than in free DNA (23% vs 18%). In addition, 51 steps out of 270 have BII populations exceeding 50% in DNA<sub>NCP</sub>, while such percentages are exceptional in free DNA according to both experimental (41, 46, 79) and computational (44, 59) data. Hence, the distributions of BII percentages show an increase of the number of very BII rich steps in bound 601-dodecamers (within NCP) compared to their free counterparts (Figure 4-A).

The BI and BII populations show a striking periodic profile along both DNA<sub>NCP</sub> strands (Figure 3-B), where BI- and BII-rich regions alternate. Such pronounced periodicity is specific to DNA<sub>NCP</sub> and, to our knowledge, not observed in free DNA, in particular in oligomers related to the 601 sequence (Figure S5). From a methodological point of view, it shows that the CHARMM36 force-field is sufficiently well-tuned for DNA to respond differently to changing environments, not a trivial achievement. No backbone conformational periodicity is observed in the main NCP X-ray structure (3MVD) used to build the model to initiate the MDs, which likely follows from the 3 Å resolution of 3MVD, not sufficient to capture all DNA details. On the other hand, 1KX5, the best resolved X-ray NCP structure formed with a DNA from the human  $\alpha$ -satellite, presents a similar alternation of BI and BII states (41, 80), implying that the profile obtained here is emerging as a likely common feature of DNA<sub>NCP</sub>.

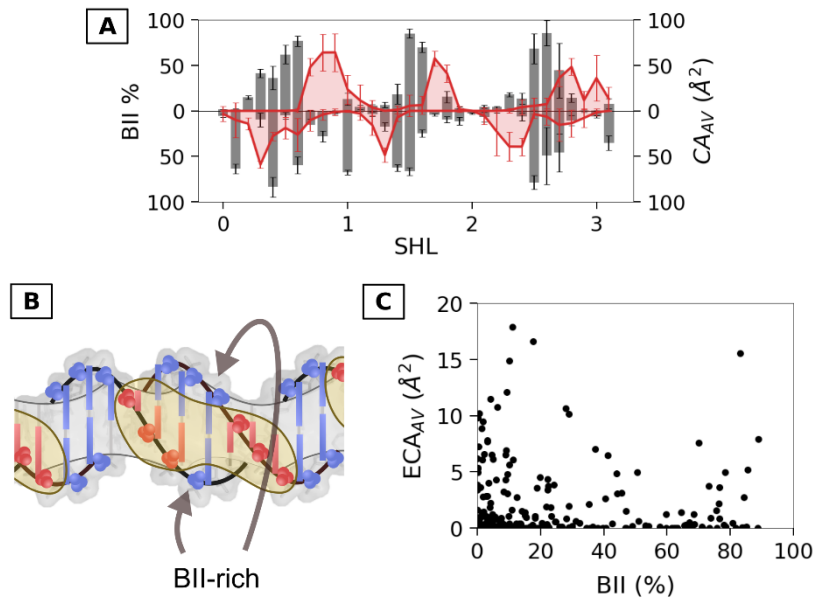
The particular distribution of BI- and BII-rich steps along the strands I and J of DNA<sub>NCP</sub> (Figure 3-B) is corroborated by the combinations explored by the facing phosphate groups, pi and pj. Thus, pi•pj tracts predominantly in BI•BI alternate with segments mainly exploring the three BII-containing combinations, BI•BII, BII•BI and BII•BII (Figure 3-C), here called BII•Bn, n = I or II. Note that the periodic alternation emerges especially clearly in the BII•BII population profile where high values punctuate the bound 601 sequence by repeatedly arising every 10th bp, at SHLs  $\pm$  5.5, 4.5, 3.5 etc... (Figure 3-C). As for BII percentages, the comparison of free and bound 601-dodecamers shows different distributions of BII•Bn percentages, more spread in DNA<sub>NCP</sub> than in free DNA (Figure 4-B).

The presence of BII hot spots in DNA<sub>NCP</sub> raises the question of a potential stabilization of this state by the histones. BII-rich regions are mainly located between two contact blocks involving the strands I and J across the minor groove (Figures 5-A, 5-B and S6). There is no notable correlation between the BII percentages and the contacts between the corresponding phosphate group atoms and amino acids (Figure 5-C). Thus, BII conformers are not promoted or stabilized by direct contacts to the histones. So, other factors must be considered to understand why the BII populations are so marked and prevalent within the NCP. The following provides answers which invoke structural couplings implicated in the wrapping requirements, backbone states and helical parameters. First, we consider inter base parameters.



**Figure 4: Distribution of BII percentages in free and bound 601 sequences.**

Distribution of BII (A) and BII•Bn (B) percentages in bound (black) and free (blue) 601-dodecamers. BII•Bn corresponds to the three BII-containing combinations adopted by facing phosphate groups, BI•BII, BII•BI and BII•BII, in paired nucleotides. Each distribution is calculated by dividing the range of the considered parameter (here from 0 to 100) into 20 slices.  $N$  is the count of occurrences in each slice.

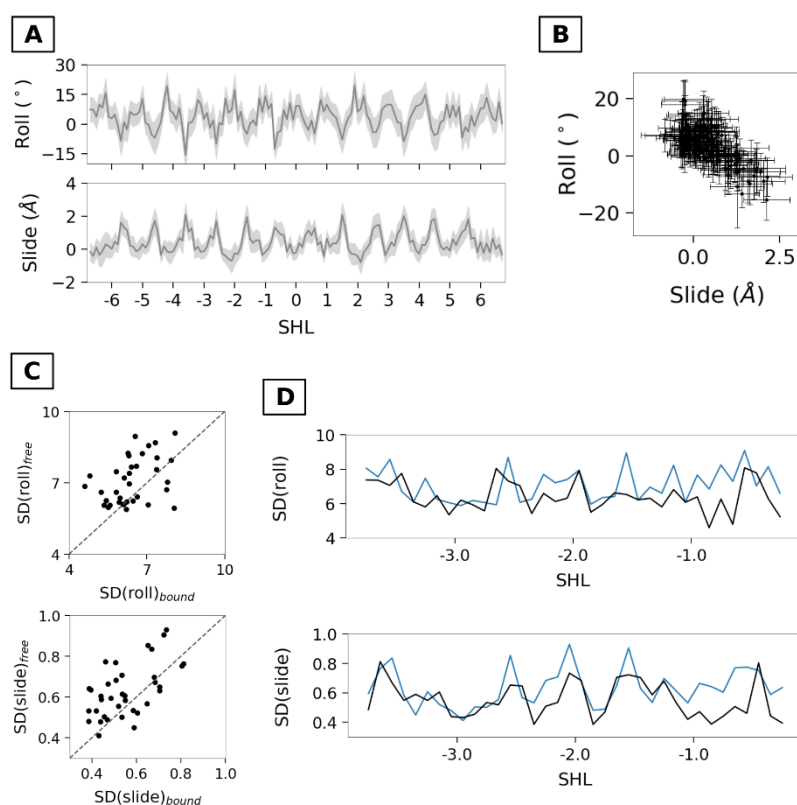


**Figure 5: BII conformers in  $DNA_{NCP}$  versus DNA-histone contacts.**

**A:** Detailed view of BII percentages of the dinucleotide steps (grey bars) and contact areas (red area, values averaged across the simulations,  $CA_{av}$ ) between these steps and the histone structured domains along strands I (y axes oriented upwards) and J (y axes downwards); the plot focuses on the interval from SHL 0 to SHL 3; the vertical bars correspond to the standard deviations. **B:** Schematic representation of the location of BII-rich regions compared to the contacts between  $DNA_{NCP}$  and the histone structured domains. Contacted and non-contacted nucleotides are colored in red and blue, respectively; contact regions of histone structured domains are represented by beige clouds. **C:** For each  $DNA_{NCP}$  phosphodiester junction, the electrostatic contact area (averaged across the simulations,  $ECA_{av}$ ) between the histones and the O3', O1P, O2P and O5' atoms was plotted against the BII percentage (BII %).

## Inter base pair parameters in DNA<sub>NCP</sub>

At the level of complementary dinucleotide, the structure is usually described by six inter base pair parameters, shift, slide, rise, tilt, roll and twist. Marked periodic, sinusoidal profiles characterize the roll and slide values along DNA<sub>NCP</sub> (Figures 6-A). The other parameters, shift, rise, tilt and twist, show more diffuse profiles (Figure S7), in line with their poor reproducibility observed across NCP X-ray structures (11, 49, 50). Overall, the parameter values do not indicate severe local distortions in simulated DNA<sub>NCP</sub> (Figures 6-A and S7), as further shown by the very similar distributions of the inter base pair parameters of bound and free 601 sequences (Figure S8-A). Only roll and slide in bound 601-dodecamers differ from their free counterparts by a slight excess of negative and positive values, respectively (Figure S8-A). As an additional remark, significant couplings exist between pairs of DNA<sub>NCP</sub> parameters, such as slide/roll (Figure 6-B), slide/twist and twist/roll (Figure S8-B). They are reinforced in DNA<sub>NCP</sub> compared to free 601-dodecamers (Figure S8-B) as previously described from large datasets of X-ray structures of DNA free or bound to proteins (39, 81). To our knowledge, this feature remains intriguing.



**Figure 6: Variability of inter base pair parameters in DNA<sub>NCP</sub>.**

**A:** Profiles of average values of roll and slide along DNA<sub>NCP</sub>; the grey areas represent the standard deviations. **B:** Relationship between roll and slide values; vertical and horizontal bars are standard deviations. **C and D:** Comparisons of the standard deviations of the slide and roll values in free ( $SD_{free}$ ) and bound ( $SD_{bound}$ ) 601-dodecamers: correlation plots, with the  $y=x$  diagonal indicated by a dashed line (C) and profiles of standard deviations along the 601-dodecamer sequence, histone free (blue) or histone bound (black) (D).

According to the standard deviations associated to the average value profiles (Figures 6-A and S7), each DNA<sub>NCP</sub> helical parameter presents some variability. Whether this variability is restrained by the presence of histones was estimated by comparing data from bound and free 601-dodecamers. The standard deviations of helical parameters are often smaller for bound steps than for free steps (Figure S9), in particular with slide and roll (Figure 6-C). Only the rise escapes this trend (Figure S9). In addition, several bound and free standard deviations significantly compare (Figures 6-C and S9-B) as shown by the corresponding correlation coefficients (Table 1). These correlations are also perceptible in the similarities existing between the profiles of bound and free standard deviations (Figure S9-A) as for those of slide and roll (Figure 6-C). So, the variability of DNA<sub>NCP</sub> helical parameters appears to reflect the intrinsic properties of DNA more than the interactions with the histones. To further test this proposal, correlation coefficients were calculated for pairs of variables, *i.e.* standard deviations of each type of inter base pair parameters *vs* DNA/histone contact areas calculated for complementary nucleotides. The very modest correlation coefficients (Table 1) confirm that the plasticity of the considered helical parameters is not primarily in response to the size of the local interfaces.

|       | Correlation coefficients |                       |
|-------|--------------------------|-----------------------|
|       | free <i>vs</i> bound SD  | Bound SD <i>vs</i> CA |
| Shift | 0.68                     | -0.35                 |
| Rise  | 0.12                     | -0.22                 |
| Tilt  | 0.57                     | -0.20                 |
| Slide | 0.57                     | -0.33                 |
| Roll  | 0.40                     | -0.20                 |
| Twist | 0.33                     | -0.18                 |

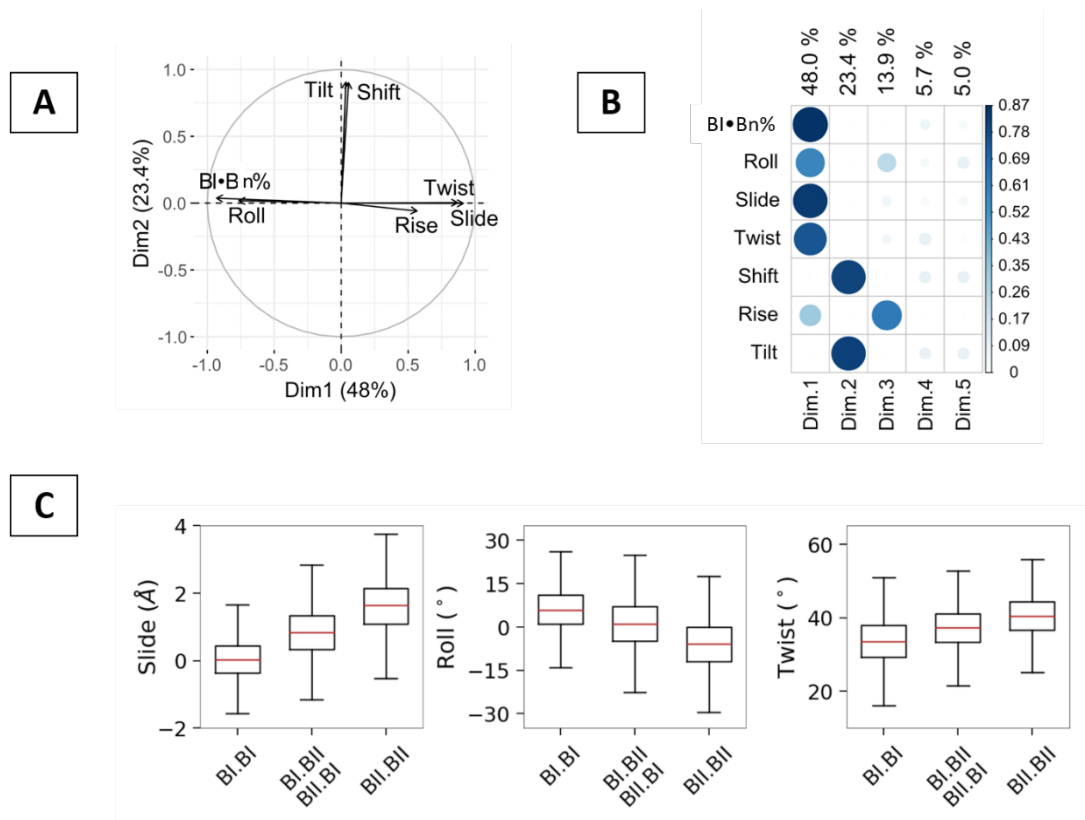
**Table 1: Comparisons involving standard deviations of helical parameters in DNA<sub>NCP</sub>.**

*The standard deviations of the six helical parameters in DNA<sub>NCP</sub> were compared to their free counterparts and to the DNA/histone contact areas (CA). This Table reports the associated correlation coefficients.*

The consideration of inter base pair parameters underlines that the specific signature of DNA<sub>NCP</sub> helical parameters is expressed by well-defined periodic profiles of roll and slide average values along the sequence; however, this typical alternation is not associated with extreme, outstanding parameter values. The variabilities of most inter base pair parameters along DNA<sub>NCP</sub>, quantified by standard deviations, parallel those of their free homologs. The next step is to examine whether the free B-DNA couplings involving inter base pair parameters and backbone conformations remain pertinent in NCP.

### **Structural couplings at the dinucleotide level in DNA<sub>NCP</sub>**

An interdependence of structural descriptors is suggested by the visual inspection of variations of BII•Bn combinations (Figure 3-C), roll and slide (Figure 6-A) along DNA<sub>NCP</sub>. A more comprehensive picture of the couplings within DNA<sub>NCP</sub> is provided by principal component analysis (PCA), which was applied to the six inter base pair parameters and the BII•Bn populations examined in the previous section.



**Figure 7: Structural couplings in  $DNA_{NCP}$  complementary dinucleotides.**

**A** and **B**: Results of PCA carried out on a data set containing the average values of the six inter base pair parameters and of the frequency of the BII-containing combinations, BI•BII, BII•BI and BII•BII (BI•Bn%).

**A**: Correlation circle plot in the plane of the first two principal components, Dim 1 on the horizontal x axis and Dim 2 on the vertical y axis. Each variable is represented by a vector; the contribution of a given variable to the total variance is related to the vector length; a cluster of parallel vectors indicates correlated (vectors with similar directions) or anti-correlated (vectors of opposite directions) variables, respectively. **B**: Details about the first five PC dimensions. The circles are associated to the variables indicated on the left; both circle size and color shade code the contribution of the considered variable to the total variance (color scale given on the right). The percentage of total variance explained by each PC is given above the boxes. **C**: Boxplots representing the roll, slide and twist values associated to the observed combinations of facing phosphates, BI•BI, BI•BII or BII•BI, and BII•BII in  $DNA_{NCP}$ .

The first two PC dimensions are highly meaningful, explaining 71 % of the total variance. The first dimension arises from a strong dependence between  $\pi_i \cdot \pi_j$  combinations, slide, roll, twist and, to a lesser extent, rise (Figures 7-A and 7-B). Indeed, the three different  $\pi_i \cdot \pi_j$  combinations are intimately associated with specific values of slide, roll and twist (Figure 7-C). The second PC dimension shows the interdependence between tilt and shift (Figures 7-A and 7-B), the shift being known to alleviate clashes from tilt (81).

The concerted changes of phosphodiester junctions, slide, roll and twist can also be illustrated by dividing the complementary dinucleotides into four categories according to the roll value, extreme negative,

intermediate negative, intermediate positive, extreme positive (Table 2). Considering in addition the DNA-histone contact areas enables us to broaden the idea developed from the preferred locations of BI and BII-very rich regions (Figure 5): the extreme roll regions occur predominantly outside the most extensive contacts with the histones (Table 2).

| Roll (°) categories | Roll (°)   | BII•Bn% | Slide (Å) | Twist (°)  | CA (Å <sup>2</sup> ). |
|---------------------|------------|---------|-----------|------------|-----------------------|
| Roll < -5           | -8.3 ± 3.3 | 90 ± 10 | 1.7 ± 0.4 | 41.1 ± 3.3 | 23 ± 28               |
| -5 ≤ Roll < 0       | -2.3 ± 1.5 | 67 ± 25 | 1.0 ± 0.5 | 37.9 ± 3.0 | 41 ± 30               |
| 0 ≤ Roll < 5        | 2.4 ± 1.4  | 26 ± 24 | 0.3 ± 0.4 | 34.4 ± 3.3 | 42 ± 30               |
| 5 ≤ Roll            | 9.5 ± 3.5  | 21 ± 20 | 0.1 ± 0.4 | 33.0 ± 3.2 | 26 ± 31               |

**Table 2. Structural parameters as a function of roll category.**

*The complementary dinucleotides in simulated DNA<sub>NCP</sub> were classified in 4 roll categories (first column). The parameter values presented in the five other columns are averages over the MD snapshots. BII•Bn correspond to the three BII-containing combinations, BI•BII, BII•BI and BII•BII. CA designates the contact areas measured between the complementary dinucleotides and the histones.*

In sum, the notable finding is that most aspects of DNA internal mechanics emerging from solution studies of free DNA (41, 42, 59) are preserved and expressed in DNA<sub>NCP</sub>. We also understand that pi•pj combinations contribute to the DNA<sub>NCP</sub> pathway around the histones because they are inherently coupled to specific adjustments of slide and roll, two helical parameters previously considered as the major factors accompanying DNA wrapping (9, 13). Thus, some places in DNA<sub>NCP</sub> facilitate wrapping if they can adopt conducive roll/slide values and adapted phosphate group combinations. Since the dinucleotide intrinsic ability to populate BII conformations depends on sequence, one expects some DNA sequences to favor wrapping in the NCP more than others.

#### **DNA intrinsic sequence effect and structural characteristics of DNA<sub>NCP</sub>.**

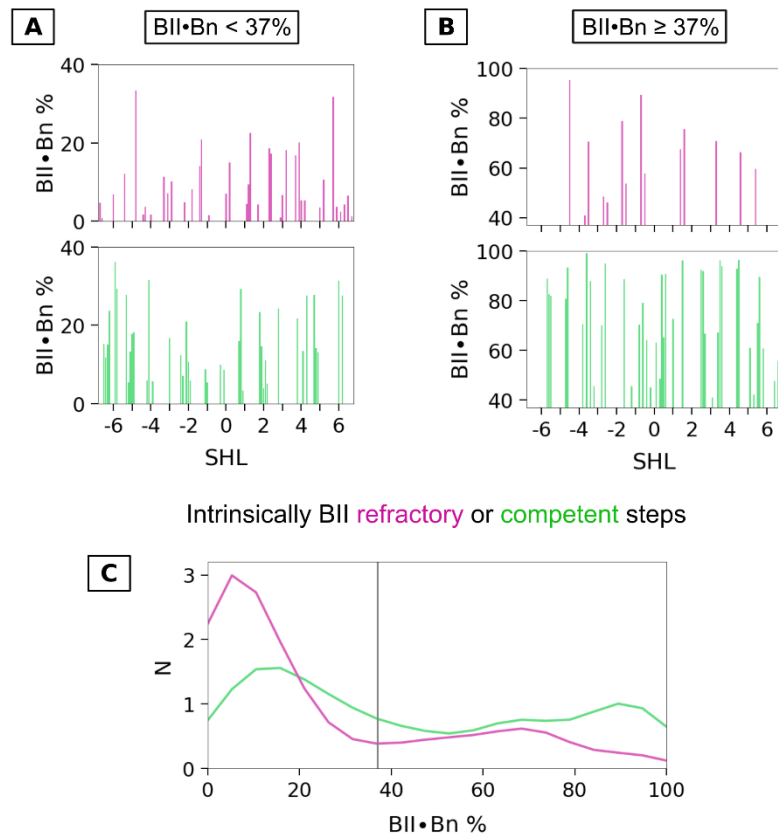
We recall that, in free DNA, NMR-inferred populations of phosphate group conformers (43) primarily depend on the dinucleotide sequence, each dinucleotide type being associated to a specific BII propensity (41, 46, 59). Keeping in mind that, globally, the BI state remains the major conformation, following experimental quantifications, the free complementary dinucleotides can be classified in i) BII competent steps that explore BII more (CpA•TpG, GpG•CpC and CpG•CpG, with at least 40% of BII conformers in both facing phosphate groups) or less (GpC•GpC, GpA•TpC and TpA•TpA, with 15 – 25% of BII conformers in both facing phosphate groups) frequently, and ii) mostly BI (or BII refractory) steps, ApT•ApT, ApC•GpT, ApA•TpT and ApG•CpT, with at most 9% of BII conformers on one strand. As already said, this classification is satisfactorily reproduced by the CHARMM36 DNA force-field despite residual biases yielding too high and too low BII propensities of TpA and GpC, respectively (59). Because of the tight relationship between backbone states and helical parameters, BII competent and mostly BI phosphate groups correspond to intrinsically flexible and stiff steps, respectively. The backbone states are in fact an effective reporter of the sequence-dependent, local

deformability of DNA and its intramolecular energetics.

It has been observed that DNA sequences which contain large regions characterized by a ~10 bp alternation of intrinsically BII competent and mostly BI segments show enhanced affinity for the histone octamer (9, 41, 46) and favor nucleosome assembly (82). Here, the simulations show that the DNA wrapping implicates successive segments with low and high BI populations (equivalent to high and low BII•Bn populations) along the sequence. The next investigation thus focuses on the composition of DNA<sub>NCP</sub> BI and BII-rich regions in terms of intrinsically BII refractory or competent steps.

Examining the actual BI and BII distributions along DNA<sub>NCP</sub> first highlights that DNA<sub>NCP</sub> BI-rich regions contain almost equally intrinsically BI- and BII-prone steps (Figure 8-A). The presence of intrinsically BII competent steps in DNA<sub>NCP</sub> BI regions is in fact not surprising since such dinucleotides spontaneously explore both BII and BI conformations, with BII percentages very rarely exceeding 50% in free DNA (41, 46, 59). In contrast, DNA<sub>NCP</sub> BII-rich regions include a very limited number of intrinsically BI dinucleotides (Figure 8-B): they represent only 23% of the steps displaying BII•Bn percentages higher than 37%, the overall average of these combinations in DNA<sub>NCP</sub>. Indeed, forcing BI prone steps to form BII rich regions in DNA<sub>NCP</sub> would imply internal stress and an energetic penalty. In line with the above, the distribution of DNA<sub>NCP</sub> BII•Bn percentage shows a clear maximum around 10% for intrinsically BI prone steps, ApT•ApT, ApC•GpT, ApA•TpT and ApG•CpT, but a more uniform profile for intrinsically BII competent steps (Figure 8-C).

In sum, these analyses strongly support that the intrinsic properties of the dinucleotides along the 601 sequence are exploited in NCP formation, as postulated before (41, 46, 82). They also suggest that a critical point may be the location of BII refractory dinucleotides along the nucleosomal sequences.



**Figure 8: Sequence dependence of BII•Bn population in  $DNA_{NCP}$ .**

The plots of both top panels were based on a double selection: i)  $DNA_{NCP}$  BII•Bn percentages (BII•Bn %) either (A) lower than, or (B) equal to or higher than the overall average of BII•Bn percentage which is 37%; ii) complementary dinucleotides intrinsically either BI prone (ApT•ApT, ApC•GpT, ApA•TpT and ApG•CpT) (pink bars) or BII competent steps (green bars). Panel C: Distribution of BII•Bn % in  $DNA_{NCP}$  for the complementary dinucleotides either intrinsically BI prone (pink line) or BII competent (green line). Each distribution was calculated by dividing the X-axis range (0 to 100) into 20 equal slices. N is the count of occurrences in each slice. The grey vertical line represents the overall average of BII•Bn % in the MD of  $DNA_{NCP}$ . BII•Bn corresponds to BI•BII, BII•BI and BII•BII.

## CONCLUSIONS

The widespread descriptions of NCP have popularized the picture of a packed DNA appearing quite different from its free state, because of its super-helical path wrapped around the histone structured domains. Such tight bending around the histones could suggest a  $DNA_{NCP}$  state with dramatically altered properties, including highly restricted dynamics. The DNA-histone interface seems *a priori* to corroborate this view, considering the dense interaction network that maintains the cohesion of the complex (47). However, the intuition of an inhibited  $DNA_{NCP}$  dynamics is not based on any data or objective analysis, so we decided to examine more closely the DNA behavior within NCP using MD simulations.

Unexpected analogies between free and bound DNA structural features were uncovered. The base pairing in  $DNA_{NCP}$  is as stable as in free DNA; the inter base pairs parameters cover similar ranges in free and



bound DNA, without recognizable outstanding hinge points. Thus, expressions such as “extreme DNA alterations” or “distorted double helix” appear as largely exaggerated for describing DNA<sub>NCP</sub>. Free DNA and DNA<sub>NCP</sub> share other characteristics, including the variability of inter base pair parameters and the interdependence of the backbone conformational states with slide, roll, and twist. DNA<sub>NCP</sub> therefore remains substantially malleable along its sequence. The absence of any significant restriction caused by the histones to the DNA local dynamics is advantageous regarding entropy. In addition, the somewhat broad conformational space accessible to some DNA<sub>NCP</sub> steps may well play a role to accommodate protein binding by the DNA while on the nucleosome.

Of course, some features are particular to DNA<sub>NCP</sub>, such as BII percentages reaching unusual extreme values. Indeed, the super-helical pathway and the afferent curvature require the alternation of two types of regions along DNA<sub>NCP</sub>, one with BII-rich conformers/negative rolls/positive slide and the other with BI-rich conformers/positive or null rolls/negative slide. So, the periodic oscillation of high and low BII percentages along the sequence participates to the DNA pathway imposed by the 3D organization of the NCP in addition to the well-known variations of roll and slide values.

Actually, this result strengthens the idea that the sequence-dependent intrinsic structural properties of DNA, comprising the 601 sequence and its derivatives, contribute to modulate the affinity for the histone core. In previous papers, we annotated DNA sequences optimized for nucleosome formation with the experimental TRX scale. We found that more or less extended regions of the examined DNAs showed a ~10 base-pair periodic alternation of intrinsically stiff, BI-rich, and flexible, BII-rich, regions (41); the extent of the regions where such alternation was observable correlated the affinity for histones (46). Recently, we also experimentally revealed that modifying this alternation by mutating some bases of the 601 sequence negatively affected the NCP assembly (82). The simulation analyses presented in this work complement those previous studies by showing that i) there is an alternation of BI and BII-rich regions in DNA<sub>NCP</sub> and ii) DNA<sub>NCP</sub> BII-rich regions mainly contain intrinsically BII-competent steps whereas DNA<sub>NCP</sub> BI-rich regions are equally composed of BII-refractory and BII-competent steps, both types of dinucleotides exploring BI conformers. Thus, we better understand the scenario of the efficiency of the 601 sequence to assemble into nucleosome, even if the correspondence between the behavior of free and bound regions is not perfect: the favorable conformational pre-organization of this DNA is well enough marked to reduce the energetic penalty by preventing deformation cost from an excess of stiff dinucleotides or entropy loss from an excess of flexible dinucleotides. This rational foundation opens up the exciting methodological opportunity to design new DNA sequences more or less adapted to binding the histones upon internal mechanics criteria, consequently to get control over the DNA compaction.

More broadly, it is of course difficult to extrapolate our findings to the situation *in vivo*. The 601 sequence is a very particular case and its high positioning character is not really compatible with the necessity to continuously assemble and disassemble nucleosomes in the cell. We therefore anticipate that sequences with similar properties of pre-organization are very rare in the genomes. Nevertheless, it seems reasonable to expect that the DNA sequence acts in conjunction with the numerous trans factors that regulate chromatin remodeler binding.

## Acknowledgments / Fundings

The simulations were carried out on the GENCI-CEA platform.

## Supporting Information Available

## Author Contributions

Romain Retureau carried out the analyses of simulations that were initially performed by Ahmad Elbahnsi, Romain Retureau, Nicolas Foloppe, Christophe Oguey and Brigitte Hartmann discussed the results and commented on the paper. Brigitte Hartmann wrote the manuscript with the help of Nicolas Foloppe and Christophe Oguey.

## REFERENCES

1. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
2. McGinty,R.K. and Tan,S. (2015) Nucleosome Structure and Function. *Chem. Rev.*, **115**, 2255–2273.
3. Kornberg,R.D. and Lorch,Y. (1999) Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell*, **98**, 285–294.
4. Zhou,K., Gaullier,G. and Luger,K. (2018) Nucleosome structure and dynamics are coming of age. *Nat. Struct. Mol. Biol.*, 10.1038/s41594-018-0166-x.
5. Cutter,A.R. and Hayes,J.J. (2015) A brief review of nucleosome structure. *FEBS Lett.*, **589**, 2914–2922.
6. Luger,K. (2001) Nucleosomes: Structure and Function. In John Wiley & Sons, Ltd (ed), *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester.
7. Luger,K. and Richmond,T.J. (1998) DNA binding within the nucleosome core. *Curr. Opin. Struct. Biol.*, **8**, 33–40.
8. Wu,B., Mohideen,K., Vasudevan,D. and Davey,C.A. (2010) Structural insight into the sequence dependence of nucleosome positioning. *Struct. Lond. Engl. 1993*, **18**, 528–536.
9. Olson,W.K. and Zhurkin,V.B. (2011) Working the kinks out of nucleosomal DNA. *Curr. Opin. Struct. Biol.*, **21**, 348–357.
10. Yang,X. and Yan,Y. (2011) Statistical investigation of position-specific deformation pattern of nucleosome DNA based on multiple conformational properties. *Bioinformatics*, **7**, 120–124.
11. Bishop,T.C. (2008) Geometry of the nucleosomal DNA superhelix. *Biophys. J.*, **95**, 1007–1017.
12. Xu,F. and Olson,W.K. (2010) DNA Architecture, Deformability, and Nucleosome Positioning. *J. Biomol. Struct. Dyn.*, **27**, 725–739.
13. Tolstorukov,M.Y., Colasanti,A.V., McCandlish,D., Olson,W.K. and Zhurkin,V.B. (2007) A Novel ‘Roll-and-Slide’ Mechanism of DNA Folding in Chromatin. Implications for Nucleosome Positioning. *J. Mol. Biol.*, **371**, 725–738.
14. Tan,S. and Davey,C.A. (2011) Nucleosome structural studies. *Curr. Opin. Struct. Biol.*, **21**, 128–136.
15. Ong,M.S., Richmond,T.J. and Davey,C.A. (2007) DNA stretching and extreme kinking in the nucleosome core. *J. Mol. Biol.*, **368**, 1067–1074.
16. Edayathumangalam,R.S., Weyermann,P., Dervan,P.B., Gottesfeld,J.M. and Luger,K. (2005) Nucleosomes in solution exist as a mixture of twist-defect states. *J. Mol. Biol.*, **345**, 103–114.
17. Muthurajan,U.M., Park,Y.-J., Edayathumangalam,R.S., Suto,R.K., Chakravarthy,S., Dyer,P.N. and Luger,K. (2003) Structure and dynamics of nucleosomal DNA. *Biopolymers*, **68**, 547–556.
18. Hughes,A.L. and Rando,O.J. (2014) Mechanisms underlying nucleosome positioning in vivo. *Annu. Rev. Biophys.*, **43**, 41–63.
19. Lieleg,C., Krietenstein,N., Walker,M. and Korber,P. (2015) Nucleosome positioning in yeasts: methods, maps, and mechanisms. *Chromosoma*, **124**, 131–151.

20. McGinty,R.K. and Tan,S. (2016) Recognition of the nucleosome by chromatin factors and enzymes. *Curr. Opin. Struct. Biol.*, **37**, 54–61.
21. Volokh,O.I., Derkacheva,N.I., Studitsky,V.M. and Sokolova,O.S. (2016) Structural studies of chromatin remodeling factors. *Mol. Biol.*, **50**, 812–822.
22. Speranzini,V., Pilotto,S., Sixma,T.K. and Mattevi,A. (2016) Touch, act and go: landing and operating on nucleosomes. *EMBO J.*, **35**, 376–388.
23. Klug,A. and Lutter,L.C. (1981) The helical periodicity of DNA on the nucleosome. *Nucleic Acids Res.*, **9**, 4267–4283.
24. Simpson,R.T. and Stafford,D.W. (1983) Structural features of a phased nucleosome core particle. *Proc. Natl. Acad. Sci. U. S. A.*, **80**, 51–55.
25. Zhong,J., Luo,K., Winter,P.S., Crawford,G.E., Iversen,E.S. and Hartemink,A.J. (2016) Mapping nucleosome positions using DNase-seq. *Genome Res.*, 10.1101/gr.195602.115.
26. Lahm,A. and Suck,D. (1991) DNase I-induced DNA conformation. 2 A structure of a DNase I-octamer complex. *J. Mol. Biol.*, **222**, 645–667.
27. Weston,S.A., Lahm,A. and Suck,D. (1992) X-ray structure of the DNase I-d(GGTATAACC)<sub>2</sub> complex at 2.3 Å resolution. *J. Mol. Biol.*, **226**, 1237–1256.
28. Djuranovic,D. and Hartmann,B. (2005) Molecular dynamics studies on free and bound targets of the bovine papillomavirus type I e2 protein: the protein binding effect on DNA and the recognition mechanism. *Biophys. J.*, **89**, 2542–2551.
29. Djuranovic,D., Oguey,C. and Hartmann,B. (2004) The role of DNA structure and dynamics in the recognition of bovine papillomavirus E2 protein target sequences. *J. Mol. Biol.*, **339**, 785–796.
30. Heddi,B., Foloppe,N., Oguey,C. and Hartmann,B. (2008) Importance of accurate DNA structures in solution: the Jun-Fos model. *J. Mol. Biol.*, **382**, 956–970.
31. Tisné,C., Delepierre,M. and Hartmann,B. (1999) How NF-kappaB can be attracted by its cognate DNA. *J. Mol. Biol.*, **293**, 139–150.
32. Azad,R.N., Zafiroopoulos,D., Ober,D., Jiang,Y., Chiu,T.-P., Sagendorf,J.M., Rohs,R. and Tullius,T.D. (2018) Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Res.*, **46**, 2636–2647.
33. Heddi,B., Abi-Ghanem,J., Lavigne,M. and Hartmann,B. (2010) Sequence-dependent DNA flexibility mediates DNase I cleavage. *J. Mol. Biol.*, **395**, 123–133.
34. Parvin,J.D., McCormick,R.J., Sharp,P.A. and Fisher,D.E. (1995) Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, **373**, 724–727.
35. Koudelka,G.B. and Carlson,P. (1992) DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature*, **355**, 89–91.
36. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307–318.
37. Wecker,K., Bonnet,M.C., Meurs,E.F. and Delepierre,M. (2002) The role of the phosphorus BI-BII transition in protein-DNA recognition: the NF-kappaB complex. *Nucleic Acids Res.*, **30**, 4452–4459.
38. Hartmann,B., Piazzola,D. and Lavery,R. (1993) BI-BII transitions in B-DNA. *Nucleic Acids Res.*, **21**, 561–568.
39. Djuranovic,D. and Hartmann,B. (2003) Conformational characteristics and correlations in crystal structures of nucleic acid oligonucleotides: evidence for sub-states. *J. Biomol. Struct. Dyn.*, **20**, 771–788.
40. Djuranovic,D. and Hartmann,B. (2004) DNA fine structure and dynamics in crystals and in solution: the impact of BI/BII backbone conformations. *Biopolymers*, **73**, 356–368.
41. Heddi,B., Oguey,C., Lavelle,C., Foloppe,N. and Hartmann,B. (2010) Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.*, **38**, 1034–1047.
42. Imeddourene,A.B., Xu,X., Zargarian,L., Oguey,C., Foloppe,N., Mauffret,O. and Hartmann,B. (2016) The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res.*, **44**, 3432–3447.
43. Heddi,B., Foloppe,N., Bouchemal,N., Hantz,E. and Hartmann,B. (2006) Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.*, **128**, 9170–9177.
44. Dršata,T., Pérez,A., Orozco,M., Morozov,A.V., Sponer,J. and Lankaš,F. (2013) Structure, Stiffness and Substates of the Dickerson-Drew Dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
45. Oguey,C., Foloppe,N. and Hartmann,B. (2010) Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions. *PLoS ONE*, **5**.

46. Xu,X., Ben Imeddourene,A., Zargarian,L., Foloppe,N., Mauffret,O. and Hartmann,B. (2014) NMR studies of DNA support the role of pre-existing minor groove variations in nucleosome indirect readout. *Biochemistry*, **53**, 5601–5612.
47. Elbahnsi,A., Retureau,R., Baaden,M., Hartmann,B. and Oguey,C. (2018) Holding the Nucleosome Together: A Quantitative Description of the DNA-Histone Interface in Solution. *J. Chem. Theory Comput.*, **14**, 1045–1058.
48. Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
49. Chua,E.Y.D., Vasudevan,D., Davey,G.E., Wu,B. and Davey,C.A. (2012) The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Res.*, **40**, 6338–6352.
50. Marathe,A. and Bansal,M. (2011) An ensemble of B-DNA dinucleotide geometries lead to characteristic nucleosomal DNA structure and provide plasticity required for gene expression. *BMC Struct. Biol.*, **11**, 1.
51. Dlakić,M., Ussery,D.W. and Brunak,S. (2005) DNA Bendability and Nucleosome Positioning in Transcriptional Regulation. In Ohyama,T. (ed), *DNA Conformation and Transcription*, Molecular Biology Intelligence Unit. Springer US, Boston, MA, pp. 189–202.
52. Roccatano,D., Barthel,A. and Zacharias,M. (2007) Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation. *Biopolymers*, **85**, 407–421.
53. Harp,J.M., Hanson,B.L., Timm,D.E. and Bunick,G.J. (2000) Asymmetries in the nucleosome core particle at 2.5 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 1513–1534.
54. Tsunaka,Y., Kajimura,N., Tate,S. and Morikawa,K. (2005) Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res.*, **33**, 3424–3434.
55. Thåström,A., Bingham,L.M. and Widom,J. (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.*, **338**, 695–709.
56. Hart,K., Foloppe,N., Baker,C.M., Denning,E.J., Nilsson,L. and Mackerell,A.D. (2012) Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.*, **8**, 348–362.
57. Esque,J., Leonard,S., de Brevern,A.G. and Oguey,C. (2013) VLDP web server: a powerful geometric tool for analysing protein structures in their environment. *Nucleic Acids Res.*, **41**, W373–W378.
58. Retureau,R., Oguey,C., Mauffret,O. and Hartmann,B. (2019) Structural explorations of NCP7-nucleic acid complexes give keys to decipher the binding process. *J. Mol. Biol.*, 10.1016/j.jmb.2019.03.002.
59. Ben Imeddourene,A., Elbahnsi,A., Guérout,M., Oguey,C., Foloppe,N. and Hartmann,B. (2015) Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics. *PLoS Comput. Biol.*, **11**, e1004631.
60. Makde,R.D., England,J.R., Yennawar,H.P. and Tan,S. (2010) Structure of RCC1 chromatin factor bound to the nucleosome core particle. *Nature*, **467**, 562–566.
61. Davey,C.A., Sargent,D.F., Luger,K., Maeder,A.W. and Richmond,T.J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
62. Park,Y.-J., Dyer,P.N., Tremethick,D.J. and Luger,K. (2004) A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. *J. Biol. Chem.*, **279**, 24274–24282.
63. Morales,V. and Richard-Foy,H. (2000) Role of histone N-terminal tails and their acetylation in nucleosome dynamics. *Mol. Cell. Biol.*, **20**, 7230–7237.
64. Krieger,E., Nielsen,J.E., Spronk,C.A.E.M. and Vriend,G. (2006) Fast empirical pKa prediction by Ewald summation. *J. Mol. Graph. Model.*, **25**, 481–486.
65. Brooks,B.R., Brooks,C.L., Mackerell,A.D., Nilsson,L., Petrella,R.J., Roux,B., Won,Y., Archontis,G., Bartels,C., Boresch,S., *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.
66. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kale,L. and Schulten,K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
67. Mackerell,A.D. (2004) Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.*, **25**, 1584–1604.
68. Beglov,D. and Roux,B. (1994) Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations. *J. Chem. Phys.*, **100**, 9050–9063.
69. Darden,T., York,D. and Pedersen,L. (1993) Particle Mesh Ewald - an N.log(n) Method for Ewald Sums in Large Systems. *J. Chem. Phys.*, **98**, 10089–10092.

70. van Gunsteren, W.F. and Berendsen, H.J.C. (1977) Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.*, **34**, 1311–1327.
71. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J.C. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem.*, **26**, 1701–1718.
72. McDonald, I.K. and Thornton, J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
73. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
74. Lu, X.-J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
75. Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.J., Neidle, S., Shakked, Z., *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
76. Lindahl, V., Villa, A. and Hess, B. (2017) Sequence dependency of canonical base pair opening in the DNA double helix. *PLOS Comput. Biol.*, **13**, e1005463.
77. Huang, N., Banavali, N.K. and MacKerell, A.D. (2003) Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase. *Proc. Natl. Acad. Sci.*, **100**, 68–73.
78. Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., *et al.* (2010) Relationship between nucleosome positioning and DNA methylation. *Nature*, **466**, 388–392.
79. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J. and Hatcher, M.E. (2009) <sup>31</sup>P NMR Investigation of Backbone Dynamics in DNA Binding Sites. *J. Phys. Chem. B*, **113**, 2596–2603.
80. Svozil, D., Kalina, J., Omelka, M. and Schneider, B. (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.
81. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 11163–11168.
82. Hatakeyama, A., Hartmann, B., Travers, A., Nogues, C. and Buckle, M. (2016) High-resolution biophysical analysis of the dynamics of nucleosome formation. *Sci. Rep.*, **6**, 27337.

# **A dynamic view of DNA structure within the nucleosome: biological implications**

Romain Retureau, Nicolas Foloppe, Ahmad Elbahns, Christophe Oguey and Brigitte Hartmann

## **Figures**

Figure S1: NCP models and the four free DNA dodecamers with respect to the 5' half of DNA<sub>NCP</sub>.

Figure S2: Root Mean Square Deviation (RMSD) of DNA<sub>NCP</sub> during the trajectories.

Figure S3: RMSFs of nucleotides in DNA<sub>NCP</sub>.

Figure S4: BII percentages in DNA<sub>NCP</sub> during the trajectories.

Figure S5: Distributions and combinations of BI and BII conformations in DNA<sub>NCP</sub>.

Figure S6: BII conformers and DNA-histone contacts in DNA<sub>NCP</sub>.

Figure S7: Inter base pair parameters in DNA<sub>NCP</sub>.

Figure S8: Inter base pair parameters in free and bound 601-dodecamers.

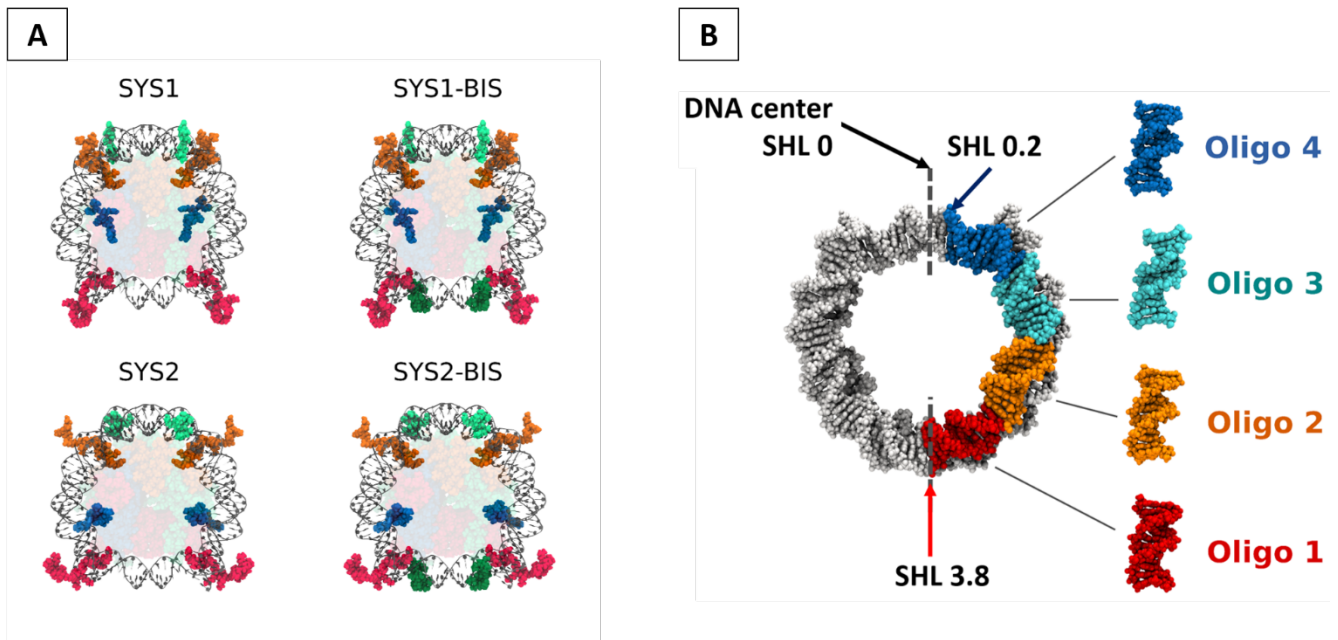
Figure S9: Variability of inter base pair parameters in free and bound 601-dodecamers.

## **Tables**

Table S1: Histone N-tails composition of the models.

Table S2: DNA sequences.

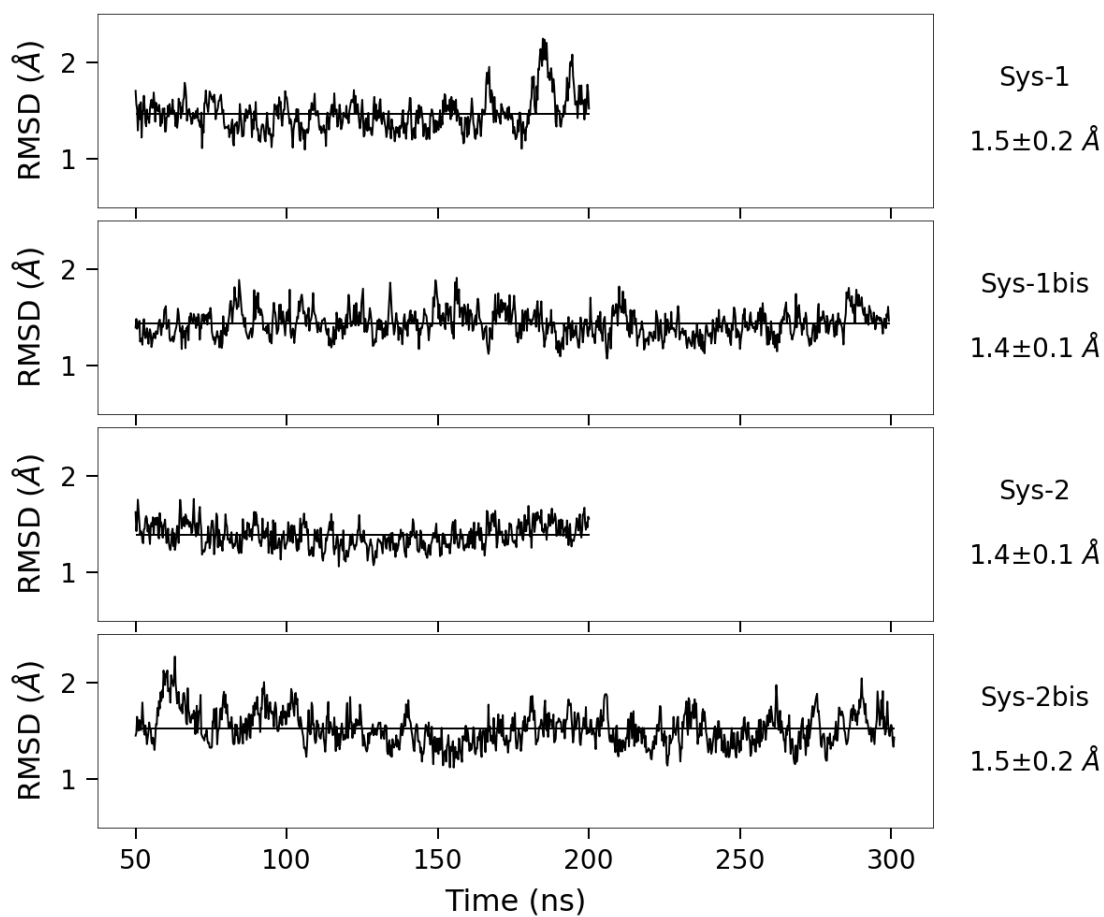
## FIGURES



**Figure S1: Studied NCP models and the four free DNA dodecamers with their copies in the 5' half of  $\text{DNA}_{\text{NCP}}$ .**

**A:** molecular representations of the four models used as starting point of simulations;  $\text{DNA}_{\text{NCP}}$  is in grey; the histone structured domains are transparent for clarity; the histones tails are represented at molecular level for the N-ter domains of H3 (red), H4 (blue), H2A (green) and H2B (orange) as well as for C-ter H2A (dark green).

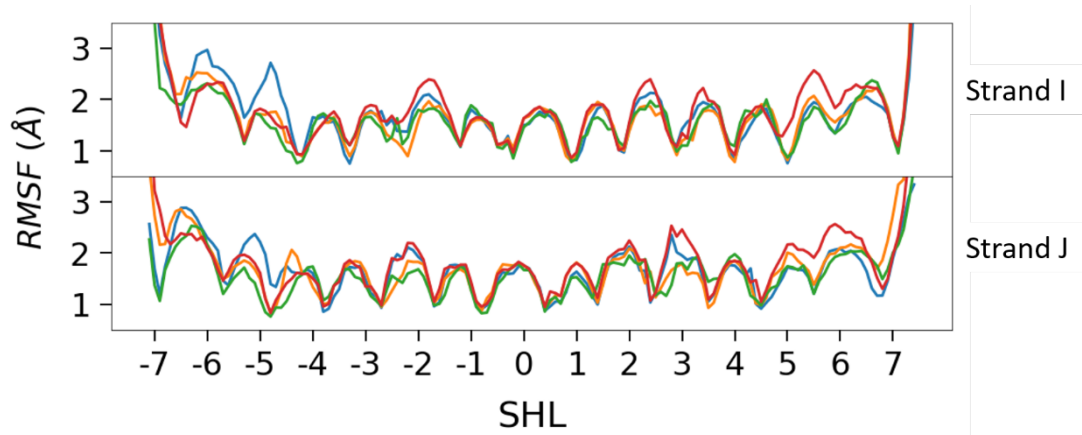
**B:** The molecular model at the left is a top view of the structure of the 601 sequence 5' half (in gray) within the nucleosome; the histones were removed from the view for clarity; the colored regions correspond to the location of the free dodecamers studied by NMR, also shown at the right of the figure.



**Figure S2: Root Mean Square Deviation (RMSD) of DNA<sub>NCP</sub> during the trajectories.**

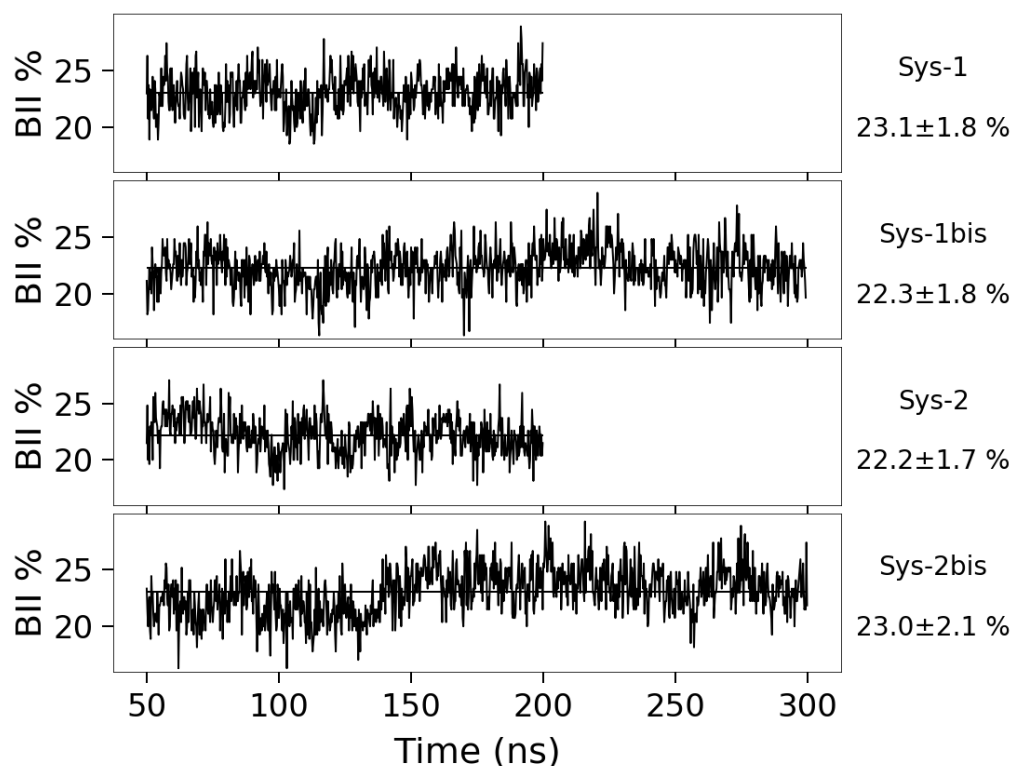
The RMSDs were calculated on the heavy atoms of the DNA<sub>NCP</sub>, excluding three base pairs at each extremity. The references were the time-average structures from each simulation, excluding the first 50 ns. The horizontal lines correspond to the average RMSD values, also reported on the right of the panels, below the simulation names.





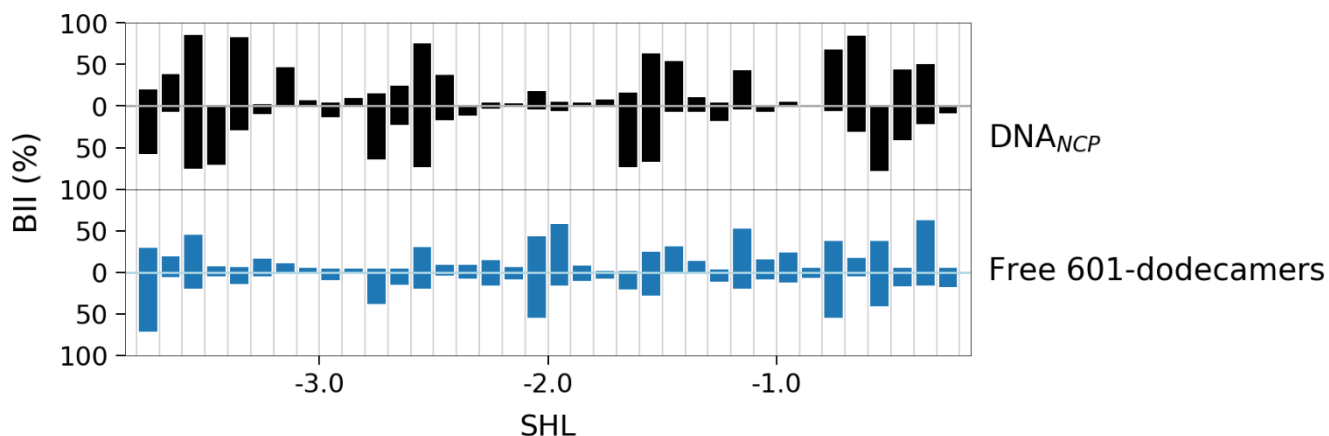
**Figure S3: RMSFs of nucleotides in DNA<sub>NCP</sub>.**

The atomic position Root Mean Square Fluctuations (RMSFs) were calculated for the nucleotides along the strands I and J of DNA<sub>NCP</sub>, excluding the three base pairs at each extremity. The different colors represent the four simulations analyzed here: SYS1 in blue, SYS1BIS in red, SYS2 in green and SYS2BIS in orange.



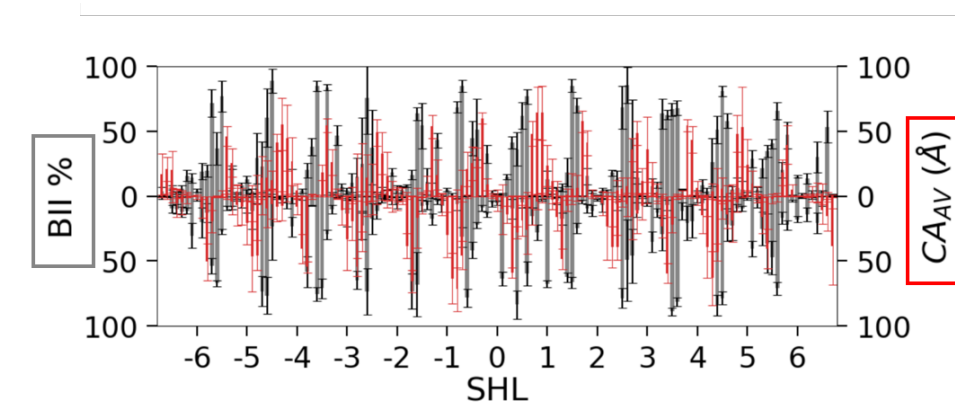
**Figure S4: BII percentages in DNA<sub>NCP</sub> during the trajectories.**

The BII percentages (BII %) were calculated on all the phosphodiester junctions of DNA<sub>NCP</sub>, excluding the two junctions (involving three base pairs) at each extremity. BI and BII were defined by  $(\epsilon-\zeta) < 30^\circ$  and  $(\epsilon-\zeta) > 30^\circ$ , respectively. The horizontal lines correspond to the average BII% values, also reported with standard deviations on the right of the panels, below the simulation names.



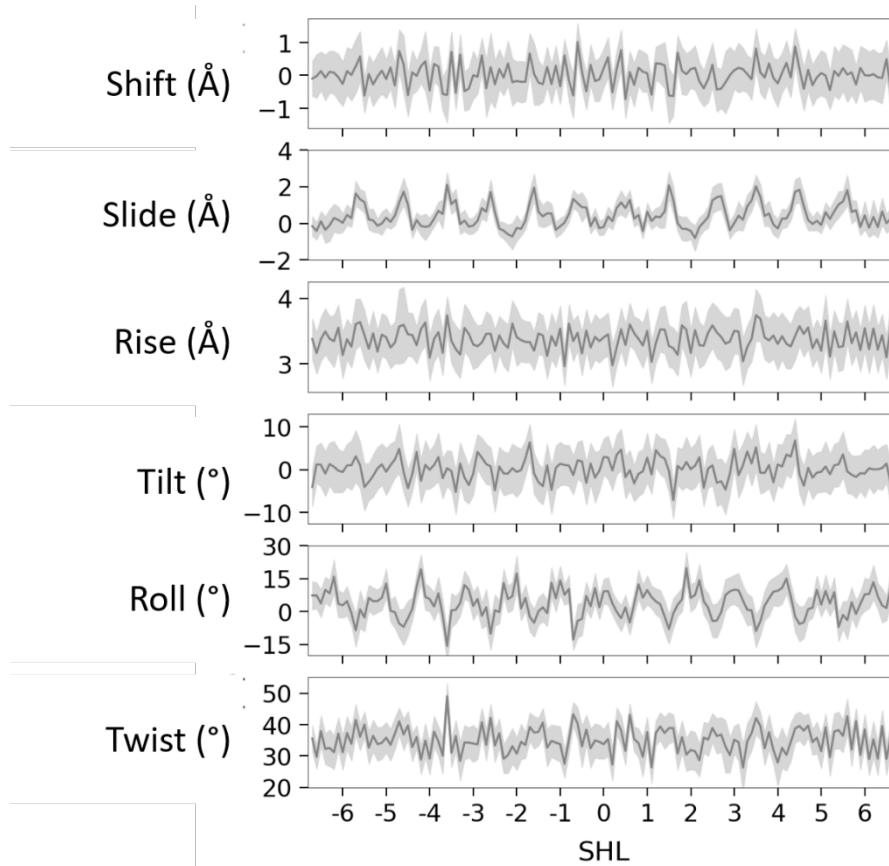
**Figure S5: Distributions and combinations of BI and BII conformations in  $\text{DNA}_{\text{NCP}}$ .**

BII percentages (BII %) of each phosphate group  $p_i$  and  $p_j$ , along the strands I ( $y$  axis oriented upwards) and J ( $y$  axis downwards) of the part  $\text{DNA}_{\text{NCP}}$  (top panel, black bars) and their free counterparts (bottom panel, blue bars).



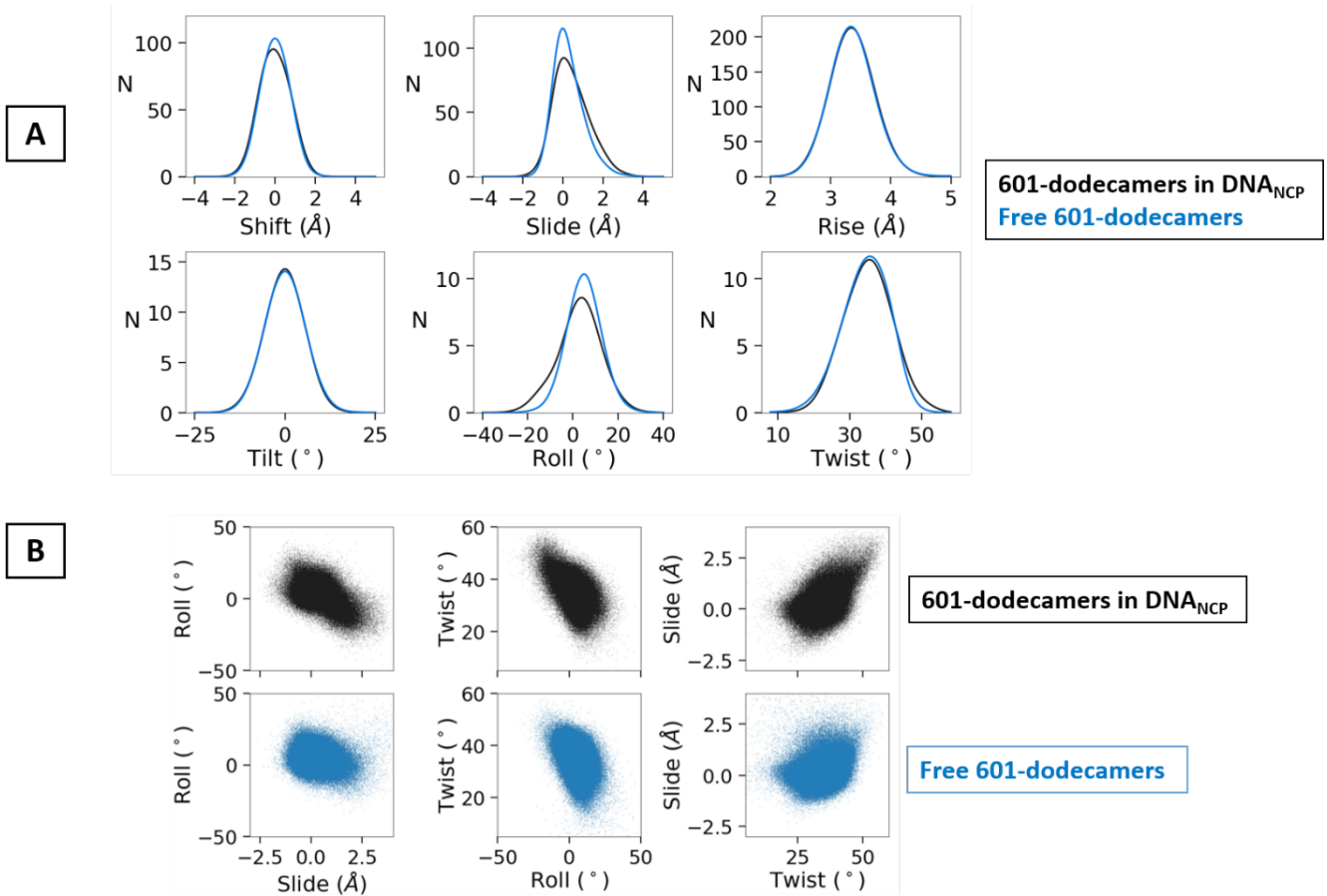
**Figure S6: BII conformers and DNA-histone contacts in  $\text{DNA}_{\text{NCP}}$ .**

Profiles of BII percentages (BII %, black) and contact area (CA, red) calculated between the dinucleotide steps  $\text{N}_p\text{N}$  and the histone structured cores along the strands I ( $y$  axis oriented upwards) and J ( $y$  axis downwards) of the 601 sequence. The BII percentages were calculated for each phosphodiester junctions of  $\text{DNA}_{\text{NCP}}$ .



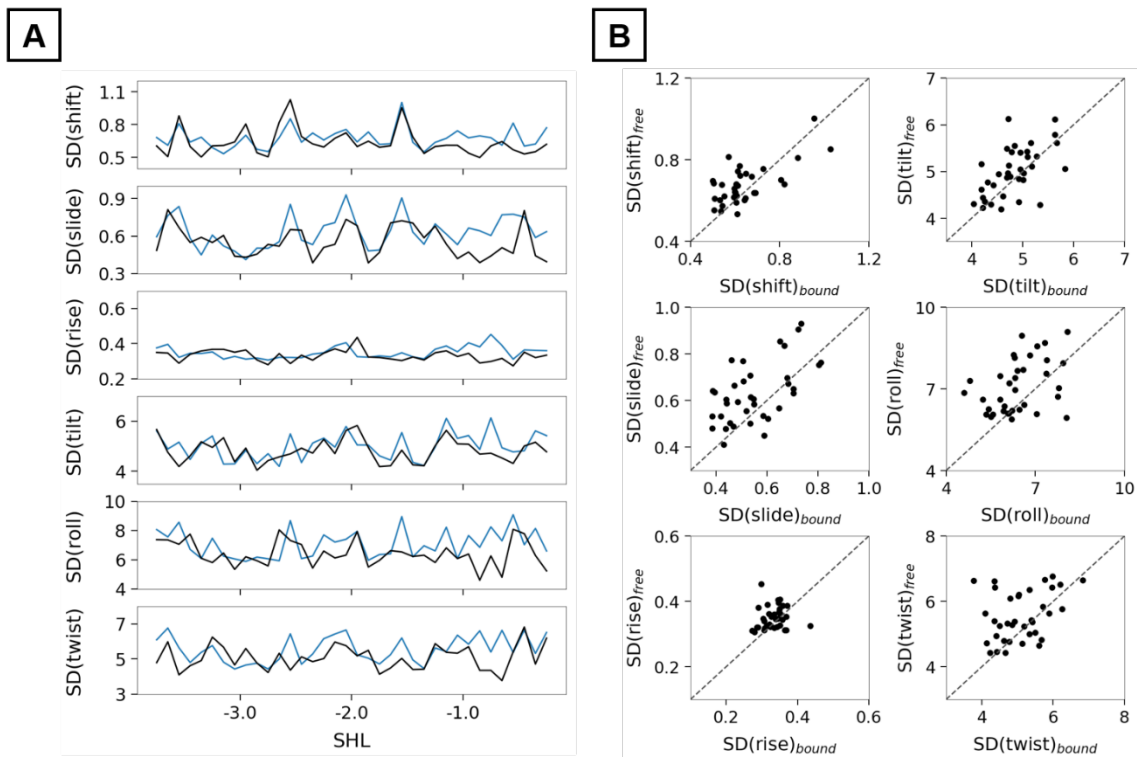
**Figure S7: Inter base pair parameters in DNA<sub>NCP</sub>.**

Profiles of average values of the six inter base pair parameters (shift, slide, rise, tilt, roll and twist) along DNA<sub>NCP</sub>; the grey areas represent the standard deviations. Roll and slide profiles are also shown in Figure 6-A in the main text.



**Figure S8: Inter base pair parameters in free and bound 601-dodecamers.**

**A:** Distribution of the inter base pair parameters in the bound (black) and free (blue) 601-dodecamers. Each distribution is calculated dividing the range of the considered parameter into 200 slices; this is the reason why the y-axis ranges are variable. *N* is the occurrence. **B:** Couplings between the helical parameters of slide, roll and twist in the bound (black) and free (blue) 601-dodecamers. The correlation coefficients are from 0.2 (roll vs slide) to 0.43 (twist vs roll) for free steps and from 0.49 (slide vs twist) to 0.54 (roll vs slide) for bound steps.



**Figure S9: Variability of inter base pair parameters in free and bound 601-dodecamers.**

**A:** Profiles of standard deviations of the six inter base pair parameters in free ( $SD_{free}$ , in blue) and bound ( $SD_{bound}$ , in black) 601-dodecamers. **B:** Comparisons between the standard deviations of the six inter base pair parameters in free ( $SD_{free}$ ) and bound ( $SD_{bound}$ ) 601-dodecamers;  $x=y$  diagonals are represented by dashed lines.

## TABLES

**Table S1: Histone N-tails composition of the models.**

The composition of the N- terminal tails is given for the four histone types in the four models studied here. These models contain added amino acids taken from the specified chains of 1KX5.

|            | SYS1 and SYS1-bis    | SYS2 and SYS2-bis    |
|------------|----------------------|----------------------|
| H3 N tail  | 24 → 44 from chain A | 24 → 44 from chain E |
| H4 N tail  | 17 → 24 from chain B | 17 → 24 from chain F |
| H2A N tail | 9 → 16 from chain C  | 9 → 16 from chain G  |
| H2B N tail | 19 → 34 from chain D | 19 → 34 from chain H |

**Table S2: DNA sequences.**

The first row gives the 5' → 3' composition of the two 146 bp strands of the 601 sequence; the centers are indicated by a star. The next rows contain the sequences of four “601 dodecamers” previously studied free in solution by NMR and modeling. Up to three-bases overlaps at the ends, juxtaposing the four dodecamers reproduces the continuous 39 bp fragment of sequence 601 highlighted in yellow.

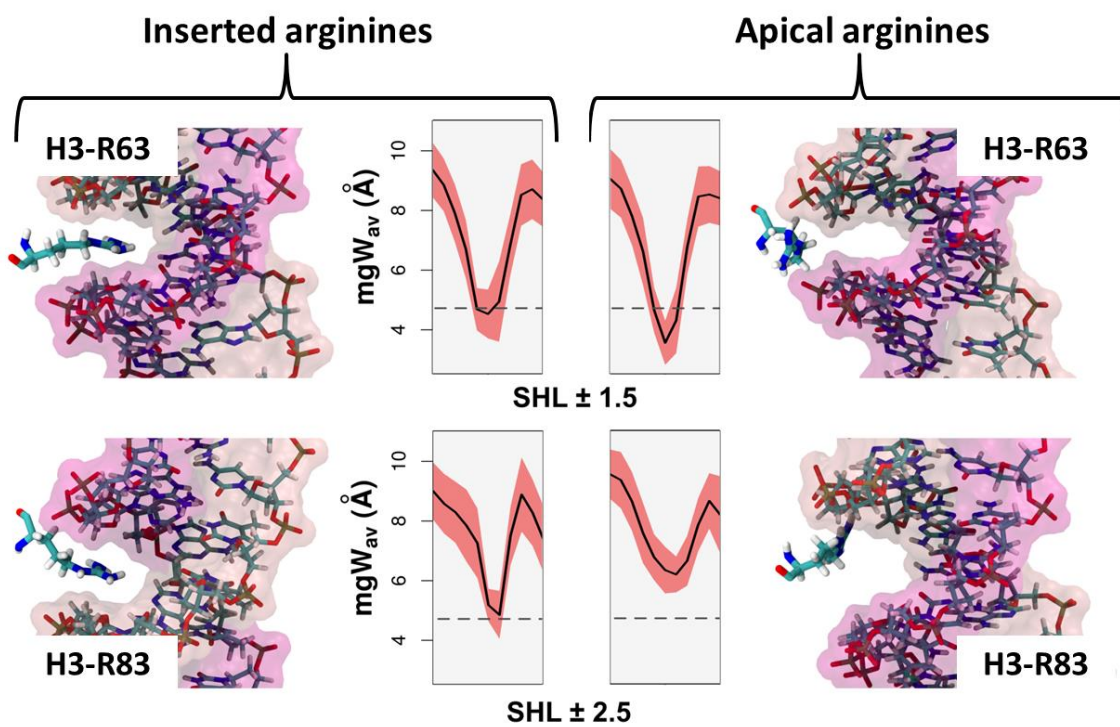
|                 |  |
|-----------------|--|
| Sequence<br>601 | Strand 1: 5'-TCGAGAATCCCGGTGCCGAGGCCGCTCAATGG<br><b>TCGTAGACAGCTCTAGCACCGCTTAAACGCTCGTACGCG</b> C*T<br>GTCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTC<br>TCCAGGCACGTGTCAGATATATACATCCGAT-3'<br>Strand 2: 5'-ATCGGATGTATATATCTGACACGTGCCTGGA<br>GACTAGGGAGTAATCCCCTTGGCGGTAAACGCGGGGGAC<br>A*G <b>CGCGTACGTGCGTTTTAAGCGGTGCTAGAGCTGTCTACGA</b><br>CCAATTGAGCGGCCTCGGCACCGGATTCTCGA-3' |
| Oligo 1         | 5'-TCGTAGCAAGCT-3'•5'-AGCTTGCTACGA-3'  |
| Oligo 2         | 5'-GCTCTAGCACCG-3'•5'-CGGTGCTAGAGC-3'  |
| Oligo 3         | 5'-CCGCTTAAACGC-3'•5'-GCGTTTTAAGCGG-3'   |
| Oligo 4         | 5'-CGCACGTACGCG-3'•5'-CGCGTACGTGCG-3'  |

## Conclusion et développements en cours

Il semble que l'ADN nucléosomal, bien que très contraint par des contacts denses avec les histones, présente une dynamique qui n'est pas très différente de celle de l'ADN libre. En effet, les distributions des valeurs associées à l'équilibre BI↔BII ou aux paramètres hélicoïdaux locaux sont très similaires dans l'ADN nucléosomal et l'ADN libre. Dans les deux cas, les régions riches en BII sont associées avec des *roll* négatifs et des *slide* positifs, alors que les régions pauvres en BII correspondent à des *roll* positifs et des *slide* négatifs.

De plus, bien que l'effet de la séquence dinucléotidique sur la propension à adopter les conformères BII soit moins marqué dans l'ADN nucléosomal que dans l'ADN libre, on observe tout de même une tendance suffisante pour renforcer l'idée que les propriétés intrinsèques de l'ADN jouent un rôle dans la capacité des ADN à former des nucléosomes.

Afin d'aller plus loin dans cette analyse, je travaille actuellement sur un autre aspect de l'ADN nucléosomal, moins local : les dimensions des sillons. Comme je l'ai indiqué pour le premier article (p. 97), j'étudie actuellement l'insertion des arginines dans le petit sillon. Dans nos dynamiques, les arginines H3-R83 et H3-R63 s'insèrent respectivement 36 et 72% du temps dans le petit sillon. Ces arginines explorent donc deux états, insérée et apical (non-insérée) par rapport au petit sillon. L'étude de la largeur de ces petits sillons en fonction de ces états (Figure 40) semble mettre en évidence qu'une taille de sillon bien précise ( $\sim 4.5\text{\AA}$ ) est requise pour optimiser l'interaction avec une arginine. Lorsque l'arginine se met en position apicale, le petit sillon reprend ses dimensions « naturelles », compte tenu évidemment de la contrainte de la courbure de l'ADN au sein du nucléosome.



*Figure 40 – Largeur moyenne du petit sillon ( $mgW_{av}$ ) pour les conformations insérées (gauche) ou non-insérées (droite) des arginines H3-R63 et H3-R83. Le panel de gauche présente une structure des arginines R83 et R63 insérées et la largeur du petit sillon ( $mgW$ ) associée moyennée sur l'ensemble des structures où ces arginines sont insérées. Le panel de droite représente les mêmes arginines et les mêmes largeurs de petit sillon pour des arginines apicales. Dans les 4 encarts centraux, l'écart-type de la largeur du petit sillon est visualisé par une zone rouge autour de la valeur moyenne.*



# Chapitre 4 : Une autre interface AN/Protéine, les complexes NCp7

## Sommaire<sup>4</sup>

La protéine 7 de la nucléocapside (NCp7) du VIH de type 1 (VIH-1) est une forme mature issue du clivage de son précurseur, la polyprotéine Gag. Il s'agit d'une petite protéine de 55 résidus, composée de deux doigts de zinc très structurés (ZF1 et ZF2) séparés par un séparateur (*linker*) flexible de 7 résidus. Elle intervient dans de nombreuses étapes du cycle du VIH-1 impliquant les acides nucléiques qu'elle lie grâce aux doigts de zinc. Cette protéine est capable de se fixer de manière non-spécifique à des AN très diverses dans leur type (ARN ou ADN), leur séquence ou encore leur repliement (simple brin, tige-boucle, *bulge*). Elle joue notamment un rôle chaperonne de certains AN en déstabilisant leur structure secondaire.

En raison de la flexibilité du *linker*, l'étude expérimentale du repliement de la NCp7 libre et du paysage conformationnel qu'elle explore a présenté un challenge important. Ce n'est que récemment que la combinaison des méthodes expérimentales SAXS et RMN a permis de mieux définir un ensemble représentatif des modes de repliements adoptés par la NCp7 libre. Concernant la NCp7 complexée avec des AN, il existe cinq études de RMN associée à des structures disponibles dans la PDB, deux avec de l'ARN fixé à une NCp7 provenant de la souche VIH-1 MB et trois avec de l'ADN fixé à une NCp7 extraite de la souche NL4-3. Chaque système comporte plusieurs modèles, les cinq systèmes représentant un ensemble de 84 modèles. Les NCp7 des deux souches ont trois résidus différents dans leur séquence (Article 4 – Figure 1).

C'est en nous basant sur ces structures expérimentales que nous nous sommes proposés de comprendre le mécanisme de la fixation de la NCp7 sur les AN. Notamment, nous voulions mettre ici en évidence les points communs et les divergences qui caractérisent le repliement de la NCp7 et l'interface NCp7/AN dans ces complexes très différents.

Dans un premier temps, nous avons montré que l'orientation de l'ensemble ZF1-ZF2 (les deux doigts de zinc ZF1 et ZF2 sans le *linker* et les parties N-ter et C-ter) dans les différents complexes est très conservée. Le mode de fixation global de la NCp7 semble donc ne pas crucialement dépendre du type de l'AN, de son repliement, de sa séquence ou encore de la séquence de la NCp7.

L'étude de l'interface des différents systèmes NCp7/AN avec VLDM a permis de mettre en évidence des schémas d'interaction différents pour ZF1 et ZF2. Pour ZF2, l'interaction est la même dans tous

---

<sup>4</sup> L'ensembles des références de cette section peuvent-être trouvées dans l'article associé au chapitre (p. 177) et dans l'introduction de ce manuscrit (Les complexes acides nucléique-NCp7, p. 40)

les systèmes : une poche protéique se forme autour d'une guanine, avec en particulier un excellent empilement avec le résidu aromatique TRP37. Concernant ZF1, on observe deux schémas d'interaction : avec les ARN, les résidus PHE16 et THR24 forment une pince autour d'une guanine, similaire à celle observée dans ZF2 ; avec les ADN, la PHE16 contacte une pyrimidine et l'empilement est moins important. Cette dualité d'interaction peut être expliquée par la séquence des AN, le type des AN ou la séquence de la NCp7 qui selon la souche présente soit une isoleucine soit une thréonine en position 24.

Enfin, l'analyse des modèles de la NCp7 libre a permis de définir deux superfamilles de conformation, dont l'une, la NCp7 dite ouverte, est caractérisée par la forte accessibilité du résidu TRP37, propice à la reconnaissance d'une guanine libre. L'autre famille, la NCp7 dite fermée, ressemble beaucoup à la forme de la NCp7 complexée.



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

Romain Retureau<sup>1</sup>, Christophe Oguey<sup>2</sup>, Olivier Mauffret<sup>1</sup> and Brigitte Hartmann<sup>1</sup>

<sup>1</sup> - LBPA, UMR 8113, ENS Paris-Saclay—CNRS, 61 avenue du Président Wilson, 94235 Cachan cedex, France

<sup>2</sup> - LPTM, CNRS UMR 8089, Université de Cergy-Pontoise, 2 avenue Adolphe Chauvin, 95031 Cergy-Pontoise, France

**Correspondence to Olivier Mauffret and Brigitte Hartmann:** [olivier.mauffret@ens-paris-saclay.fr](mailto:olivier.mauffret@ens-paris-saclay.fr), [bhartman@ens-paris-saclay.fr](mailto:bhartman@ens-paris-saclay.fr)

<https://doi.org/10.1016/j.jmb.2019.03.002>

Edited by John Johnson

## 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 (NAs), 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.

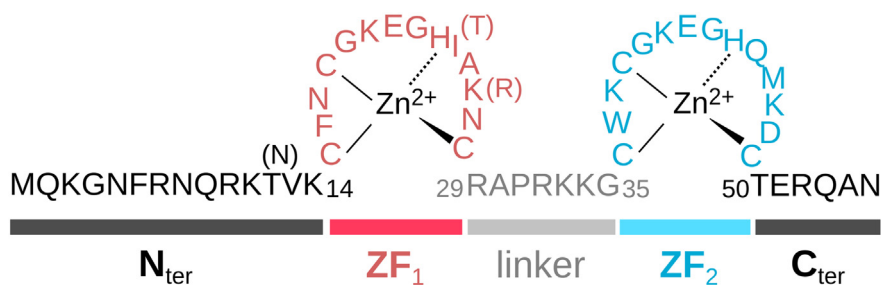
© 2019 Elsevier Ltd. All rights reserved.

## 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 seven residues as earlier shown by a structural study [1]. Owing to their NC domain these proteins form complexes with nucleic acids (NAs) and mediate many stages of the HIV-1 cycle, from viral genome recognition to RNA packaging [2–10]. They are able to bind NAs 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 NAs.

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 NAs) 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



**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 (gray) 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).

that, taken together, fit at best the whole experimental data sets. 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].

Three-dimensional (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 with 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 four hydrogen bonds; in addition, compared to any pyrimidine, a guanine can maximize the stacking with aromatic residues, PHE 16 in ZF1 and TRP 37 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 consists 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 Voronoi Laguerre Delaunay for Macromolecules (VLDM) [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 (CA) 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–NA complexes belonging to five different systems (Table 1) were analyzed. Before reporting the results of our analyses, we need to make a few comments regarding both the proteins and NAs in these systems.

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 an 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 1A1T, 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 (1A1T) 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, that is, 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 an 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, that is, 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 data set by calculating the cross-RMSDs on

**Table 1.** Summary of the studied NCp7–NA complexes

| PDB code | NCp7  | NA   | $N_{\text{models}}$ |
|----------|---|--|---------------------|
| 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 (stem–pentaloop)       | 19                  |
| 2L4L     | 45 residues: truncation of the first 10 N-terminal residues; M-B strain                             | DNA of 4 nt, 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                  |

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 ( $N_{\text{models}}$ ) in each system is given in the last column, summing up to 84. “nt” stands for nucleotide.



the backbone heavy atoms of either ZF1 (CYS 15 → CYS 28) or ZF2 (CYS 36 → CYS 49). The very low cross-RMSD values,  $0.7 \pm 0.25$  Å for ZF1 and  $0.85 \pm 0.6$  Å for ZF2 on average, confirm 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  $\Phi$ ,  $\Psi$  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  $1.4 \pm 0.3$  Å) associated with 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  $3.8 \pm 0.2$  Å on average for 1A1T *versus* 1F6U) or between NCp7–RNA and NCp7–DNA systems (Supplemental Fig. S1-A, cross-RMSD values of  $3.4 \pm 0.3$  Å on average for 1A1T or 1F6U *versus* 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  $D_{ZnZF1-ZnZF2}$  between the two zinc atoms in ZF1 and ZF2 and pseudo-dihedral angles reflecting the relative orientations of ZF1 and ZF2. The values of  $D_{ZnZF1-ZnZF2}$  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).

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 were then scrutinized through three  $\text{Ca-Zn}_{ZF1}\text{-Zn}_{ZF2}\text{-Ca}$  pseudo-dihedral angles involving the two Zn atoms and two Ca 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  $D_{ZnZF1-ZnZF2}$ , 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 of a total of 19 models show alterations in the ZF1 and ZF2 folding and relative positioning, at least according to the  $\text{Ca}_{\text{CYS28}}\text{-Zn}_{\text{ZF1}}\text{-Zn}_{\text{ZF2}}\text{-Ca}_{\text{CYS36}}$  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,  $\text{Zn}_{\text{ZF1}} \rightarrow \text{Ca}_{\text{LYS20}}$  and  $\text{Zn}_{\text{ZF2}} \rightarrow \text{Ca}_{\text{GLY43}}$  point toward opposite directions conversely to  $\text{Zn}_{\text{ZF1}} \rightarrow \text{Ca}_{\text{CYS15}}$  and  $\text{Zn}_{\text{ZF2}} \rightarrow \text{Ca}_{\text{CYS49}}$  that are directed toward similar directions as well as  $\text{Zn}_{\text{ZF1}} \rightarrow \text{Ca}_{\text{CYS28}}$  and  $\text{Zn}_{\text{ZF2}} \rightarrow \text{Ca}_{\text{CYS36}}$  (Fig. 2).

In summary, each zinc finger, ZF1 or ZF2, behaves as a quasi-rigid body in our data set of NCp7–NA structures. Despite 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.

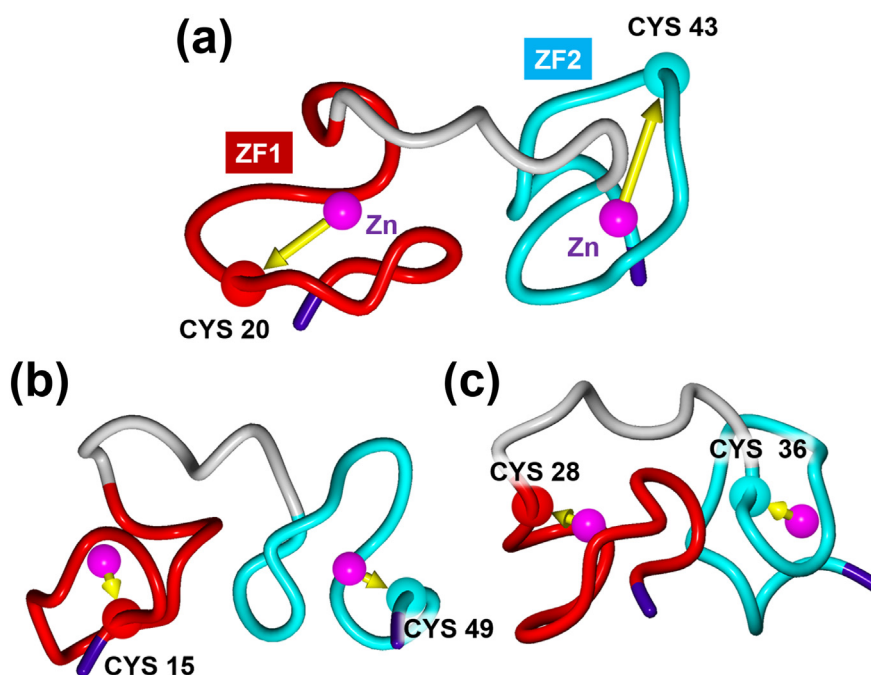
### Analysis of free ZF1–ZF2 ensemble and comparison with their bound counterparts

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

**Table 2.** Structural parameters related to the relative positioning of ZF1 and ZF2 in bound NCp7

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

The spatial positioning of ZF1 and ZF2 with respect to each other was monitored in each system by calculating  $D_{ZnZF1-ZnZF2}$ , 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 Ca 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.



**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 (gray). 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:  $Zn_{ZF1} \rightarrow Ca_{LYS20}/Zn_{ZF2} \rightarrow Ca_{GLY43}$  (a),  $Zn_{ZF1} \rightarrow Ca_{CYS15}/Zn_{ZF2} \rightarrow Ca_{CYS28}$  (b) and  $Zn_{ZF1} \rightarrow Ca_{CYS28}/Zn_{ZF2} \rightarrow Ca_{CYS36}$  (c).

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.

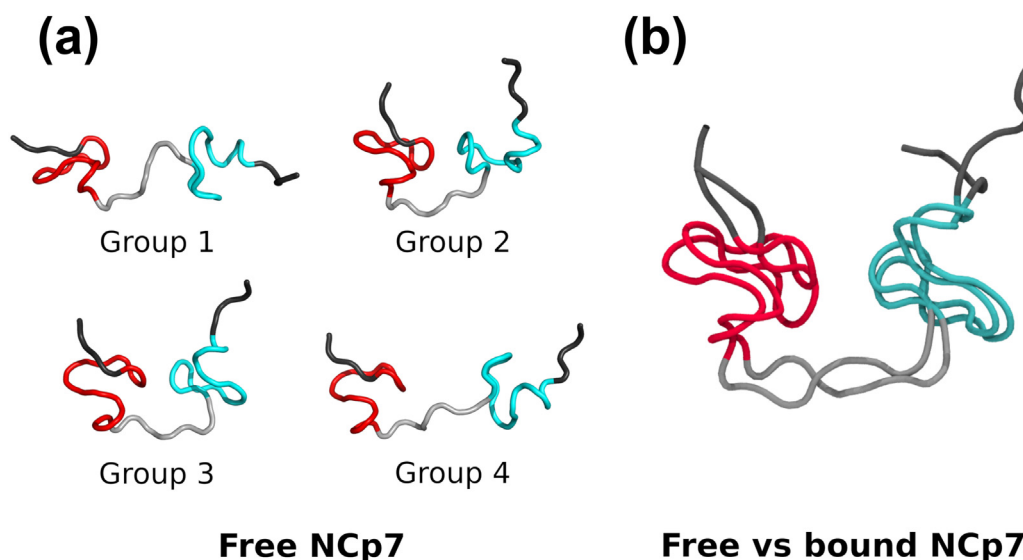
Let us look at the 5I1R models categorized using the  $D_{Zn_{ZF1}-Zn_{ZF2}}$  distance and  $Ca-Zn_{ZF1}-Zn_{ZF2}-Ca$  angles. The existence of four distinct and homoge-

neous groups (Table 3 and Fig. 3a) 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  $Ca-Zn_{ZF1}-Zn_{ZF2}-Ca$  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  $\Phi$ ,  $\Psi$  angles along the linker stresses the flexibility of two linker residues, LYS 34 and GLY 35, as previously described [32]. Indeed, various  $\Phi$ ,  $\Psi$  combinations of LYS 34 and GLY 35 are associated with each type of ZF1 and ZF2 relative positioning, except from  $\Phi_{LYS 34}$  in *g-* in all open models.

**Table 3.** Structural parameters related to the relative positioning of ZF1 and ZF2 in free NCp7

| Group | Models in 5I1R    | $D_{Zn_{ZF1}-Zn_{ZF2}}$ | $(Ca_{15}-Zn_{ZF1}-Zn_{ZF2}-Ca_{49})$ | $(Ca_{20}-Zn_{ZF1}-Zn_{ZF2}-Ca_{43})$ | $(Ca_{28}-Zn_{ZF1}-Zn_{ZF2}-Ca_{36})$ | Free versus bound NCp7: RMSD <sub>av</sub> |
|-------|-------------------|-------------------------|---------------------------------------|---------------------------------------|---------------------------------------|--|
| 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)                                  |

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  $D_{Zn_{ZF1}-Zn_{ZF2}}$  (Å) and three pseudo-dihedral angles (°) are given in terms 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).



**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 gray 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 5l1R; models 1, 2, 3 and 10 are representative of groups 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 5l1R) and bound (model 16 of 2JZW) NCp7; the corresponding RMSD is 2.2 Å.

The structures of group 3, characterized by moderate  $D_{ZnF1-ZnF2}$  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. 3b), 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 CA 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. 4a). 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 (Fig. 4a and 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. 4b). More globally, the residues of ZF2 are more accessible than those of ZF1 in the four conformational groups that we identified (Fig. 4a).

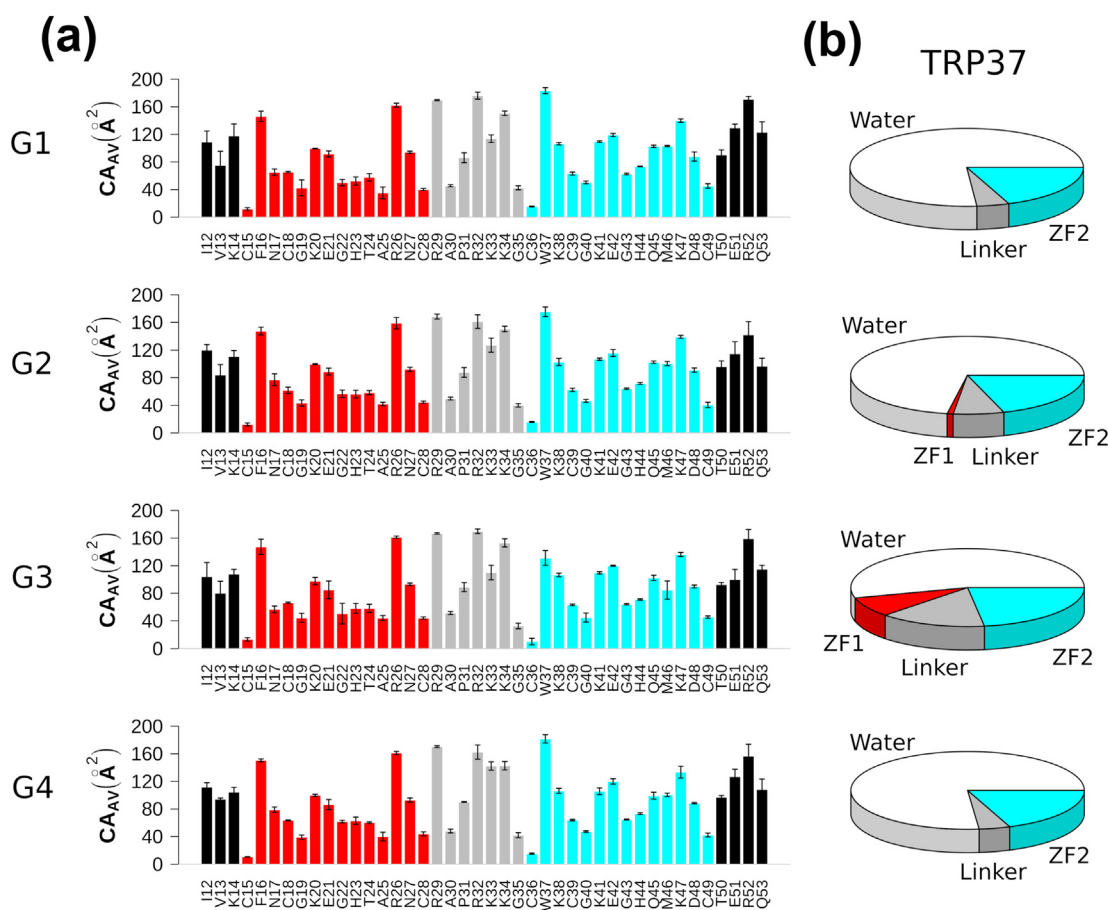
These analyses support the idea that both open and closed conformations sampled by NCp7 in its free state are exploited during the NA binding

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 CAs. Despite the fact that CAs do not measure energies directly, comparison of surfaces of similar electrostatic/non-electrostatic balance does 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. 5a, 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.



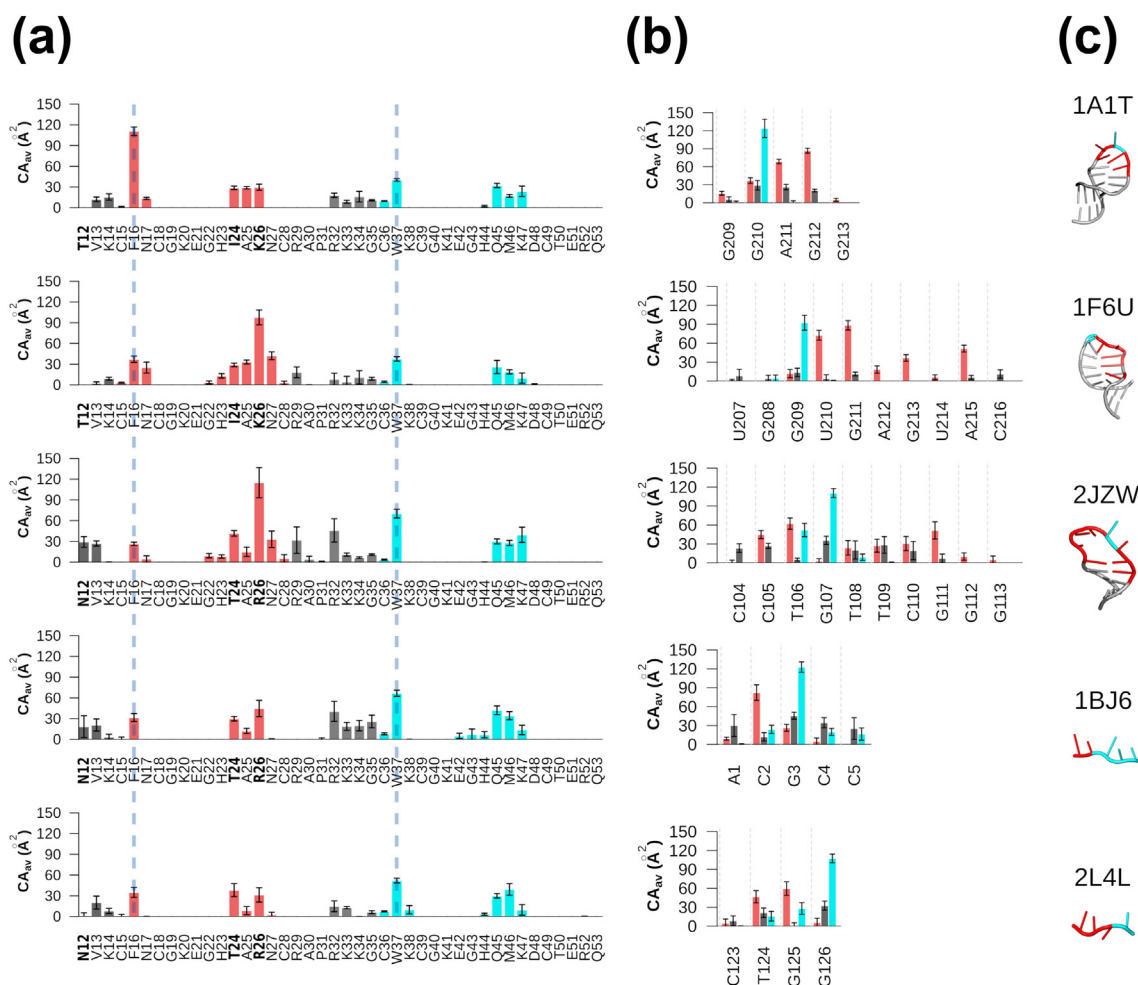


**Fig. 4.** Accessibility to water molecules of amino acids in free NCp7 and the particular case of TRP 37. (a) The contact areas ( $CA_{av}$ ) between amino acids and water molecules are plotted along the free NCp7 sequence for the four structural groups identified in Table 3. The data were averaged over the models that constitute each group; the vertical error bars correspond to standard deviations. (b) These pie charts represent the CAs ( $\text{\AA}^2$ ) between TRP 37 and either amino acids or water molecules in the same four structural groups. In both panels, the data associated with ZF1, linker and ZF2 residues are in red, grey and cyan, respectively.

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 and 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  $64\% \pm 4\%$  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';  $23\% \pm 5\%$  of the total CAs) and phosphate groups (P, O1P, O2P, O5' and O3' atoms;  $13\% \pm 4\%$  of the total CAs) (Supplemental Fig. S4).

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. 5b), extruded from the NA loops in 1A1T, 1F6U and 2JZW and intrinsically accessible in the single-strand DNAs in 1BJ6 and 2L4L. These particular guanines are enclosed in pockets of very



**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 panels a and b, the data associated with N-terminal domain, ZF1, linker and ZF2 residues are in gray, red, gray and cyan, respectively. The data were averaged over the models that constitute each system; the vertical error bars correspond to standard deviations. In panel a, the vertical dashed lines point out the contacts involving PHE 16 and TRP 37, central in the ZF1 and ZF2 interfaces.

**Table 4.** Components of NCp7/NA CAs

| 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)  |

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.

**Table 5.** CAs associated with key guanines in the linker-ZF2/NA interfaces

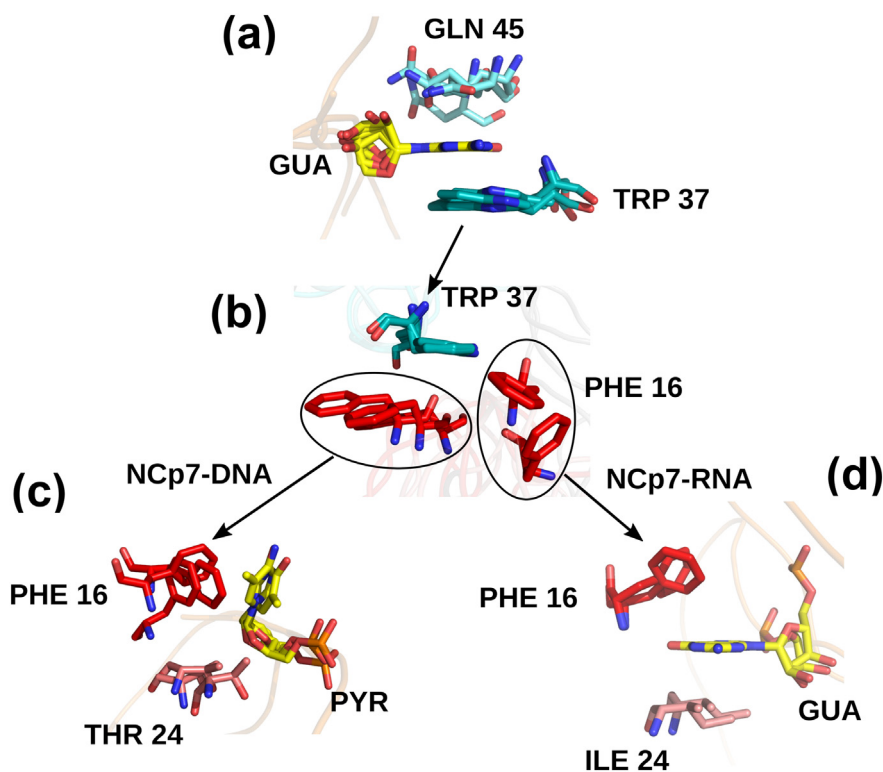
| System | Base | CA(linker/NA) |           |           |           | CA(ZF2/NA) |           |           |           |           |           | Sum     |
|--------|------|---------------|-----------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|---------|
|        |      | ARG<br>32     | LYS<br>33 | LYS<br>34 | GLY<br>35 | CYS<br>36  | TRP<br>37 | HIS<br>44 | GLN<br>45 | MET<br>46 | LYS<br>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) |

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 CA ( $\text{\AA}^2$ ) 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.

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. 6a) 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. 6a). 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].

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



**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/ILE 24 interactions observed in NCp7-RNA models; the superimposition was made on the guanines.

residues, apart from the noticeable case of a recurrent hydrogen bond engaging the backbone of GLY 35. However, 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. 6b). 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,  $11.7 \pm 1.5 \text{ \AA}^2$  on average. The remaining interaction in 1F6U has a very weak CA of  $1.3 \pm 5 \text{ \AA}^2$ . 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. 5b). 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  $CA_{\text{LYS or ARG 26}}$  value (Fig. 5b 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 for proximity interactions that increase the total CAs of the ZF1/NA interfaces ( $CA_{\text{av-ZF1/NA}} = 313 \pm 18 \text{ \AA}^2$  for 1F6U,  $348 \pm 35 \text{ \AA}^2$  for 2JZW;  $242 \pm 10 \text{ \AA}^2$  for 1A1T;  $161 \pm 23 \text{ \AA}^2$  for 1BJ6 and  $147 \pm 20 \text{ \AA}^2$  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 with guanines are significantly lower with ZF1 (Table 6) than with ZF2 (Table 5).

Observing that the electrostatic/non-electrostatic CA ratios are comparable ( $0.25 \pm 0.04$  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. 5b and Table 6). However, the presence of a pyrimidine clearly disfavors the electrostatic component, and the electrostatic/non-electrostatic CA ratio decreases down to 0.06 against 0.25 for a guanine.

In the three NCp7–DNA complexes (Fig. 6c), 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. 6d) in a manner that evokes the ZF2/NA interface (Fig. 6a), 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. 6d and Table 6), comprising CH/ $\pi$  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

**Table 6.** CAs associated with key nucleotides in extended ZF1/NA interfaces

| System   | Base | N-ter domain |           | ZF1       |           |           |           |           |           | Sum    |         |
|----------|------|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------|---------|
|          |      | VAL<br>13    | LYS<br>14 | CYS<br>15 | PHE<br>16 | ILE<br>24 | THR<br>24 | ALA<br>25 | LYS<br>26 |        |         |
| NCp7-RNA | 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) |
|          | 2JZW | T106         | 5 (2)     | No        | No        | 27 (3)    | –         | 13 (2)    | 4 (2)     | 18 (8) | 67 (3)  |
| NCp7-DNA | 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)  |

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 ( $CA_{\text{av}}$ ,  $\text{\AA}^2$ ) calculated and averaged on the models of each system; the last column, “Sum,” gives the total  $CA_{\text{av}}$  of each considered nucleotide. Standard deviations are given in brackets. “No” stands for not observed.



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 with PHE 16/nucleotide/amino acid 24, which show similar values,  $45 \pm 6$  and  $51 \pm 6 \text{ \AA}^2$  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  $73 \pm 8 \text{ \AA}^2$ .

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. 5b) 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 *versus* THR) and 26 (ARG *versus* LYS), (ii) NA nature (RNA *versus* 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 *versus* DNA), the NA folding (stem–loop *versus* 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 necessarily 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 CAs. 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. 4a) 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. 4b), likely facilitate the detection and the fixation of the targeted guanine by this residue. This guanine in fact concentrates most contacts (Fig. 5b) 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. 6a).

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. 5a), 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. 3b), 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 CAs 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 toward 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, HIS 23, THR 24, ARG 26 and ASN 27) engage substantial contacts with the first C<sub>105</sub>:G<sub>111</sub> 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. 6c) 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 (LYS or ARG). 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 with 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 NAs and protein contained in each complex is supplied

in 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) become 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, a software originally developed for proteins [44] and recently extended to NAs [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). The protein or NA accessibility was represented by the CAs between water molecules and solute residues. The NCp7/NA interfaces were quantified by the CAs 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°.

### Acknowledgments

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

### CRedit authorship contribution statement

**Romain Retureau:** Investigation, Visualization. **Christophe Oguey:** Methodology, Software, Writing - review & editing. **Olivier Mauffret:** Funding acquisition, Validation. **Brigitte Hartmann:** Conceptualization, Supervision, Writing - original draft, Writing - review & editing.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmb.2019.03.002>.

Received 7 January 2019;

Received in revised form 25 February 2019;

Accepted 2 March 2019

Available online 12 March 2019

#### Keywords:

nucleo-capsid protein;

NCp7;

protein–nucleic acid interactions;

protein–nucleic acid binding process;

Voronoi tessellation

#### Abbreviations used:

HIV-1, human immunodeficiency virus type 1; NA, nucleic acid; 3D, three-dimensional; VLDM, Voronoi Laguerre Delaunay for Macromolecules; CA, contact area.

### 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, <https://doi.org/10.1002/pro.5560010502>.
- [2] E.O. Freed, HIV-1 assembly, release and maturation, *Nat. Rev. Microbiol.* 13 (2015) 484–496, <https://doi.org/10.1038/nrmicro3490>.
- [3] E. Mailler, S. Bemacchi, R. Marquet, J.-C. Paillart, V. Vivet-Boudou, R.P. Smyth, The life-cycle of the HIV-1 Gag–RNA complex, *Viruses*. 8 (2016), <https://doi.org/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), <https://doi.org/10.3390/v8090246>.
- [5] J.A. Thomas, R.J. Gorelick, Nucleocapsid protein function in early infection processes, *Virus Res.* 134 (2008) 39–63, <https://doi.org/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, [https://doi.org/10.1016/S1054-3589\(07\)55009-X](https://doi.org/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. <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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. <https://doi.org/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, <https://doi.org/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. <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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. <https://doi.org/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, <https://doi.org/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. Lyonais, 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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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 <sup>1</sup>H 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, <https://doi.org/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* 17 (2016) 1548–1552, <https://doi.org/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. <https://doi.org/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, <https://doi.org/10.1074/jbc.M303819200>.
- [36] P.-J. Racine, C. Chamontin, H. de Rocquigny, S. Bernacchi, J.-C. Paillart, M. Mougél, Requirements for nucleocapsid-mediated regulation of reverse transcription during the late steps of HIV-1 assembly, *Sci. Rep.* 6 (2016), 27536. <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1093/nar/gkt509>.
- [45] A. Elbahsi, R. Rétureau, 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, <https://doi.org/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, <https://doi.org/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. <https://doi.org/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* 19 (2013) 1078–1088, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1134/S0006297914070062>.
- [53] S.Z. Borozan, B.P. Dimitrijević, S.Đ. Stojanović, Cation– $\pi$  interactions in high resolution protein–RNA complex crystal structures, *Comput. Biol. Chem.* 47 (2013) 105–112, <https://doi.org/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, <https://doi.org/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, <https://doi.org/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. <https://doi.org/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, <https://doi.org/10.1006/jmbi.1994.1334>.

## SUPPLEMENTARY DATA

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

Romain Retureau, Christophe Oguey, Olivier Mauffret and Brigitte Hartmann

**Figure S1:** Comparison of the conformations of the ZF1-ZF2 ensemble across the NCp7-NA systems.

**Figure S2:** Comparison of the conformations of the free and bound structures of the ZF1-ZF2 ensemble.

**Figure S3:** Hydrogen bonds at the NCp7/NA interfaces.

**Figure S4:** NA elements composing the NCp7/NA interface.

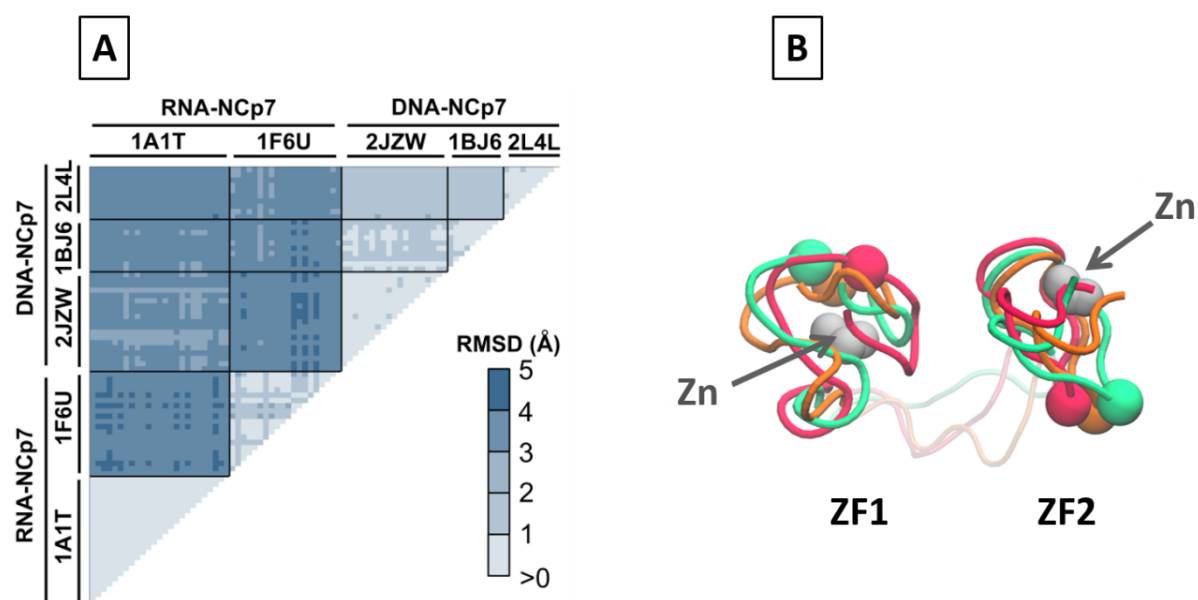
**Figure S5:** ZF2-linker/NA interfaces.

**Figure S6:** Contact areas of the main interactions at NCp7/NA interfaces.

**Tables S1-1, -2 and -3.** Amino acid and nucleotide composition and contribution to NCp7/NA interfaces.

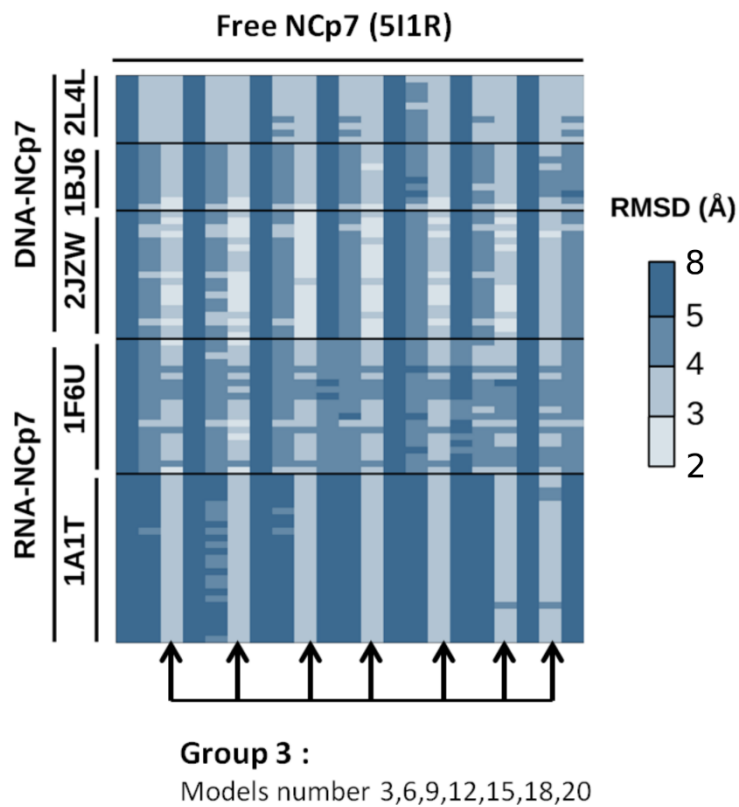
**Figure S1:** Comparison of the conformations of the ZF1-ZF2 ensemble across the NCp7-NA systems.

**A:** The structures of the ZF1-ZF2 ensemble was compared across the models of the five systems studied here by calculating cross-RMSDs on the backbone heavy atoms of the CYS 15 → CYS 28 and CYS 36 → CYS 49 segments considered together. The RMSD values are colored according to the code given on the bottom right. **B:** The superimposition of the backbone trace of models 1 from 1A1T (red), 18 from 1F6U (orange) and 8 from 2JZW (green) illustrates the conformational similarity of the ZF1-ZF2 ensemble. The associated cross-RMSDs are  $2.0 \pm 0.3$  Å. The linkers are represented by transparent backbone traces. To better depict the zinc finger backbone courses, the C $\alpha$  atoms of residues 20 and 43, located at the center of the ZF1 and ZF2 sequences, respectively, were represented by balls.



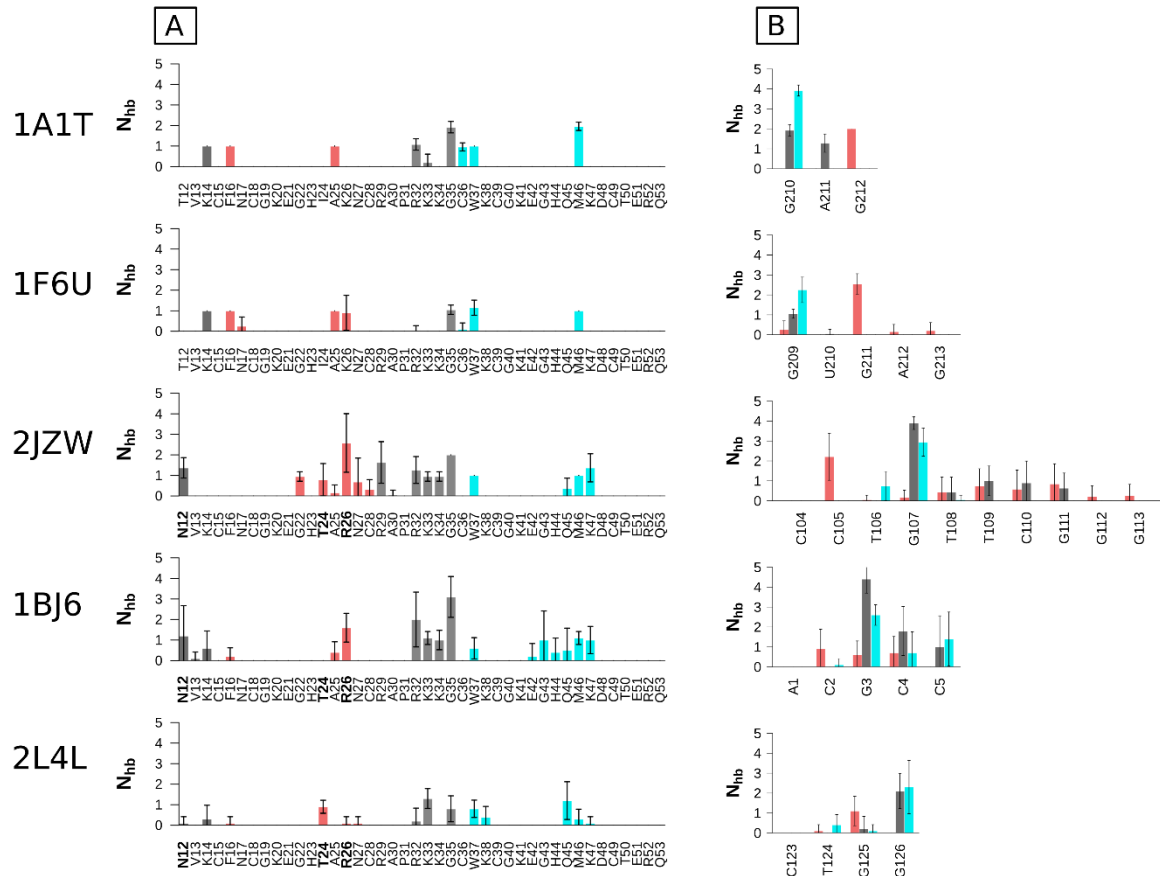
**Figure S2:** Comparison of the conformations of the free and bound structures of the ZF1-ZF2 ensemble.

The structures of the ZF1-ZF2 ensemble was compared between the models of the five NCp7-NA systems and the free NCp7 models by calculating cross-RMSDs on the backbone heavy atoms of the CYS 15 → CYS 28 and CYS 36 → CYS 49 segments considered together. The RMSD values are colored according to the code given on the right.



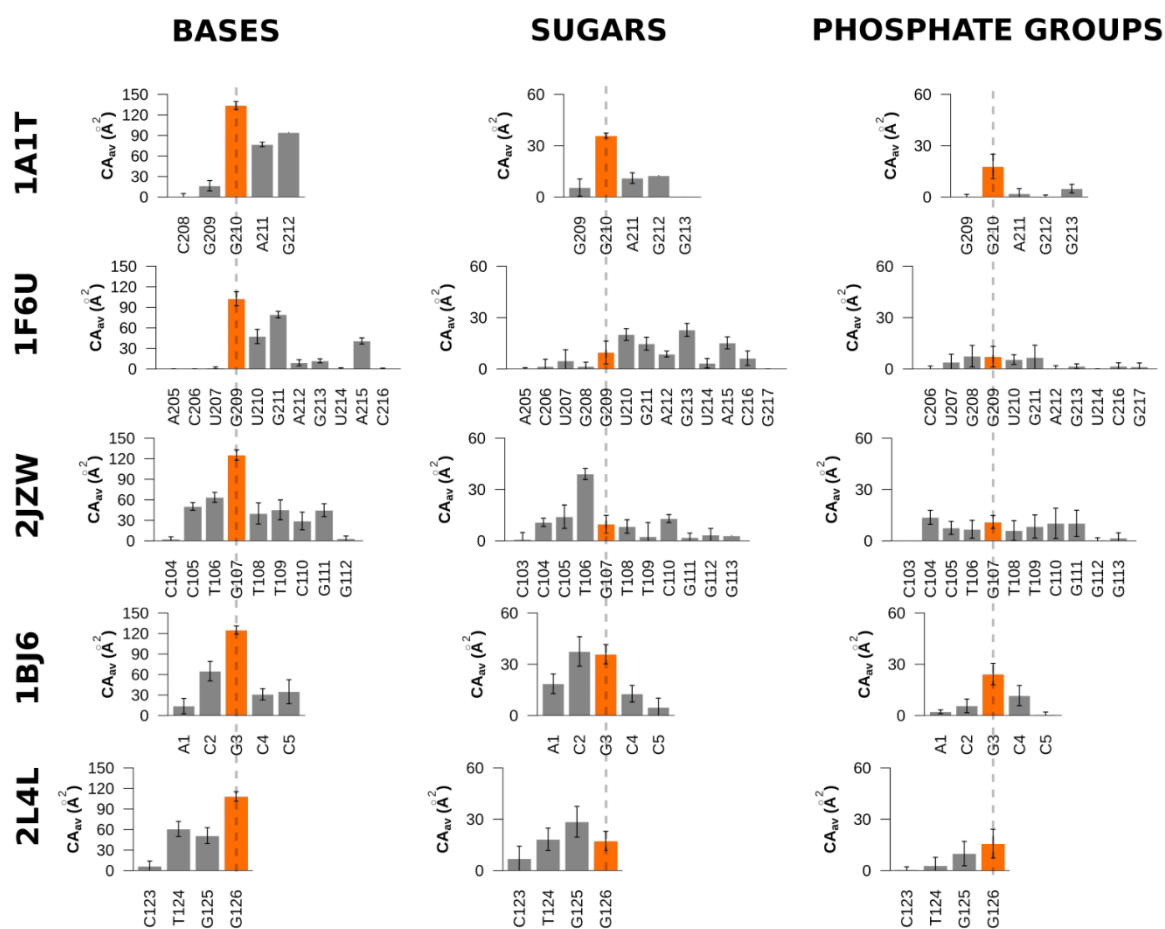
**Figure S3:** Hydrogen bonds at the NCp7/NA interfaces.

The number ( $N_{av}$ ) of hydrogen bonds between NCp7 and NA residues extracted from HBPlus analysis was plotted along the NCp7 (A) and NA (B) sequences, considering independently the five systems studied here.  $N_{av}$  values were averaged on the ensemble of models constituting each system; the vertical error bars represent standard deviations. The data associated to ZF1 is in red, ZF2 in cyan, the linker in grey and the N- and C-terminal domains in black. The hydrogen bonds were defined by  $D-H-A > 120^\circ$  and  $D-A$  distance  $< 3.9 \text{ \AA}$ ; more permissive criteria ( $D-H-A > 60^\circ$  and  $D-A$  distance  $< 4.1 \text{ \AA}$ ) led to very similar results.



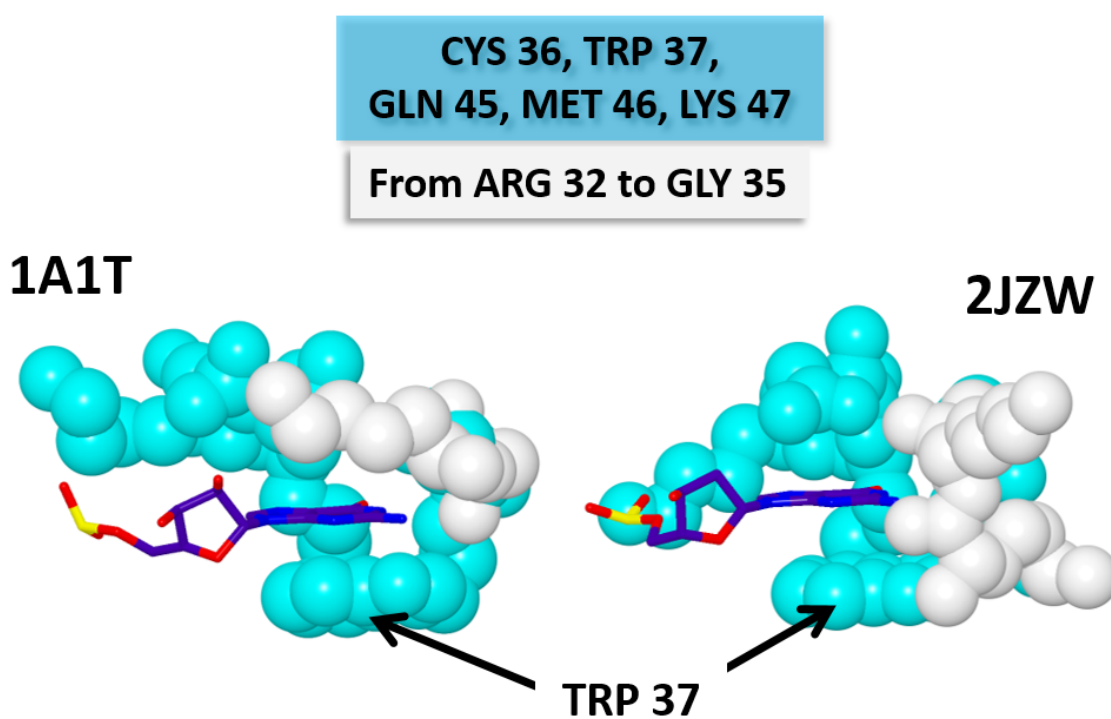
**Figure S4:** NA elements composing the NCp7/NA interface.

The contribution of the atoms of bases, sugars (C1', C2', C3', C4', C5', O4' and, if relevant, O2' atoms) and phosphate groups (P, O1P, O2P, O5' and O3' atoms) to the interfaces is given in term of CAs ( $\text{\AA}^2$ ) along the NA sequences, for the five systems studied here. The CA values were calculated on the ensemble of models constituting each system and averaged; the vertical error bars correspond to standard deviations. The whole ensemble of NCp7 amino acids (residues 12  $\rightarrow$  53) and nucleotides were considered. Note the different CA ranges in the vertical axes, more extended for the CAs associated to bases. The NA sequences are aligned on the guanines (orange) contacted by TRP 37 of ZF2, a very robust interaction point in all systems. The other contacts involving both bases and sugars are mainly located 3' this specific guanine in NCp7-RNA complexes.



**Figure S5:** ZF2-linker/NA interfaces.

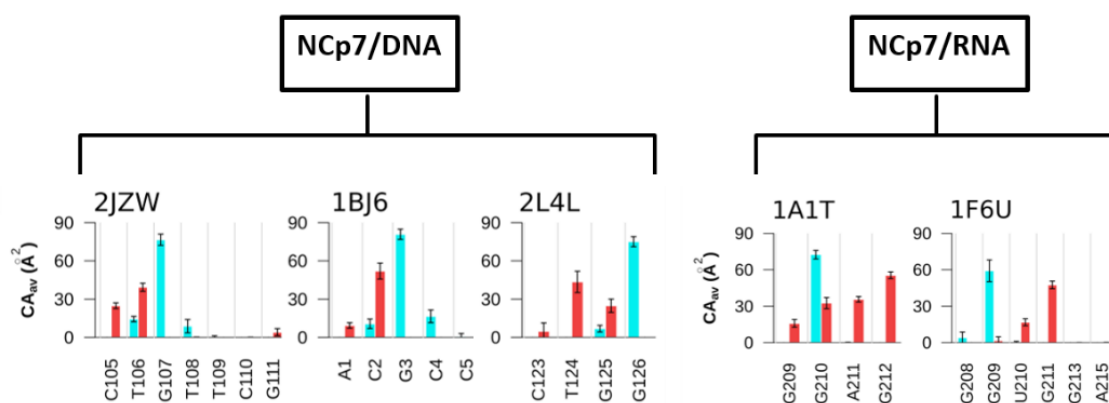
Typical structures representing pockets formed by ZF2 (cyan) and linker (grey) residues surrounding a single guanine belonging to RNA (1A1T) or DNA (2JZW) loops.





**Figure S6:** Contact areas of the main interactions at NCp7/NA interfaces.

The contact areas ( $CA_{av}$ ) are plotted along the NA sequences for each studied complex, focusing on the interactions with PHE 16 and residue 24 of ZF1 (red) on one hand and TRP 37 and GLN 45 of ZF2 (cyan) on the other hand. The data were averaged over the models that constitute each system; the vertical thin error bars correspond to standard deviations.



**Tables S1-1, -2 and -3.** Amino acid and nucleotide composition and contribution to NCp7/NA interfaces.

The three tables below (pages 8, 9 and 10) present the amino acids and nucleotides involved in the NCp7/NA interfaces. The data were obtained by VLDM analyses of the five systems considered in this study. Tables S2-1, -2 and -3 report the data of the interfaces involving ZF1, the linker and ZF2, respectively. The interaction of each individual couple of residues is characterized by the contact area (in Å<sup>2</sup>) calculated and averaged on the models of each system. The column “Sum<sub>aa</sub>” and the rows “Sum<sub>nuc</sub>” correspond to the contribution sum of each amino acid or nucleotide to the interface; yellow backgrounds highlighted the total CAs (sum of Sum<sub>aa</sub> = sum of Sum<sub>nuc</sub>) associated to each system. The standard deviations are not given explicitly at each CA value to preserve the Table readability; however, to distinguish interactions recurrently encountered across the models of each system, CAs of Sum<sub>aa</sub> and Sum<sub>nuc</sub> associated to standard deviations < 20% of their value are marked in bold on a grey background; otherwise, they are in italic on a light grey background. The amino acid and nucleotide numberings follow the PDB files.

Table S1-1. ZF1/NA interfaces.

|                    |                    | THR<br>12 | VAL<br>13 | LYS<br>14 | CYS<br>15 | PHE<br>16 | ASN<br>17 | GLY<br>22 | HSD<br>23 | ILE<br>24 | ALA<br>25 | LYS<br>26 | ASN<br>27 | CYS<br>28 | Sum <sub>aa</sub> |
|--------------------|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-------------------|
| 1A1T               | G209               |           | 4         | 1         |           | 16        |           |           |           |           |           |           |           |           | 22                |
|                    | G210               |           |           | 2         |           | 33        | 4         |           |           |           |           |           |           |           | 39                |
|                    | A211               |           |           |           |           | 36        | 9         |           |           |           | 17        | 7         |           |           | 69                |
|                    | G212               |           | 8         | 13        | 1         | 27        |           |           |           | 29        | 12        | 18        |           |           | 107               |
|                    | G213               |           |           |           |           |           |           |           |           |           |           | 5         |           |           | 5                 |
|                    | Sum <sub>nuc</sub> |           | 12        | 15        | 1         | 111       | 14        |           |           | 29        | 29        | 30        |           |           | 242               |
| 1F6U               | G209               |           |           |           |           | 2         | 10        |           |           |           |           |           |           |           | 12                |
|                    | U210               |           |           |           |           | 17        | 15        |           |           |           | 23        | 16        |           | 1         | 73                |
|                    | G211               |           | 2         | 9         | 3         | 19        |           |           |           | 29        | 10        | 28        |           |           | 100               |
|                    | A212               |           |           |           |           |           |           |           |           |           |           | 18        |           |           | 18                |
|                    | G213               |           |           |           |           |           |           |           |           |           |           | 32        | 4         |           | 37                |
|                    | U214               |           |           |           |           |           |           |           |           |           |           | 1         | 5         |           | 6                 |
|                    | A215               |           |           |           |           |           |           | 4         | 13        |           |           |           | 33        | 1         | 51                |
|                    | Sum <sub>nuc</sub> |           | 2         | 9         | 3         | 37        | 25        | 4         | 13        | 29        | 33        | 96        | 42        | 2         | 297               |
|                    |                    | ASP<br>12 | VAL<br>13 | LYS<br>14 | CYS<br>15 | PHE<br>16 | ASN<br>17 | GLY<br>22 | HSD<br>23 | THR<br>24 | ALA<br>25 | ARG<br>26 | ASN<br>27 | CYS<br>28 | Sum <sub>aa</sub> |
| 2JZW               | C104               | 22        | 1         |           |           |           |           | 1         |           |           |           |           |           |           | 25                |
|                    | C105               | 6         | 21        |           |           |           |           | 7         | 6         | 25        |           | 6         | 1         |           | 72                |
|                    | T106               |           | 5         |           |           | 27        |           |           |           | 13        | 4         | 18        |           |           | 67                |
|                    | G107               |           |           |           |           |           |           |           |           |           |           | 3         |           |           | 3                 |
|                    | T108               |           |           |           |           |           | 1         |           |           |           | 6         | 17        |           |           | 24                |
|                    | T109               |           |           |           |           |           | 4         |           |           |           | 4         | 15        |           | 4         | 27                |
|                    | C110               |           |           |           |           |           |           |           |           |           | 1         | 22        | 6         | 1         | 30                |
|                    | G111               |           |           |           |           |           |           |           | 2         | 4         |           | 18        | 26        |           | 51                |
|                    | G112               |           |           |           |           |           |           |           |           |           |           | 10        |           |           | 10                |
|                    | G113               |           |           |           |           |           |           |           |           |           |           | 5         |           |           | 5                 |
| Sum <sub>nuc</sub> | 28                 | 27        |           |           | 27        | 5         | 9         | 8         | 42        | 15        | 115       | 33        | 5         | 314       |                   |
| 1BJ6               | A1                 | 18        | 12        |           |           |           |           |           |           | 9         |           |           |           |           | 39                |
|                    | C2                 |           | 9         | 3         | 1         | 31        |           |           |           | 21        | 10        | 16        |           |           | 91                |
|                    | G3                 |           |           |           |           |           |           |           |           |           | 2         | 24        |           |           | 26                |
|                    | C4                 |           |           |           |           |           |           |           |           |           |           | 5         |           |           | 5                 |
|                    | Sum <sub>nuc</sub> | 18        | 21        | 3         | 1         | 31        |           |           |           | 30        | 12        | 45        |           |           | 161               |
| 2L4L               | C123               | 2         | 7         |           |           |           |           |           |           | 5         |           |           |           |           | 13                |
|                    | T124               |           | 13        | 8         | 1         | 32        |           |           |           | 12        | 2         |           |           |           | 68                |
|                    | G125               |           |           |           |           | 3         |           |           |           | 22        | 6         | 26        | 3         |           | 60                |
|                    | G126               |           |           |           |           |           |           |           |           |           |           | 5         |           |           | 6                 |
|                    | Sum <sub>nuc</sub> | 2         | 20        | 8         | 1         | 35        |           |           |           | 38        | 9         | 31        | 3         |           | 147               |

Table S1-2. Linker/NA interfaces.

|                    |                    | ARG<br>29 | ALA<br>30 | ARG<br>32 | LYS<br>33 | LYS<br>34 | GLY<br>35 | Sum <sub>aa</sub> |
|--------------------|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|-------------------|
| 1A1T               | G210               |           |           |           | 2         | 14        | 11        | 28                |
|                    | A211               |           |           | 18        | 6         | 2         |           | 26                |
|                    | Sum <sub>nuc</sub> |           |           | 18        | 9         | 16        | 11        | 54                |
| 1F6U               | C206               |           |           | 1         | 1         |           |           | 2                 |
|                    | U207               |           |           | 2         | 3         | 4         |           | 8                 |
|                    | G208               |           |           |           |           | 3         |           | 4                 |
|                    | G209               |           |           | 3         |           | 2         | 9         | 14                |
|                    | U210               |           |           | 2         |           | 2         |           | 4                 |
|                    | A215               | 6         |           |           |           |           |           | 6                 |
|                    | C216               | 11        |           |           |           |           |           | 11                |
|                    | G217               | 1         |           |           |           |           |           | 1                 |
| Sum <sub>nuc</sub> | 18                 |           | 8         | 4         | 11        | 9         | 50        |                   |
| 2JZW               | G107               |           |           | 7         | 11        | 7         | 11        | 36                |
|                    | T108               |           |           | 20        |           |           |           | 20                |
|                    | T109               | 6         | 4         | 18        |           |           |           | 28                |
|                    | C110               | 18        | 1         |           |           |           |           | 19                |
|                    | G111               | 7         |           |           |           |           |           | 7                 |
|                    | Sum <sub>nuc</sub> | 32        | 4         | 46        | 11        | 7         | 11        | 110               |
| 1BJ6               | G3                 |           |           | 18        | 12        | 7         | 10        | 46                |
|                    | C4                 |           |           | 18        | 6         | 3         | 8         | 34                |
|                    | C5                 |           |           | 5         | 2         | 10        | 8         | 25                |
|                    | Sum <sub>nuc</sub> |           |           | 40        | 19        | 20        | 26        | 105               |
| 2L4L               | G125               |           |           | 1         |           |           |           | 1                 |
|                    | G126               |           |           | 13        | 13        |           | 6         | 33                |
|                    | Sum <sub>nuc</sub> |           |           | 15        | 13        |           | 6         | 34                |

Table S1-3. ZF2/NA interfaces.

|      |                    | CYS<br>36 | TRP<br>37 | LYS<br>38 | GLU<br>42 | GLY<br>43 | HIS<br>44 | GLN<br>45 | MET<br>46 | LYS<br>47 | Sum <sub>aa</sub> |
|------|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-------------------|
| 1A1T | G209               |           |           |           |           |           |           |           |           | 1         | 1                 |
|      | G210               | 10        | 40        |           |           |           | 2         | 32        | 17        | 22        | 124               |
|      | A211               |           |           |           |           |           |           |           |           | 1         | 1                 |
|      | Sum <sub>nuc</sub> | 10        | 40        |           |           |           | 2         | 32        | 17        | 24        | 126               |
| 1F6U | G208               |           |           |           |           |           |           | 4         |           |           | 5                 |
|      | G209               | 4         | 37        |           |           |           |           | 22        | 19        | 10        | 92                |
|      | Sum <sub>nuc</sub> | 4         | 37        |           |           |           |           | 26        | 19        | 10        | 97                |
| 2JZW | T106               |           | 14        |           |           |           |           |           | 14        | 23        | 52                |
|      | G107               | 4         | 48        |           |           |           |           | 29        | 14        | 16        | 110               |
|      | T108               |           | 7         |           |           |           |           | 1         |           |           | 9                 |
|      | Sum <sub>nuc</sub> | 4         | 70        |           |           |           |           | 30        | 28        | 40        | 171               |
| 1BJ6 | C2                 |           | 11        |           |           |           |           |           | 12        | 1         | 24                |
|      | G3                 | 8         | 56        |           |           |           |           | 25        | 22        | 12        | 123               |
|      | C4                 |           |           |           |           |           | 3         | 17        |           |           | 20                |
|      | C5                 |           |           |           | 5         | 7         | 4         | 1         |           |           | 17                |
|      | Sum <sub>nuc</sub> | 8         | 67        |           | 5         | 7         | 7         | 42        | 34        | 14        | 184               |
| 2L4L | T124               |           |           | 9         |           |           |           |           | 4         | 3         | 16                |
|      | G125               |           | 7         | 1         |           |           |           |           | 15        | 6         | 28                |
|      | G126               | 7         | 45        |           |           |           | 3         | 30        | 21        | 1         | 107               |
|      | Sum <sub>nuc</sub> | 7         | 52        | 10        |           |           | 3         | 30        | 39        | 10        | 151               |

## Conclusion

Cette analyse exhaustive des complexes NCp7/NA nous permet de proposer un mécanisme pour la fixation de la NCp7 sur des AN. Dans un premier temps, la TRP37 accessible de la NCp7 en forme ouverte reconnaît une guanine accessible (comme dans un simple brin) d'un AN. La forme fermée de la NCp7 et la charge portée par certains résidus du *linker* et ZF1, facilitent ensuite l'interaction de ZF1/AN qui dépend du type d'AN, de la séquence de l'AN ou de la séquence de la NCp7. Ceci suggère que ZF1 est capable de s'adapter à plusieurs facteurs.

Au laboratoire, l'équipe caractérise actuellement des nouveaux systèmes par RMN où les NCp7 des deux souches NL4-3 et MB sont complexés à une série d'ADN de séquences différentes. Cette étude permettra d'étudier les effets de la mutation I24T de la NCp7 et de la séquence d'ADN sur l'interface ZF1/ADN. Pour aller plus loin, il serait intéressant ensuite d'étudier des séquences ARN et ADN identiques ou très semblables pour mieux comprendre comment la NCp7 peut distinguer ARN et ADN.

# Discussion et conclusion

Le nucléosome et les complexes NCp7/AN sont donc deux exemples qui mettent en lumière la diversité structurale des interfaces AN/protéines non spécifiques et des mécanismes de fixation qui les caractérisent. Au cours de mes travaux, à l'aide de méthodes bio-informatiques j'ai pu apporter de nouvelles connaissances structurales et ainsi que de nouvelles pistes d'explorations. J'ai aussi eu la chance de pouvoir travailler avec des expérimentalistes et d'utiliser mes résultats théoriques pour aider à interpréter leurs résultats.

## Le nucléosome

Grâce aux évolutions des méthodes de simulations et des performances des ordinateurs, j'ai pu travailler sur des dynamiques moléculaires de plus de 1 $\mu$ s du nucléosome, un très gros complexe, au niveau atomique, en solution. J'ai analysé ces simulations, et grâce à un nouveau logiciel, VLDM, j'ai pu apporter une description précise de l'interface ADN/histone, caractérisée par des contacts très denses et très stables. Pour maintenir l'ADN enroulé autour du cœur d'histones, nous avons mis en évidence la forte contribution des contacts hydrophobes qui viennent compléter les contacts électrostatiques. Nous montrons également que les queues d'histones représentent près d'un tiers des contacts à l'interface et que la présence d'ions entre les deux doubles hélices d'ADN superposées permet d'écranter les groupements phosphates proches. En collaborant avec l'équipe de Malcolm Buckle, j'ai pu mettre directement à contribution ces travaux sur l'interface ADN/histone afin d'interpréter des résultats expérimentaux d'assemblage et de désassemblage du nucléosome. Nous avons ainsi pu éclaircir les étapes structurales successives qui mènent à la formation et à la dissociation du nucléosome. Nous avons également mis en évidence l'effet de la séquence ADN sur ces étapes.

En parallèle, et dans la continuité des travaux menés au laboratoire sur la dynamique de l'ADN, j'ai analysé le comportement de l'ADN nucléosomal dans nos simulations. Cette analyse révèle que malgré l'interface très dense et stable, l'ADN nucléosomal garde certaines propriétés des ADN libres. Notamment, il conserve et même amplifie le couplage entre l'équilibre BI $\leftrightarrow$ BII et les paramètres locaux *roll*, *slide* et *twist*. Cependant, la distribution périodique de ces paramètres long de la séquence caractérise l'ADN nucléosomal. On observe également que l'effet de séquence sur la flexibilité des dinucléotides mis en évidence sur des ADN libres persiste dans l'ADN nucléosomal. Bien qu'imparfaite, cette persistance est suffisante pour signifier que les propriétés de l'ADN libre, et dans notre cas celles de la séquence 601, sont bel est bien exploitées dans le cadre de la formation du nucléosome.



A la suite de ces travaux, j'ai cherché à définir plus finement ce qui caractérise chacun des sites de fixation ADN/histone. Toujours grâce aux dynamiques moléculaires, j'ai pu analyser l'insertion des arginines du cœur d'histone dans le petit sillon de l'ADN. Je constate que ces insertions ne dépendent pas uniquement des composantes électrostatiques offertes par le petit sillon mais sont également caractérisées par i) un réseau protéique autour de l'arginine en contact avec l'ADN et ii) une structure de petit sillon adaptable.

Ces résultats sur le comportement des arginines, même préliminaires, sont particulièrement intéressants dans le cadre de l'étude des modifications post-traductionnelles du cœur d'histone déjà répertoriées (Bowman et Poirier 2015) et de leur possible impact sur l'interface. L'exemple le plus probant est celui de la phosphorylation de la thréonine 118 de l'histone H3 qui entraîne une déstabilisation du nucléosome (North et al. 2014). Cette H3-T118 correspond bien à un résidu qui participe à maintenir l'arginine H4-R45 à l'intérieur du petit sillon au SHL +/- 1.5. De plus, en identifiant des réseaux d'interaction similaires, je peux proposer de nouveaux sites de modification potentiels, notamment au niveau des histones H2A et H2B pour lesquelles il n'existe aucun MPT connu vers l'arginine insérée. Bien que Lüger ait déjà proposé un site équivalent sur l'histone H4 au niveau de la thréonine 80, notre modèle nous permet d'en proposer deux supplémentaires, les thréonines 76 et 85 des histones H2A et H2B respectivement. De la même manière, nous avons montré la forte contribution des queues d'histones, et plus particulièrement de leur racine, dans les contacts à l'interface. Or, il existe également des sites connus de MPT dans ces régions, lesquelles sont impliquées dans la déstabilisation de la chromatine (Bowman et Poirier 2015). D'autres analyses préliminaires que j'ai entreprises montrent des fluctuations plus importantes du squelette peptidique et des variations dans le positionnement des chaînes latérales des résidus dans ces régions. On remarque notamment au niveau de la queue N-terminale de H2B qui se situe entre les deux doubles hélices d'ADN superposées que les positions de trois arginines fluctuent beaucoup. De plus, ces résidus sont plus proches de l'extérieur de la structure nucléosomal et seraient donc plus accessibles aux agents responsables des MPT.

## Les complexes AN/NCp7

La NCp7 du VIH-1 est une protéine très étudiée au laboratoire et c'est dans ce cadre que j'ai entrepris son étude afin de mieux comprendre la fixation de la NCp7 sur les ADN et sur les ARN. Ce système, beaucoup plus petit que le nucléosome, m'a permis d'appliquer VLDM et les scripts que j'avais développé précédemment. En analysant les cinq complexes AN/NCp7 expérimentaux disponibles et en quantifiant les interfaces, il m'est apparu très rapidement que comme pour le nucléosome, les contacts hydrophobes sont une composante importante de l'interface. J'ai pu faire ressortir les résidus qui sont impliqués de manière systématique dans l'interface et extraire des premières informations sur le mécanisme de fixation de la NCp7. Dans ZF2, l'empilement de TRP37 avec une guanine dont l'autre face était encadrée par le GLN45 était très clairement observé dans tous les systèmes alors que la fixation de ZF1 était différente selon le type d'AN (ADN versus ARN) et la souche d'origine de la NCp7. Cette cohérence très forte du mode de fixation de ZF2 a permis de proposer que la fixation de la NCp7 sur l'AN passe d'abord par ce doigt de zinc. J'ai ensuite poussé l'analyse en étudiant le comportement de la NCp7 libre, et en caractérisant deux types de conformations explorées, l'une ouverte et l'autre fermée ; j'ai mis en évidence le fait que l'accessibilité de ZF2 était plus forte dans sa conformation ouverte que dans sa conformation fermée. La compilation de ces résultats m'a donc permis de proposer un mécanisme de fixation séquentielle de la NCp7 : d'abord la NCp7 dans sa conformation ouverte se fixe sur l'AN par le biais de ZF2 ; ensuite l'équilibre ouvert/fermé de la protéine permet à ZF1 de se fixer selon différents modes qui peuvent dépendre de la séquence et du type d'AN et de la séquence de la NCp7.

Au laboratoire, à la suite de cette étude, deux complexes avec la même séquence ADN mais deux souches NCp7 différentes sont en train d'être étudiés afin de clarifier l'impact de la séquence NCp7 sur ces complexes.

## Conclusion générale

Les pistes que nous avons ouvertes pour des travaux futurs montrent l'intérêt de ce genre d'étude pour comprendre les mécanismes qui régissent le nucléosome. Expérimentalement, c'est essentiellement la cristallographie aux rayons X qui a permis d'obtenir des informations structurales sur le nucléosome, trop gros pour être accessible au niveau atomique à une approche en solution, comme la RMN. Dans le cas de la NCp7 et de ses complexes, c'est l'inverse : cette petite protéine, même complexée aux AN, est trop flexible pour cristalliser et est donc étudiée par RMN, avec toutes les difficultés que cela comporte de travailler en solution sur des complexes non spécifiques, peu stables. La dynamique moléculaire associée à l'utilisation d'un outil d'analyse performant comme VLDM nous offre un regard complémentaire aux expériences et nous permet d'enrichir nos connaissances des mécanismes qui régissent les biomolécules.

- Abd El-Wahab, Ekram W., Redmond P. Smyth, Elodie Mailler, Serena Bernacchi, Valérie Vivet-Boudou, Marcel Hijnen, Fabrice Jossinet, Johnson Mak, Jean-Christophe Paillart, et Roland Marquet. 2014. « Specific Recognition of the HIV-1 Genomic RNA by the Gag Precursor ». *Nature Communications* 5 (juillet): 4304. <https://doi.org/10.1038/ncomms5304>.
- Albert, Istvan, Travis N. Mavrich, Lynn P. Tomsho, Ji Qi, Sara J. Zanton, Stephan C. Schuster, et B. Franklin Pugh. 2007. « Translational and Rotational Settings of H2A.Z Nucleosomes across the *Saccharomyces Cerevisiae* Genome ». *Nature* 446 (7135): 572-76. <https://doi.org/10.1038/nature05632>.
- Allers, J., et Y. Shamoo. 2001. « Structure-Based Analysis of Protein-RNA Interactions Using the Program ENTANGLE ». *Journal of Molecular Biology* 311 (1): 75-86. <https://doi.org/10.1006/jmbi.2001.4857>.
- Amarasinghe, G. K., R. N. De Guzman, R. B. Turner, K. J. Chancellor, Z. R. Wu, et M. F. Summers. 2000. « NMR Structure of the HIV-1 Nucleocapsid Protein Bound to Stem-Loop SL2 of the Psi-RNA Packaging Signal. Implications for Genome Recognition ». *Journal of Molecular Biology* 301 (2): 491-511. <https://doi.org/10.1006/jmbi.2000.3979>.
- Anton, Halina, Nedal Taha, Emmanuel Boutant, Ludovic Richert, Heena Khatter, Bruno Klaholz, Philippe Rondé, Eléonore Réal, Hugues de Rocquigny, et Yves Mély. 2015. « Investigating the Cellular Distribution and Interactions of HIV-1 Nucleocapsid Protein by Quantitative Fluorescence Microscopy ». *PLoS ONE* 10 (2). <https://doi.org/10.1371/journal.pone.0116921>.
- Avilov, Sergiy V., Julien Godet, Etienne Piémont, et Yves Mély. 2009. « Site-Specific Characterization of HIV-1 Nucleocapsid Protein Binding to Oligonucleotides with Two Binding Sites <sup>†</sup> ». *Biochemistry* 48 (11): 2422-30. <https://doi.org/10.1021/bi8022366>.
- Bai, Lu, et Alexandre V. Morozov. 2010. « Gene Regulation by Nucleosome Positioning ». *Trends in Genetics* 26 (11): 476-83. <https://doi.org/10.1016/j.tig.2010.08.003>.
- Baldi, Sandro. 2019. « Nucleosome Positioning and Spacing: From Genome-Wide Maps to Single Arrays ». *Essays In Biochemistry* 63 (1): 5-14. <https://doi.org/10.1042/EBC20180058>.
- Barré-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, et al. 1983. « Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS) ». *Science (New York, N.Y.)* 220 (4599): 868-71. <https://doi.org/10.1126/science.6189183>.
- Batabyal, Subrata, Susobhan Choudhury, Dilip Sao, Tanumoy Mondol, et Pal Samir Kumar. 2014. « Dynamical perspective of protein-DNA interaction ». *BioMolecular Concepts* 5 (1): 21-43. <https://doi.org/10.1515/bmc-2013-0037>.
- Baxevanis, Andreas D., et David Landsman. 1996. « Histone Sequence Database: A Compilation of Highly-Conserved Nucleoprotein Sequences ». *Nucleic Acids Research* 24 (1): 245-47. <https://doi.org/10.1093/nar/24.1.245>.
- Bazzi, Ali, Loussiné Zargarian, Françoise Chaminade, Christian Boudier, Hughes De Rocquigny, Brigitte René, Yves Mély, Philippe Fossé, et Olivier Mauffret. 2011. « 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 Research* 39 (9): 3903-16. <https://doi.org/10.1093/nar/gkq1290>.
- Becker, Peter B., et Jerry L. Workman. 2013. « Nucleosome Remodeling and Epigenetics ». *Cold Spring Harbor Perspectives in Biology* 5 (9). <https://doi.org/10.1101/cshperspect.a017905>.
- Beltz, Hervé, Joel Azoulay, Serena Bernacchi, Jean-Pierre Clamme, Damien Ficheux, Bernard Roques, Jean-Luc Darlix, et Yves Mély. 2003. « Impact of the Terminal Bulges of HIV-1 CTAR DNA on Its Stability and the Destabilizing Activity of the Nucleocapsid Protein NCp7 ». *Journal of Molecular Biology* 328 (1): 95-108. [https://doi.org/10.1016/s0022-2836\(03\)00244-4](https://doi.org/10.1016/s0022-2836(03)00244-4).
- Beltz, Hervé, Céline Clauss, Etienne Piémont, Damien Ficheux, Robert J. Gorelick, Bernard Roques, Caroline Gabus, Jean-Luc Darlix, Hugues de Rocquigny, et Yves Mély. 2005. « Structural Determinants of HIV-1 Nucleocapsid Protein for CTAR DNA Binding and

- Destabilization, and Correlation with Inhibition of Self-Primed DNA Synthesis ». *Journal of Molecular Biology* 348 (5): 1113-26. <https://doi.org/10.1016/j.jmb.2005.02.042>.
- Ben Imeddourene, Akli, Ahmad Elbahnsi, Marc Gu eroult, Christophe Oguey, Nicolas Foloppe, et Brigitte Hartmann. 2015. « Simulations Meet Experiment to Reveal New Insights into DNA Intrinsic Mechanics ». *PLoS Computational Biology* 11 (12): e1004631. <https://doi.org/10.1371/journal.pcbi.1004631>.
- Berman, H. M. 1997. « Crystal Studies of B-DNA: The Answers and the Questions ». *Biopolymers* 44 (1): 23-44. [https://doi.org/10.1002/\(SICI\)1097-0282\(1997\)44:1<23::AID-BIP3>3.0.CO;2-1](https://doi.org/10.1002/(SICI)1097-0282(1997)44:1<23::AID-BIP3>3.0.CO;2-1).
- Bernacchi, Serena, Stoyl Stoylov, Etienne Pi mont, Damien Ficheux, Bernard P. Roques, Jean Luc Darlix, et Yves M ly. 2002. « HIV-1 Nucleocapsid Protein Activates Transient Melting of Least Stable Parts of the Secondary Structure of TAR and Its Complementary Sequence ». *Journal of Molecular Biology* 317 (3): 385-99. <https://doi.org/10.1006/jmbi.2002.5429>.
- Bernauer, J., J. Az , J. Janin, et A. Poupon. 2007. « A New Protein-Protein Docking Scoring Function Based on Interface Residue Properties ». *Bioinformatics (Oxford, England)* 23 (5): 555-62. <https://doi.org/10.1093/bioinformatics/btl654>.
- Best, Robert B., Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, et Alexander D. Mackerell. 2012. « Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone  $\phi$ ,  $\psi$  and Side-Chain  $\chi(1)$  and  $\chi(2)$  Dihedral Angles ». *Journal of Chemical Theory and Computation* 8 (9): 3257-73. <https://doi.org/10.1021/ct300400x>.
- Bjeli , Sasa, Silke Wieninger, Ilian Jelesarov, et Andrey Karshikoff. 2008. « Electrostatic Contribution to the Thermodynamic and Kinetic Stability of the Homotrimeric Coiled Coil Lpp-56: A Computational Study ». *Proteins* 70 (3): 810-22. <https://doi.org/10.1002/prot.21585>.
- Blanchet, Christophe, Marco Pasi, Krystyna Zakrzewska, et Richard Lavery. 2011. « CURVES+ Web Server for Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures ». *Nucleic Acids Research* 39 (Web Server issue): W68-73. <https://doi.org/10.1093/nar/gkr316>.
- B hm, Vera, Aaron R. Hieb, Andrew J. Andrews, Alexander Gansen, Andrea Rocker, Katalin T th, Karolin Luger, et J rg Langowski. 2011. « Nucleosome Accessibility Governed by the Dimer/Tetramer Interface ». *Nucleic Acids Research* 39 (8): 3093-3102. <https://doi.org/10.1093/nar/gkq1279>.
- Bombarda, E., A. Ababou, C. Vuilleumier, D. G rard, B. P. Roques, E. Pi mont, et Y. M ly. 1999. « Time-Resolved Fluorescence Investigation of the Human Immunodeficiency Virus Type 1 Nucleocapsid Protein: Influence of the Binding of Nucleic Acids ». *Biophysical Journal* 76 (3): 1561-70. [https://doi.org/10.1016/S0006-3495\(99\)77315-7](https://doi.org/10.1016/S0006-3495(99)77315-7).
- B rner, Richard, Danny Kowerko, Helena Guiset Miserachs, Michelle F. Schaffer, et Roland K. O. Sigel. 2016. « Metal ion induced heterogeneity in RNA folding studied by smFRET ». *Coordination Chemistry Reviews, A Special Issue in Honor of Professor Henryk Kozlowski*, 327-328 (novembre): 123-42. <https://doi.org/10.1016/j.ccr.2016.06.002>.
- Bourbigot, Sarah, Nick Ramalanjaona, Christian Boudier, Gilmar F. J. Salgado, Bernard P. Roques, Yves M ly, Serge Bouaziz, et Nelly Morellet. 2008. « How the HIV-1 Nucleocapsid Protein Binds and Destabilises the (-)Primer Binding Site during Reverse Transcription ». *Journal of Molecular Biology* 383 (5): 1112-28. <https://doi.org/10.1016/j.jmb.2008.08.046>.
- Bowman, Gregory D., et Michael G. Poirier. 2015. « Post-Translational Modifications of Histones That Influence Nucleosome Dynamics ». *Chemical Reviews* 115 (6): 2274-95. <https://doi.org/10.1021/cr500350x>.
- Brogaard, Kristin, Liqun Xi, Ji-Ping Wang, et Jonathan Widom. 2012. « A base pair resolution map of nucleosome positions in yeast ». *Nature* 486 (7404): 496-501. <https://doi.org/10.1038/nature11142>.
- Brooks, B.R., C.L. Brooks, A.D. MacKerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, et al. 2009.

- « CHARMM: The Biomolecular Simulation Program ». *Journal of computational chemistry* 30 (10): 1545-1614. <https://doi.org/10.1002/jcc.21287>.
- Brooks, C. L., et Martin Karplus. 1983. « Deformable stochastic boundaries in molecular dynamics ». *The Journal of Chemical Physics* 79 (12): 6312-25. <https://doi.org/10.1063/1.445724>.
- Brower-Toland, Brent D., Corey L. Smith, Richard C. Yeh, John T. Lis, Craig L. Peterson, et Michelle D. Wang. 2002. « Mechanical Disruption of Individual Nucleosomes Reveals a Reversible Multistage Release of DNA ». *Proceedings of the National Academy of Sciences of the United States of America* 99 (4): 1960-65. <https://doi.org/10.1073/pnas.022638399>.
- Burgess, Rebecca J., et Zhiguo Zhang. 2013. « Histone Chaperones in Nucleosome Assembly and Human Disease ». *Nature Structural & Molecular Biology* 20 (1): 14-22. <https://doi.org/10.1038/nsmb.2461>.
- Burley, S. K., et G. A. Petsko. 1985. « Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization ». *Science (New York, N.Y.)* 229 (4708): 23-28. <https://doi.org/10.1126/science.3892686>.
- Carteau, S., R. J. Gorelick, et F. D. Bushman. 1999. « Coupled Integration of Human Immunodeficiency Virus Type 1 CDNA Ends by Purified Integrase in Vitro: Stimulation by the Viral Nucleocapsid Protein ». *Journal of Virology* 73 (8): 6670-79.
- Chaudhury, Arindam, Praveen Chander, et Philip H. Howe. 2010. « Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles ». *RNA* 16 (8): 1449-62. <https://doi.org/10.1261/rna.2254110>.
- Chen, W., Q. Zhu, Y. Liu, et Q. Zhang. 2017. « Chapter Nine - Chromatin Remodeling and Plant Immunity ». In *Advances in Protein Chemistry and Structural Biology*, édité par Rossen Donev, 106:243-60. Chromatin Remodelling and Immunity. Academic Press. <https://doi.org/10.1016/bs.apcsb.2016.08.006>.
- Chen, Yujie, Joshua M. Tokuda, Traci Topping, Steve P. Meisburger, Suzette A. Pabit, Lisa M. Gloss, et Lois Pollack. 2017. « Asymmetric Unwrapping of Nucleosomal DNA Propagates Asymmetric Opening and Dissociation of the Histone Core ». *Proceedings of the National Academy of Sciences of the United States of America* 114 (2): 334-39. <https://doi.org/10.1073/pnas.1611118114>.
- Chu, Xiakun, et Jin Wang. 2014. « Specificity and Affinity Quantification of Flexible Recognition from Underlying Energy Landscape Topography ». *PLoS Computational Biology* 10 (8). <https://doi.org/10.1371/journal.pcbi.1003782>.
- Chua, Eugene Y. D., Dileep Vasudevan, Gabriela E. Davey, Bin Wu, et Curt A. Davey. 2012. « The Mechanics behind DNA Sequence-Dependent Properties of the Nucleosome ». *Nucleic Acids Research* 40 (13): 6338-52. <https://doi.org/10.1093/nar/gks261>.
- Crémisi, C., et M. Yaniv. 1980. « Sequential Assembly of Newly Synthesized Histones on Replicating SV40 DNA ». *Biochemical and Biophysical Research Communications* 92 (4): 1117-23. [https://doi.org/10.1016/0006-291x\(80\)90402-7](https://doi.org/10.1016/0006-291x(80)90402-7).
- Cui, Feng, et Victor B. Zhurkin. 2010. « Structure-based Analysis of DNA Sequence Patterns Guiding Nucleosome Positioning in vitro ». *Journal of biomolecular structure & dynamics* 27 (6): 821-41.
- Darden, Tom, Darrin York, et Lee Pedersen. 1993. « Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems ». *The Journal of Chemical Physics* 98 (12): 10089-92. <https://doi.org/10.1063/1.464397>.
- Darlix, J. L., M. Lapadat-Tapolsky, H. de Rocquigny, et B. P. Roques. 1995. « First Glimpses at Structure-Function Relationships of the Nucleocapsid Protein of Retroviruses ». *Journal of Molecular Biology* 254 (4): 523-37. <https://doi.org/10.1006/jmbi.1995.0635>.
- Darlix, J. L., A. Vincent, C. Gabus, H. de Rocquigny, et B. Roques. 1993. « Trans-Activation of the 5' to 3' Viral DNA Strand Transfer by Nucleocapsid Protein during Reverse Transcription of HIV1 RNA ». *Comptes Rendus de l'Academie Des Sciences. Serie III, Sciences de La Vie* 316 (8): 763-71.

- Darlix, Jean-Luc, José Luis Garrido, Nelly Morellet, Yves Mély, et Hugues de Rocquigny. 2007. « Properties, Functions, and Drug Targeting of the Multifunctional Nucleocapsid Protein of the Human Immunodeficiency Virus ». *Advances in Pharmacology (San Diego, Calif.)* 55: 299-346. [https://doi.org/10.1016/S1054-3589\(07\)55009-X](https://doi.org/10.1016/S1054-3589(07)55009-X).
- Darlix, Jean-Luc, Julien Godet, Roland Ivanyi-Nagy, Philippe Fossé, Olivier Mauffret, et Yves Mély. 2011. « Flexible Nature and Specific Functions of the HIV-1 Nucleocapsid Protein ». *Journal of Molecular Biology* 410 (4): 565-81. <https://doi.org/10.1016/j.jmb.2011.03.037>.
- Darlix, Jean-Luc, Hugues de Rocquigny, Olivier Mauffret, et Yves Mély. 2014. « Retrospective on the All-in-One Retroviral Nucleocapsid Protein ». *Virus Research* 193 (novembre): 2-15. <https://doi.org/10.1016/j.virusres.2014.05.011>.
- Davey, Curt A., David F. Sargent, Karolin Luger, Armin W. Maeder, et Timothy J. Richmond. 2002. « Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study. » *Journal of Molecular Biology* 319 (5): 1097-1113. [https://doi.org/10.1016/S0022-2836\(02\)00386-8](https://doi.org/10.1016/S0022-2836(02)00386-8).
- Davidovich, Chen, Xueyin Wang, Catherine Cifuentes-Rojas, Karen J. Goodrich, Anne R. Gooding, Jeannie T. Lee, et Thomas R. Cech. 2015. « Toward a Consensus on the Binding Specificity and Promiscuity of PRC2 for RNA ». *Molecular Cell* 57 (3): 552-58. <https://doi.org/10.1016/j.molcel.2014.12.017>.
- De Guzman, R. N., Z. R. Wu, C. C. Stalling, L. Pappalardo, P. N. Borer, et M. F. Summers. 1998. « Structure of the HIV-1 Nucleocapsid Protein Bound to the SL3 Psi-RNA Recognition Element ». *Science (New York, N.Y.)* 279 (5349): 384-88.
- Deshmukh, Lalit, Charles D. Schwieters, Alexander Grishaev, et G. Marius Clore. 2016. « Quantitative Characterization of Configurational Space Sampled by HIV-1 Nucleocapsid Using Solution NMR, X-Ray Scattering and Protein Engineering ». *Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry* 17 (11): 1548-52. <https://doi.org/10.1002/cphc.201600212>.
- Dingwall, Colin, George P. Lomonosoff, et Ronald A. Laskey. 1981. « High Sequence Specificity of Micrococcal Nuclease ». *Nucleic Acids Research* 9 (12): 2659-74. <https://doi.org/10.1093/nar/9.12.2659>.
- Djikaev, Y. S., et E. Ruckenstein. 2016. « Recent Developments in the Theoretical, Simulational, and Experimental Studies of the Role of Water Hydrogen Bonding in Hydrophobic Phenomena ». *Advances in Colloid and Interface Science* 235 (septembre): 23-45. <https://doi.org/10.1016/j.cis.2016.05.006>.
- Djuranovic, D., et B. Hartmann. 2003. « Conformational Characteristics and Correlations in Crystal Structures of Nucleic Acid Oligonucleotides: Evidence for Sub-States ». *Journal of Biomolecular Structure & Dynamics* 20 (6): 771-88. <https://doi.org/10.1080/07391102.2003.10506894>.
- . 2004. « DNA Fine Structure and Dynamics in Crystals and in Solution: The Impact of BI/BII Backbone Conformations ». *Biopolymers* 73 (3): 356-68. <https://doi.org/10.1002/bip.10528>.
- Djuranovic, D., C. Oguey, et B. Hartmann. 2004. « The Role of DNA Structure and Dynamics in the Recognition of Bovine Papillomavirus E2 Protein Target Sequences ». *Journal of Molecular Biology* 339 (4): 785-96. <https://doi.org/10.1016/j.jmb.2004.03.078>.
- Dorfman, T., J. Luban, S. P. Goff, W. A. Haseltine, et H. G. Göttlinger. 1993. « Mapping of Functionally Important Residues of a Cysteine-Histidine Box in the Human Immunodeficiency Virus Type 1 Nucleocapsid Protein ». *Journal of Virology* 67 (10): 6159-69.
- Doyle, Michael, et Michael F. Jantsch. 2002. « New and Old Roles of the Double-Stranded RNA-Binding Domain ». *Journal of Structural Biology* 140 (1-3): 147-53.
- Draizen, Eli J., Alexey K. Shaytan, Leonardo Mariño-Ramírez, Paul B. Talbert, David Landsman, et Anna R. Panchenko. 2016. « HistoneDB 2.0: A Histone Database with Variants--an

- Integrated Resource to Explore Histones and Their Variants ». *Database: The Journal of Biological Databases and Curation* 2016. <https://doi.org/10.1093/database/baw014>.
- Drew, H. R., et C. R. Calladine. 1987. « Sequence-Specific Positioning of Core Histones on an 860 Base-Pair DNA. Experiment and Theory ». *Journal of Molecular Biology* 195 (1): 143-73.
- Dupuis, Franck, Jean-François Sadoc, Rémi Jullien, Borislav Angelov, et Jean-Paul Mornon. 2005. « Voro3D: 3D Voronoi Tessellations Applied to Protein Structures ». *Bioinformatics (Oxford, England)* 21 (8): 1715-16. <https://doi.org/10.1093/bioinformatics/bth365>.
- Eaton, Bruce E., Larry Gold, et Dominic A. Zichi. 1995. « Let's get specific: the relationship between specificity and affinity ». *Chemistry & Biology* 2 (10): 633-38. [https://doi.org/10.1016/1074-5521\(95\)90023-3](https://doi.org/10.1016/1074-5521(95)90023-3).
- Egelé, Caroline, Emmanuel Schaub, Nick Ramalanjaona, Etienne Piémont, Damien Ficheux, Bernard Roques, Jean-Luc Darlix, et Yves Mély. 2004. « HIV-1 Nucleocapsid Protein Binds to the Viral DNA Initiation Sequences and Chaperones their Kissing Interactions ». *Journal of Molecular Biology* 342 (2): 453-66. <https://doi.org/10.1016/j.jmb.2004.07.059>.
- Erler, Jochen, Ruihan Zhang, Loukas Petridis, Xiaolin Cheng, Jeremy C. Smith, et Jörg Langowski. 2014. « The Role of Histone Tails in the Nucleosome: A Computational Study ». *Biophysical Journal* 107 (12): 2911-22. <https://doi.org/10.1016/j.bpj.2014.10.065>.
- Esque, Jeremy, Christophe Oguey, et Alexandre G. de Brevern. 2010. « A Novel Evaluation of Residue and Protein Volumes by Means of Laguerre Tessellation ». *Journal of Chemical Information and Modeling* 50 (5): 947-60. <https://doi.org/10.1021/ci9004892>.
- . 2011. « Comparative Analysis of Threshold and Tessellation Methods for Determining Protein Contacts ». *Journal of Chemical Information and Modeling* 51 (2): 493-507. <https://doi.org/10.1021/ci100195t>.
- Feller, Scott E., Yuhong Zhang, Richard W. Pastor, et Bernard R. Brooks. 1995. « Constant pressure molecular dynamics simulation: The Langevin piston method ». *The Journal of Chemical Physics* 103 (11): 4613-21. <https://doi.org/10.1063/1.470648>.
- Felsenfeld, Gary, et Mark Groudine. 2003. « Controlling the Double Helix ». *Nature* 421 (6921): 448-53. <https://doi.org/10.1038/nature01411>.
- Ferré-D'Amaré, Adrian R. 2016. « RNA binding: getting specific about specificity ». *Cell chemical biology* 23 (10): 1177-78. <https://doi.org/10.1016/j.chembiol.2016.10.001>.
- Finch, J. T., L. C. Lutter, D. Rhodes, R. S. Brown, B. Rushton, M. Levitt, et A. Klug. 1977. « Structure of Nucleosome Core Particles of Chromatin ». *Nature* 269 (5623): 29-36. <https://doi.org/10.1038/269029a0>.
- Fisher, R. J., A. Rein, M. Fivash, M. A. Urbaneja, J. R. Casas-Finet, M. Medaglia, et L. E. Henderson. 1998. « Sequence-Specific Binding of Human Immunodeficiency Virus Type 1 Nucleocapsid Protein to Short Oligonucleotides ». *Journal of Virology* 72 (3): 1902-9.
- Formosa, Tim. 2012. « The Role of FACT in Making and Breaking Nucleosomes ». *Biochimica Et Biophysica Acta* 1819 (3-4): 247-55. <https://doi.org/10.1016/j.bbagr.2011.07.009>.
- Fratini, A. V., M. L. Kopka, H. R. Drew, et R. E. Dickerson. 1982. « Reversible Bending and Helix Geometry in a B-DNA Dodecamer: CGCGAATTBrCGCG ». *The Journal of Biological Chemistry* 257 (24): 14686-707.
- Fujii, Satoshi, Hidetoshi Kono, Shigeori Takenaka, Nobuhiro Go, et Akinori Sarai. 2007. « Sequence-Dependent DNA Deformability Studied Using Molecular Dynamics Simulations ». *Nucleic Acids Research* 35 (18): 6063-74. <https://doi.org/10.1093/nar/gkm627>.
- Gansen, Alexander, Suren Felekyan, Ralf Kühnemuth, Kathrin Lehmann, Katalin Tóth, Claus A. M. Seidel, et Jörg Langowski. 2018. « High Precision FRET Studies Reveal Reversible Transitions in Nucleosomes between Microseconds and Minutes ». *Nature Communications* 9 (1): 4628. <https://doi.org/10.1038/s41467-018-06758-1>.
- Gansen, Alexander, Alessandro Valeri, Florian Hauger, Suren Felekyan, Stanislav Kalinin, Katalin Tóth, Jörg Langowski, et Claus A. M. Seidel. 2009. « Nucleosome Disassembly Intermediates Characterized by Single-Molecule FRET ». *Proceedings of the National*



- Academy of Sciences of the United States of America* 106 (36): 15308-13.  
<https://doi.org/10.1073/pnas.0903005106>.
- Gao, Mu, et Jeffrey Skolnick. 2009. « From Nonspecific DNA–Protein Encounter Complexes to the Prediction of DNA–Protein Interactions ». *PLOS Computational Biology* 5 (3): e1000341.  
<https://doi.org/10.1371/journal.pcbi.1000341>.
- Gehring, K., J. L. Leroy, et M. Guéron. 1993. « A Tetrameric DNA Structure with Protonated Cytosine.Cytosine Base Pairs ». *Nature* 363 (6429): 561-65.  
<https://doi.org/10.1038/363561a0>.
- Genheden, Samuel, et Ulf Ryde. 2015. « The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities ». *Expert Opinion on Drug Discovery* 10 (5): 449-61.  
<https://doi.org/10.1517/17460441.2015.1032936>.
- Godet, Julien, Hugues de Rocquigny, Chinnappan Raja, Nicole Glasser, Damien Ficheux, Jean-Luc Darlix, et Yves Mély. 2006. « During the Early Phase of HIV-1 DNA Synthesis, Nucleocapsid Protein Directs Hybridization of the TAR Complementary Sequences via the Ends of Their Double-Stranded Stem ». *Journal of Molecular Biology* 356 (5): 1180-92.  
<https://doi.org/10.1016/j.jmb.2005.12.038>.
- Gorelick, R. J., T. D. Gagliardi, W. J. Bosche, T. A. Wiltrot, L. V. Coren, D. J. Chabot, J. D. Lifson, L. E. Henderson, et L. O. Arthur. 1999. « Strict Conservation of the Retroviral Nucleocapsid Protein Zinc Finger Is Strongly Influenced by Its Role in Viral Infection Processes: Characterization of HIV-1 Particles Containing Mutant Nucleocapsid Zinc-Coordinating Sequences ». *Virology* 256 (1): 92-104.  
<https://doi.org/10.1006/viro.1999.9629>.
- Hagerman, Thomas A., Qiang Fu, Benoit Molinié, James Denvir, Stuart Lindsay, et Philippe T. Georgel. 2009. « Chromatin Stability at Low Concentration Depends on Histone Octamer Saturation Levels ». *Biophysical Journal* 96 (5): 1944-51.  
<https://doi.org/10.1016/j.bpj.2008.10.070>.
- Hall, Michael A., Alla Shundrovsky, Lu Bai, Robert M. Fulbright, John T. Lis, et Michelle D. Wang. 2009. « High-Resolution Dynamic Mapping of Histone-DNA Interactions in a Nucleosome ». *Nature Structural & Molecular Biology* 16 (2): 124-29.  
<https://doi.org/10.1038/nsmb.1526>.
- Hargittai, Michele R. S., Robert J. Gorelick, Ioulia Rouzina, et Karin Musier-Forsyth. 2004. « Mechanistic Insights into the Kinetics of HIV-1 Nucleocapsid Protein-Facilitated TRNA Annealing to the Primer Binding Site ». *Journal of Molecular Biology* 337 (4): 951-68.  
<https://doi.org/10.1016/j.jmb.2004.01.054>.
- Harrison, G. P., G. Miele, E. Hunter, et A. M. Lever. 1998. « Functional Analysis of the Core Human Immunodeficiency Virus Type 1 Packaging Signal in a Permissive Cell Line ». *Journal of Virology* 72 (7): 5886-96.
- Hartmann, Brigitte, Michael R. Sullivan, et Lester F. Harris. 2003. « Operator Recognition by the Phage 434 CI Repressor: MD Simulations of Free and Bound 50-Bp DNA Reveal Important Differences between the OR1 and OR2 Sites ». *Biopolymers* 68 (2): 250-64.  
<https://doi.org/10.1002/bip.10243>.
- Hatakeyama, Akiko, Brigitte Hartmann, Andrew Travers, Claude Nogues, et Malcolm Buckle. 2016. « High-Resolution Biophysical Analysis of the Dynamics of Nucleosome Formation ». *Scientific Reports* 6: 27337. <https://doi.org/10.1038/srep27337>.
- Heddi, Brahim, Josephine Abi-Ghanem, Marc Lavigne, et Brigitte Hartmann. 2010. « Sequence-Dependent DNA Flexibility Mediates DNase I Cleavage ». *Journal of Molecular Biology* 395 (1): 123-33. <https://doi.org/10.1016/j.jmb.2009.10.023>.
- Heddi, Brahim, Nicolas Foloppe, Christophe Oguey, et Brigitte Hartmann. 2008. « Importance of Accurate DNA Structures in Solution: The Jun-Fos Model ». *Journal of Molecular Biology* 382 (4): 956-70. <https://doi.org/10.1016/j.jmb.2008.07.047>.
- Heddi, Brahim, Christophe Oguey, Christophe Lavelle, Nicolas Foloppe, et Brigitte Hartmann. 2010. « Intrinsic Flexibility of B-DNA: The Experimental TRX Scale ». *Nucleic Acids*

- Research* 38 (3): 1034-47. <https://doi.org/10.1093/nar/gkp962>.
- Hemelaar, Joris. 2012. « The Origin and Diversity of the HIV-1 Pandemic ». *Trends in Molecular Medicine* 18 (3): 182-92. <https://doi.org/10.1016/j.molmed.2011.12.001>.
- Herbert, Alan. 2019. « Z-DNA and Z-RNA in Human Disease ». *Communications Biology* 2 (1): 1-10. <https://doi.org/10.1038/s42003-018-0237-x>.
- Hippel, P H von, et O G Berg. 1986. « On the specificity of DNA-protein interactions. » *Proceedings of the National Academy of Sciences of the United States of America* 83 (6): 1608-12.
- Hoch, Duane A., Jessica J. Stratton, et Lisa M. Gloss. 2007. « Protein-Protein Förster Resonance Energy Transfer Analysis of Nucleosome Core Particles Containing H2A and H2A.Z ». *Journal of Molecular Biology* 371 (4): 971-88. <https://doi.org/10.1016/j.jmb.2007.05.075>.
- Holde, Kensal E. van. 1989. *Chromatin*. Springer Science & Business Media.
- Houzet, Laurent, Jean Christophe Paillart, Fatima Smagulova, Stephan Maurel, Zakia Morichaud, Roland Marquet, et Marylène Mougel. 2007. « HIV Controls the Selective Packaging of Genomic, Spliced Viral and Cellular RNAs into Virions through Different Mechanisms ». *Nucleic Acids Research* 35 (8): 2695-2704. <https://doi.org/10.1093/nar/gkm153>.
- Hu, Wei-Shau, et Stephen H. Hughes. 2012. « HIV-1 Reverse Transcription ». *Cold Spring Harbor Perspectives in Medicine* 2 (10). <https://doi.org/10.1101/cshperspect.a006882>.
- Huan, J., W. Wang, A. Washington, J. Prins, R. Shah, et A. Tropsha. 2004. « Accurate Classification of Protein Structural Families Using Coherent Subgraph Analysis ». *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 411-22.
- Iilina, Tatiana, Krystal LaBarge, Stefan G. Sarafianos, Rieko Ishima, et Michael A. Parniak. 2012. « Inhibitors of HIV-1 Reverse Transcriptase—Associated Ribonuclease H Activity ». *Biology* 1 (3): 521-41. <https://doi.org/10.3390/biology1030521>.
- Iraci, Nunzio, Oriana Tabarrini, Claudio Santi, et Luca Sancineto. 2018. « NCp7: Targeting a Multitask Protein for next-Generation Anti-HIV Drug Development Part 2. Noncovalent Inhibitors and Nucleic Acid Binders ». *Drug Discovery Today* 23 (3): 687-95. <https://doi.org/10.1016/j.drudis.2018.01.022>.
- Iwahara, Junji, Charles D. Schwieters, et G. Marius Clore. 2004. « Characterization of Nonspecific Protein–DNA Interactions by 1H Paramagnetic Relaxation Enhancement ». *Journal of the American Chemical Society* 126 (40): 12800-808. <https://doi.org/10.1021/ja046246b>.
- Iwahara, Junji, Markus Zweckstetter, et G. Marius Clore. 2006. « NMR Structural and Kinetic Characterization of a Homeodomain Diffusing and Hopping on Nonspecific DNA ». *Proceedings of the National Academy of Sciences* 103 (41): 15062-67. <https://doi.org/10.1073/pnas.0605868103>.
- Jankowsky, Eckhard, et Michael E. Harris. 2015. « Specificity and non-specificity in RNA–protein interactions ». *Nature reviews. Molecular cell biology* 16 (9): 533-44. <https://doi.org/10.1038/nrm4032>.
- Jenuwein, T., et C. D. Allis. 2001. « Translating the Histone Code ». *Science (New York, N.Y.)* 293 (5532): 1074-80. <https://doi.org/10.1126/science.1063127>.
- Jeong, Euna, Hyunwoo Kim, Seong-Wook Lee, et Kyungsook Han. 2003. « Discovering the Interaction Propensities of Amino Acids and Nucleotides from Protein-RNA Complexes ». *Molecules and Cells* 16 (2): 161-67.
- Jiang, Cizhong, et B. Franklin Pugh. 2009. « Nucleosome Positioning and Gene Regulation: Advances through Genomics ». *Nature Reviews Genetics* 10 (3): 161-72. <https://doi.org/10.1038/nrg2522>.
- Jones, Susan, Paul van Heyningen, Helen M. Berman, et Janet M. Thornton. 1999. « Protein-DNA interactions: a structural analysis 11 Edited by K. Nagai ». *Journal of Molecular Biology* 287 (5): 877-96. <https://doi.org/10.1006/jmbi.1999.2659>.
- Jorgensen, William L., et Jeffry D. Madura. 1983. « Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water ». *Journal of the American Chemical Society* 105 (6): 1407-13. <https://doi.org/10.1021/ja00344a001>.

- Kalodimos, Charalampos G., Nikolaos Biris, Alexandre M. J. J. Bonvin, Marc M. Levandoski, Marc Guennegues, Rolf Boelens, et Robert Kaptein. 2004. « Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes ». *Science* 305 (5682): 386-89. <https://doi.org/10.1126/science.1097064>.
- Kaplan, Noam, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, et al. 2009. « The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome ». *Nature* 458 (7236): 362-66. <https://doi.org/10.1038/nature07667>.
- Kenan, Daniel J., Charles C. Query, et Jack D. Keene. 1991. « RNA recognition: towards identifying determinants of specificity ». *Trends in Biochemical Sciences* 16 (janvier): 214-20. [https://doi.org/10.1016/0968-0004\(91\)90088-D](https://doi.org/10.1016/0968-0004(91)90088-D).
- Kent, Nicholas A., Steffan Adams, Alex Moorhouse, et Konrad Paszkiewicz. 2011. « Chromatin Particle Spectrum Analysis: A Method for Comparative Chromatin Structure Analysis Using Paired-End Mode next-Generation DNA Sequencing ». *Nucleic Acids Research* 39 (5): e26-e26. <https://doi.org/10.1093/nar/gkq1183>.
- Kiefer, Christine M., Chunhui Hou, Jane Little, et Ann Dean. 2008. « Epigenetics of  $\beta$ -globin gene regulation ». *Mutation research* 647 (1-2): 68-76. <https://doi.org/10.1016/j.mrfmmm.2008.07.014>.
- Kobayashi, N., et N. Go. 1997. « A Method to Search for Similar Protein Local Structures at Ligand Binding Sites and Its Application to Adenine Recognition ». *European Biophysics Journal: EBJ* 26 (2): 135-44.
- Krishnamoorthy, G., Bernard Roques, Jean-Luc Darlix, et Yves Mély. 2003. « DNA Condensation by the Nucleocapsid Protein of HIV-1: A Mechanism Ensuring DNA Protection ». *Nucleic Acids Research* 31 (18): 5425-32. <https://doi.org/10.1093/nar/gkg738>.
- Lam, W. C., A. H. Maki, J. R. Casas-Finet, J. W. Erickson, B. P. Kane, R. C. Sowder, et L. E. Henderson. 1994. « Phosphorescence and Optically Detected Magnetic Resonance Investigation of the Binding of the Nucleocapsid Protein of the Human Immunodeficiency Virus Type 1 and Related Peptides to RNA ». *Biochemistry* 33 (35): 10693-700. <https://doi.org/10.1021/bi00201a017>.
- Längst, Gernot, et Peter B. Becker. 2004. « Nucleosome remodeling: one mechanism, many phenomena? ». *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression, Chromatin structure and function*, 1677 (1): 58-63. <https://doi.org/10.1016/j.bbaexp.2003.10.011>.
- Lankas, Filip. 2004. « DNA Sequence-Dependent Deformability--Insights from Computer Simulations ». *Biopolymers* 73 (3): 327-39. <https://doi.org/10.1002/bip.10542>.
- Lankaš, Filip, Jiří Šponer, Jörg Langowski, et Thomas E. Cheatham. 2004. « DNA Deformability at the Base Pair Level ». *Journal of the American Chemical Society* 126 (13): 4124-25. <https://doi.org/10.1021/ja0390449>.
- Lantermann, Alexandra B., Tobias Straub, Annelie Strålfors, Guo-Cheng Yuan, Karl Ekwall, et Philipp Korber. 2010. « *Schizosaccharomyces Pombe* Genome-Wide Nucleosome Mapping Reveals Positioning Mechanisms Distinct from Those of *Saccharomyces Cerevisiae* ». *Nature Structural & Molecular Biology* 17 (2): 251-57. <https://doi.org/10.1038/nsmb.1741>.
- Latchman, D. S. 1993. « Transcription factors: an overview. » *International Journal of Experimental Pathology* 74 (5): 417-22.
- Le Guilloux, Vincent, Peter Schmidtke, et Pierre Tuffery. 2009. « Fpocket: An Open Source Platform for Ligand Pocket Detection ». *BMC Bioinformatics* 10 (juin): 168. <https://doi.org/10.1186/1471-2105-10-168>.
- Lee, Min-Ho, et Tim Schedl. 2006. *RNA-Binding Proteins*. WormBook. <https://www.ncbi.nlm.nih.gov/books/NBK19709/>.
- Lee, William, Desiree Tillo, Nicolas Bray, Randall H. Morse, Ronald W. Davis, Timothy R. Hughes, et Corey Nislow. 2007. « A High-Resolution Atlas of Nucleosome Occupancy in Yeast ». *Nature Genetics* 39 (10): 1235-44. <https://doi.org/10.1038/ng2117>.
- Lejeune, Diane, Nicolas Delsaux, Benoît Charloreaux, Annick Thomas, et Robert Bresseur. 2005.

- « Protein–nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure ». *Proteins: Structure, Function, and Bioinformatics* 61 (2): 258-71. <https://doi.org/10.1002/prot.20607>.
- Leprivier, Gabriel, David Baillat, Agnès Begue, Brigitte Hartmann, et Marc Aumercier. 2009. « Ets-1 P51 and P42 Isoforms Differentially Modulate Stromelysin-1 Promoter According to Induced DNA Bend Orientation ». *Nucleic Acids Research* 37 (13): 4341-52. <https://doi.org/10.1093/nar/gkp307>.
- Levin, Judith G., Jianhui Guo, Ioulia Rouzina, et Karin Musier-Forsyth. 2005. « Nucleic Acid Chaperone Activity of HIV-1 Nucleocapsid Protein: Critical Role in Reverse Transcription and Molecular Mechanism ». *Progress in Nucleic Acid Research and Molecular Biology* 80: 217-86. [https://doi.org/10.1016/S0079-6603\(05\)80006-6](https://doi.org/10.1016/S0079-6603(05)80006-6).
- Levin, Judith G., Mithun Mitra, Anjali Mascarenhas, et Karin Musier-Forsyth. 2010. « Role of HIV-1 Nucleocapsid Protein in HIV-1 Reverse Transcription ». *RNA Biology* 7 (6): 754-74.
- Liang, J., H. Edelsbrunner, et C. Woodward. 1998. « Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design ». *Protein Science: A Publication of the Protein Society* 7 (9): 1884-97. <https://doi.org/10.1002/pro.5560070905>.
- Lin, Hsuan-Chun, Jing Zhao, Courtney N. Niland, Brandon Tran, Eckhard Jankowsky, et Michael E. Harris. 2016. « Analysis of the RNA binding specificity landscape of C5 protein reveals structure and sequence preferences that direct ribonuclease P specificity ». *Cell chemical biology* 23 (10): 1271-81. <https://doi.org/10.1016/j.chembiol.2016.09.002>.
- Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, et T. J. Richmond. 1997. « Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution ». *Nature* 389 (6648): 251-60. <https://doi.org/10.1038/38444>.
- Luger, Karolin. 2003. « Structure and Dynamic Behavior of Nucleosomes ». *Current Opinion in Genetics & Development* 13 (2): 127-35.
- Luscombe, Nicholas M., Roman A. Laskowski, et Janet M. Thornton. 2001. « Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level ». *Nucleic Acids Research* 29 (13): 2860-74.
- Mackerell, Alexander D., Michael Feig, et Charles L. Brooks. 2004. « Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations ». *Journal of Computational Chemistry* 25 (11): 1400-1415. <https://doi.org/10.1002/jcc.20065>.
- Maele, Bénédicte Van, Katrien Busschots, Linos Vandekerckhove, Frauke Christ, et Zeger Debyser. 2006. « Cellular Co-Factors of HIV-1 Integration ». *Trends in Biochemical Sciences* 31 (2): 98-105. <https://doi.org/10.1016/j.tibs.2005.12.002>.
- Makde, Ravindra D., Joseph R. England, Hemant P. Yennawar, et Song Tan. 2010. « Structure of RCC1 chromatin factor bound to the nucleosome core particle ». *Nature* 467 (7315): 562-66. <https://doi.org/10.1038/nature09321>.
- Mariño-Ramírez, Leonardo, Maricel G. Kann, Benjamin A. Shoemaker, et David Landsman. 2005. « Histone Structure and Nucleosome Stability ». *Expert Review of Proteomics* 2 (5): 719-29. <https://doi.org/10.1586/14789450.2.5.719>.
- Maris, Christophe, Cyril Dominguez, et Frédéric H.-T. Allain. 2005. « The RNA Recognition Motif, a Plastic RNA-Binding Platform to Regulate Post-Transcriptional Gene Expression ». *The FEBS Journal* 272 (9): 2118-31. <https://doi.org/10.1111/j.1742-4658.2005.04653.x>.
- Matta, Chérif F., Norberto Castillo, et Russell J. Boyd. 2006. « Extended Weak Bonding Interactions in DNA: Pi-Stacking (Base-Base), Base-Backbone, and Backbone-Backbone Interactions ». *The Journal of Physical Chemistry. B* 110 (1): 563-78. <https://doi.org/10.1021/jp054986g>.
- Mattiroli, Francesca, Sheena D'Arcy, et Karolin Luger. 2015. « The Right Place at the Right Time: Chaperoning Core Histone Variants ». *EMBO Reports* 16 (11): 1454-66. <https://doi.org/10.15252/embr.201540840>.

- Mavrigh, Travis N., Cizhong Jiang, Ilya P. Ioshikhes, Xiaoyong Li, Bryan J. Venters, Sara J. Zanton, Lynn P. Tomsho, et al. 2008. « Nucleosome Organization in the *Drosophila* Genome ». *Nature* 453 (7193): 358-62. <https://doi.org/10.1038/nature06929>.
- McGinty, Robert K., et Song Tan. 2015. « Nucleosome Structure and Function ». *Chemical Reviews* 115 (6): 2255-73. <https://doi.org/10.1021/cr500373h>.
- Mély, Y., N. Jullian, N. Morellet, H. De Rocquigny, C. Z. Dong, E. Piémont, B. P. Roques, et D. Gérard. 1994. « Spatial Proximity of the HIV-1 Nucleocapsid Protein Zinc Fingers Investigated by Time-Resolved Fluorescence and Fluorescence Resonance Energy Transfer ». *Biochemistry* 33 (40): 12085-91.
- Mély, Y., E. Piémont, M. Sorinas-Jimeno, H. de Rocquigny, N. Jullian, N. Morellet, B. P. Roques, et D. Gérard. 1993. « Structural and Dynamic Characterization of the Aromatic Amino Acids of the Human Immunodeficiency Virus Type I Nucleocapsid Protein Zinc Fingers and Their Involvement in Heterologous TRNA(Phe) Binding: A Steady-State and Time-Resolved Fluorescence Study ». *Biophysical Journal* 65 (4): 1513-22. [https://doi.org/10.1016/S0006-3495\(93\)81222-0](https://doi.org/10.1016/S0006-3495(93)81222-0).
- Mély, Y., H. de Rocquigny, M. Sorinas-Jimeno, G. Keith, B. P. Roques, R. Marquet, et D. Gérard. 1995. « Binding of the HIV-1 Nucleocapsid Protein to the Primer TRNA(3Lys), in Vitro, Is Essentially Not Specific ». *The Journal of Biological Chemistry* 270 (4): 1650-56. <https://doi.org/10.1074/jbc.270.4.1650>.
- Miyagi, Atsushi, Toshio Ando, et Yuri L. Lyubchenko. 2011. « Dynamics of Nucleosomes Assessed with Time-Lapse High-Speed Atomic Force Microscopy ». *Biochemistry* 50 (37): 7901-8. <https://doi.org/10.1021/bi200946z>.
- Modrow, Susanne, Dietrich Falke, Uwe Truyen, et Hermann Schätzl. 2013. « Viruses with a Single-Stranded DNA Genome ». In *Molecular Virology*, édité par Susanne Modrow, Dietrich Falke, Uwe Truyen, et Hermann Schätzl, 875-918. Berlin, Heidelberg: Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-20718-1\\_20](https://doi.org/10.1007/978-3-642-20718-1_20).
- Morellet, N., H. Déméné, V. Teilleux, T. Huynh-Dinh, H. de Rocquigny, M. C. Fournié-Zaluski, et B. P. Roques. 1998. « Structure of the Complex between the HIV-1 Nucleocapsid Protein NCp7 and the Single-Stranded Pentanucleotide d(ACGCC) ». *Journal of Molecular Biology* 283 (2): 419-34. <https://doi.org/10.1006/jmbi.1998.2098>.
- Morellet, N., N. Jullian, H. De Rocquigny, B. Maigret, J. L. Darlix, et B. P. Roques. 1992. « Determination of the Structure of the Nucleocapsid Protein NCp7 from the Human Immunodeficiency Virus Type 1 by 1H NMR ». *The EMBO Journal* 11 (8): 3059-65.
- Morellet, N., H. de Rocquigny, Y. Mély, N. Jullian, H. Déméné, M. Ottmann, D. Gérard, J. L. Darlix, M. C. Fournie-Zaluski, et B. P. Roques. 1994. « Conformational Behaviour of the Active and Inactive Forms of the Nucleocapsid NCp7 of HIV-1 Studied by 1H NMR ». *Journal of Molecular Biology* 235 (1): 287-301.
- Morozova, N., J. Allers, J. Myers, et Y. Shamoo. 2006. « Protein–RNA Interactions: Exploring Binding Patterns with a Three-Dimensional Superposition Analysis of High Resolution Structures ». *Bioinformatics* 22 (22): 2746-52. <https://doi.org/10.1093/bioinformatics/btl470>.
- Muriaux, D., H. De Rocquigny, B. P. Roques, et J. Paoletti. 1996. « NCp7 Activates HIV-1Lai RNA Dimerization by Converting a Transient Loop-Loop Complex into a Stable Dimer ». *The Journal of Biological Chemistry* 271 (52): 33686-92. <https://doi.org/10.1074/jbc.271.52.33686>.
- Murzin, A G. 1993. « OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. » *The EMBO Journal* 12 (3): 861-67.
- Nadassy, K., S. J. Wodak, et J. Janin. 1999. « Structural Features of Protein-Nucleic Acid Recognition Sites ». *Biochemistry* 38 (7): 1999-2017. <https://doi.org/10.1021/bi982362d>.
- North, Justin A., Marek Šimon, Michelle B. Ferdinand, Matthew A. Shoffner, Jonathan W. Picking, Cecil J. Howard, Alex M. Mooney, John van Noort, Michael G. Poirier, et Jennifer J. Ottesen. 2014. « Histone H3 phosphorylation near the nucleosome dyad alters chromatin structure ». *Nucleic Acids Research* 42 (8): 4922-33. <https://doi.org/10.1093/nar/gku150>.

- Oguey, Christophe, Nicolas Foloppe, et Brigitte Hartmann. 2010. « Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions ». *PLoS One* 5 (12): e15931. <https://doi.org/10.1371/journal.pone.0015931>.
- Omichinski, J. G., G. M. Clore, K. Sakaguchi, E. Appella, et A. M. Gronenborn. 1991. « 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 Letters* 292 (1-2): 25-30.
- Pace, C. Nick, Hailong Fu, Katrina Lee Fryar, John Landua, Saul R. Trevino, Bret A. Shirley, Marsha McNutt Hendricks, et al. 2011. « Contribution of Hydrophobic Interactions to Protein Stability ». *Journal of molecular biology* 408 (3): 514-28. <https://doi.org/10.1016/j.jmb.2011.02.053>.
- Park, Young-Jun, Jayanth V. Chodaparambil, Yunhe Bao, Steven J. McBryant, et Karolin Luger. 2005. « Nucleosome Assembly Protein 1 Exchanges Histone H2A-H2B Dimers and Assists Nucleosome Sliding ». *The Journal of Biological Chemistry* 280 (3): 1817-25. <https://doi.org/10.1074/jbc.M411347200>.
- Parkinson, Gary N., Michael P. H. Lee, et Stephen Neidle. 2002. « Crystal Structure of Parallel Quadruplexes from Human Telomeric DNA ». *Nature* 417 (6891): 876. <https://doi.org/10.1038/nature755>.
- Pasi, Marco, Dimitar Angelov, Jan Bednar, Stefan Dimitrov, et Richard Lavery. 2016. « Extra views on structure and dynamics of DNA loops on nucleosomes studied with molecular simulations ». *Nucleus* 7 (6): 554-59. <https://doi.org/10.1080/19491034.2016.1260800>.
- Percipalle, P., A. Simoncsits, S. Zakhariyev, C. Guarnaccia, R. Sánchez, et S. Pongor. 1995. « Rationally Designed Helix-Turn-Helix Proteins and Their Conformational Changes upon DNA Binding ». *The EMBO Journal* 14 (13): 3200-3205.
- Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, et Klaus Schulten. 2005. « Scalable Molecular Dynamics with NAMD ». *Journal of computational chemistry* 26 (16): 1781-1802. <https://doi.org/10.1002/jcc.20289>.
- Poljak, L., S. M. Batson, D. Ficheux, B. P. Roques, J. L. Darlix, et E. Kas. 2003. « Analysis of NCp7 dependent activation of HIV-1 cDNA integration and its conservation among retroviral nucleocapsid proteins ». *Journal of Molecular Biology* 3: 411-21.
- Poupon, Anne. 2004. « Voronoi and Voronoi-Related Tessellations in Studies of Protein Structure and Interaction ». *Current Opinion in Structural Biology* 14 (2): 233-41. <https://doi.org/10.1016/j.sbi.2004.03.010>.
- Raveh-Sadka, Tali, Michal Levo, Uri Shabi, Boaz Shany, Leeat Keren, Maya Lotan-Pompan, Danny Zeevi, Eilon Sharon, Adina Weinberger, et Eran Segal. 2012. « Manipulating Nucleosome Disfavoring Sequences Allows Fine-Tune Regulation of Gene Expression in Yeast ». *Nature Genetics* 44 (7): 743-50. <https://doi.org/10.1038/ng.2305>.
- Richmond, T. J., J. T. Finch, B. Rushton, D. Rhodes, et A. Klug. 1984. « Structure of the Nucleosome Core Particle at 7 Å Resolution ». *Nature* 311 (5986): 532. <https://doi.org/10.1038/311532a0>.
- Richmond, T. J., M. A. Searles, et R. T. Simpson. 1988. « Crystals of a Nucleosome Core Particle Containing Defined Sequence DNA ». *Journal of Molecular Biology* 199 (1): 161-70. [https://doi.org/10.1016/0022-2836\(88\)90386-5](https://doi.org/10.1016/0022-2836(88)90386-5).
- Richmond, Timothy J., et Curt A. Davey. 2003. « The Structure of DNA in the Nucleosome Core ». *Nature* 423 (6936): 145-50. <https://doi.org/10.1038/nature01595>.
- Rohs, Remo, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, et Richard S. Mann. 2010. « Origins of specificity in protein-DNA recognition ». *Annual Review of Biochemistry* 79: 233-69. <https://doi.org/10.1146/annurev-biochem-060408-091030>.
- Rohs, Remo, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, et Barry Honig. 2009. « The role of DNA shape in protein-DNA recognition ». *Nature* 461 (7268): 1248-53. <https://doi.org/10.1038/nature08473>.

- Salazar, M., O. Y. Fedoroff, J. M. Miller, N. S. Ribeiro, et B. R. Reid. 1993. « The DNA Strand in DNA.RNA Hybrid Duplexes Is Neither B-Form nor A-Form in Solution ». *Biochemistry* 32 (16): 4207-15. <https://doi.org/10.1021/bi00067a007>.
- Sancineto, Luca, Nunzio Iraci, Oriana Tabarrini, et Claudio Santi. 2018. « NCp7: Targeting a Multitasking Protein for next-Generation Anti-HIV Drug Development Part 1: Covalent Inhibitors ». *Drug Discovery Today* 23 (2): 260-71. <https://doi.org/10.1016/j.drudis.2017.10.017>.
- Saunders, Laura R., et Glen N. Barber. 2003. « The DsRNA Binding Protein Family: Critical Roles, Diverse Cellular Functions ». *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 17 (9): 961-83. <https://doi.org/10.1096/fj.02-0958rev>.
- Schneider, B., S. Neidle, et H. M. Berman. 1997. « Conformations of the Sugar-Phosphate Backbone in Helical DNA Crystal Structures ». *Biopolymers* 42 (1): 113-24.
- Schones, Dustin E., Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, et Keji Zhao. 2008. « Dynamic Regulation of Nucleosome Positioning in the Human Genome ». *Cell* 132 (5): 887-98. <https://doi.org/10.1016/j.cell.2008.02.022>.
- Seeman, N. C., J. M. Rosenberg, et A. Rich. 1976. « Sequence-Specific Recognition of Double Helical Nucleic Acids by Proteins ». *Proceedings of the National Academy of Sciences of the United States of America* 73 (3): 804-8. <https://doi.org/10.1073/pnas.73.3.804>.
- Shereda, R. D., A. G. Kozlov, T. M. Lohman, M. M. Cox, et J. L. Keck. 2008. « SSB as an Organizer/Mobilizer of Genome Maintenance Complexes. » *Critical Reviews in Biochemistry and Molecular Biology* 43 (5): 289-318. <https://doi.org/10.1080/10409230802341296>.
- Silva, Micael, Angelo Miguel Figueiredo, et Eurico J. Cabrita. 2014. « Epitope Mapping of Imidazolium Cations in Ionic Liquid-Protein Interactions Unveils the Balance between Hydrophobicity and Electrostatics towards Protein Destabilisation ». *Physical Chemistry Chemical Physics: PCCP* 16 (42): 23394-403. <https://doi.org/10.1039/c4cp03534h>.
- Sleiman, Dona, Valérie Goldschmidt, Pierre Barraud, Roland Marquet, Jean-Christophe Paillart, et Carine Tisné. 2012. « Initiation of HIV-1 Reverse Transcription and Functional Role of Nucleocapsid-Mediated tRNA/Viral Genome Interactions ». *Virus Research* 169 (2): 324-39. <https://doi.org/10.1016/j.virusres.2012.06.006>.
- Smith, Adam J. T., Xiyun Zhang, Andrew G. Leach, et K. N. Houk. 2009. « Beyond Picomolar Affinities: Quantitative Aspects of Noncovalent and Covalent Binding of Drugs to Proteins ». *Journal of medicinal chemistry* 52 (2): 225-33. <https://doi.org/10.1021/jm800498e>.
- Struhl, Kevin, et Eran Segal. 2013. « Determinants of nucleosome positioning ». *Nature structural & molecular biology* 20 (3): 267-73. <https://doi.org/10.1038/nsmb.2506>.
- Summers, M. F., L. E. Henderson, M. R. Chance, J. W. Bess, T. L. South, P. R. Blake, I. Sagi, G. Perez-Alvarado, R. C. Sowder, et D. R. Hare. 1992. « Nucleocapsid Zinc Fingers Detected in Retroviruses: EXAFS Studies of Intact Viruses and the Solution-State Structure of the Nucleocapsid Protein from HIV-1 ». *Protein Science: A Publication of the Protein Society* 1 (5): 563-74. <https://doi.org/10.1002/pro.5560010502>.
- Talbert, Paul B., et Steven Henikoff. 2014. « Environmental Responses Mediated by Histone Variants ». *Trends in Cell Biology* 24 (11): 642-50. <https://doi.org/10.1016/j.tcb.2014.07.006>.
- Tanchou, V., C. Gabus, V. Rogemond, et J. L. Darlix. 1995. « Formation of Stable and Functional HIV-1 Nucleoprotein Complexes in Vitro ». *Journal of Molecular Biology* 252 (5): 563-71. <https://doi.org/10.1006/jmbi.1995.0520>.
- Taylor, Todd, Margarita Rivera, Glenda Wilson, et Iosif I. Vaisman. 2005. « New Method for Protein Secondary Structure Assignment Based on a Simple Topological Descriptor ». *Proteins* 60 (3): 513-24. <https://doi.org/10.1002/prot.20471>.
- Taylor, W. R. 1999. « Protein Structural Domain Identification ». *Protein Engineering* 12 (3):

- Thåström, A., P. T. Lowary, H. R. Widlund, H. Cao, M. Kubista, et J. Widom. 1999. « Sequence Motifs and Free Energies of Selected Natural and Non-Natural Nucleosome Positioning DNA Sequences ». *Journal of Molecular Biology* 288 (2): 213-29. <https://doi.org/10.1006/jmbi.1999.2686>.
- Theobald, Douglas L., Rachel M. Mitton-Fry, et Deborah S. Wuttke. 2003. « Nucleic Acid Recognition by OB-Fold Proteins ». *Annual Review of Biophysics and Biomolecular Structure* 32 (1): 115-33. <https://doi.org/10.1146/annurev.biophys.32.110601.142506>.
- Thomas, J. O. 2001. « HMG1 and 2: Architectural DNA-Binding Proteins ». *Biochemical Society Transactions* 29 (4): 395-401. <https://doi.org/10.1042/bst0290395>.
- Thomas, James A., Tracy D. Gagliardi, W. Gregory Alvord, Mariusz Lubomirski, William J. Bosche, et Robert J. Gorelick. 2006. « Human Immunodeficiency Virus Type 1 Nucleocapsid Zinc-Finger Mutations Cause Defects in Reverse Transcription and Integration ». *Virology* 353 (1): 41-51. <https://doi.org/10.1016/j.virol.2006.05.014>.
- Tisné, C., M. Delepierre, et B. Hartmann. 1999. « How NF-KappaB Can Be Attracted by Its Cognate DNA ». *Journal of Molecular Biology* 293 (1): 139-50. <https://doi.org/10.1006/jmbi.1999.3157>.
- Travers, A. 2001. « DNA-Binding Proteins ». In *Encyclopedia of Genetics*, édité par Sydney Brenner et Jefferey H. Miller, 541-44. New York: Academic Press. <https://doi.org/10.1006/rwgn.2001.0348>.
- Travers, Andrew, Edwige Hiriart, Mark Churcher, Micaela Caserta, et Ernesto Di Mauro. 2010. « The DNA Sequence-Dependence of Nucleosome Positioning in Vivo and in Vitro ». *Journal of Biomolecular Structure & Dynamics* 27 (6): 713-24. <https://doi.org/10.1080/073911010010524942>.
- Tyler, Jessica K. 2002. « Chromatin Assembly. Cooperation between Histone Chaperones and ATP-Dependent Nucleosome Remodeling Machines ». *European Journal of Biochemistry* 269 (9): 2268-74. <https://doi.org/10.1046/j.1432-1033.2002.02890.x>.
- Ucci, Jason W., et James L. Cole. 2004. « Global Analysis of Nonspecific Protein-Nucleic Interactions by Sedimentation Equilibrium ». *Biophysical chemistry* 108 (1-3): 127-40. <https://doi.org/10.1016/j.bpc.2003.10.033>.
- Urbaneja, M. A., B. P. Kane, D. G. Johnson, R. J. Gorelick, L. E. Henderson, et J. R. Casas-Finet. 1999. « Binding Properties of the Human Immunodeficiency Virus Type 1 Nucleocapsid Protein P7 to a Model RNA: Elucidation of the Structural Determinants for Function ». *Journal of Molecular Biology* 287 (1): 59-75. <https://doi.org/10.1006/jmbi.1998.2521>.
- Valieva, M. E., A. V. Feofanov, et V. M. Studitsky. 2016. « Histone Chaperones: Variety and Functions ». *Moscow University Biological Sciences Bulletin* 71 (3): 165-69. <https://doi.org/10.3103/S0096392516030123>.
- Valouev, Anton, Steven M. Johnson, Scott D. Boyd, Cheryl L. Smith, Andrew Z. Fire, et Arend Sidow. 2011. « Determinants of nucleosome organization in primary human cells ». *Nature* 474 (7352): 516-20. <https://doi.org/10.1038/nature10002>.
- Vandevenne, Marylène, David A. Jacques, Crisbel Artuz, Cuong Dinh Nguyen, Ann H. Y. Kwan, David J. Segal, Jacqueline M. Matthews, Merlin Crossley, J. Mitchell Guss, et Joel P. Mackay. 2013. « New Insights into DNA Recognition by Zinc Fingers Revealed by Structural Analysis of the Oncoprotein ZNF217 ». *The Journal of Biological Chemistry* 288 (15): 10616-27. <https://doi.org/10.1074/jbc.M112.441451>.
- Venkatesh, Swaminathan, et Jerry L. Workman. 2015. « Histone Exchange, Chromatin Structure and the Regulation of Transcription ». *Nature Reviews. Molecular Cell Biology* 16 (3): 178-89. <https://doi.org/10.1038/nrm3941>.
- Vignali, M., A. H. Hassan, K. E. Neely, et J. L. Workman. 2000. « ATP-Dependent Chromatin-Remodeling Complexes ». *Molecular and Cellular Biology* 20 (6): 1899-1910. <https://doi.org/10.1128/mcb.20.6.1899-1910.2000>.
- Vogel, Maartje J., Ludo Pagie, Wendy Talhout, Marja Nieuwland, Ron M. Kerkhoven, et Bas van



- Steensel. 2009. « High-resolution mapping of heterochromatin redistribution in a *Drosophila* position-effect variegation model ». *Epigenetics & Chromatin* 2 (1): 1. <https://doi.org/10.1186/1756-8935-2-1>.
- Volokh, O. I., N. I. Derkacheva, V. M. Studitsky, et O. S. Sokolova. 2016. « Structural Studies of Chromatin Remodeling Factors ». *Molecular Biology* 50 (6): 812-22. <https://doi.org/10.1134/S0026893316060212>.
- Wako, H., et T. Yamato. 1998. « Novel Method to Detect a Motif of Local Structures in Different Protein Conformations ». *Protein Engineering* 11 (11): 981-90.
- Wang, Difei, Nikolai B. Ulyanov, et Victor B. Zhurkin. 2010. « Sequence-Dependent Kink-and-Slide Deformations of Nucleosomal DNA Facilitated by Histone Arginines Bound in the Minor Groove ». *Journal of Biomolecular Structure & Dynamics* 27 (6): 843-59. <https://doi.org/10.1080/07391102.2010.10508586>.
- Weiner, Assaf, Amanda Hughes, Moran Yassour, Oliver J. Rando, et Nir Friedman. 2010. « High-Resolution Nucleosome Mapping Reveals Transcription-Dependent Promoter Packaging ». *Genome Research* 20 (1): 90-100. <https://doi.org/10.1101/gr.098509.109>.
- West, Sean M., Remo Rohs, Richard S. Mann, et Barry Honig. 2010. « Electrostatic Interactions between Arginines and the Minor Groove in the Nucleosome ». *Journal of Biomolecular Structure & Dynamics* 27 (6): 861-66. <https://doi.org/10.1080/07391102.2010.10508587>.
- Wickner, R B. 1996. « Double-stranded RNA viruses of *Saccharomyces cerevisiae*. » *Microbiological Reviews* 60 (1): 250-65.
- Wiechens, Nicola, Vijender Singh, Triantafyllos Gkikopoulos, Pieta Schofield, Sonia Rocha, et Tom Owen-Hughes. 2016. « The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes Adjacent to CTCF and Other Transcription Factors ». *PLoS Genetics* 12 (3): e1005940. <https://doi.org/10.1371/journal.pgen.1005940>.
- Wilkinson, Kevin A., Robert J. Gorelick, Suzy M. Vasa, Nicolas Guex, Alan Rein, David H. Mathews, Morgan C. Giddings, et Kevin M. Weeks. 2008. « High-Throughput SHAPE Analysis Reveals Structures in HIV-1 Genomic RNA Strongly Conserved across Distinct Biological States ». *PLoS Biology* 6 (4): e96. <https://doi.org/10.1371/journal.pbio.0060096>.
- Winston, F., et M. Carlson. 1992. « Yeast SNF/SWI Transcriptional Activators and the SPT/SIN Chromatin Connection ». *Trends in Genetics: TIG* 8 (11): 387-91.
- Wintjens, René, et Marianne Rooman. 1996. « Structural Classification of HTH DNA-binding Domains and Protein – DNA Interaction Modes ». *Journal of Molecular Biology* 262 (2): 294-313. <https://doi.org/10.1006/jmbi.1996.0514>.
- Woodcock, Christopher L., et Rajarshi P. Ghosh. 2010. « Chromatin Higher-Order Structure and Dynamics ». *Cold Spring Harbor Perspectives in Biology* 2 (5): a000596. <https://doi.org/10.1101/cshperspect.a000596>.
- Worcel, A., S. Han, et M. L. Wong. 1978. « Assembly of Newly Replicated Chromatin ». *Cell* 15 (3): 969-77. [https://doi.org/10.1016/0092-8674\(78\)90280-5](https://doi.org/10.1016/0092-8674(78)90280-5).
- Wu, Bin, Kareem Mohideen, Dileep Vasudevan, et Curt A. Davey. 2010. « Structural Insight into the Sequence Dependence of Nucleosome Positioning ». *Structure* 18 (4): 528-36. <https://doi.org/10.1016/j.str.2010.01.015>.
- Xu, Fei, et Wilma K. Olson. 2010. « DNA Architecture, Deformability, and Nucleosome Positioning ». *Journal of Biomolecular Structure & Dynamics* 27 (6): 725-39. <https://doi.org/10.1080/073911010010524943>.
- Xu, Xiaoqian, Akli Ben Imeddourene, Loussiné Zargarian, Nicolas Foloppe, Olivier Mauffret, et Brigitte Hartmann. 2014. « NMR Studies of DNA Support the Role of Pre-Existing Minor Groove Variations in Nucleosome Indirect Readout ». *Biochemistry* 53 (35): 5601-12. <https://doi.org/10.1021/bi500504y>.
- Yan, Zhiqiang, Xiliang Zheng, Erkang Wang, et Jin Wang. 2013. « Thermodynamic and Kinetic Specificities of Ligand Binding ». *Chemical Science* 4 (6): 2387-95. <https://doi.org/10.1039/C3SC50478F>.
- Yesudhas, Dhanusha, Maria Batoool, Muhammad Ayaz Anwar, Suresh Panneerselvam, et Sangdun

- Choi. 2017. « Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors ». *Genes* 8 (8). <https://doi.org/10.3390/genes8080192>.
- Zahler, A. M., J. R. Williamson, T. R. Cech, et D. M. Prescott. 1991. « Inhibition of Telomerase by G-Quartet DNA Structures ». *Nature* 350 (6320): 718-20. <https://doi.org/10.1038/350718a0>.
- Zargarian, Loussiné, Igor Kanevsky, Ali Bazzi, Jonathan Boynard, Françoise Chaminade, Philippe Fossé, et Olivier Mauffret. 2009. « Structural and Dynamic Characterization of the Upper Part of the HIV-1 CTAR DNA Hairpin ». *Nucleic Acids Research* 37 (12): 4043-54. <https://doi.org/10.1093/nar/gkp297>.
- Zargarian, Loussiné, Carine Tisné, Pierre Barraud, Xiaoqian Xu, Nelly Morellet, Brigitte René, Yves Mély, Philippe Fossé, et Olivier Mauffret. 2014. « Dynamics of Linker Residues Modulate the Nucleic Acid Binding Properties of the HIV-1 Nucleocapsid Protein Zinc Fingers ». *PLoS One* 9 (7): e102150. <https://doi.org/10.1371/journal.pone.0102150>.
- Zeraati, Mahdi, David B. Langley, Peter Schofield, Aaron L. Moye, Romain Rouet, William E. Hughes, Tracy M. Bryan, Marcel E. Dinger, et Daniel Christ. 2018. « I-Motif DNA Structures Are Formed in the Nuclei of Human Cells ». *Nature Chemistry* 10 (6): 631-37. <https://doi.org/10.1038/s41557-018-0046-3>.
- Zhang, Y., et E. Barklis. 1995. « Nucleocapsid Protein Effects on the Specificity of Retrovirus RNA Encapsulation ». *Journal of Virology* 69 (9): 5716-22.
- Zhang, Yong, Zarmik Moqtaderi, Barbara P. Rattner, Ghia Euskirchen, Michael Snyder, James T. Kadonaga, X. Shirley Liu, et Kevin Struhl. 2009. « Intrinsic Histone-DNA Interactions Are Not the Major Determinant of Nucleosome Positions in Vivo ». *Nature Structural & Molecular Biology* 16 (8): 847-52. <https://doi.org/10.1038/nsmb.1636>.
- Zhao, Yingming, et Benjamin A. Garcia. 2015. « Comprehensive Catalog of Currently Documented Histone Modifications ». *Cold Spring Harbor Perspectives in Biology* 7 (9): a025064. <https://doi.org/10.1101/cshperspect.a025064>.
- Zheng, Chunyang, et Jeffrey J. Hayes. 2003. « Structures and Interactions of the Core Histone Tail Domains ». *Biopolymers* 68 (4): 539-46. <https://doi.org/10.1002/bip.10303>.
- Zhimulev, I. F., E. S. Belyaeva, O. V. Fomina, M. O. Protopopov, et V. N. Bolshakov. 1986. « Cytogenetic and Molecular Aspects of Position Effect Variegation in *Drosophila melanogaster* ». *Chromosoma* 94 (6): 492-504. <https://doi.org/10.1007/BF00292759>.
- Zhurkin, V. B., Y. P. Lysov, et V. I. Ivanov. 1979. « Anisotropic Flexibility of DNA and the Nucleosomal Structure ». *Nucleic Acids Research* 6 (3): 1081-96.

**Titre :** Interactions acides nucléiques/protéines non spécifiques : le nucléosome et les complexes de la NCp7

**Mots clés :** Interface acides nucléiques/protéine, Nucléosome, NCp7, Dynamique Moléculaire

**Résumé :** Les protéines régulent et exécutent l'ensemble des fonctions vitales des organismes en interagissant notamment avec les acides nucléiques (AN), dont l'ADN, support de l'information génétique. Appréhender la nature de ces types d'interactions est central en biologie.

Le nucléosome, qui est l'unité élémentaire de la compaction de l'ADN chez les cellules eucaryotes, est formé d'un d'ADN enroulé autour d'un cœur protéique d'histone ; il contrôle l'accessibilité de l'ADN en se formant et en se dissociant le long des génomes. Ici, le nucléosome a été modélisé par dynamique moléculaire en solution. L'analyse de l'interface ADN-histone par une méthode géométrique innovante a permis de comprendre comment la forte cohésion de ce complexe était assurée. La description de l'interface a aussi servi à interpréter des expériences d'assemblage et de

désassemblage du nucléosome qui ont par ailleurs démontré l'effet de la séquence d'ADN sur ces processus. Enfin, j'ai comparé les dynamiques de l'ADN nucléosomal et de l'ADN nu, et montré quelles propriétés structurales étaient conservées au sein du nucléosome et comment elles sont utilisées pour moduler l'efficacité de l'association ADN-histones.

Une stratégie semblable a été appliquée à des structures expérimentales de complexes entre ADN ou ARN et NCp7, une protéine du VIH-1 chaperon des AN. Cette dernière étude propose un mécanisme d'association entre les partenaires sur des bases rationnelles.

Dans ces deux études, je mets en évidence des mécanismes de formation des complexes en plusieurs étapes et j'illustre les préférences de structure et de séquence des AN chez des protéines dites non-spécifiques.

**Title :** Non-specific nucleic acids/protein interactions : nucleosome and NCp7 complexes

**Keywords :** nucleic acids/protein interface, Nucleosome, NCp7, Molecular Dynamics

**Abstract :** Proteins regulate and perform the vital functions of organisms, in particular by interacting with nucleic acids (NA), including DNA which carries the genetic information. Understanding the nature of these interactions is central in biology.

The nucleosome is the basic unit of DNA compaction in eukaryotes. Composed of a DNA wrapped around a histone core, this complex regulates the DNA accessibility by assembling and disassembling along the genome. Here, we carried out molecular dynamic simulations of the nucleosome in solution. The analysis of the DNA-histone interface with an innovative geometrical method highlighted the strong cohesion of the complex. Such an in-depth description of the interface was also used to interpret nucleosome assembly and disassembly experiments. Those experiments emphasized in

particular the DNA sequence effect in both assembly and disassembly processes. Finally, the comparison between nucleosomal and free DNA dynamics showed which structural properties were conserved in the complex and how they contributed to the DNA-histone assembly efficiency.

A similar strategy was used on experimental structures of NCp7, a HIV-1 NA chaperone protein, complexed with either DNA or RNA. The latter analysis suggested a rational basis to describe the mechanism of partner assembly.

In both studies, I evidenced stepwise mechanisms of complex assemblies and I illustrated NA structure and sequence preferences of some so-called non-specific proteins.

