Étude de l'assemblage, de la mécanique et de la dynamique des complexes ADN-protéine impliquant le développement d'un modèle " gros grains "

Loic Éthève

N°d'ordre NNT : 2016 LYSE1242

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 205
(Ecole Doctorale Interdisciplinaire Science-Santé)

Spécialité de doctorat : Aspects moléculaires et cellulaires de la biologie
Discipline : (Eventuellement)

Soutenue publiquement le 01/12/2016, par :
Loïc ETHEVE

Étude de l'assemblage, de la mécanique et de la dynamique des complexes ADN-protéine impliquant le développement d'un modèle « gros grains »

Devant le jury composé de :

Pasquali, Samuela Professeur Université Paris Descartes Rapporteur.e

Guerois, Raphael Chercheur CEA Université Paris Saclay Rapporteur.e

Dejaegere, Annick Professeur Université de Strasbourg Examinateur.rice

Gouet, Patrice Professeur Université Claude Bernard Lyon 1 Examinateur.rice

Lavery, Richard Directeur de Recherche CNRS Université Claude Bernard Lyon 1
Directeur de thèse

Martin, Juliette Chargée de Recherche Université Claude Bernard Lyon 1 Co-directrice de thèse

Remerciements

Je remercie Richard Lavery et Juliette Martin pour la confiance qu'ils m'ont accordée en acceptant d'encadrer ce travail doctoral, pour leurs multiples conseils et pour toutes les heures qu'ils ont consacré à diriger cette recherche. Un grand merci pour les nombreuses et fastidieuses relectures du manuscrit de thèse qui est et sera surement le plus long que je réaliserai.

Merci à Samuela Pasquali et Raphaël Guérois d'avoir accepté de lire et juger mon travail de thèse, à Patrice Gouet qui a présidé ce jury ainsi qu'à Annick Dejaegere pour l'intérêt qu'elle a porté à mon travail en tant qu'examinatrice.

Merci à toute l'équipe BISI : Elisa Frezza, Nicoletta Ceres, Luca Monticelli, Guillaume Launay et Mélanie Garnier qui ont fait que l'ambiance a été sympathique, décontractée et chaleureuse pendant ces trois années. Merci pour ces bons moments partagés autour de pâtisseries, gâteaux, bières et autres mets.

Merci à Marco Pasi, Anna Reymer et Jonathan Barnoud avec qui j'ai pu partager et approfondir mes connaissances en modélisation et en programmation.

Enfin je remercierai mes parents qui m'ont toujours poussé à me dépasser pour toujours aller plus loin et atteindre le titre de Docteur. Merci à mon épouse Séverine, pour son soutien, son écoute et sa patience tout au long de ces trois années de thèse.

Résumé

Les interactions ADN-protéine sont fondamentales dans de nombreux processus biologiques tels que la régulation des gènes et la réparation de l'ADN. Cette thèse est centrée sur l'analyse des propriétés physiques et dynamiques des interfaces ADN-protéine. À partir de l'étude de quatre complexes ADN-protéine, nous avons montré que l'interface ADN-protéine est dynamique et que les ponts salins et liaisons hydrogène se forment et se rompent dans une échelle de temps de l'ordre de la centaine de picosecondes. L'oscillation des chaînes latérales des résidus est dans certains cas capable de moduler la spécificité d'interaction. Nous avons ensuite développé un modèle de protéine gros grains dans le but de décomposer les interactions ADN-protéine en identifiant les facteurs qui modulent la stabilité et la conformation de l'ADN ainsi que les facteurs responsables de la spécificité de reconnaissance ADN-protéine. Notre modèle est adaptable, allant d'un simple volume mimant une protéine à une représentation plus complexe comportant des charges formelles sur les résidus polaires, ou des chaînes latérales à l'échelle atomique dans le cas de résidus clés ayant des comportements particuliers, tels que les cycles aromatiques qui s'intercalent entre les paires de base de l'acide nucléique.

Mots-clés : bio-informatique structurale, interactions ADN-protéine, sélectivité de séquence, lecture directe, lecture indirecte, dynamique moléculaire, enfilage moléculaire, facteurs de transcription

Abstract

DNA-protein interactions are fundamental in many biological processes such as gene regulation and DNA repair. This thesis is focused on an analysis of the physical and dynamic properties of DNA-protein interfaces. In a study of four DNA-protein complexes, we have shown that DNA-protein interfaces are dynamic and that the salt bridges and hydrogen bonds break and reform over a time scale of hundreds of picoseconds. In certain cases, this oscillation of protein side chains is able to modulate interaction specificity. We have also developed a coarse-grain model of proteins in order to deconvolute the nature of protein-DNA interactions, identifying factors that modulate the stability and conformation of DNA and factors responsible for the protein-DNA recognition specificity. The design of our model can be changed from a simple volume mimicking the protein to a more complicated representation by the addition of formal charges on polar residues, or by adding atomic-scale side chains in the case of key residues with more precise behaviors, such as aromatic rings that intercalate between DNA base pairs.

Keywords : structural bioinformatics, DNA-protein interactions, sequence selectivity, direct readout, indirect readout, molecular dynamics, threading, transcription factors

Introduction

En 1869, le biologiste suisse Johann Friedrich Miescher détecte une substance riche en phosphate qu'il nommera nucléine et dont il démontre la présence dans les noyaux des différents types de cellules, émettant l'hypothèse que celle-ci joue un rôle dans la transmission de l'hérédité. Il faudra attendre le début du XXe siècle pour que cette substance soit identifiée et appelée acide désoxyribonucléique, ou ADN. Cette substance est une macromolécule biologique présente dans toutes les cellules animales, végétales, bactériennes ainsi que chez de nombreux virus. Cette molécule est le support quasi universel de l'information génétique (1). Cette information est stockée dans les noyaux eucaryotes sous forme de chromosomes, dans le cytosol chez les procaryotes. La taille du génome varie selon les organismes, de 13 000 paires de bases pour le virus de la grippe, à 3 milliards de paires de bases pour le génome humain. À ce jour, le génome le plus grand connu est celui de la plante *Paris japonica*, qui contient environ 150 milliards de paires de bases soit 50 fois la taille du génome humain.

Dans une cellule, d'autres macromolécules biologiques sont essentielles dans la régulation de l'homéostasie : les protéines. Une protéine est un biopolymère codé par l'ADN au niveau de régions appelées gènes et possède une séquence menant à une structure tridimensionnelle et une fonction spécifique. Les protéines sont impliquées dans plusieurs aspects du fonctionnement de la machinerie cellulaire. Elles ont un rôle dans le métabolisme, notamment par les enzymes qui vont catalyser de nombreuses réactions biochimiques. Elles assurent également un rôle dans le processus de signalisation au sein de la cellule et avec l'extérieur, et également dans la régulation de l'expression génétique grâce à l'intervention de protéines qui vont se lier à l'ADN : les facteurs de transcription.

Ces dernières décennies, les progrès réalisés en biologie structurale ont permis de nombreuses avancées, grâce notamment au développement de méthodes comme la cristallographie aux rayons X et la spectroscopie RMN (Résonance Magnétique Nucléaire) qui ont enrichi les bases de données de structures 3D. L'accumulation de données dans la Protein Data Bank (PDB) (2), ainsi que dans la Nucleic Acid Database (NDB) (3) a révélé que les complexes ADN-protéine possèdent une large gamme de structures complexes. Ces structures ont été classifiées et regroupées en différentes familles selon le motif de reconnaissance utilisé par la protéine pour se lier à l'ADN (4). L 'analyse de ces complexes a amélioré notre compréhension des mécanismes mises en jeu lors de la reconnaissance spécifique entre les deux partenaires. Cependant, notre

compréhension de ces mécanismes reste partielle, puisque les structures ADN-protéine sur lesquelles ont été faites ces observations sont, en générale, statiques. Depuis quelques années, plusieurs groupes s'intéressent à l'aspect dynamique des interfaces ADN-protéine dans le but comprendre les mécanismes qui permettent aux protéines de trouver et de se lier à leur cible de manière spécifique. L'utilisation de la RMN, ainsi que de la modélisation moléculaire permet actuellement de mieux comprendre le processus de recherche spatiale des protéines, et notamment leurs mouvements, une fois liées à l'ADN de façon non spécifiques (5, 6). Ensuite, il faut aussi comprendre comment le bon site liaison est reconnu pour passer à un complexe spécifique.

L'objectif de cette thèse consiste à employer la méthode de la dynamique moléculaire pour étudier différentes familles de complexes ADN-protéine dans le but de comprendre le rôle de la dynamique au niveau des interfaces. Cette étude sera couplée à un protocole de « clustering » et d'enfilage moléculaire qui permettra d'analyser l'impact potentiel de changements dans la conformation des interfaces sur la reconnaissance du site de liaison. En parallèle, un modèle simplifié des protéines (modèle gros grains) sera développé afin de décomposer les facteurs (volume de la protéine, effets électrostatiques, représentation gros grains ou atomique) responsables de la reconnaissance.

Cette thèse sera découpée en cinq chapitres. Les deux premiers chapitres décriront la structure tridimensionnelle des complexes ADN-protéine ainsi que l'état des connaissances actuelles sur les mécanismes de la reconnaissance. Le chapitre trois sera consacré à la méthodologie utilisée pour mener à bien cette thèse. Les chapitres quatre et cinq présenteront les résultats sur la reconnaissance spécifique ADN-protéine de quatre complexes étudiés grâce à la dynamique moléculaire tout atome et aux modèles gros grains.

Chapitre I. Introduction à la structure de l'ADN et des protéines

Les deux premières parties de ce chapitre 1 présentent une description fine de la structure de l'ADN ainsi que des protéines et la seconde partie présente I) l'implication des complexes ADN-protéine au niveau biologique et II) une vue d'ensemble des familles protéiques qui se lient à l'ADN.

I. Structure de l'acide désoxyribonucléique (ADN)

A. Éléments constituant l'ADN

1. Le nucléotide

L'ADN est constitué de deux brins antiparallèles qui forment une double hélice. Chacun de ces brins est composé d'un enchaînement de nucléotides, le nucléotide étant lui-même constitué d'une base azotée liée à un ose ayant un cycle de cinq atomes (quatre atomes de carbone et un oxygène) en position 1' et d'un groupe phosphate (cf. Figure 1). Chaque nucléotide ayant la même structure au niveau de son ose et du groupement phosphate, la différence entre nucléotides s'effectue au niveau de la base azotée (cf. Figure 1).

2. Les bases azotées

Il en existe quatre différentes, l'adénine (A) et la guanine (G) qui sont des purines (molécule azotée hétérocyclique ($C_5H_4N_4$) constituée d'un cycle pyrimidine et d'un cycle imidazole) et la thymine (T) et la cytosine (C) qui sont des pyrimidines (molécule azotée hétérocyclique $C_4H_4N_2$) (cf. Figure 1).

3. Le désoxyribose et le groupe phosphate

L'ose présent dans la molécule d'ADN est appelé β-d-2'-désoxyribose en raison de l'absence du groupement hydroxyle (2'-OH) du sucre ribose remplacé par un atome hydrogène. La formation de liaisons phosphodiester entre les atomes C3' et C5' du désoxyribose et le groupement OH de l'acide phosphorique va permettre la formation du brin de la double hélice.

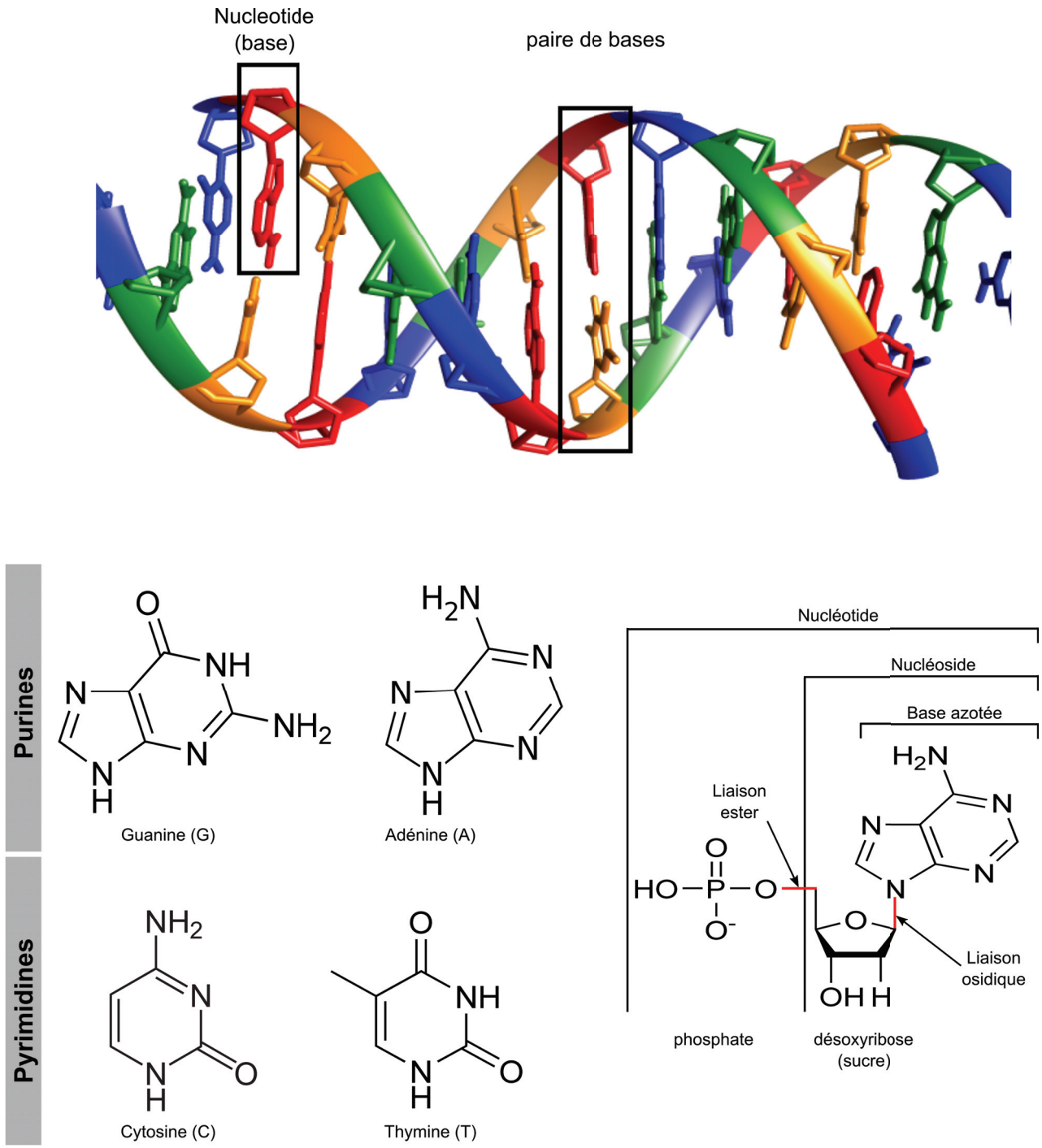


Figure 1 : Représentation d'un fragment d'ADN (succession de GCAT) coloré en fonction du type de base (A = rouge, T =orange, C = bleu et G = vert). La structure de chacune des quatre bases azotées et d'un nucléotide est donnée dans le panneau du bas de la figure.

B. Stabilité de la double hélice

En 1953, l'association par complémentarité de bases A-T et G-C a été postulée par Francis Crick et James Watson lors de la résolution de la structure de l'ADN (7). Cette complémentarité de bases avait également été décrite de manière indirecte dans les travaux

de Erwin Chargaff qui en 1950 (8) avait annoncé que le ratio des nucléotides A/T et C/G était proche de 1.

La double hélice d'ADN est stabilisée par la formation de liaisons hydrogène entre bases d'une même paire et par des interactions d'empilement entre bases successives, et également par l'environnement composé des molécules d'eau et des contre ions. Premièrement, la paire A-T forme deux liaisons hydrogène alors que la paire G-C en forme trois. Dans la majorité des cas, les bases A, T, G et C sont appariées dans la forme dite Watson et Crick (W&C) (cf. Figure 2). Cependant, il est possible de rencontrer un second type d'appariement dit de Hoogsteen (cf. figure 2)(9). Ce type d'appariement est stable lorsque le pH est bas (pH <=5), car l'azote N3 de la cytosine doit être protoné. Des calculs d'énergie libre en dynamique moléculaire réalisés dans un environnement aqueux ont mis en évidence que les liaisons hydrogène pour les paires W&C ont une énergie de -4,3 kcal.mol^{-1} pour la paire A-T et de -5,8 kcal.mol^{-1} pour G-C (10).

Deuxièmement, l'empilement ou « π-π stacking » des cycles aromatiques (de nature hydrophobe) des bases azotées permet de minimiser l'interaction avec les molécules d'eau qui sont présentes dans l'environnement. L'orientation du brin d'ADN et les interactions électrostatiques entre les bases (11) permettent de maximiser la surface d'interaction entre deux paires successives du même brin mais également entre bases du brin Watson et bases du brin Crick. Des calculs *ab initio* de chimie quantique réalisé en milieu *in vacuum* ont montré que l'énergie de d'empilement π-π varie entre -9,5 kcal.mol^{-1} pour le pas dinucléotidique GpG, le moins stable à -13,2 kcal.mol^{-1} pour le pas GpC, le plus stable (12).

La présence de l'eau et des cations dans l'environnement réduit le phénomène de répulsion entre groupements phosphates adjacents. Néanmoins, dans un milieu composé exclusivement d'eau pure, c'est-à-dire sans ions, l'ADN n'est jamais retrouvé sous la forme de double brin, mais sous sa forme simple brin (13, 14).

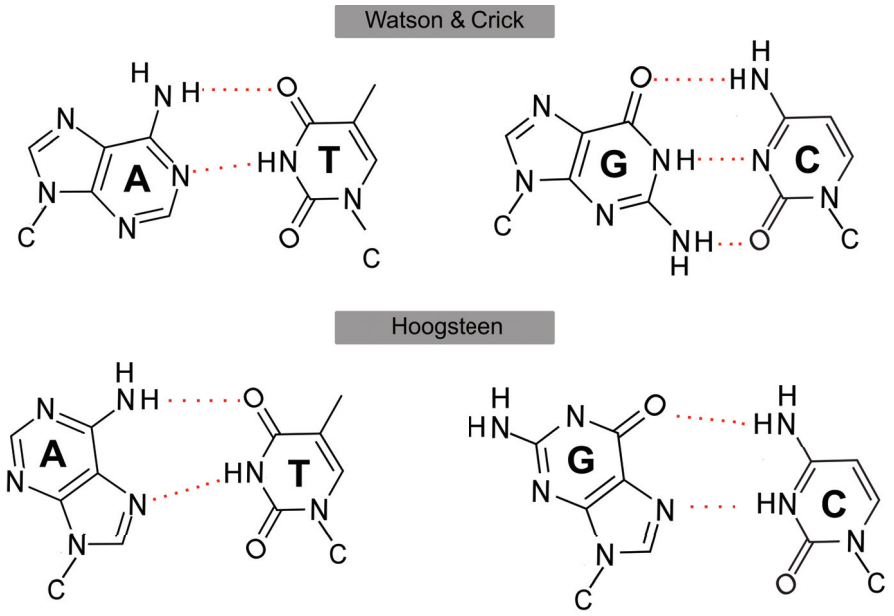


Figure 2 : Appariement des paires de bases selon Watson & Crick et selon Hoogsteen (15).

La stabilité de la molécule d'ADN provient essentiellement du phénomène d'empilement π-π des bases nucléiques (16, 17), car la nature électrostatique des liaisons hydrogène n'est pas significativement stabilisante lorsque la concentration saline dépasse les concentrations physiologiques (18). Ce phénomène d'empilement des bases azotées facilite les changements de conformations locales (grâce à des mouvements de rotations et translations des bases les unes par rapport aux autres) et permet une augmentation de flexibilité de la molécule d'ADN et sera développé dans le chapitre 2.

Afin de former un brin continu d'ADN, chaque nucléotide est relié au nucléotide suivant par une liaison phosphodiester. Le groupement phosphate hautement acide (pKa ~ 1,5) à pH neutre, porte une charge nette de -1, distribuée principalement sur deux oxygènes non ester : OP1 et OP2. Dans des conditions physiologiques, chaque groupement phosphate possède une charge effective comprise entre -0,3 et -0,6, car on estime qu'environ 40 à 80 % de sa charge est contrebalancée par la présence de cations (19, 20).

C. Le brin d'ADN

1. Angles de torsions du brin

On décrit le brin suivant six angles de torsions. L'angle α défini par O3'(i-1)-P-O5'-C5', β par P-O5'-C5'-C4', γ par O5'-C5'-C4'-C3', δ par C5'-C4'-C3'-O3', ε par C4'-C3'-O3'-P(i+1) et ζ par C3'-O3'-P(i+1)-O5'(i+1) (cf. Figure 3).

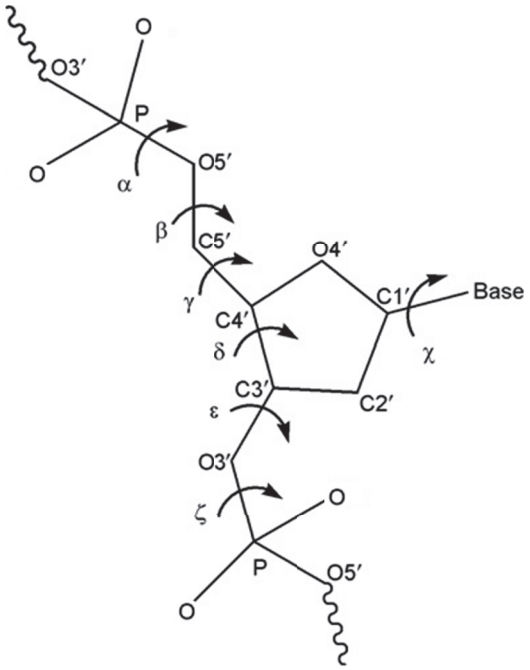


Figure 3 : Définition des angles α, β, γ, δ, ε, ζ et χ sur le brin d'ADN.

L'analyse des angles de torsions α et γ des structures cristallographiques montre une préférence pour la conformation α/γ : g-/g+ (angle de -60° et +60° respectivement) pour des ADN-B (description chapitre 1/D/1). Les conformations α/γ : g-/g+ et α/γ : t/t (angle de 180°) ont été observées au sein des ADN-A (description chapitre 1/D/2) et des ARN (21–23) et des conformations non canoniques ont été observées dans les structures de complexes ADN-protéine (24). Les angles dièdres ε et ζ décrivent deux types de conformations appelées BI : t/g- et BII : g-/t (cf. Figure 4). Dans le cas BI, la différence ε-ζ est négative avec une valeur moyenne de -90° alors que dans la conformation BII, cette différence est positive avec une valeur moyenne de +90°. L'étude menée par Madhumalar et Bansal en 2005 (25) montre que les dinucléotides GpC, CpG, CpA, TpG et TpA adoptent plus fréquemment la conformation BII alors que les dinucléotides TpT, TpC, CpT, CpC n'adoptent que rarement la conformation BII. D'autres études tendent à montrer que la transition BI-BII aurait un rôle essentiel dans plusieurs processus biologiques (26–29).

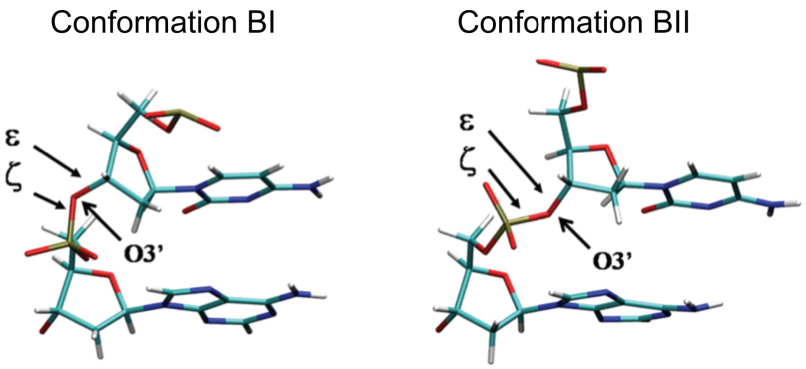


Figure 4 : Représentation des conformations BI et BII (Source : http://mmb.irbbarcelona.org/).

2. *Conformations adoptées par le β-d-2'-désoxyribose*

On décrit la géométrie du désoxyribose selon cinq angles de torsion (cf. Figure 5).

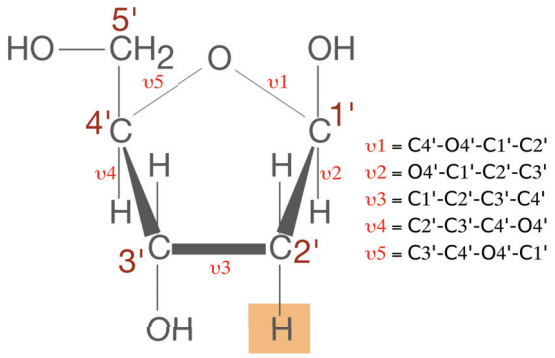


Figure 5 : Angles définissant le désoxyribose

Dans la majorité des cas, le cycle furanique du désoxyribose présente une structure non planaire et à cause des contraintes liées à la forme cyclique du sucre, sa structure peut être déterminée approximativement par deux paramètres. Dans le cas présent, notre β-d-2'-désoxyribose est décrit par deux paramètres que sont l'amplitude (τ) et son angle de phase de pseudorotation (P) décrite dans les travaux de Westhof et Sundaralingam en 1983 (30).

$$\tau = \sqrt{a^2 + b^2}$$

$$P = \cos^{-1}(a/\tau)$$

Ou $a = 0{,}4 \sum_{i=1}^{5} \nu_i \cos[0{,}8\pi(i = 1)]$ et $b = -0{,}4 \sum_{i=1}^{5} \nu_i \sin[0{,}8\pi(i - 1)]$.

Les formes C3'-endo et C2'-endo sont favorisées dans les différentes formes d'ADN physiologiques A et B respectivement. Ces deux conformations ont un impact important sur la structure de l'ADN et contribuent à définir la distance entre deux phosphates successifs du

même brin : 5,9 Å et 7 Å pour la conformation C3'-endo et C2'-endo respectivement (cf. Figure 6).

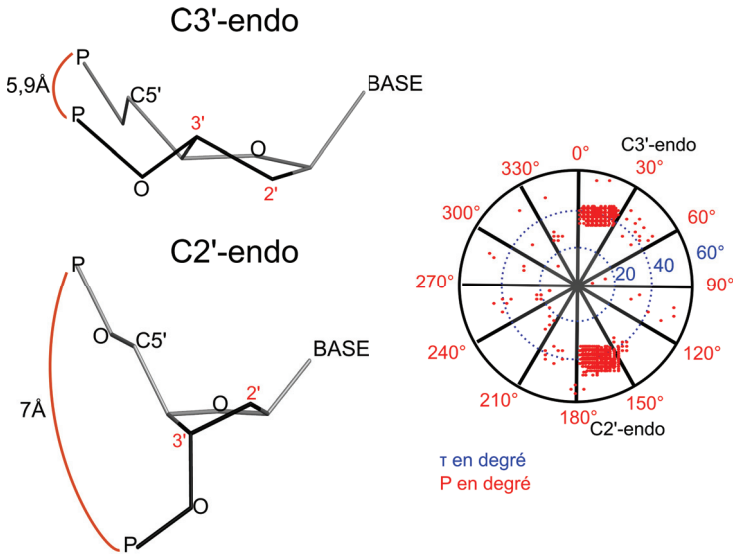


Figure 6 : Conformations principales du désoxyribose. C3'-endo retrouvé pour l'ADN A et C2'-endo pour l'ADN B. Diagramme des valeurs de τ et P retrouvées dans les structures ADN A et B (description chapitre 1/1/E) Adaptation du site : cactus.nci.nih.gov.

Les bases azotées sont reliées au pentose par une liaison glycosidique. Cette liaison glycosidique est définie par les atomes O4'-C1'-N9-C4 (pour les purines) ou par les atomes O4'-C1'-N1-C2 (pour les pyrimidines) et son angle est fortement corrélé aux conformations adoptées par le désoxyribose. La liaison glycosidique peut être de deux natures : soit **syn** si les valeurs d'angles sont comprises entre -90° et +90°, soit **anti** pour des valeurs d'angles comprises entre +90° et +180°. Dans un appariement Watson et Crick, la liaison glycosidique adopte plus fréquemment la conformation **anti**. Dans un ADN type Z (voir description ci-après), la guanine adopte préférentiellement une conformation **syn** alors que la cytosine adopte une conformation **anti**, lui conférant ainsi sa forme particulière dite zigzag.

3. Paramètres hélicoïdaux

La structure des acides nucléiques peut être décrite également par des paramètres géométriques qui définissent la position relative entre deux bases et entre deux paires de bases (pb) consécutives suivant la nomenclature établie lors des accords de Tsukuba (31). Si l'on considère les bases comme des blocs rigides, il faut six paramètres pour décrire le mouvement de deux bases au sein d'une paire de bases, six paramètres pour décrire les mouvements entre deux paires de bases successives et quatre paramètres pour décrire la

position d'une paire de bases par rapport à l'axe hélicoïdal. Nous reviendrons plus en détails sur la définition de ces paramètres au cours de la présentation de l'algorithme Curves+ (cf Chapitre III.II.B) utilisé pour caractériser l'ensemble des doubles hélices d'ADN étudiées durant cette thèse

(a) Les paramètres intra paire de bases

Il existe trois paramètres de rotation, Buckle (κ), Propeller-Twist (ω) et Opening (σ) et trois paramètres de translation Shear (Sx), Stretch (Sy) et Stagger (Sz) (cf. Figure 7). Selon la nomenclature définie par les accords Tsukuba, une valeur différente de 0 pour ces paramètres décrit une déformation dans l'appariement des bases par rapport à un ADN canonique.

(b) Position des bases par rapport à l'axe hélicoïdal

Si on considère une paire de bases comme un seul élément et non plus comme une association de deux bases, on peut mesurer deux paramètres de translation, displacement X (dx) et Y(dy), et deux paramètres de rotation, inclination (η) et tip (θ) (notés paramètres intra pb axe sur la Figure 7).

(c) Les paramètres inter paire de bases

Trois paramètres rotationnels Tilt (τ), Roll (ρ) et Twist (Ω) et trois paramètres translationnels Shift (Dx), Slide (Dy) et Rise (Dz) permettent de décrire la position relative de deux paires de bases successives (cf. Figure 7).

(d) Les paramètres des sillons

Dû à sa forme en double hélice, l'ADN présente deux sillons que l'on a coutume d'appeler grand et petit sillon. L'orientation de certains atomes ou groupements d'atomes permet de localiser la position du grand et du petit sillon. Pour les purines les atomes N7 et C6 pointent en direction du grand sillon alors que pour les pyrimidines ce sont les atomes C4 et C5 qui sont orientés vers le grand sillon. On note également que les liaisons glycosidiques se trouvent dans le petit sillon. Une fois chaque sillon déterminé, on peut obtenir une mesure de leur largeur et de leur profondeur grâce à l'outil Curves+ définie dans le chapitre 3 : Méthodologie.

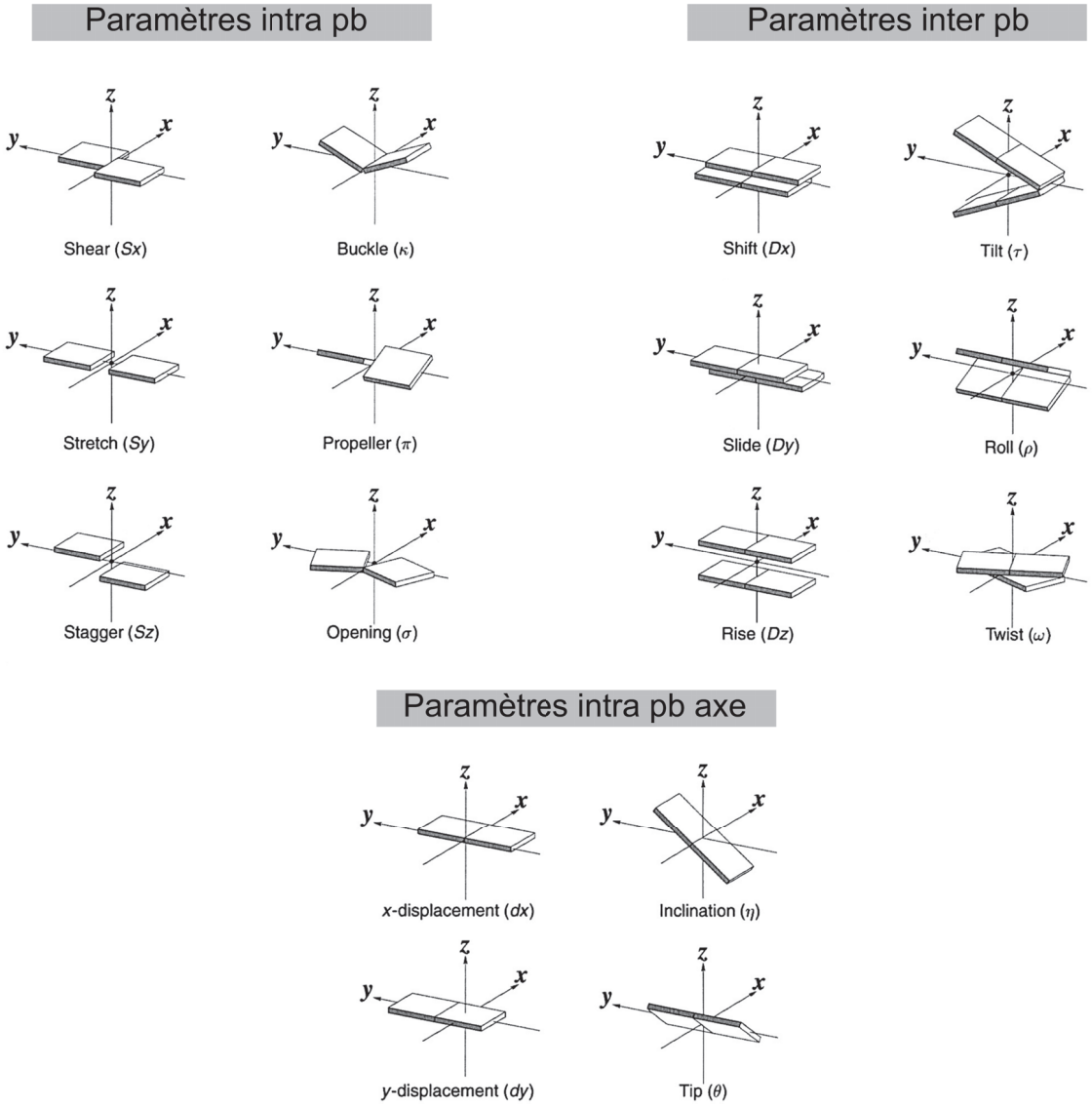


Figure 7 : Représentation schématique des paramètres hélicoïdaux intra et inter paires de bases, position des bases par rapport à l'axe hélicoïdal (intra pb axe) définissant la structure de l'ADN selon la nomenclature de Tsukuba (31). Image adaptée de (32).

D. Polymorphisme de la double hélice

La double hélice est une macromolécule très polymorphe. Sa structure hélicoïdale est capable de subir des changements conformationnels importants (cf. Figure 8). La force ionique, la déshydratation de l'interface de l'ADN ainsi que sa séquence en nucléotides sont des facteurs qui déterminent la structure de la double hélice. A partir des résultats expérimentaux, les formes allomorphes de l'ADN ont pu être catégorisées en trois classes principales B, a et Z. Depuis les premières études aux rayons X des fibres d'ADN, on sait que l'ADN adopte principalement les formes B et A (33, 34). Quelques années plus tard, Wang a mis en évidence une troisième famille, l'ADN Z, cristallisé grâce à une forte concentration ionique et pour des séquences particulières (35) Dans cette section, nous donnerons une

brève description de ces différentes formes d'ADN que l'on retrouve dans une cellule vivante. Depuis d'autres formes ont été caractérisées mais ont un intérêt biologique trop spécifique qui n'est pas présenté dans le cadre de cette thèse.

1. ADN type B

L'ADN de type B est l'ADN le plus communément retrouvé dans les conditions physiologiques (36, 37). C'est une hélice droite qui effectue un tour complet tous les 36 Å soit environ toutes les 10,5 paires de bases. Les paires de bases sont quasi perpendiculaires à l'axe hélicoïdal. L'hélice de type B possède un diamètre de 23 Å et présente un petit sillon étroit (11,7 Å) et profond, un grand sillon large (17 Å) et peu profond. L'hélice de type B présente un motif répété toutes les bases, c'est-à-dire qu'en appliquant des mouvements de rotations et de translations au mononucléotide i, on obtient la position du mononucléotide i+1.

2. ADN type A

Les milieux déshydratés ainsi que des séquences riches en GC favorisent l'apparition de cette alternative structurale de la forme B. L'ADN de type A est une hélice droite de 24 Å de diamètre dont les paires de bases sont déplacées en direction du petit sillon. La forme A diffère essentiellement de la forme B par la position des paires de bases, excentrées par rapport à l'axe hélicoïdal et par l'inclinaison de ces dernières (environ 10 à 20°) (38). En termes de paramètres hélicoïdaux, le Rise et le Twist sont diminués par rapport à la forme B, ce qui se caractérise par la présence d'un grand sillon étroit (11,1 Å) et profond alors que le petit sillon est large (16,7 Å) et peu profond. L'hélice effectue un tour complet toutes les 11 paires de bases et possède un motif répété toutes les paires de bases.

3. ADN type Z

Cette conformation retrouvée *in vivo (39, 40)* est favorisée lorsque la séquence en nucléotide décrit une alternance purine-pyrimidine (exemple poly [GC]) dans des conditions où la concentration saline est extrêmement élevée (2,7 M NaCl ou 700mM $MgCl_2$) (41, 42)) lorsque l'ADN est méthylé (42) et également en présence des protéines type « *Z-DNA binding protein 1* » (43). Contrairement à l'ADN de type A et B, l'ADN de type Z est une hélice gauche de 1,8 nm de diamètre dont les paires de bases sont inclinées de -9° par rapport à l'axe hélicoïdal. Il faut 12 paires de bases pour effectuer un tour complet d'hélice. Dans un ADN Z

les bases sont déplacées du côté du grand sillon lui conférant une forme convexe (bases très exposées au solvant) et chaque base nucléique a effectué une rotation de 180°, c'est-à-dire que la face de la base qui est orientée vers le haut de l'hélice dans un ADN B pointera vers le bas dans un ADN Z. Le petit sillon est très profond alors que le grand sillon présente une forme convexe peu profonde. Le double brin d'ADN adopte une forme dite zigzag lié au fait que le nucléotide guanine adopte une conformation -syn- au niveau de la liaison glycosidique alors que le nucléotide cytosine adopte une conformation -anti-. Contrairement à l'ADN A et B le motif de répété dans un ADN Z est di-nucléotidique, c'est-à-dire qu'en appliquant des rotations et des translations sur le dinucléotide **i** on obtient la position du dinucléotide suivant et qu'au sein d'un motif dinucléotidique les paramètres de translations et rotations seront différents pour les deux nucléotides qui le constituent. Ainsi, au sein d'un motif ADN Z, les deux angles de Twist successifs sont de -9° et de -51° respectivement (44)

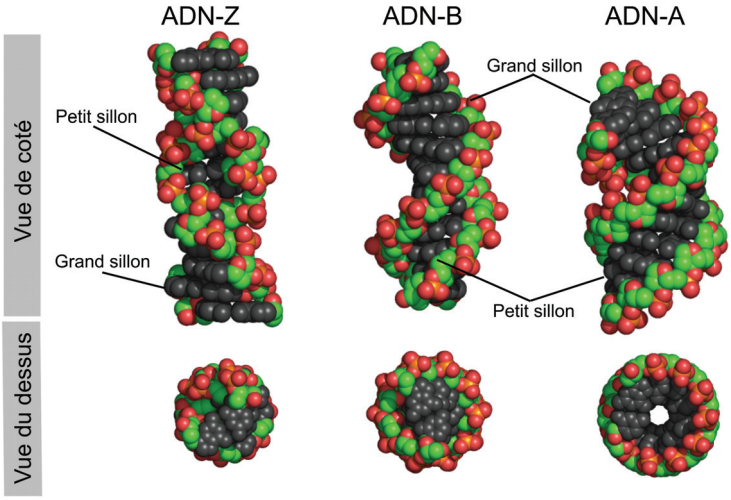


Figure 8 : Représentation en mode sphère des 3 grands types d'ADN. Les bases des différents ADN Z, B et A sont colorées en gris et le brin phosphodiester est en couleur.

II. Structure des protéines

La structure des protéines suit une organisation hiérarchique présentée figure 11 :

- La structure primaire ou séquence de la protéine correspond à la succession linéaire des acides aminés.
- La structure secondaire décrit le repliement local de la protéine en hélice alpha ou brin bêta.
- La structure tertiaire correspond au repliement spatial de la chaîne polypeptidique (structure tridimensionnelle).

- La structure quaternaire correspond à l'association de plusieurs chaînes polypeptidiques pour former un complexe.

A. Structure primaire

Les protéines sont des polymères issus de la traduction des gènes dont la brique élémentaire est l'acide aminé. L'acide aminé se compose d'un carbone asymétrique lié à un hydrogène, un groupement amine (NH2), un groupement carboxyle (COOH) et à une chaîne latérale. La traduction du code génétique conduit à la formation de 20 acides aminés protéinogène standard qui possèdent une chaîne latérale avec des propriétés physico-chimiques uniques. Les acides aminés peuvent être classés en fonction des propriétés physico-chimiques de leur chaîne latérale. En 1986, Taylor propose une classification qui permet de regrouper les acides aminés en fonction de la taille, de l'hydrophobicité ou de la polarité de la chaîne latérale (cf. Figure 9). Cependant, certains acides aminés sont difficiles à catégoriser, car leurs propriétés peuvent changer en fonction de leur environnement. Ainsi à pH = 7, l'azote du cycle imidazole peut être protoné (pKa=6.8) et donc chargé positivement alors que des pH supérieurs à 7 conduisent à une forme non chargée, permettant à l'histidine d'être classée à la fois dans les polaires chargés ou dans les polaires non-chargés. La cystéine est un second exemple de ce type d'ambigüité. Au sein d'une protéine ou lors d'interaction entre deux protéines, la cystéine peut former un pont disulfure (liaison covalente entre deux cystéines). Dans cette configuration, la cystéine n'est plus un acide polaire, mais devient apolaire.

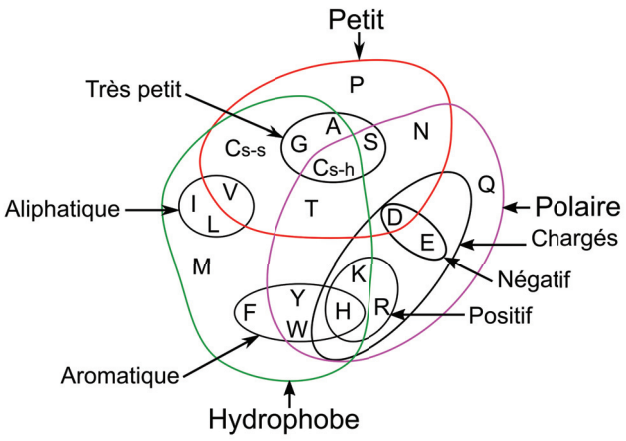


Figure 9 : Diagramme de Taylor indiquant la classification des acides aminé d'après leurs propriétés physico-chimiques. La cystéine peut être sous sa forme appariée Cs-h ou non appariée Cs-s. Les acides aminés sont présentés sous leur code 1 lettre.

Lors de la polymérisation, l'acide aminé forme une liaison covalente, nommée liaison peptidique par réaction entre le groupement carboxyle COOH et le groupement amine NH2 de l'acide aminé suivant de la chaîne peptidique ayant comme produit de réaction, la libération d'une molécule d'eau. La liaison peptidique est stabilisée par mésomérie, c'est à dire une délocalisation des électrons au niveau de la liaison conjuguée, et ne peut donc facilement subir de rotation. Cette propriété rend la liaison généralement planaire. La liaison peptidique peut être *cis* ou *trans*. Afin d'éviter les encombrements stériques, c'est dans la configuration *trans* que l'on retrouve la plus plupart des résidus.

B. Structure secondaire

Lors du repliement local de la chaîne polypeptidique, les angles dièdres phi (ϕ) et psi (φ) permettent de définir la structure secondaire la protéine. L'angle ϕ est défini par les atomes CO_{n-1}-NH_n-$C\alpha_n$-CO_n et l'angle φ par les atomes NH_n-$C\alpha_n$-CO_n-NH_{n+1} (cf. Figure 10). L'encombrement stérique entre chaînes latérales rend énergétiquement défavorables certains couples de valeurs d'angle ϕ/φ. Le biologiste et physicien indien Gopalasamudram Narayana Ramachandran a décrit trois régions d'angles énergétiquement favorables par le couple ϕ/φ en 1963 (45). Parmi ces trois régions, on trouve les couples d'angles ϕ/φ qui correspondent aux structures régulières (les hélices α et les feuillets β) décrites par Pauling en 1951 (46).

.

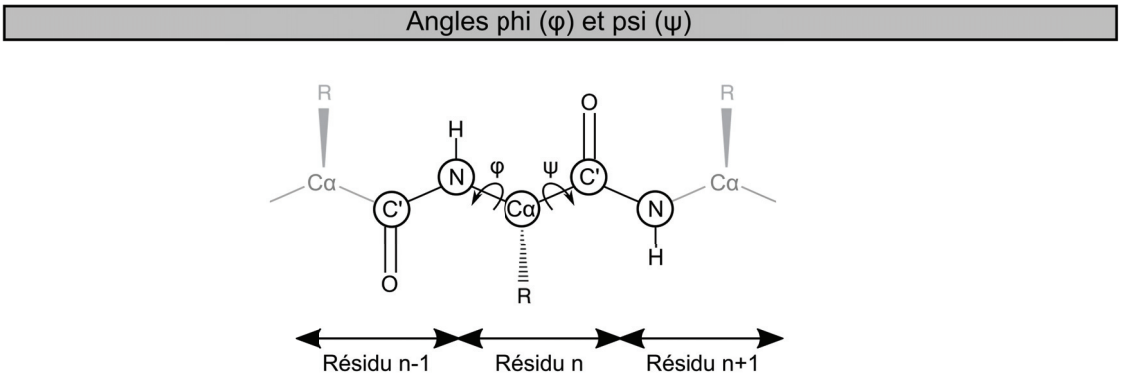


Figure 10 : Représentation schématique des angles ϕ et ψ.

1. L'hélice α

L'hélice α est la structure secondaire la plus abondante dans les protéines. La valeur moyenne des angles ϕ/φ est en moyenne de -57° et -47° respectivement. Il en existe deux types, droite (la plus courante) et gauche. Un tour moyen d'hélice contient 3,6 résidus et se

caractérise par la formation de liaisons hydrogène entre le groupement carbonyle CO d'un résidu *i* et le groupement amide NH d'un résidu *i+4*.

2. Les brins et feuillets β

Le brin β est une structure périodique étendue. La valeur des angles ϕ/φ est en moyenne de -119°/+113° et -139°/+135° pour le feuillet parallèle et antiparallèle respectivement. Dans cette structure, les liaisons hydrogène qui la stabilisent se font entre résidus distants alors que dans le cas de l'hélice α les liaisons hydrogène sont entre résidus proches. L'association de plusieurs brins β forme un feuillet β. L'association de brins β peut suivre deux topologies différentes. Si deux brins sont orientés dans la même direction, on parle de feuillet parallèle. Dans le cas où les deux brins sont orientés dans deux directions différentes, on parle de feuillet antiparallèle. Dans un feuillet parallèle, le groupement amine NH et carbonyle CO d'un résidu *i* du premier brin β forment des liaisons hydrogène avec le groupement carbonyle CO d'un résidu j et amine NH2 d'un résidu *j+2* appartenant au second brin alors que dans un feuillet antiparallèle, les liaisons hydrogène, s'effectuent entre l'amine et le carbonyle du résidu i du premier brin et le résidu j du second brin.

C. Structure tertiaire

La structure tertiaire ou structure tridimensionnelle (3D) d'une protéine, est définie par son repliement dans l'espace et l'association des structures secondaires (hélice et brin) grâce à des éléments moins structurés (coudes, boucles, β-turns) pour former des motifs qui permettront à la protéine d'assurer sa fonction. Différentes méthodes permettent d'obtenir la structure tertiaire d'une protéine. Les deux principales méthodes actuelles sont la cristallographie aux rayons X (~89 % des structures de la PDB) et la spectroscopie par résonance magnétique nucléaire (RMN) (~10 % des structures de la PDB). Une fois la structure connue, celle-ci est déposée dans la PDB.

D. Structure quaternaire

La structure quaternaire correspond à l'association de plusieurs domaines protéiques par des liaisons le plus souvent non covalentes pour former un complexe multi domaines.

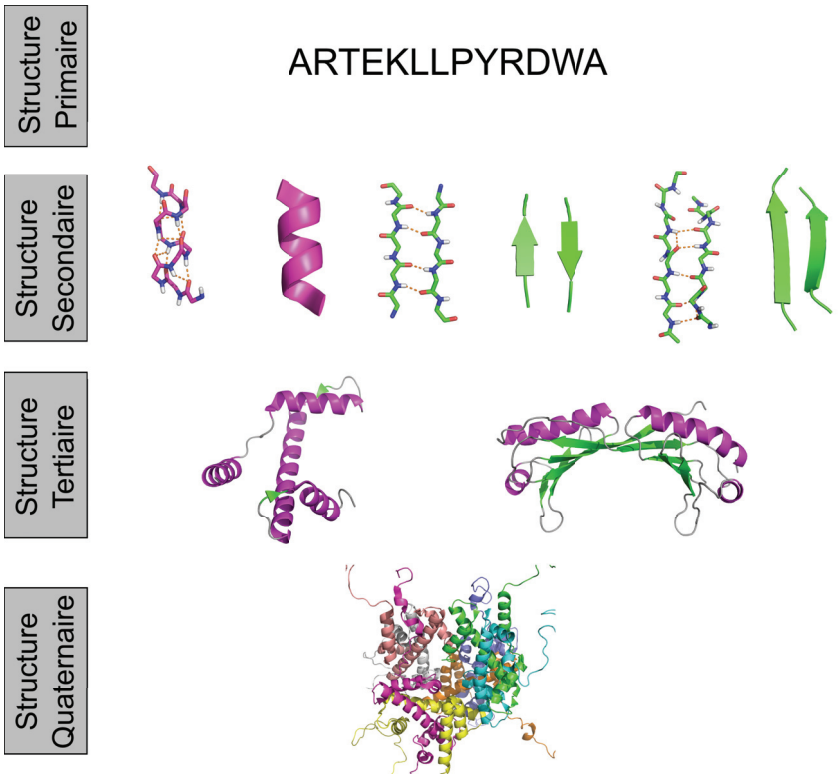


Figure 11 : Organisation structurale des protéines de la structure primaire à la structure quaternaire. La structure quaternaire présente les histones H2A, H2B, H3 etH4 qui forment un complexe impliqué dans la compaction de l'ADN.

E. Forces régulant la stabilité des protéines

La structure 3D d'une protéine est stabilisée par différentes interactions non covalentes et de faibles énergies comme les interactions de van der Waals, les liaisons hydrogène, les ponts salins ou interactions hydrophobes. On retrouve parfois des interactions covalentes grâce aux ponts disulfures formés entre deux cystéines.

1. Les interactions électrostatiques

(a) Forces de van der Waals

Les interactions de van der Waals sont des interactions électrostatiques de faible énergie (typiquement moins de 1 kcal.mol^{-1}) qui existent entre deux atomes à courte distance (de 3 à 4 Å). Au vu de leur nombre important dans les protéines, elles jouent un rôle primordial dans l'attraction et la répulsion entre particules.

(b) Les liaisons hydrogène

Les liaisons hydrogène qui se forment dans le milieu aqueux entre deux acides aminés polaires ou entre atomes du squelette peptidique possèdent une énergie d'environ 1 à 2

kcal.mol^{-1}. Ces interactions peuvent être médiées par des molécules d'eau et donc se faire et se défaire rapidement. Ces interactions sont très importantes dans le maintien des structures secondaires (hélices et brins). Du fait de leur labilité notamment dans les régions exposées au solvant, ces interactions contribuent peu à la stabilité, mais sont néanmoins importantes pour le repliement des protéines (47).

(c) Les ponts salins

Il existe aussi un type d'interaction entre résidus négativement chargés (aspartate, glutamate) et les résidus positivement chargés (arginine, lysine et histidine selon la valeur du pH) nommés ponts salins. Le rôle de ce type d'interaction électrostatique fait débat dans la littérature, car dans certains cas la formation de ponts salins stabilise la structure de la protéine (exemple : lysozyme T4) (48, 49) alors que dans d'autres cas (50, 51) la présence de ponts salins a un effet déstabilisateur.

2. Les interactions hydrophobes

Outre les interactions électrostatiques, les interactions hydrophobes sont un facteur déterminant dans le repliement et la stabilité des protéines. Cette interaction se produit entre résidus apolaires afin de limiter la surface de contact avec l'eau. En général, les résidus apolaires sont enfouis au centre de la protéine et forment le cœur hydrophobe. L'effet hydrophobe est principalement dirigé par l'entropie. Si les molécules hydrophobes sont rassemblées, l'entropie liée aux groupes hydrophobes est minimale, ce qui n'est en principe pas favorable, mais le nombre de molécules d'eau contraintes de s'orienter est nettement inférieur : le rassemblement des molécules hydrophobes permet de maximiser l'entropie de l'eau, qui représente la part majoritaire de l'entropie du système.

3. Exemple de liaison covalente

Dans le cas de certaines protéines, le rapprochement de deux cystéines et leur oxydation permet la formation d'une liaison covalente (S-S). L'énergie libre qui résulte de la formation de cette liaison est d'environ -3,5 kcal.mol^{-1} (52). Cette énergie relativement élevée permet de stabiliser la structure de la protéine.

III. Rôle biologique des interactions ADN-protéine

L'ADN et les protéines forment des interactions afin de réguler différents processus biologiques tout au long de la vie de la cellule. Chez les organismes eucaryotes, l'ADN est retrouvé sous deux formes : l'hétérochromatine et l'euchromatine. L'hétérochromatine est la structure de l'ADN la plus fréquemment rencontrée. Elle nécessite la présence de complexes protéiques (exemple histones) qui sont liés de manière prolongée dans le temps et qui vont permettre entre autre de compacter l'ADN dans le noyau. En revanche, dans le cas de l'euchromatine, l'ADN est moins dense et les gènes sont accessibles afin d'être réparés ou bien transcrits en réponse à un besoin physiologique ou en réponse à un stimulus externe pendant une durée éphémère. De ce fait, on dénombre deux types d'interactions : les permanentes ou semi-permanentes qui vont jouer un rôle dans l'architecture et le stockage de la molécule d'ADN et les interactions éphémères qui vont se produire en réponse à un besoin physiologique de la cellule à un instant donné.

A. Le nucléosome

Dans les années 1970, les premières observations au microscope électronique révèlent l'existence de complexes protéines-ADN, les nucléosomes. Ce complexe ADN-protéine résolu par cristallographie aux rayons X en 1997 (53) est la brique élémentaire de la hétérochromatine (54). Le nucléosome est une structure semi-permanente composée de deux paires de quatre histones H2A, H2B, H3 et H4 autour desquelles s'enroulent environ 147 paires de bases d'ADN sur un tour trois quarts formant ainsi le premier niveau de compaction l'ADN. Les nucléosomes s'organisent ensuite afin de former des niveaux de compaction de plus en plus dense permettant ainsi de stocker un ADN qui dans sa forme déroulée mesurerait plus de deux mètres en un chromosome de 1,4 μm de large pouvant être contenu dans un noyau de 3 à 10 μm de diamètre.

B. Réplication et réparation de l'ADN

Lorsqu'une cellule entre en division, la molécule d'ADN doit être dupliquée afin que chaque cellule fille dispose de la même information que la cellule mère. Ce processus biologique est catalysé par un groupe d'enzymes : le réplisome. Les topoisomérases sont des enzymes qui vont se lier non spécifiquement à la double hélice et le dérouler (55–57) et rendre les bases nucléiques accessibles à d'autres protéines. Les hélicases sont des protéines qui hydrolysent l'adénosine-5'-triphosphate (ATP). Ces protéines vont se fixer à l'ADN, le désenrouler et rompre les liaisons hydrogène entre les paires de bases pour libérer les deux brins (58). Lors de la dernière étape de la réplication, les polymérases se fixent de manière non spécifique à l'ADN et dupliquent les deux brins dans le sens 5' vers 3' (59). Lors de la duplication ou lorsque notre ADN est soumis à des facteurs environnementaux chimiques (radicaux libres, agents alkylants) ou physiques (ultraviolets), notre ADN peut être endommagé. La réparation de l'ADN est rendue possible par l'interaction d'endonucléases (nucléases capables d'hydrolyser une chaîne nucléotidique en son milieu) ou d'exonucléases (nucléases capables d'hydrolyser les nucléotides situés uniquement aux extrémités de l'ADN) qui vont permettre l'excision de base ou de nucléotide. Chez certains organismes, en particulier les bactéries, la formation de complexes ADN-protéine est utilisée comme système de protection contre les génomes étrangers. Cette fonction est assurée par des nucléases de restriction (ou enzymes de restriction), qui vont se fixer sur des régions de l'ADN étranger comportant une séquence spécifique (par exemple : EcoRI (60), EcoRV (61), BamHI (62) coupent respectivement les sites 5'-GAATTC-3', 5'-GATATC-3' et 5'-GGATCC-3').

C. La régulation génétique

Enfin les interactions ADN-protéine permettent de réguler l'expression génétique, c'est-à-dire traduire l'information génétique contenue dans une séquence d'ADN. Ce mécanisme de régulation permet d'augmenter ou de diminuer la quantité d'ARN transcrits en fonction des besoins de la cellule pour que son métabolisme soit en adéquation avec son environnement. Cette régulation est un mécanisme fondamental lors de la différenciation cellulaire. Que ce soit chez les cellules eucaryotes ou procaryotes, cette régulation a lieu par des protéines capables d'activer ou d'inhiber spécifiquement la transcription de certains gènes. On nomme ces protéines des facteurs de transcription.

Chez l'homme on estime le nombre de facteurs de transcription à environ 1800 (63) et chez la levure *Saccharomyces cerevisisae* on dénombre environ 300 facteurs de transcription . Chez l'homme, les protéines impliquées dans les maladies font l'objet d'études intensives (par exemple les protéines p53, ERS1/ERS2 FOS, MYC, JUN). Sur les 1800 facteurs de transcription estimés, environ deux tiers ne sont pas caractérisés et dans le tiers restant, plusieurs protéines ont été caractérisées grâce à des protéines orthologues présentes chez d'autres espèces (63). La levure possède un génome plus petit et la fonction des facteurs de transcription a été très étudiée et chaque site de liaison a été caractérisé expérimentalement (64).

Ces protéines se fixent en amont ou en aval du site d'initiation de la transcription du gène, et ce parfois jusqu'à plusieurs centaines voire de milliers de paires de bases avant ou après le site d'initiation au niveau de séquence dite cisrégulatrice (65–68). En règle générale, les facteurs de transcription qui activent les gènes recrutent le complexe d'initiation de la transcription et l'ARN polymérase qui va permettre la synthèse de l'ARN messager. Plusieurs études ont démontré que des mutations de l'ADN ou des facteurs de transcription, ou des modifications des régions cisrégulatrices sont responsables de maladies telles que l'hémophilie B, la leucémie, la bêta thalassémie ou encore l'épilepsie myoclonique (69–73).

IV. Familles structurales des complexes ADN-protéine

Actuellement la PDB contient près de 107 000 structures de protéines et 5500 structures de complexes protéines-acide nucléiques (ADN et/ou ARN). Dans ces structures l'hélice α est la structure secondaire la plus fréquemment rencontrée. Dans les complexes ADN-protéine, la structure en hélice α permet de s'enfouir aisément dans le grand sillon de l'ADN. Malgré que le grand sillon soit le sillon privilégié pour l'enfouissement des hélices lors de la reconnaissance, certaines protéines comme le répresseur Lac possèdent une structure en hélices capable d'interagir au niveau du petit sillon de l'ADN (74, 75). Malgré que l'hélice alpha soit la structure secondaire la plus observée au niveau des interfaces d'interactions, ces hélices sont souvent associées avec d'autres éléments structuraux formant des motifs. Une première classification des structures des motifs protéiques utilisés dans l'interaction ADN-protéine a été proposée en 1991 par Harrison (76) puis étendue par Luisi (77) et Luscombe (4) et est développée ci-après.

A. Le motif hélix-turn-hélix (HTH)

Le motif HTH (cf. Figure 12) est l'élément de reconnaissance le plus commun parmi les facteurs de transcription procaryotes ou eucaryotes (76–79). Il est formé d'environ 20 acides aminés et se caractérise par deux hélices α séparées par quatre résidus en moyenne qui forment une structure dite β-turn. Une des deux hélices α (l'hélice de reconnaissance) va se lier à l'ADN dans le grand sillon. La seconde hélice peut former des contacts avec le brin d'ADN ou former des interactions médiées par des molécules d'eau afin de stabiliser le complexe comme c'est le cas pour le répresseur tryptophane (80). Ce motif HTH est très conservé structuralement (81) malgré une variabilité importante des séquences et des modes d'interactions. Il existe une extension du motif HTH : winged HTH qui se caractérise par la présence d'une troisième hélice et d'un feuillet β antiparallèle permettant des contacts avec le brin phosphodiester (82–84) (cf. Figure 12).

B. Le motif fermeture (Zipper)

Cette famille se compose de deux sous-familles : les leucines zipper et les motifs helix-loop-helix. Ces protéines ont comme particularité de se dimériser soit en homodimère soit en hétérodimère.

1. Le motif Leucine zipper

Les leucines zipper (cf. Figure 12) se caractérisent par deux grandes hélices α d'environ 60 résidus, qui se dimérisent au niveau d'une région riche en acides aminés hydrophobes (souvent des leucines retrouvées tous les 8 résidus environ). Les deux hélices se lient à l'ADN côté grand sillon dans des directions opposées par la partie basique (riche en acides aminés chargés positivement). Chacune des hélices reconnaît la moitié du site de reconnaissance (85).

2. Le motif Helix-loop-Helix

Les motifs de cette sous-famille sont constitués de deux hélices α séparées par une boucle. La dimérisation s'effectue de la même façon que pour les leucines zipper et l'interaction s'effectue dans le grand sillon de l'ADN (86) (cf. Figure 12).

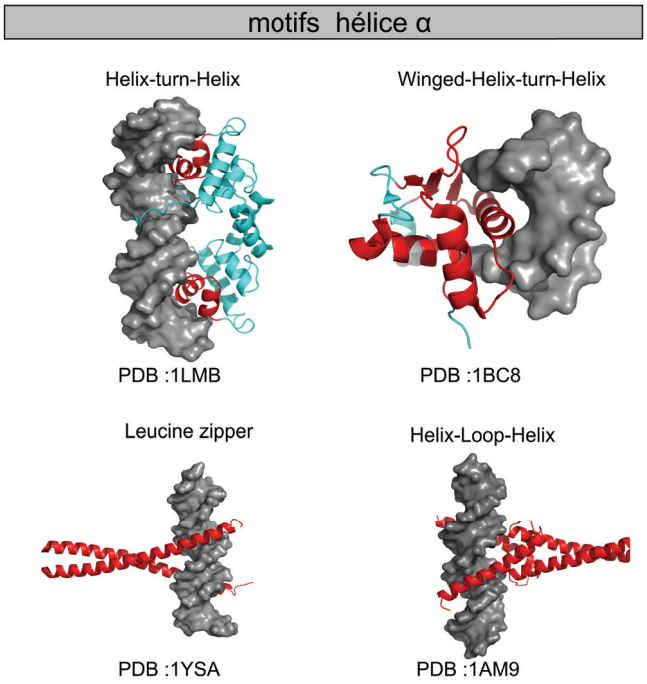


Figure 12 : Motifs alpha les plus rencontrés dans l'interaction ADN-protéine. Les protéines sont représentées en mode cartoon avec le motif d'interaction avec l'ADN coloré en rouge. L'ADN en gris est représenté en surface.

Bien que moins représentées, certaines protéines interagissent avec l'ADN grâce à des brins et feuillets β.

C. Le motif brin β

1. Le motif TBP

Les représentants les plus étudiés de ce groupe sont les protéines TATA (ou TATA-binding protein [TBP]) dont un large feuillet β interagit avec le petit sillon de l'ADN provoquant une importante déformation de la double hélice (87, 88). Ces protéines participent au complexe multiprotéique qui se forme au moment du processus de transcription de l'ADN. Le motif de liaison de la TBP est constitué d'un feuillet antiparallèle de 10 brins (cf. Figure 13). L'interaction de cette protéine forme des coudes au niveau de l'ADN à cause de l'intercalation de deux phénylalanines.

2. Le motif β ruban

Les protéines ayant un motif β ruban se lient à l'ADN grâce à deux ou trois petits brins β (cf. figure 13) (89, 90). Contrairement aux protéines TBP, les protéines possédant un β ruban peuvent former des interactions avec les deux sillons de l'ADN.

3. *Les autres motifs β*

Moins souvent rencontrés, on trouve aussi des protéines Immunoglobulin-like β sandwich tel que le facteur de transcription P53-like (91) ou encore le domaine E-set (92) (cf. Figure 13). Les membres constituant cette famille ont une séquence variable, mais une structure conservée du domaine d'interaction avec l'ADN. Le motif en sandwich β se compose de deux feuillets de brin β antiparallèles. Malgré que le domaine d'interaction soit composé principalement structuré en sandwich β, la reconnaissance ADN-protéine s'effectue grâce aux boucles.

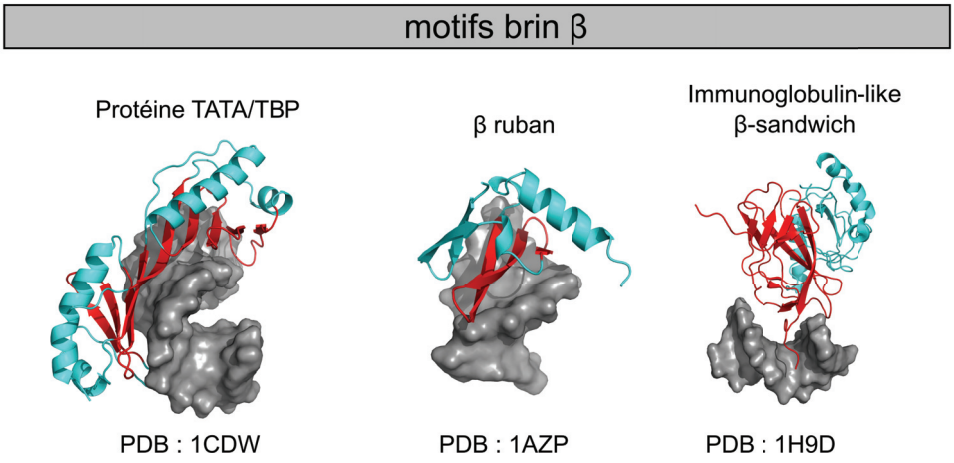


Figure 13 : Motifs bêta les plus rencontrés. Les protéines sont représentées en mode cartoon avec le motif d'interaction coloré en rouge. L'ADN en gris est représenté en surface.

D. Les motifs α et β

Parmi les nombreuses protéines qui interagissent avec l'ADN, certaines possèdent un motif avec des structures α, mais également des structures secondaires en brin/feuillet β.

1. *Le motif doigt de zinc*

Le premier motif doigt de zinc a été mis en évidence en 1983 chez la grenouille *Xenopus laevis* pour le facteur de transcription IIIA (TFIIIA)(93, 94). Le motif décrit chez cette grenouille contient une petite hélice α, deux brins β antiparallèles et un ion zinc (Zn^{2+}) stabilisé par la présence de deux cystéines et deux histidines (76, 77, 94). Ce motif correspond à la séquence : X_2-Cys-$X_{2,4}$-Cys-X_{12}-His-$X_{3,4,5}$-His ou X représente n'importe quel acide aminé et les chiffres le nombre d'acides aminés. Cependant d'autres types de doigts de zinc ont été observés depuis, en fonction de la position relative de l'ion Zn^{2+}. On trouve aussi les doigts de

zinc Cys_4 ou Cys_6 (2 ions Zn^{2+} en interaction avec six cystéines) (95). Lorsque les doigts de zinc sont au nombre de deux ou plus, l'espace séparant deux sites de reconnaissance sur l'ADN est de trois paires de bases en moyenne. Chaque hélice forme des interactions avec environ quatre bases de l'ADN (cf. Figure 14).

2. Le motif Ribbon-Helix-Helix (RHH)

Ce motif est constitué de deux brins antiparallèles suivis de deux hélices α. Dans ce motif, les brins β s'intercalent dans le grand sillon de l'ADN alors que les hélices α permettent la dimérisation de la protéine (cf. figure 14) (96).

E. Cas particulier : les enzymes

Les enzymes forment une famille particulière, en effet toutes les protéines qui possèdent un domaine catalytique et donc une activité enzymatique sont classées dans cette catégorie quelque soit le motif d'interaction avec l'ADN (parmi les motifs α ou motifs β décrit précédemment) (4). Ces protéines ont toutes la particularité d'altérer la structure de l'ADN que ce soit pour le réparer (glycolases, méthyltransférases), le couper (enzymes de restriction) ou encore lors du processus de transcription (hélicases, ligases, topoisomérases). Ces protéines utilisent souvent une combinaison d'hélices α, de brins β et même de boucle pour reconnaître et interagir avec l'ADN.

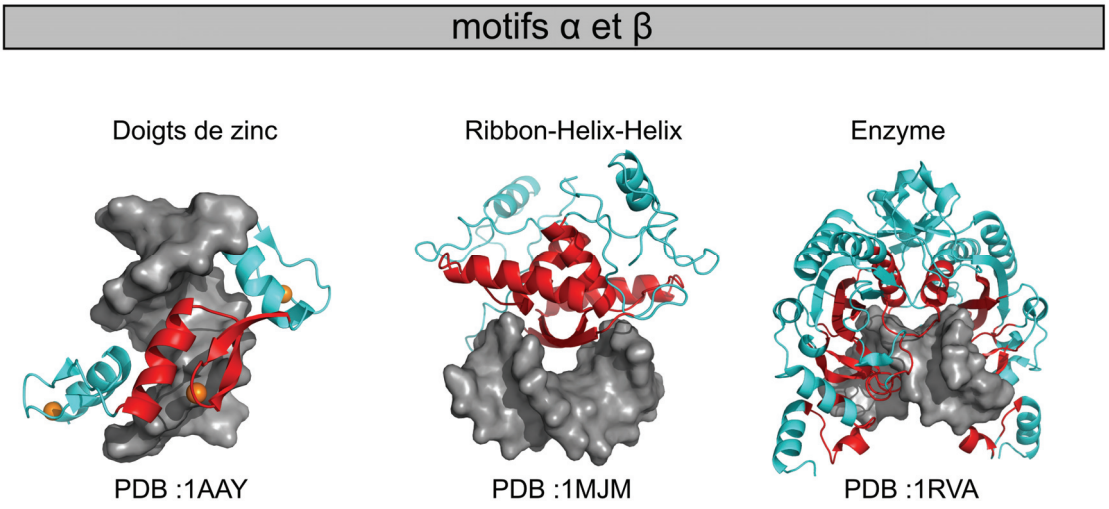


Figure 14 : Structures présentant une composition mixte en hélice α et brins β. Les protéines sont représentées en mode cartoon avec le motif d'interaction coloré en rouge. L'ADN en gris est représenté en surface.

Dans de nombreux complexes, l'interaction ADN-protéine fait intervenir plusieurs domaines identiques ou différents. C'est le cas par exemple de la protéine MarA qui possède deux motifs HTH qui forment des interactions avec le grand sillon de l'ADN (97).

V. Conclusion

Ce chapitre a décrit les bases fondamentales de la structure de l'ADN et des protéines. L'ADN est un macromolécule complexe dont la structure interne des nucléotides lui permet de subir d'importants changements de conformations globales (conformations ADN type A, B ou Z) ou locales par modification des paramètres hélicoïdaux inter et intra paire de bases, par des modifications dans la structure du désoxyribose et dans la flexibilité du brin phosphodiester. Les protéines sont un élément essentiel dans le maintien de l'homéostasie cellulaire. Les interactions protéines-ADN régulent l'activité génétique en réponse à un stress ou stimulus, mais sont également impliquées dans des phénomènes vitaux comme la réparation et la transmission de l'information génétique à la descendance.

L'accumulation de structures dans les bases de données telle que la PDB révèle que l'interface ADN-protéine fait intervenir plusieurs motifs qu'il est difficile de caractériser et qui ne cessent d'évoluer et de s'affiner avec le temps. Cette classification est une première étape dans la caractérisation des mécanismes de reconnaissances employés par les protéines pour interagir avec l'ADN. Le Chapitre 2 va décrire de manières plus fines les méthodes qui permettent de détecter, de prédire et de caractériser les facteurs responsables de la reconnaissance directe (liaison hydrogène, van der Waals, hydrophobe) et indirecte (effet de la séquence, courbure).

Chapitre II. Les mécanismes de la reconnaissance ADN-protéine

L'interaction entre l'ADN et les protéines est au centre de plusieurs processus biologiques. La présence de protéines sur notre matériel génétique permet de le réparer, de le compacter ou de le transcrire. Depuis ces deux dernières décennies, l'accumulation de données expérimentales génomiques ou structurales a montré que les protéines se fixent sur des séquences mesurant en moyenne 5 à 20 paires de bases. Un génome bactérien a une taille de l'ordre de 1 million de paire de bases (Mpb) (*Mycoplasme pneumoniae*) à quelques millions de paires de bases (*Escherichia coli* avec 4,6 Mpb) alors que le génome humain en compte près de 3 milliards. Une des questions qui fait l'objet de nombreuses recherches et investigations est de comprendre comment une protéine est capable d'identifier son site de fixation parmi ces millions voire milliards de possibilités qu'offre le génome.

Tout comme le code qui permet la traduction des acides aminés, on s'attendait à ce que les processus qui régissent l'interaction ADN-protéine suivent un code simple et unique. Or avec l'accumulation de données provenant d'expériences « omiques » ainsi que de la biologie structurale notamment via les données cristallographiques, RMN et de modélisation *in silico*, on sait désormais que plusieurs facteurs vont contribuer à la reconnaissance spécifique au sein des complexes ADN-protéine. La communauté scientifique s'entend à dire qu'il existe deux grands mécanismes responsables de la reconnaissance spécifique : la lecture directe correspond à l'interaction directe entre les acides aminés et les bases de l'ADN (liaison hydrogène et hydrophobes) et la lecture indirecte qui correspond aux variations structurales de l'ADN dépendent de la séquence de bases et qui vont permettre aux protéines d'effectuer de nouveaux contacts.

Le premier connu sous le terme lecture directe a été proposé suite aux premières analyses sur des structures cristallographiques. L'ADN possède des donneurs et accepteurs de liaisons hydrogène et des groupes hydrophobes capables d'être reconnus de manière complémentaire par les chaînes latérales des acides aminés. Cependant avec plus de 1500 complexes résolus à ce jour dans la *Protein Data Bank*, on remarque qu'il n'y a pas de code simple (1 résidu = 1 base) entre les séquences protéiques et nucléiques. Bien que ce mécanisme de lecture directe soit une part importante de la reconnaissance, elle n'est pas suffisante pour expliquer à elle seule la spécificité d'interaction ADN-protéine. Ces dernières années, le terme de lecture indirecte a été proposé comme nouveau mécanisme de reconnaissance. Bien qu'ayant encore une vue fragmenté de ce mécanisme, il est désormais évident que la déformation de l'ADN localement grâce aux modifications des paramètres

hélicoïdaux ou globalement (forme et courbure de l'ADN) permet aux protéines d'identifier leur site de fixation sur le génome (98–100).

Dans ce chapitre, nous aborderons dans un premier temps les méthodes permettant d'identifier, d'analyser et prédire les interactions ADN-protéine, puis les mécanismes qui permettent aux protéines de se déplacer sur le génome afin d'identifier leur site d'interaction et enfin les différents mécanismes de la reconnaissance directe et indirecte seront décrits.

I. Méthodes d'analyses, de détections et de prédictions des interactions ADN-protéine

Depuis le début des années 2000, de nombreuses méthodes *in vivo*, *in vitro* et *in silico* ont été développées afin de caractériser, de déterminer et de prédire la spécificité des complexes ADN-protéine. Ces méthodes ont permis de déterminer les éléments structuraux et physico-chimiques responsables de la reconnaissance. Les méthodes *in silico* permettent d'ajouter un aspect dynamique à cette reconnaissance, mais également d'effectuer des prédictions de reconnaissance grâce à toutes les connaissances acquises ces deux dernières décennies.

Afin de caractériser la façon dont une protéine interagit spécifiquement l'ADN, deux grandes approches de la biologie expérimentale sont utilisées : les méthodes *in vivo* et les méthodes *in vitro*. Les expériences *in vivo* permettent d'obtenir des informations sur les séquences reconnues par les protéines dans un contexte biologique donné alors que les approches *in vitro* cherchent généralement à identifier le ou les sites de fixations (101) ou l'impact de mutations sur la reconnaissance (102). Le choix de la méthode à utiliser dépend du type d'informations que l'on souhaite obtenir ainsi que du nombre de séquences que l'on souhaite analyser, mais également de la quantité de matière biologique à disposition (103).

A. Méthodes de détections des interactions

1. Approches In vivo

Les premières expériences permettant la caractérisation des interactions ADN-protéine étaient réalisées grâce à la méthode EMSA (electrophoretic mobility shift essay) (104, 105). Cette méthode utilise un gel d'électrophorèse sur lequel la protéine d'intérêt est

marquée de manière radioactive et incubée avec un fragment d'ADN. Si la protéine interagit avec l'ADN, la migration sera retardée par rapport à un ADN nu. La caractérisation à l'échelle du génome n'est cependant pas possible avec cette méthode et il faut attendre le développement de méthodes utilisant les puces à ADN (106–110)et à protéines (111–113) pour avoir une caractérisation systématique. Actuellement une méthode est fréquemment utilisée : la méthode d'Immunoprécipitation de la Chromatine (sigle anglais ChIP) développée en 1984 par John T. Lis et David Gilmour (114).

(a) Principe de l'immunoprécipitation de la chromatine

Cette méthode permet de séquencer, ou de détecter par des puces à ADN des interactions entre des protéines liées de manière covalente à l'ADN au sein de cellule. Il existe plusieurs variantes de la méthode ChIP (**Ch**romatine **I**mmuno**P**recipitation) telle que ChIP-chip (**Ch**romatine **I**muno**P**recipitation on **chip**), ChIP-seq (**Ch**romatine **I**mmuno**P**recipitation **seq**uencing) et ChIP-exo (**Ch**romatine **I**mmuno**P**recipitation with **exo**nuclease). Lors d'une expérience de ChIP, les protéines vont être liées à l'ADN de manière covalente (crosslinking) grâce à l'action d'un agent de réticulation tel que le formaldéhyde. La cellule est ensuite lysée puis l'ADN coupé par sonication ou digestion enzymatique. À l'aide d'anticorps qui vont se lier spécifiquement aux protéines étudiées, on va précipiter le complexe et le purifier. L'échantillon obtenu est ensuite chauffé afin de séparer la protéine du fragment d'ADN qui est ensuite amplifié. Cette partie est commune à toutes les méthodes dites ChIP. En revanche la seconde partie du nom de la méthode détermine la suite du protocole.

La méthode ChIP-chip (106, 115, 116) utilise la méthode de la puce à ADN en seconde partie pour accélérer le processus d'identification afin que plusieurs milliers de séquences puissent être analysés simultanément.

La méthode ChIP-seq (107, 117) amplifie chaque échantillon en phase solide puis analyse les fragments d'ADN par des méthodes de séquençage nouvelle génération : le séquençage par synthèse, par ligation par simple molécule ou par pyroséquençage. Récemment, la méthode ChIP-seq a été utilisée sur plusieurs lignées cellulaires humaines et de chimpanzés afin de déterminer l'impact des variations génétiques entre individus sur la présence de facteurs de transcription (118).

Une méthode *in vitro*, DIP-chip (immunoprécipitation de l'ADN) basée sur le même principe a été développée en 2005 (119) (cf. Figure 15). Contrairement à la chromatine utilisée dans les méthodes ChIP, l'ADN génomique utilisé dans les expériences de DIP est nu,

c'est à dire sans toutes les protéines (nucléosomes, enzymes). La protéine dont on souhaite tester la spécificité d'interaction est purifiée et incubée avec l'ADN nu. On procède ensuite au séquençage ou à la détection des fragments grâce aux puces à ADN des fragments d'ADN où la protéine s'est fixée.

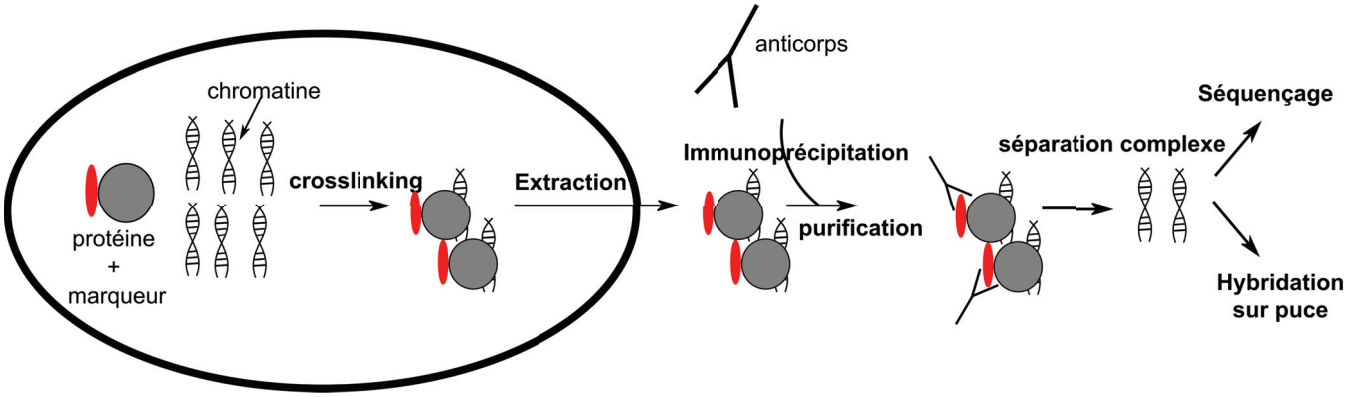


Figure 15 : Procédure *in vivo* pour la caractérisation des sites de liaisons ADN-protéine. Procédure de la méthode ChIP. La partie séquençage ou hybridation sur puce dépend du second terme de la méthode ChIP (exemple ChIP-seq conduit au séquençage alors que ChIP-chip conduit à l'hybridation sur puce

2. Approches In vitro

(a) Principe méthode d'empreinte ADN (DNA footprinting)

La méthode d'empreinte ADN (120) est une méthode de la biologie moléculaire qui détecte les interactions ADN-protéine en utilisant le fait qu'une protéine qui se lie à l'ADN lui confère une protection contre le clivage enzymatique (121). La première étape consiste à lyser la cellule puis amplifier les fragments d'ADN génomique (marqué à l'extrémité 3' ou 5' par une marque radioactive ou fluorescente) par PCR. On réalise ensuite deux échantillons. Sur les deux échantillons, un seul sera incubé avec la protéine d'intérêt. Les échantillons sont ensuite digéré grâce à des enzymes (type DNAse1) ou chimiquement. Une protéine qui se serait liée à l'ADN protège donc le site de liaisons de la dégradation chimique ou enzymatique. L'analyse de la séquence se fait ensuite par séquençage (automates) ou sur un gel de polyacrylamide en conditions dénaturantes (cf. Figure 16C). L'échantillon qui n'a pas été mis en présence de la protéine permettra d'identifier la séquence qui aura été masquée par la protéine.

(b) SELEX et HT SELEX

L'approche SELEX (cf. Figure 16B) (systematic evolution of ligands by exponential enrichment) est basée sur la méthode suivante : dans un premier temps la protéine purifiée est incubée avec un ensemble de fragments d'ADN aléatoire. Les fragments où la protéine s'est liée sont ensuite amplifiés par PCR et réincubés avec la même protéine afin d'identifier des interactions avec une haute affinité (101, 122). Après plusieurs cycles de réincubation, les fragments sont analysés par séquençage. Cette méthode est très pratique, car elle permet l'étude systématique des interactions ADN-protéine sans a priori sur la séquence de fixation (123). La méthode HT SELEX (high-throughput systematic evolution of ligands by exponential enrichment) est une version plus récente permettant une analyse à haut débit des interactions ADN-protéine (124).

(c) Puce à protéines

Cette méthode permet de mesurer la spécificité de liaison ADN-protéine pour des fragments d'ADN de 8 à 10 paires de bases. Avec cette technique, il est possible de tester 4^N séquences (soit 1 048 576 séquences pour un fragment d'ADN de 10 paires de bases) (108). Le principe est simple, sur une puce on fixe toutes les combinaisons de séquences d'ADN possibles que l'on incube ensuite avec la protéine d'intérêt. On ajoute ensuite un anticorps spécifique marqué d'un fluorophore qui va se lier à la protéine étudiée. On excite ensuite la plaque avec un laser qui renverra un signal lumineux si la protéine est fixée au fragment d'ADN (cf. Figure 16C).

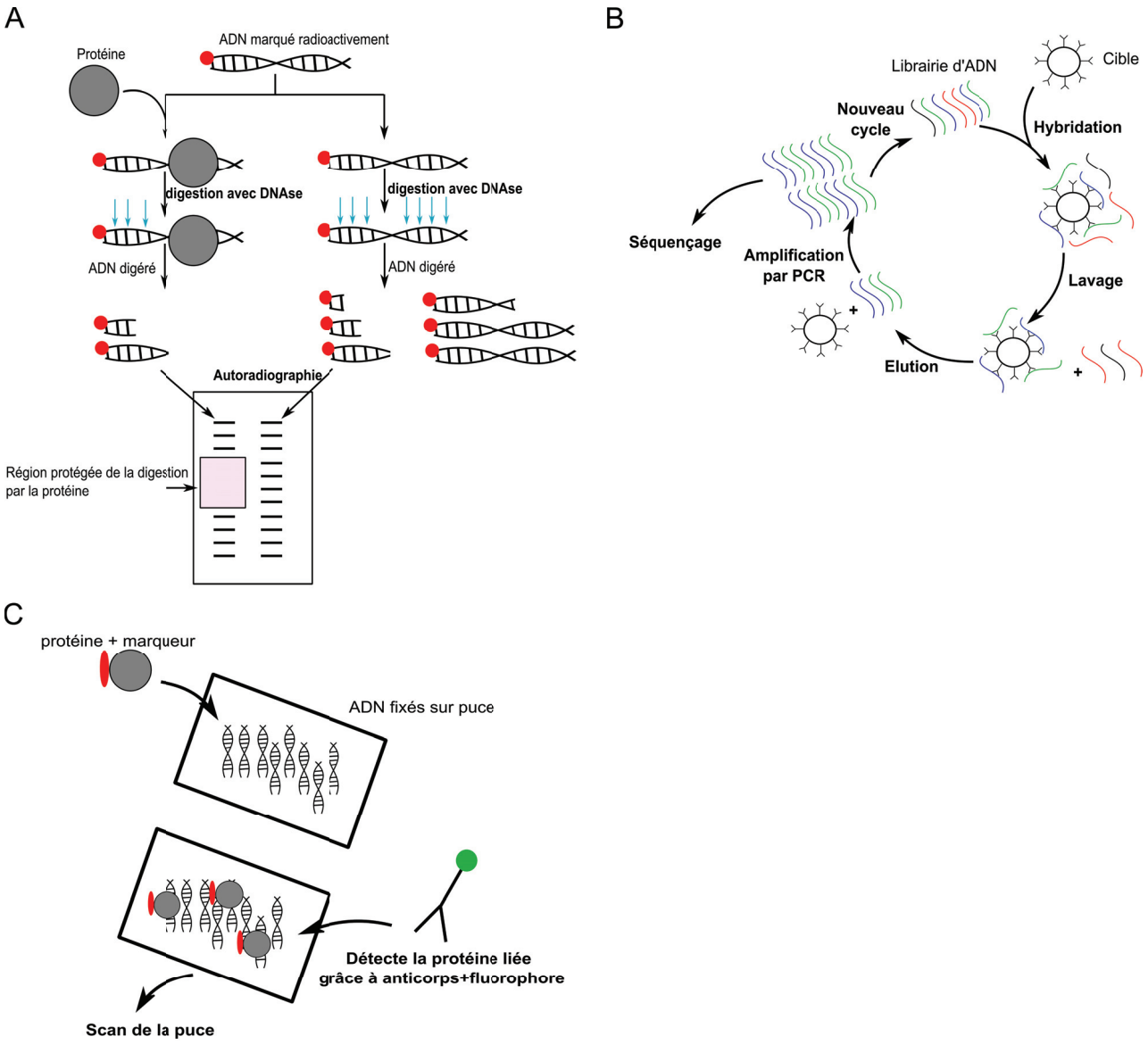


Figure 16 : Procédure *in vitro* pour la caractérisation des sites de liaisons ADN-protéine. A) procédure de la méthode d'empreinte ADN. B) procédure *in vitro* de la méthode SELEX ; C) procédure de la méthode des puces à protéines

B. Méthodes de prédiction de la spécificité de reconnaissance *in silico*

Grâce au nombre croissant de données structurales et à la bioinformatique, plusieurs méthodes computationnelles prédisant la spécificité de liaison ADN-protéine ont émergé ces deux dernières décennies. Pour un ADN où le site d'interaction comprend N paires de bases (N contient entre 5 à 25 pb), il y a 4^N séquences possibles à laquelle une protéine ou un facteur de transcriptions peut se lier. Pour un site de liaison contenant 25 paires de bases, il y a $4^{25} \approx 10^{15}$ combinaisons potentielle. Les méthodes *in vivo* ou *in vitro* actuellement

pratiquées ne peuvent tester un aussi grand nombre de séquences et donc déterminer avec précision les sites d'interactions. Afin d'étudier un plus grand nombre de séquences, des méthodes de prédiction *in silico* ont été développées. À partir de structures atomistiques provenant de la PDB, ces méthodes vont effectuer des mutations des bases nucléiques (par enfilage moléculaire) aux différentes positions du site d'interaction et mesurer l'énergie qui résulte de la formation du complexe. La création de mutants va permettre de répondre à deux principales questions ; 1) Qu'advient-il de la stabilité du complexe après mutation de certaines bases de l'ADN ? et 2) Grâce à la mutation systématique de l'ensemble des bases du site de liaison, quelles sont les séquences sélectionnées par la protéine pour effectuer des interactions ?

1. Fonctions énergétiques : Approches statistiques et physiques

Il existe plusieurs méthodes, qui utilisent des fonctions d'énergies et des approches d'échantillonnages des structures différentes. Actuellement on peut diviser les différentes approches en deux groupes en fonction de la méthode employée pour calculer l'énergie libre du complexe. Les méthodes basées sur le potentiel statistique (knowledge-based) utilisent les données expérimentales et les structures des bases de données pour estimer l'énergie entre atomes lourds du complexe ADN-protéine d'une part et d'autre part, il y a les méthodes de mécanique moléculaire qui calculent l'énergie par la sommation de termes énergétiques indépendants (énergies de Lennard-Jones et électrostatique).

(a) Modèles knowledge-based

Le modèle knowledge-based est un modèle mathématique paramétrique qui va calculer l'énergie d'un complexe à partir de données statistiques provenant de structures de complexes ADN-protéine. Dans ce type de modèle, on considère le nombre d'interactions formées entre l'ADN et les protéines pour estimer l'énergie libre de formation du complexe. Au sein des modèles knowledge-based, il y a deux approches différentes pour calculer le nombre d'interactions. Certaines méthodes considèrent les interactions à l'échelle de l'atome c'est-à-dire entre les atomes lourds situés à l'interface ADN-protéine (125, 126) pour générer une énergie d'interaction dérivée d'une matrice de contacts alors que d'autres méthodes considèrent les interactions à l'échelle du résidu et de l'acide nucléique (127–131). Comme les modèles knowledge-based s'appuient sur des probabilités statistiques pour calculer l'énergie

libre d'un complexe à partir de matrices de contacts, il est nécessaire que celles-ci soient créées à partir d'un jeu de données de structures important et non redondant afin d'être sûr que statistiquement la matrice de contact soit la plus représentative possible. Dans ce type de modèle, lorsqu'une interaction est surreprésentée par rapport à ce qui est attendu statistiquement par le hasard, l'énergie associée à la paire sera favorable (énergie d'interaction <0) alors qu'une interaction sous-représentée sera énergétiquement pénalisée par des valeurs d'énergie positives, car elles ont moins de chances de se produire dans la nature. En fonction de la méthode knowledge-based, la fonction qui dérive l'énergie à partir de la matrice de contacts va changer. Actuellement on retrouve trois fonctions : (132), DFIRE (129, 130) et fonction μ (133–135). La méthode quasichimique calcule l'énergie d'interaction ε entre les atomes j de la protéine et i de l'ADN à une distance d selon l'équation :

$$\varepsilon(i,j,d) = -RT * ln\frac{N(i,j,d)}{N(d)\chi_i\chi_j}$$

Où R est la constante des gaz parfaits, T la température, N(i, j, d) est le nombre de contacts observés dans le jeu de données, N(d) est le nombre total de contacts dans les structures de référence pour la gamme de distance d et χ_i , χ_j est la fraction d'atome du type i, j pour la protéine et l'ADN respectivement. Le dénominateur $N(d)\chi_i\chi_j$ décrit le nombre de contacts attendu entre les deux types d'atomes pour la distance d et sert d'état de référence.

La méthode DFIRE utilise un autre état de référence pour calculer l'énergie d'interaction :

$$\varepsilon(i,j,d) = -RT * ln\frac{N(i,j,d)}{N(i,j,d_{cut})\left(\frac{r(d)}{r(d_{cut})}\right)^{\alpha}\left(\frac{\Delta r(d)}{\Delta r(d_{cut})}\right)}$$

Où d_{cut} est la distance de référence d'une gamme de valeurs, $r(d)$ est la distance moyenne de la gamme de valeurs d_{cut}. $\Delta r(d)$ est la largeur de la gamme de valeurs d_{cut}. L'état de référence de la fonction DFIRE suppose que les interactions soient à courtes distances et que au delà de la valeur d_{cut} il n'y ait plus d'interaction entre les atomes. Le nombre de contacts est ensuite normalisé dans les différentes gammes de valeurs de distances.

Enfin la fonction μ est une fonction très utilisée dans la modélisation du repliement des protéines (133–135) qui part du principe que la structure native est la structure la moins frustrée et correspond au minimum global :

$$\varepsilon(i,j,d) = \frac{-\mu(d)N(i,j,d) + \big(1-\mu(d)\big)N*(i,j,d)}{\mu(d)N(i,j,d) + \big(1-\mu(d)\big)N*(i,j,d)}$$

Où $N * (i, j, d)$ est le nombre de non-contacts c'est-à-dire le nombre total de paires (i,j) dans le complexe moins le nombre de contacts $N(i, j, d)$ dans la gamme de valeurs d. Le paramètre μ étant choisi pour que $\varepsilon(i, j, d)$ soit égale à 0 pour chaque gamme de valeur d.

La plupart des modèles knowledge-based utilisent une matrice de contact qui regroupe l'ensemble des interactions entre atomes lourds sans distinctions des interactions entre atomes lourds pour former des liaisons hydrogène ou interactions entre atomes lourds pour former des interactions hydrophobes. AlQuraishi et McAdams ont récemment développé une méthode *de novo* (136) permettant de calculer l'énergie interatomique à partir de l'interface des structures ADN-protéines et des énergies de liaisons obtenues expérimentalement. Dans cette approche, l'interaction ADN-protéine est étudiée comme un capteur mésoscopique qui reçoit un signal microscopique. Le potentiel d'interaction atomique est ensuite adapté par une régression logistique qui tient compte de la faible représentation des atomes pour éviter un sur-apprentissage du modèle.

(b) Modèles en mécanique moléculaire

Contrairement aux fonctions d'énergies des méthodes knowledge-based, qui utilisent la probabilité de rencontrer une interaction entre deux atomes ou entre un résidu et une base pour évaluer l'énergie de liaison, les méthodes de mécanique moléculaire utilisent des modèles physiques plus complexes faisant intervenir plusieurs termes énergétiques (énergies de Lennard-Jones, énergies des interactions Coulombiennes, changements de torsions et angles de valence, …) pour évaluer l'énergie du complexe ADN-protéine, tout en tenant compte de la contribution énergétique des changements de conformation des partenaires lors de la formation d'une interaction. Les méthodes de mécanique moléculaire calculent l'énergie de formation du complexe dans des conditions qui se rapprochent le plus de l'environnement physiologique, notamment grâce à la présence d'un environnement ionique et aqueux implicite (137), (138), (139), (140) ou explicite (141), (142), (143) qui va permettre de moduler l'interaction électrostatique entre la protéine et l'ADN. Ces méthodes nécessitent donc un nombre de paramètres important (angles, distances atomiques, charges, constantes de forces) afin de reproduire des valeurs expérimentales ou théoriques issues de calculs en mécanique quantique.

La dynamique moléculaire est très utilisée pour étudier l'énergie des liaisons des macromolécules et plusieurs groupes ont employé la simulation en dynamique moléculaire et les calculs d'énergie libre pour estimer la spécificité d'interaction des protéines pour

différents sites d'ADN. La méthode classique de dynamique moléculaire ne permet pas d'obtenir directement des informations sur l'énergie libre compte tenu des difficultés d'échantillonnage. C'est pourquoi des variantes de dynamique ou des outils complémentaires sont employés lors du calcul d'énergie libre.

Dans les simulations pour obtenir l'énergie libre de façon « alchimique », la différence d'énergie libre entre deux états est calculée en échantillonnant le long d'une transformation artificielle, contrôlée par une variable λ, liant les deux états (structure native et mutant par exemple). Cette transformation peut s'effectuer lors de simulation où λ est constant ou avec λ qui varient continuellement entre les deux états. Liu et Bader (141) ont utilisé l'approche d'intégration thermodynamique permettant de calculer la différence d'énergie entre la forme native et la forme mutée en effectuant des simulations avec différentes valeurs discrètes de λ. Cette méthode est particulièrement intéressante car elle permet de séparer l'énergie libre en contributions individuelles (énergie provenant de l'ADN, de la protéine ou du solvant par exemple) (144). L'analyse des complexes MAT-α2 et de la leucine zipper GCN4 en eau explicite par cette méthode a montré que les liaisons hydrogènes ADN-protéines médiées par les molécules d'eau étaient prédominantes par rapport aux contacts directs et que leur rôle est essentiel pour améliorer la prédiction du site de liaison (141). Néanmoins, Moroni a démontré qu'utiliser des modèles d'eau implicite (modèle où la constante diélectrique dépend de la distance, modèle dépendant de la surface accessible au solvant (Accessible surface area-based method) ou encore la méthode Generalized Born) permettait aussi d'obtenir une bonne prédiction du site de liaison. Cependant la meilleure prédiction en accord avec les données expérimentales est obtenue en utilisant la fonction d'énergie utilisée en mécanique moléculaire (contenant les termes énergétiques des interactions liées et non liées) à laquelle est ajouté le modèle d'eau Generalized Born et l'accessibilité au solvant (MM/GBSA).

En 2009, Temiz et Camacho (145) ont élaboré une nouvelle approche permettant de décrire l'affinité de liaison des protéines en doigt de zinc de la famille C_2H_2 qui utilise la dynamique moléculaire couplée à de la modélisation par homologie pour générer un score empirique basé sur l'analyse des liaisons hydrogène ADN-protéine et la désolvatation de l'interface. Dans ce modèle, la force de l'interaction des liaisons hydrogène est modulée afin de tenir compte de l'accessibilité au solvant des atomes. Les énergies calculées par cette approche montrent de très bons résultats pour différentes protéines en doigt de zinc mais restent encore à être testés pour d'autres familles protéiques.

2. Déconvolution des mécanismes de reconnaissance par les méthodes in silico

La reconnaissance peut être divisée en deux composantes énergétiques connues sous le terme de lecture directe et indirecte, décrites de manière plus approfondie dans la section III et IV de ce chapitre. Le terme de lecture directe fait référence aux interactions qui ont lieu entre la protéine et les bases de l'ADN par l'intermédiaire de liaisons hydrogène, de ponts salins ou d'interactions hydrophobes. Le terme indirect est apparu à la fin des années 80 (80) et fait référence aux déformations de l'ADN qui vont favoriser l'interaction ADN-protéine. Si l'on souhaite caractériser l'impact de la déformation sur la reconnaissance, il est important d'avoir un modèle capable de donner la contribution énergétique provenant de la reconnaissance directe (énergie d'interaction) et de la reconnaissance indirecte (déformation de l'ADN). En 1998, Olson et coll. ont développé une fonction d'énergie empirique qui décrit les préférences géométriques et la déformation entre les paires de bases grâce à six paramètres hélicoïdaux (Twist, Roll, Tilt, Shift, Slide et Rise) à partir de données observées dans 92 complexes ADN-protéine (146). Cette fonction empirique peut être ensuite ajoutée aux fonctions d'énergies développées par les méthodes statistiques pour obtenir une approximation de l'impact de la déformation de l'ADN dans le mécanisme de reconnaissance. La méthode knowledge-based DNAPROT utilise cette estimation empirique d'Olson couplée à trois matrices de contacts (matrice des liaisons hydrogène, des interactions hydrophobes et des interactions médiées par les molécules d'eau) (147) afin d'obtenir une valeur quantitative du mécanisme qui prédomine dans la spécificité de reconnaissance. Ce type de modèle fourni de bonnes prédictions (147), (132), (136) et a montré de bons résultats de prédiction pour deux protéines modélisées par homologie.

3. Dépendance énergétique entre paires de bases adjacentes

Les modèles de prédiction actuels considèrent chaque position comme indépendante ; c'est-à-dire que la modification d'une base nucléique du site de liaison n'aura pas d'incidence sur la structure des bases environnantes et que l'énergie libre du complexe ΔG peut être obtenue comme la résultante des énergies libres à chaque position i de l'ADN ($\Delta G = \sum_{i=1}^{N} \Delta G_i$ ou N est le nombre de paires de bases). La plupart des méthodes knowledge-based et de

mécanique moléculaire utilisent cette notion d'indépendance entre les bases pour réduire l'espace des séquences à explorer. Ainsi au lieu de tester les 4^N possibilités pour un ADN de N paires de bases on effectue les calculs uniquement sur 4N séquences. Par exemple pour un ADN de 10 paires de bases, on réduit le nombre de séquences de 4^{10} à 40. L'étude de la spécificité ADN-protéine en utilisant une approche par indépendance des sites montre de bons résultats (148), mais atteint rapidement ses limites, notamment dans les cas où la structure de l'ADN dépend de la séquence en nucléotides (149). Pour surmonter ce problème, la méthode d'enfilage moléculaire ADAPT (150) calcule l'énergie de liaison pour toutes les combinaisons de séquences possibles (4^N possibilités) pour le site de liaison afin de capturer la corrélation énergétique entre les nucléotides de chaque position (plus de détails sur la méthode ADAPT donnés dans le chapitre Méthodologie). Les résultats obtenus avec ADAPT sont cohérents avec les données expérimentales pour plusieurs familles et complexes ADN-protéine (151, 152).

4. Flexibilité accordée au modèle

On peut également classer ces méthodes d'étude des complexes protéine-ADN en fonction de la flexibilité accordée au modèle. Lorsqu'une base est remplacée pour effectuer une mutation, la nouvelle base est introduite en préservant la planéité de la base, la géométrie du désoxyribose et de la liaison glycosidique ainsi que le brin phosphodiester.

Si on ne permet aucune relaxation au niveau de l'interface ADN-protéine, l'ajout d'une nouvelle base peut induire des encombrements stériques ou une interaction électrostatique défavorable qui biaisera les calculs énergétiques. Les méthodes knowledge-based n'autorisent aucune flexibilité que ce soit pour la protéine ou pour l'ADN. Ce manque de flexibilité dans le modèle a tendance à favoriser la base présente dans la structure au détriment des séquences testées. Les méthodes de mécanique moléculaire permettent d'introduire différents niveaux de flexibilité au sein du complexe ADN-protéine.

Après l'introduction de la mutation, certaines méthodes vont permettre un réarrangement des chaines latérales de la protéine grâce à des cycles de minimisation (recherche d'un minimum local d'énergie) (150, 153) ou en utilisant des librairies contenant des rotamères (pour échantillonner des réarrangements non accessibles par la minimisation) tout en maintenant le squelette peptidique rigide (138),(154), (155).

Du côté de l'ADN, on distingue trois niveaux de flexibilité :

1) le brin phosphodiester est figé, mais on autorise un réarrangement des bases grâce à des cycles de minimisation et/ou grâce à des librairies de rotamères (137), (138), (139). En 2010 Serrano a utilisé cette approche avec le champ de force FoldX pour prédire le site de fixation de la protéine Pax6 lors de mutations (137).

2) Aucune contrainte n'est appliquée sur l'ensemble de l'ADN qui est minimisé (155), (150, 153) ;

3) Exploration par dynamique moléculaire de l'espace conformationnel après la mutation grâce à de la dynamique moléculaire (140), (141), (142), (143). L'apport de la flexibilité et l'introduction de cycles de minimisation dans les structures cristallographiques sur la qualité des prédictions restent encore controversés. En 2004, Endres et coll. ont montré que relaxer la structure diminue la précision des prédictions (156) alors qu'en 2007, Donald et ses collaborateurs démontrent que la minimisation permet d'augmenter le pouvoir prédictif des méthodes de mécanique moléculaire (132).

5. Création des matrices de poids par position à partir des prédictions d'énergies de libres

Nous avons vu précédemment que les différents modèles de prédiction partaient de l'hypothèse que l'énergie libre d'un complexe était la résultante de l'énergie libre issue de chaque paire de bases et que donc l'énergie libre de formation du complexe pouvait être calculé grâce à la relation :

$$\Delta G = \sum_{i=1}^{N} \Delta G_i$$

où N est le nombre de paires de bases du site de liaison et ΔG_i est la contribution énergétique de la paire de bases à la position i.

En utilisant la formule de Boltzmann (157), on peut calculer la probabilité d'observer le nucléotide m à la position i avec l'équation suivante :

$$p_m^i = \frac{exp^{-\beta \Delta G_m^i}}{\sum_{k \in \{A,C,G,T\}} exp^{-\beta \Delta G_k^i}}$$

où ΔG_m^i est l'énergie de liaison de la structure à la position i quand le nucléotide m est présent et $\beta = 1/RT$ avec R correspondant à la constante des gaz parfaits et T à la température. En calculant p_m^i pour chaque position et chaque nucléotide, on obtient une matrice de poids par position permettant de visualiser la spécificité de séquence d'une protéine.

II. Recherche du bon site de liaison par les protéines

Une fois la séquence d'ADN identifiée pour une protéine donnée, deux questions restent encore sans réponse ; I) Comment une protéine est-elle capable de retrouver sa cible parmi des génomes de 10^6 à 10^9 paires de bases ? ; II) Une fois le site d'interaction atteint quels sont les mécanismes permettant la reconnaissance spécifique ADN-protéine ? Dans cette section les mécanismes utilisés pour trouver le bon site de liaison vont être abordés et les mécanismes permettant la reconnaissance spécifique seront traités dans la section III et IV de ce chapitre.

A. Les mécanismes de recherche

Le répresseur lactose I (LacI) possède une constante d'association d'environ 10^{10} $M^{-1}.s^{-1}$ soit 100 fois supérieures à la constante d'association (diffusion limite $\sim 10^8$ $M^{-1}.s^{-1}$ (158)) attendue pour les biomolécules par le modèle de diffusion 3D. Toutes les protéines se liant à l'ADN ne possèdent pas une constante d'association supérieure à la constante de diffusion limite (159–161), mais c'est un critère retrouvé chez plusieurs protéines appartenant à la même famille que LacI (162, 163). Pour expliquer cette divergence, Riggs propose un nouveau mécanisme, selon lequel la protéine n'utiliserait pas uniquement une diffusion 3D, mais également en se déplaçant le long des sillons de l'ADN de manière curviligne (164, 165) grâce aux interactions non spécifiques suite à une fixation aléatoire que l'on nomme diffusion 1D (diffusion Brownienne ou coulissante) (166). Cette association entre diffusion 1D et 3D est couramment appelée diffusion facilitée. Depuis le modèle a évolué et maintenant la diffusion facilitée contient quatre mécanismes (cf. Figure 17) : 1) diffusion 1D coulissante sans dissociation du complexe, 2) diffusion 1D par saut de la protéine quelques bases plus loin, 3) diffusion 3D par saut de la protéine par transfert entre deux segments d'ADN et 4) diffusion 3D classique (167, 168). Plusieurs expériences *in vivo* (169, 170), *in vitro* (165, 171, 172) et *in silico* (173–176) ont été menées pour étudier le modèle de diffusion facilitée.

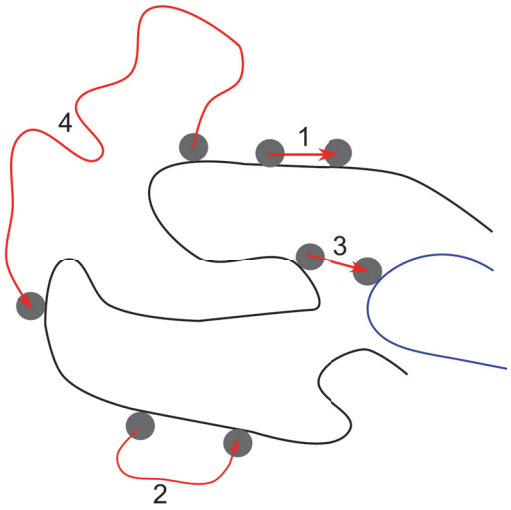


Figure 17 : Représentation schématique des quatre mécanismes utilisés par les protéines pour rechercher leur site de liaison. 1) diffusion 1D glissant ; 2) Diffusion 1D par saut de la protéine ; 3) transfert entre deux segments d'ADN différents représentés en noir et bleu et 4) diffusion 3D.

B. Effet de la concentration ionique

Plusieurs études réalisées ces dernières années montrent que le modèle de diffusion facilitée dépend de la concentration saline du milieu. Lorsque la concentration saline augmente, le mécanisme de diffusion 1D devient moins efficace que la diffusion 3D, car la force ionique du milieu diminue l'attractivité électrostatique ADN-protéine (177) et facilite la dissociation entre les deux partenaires (cf. Figure 18).

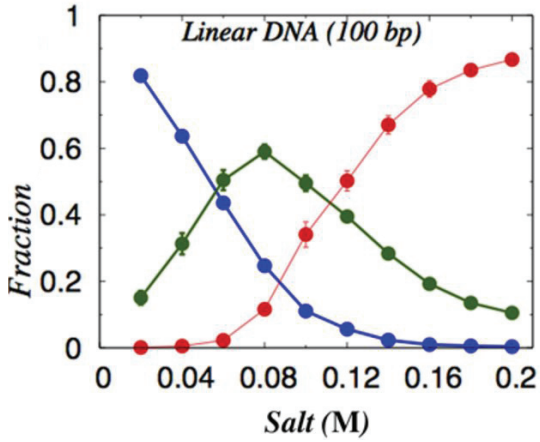


Figure 18 : Effet de la concentration saline sur les mécanismes de diffusion. En rouge le mécanisme de diffusion 3D, en bleu la diffusion 1D glissante et en vert la diffusion 1D par saut de la protéine. Image adaptée de (176).

C. Effet de la forme de l'ADN

La forme de l'ADN joue un rôle important dans le mécanisme de recherche. Une étude en dynamique moléculaire sur la protéine Sap-1 montre que la courbure de l'ADN influence le mode de diffusion utilisé par la protéine. Plus l'ADN est courbé et moins la diffusion 1D

glissante est favorisée au profit de la diffusion 1D par saut de la protéine quelques bases plus loin. Dans l'expérience, la concentration ionique de 0,02 M (concentration maximale déterminée dans l'étude de Sap-1 où la protéine reste associée à l'ADN) ne permet pas de voir l'impact de la courbure de l'ADN sur la diffusion 3D qui ne se produit que rarement (environ 1% du temps) à cette concentration saline (cf. Figure 19A). Lorsque le niveau de courbure devient plus important, le potentiel électrostatique négatif augmente sur la face de l'ADN où les phosphates se rapprochent rendant plus difficile la diffusion glissante de la protéine de l'intérieur vers l'extérieur de la courbure formée par l'ADN (176, 178) au détriment du mécanisme de diffusion 1D par saut de la protéine(179).

Le niveau d'enroulement de l'ADN (grâce au twist) influence également le mode de recherche. Pour la protéine Sap-1, un ADN faiblement twisté présente un grand sillon plus large et favorise le phénomène de diffusion 1D glissant alors qu'un ADN fortement twisté possède un grand sillon très étroit ce qui favorise la diffusion 1D par saut de la protéine. (cf. Figure 19) [172].

Enfin la plupart des expériences réalisées tendent à conclure que le phénomène de diffusion 1D glissant est un mécanisme lent et qu'il contribue peu (environ 20 %) à la recherche totale du site de liaison. La majeure partie de la recherche s'effectue davantage avec la diffusion 1D par saut de la protéine et par la diffusion 3D. La protéine SAP-1 présenté ci avant est un cas particulier car la diffusion 3D ne contribue que pour 1% de la recherche.

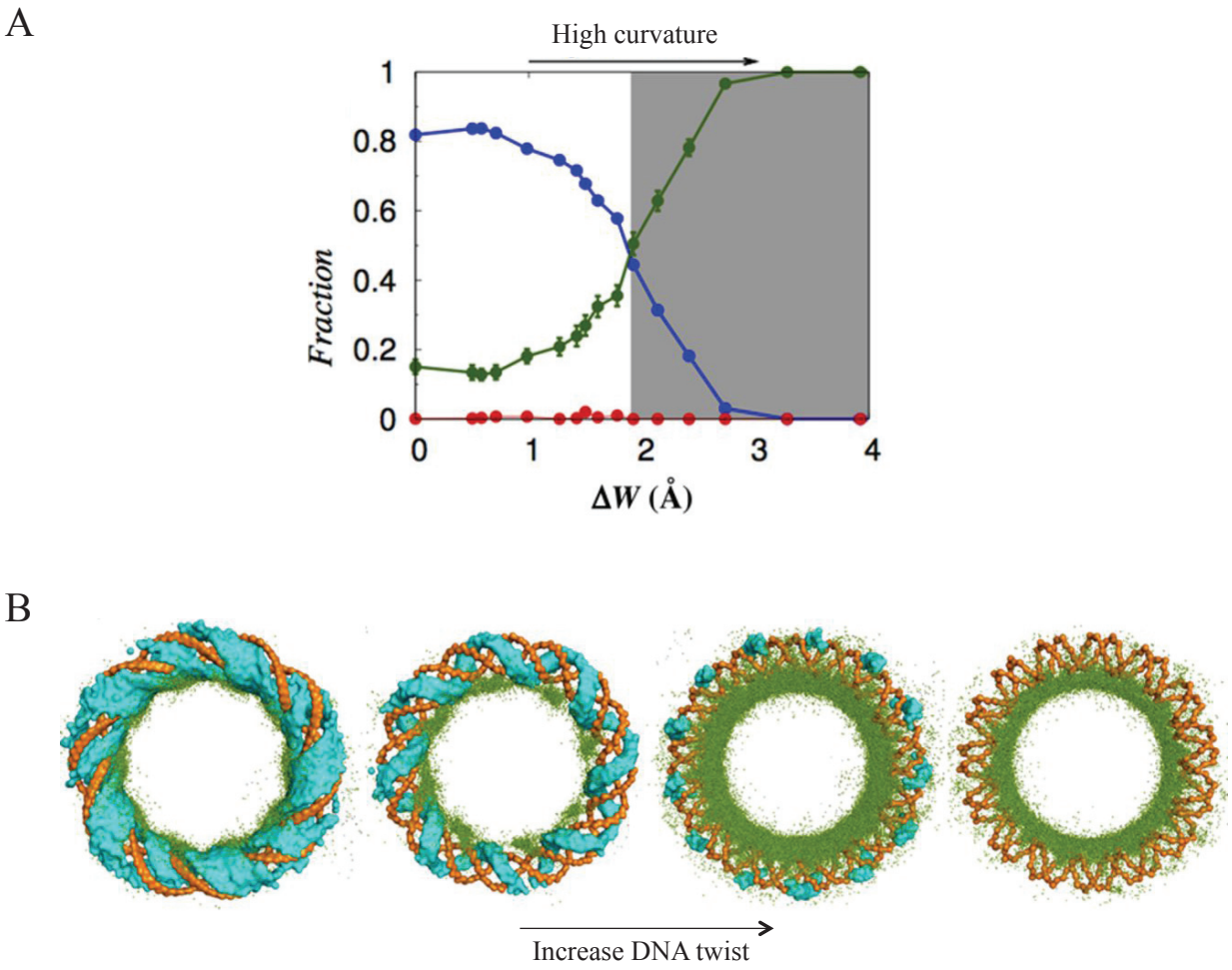


Figure 19 : Effet de la forme de l'ADN sur les mécanismes de diffusion. A) Impact du niveau de courbure de l'ADN avec en rouge le mécanisme de diffusion 3D, en bleu la diffusion 1D glissante et en vert la diffusion 1D par saut de la protéine. B) Effet du Twist sur les mécanismes de diffusion 1D glissante (représentée par la surface cyan) et sur diffusion 1D par saut de protéine (représentée par les points verts) définit par la position de la protéine Sap-1 le long de l'ADN lors de la dynamique moléculaire. Image adaptée de (176).

III. Mécanisme de la lecture directe

Une étude réalisée en 2007 sur des complexes de la PDB (180), montre qu'une interface ADN-protéine, est constituée en moyenne de 24 résidus protéiques et de 12 nucléotides pour une surface de contact de 1600 ± 400 Å^2 en moyenne. Cette interface se compose d'interactions spécifiques de différentes natures : électrostatiques (liaisons hydrogène) ou interactions entre groupements hydrophobes.

A. Les liaisons hydrogène

La répartition et le nombre des donneurs et accepteurs de liaisons hydrogène ne sont pas les mêmes selon la paire AT ou GC et dépendent du sillon. Dans une paire AT, il y a possibilité d'effectuer trois liaisons hydrogène (deux accepteurs et un donneur) dans le grand sillon contre deux dans le petit sillon (deux accepteurs) (Figure 20A). Pour la paire GC, il y a aussi deux accepteurs et un donneur de liaison hydrogène dans le grand sillon, en revanche au niveau du petit sillon, la paire GC peut effectuer trois liaisons grâce à deux accepteurs et un donneur (Figure 20A). Lorsque l'on regarde les donneurs et accepteurs de liaisons hydrogène pour les paires Watson & Crick AT, TA, CG ou GC, on s'aperçoit que la position spatiale des groupements fonctionnels dans le grand sillon de l'ADN offre quatre possibilités d'appariement distinctes alors que le petit sillon en offre deux (78, 181). L'ordre des groupements fonctionnels dans le petit sillon est tel que la paire TA semble identique à la paire AT et que la paire GC semble identique à la paire CG (cf. Figure 20B).

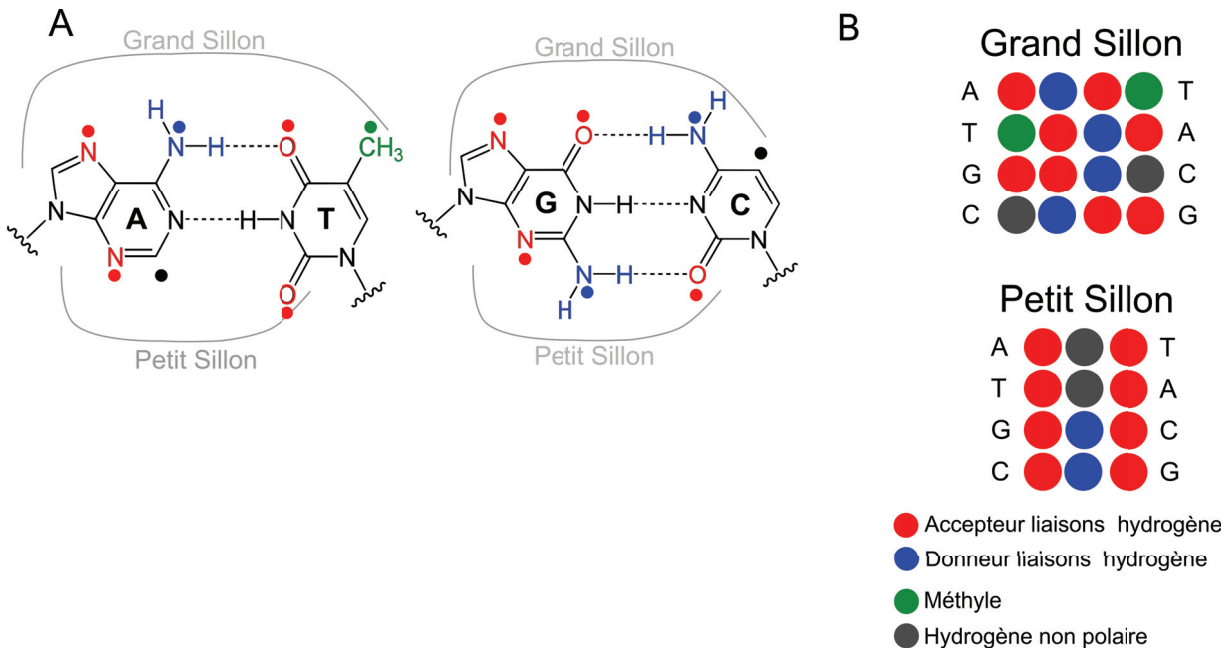


Figure 20 : Groupements fonctionnels vus dans le grand et petit sillon. A) Groupes fonctionnels des paires de bases AT et GC observés dans le grand et petit sillon de l'ADN avec en rouge les accepteurs et en bleu les donneurs de liaisons hydrogène. En vert le groupement méthyle disponible pour former des interactions hydrophobes. B) Représentation schématique de la localisation spatiale des groupes fonctionnels des paires AT, TA, GC et CG dans le grand et petit sillon de l'ADN. Le code couleur utilisé dans la partie A) de la figure est réutilisé dans la partie B).

Malgré la présence de système de reconnaissance spécifique entre le grand et le petit sillon et une organisation de groupes fonctionnels unique, aucun code simple de reconnaissance (entre les chaînes latérales des acides aminés et les bases de l'ADN) n'a été élucidé (182–184).

L'analyse de complexes non redondants réalisée par Luscombe et coll.(185), par Kono et Sarai (125) en 2001 et par Lejeune en 2005(186), révèle tout de même que certains acides aminés ont une préférence pour certains nucléotides (cf. Tableau 1) et certaines parties du nucléotide en particulier les groupements phosphates et les bases. Une interaction bidentée, c'est-à-dire une chaîne latérale qui forme deux liaisons hydrogène avec l'ADN confère un niveau de spécificité supérieure par rapport à une liaison hydrogène simple (181, 187). La guanine est la seule base qui possède deux accepteurs de liaisons hydrogène via les atomes N7 et O6. La substitution de cette base réduit l'affinité de liaison (181). Les acides aminés arginine et lysine ont une préférence pour la base guanine lorsqu'ils forment une interaction bidentée ou bifurquée (un seul donneur interagit avec deux accepteurs) (185). Ce type d'interaction a été observé dans de nombreux complexes biologiques comme l'endonucléase EcoRI (60), le répresseur tryptophane (80), le répresseur λcro et la protéine CAP(188). Dans le cas d'une interaction simple, il n'y a plus de distinction entre guanine et adénine, car les deux bases possèdent l'atome N7 (60).

Les acides aminés asparagine et glutamine ont une préférence d'interaction avec la base adénine ; base qui accepte un hydrogène avec l'atome N7 et donne un hydrogène avec l'atome N6 pour effectuer une double liaison hydrogène (185).

En revanche les acides aminés aspartate et glutamate n'effectuent que rarement des interactions avec les bases en raison d'un environnement électrostatique défavorable. Cependant elles ne sont pas totalement absentes et certains complexes comme RAP1 (189) et la protéine Max (190) possèdent un aspartate et un glutamate enfouis dans le grand sillon qui forment respectivement des liaisons hydrogène avec les bases A/C et C/A respectivement.

Tableau 1 : Préférences des contacts directs entre acides aminés et bases nucléiques comptabilisés sur 139 complexes ADN-protéines(186).

	A	C	G	T
Ala	↓	↓	↓	↓
Ile	↓	↓	↓	↓
Leu	↓	↓	↓	↓
Met	–	–	–	–
Val	↓	↓	↓	
Phe	↓	–	–	–
Trp	–	↑	–	–
Tyr	–	–	↑	↑
Cys	–	–	–	–
Pro	–	–	–	–
Arg	↑	↑	↑	↑
Asp	↓	–	↓	↓
Glu	↓	↓	↓	↓
Lys	↑	↑	↑	↑
Asn	–	↑	↑	↑
Gln	–	–	–	–
His	–	–	↑	↑
Ser	–	–	↑	↑
Thr	–	↑	↑	↑

*La direction de la flèche indique si le nombre de contacts est surreprésenté (flèche verte) ou sous-représenté (flèche rouge) par rapport à ce qui est attendu par le hasard. Les paires n'ayant pas de différence significative entre observé et attendu sont représentées par un tiret.

La position des chaînes latérales des acides aminés impliqués dans le réseau de liaisons hydrogène n'est pas aléatoire. Dans les structures cristallographiques, on distingue des préférences de positionnement dans le grand sillon et dans le petit sillon, la position des chaînes latérales semble plus diffuse (125, 185). La majorité des protéines se lient coté grand sillon de l'ADN, cependant certaines protéines forment des liaisons hydrogène avec le petit sillon malgré le fait que la distribution donneurs/accepteurs ne permettent pas de distinguer facilement une paire AT d'une paire TA ou une paire CG d'une paire GC (181). Certaines protéines en doigt de zinc Cys2Cys2 ayant un domaine GATA-like (protéine THAP) et certaines protéines de la famille des HMG (high mobilty groupe) forment des liaisons côté petit sillon de l'ADN (191–195). Cette interaction qui s'effectue dans le petit sillon n'est pas suffisante pour expliquer la spécificité d'interaction. Le complexe TBP révèle la formation de quatre liaisons hydrogène au niveau du petit sillon sur la boîte TATA. Un si petit nombre de liaisons hydrogène au

niveau du petit sillon n'est pas suffisant pour expliquer la spécificité d'interaction de ce complexe (87, 88, 195). D'autres facteurs liés à la flexibilité et à la forme de l'ADN sont nécessaires et sont développés dans la partie III de ce chapitre.

B. Liaisons hydrogène médiées par des molécules H2O

Plusieurs études menées sur des complexes cristallographiques (185, 196) révèlent que les molécules d'eau sont impliquées dans le réseau de liaisons hydrogène formé entre ADN (base et brin phosphodiester) et protéine et seraient partie intégrante de la reconnaissance. Les molécules d'eau présentes à l'interface forment des liaisons hydrogène préférentiellement avec les atomes accepteurs protéique et nucléique (196). Au sein des protéines ce sont essentiellement les oxygènes du squelette peptidique et les oxygènes des chaînes latérales aspartate et glutamate qui sont les principaux contributeurs, alors que pour l'ADN, les oxygènes du groupement phosphate ainsi que les azotes des bases sont équitablement sollicités. Ce type d'interaction a souvent été observé chez les enzymes interagissant avec l'ADN (197) et pour le répresseur tryptophane (80). Dans la structure cristallographique du répresseur tryptophane, la plupart des contacts directs entre la protéine et l'ADN s'effectuent au niveau des groupes phosphates et l'ensemble des contacts avec les bases CTAG du demi-site de reconnaissance fait intervenir des molécules d'eau. Le rôle des molécules d'eau dans la stabilité et la spécificité ADN-protéine a également été décrit dans l'homéodomaine Pax (198) où 18 molécules se trouvent à l'interface ADN-protéine. Certaines études montrent cependant que toutes les molécules d'eau présentes à l'interface ne sont pas essentielles (199). De ces observations est née l'idée que les chaînes latérales d'une protéine à l'interface occupent les positions de molécules d'eau présentes dans l'ADN dans sa forme non liée (200).

Plusieurs simulations de dynamique moléculaire supportent les conclusions obtenues sur les structures cristallographiques et RMN sur l'importance des liaisons hydrogène médiées par les molécules d'eau (201–205).

C. Les contacts hydrophobes

Bien que les liaisons hydrogène permettent une reconnaissance spécifique des bases en particulier pour discriminer les purines, les contacts hydrophobes sont particulièrement utilisés dans la reconnaissance des pyrimidines (78). Dans le site de

reconnaissance du répresseur P22 c2 du bactériophage lambdoïde P22, quatre groupements méthyles forment une cavité reconnue spécifiquement par une valine (206). Les protéines qui se lient dans le petit sillon de l'ADN (par exemple TBP (87), SRY (207)) élargissent le petit sillon et le déshydratent. Ces protéines forment des contacts hydrophobes au niveau de leur surface d'interaction. Dans le cas du complexe TBP, quatre liaisons hydrogène sont formées sur les 16 possibles suggérant que les contacts hydrophobes contribuent à la spécificité d'interaction (87, 88, 195).

Les acides aminés aromatiques (Phe, Tyr, Trp et His) sont également importants dans la discrimination des bases azotées. Ces acides aminés ont la possibilité de former des interactions d'empilements π. Les interactions π sont de deux types π-π face à face ou π-π en forme T (cf. Figure 21A) (208–211). L'analyse de 428 structures cristallographiques, révèle que près de 41 % des complexes possèdent au moins une ou deux interactions d'empilement π (212). Les interactions d'empilement π impliquent principalement les pyrimidines. Les phénylalanines et les tyrosines sont plus souvent sollicitées pour former ces interactions que les acides aminés histidines et tryptophane (cf. Figure 21). Les interactions d'empilement π sont particulièrement utilisées par les enzymes qui réparent les bases endommagées de l'ADN (213–215).

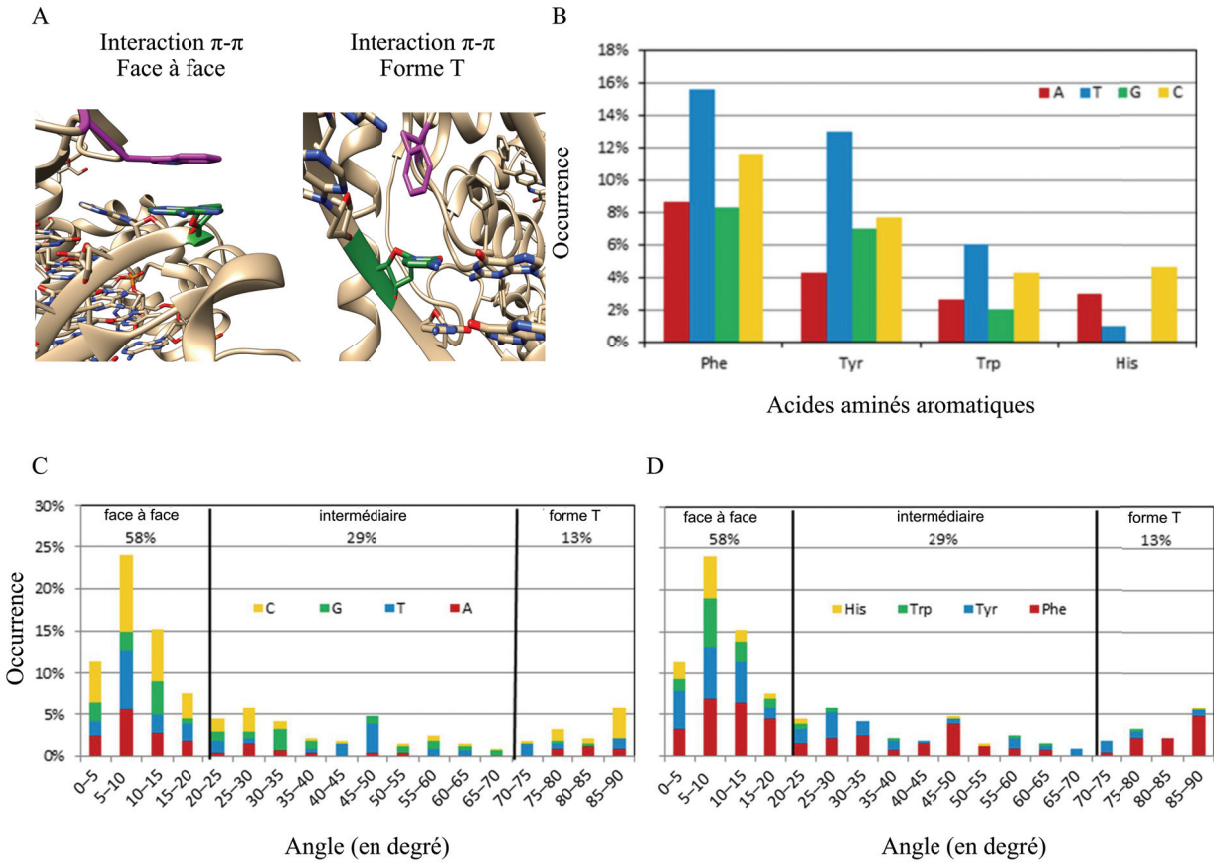


Figure 21 : Structure et statistiques des interactions π-π réalisé sur 482 complexes ADN-protéine. A) Structure des interactions d'empilement π-π face à face (PDB : 3MR5) et en forme de T (PDB : 2WQ7) (avec en rose l'acide aminé et en vert le nucléotide impliqué). B) Répartition des interactions π-π (sans distinction de conformation) en fonction des acides aminés aromatiques et des différentes bases. Répartition des interactions par type d'empilement π : C) par base nucléique D) par acide aminé (212). Les interactions d'empilement ont été découpées en fonction de l'angle formé entre les deux cycles par intervalle de 5 degrés. La forme intermédiaire désigne des configurations d'angles qui ne sont ni la forme face à face ni la forme T.

IV. Mécanismes de la lecture indirecte

Dans la plupart des complexes ADN-protéine, les interactions électrostatiques (liaisons hydrogène) et les interactions hydrophobes ne suffisent pas à expliquer la spécificité d'interaction. Durant ces dernières décennies, il est devenu évident que la forme globale et les variations locales de structure de l'ADN influencent l'interaction ADN-protéine et leur spécificité. Ce type d'interaction a été nommé reconnaissance indirecte (80) ou reconnaissance analogue (216). Désormais, la structure de l'ADN n'est plus considérée comme statique, mais comme une structure flexible et dynamique dont la structure moyenne est un mélange de plusieurs sous états (217–219). La structure de l'ADN et sa capacité à être plus ou moins déformé dépendent de la séquence en nucléotides. La plupart des protéines se fixent sur des ADN de type B dont la structure locale ou globale dévie souvent d'un ADN B idéal.

A. Déformations globales

Dans cette partie, sont définies comme déformations globales ; I) les ADN ayant une forte courbure ; II) une structure d'ADN de type A ou Z ou B.

1. Courbure de l'ADN.

Certaines séquences nucléiques ont une tendance naturelle à se courber. Dans des conditions physiologiques, les séquences qui présentent une succession de nucléotide A et T (tels que A_nT_m), de courtes successions d'adénine (A-tracts) [172], [173], [174] ou de courtes répétitions du nucléotide guanine (G-tracts) (223–225) répétées en phase avec le pas de l'hélice, confère une courbure intrinsèque à la molécule d'ADN en l'absence de contraintes extérieures suggérant qu'un positionnement approprié de ce type de séquence (séquence de fixation de la protéine, séquences séparant deux demi-sites de reconnaissance) peut permettre à l'ADN d'adopter des configurations privilégiées lors de l'interaction ADN-protéine. Dans certains cas, cette courbure intrinsèque de la molécule d'ADN joue un rôle dans la reconnaissance.

La protéine E2 du papillomavirus interagit avec l'ADN sous forme dimérique ou chaque monomère reconnaît un demi-site d'interaction séparé par quatre paires de bases ($ACCGN_4CGGT$) où N_4 correspond aux quatre bases séparant les demi-sites (226, 227). Chez l'homme, la protéine E2 se lie préférentiellement sur les séquences où N_4 sont les séquences AATT et AAAA (227, 228). Ces séquences sont naturellement courbées en l'absence de la protéine et des études ont montré que la protéine E2 a une affinité de liaison plus importante pour les séquences ayant un ADN précourbé que pour les ADN où la protéine doit induire la courbure (227, 229).

Certaines séquences ne se courbent pas spontanément et c'est l'interaction avec la protéine qui va induire la courbure afin de créer de nouveaux contacts. Maher a travaillé sur la courbure de l'ADN en utilisant le concept de protéine fantôme. Dans ces expériences, Maher a simulé l'interaction entre les acides aminés arginine/lysine et le brin phosphodiester en diminuant la charge nette du groupement phosphate afin de former des « patchs » neutres. Les résultats de ces expériences montrent que la neutralisation des groupements phosphate sur une face de l'ADN réduit la répulsion phosphate-phosphate et induit une courbure(223, 230–233). Ce mécanisme de neutralisation d'une face de l'ADN a été observé pour plusieurs protéines(234, 235). On

estime que la neutralisation engendre une courbure moyenne de 3,5° par phosphate neutralisé.

On distingue deux classes de protéines qui courbent l'ADN. D'une part les protéines qui courbent l'ADN de manière concave (par rapport au site de fixation de la protéine) et d'autre part les protéines qui courbent l'ADN de manière convexe (cf. Figure 22). Parmi ces protéines le niveau de courbure peut varier de manière très importante ; 8° pour la structure cristallographique de PU.1 (234) jusqu'à 100° pour le complexe TBP (87, 88, 236). A noter que seule une courbure concave peut être expliquée par le mécanisme étudié par Maher (231).

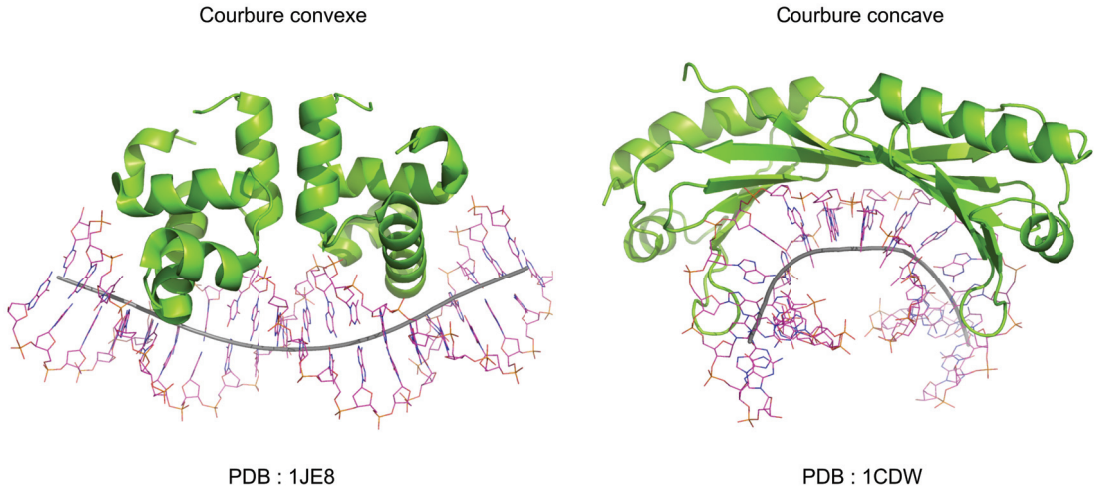


Figure 22 : Exemple de courbure convexe et de courbure concave. La direction de la courbure est représentée par l'axe hélicoïdal de l'ADN (représentation en mode bâton gris).

2. Formes de l'ADN

La forme globale de l'ADN est aussi très importante et définit le type d'interaction et les acides aminés qui vont prédominer au niveau d'une interface protéine-ADN. Lors d'une étude réalisée en 2005 sur 139 complexes ADN-protéines, chaque nucléotide (sur un total de 1195) a été catégorisé comme appartenant à la classe ADN-A ou ADN-B en fonction de son couple d'angles δ (entre les atomes C5'-C4'C3-'O3')-χ (entre les atomes O4'-C1'-N1-C2 pour les pyrimidines et O4'-C1'-N9-C4 pour les purines). Selon la définition de Lu et coll. (237), un nucléotide est catégorisé en forme A si δ est compris entre 60° et 110° et χ entre 150° et -140° alors qu'il sera considéré en forme B si δ est compris entre 70° et 180° et χ entre -140° et -60°. Dans ces 139 complexes, les nucléotides adoptent

préférentiellement une conformation B (93% des 1195 nucléotides). Cependant, les nucléotides qui sont en contact direct avec la protéine au niveau de la surface d'interaction adoptent préférentiellement des conformations retrouvées dans les ADN A (cf. Figure 23) (238), (186).

De plus, en fonction de la forme de l'ADN, la nature des acides aminés présents au niveau de l'interface d'interaction n'est pas la même (186). Par exemple, on trouve davantage de résidus négatifs à l'interface dans un complexe protéine-ADN de type A (environ 12%) que dans un complexe protéine-ADN de type B (environ 4,5%) ; les acides aminés arginine/lysine interagissent plus fréquemment avec les nucléotides dans un ADN en conformation B (36% contre 14% dans un ADN A) alors que les acides aminés polaires ont une préférence pour la forme A (40% contre 26% dans la forme B) (239) (cf. Figure 24).

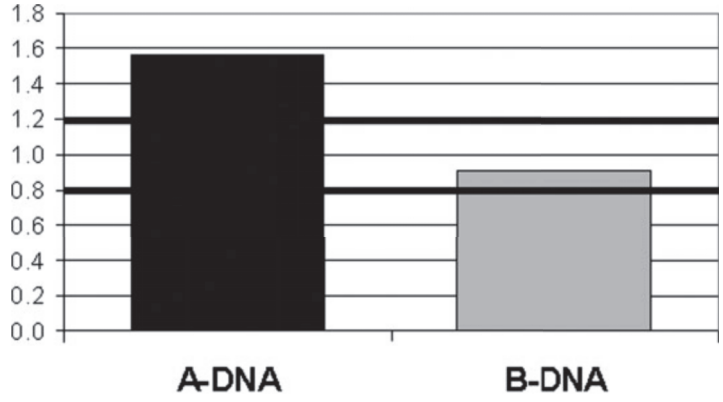


Figure 23 : Propension des protéines à interagir avec l'ADN B et l'ADN A. La propension correspond à la fréquence du type d'ADN au niveau du site d'interaction par rapport à la fréquence du même type d'ADN dans l'ensemble de la base de données et se calcule $P_i = \left(\frac{I_i}{\sum_i I_i}\right) / \left(\frac{T_i}{\sum_i T_i}\right)$. Avec I_i est le nombre de nucléotide dans la conformation i (i étant A ou B) en interaction directe avec les protéines, T_i le nombre de nucléotide dans la même conformation dans l'ensemble des complexes qu'il soit en interaction ou non avec les protéines. Une valeur supérieure à 1,2 correspond à un type de structure favorisée alors qu'une valeur inférieure à 0,8 correspond à un type de structure défavorisée (186).

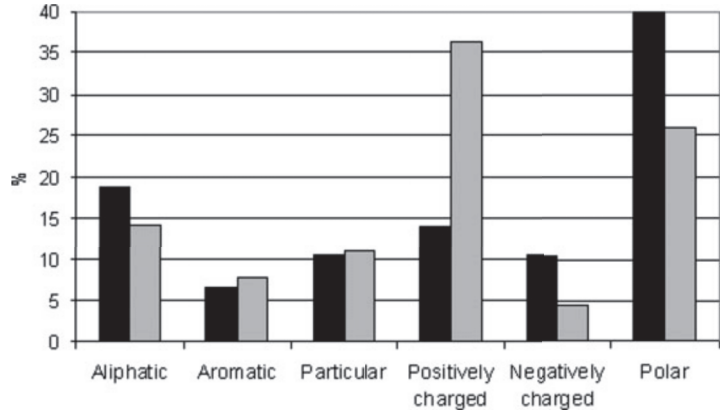


Figure 24 : Distribution des familles d'acides aminés au niveau du site d'interaction avec l'ADN réalisée sur 139 complexes (résolution < 2Å) (186). En noir sont représentés les acides aminés qui interagissent avec un ADN de type A et en gris un ADN de type B

La forme de l'ADN joue également un rôle dans l'exposition au solvant de certains groupes fonctionnels. Par exemple dans un ADN de type A, le désoxyribose qui est en conformation C3'-endo est d'avantage exposé au solvant par rapport à un ADN B où le sucre est en conformation C2'-endo (239). L'exposition du désoxyribose dans la forme A de l'ADN contribue à la spécificité d'interaction des protéines en doigt de zinc pour les séquences riches en nucléotides GC (240). L'exposition au solvant du désoxyribose semble jouer un rôle essentiellement dans le complexe TBP où 50 % de la surface d'interface ADN-protéine est formé par les désoxyriboses (87, 241). D'autres protéines reconnaissent des formes d'ADN intermédiaires A/B ; c'est le cas par exemple de certaines protéines en doigts de zinc (242–244)ou du facteur de transcription TFIIIA (245). L'ADN de type Z est une forme peu commune et qui possède une structure particulière (cf. Chapitre I.I.D). Les enzymes de la famille ADAR (enzyme qui convertit l'adénosine en inosine par désamination) et la protéine DLM-1 sont deux exemples décrits dans la littérature comme ayant la capacité d'interagir avec un ADN de type Z (246, 247).

La forme générale de l'ADN a un impact sur l'exposition au solvant du désoxyribose au niveau du petit sillon de l'ADN, mais a également une conséquence directe sur le potentiel électrostatique. La Figure 25 présente le potentiel électrostatique obtenu sur les formes A, B et Z avec la même séquence (succession de nucléotides GC). Malgré une séquence identique, le potentiel électrostatique de chacune des formes est très différent, ce qui va induire une différence significative au niveau des interactions électrostatiques (ponts salins et liaisons hydrogène) disponibles et par conséquent avoir un rôle important dans la reconnaissance.

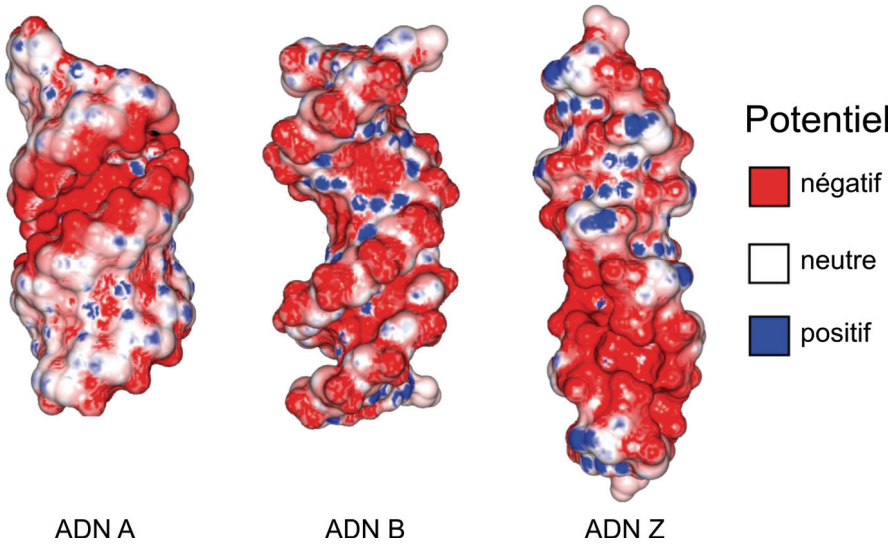


Figure 25 : Potentiel électrostatique sur la surface de l'ADN en fonction des formes A, B et Z (répétition de nucléotides GC) (248) calculé par résolution de l'équation de Poisson-Boltzmann avec Delphi (249, 250). Un Rouge le potentiel est négatif -15kT/e et en bleu un potentiel positif +15kT/e.

B. Déformations locales

Dans cette partie, les déformations locales sont les déformations qui touchent seulement un nombre restreint de nucléotides (typiquement de deux à cinq nucléotides).

1. Variabilité des paramètres hélicoïdaux

Depuis la résolution de la première structure de l'ADN, il est devenu évident que la séquence en nucléotides peut induire des changements de conformation dans la double hélice. Malgré l'accumulation de données structurales, il est difficile de quantifier l'impact de la séquence sur la structure de l'ADN, car même au niveau de courts fragments (par exemple, les tétranucléotides) toutes les séquences ne sont pas représentées dans la PDB (≈ 50% dans le cas des tétranucléotides). Pour observer l'impact de la séquence sur la structure et la dynamique de l'ADN, plusieurs travaux de dynamique moléculaire ont été réalisés par le « *Ascona B-DNA Consortium* » (ABC) (251–254) et par d'autres groupes (25, 255–257).

La séquence en nucléotide a une influence sur les paramètres structuraux de l'ADN à plusieurs niveaux. Les purines (A et G) et les pyrimidines (T et C) se distinguent au niveau des paramètres du brin phosphodiester (angle de torsions χ et des valeurs d'angle de phase de pseudorotation), mais également en termes de Propeller-twist, Opening et Buckle.

L'analyse des paramètres entre paires de bases, révèle que la nature des pas dinucléotidique, YpR (pyrimidine-purine), YpY (pyrimidine-pyrimidine), RpY (purine-pyrimidine) ou RpR (purine-purine) influence les propriétés structurales de l'ADN. Par exemple un pas nucléotidique YpR (TpG, TpA ou CpG) possède des valeurs de Twist plus faible, un fort angle de Roll positif par rapport aux autres associations possibles. Les mouvements des groupements phosphates associés à la transition BI/BII ont été observés quelque soit la nature du pas dinucléotidique, mais semblent facilités pour les dinucléotides pyrimidine-purine (258, 259). Cette transition BI vers BII influence localement la largeur des sillons (260), est associée à une diminution de l'empilement des paires de bases et semble favorisée lorsque l'activité de l'eau est réduite (réduction du nombre de molécule d'eau présent à l'interface (261).

Jusqu'à présent, la plupart des travaux menés sur la dépendance entre séquence et structure utilisaient le modèle du pas dinucléotidique (25, 255–257). Le groupe ABC est allé plus loin dans ces analyses, en étudiant également l'impact des bases qui encadrent ces pas dinucléotidiques (251). La présence des bases flanquantes et leur nature (purine ou pyrimidine) altèrent de manière significative l'ensemble des paramètres structuraux de l'ADN (cf. Figure 26).

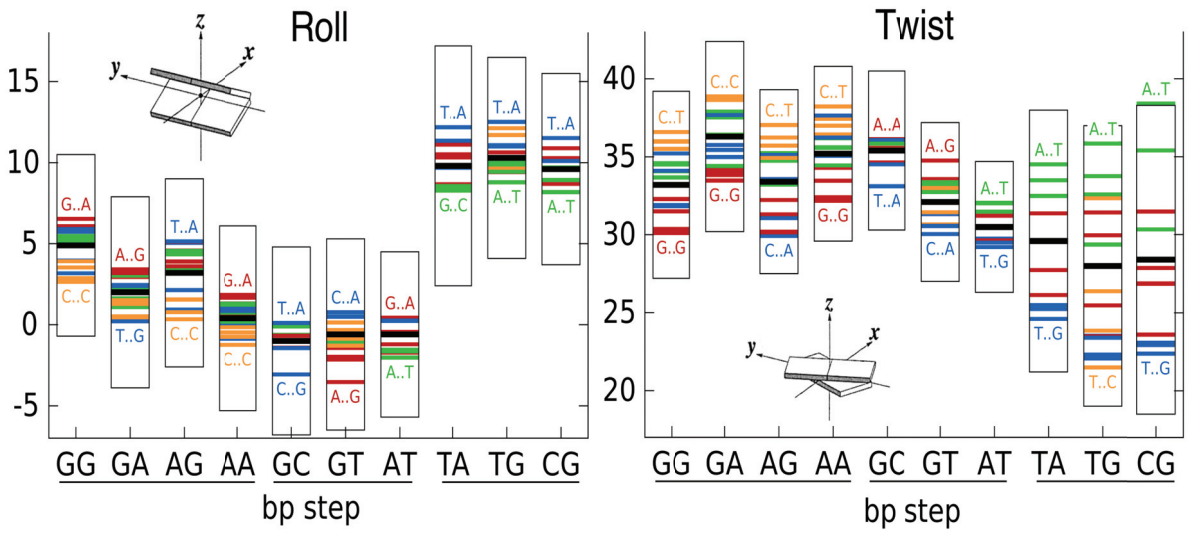


Figure 26 : Effet de la séquence dinucléotidique sur les paramètres Roll et Twist. Pour chacun des deux paramètres, la valeur moyenne est donnée par la ligne noire et l'écart type est délimité par la boîte pour chacun des 10 dinucléotides. L'impact des bases « flanquantes » est décrit par les traits de couleurs, avec en rouge lorsque les bases flanquantes sont du type purine-purine, en vert purine-pyrimidine, en bleu pyrimidine-purine et en orange pyrimidine-pyrimidine. Figure adaptée de (251).

Bandyopadhyay et ses collaborateurs ont démontré l 'effet des paires bases adjacentes sur l'empilement (262). Le pas dinucléotidique ApA a une structure relativement stable, alors que le pas CpA présente un comportement bimodal. Des études complémentaires

ont montré que les pas pyrimidine-purine en général sont moins stables et plus flexibles (263, 264). Enfin le groupe de Norberg et Nilsson a calculé le potentiel associé au processus d'empilement (265) permettant de définir dans une certaine mesure l'ordre de préférence des pas dinucléotidiques vis à vis du mécanisme d'empilement : RpR > RpY >YpR > YpY.

2. Déformation locale des sillons

La structure des sillons joue un rôle essentiel dans la spécificité d'interaction et a particulièrement été étudiée pour les facteurs de transcriptions de la famille Hox (266, 267) et pour le répresseur 434 (268).

L'analyse du complexe de la protéine Scr (Sex combs reduced) et de son cofacteur Extradenticle révèle l'importance de la largeur du sillon dans le phénomène de reconnaissance. Ce complexe protéique a été résolu par cristallographie avec le fragment d'ADN fkh250 et fkh250^{cons*} (identique à fkh250 à l'exception de trois paires de bases). La modification de ces trois paires de bases induit un sillon plus étroit dans le fragment fkh250 (séquence reconnue *in vivo* par Scr) que pour le fragment fkh250^{cons*} (séquence consensus de la famille Hox)(266, 267). Cette variation locale du petit sillon augmente le potentiel électrostatique à des positions spécifiques attirant l'arginine 3 et l'histidine 12 dans le petit sillon aidant ainsi à la reconnaissance (266, 267, 269). En règle générale, une séquence riche en dinucléotides AT crée un petit sillon plus étroit et donc un petit sillon plus électronégatif, exploité par les arginines ou lysines afin de compléter le mécanisme de reconnaissance spécifique grâce à la forme de l'ADN (270–272). Le complexe p53 possède deux dinucléotides ApT dont la géométrie est celle décrite par Hoogsteen (273). Dans ce complexe, la géométrie Hoogsteen rend le petit sillon plus étroit et plus électronégatif permettant ainsi à l'arginine 248 d'interagir avec le sillon pour la reconnaissance.

La forme du grand sillon peut également être utilisée pour améliorer la spécificité d'interaction. Généralement, les géométries du grand et du petit sillon sont étroitement liées ; la modification d'un sillon induit la modification du second. Par exemple pour la protéine humaine hRFX1 et la protéine extra cytoplasmique σ^E présente chez la bactérie E.*coli*, la déformation du petit sillon permet une meilleure complémentarité de forme entre les protéines et le grand sillon de l'ADN (82, 274, 275).

3. Formation de coudes dans l'ADN

Les coudes sont des cassures provoquées par rupture d'empilement entre deux nucléotides consécutifs, le plus souvent YpR (car plus flexible), et peuvent permettre d'optimiser le nombre de contacts ADN-protéine, mais également la complémentarité de forme. Dans de nombreux complexes c'est le pas TpA qui va être impliqué dans la formation de coudes (276–279) et permettre à différents acides aminés de s'intercaler pour former des liaisons hydrogène ou des interactions hydrophobes supplémentaires (75, 78, 88, 207, 280–282).

V. Représentation de la spécificité de séquence ADN-protéine

A. Création des logos

On visualise généralement la spécificité de séquence d'une protéine (ou séquence consensus) sous la forme d'un logo (283, 284) créé à partir de matrices de poids par position (PWM) (285, 286). Ce modèle très simple et intuitif permet de représenter les données expérimentales issues des méthodes *in vivo* et *in vitro.* Un logo est composé d'un empilement de lettres A, T, G et C pour chaque position de la séquence d'ADN étudiée. La taille relative de chaque lettre représente la fréquence d'apparition du nucléotide lors de l'alignement des séquences expérimentales où la protéine s'est liée. La hauteur totale (niveau d'information) à chaque position est exprimée en bits (cf. Figure 27). Dans le cas des acides nucléiques, le niveau d'information ne peut dépasser la valeur de 2 et se calcule pour la position i selon l'équation :

$$R_i = log_2(N) - \left(- \sum_{n=1}^{N} p_n log_2 p_n + e_n \right)$$

Avec **N** = 4 (nombre de nucléotides distinct), $\mathbf{p_n}$ le nombre de fois où le nucléotide **n** (A, T, C ou G) est observé et $\mathbf{e_n}$ est une correction apportée lorsque le nombre de séquences alignées est peu important. Dans le logo, la hauteur relative de chaque lettre à la position i est calculée par la relation :

$$Hauteur_i = p_n * Ri$$

Plusieurs groupes ont analysé des bases de données de sites de liaison de facteurs de transcriptions tels que TRANSFAC et JASPAR (287, 288), (289, 290) et ont mis en évidence une corrélation entre les différentes positions qu'il n'est pas possible de visualiser avec un logo simple, car avec cette convention de représentation, chaque position est considérée indépendante des autres positions du logo.

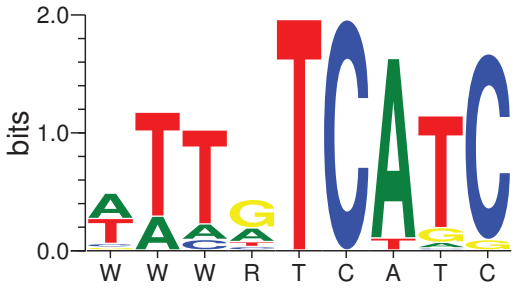


Figure 27 : Exemple de logo pour la protéine SKN-1 provenant de la base de données JASPAR (290). Sous chaque position du site de liaison, une lettre indique la préférence de nucléotide. Les lettres W et R indique une préférence pour les nucléotide A/T et A/G respectivement (nomenclature IUPAC (291))

B. Limitations de la représentation

Jusqu'à présent, la plupart des études menées sur les mécanismes de la reconnaissance ADN-protéine étaient basées sur la présomption que l'interaction était majoritairement contrôlée par les mécanismes d'interactions spécifiques entre les chaînes latérales de la protéine et les bases nucléiques de l'ADN. Il est maintenant évident que la reconnaissance acide aminé-base, la flexibilité et les propriétés structurales intrinsèques aux différentes séquences ne suffisent pas à expliquer entièrement la spécificité de reconnaissance. Très souvent les membres d'une même famille protéique partagent *in vitro* la même préférence de séquence et donc le même logo (292). Pourtant *in vivo*, ces protéines ne se lient pas sur les mêmes sites du génome. Par exemple, les facteurs de transcription Cbf1 et Tye7 se lient *in vitro* sur la même séquence d'ADN TCACTG (boîte E), mais *in vivo*, ces deux protéines ne se lient pas aux mêmes emplacements sur le génome(293, 294). En 2013, Bulyk et ces collègues (98) mettent en évidence que les bases qui entourent la boîte E influencent la structure 3D du domaine d'interaction. Les bases situées de part et d'autre du motif TCACTG modifient la largeur du petit sillon, le propeller twist, le Twist ainsi que les angles de Roll conférant une forme unique à la boîte E permettant à Cbf1 et Tye7 d'identifier la région du génome où ils doivent se fixer.

D'autres études réalisées sur les facteurs de transcription c-Myc, Max et Mad2 (protéines qui *in vitro* interagissent toutes les trois avec la même séquence CACGTG)

montrent que les bases flanquantes affectent la spécificité ADN-protéine au niveau du site consensus (98, 99, 295). Ce mécanisme a également été mis en évidence chez d'autres familles protéiques (296–298). L'apparition de ce nouveau concept de mécanisme de lecture indirecte et l'absence de code universel dans le mécanisme de lecture directe offre de nouvelles voies d'investigation et nécessite le développement de nouveaux outils capables de dissocier la partie reconnaissance directe de la reconnaissance indirecte afin de modéliser et prédire les interactions ADN-protéine avec plus de précision.

VI. Conclusion

Les études systématiques pour découvrir un code universel dans la reconnaissance ADN-protéine ont été réalisées au début des années 2000. Le manque de résultats concluants a mené les différentes équipes à réaliser non plus des études systématiques sur les mécanismes de la reconnaissance, mais à étudier les systèmes ADN-protéine au cas par cas, car chaque système semble unique ou du moins chaque famille protéique semble avoir un système qui lui est propre. Ces analyses améliorent notre compréhension des mécanismes qui régissent la reconnaissance spécifique mais reste incomplètes car elle nous provient d'une structure statique pour lequel il nous manque des informations sur la dynamique du complexe et de son interface

Chapitre III. Méthodologie

Ce chapitre va décrire les différentes méthodologies utilisées lors de la thèse afin d'étudier les propriétés mécaniques et dynamiques des complexes ADN-protéine. Selon les informations et le niveau de décomposition des interactions que l'on souhaite obtenir, il est nécessaire d'utiliser des modèles plus ou moins précis. Ces modèles vont du niveau atomique, où chaque atome est représenté, à des niveaux plus grossiers appelés « gros grains », où un groupe d'atomes est représenté par un pseudo-atome. En fonction de la taille du système à analyser, il est aussi important d'adapter les outils d'analyse. Ainsi pour de très petits systèmes (de l'ordre d'une dizaine d'atomes), il faut privilégier la méthode de mécanique quantique qui permet des calculs très précis. Cependant cette méthode ne sera pas utilisée au cours de cette thèse car les systèmes étudiés contiennent plusieurs milliers d'atomes. Pour des systèmes complexes et de taille plus importante, on utilisera des méthodes de mécanique moléculaire et de dynamique moléculaire tel que le logiciel AMBER (Assisted Model Building with Energy Refinement) (299). Ce programme permet d'étudier un système où chaque atome est décrit par ses coordonnées x, y et z dans un repère orthonormé.

I. La dynamique moléculaire

A. Les champs de force

En dynamique moléculaire, le modèle physique utilisé pour décrire les interactions entre atomes au sein d'un système est appelé champ de force. Le champ de force permet de décrire l'énergie potentielle d'un système à un instant donné lors de la dynamique. Dans la plupart des champs de force existants, l'énergie potentielle est décrite selon une équation qui regroupe les interactions entre atomes liés (c'est-à-dire les interactions représentant la déformation des liaisons, des angles de valences, des angles dièdres propres et impropres) ainsi que des termes faisant référence aux atomes non liés (l'électrostatique et les interactions de Lennard-Jones). Dans le cas du champ de force AMBER, champ de force utilisé lors de ces travaux, l'énergie potentielle est calculée selon l'équation ci-dessous :

$$E_{potentielle} = E_{termes\ liés} + E_{termes\ non\ liés}$$

$$\Leftrightarrow E_{potentielle} = E_{liaisons} + E_{angles} + E_{torsions} + E_{électrostatique} + E_{Lennard-Jones}$$

1. Les termes liés

(a) Le terme $E_{liaisons}$

Ce terme énergétique tient compte de la déformation des longueurs de liaisons entre atomes liés de manière covalente. Dans cette équation, k représente la constante de force de la liaison, r_{eq} la valeur d'équilibre et r la valeur à un instant donné.

$$E_{liaisons} = \sum_{liaisons} k_{liaison}(r - r_{eq})^2$$

(b) Le terme E_{angles}

Le terme E angles se calcule suivant l'équation ci-dessous et représente l'énergie de déformations des angles de valence entre trois atomes liés. Cette équation quadratique prend en paramètre la valeur de l'angle θ_{eq} qui est la valeur d'angle d'équilibre, k qui est la constante de force associée à la déformation de l'angle et θ l'angle calculé à un instant donné.

$$E_{angles} = \sum_{angles} k_{angle}(\theta - \theta_{eq})^2$$

(c) Le terme $E_{torsions}$

Ce terme représente la déformation des angles dièdres propres et impropres. Les angles dièdres propres correspondent aux angles formés par quatre atomes liés séquentiellement alors que les angles impropres décrivent quatre atomes qui ne sont pas nécessairement liés et permettent de maintenir la planéité de certains groupements. Chaque angle dièdre est décrit dans une ou plusieurs fonctions développées en série de Fourier. La valeur **n** correspond à l'ordre de la série de Fourier pris en compte pour l'angle ϕ considéré. Dans AMBER, cette variable n peut prendre la valeur de 1, 2 ou 3 en fonction de si elle décrit une conformation –cis-/-trans-, une double liaison planaire/non planaire ou une forme décalée/éclipsée respectivement (comme dans une liaison de type C-C d'un alkane). $V_{n,\phi}$ et $V_{n,}$ $\varphi_{impropre}$ correspondent à la constante de force associée à l'angle dièdre et γ est la valeur de la phase associée.

$$E_{torsions} = \sum_{\substack{dièdre \\ propre}} \frac{Vn,\phi}{2}[1 + \cos(n\phi + \gamma)] + \sum_{\substack{dièdre \\ impropre}} \frac{Vn,\phi_{impropre}}{2}\left[1 + \cos(n\phi_{impropre} + \gamma)\right]$$

2. *Les termes non liés*

Comme leur nom l'indique, les termes non liés vont calculer des énergies d'interaction entre deux atomes qui sont séparés par plus de deux liaisons. Ces termes sont au nombre de deux et sont l'énergie électrostatique et l'énergie de Lennard Jones.

(a) Le terme $E_{électrostatique}$

L'énergie électrostatique se calcule suivant la loi de Coulomb. Chaque atome i et j possède une charge partielle nommée qi et q_j, R_{ij} est la distance séparant les deux entités et $\varepsilon_{élec}$ est la constante diélectrique du milieu ($\varepsilon_{élec}$ = 1 avec des molécules d'eau explicites, $\varepsilon_{élec}$=80 dans un modèle d'eau implicite et typiquement $\varepsilon_{élec}$= 2 à 4 au sein des biomacromolécules).

$$E_{électrostatique} = \sum_{i,j} \left[\frac{q_i q_j}{\varepsilon_{élec} R_{ij}} \right]$$

(b) Le terme $E_{Lennard\text{-}Jones}$

L'énergie de Lennard Jones correspond à la résultante énergétique de l'attractivité et de la répulsion entre deux atomes (cf. Figure 28). Cette énergie est décrite par un potentiel dit 6-12. Le terme en puissance de 6 correspond à la force attractive, c'est-à-dire les forces de dispersions résultant de l'induction et de l'attraction de dipôles instantanés, alors que le terme 12 correspond à la force de répulsion entre les nuages électroniques. Le paramètre R_0 correspond à la distance la plus favorable énergétiquement entre les deux atomes et ϵ_{pot} correspond à la profondeur du puits de potentiel.

$$E_{Lennard-Jones} = \sum_{i,j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right]$$

$$A_{ij} = \epsilon_{pot} R_0^{12} \; et \; B_{ij} = 2\epsilon_{pot} R_0^6$$

$$R_0 = R_i + R_j \; et \; \epsilon_{pot} = \sqrt{\epsilon_{pot\,i} \epsilon_{pot\,j}}$$

Avec Ri et Rj correspondant aux rayons des atomes i et j, et $\epsilon_{pot\,i}$, $\epsilon_{pot\,j}$ correspondant au puits de potentiel pour les atomes i et j respectivement.

Il existe une seconde méthode équivalente pour calculer l'énergie de Lennard-Jones se calculant à partir de la distance minimale entre deux atomes D_{min}, correspondant à la distance ou l'énergie Lennard-Jones est égale à 0 :

$$D_{min} = \frac{R_0}{\sqrt[6]{2}}$$

La première méthode est utilisée par la suite de programme Amber lors des calculs des énergies Lennard-Jones alors que la seconde méthode a été utilisée afin de paramétrer les rayons de Van der Waals des pseudos-atomes nécessaires à la création des modèles de protéines modulables développées dans le Chapitre V.

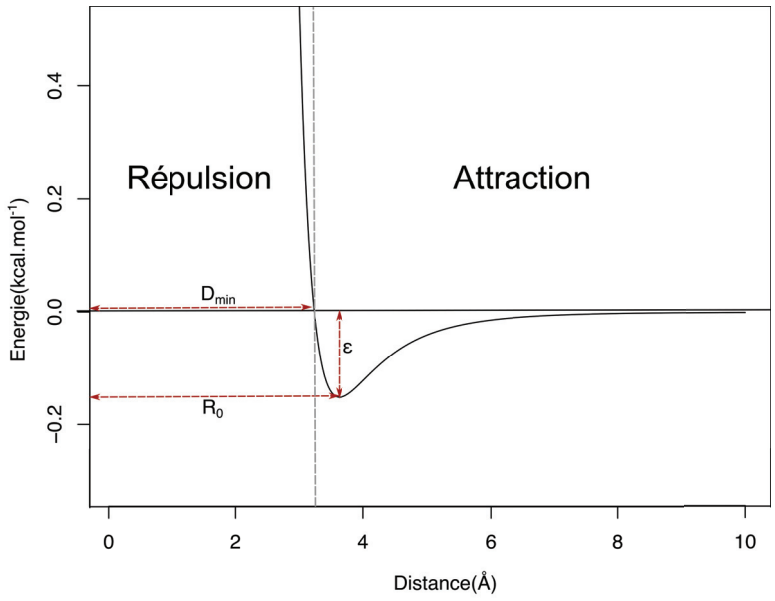


Figure 28 : Potentiel de Lennard-Jones. Représentation des paramètres R_0, ε_{pot} et, D_{min}. Une distance inférieure à D_{min} provoque une répulsion entre les atomes alors qu'une distance supérieure à D_{min} entraîne une attraction entre les atomes.

Afin de ne pas surestimer les interactions qui ont lieu à faible distance entre les atomes i et i+3, on applique des facteurs de mise à l'échelle de 1,2 pour les énergies électrostatiques et de 2 pour les énergies Lennard Jones, c'est à dire que l'on divise l'énergie calculée par le facteur de mise à l'échelle. Pour toutes les autres paires d'atomes (c'est-à-dire tout ce qui n'est pas i/i+3), ce facteur est de 1. Les énergies d'angles et de liaisons possèdent un facteur de remise à l'échelle de 1. Enfin, les énergies Lennard Jones et électrostatiques ne sont pas calculées entre les atomes liés (i et i+1) et entre les atomes séparés par un angle (i et i+2). Tous ces paramètres sont déterminés afin de reproduire au mieux les données expérimentales provenant de structures cristallographiques ou RMN ou les données théoriques issues de calculs de mécanique quantique sur des petites molécules.

B. La minimisation

Avant d'entreprendre une dynamique moléculaire, il est important de minimiser l'énergie du système étudié. Une structure minimisée correspond à un minimum énergétique local, souvent proche de la structure initiale. Cette étape est primordiale puisqu'elle permet de relaxer la structure et d'éliminer tous les encombrements stériques pouvant être présents et ainsi éviter des changements de conformation brusques lors du début de la dynamique moléculaire. Dans AMBER une structure peut être minimisée selon deux algorithmes dits de descente de gradient (steepest descent) (300) ou de gradient conjugué (301). Pour chaque pas de minimisation, on détermine les nouvelles positions des atomes $\boldsymbol{r}(k+1)$ du pas k+1 à partir des positions des atomes $\boldsymbol{r}(k)$ du pas k, de la direction de recherche $\boldsymbol{S}(k)$ et de la valeur du déplacement $\alpha(k)$ grâce à la formule :

$$\boldsymbol{r}(k+1) = \boldsymbol{r}(k) + \alpha(k)\boldsymbol{S}(k)$$

Les deux méthodes se différencient dans le choix du vecteur **S** et dans leur vitesse de convergence. La méthode de descente de gradient utilise une direction de recherche correspondante à –**g**(k), c'est-à-dire à l'opposé du gradient de l'énergie de l'itération actuelle. La recherche se fait donc en suivant la direction où la fonction d'énergie potentielle décroît le plus. La valeur du pas $\alpha(k)$ est ajustée à chaque itération, c'est-à-dire que si la valeur d'énergie diminue, $\alpha(k)$ est augmenté alors que si l'énergie augmente, $\alpha(k)$ est diminué. Cette méthode est très efficace et rapide lorsqu'on se trouve loin du minimum local, mais une fois proche du minimum local (au fond du puits de potentiel) l'algorithme converge très lentement. C'est pourquoi il est souvent couplé à l'algorithme de gradient conjugué Fletcher-Reeves qui lui converge rapidement lorsqu'on est proche du minimum. À la première itération, la direction de recherche est choisie comme précédemment. Ensuite, la direction de recherche s'écrit comme une combinaison linéaire des directions du gradient de l'itération actuelle et de l'itération précédente.

$$\boldsymbol{S}(k) = -\boldsymbol{g}(k) + b(k)\boldsymbol{S}(k-1) \; avec \; b(k) = \frac{\boldsymbol{g}(k).\boldsymbol{g}(k)}{\boldsymbol{g}(k-1).\boldsymbol{g}(k-1)}$$

La convergence est atteinte lorsque la différence d'énergie estimée pour le pas k+1 et celle du pas k est plus petite qu'une valeur définie par l'utilisateur (soit 10^{-4} kcal.mol^{-1}.Å^{-1} par défaut dans le programme AMBER).

C. Théorie de la dynamique moléculaire

1. Équation de mouvement

On peut étudier le comportement dynamique d'une macromolécule ou d'un complexe dans le temps grâce à la dynamique moléculaire. Cette méthode *in silico*, permet de suivre des changements de conformations qui se produisent typiquement jusqu'à l'échelle de la microseconde. Lors d'une dynamique moléculaire, la position des atomes d'un système et leurs vitesses en fonction du temps sont obtenues grâce à la résolution de l'équation différentielle d'ordre 2 issue de la seconde loi de mouvement de Newton.

$$\frac{\partial^2 \overrightarrow{x_i}}{\partial t^2} = \frac{\overrightarrow{F_i}}{m_i}$$

Dans cette équation, $\overrightarrow{x_i}$ correspond aux coordonnées x, y et z de l'atome **i**, **t** au temps et $\mathbf{m_i}$ à la masse de l'atome **i**. $\overrightarrow{F_i}$ correspond à la force exercée par l'ensemble des atomes du système sur l'atome **i**. Le programme AMBER utilise l'algorithme de Verlet (302) pour intégrer les équations du mouvement de Newton et calculer les trajectoires des atomes. C'est une méthode à l'ordre 3 dont le principe est le suivant.

On calcule la position de l'atome à l'instant t+Δt et t-Δt où Δt correspond au pas d'intégration.

$$\overrightarrow{x_i}(t+\Delta t) = \overrightarrow{x_i}(t) + \frac{\partial \overrightarrow{x_i}(t)}{\partial t}\Delta t + \frac{\partial^2 \overrightarrow{x_i}(t)}{2\partial t^2}\Delta t^2 + \frac{\partial^3 \overrightarrow{x_i}(t)}{6\partial t^3}\Delta t^3 + O(\Delta t^4)$$

$$\overrightarrow{x_i}(t-\Delta t) = \overrightarrow{x_i}(t) - \frac{\partial \overrightarrow{x_i}(t)}{\partial t}\Delta t + \frac{\partial^2 \overrightarrow{x_i}(t)}{2\partial t^2}\Delta t^2 - \frac{\partial^3 \overrightarrow{x_i}(t)}{6\partial t^3}\Delta t^3 + O(\Delta t^4)$$

En sommant les deux précédents termes, on obtient les coordonnées à l'instant t+Δt :

$$\overrightarrow{x_i}(t+\Delta t) = 2\overrightarrow{x_i}(t) - \overrightarrow{x_i}(t-\Delta t) + \Delta t^2 * \frac{\partial^2 \overrightarrow{x_i}(t)}{\partial t^2} + O(\Delta t^4)$$

Le problème avec cet algorithme de Verlet est qu'il ne génère pas directement la vitesse des particules. Même si la vitesse des particules n'est pas nécessaire pour connaître la position de la particule à t+Δt, il est nécessaire de calculer la vitesse notamment pour le calcul d'énergie cinétique ($\frac{1}{2}mv^2$). On peut utiliser la formule ci-dessous pour calculer la vitesse à l'instant t :

$$\frac{\partial \overrightarrow{x_i}(t)}{\partial t} = \left[\frac{\overrightarrow{x_i}(t+\Delta t) - \overrightarrow{x_i}(t-\Delta t)}{2\Delta t}\right] + O(\Delta t^2)$$

Au début de la simulation, la vitesse initiale de chacun des atomes est attribuée selon une distribution de Maxwell-Boltzmann pour une température donnée.

2. Choix du pas d'intégration Δt

Le pas d'intégration doit être en mesure de représenter le mouvement le plus rapide lors d'un mouvement moléculaire afin de garder l'énergie du système constante. Le mouvement le plus rapide correspond à la vibration intramoléculaire de la liaison d'un atome lourd avec un atome d'hydrogène et est de l'ordre de 1 femtoseconde. C'est pourquoi le pas d'intégration Δt dans les algorithmes ne peut pas être supérieur à cette valeur. Une possibilité pour l'augmenter est de figer les liaisons X-H dans la molécule en utilisant l'algorithme SHAKE (303). Le choix du pas d'intégration est primordial, car plus Δt est petit, plus la précision est grande, puisque l'erreur diminue en $O(\Delta t^4)$. En revanche en choisissant un Δt trop petit, les erreurs d'arrondis affectent le résultat, car l'on se rapproche de la précision machine. Or lors de l'intégration, les erreurs s'accumulent en proportion du nombre de pas effectués lors de la dynamique. Ce nombre de pas commande également la vitesse et le temps de calcul. En conclusion, il est nécessaire de choisir un pas d'intégration ni trop grand, ni trop petit afin d'optimiser la précision et le temps de calcul. En dynamique moléculaire atomique, ce pas d'intégration est généralement de 2 femtosecondes.

3. Condition thermodynamique de la simulation

En dynamique moléculaire, on travaille avec différentes informations comme le nombre de particules (N), l'énergie du système (E), le volume (V) qu'il occupe, la pression (P) ou encore la température de simulation (T). À partir de ces unités, trois grands ensembles thermodynamiques sont utilisés :

- l'ensemble microcanonique NVE,
- l'ensemble canonique NVT,
- l'ensemble isotherme-isobare NPT.

Dans un ensemble thermodynamique donné, trois paramètres sont maintenus constants. Dans le cas NPT, le nombre de particules N, la température T et la pression P sont maintenus constants tout au long de la simulation alors que dans NVT, c'est le nombre de particules N, le volume V et la température T qui sont constants. Dans notre cas, c'est l'ensemble NVT qui a été choisi pour les étapes de minimisation et d'équilibration et c'est l'ensemble NPT qui a été utilisé pour la phase de production de la simulation, car en condition expérimentale, les expériences se déroulent la plupart du temps à pression et température constante. Cela permet également de contrôler les dérives du couple température/pression provenant des erreurs de troncature des interactions à longue distance.

(a) Contrôle de la température T

Dans un système en mouvement, la vitesse d'un atome est corrélée à la température. D'après la théorie cinétique des gaz, l'énergie cinétique d'une particule est liée à la température et peut être obtenue par l'équation :

$$E_{cinétique} = \frac{1}{2}mv^2 = \frac{3}{2}k_B T$$

Où m est la masse de la particule, v^2 sa vitesse au carré, T la température et k_B est la constante de Boltzmann, qui vaut $k_B = 3{,}29.10^{-27} kcal.K^{-1}$. À partir de cette équation, il est possible d'obtenir la température pour un système donné en fonction de la vitesse des particules qui le compose :

$$T = \frac{\sum_{i=1}^{N} m_i \, {v_i}^2}{3Nk_B}$$

Avec m_i la masse de chaque atome, $\overrightarrow{v_i}^2$ la vitesse de chaque atome au carré, N le nombre de particules et k_B la constante de Boltzmann. Pour maintenir cette température constante au cours du temps, on utilise le thermostat de Berendsen (304) qui consiste à coupler le système à un réservoir externe de température. Ce thermostat permet au système d'échanger de la chaleur pour rééquilibrer la température en cas de gain ou de perte de chaleur. Avec ce thermostat, l'équation de mouvement de Newton va être modifiée afin d'avoir une remise à l'échelle des vitesses des atomes. Ce facteur d'échelle est obtenu par l'équation ci-dessous, où λ est le facteur d'échelle, τ_T le paramètre de couplage, Δt le pas d'intégration, T_o la température du bain de couplage, T la température choisie.

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)\right]^{1/2}$$

(b) Contrôle de la pression P

Le contrôle de la pression dans un système thermodynamique NTP, peut également se faire par le thermostat de Berendsen. Le principe est le même que celui de la température. Pour maintenir une pression constante, dans un environnement NPT, c'est le volume de V qui va devoir être remis à l'échelle. L'équation du contrôle de la pression est donnée ci-dessous.

$$\lambda = 1 - \frac{\beta \Delta t}{3\tau_P}(P_0 - P)$$

Où λ est le facteur d'échelle, τ_P le paramètre de couplage, Δt le pas d'intégration, P_o la pression du bain de couplage, P la pression désirée et $\frac{\beta}{3}$ le coefficient de compression isothermique.

4. Un environnement de simulation réaliste

(a) Solvant et ions

En dynamique moléculaire, il est possible de simuler des molécules dans le vide. Cependant, il est normalement important de tenir compte des effets du solvant lors de l'étude de macromolécules biologiques telles que les complexes ADN-protéine. En effet, les molécules d'eau jouent un rôle important pour diminuer la répulsion entre charges négatives portées par les groupements phosphates et pour stabiliser les interactions entre ADN-protéine. Chaque système ADN-protéine étudié lors de cette thèse a été simulé dans des conditions se rapprochant de la réalité, c'est-à-dire que le milieu a été hydraté avec des molécules d'eau explicite, en plus, en tenant compte de l'influence des sels, en ajoutant des ions à concentration physiologique (environ 150 mM).

Le modèle d'eau utilisé durant cette thèse est le modèle SPC/E (305). Ce modèle d'eau dit à trois sites se compose d'un oxygène chargé négativement : -0,8476e (ou e est la charge élémentaire égale à $1{,}6.10^{-19}$ Coulomb) et deux atomes hydrogènes possédant une charge de +0,4238e chacun. La géométrie de la molécule d'eau est maintenue rigide grâce à la présence de deux liaisons O-H de 1 Å et une liaison H-H de 1,6330 Å de constante de liaison 553 kcal.mol^{-1} et un angle HOH de 109,47° (angle d'un tétraèdre et non l'angle observé de 104,5°). Ce modèle SCP/E développé par Berendsen est en accord avec les

données expérimentales notamment en termes de densité et de constante de diffusion. De plus, le modèle SPC/E apporte une correction de la polarisation à la fonction d'énergie potentielle par l'ajout de 1,25 kcal.mol^{-1}. Enfin, ce modèle d'eau a été choisi et comparé à d'autres modèles d'eau plus sophistiqués tels que les modèles à quatre (TIP4P (306)) ou cinq (TIP5P (307)) sites en raison des temps de calcul plus longs nécessaires pour les modèles TIP4P ou TIP5P. Ce modèle d'eau a également été testé sur des simulations d'ADN libre ou en complexe avec des protéines et montre de bonnes corrélations avec les données issues de la PDB (308).

Le modèle de sel utilisé pour les différentes études réalisées ici est le chlorure de potassium développé par Dang (309). Ce modèle développé avec le modèle d'eau SPC/E montre des résultats en accord avec les données expérimentales. L'ion potassium a été retenu pour cette étude puisqu'il est présent à plus forte concentration au sein des cellules (310, 311). On note en passant que les ions K^+ interagissent plus fortement avec les sillons de l'ADN que les ions Na+ puisqu'ils se déshydratent plus facilement (312–315).

(b) Les conditions périodiques aux limites

Si on simule notre système sans contraintes, isolée dans l'espace, les atomes situés en périphérie vont se vaporiser. Si, au contraire, on introduit une boîte avec des murs impénétrables, on va créer des effets de bord artificiels qui peuvent perturber le système. Afin d'éviter ces effets et de simuler l'aspect infini de la solution dans laquelle est placé notre complexe ADN-protéine, on va répliquer la boîte dans toutes les directions et mimer un système sans limites. Ainsi une molécule sortant d'un côté de la boîte entrera par le côté opposé (cf. Figure 29).

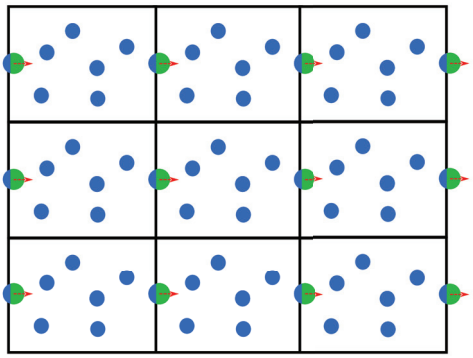


Figure 29 : Exemple de condition périodique en 2D. La molécule verte sort de la boîte centrale vers la boîte image à droite, mais rentre de nouveau dans la boîte centre par l'image de gauche.

Dans les conditions périodiques aux limites, les forces qui s'exercent sur un atome tiennent compte des atomes présents dans sa boîte, mais également des atomes présents dans la ou les boîtes adjacentes. Toutefois, il faut faire attention à ce qu'une molécule n'interagisse pas avec son image dans la boîte image adjacente. Pour cela, les interactions de chaque molécule avec son environnement doivent être tronquées à une distance R_c qui doit être au minimum égale à la moitié de la plus petite distance entre deux côtés de la boîte ou plus communément appelée la convention d'image minimum. Au-delà du rayon de coupure R_c, les interactions Lennard-Jones ne seront plus calculées. Cette méthode permet une bonne approximation dans le cas des interactions de Lennard Jones puisque l'énergie décroît rapidement avec la distance. En revanche, pour les interactions de très longue portée comme l'électrostatique, cette méthode n'est pas recommandée. Le problème de cette méthode est que le changement brutal d'énergie conduit à des artefacts dans la simulation puisque l'énergie ne peut pas tomber brusquement à 0 naturellement. Pour contourner le problème, on utilise une fonction qui va diminuer progressivement l'énergie vers 0 une fois le rayon de coupure dépassé.

Dans le cas des interactions électrostatiques, on utilise la méthode PME (« particle mesh Ewald ») (316–318). L'idée de l'algorithme est que chaque particule interagit avec toutes les autres particules de la boîte de simulation mais également avec les particules des boîtes images. Cette méthode permet de calculer l'énergie d'interaction électrostatique des systèmes périodiques efficacement. Pour calculer le potentiel électrostatique, on utilise la loi de Coulomb. Cependant, cette équation ne converge pas rapidement et un calcul entre paire d'atomes possède une complexité $O(N^2)$. Afin de palier à ce problème, l'algorithme PME utilise une astuce qui consiste à appliquer à chaque charge réelle du système une charge opposée suivant une fonction Gaussienne dont le pic est de même magnitude que la charge réelle. L'ajout de ce nuage électronique autour de la charge réelle permet d'observer de rapides variations de l'électrostatique pour de faibles valeurs de distance mais de faibles variations pour de grandes valeurs de distance. On peut alors définir un rayon de coupure. En dessous du rayon de coupure qu'on nomme espace réel l'énergie électrostatique est calculée par une équation de Poisson. L'énergie électrostatique dans cet espace est calculée entre chaque paire d'atomes en tenant compte de la charge réelle de chaque atome et du nuage électronique (défini par la fonction Gaussienne) et permet une convergence rapide et une erreur proche de 0 lorsque l'on se rapproche du rayon de coupure. Au-delà du rayon de coupure l'énergie est

approximée dans un espace de Fourier (ou espace réciproque) où les charges sont interpolées aux points de la grille grâce à la transformée de Fourier rapide. Le calcul dans l'espace de Fourier n'utilise pas les charges réelles mais les charges provenant de la distribution Gaussienne et permet de calculer efficacement l'énergie électrostatique. Cependant à ce stade, la somme des Gaussiennes dans l'espace réel a été calculée et comprend l'interaction de chaque Gaussienne avec elle-même et doit donc être retirée de l'énergie totale. Avec cette astuce de la sommation et du maillage d'Ewald, on réduit la complexité du calcul de l'énergie électrostatique de $O(N^2)$ à O(NlogN) diminuant ainsi considérablement les temps de calculs.

5. Protocole de simulation

Le protocole de simulation est celui utilisé par le consortium Ascona B-DNA pour les simulations d'acides nucléiques (251–254). Le système étudié (ADN seul ou ADN+protéine) est hydraté par des molécules d'eau de type SPC/E (305) et mis dans une boîte octaédrique tronquée. Le nombre de molécules d'eau ajouté doit assurer une couche d'hydratation d'au moins 10 Å. On ajoute ensuite suffisamment d'ions K^+ pour assurer l'électroneutralité du système puis on ajoute 150mM d'ions K^+/Cl^- pour être dans des conditions ioniques physiologiques. Ces ions sont ensuite placés de manière aléatoire dans la boîte et à 3,5 Å minimum les uns des autres.

Les simulations ont été réalisées avec la suite de programme AMBER 12 (299), en utilisant le champ de force PARM99 (319) et la modification BSC0 (320) qui corrige la surreprésentation des populations des angles α/γ=(g+/t) dans les acides nucléiques. Les paramètres de Dang (309) ont été utilisés pour les ions. La simulation utilise les conditions périodiques aux limites et les interactions électrostatiques sont traitées par l'algorithme PME (317) avec un rayon de coupure de 9 Å. Les interactions de Lennard Jones ont également été tronquées à 9 Å.

Le système est initialement minimisé avec une contrainte harmonique de 25 $kcal.mol^{-1}.Å^2$ sur les atomes de la protéine et de l'ADN. Le système est ensuite chauffé jusqu'à 300 K à volume constant pendant 100 ps. Ces contraintes sont ensuite relaxées de 5 à 1 $kcal.mol^{-1}.Å^2$ dans une série de minimisation (500 pas de minimisation steepest descent et 500 de gradient conjugué) et d'équilibration de 50 ps en terminant par deux équilibrations de 50 ps avec 0,5 $kcal.mol^{-1}.Å^2$ de contrainte et sans contrainte.

Chaque système est ensuite mis en simulation pendant 500 ns à température et pression constante (300 K et 1 bar) avec un pas d'intégration de 2 fs. L'environnement thermodynamique NPT est maintenu grâce à l'algorithme de Berendsen (304) et les liaisons X-H sont figées grâce à SHAKE (303).

II. Analyses des trajectoires en dynamique moléculaire

A. Analyses des liaisons hydrogène

À l'issue des 500 ns de simulation, les liaisons hydrogène formées entre les protéines et l'ADN ont été analysées avec la suite de programmes AmberTool15 (321). Un rayon de coupure de distance maximale de 3,5 Å entre les atomes lourds donneurs et accepteurs de liaisons hydrogène et un angle supérieur à 135° entre le donneur, l'hydrogène et l'accepteur ont été utilisés pour définir la géométrie d'une liaison hydrogène. Ensuite, le temps de vie de chaque liaison hydrogène a été calculé avec le programme Lifetime. Toute coupure de liaison hydrogène inférieure à 2ps était ignorée (322).

B. Analyses des paramètres hélicoïdaux : Curves+

Curves+ est un programme développé dans notre équipe permettant l'analyse des paramètres hélicoïdaux d'une molécule d'ADN provenant d'une structure unique ou d'un ensemble de structures issues de dynamique moléculaire (323, 324). Ce programme permet de calculer les paramètres hélicoïdaux, les angles du brin phosphodiester, mais également l'axe d'un fragment d'ADN et la géométrie des sillons. Curves+ définit un axe

hélicoïdal permettant de décrire le fragment d'ADN. Dans le cas d'une hélice parfaite comme décrite par le modèle de Watson et Crick ou des modèles basés sur la diffraction de fibre d'ADN, l'axe hélicoïdal est représenté par une ligne parfaitement droite. Chaque paire de bases peut donc être placée par rapport à la paire de bases précédente grâce à un simple mouvement de rotation et de translation le long de cet axe. Cependant ces modèles idéaux sont loin d'être majoritaires dans les complexes ADN-protéine, et Curves+ définit une fonction qui décrit l'irrégularité dans le positionnement relatif des bases pour analyser des fragments d'ADN plus ou moins déformés.

1. Système de référence

Dans Curves+ chaque base n'est pas définie par ses coordonnées cartésiennes, mais par un système de référence composé d'un point d'origine (E) et de 3 vecteurs (**bN**, **bL** et **bD**) définis à la conférence de Tsukuba (31). Le point d'origine E est placé à 4,702 Å de l'atome N1 pour les pyrimidines ou N9 pour les purines et forme un angle de 141,47° par rapport à la liaison glycosidique N-C1' (où N =N1 pour les pyrimidines et N=N9 pour les purines). Le premier vecteur (**bN**) de la référence est positionné sur le point d'origine et est orthogonal au plan formé par la produit scalaire (N1-C1') x (N1-C2) pour les pyrimidines ou (N9-C1') x (N9-C4) pour les purines et est orienté dans le sens 5' vers 3'. Le second vecteur (**bL**) du système forme un angle de -54,41° par rapport au vecteur **N-E** (ou N =N1 pour les pyrimidines et N=N9 pour les purines) et pointe en direction du grand sillon. Enfin le dernier vecteur (**bD**) est obtenu en effectuant le produit scalaire du **bN** x **bL** (cf. Figure 30). Lorsque la base est déformée, une base standard plane est ajustée sur les coordonnées cartésiennes à l'aide d'une procédure d'approximation aux moindres carrés avant de construire le système de référence (325).

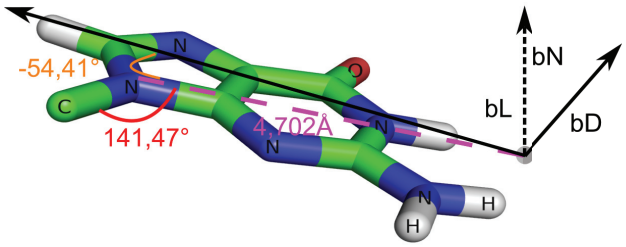


Figure 30 : système de référence pour une base dans Curves. L'origine du système de référence est représentée par le cercle gris et les trois axes bN, bL et bD sont représentés par des flèches noires pleines indiquant que l'axe est dans le plan de la base et par une flèche noire en pointillé lorsque l'axe n'est pas dans le plan de la base.

2. *Les paramètres intra paire de bases*

Chaque base du double brin d'ADN est désormais définie par un système de référence. Pour obtenir les paramètres hélicoïdaux pour une paire de bases, il faut définir la transformation géométrique (ou axe vissé) permettant de superposer le système de référence de la base i sur le système de référence de sa base complémentaire. Dans une hélice, il faut tenir compte de la pseudosymétrie entre les bases du brin Watson orientées 5' vers 3' et les bases Crick orientées 3' vers 5' en effectuant une rotation du système de référence de la base Crick de 180° autour du vecteur **bD**. La transformation géométrique se compose de mouvement de translation et de rotation. En effectuant une demi-rotation et une demi-translation autour de l'axe vissé, on crée un système de référence B_{intra} pour la paire de bases. La projection de l'axe vissé sur B_{intra} permet de définir les 3 paramètres de rotations (Buckle, Propeller et Opening) et les 3 paramètres de translations (Shear, Stretch et Stagger).

3. *Les paramètres inter paire de bases*

Chaque paire de bases i, i+1…i+n dispose désormais d'un système de référence B_{intra}. En suivant la même opération que précédemment, on calcule désormais la transformation géométrique entre deux paires de bases consécutives créant un nouveau système de référence B_{inter}. En projetant l'axe vissé sur B_{inter} on obtient les 3 paramètres de translations (Shift, Slide et Rise) et 3 rotations (Tilt, Roll, Twist) entre la paire de bases i et la paire de bases i+1.

4. *Les paramètres des sillons*

Pour calculer la largeur des sillons, Curves+ commence par remplacer chaque groupe phosphodiester par un segment de courbe lisse qui passe par tous les atomes de phosphore défini par un polynôme de degré 3 (spline cubique). Ensuite le programme calcule la distance entre tous les points uniformément répartis le long de chacun des deux brins permettant d'obtenir une matrice de distance comme sur la Figure 31. Sur cette figure la diagonale de la matrice représente la valeur de deux points à égale distance. De part et d'autre de cette diagonale, on observe des valeurs minimales formant des « vallées » qui vont définir la largeur des sillons. À gauche de la diagonale, on obtient la largeur du petit sillon et à droite de la diagonale celle du grand sillon (pour une double hélice main droite).

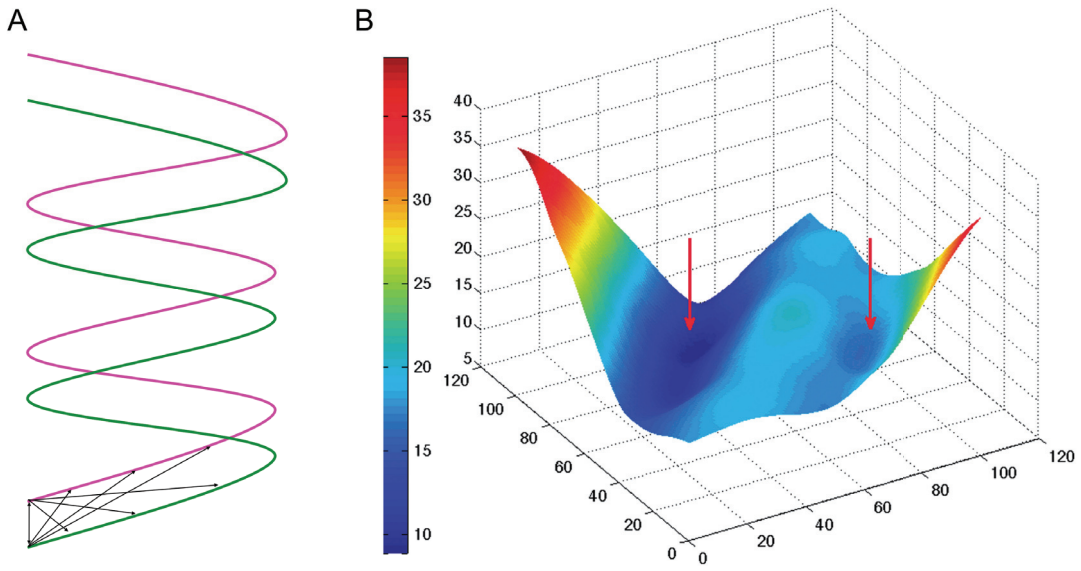


Figure 31 : Détermination de la largeur des sillons par Curves+. A) Segment de courbe lisse passant par tous les atomes de phosphore pour chacun des brins d'ADN (rose et vert) sur lesquels on va calculer toutes les distances entre les points marqués par les flèches. B) Matrice de distance (axe vertical en Å) entre les points répartis uniformément sur les deux brins de l'ADN (323). La diagonale de la matrice représente les points qui sont à égale distance. La largeur des sillons correspond au minimum (indiqué par les flèches rouges) de part et d'autre de la diagonale. En se déplaçant dans le sens horaire 5'->3' on mesure la largeur du petit sillon (flèche de gauche) alors qu'en se déplaçant dans le sens antihoraire 3'->5' on mesure la largeur du grand sillon (flèche de droite). La variation de distance est représentée par une échelle de distance de 9 Å (bleu foncé) à 38 Å (rouge foncé).

Maintenant que l'on possède la largeur du sillon au niveau d'une paire de bases, il est possible de déterminer sa profondeur. Au niveau d'une paire de bases, la profondeur du sillon est définie comme la distance entre le centre du vecteur qui définit la largeur phosphate-phosphate et le centre du vecteur définissant la paire de bases. Le vecteur définissant la paire de bases est construit en utilisant l'atome C8 des purines et C6 des pyrimidines. Afin de tenir compte du volume occupé par les atomes du brin

phosphodiester et des bases, la largeur et la profondeur des sillons sont réajustées en soustrayant 5,8 Å à la largeur du sillon et 3,5 Å à la profondeur du sillon.

Une dynamique moléculaire crée des dizaines de milliers de structures qu'il faut analyser par la suite. Curves+ est capable de lire une trajectoire de dynamique moléculaire et de calculer l'ensemble des paramètres hélicoïdaux d'une molécule d'ADN rapidement et de les stocker dans un fichier dans un format qui n'est pas très pratique à exploiter pour produire des graphiques. Le programme Canal utilise les sorties de Curves+ pour générer des fichiers plus lisibles pour générer des graphiques à l'aide du logiciel R (326).

C. Analyses ioniques : Canion

La nouvelle version de Curves+ permet d'enregistrer la position d'ions, des molécules d'eau ou de toute autre molécule autour de l'axe hélicoïdal de l'ADN lors d'une dynamique moléculaire. Curves+ fournit les coordonnées hélicoïdales curvilignes (327) de chaque ion qui sont ensuite analysées avec le programme Canion permettant de visualiser les résultats graphiquement en 1D, 2D et 3D. Dans ce système, la position d'un ion est définie en termes de position longitudinale (D), radiale (R) ou angulaire (A) (cf. Figure 32). La distribution D, définit la position d'un ion le long de la séquence d'ADN (en unité de paire de bases). La valeur R définit la distance (en Å) à laquelle se situe l'ion de l'axe hélicoïdal. La valeur A correspond à l'angle A (en degré) formé entre le vecteur formé par l'ion avec l'axe hélicoïdal et le vecteur qui pointe vers le brin 5' vers 3'. Dans Canion, la position des atomes C1' définit la région angulaire A correspondant au petit sillon (33°<A<147°) ou au grand sillon (147°<A et A<33°).

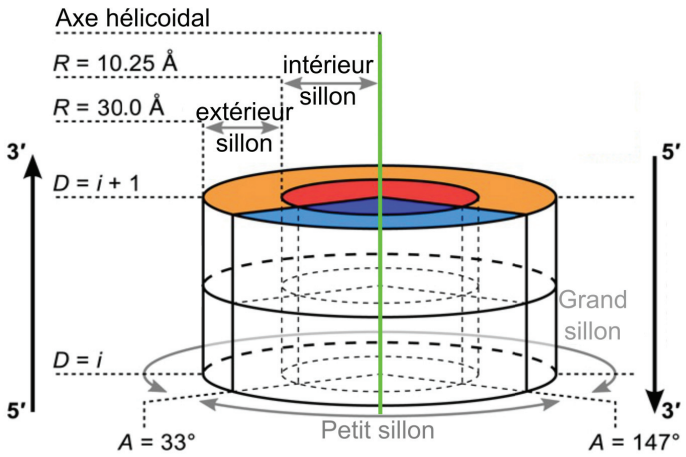


Figure 32 : Vu schématique d'un pas dinucléotidique formé par deux paires de bases i et i+1 (328). Visualisation spatiale des différentes composantes D, R et A permettant l'analyse des ions. L'espace angulaire est coupé en grand (rouge et orange) et petit sillon (bleu et cyan). L'analyse des ions peut être découpée en deux régions en fonction de la distance avec l'axe hélicoïdal ; la zone interne où la distance R entre les phosphates et l'axe hélicoïdal est inférieure à 10,25 Å et la zone externe R>10,25 Å.

L'ensemble des programmes Curves+, Canal et Canion utilisés au cours de la thèse sont disponibles gratuitement sur internet (https://bisi.ibcp.fr/tools/curves_plus/).

D. Clustering basé sur les distances atomiques de l'interface ADN-protéine

Afin d'étudier la dynamique des résidus à l'interface, un clustering des conformations obtenues durant la dynamique moléculaire a été réalisé à partir du nombre de contacts entre atomes lourds ADN-protéine formant l'interface. Les structures ont été extraites de la simulation toutes les 200 ps.

1. Calcul de la matrice de score pour une structure

Pour chaque structure extraite de la simulation, un score a été calculé entre chaque acide aminé i (avec i=1 à N) et chaque base j (avec j=1 à M) de la molécule l'ADN. Le score entre i et j est calculé en utilisant la distance entre atomes lourds grâce à la fonction sigmoïde ci-dessous. La fonction sigmoïde permet de lisser les résultats afin d'éviter d'introduire un rayon de coupure abrupte (1 en dessous du rayon de coupure et 0 au-delà). Ainsi une petite variation de distance de 0,1 Å supérieure par rapport au rayon de coupure qui pourrait être corrigé après une minimisation ne se verra pas attribuer un score nul, mais sera lissée par la fonction sigmoïdale.

$$s(i,j) = \frac{1}{1 + e^{a*(r_{ij}-b)}}$$

Où a=10 est le facteur qui contrôle la pente de la sigmoïde, et b=4,5 est le point d'inflexion où la dérivée seconde est nulle.

Chaque contact entre les atomes lourds obtient un score proche de 1 pour une distance r_{ij} inférieure à 4 Å et décroît pour atteindre environ 0 à 5 Å. Pour obtenir le score total pour une paire ij, il suffit de sommer tous les scores entre atomes lourds de la paire ij. Cette approche construit une matrice NxM pour chaque K structures issue de la trajectoire en dynamique moléculaire.

2. Calcul de la matrice de distance

Une fois les matrices NxM calculées pour chaque structure, un calcul de distance entre chaque matrice est effectué grâce à l'algorithme de Manhattan (329) (330–332) et le résultat est visualisé (cf. Figure 33). Afin d'identifier quels acides aminés sont impliqués dans la création des sous-états (blocs autour de la diagonale dans la Figure 33), il faut reproduire le calcul de distance acide aminé par acide aminé IxM, ou I est l'acide aminé avec I alant de 1 à N. Si l'interface ADN-protéine implique P acide aminé, il faut calcule 1 matrice NxM et P matrices individuelles.

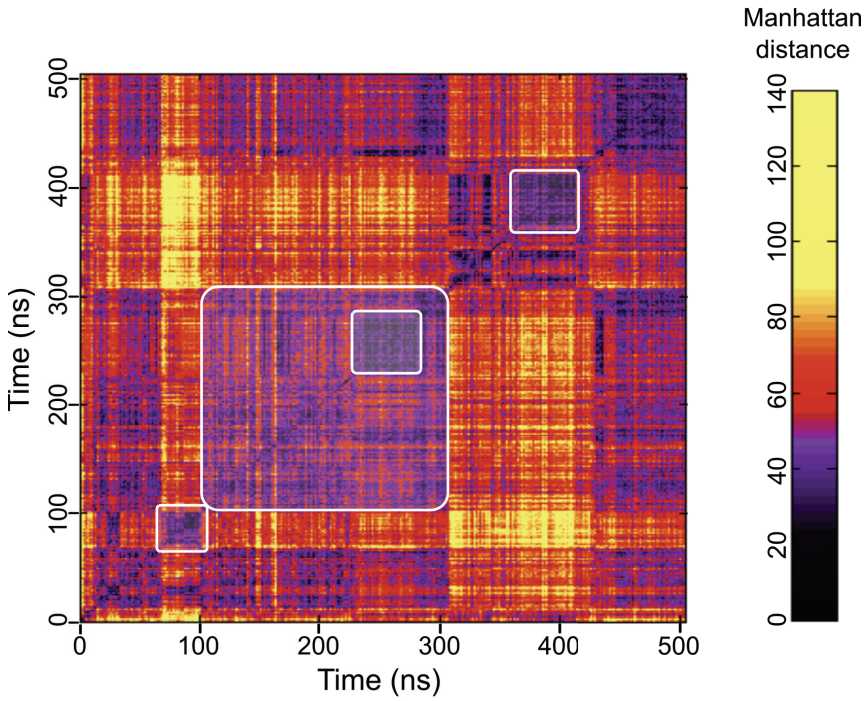


Figure 33 : Exemple de matrice de distance obtenue par l'algorithme de Manhattan. Les blocs blancs illustre des sous-états structuraux pour le complexe SKN-1.

3. Clustering des conformations

Grâce à la matrice de distance calculée précédemment, il est possible de constituer des groupes de structure ADN-protéine qui possèdent une interface similaire. Pour obtenir ces groupes, l'algorithme Ward (332) présent dans le logiciel R a été utilisé afin de minimiser la variance intra groupe. Une observation des structures est ensuite nécessaire pour valider le nombre final de groupes.

III. Analyses de la spécificité d'interaction ADN-protéine

A. Modélisation des acides nucléiques : JUMNA

JUMNA (**JU**nction **M**inimisation of **N**ucleic **A**cids) est un logiciel développé par Richard Lavery et Krystyna Zakrzewska permettant de modéliser et de minimiser l'énergie conformationnelle des acides nucléiques (333). Contrairement à de nombreux logiciels qui utilisent les coordonnées cartésiennes des atomes pour modéliser et minimiser le système, JUMNA utilise un système de coordonnées internes et hélicoïdaux permettant de réduire le nombre de degrés de liberté de la molécule afin d'accélérer les calculs et améliorer l'efficacité de la minimisation. Dans JUMNA, chaque brin d'acide nucléique est rompu en série de nucléotides 3'-monophosphate. Cette séparation des nucléotides s'effectue sur le brin phosphodiester au niveau de la liaison O5'-C5'. Ensuite chaque nucléotide est positionné dans l'espace suivant l'axe hélicoïdal selon 3 variables indépendantes de translations (Xdisp, Ydisp et Rise) et 3 variables indépendantes de rotations (Inclination, Tip et Twist). Dans JUMNA, les bases nucléiques sont rigides et les longueurs des liaisons sont fixes. La flexibilité interne de chaque nucléotide est représentée par la rotation de la liaison glycosidique, les angles dièdres ε et ζ du brin phosphodiester et la flexibilité du désoxyribose comportant deux angles dièdres (O4'-C1'-C2'-C3' et C1'-C2'-C3'-C4') et trois angles de valences (O4'-C1'-C2', C1'-C2'-C3' et C2'-C3'-C4') qui sont toutes des variables indépendantes. Enfin, les longueurs de liaison C4-O4' et O5'-C5' sont soumises à des contraintes harmoniques.

Rompre l'acide nucléique en nucléotides présente plusieurs avantages. Tout d'abord, les jonctions entre nucléotides consécutifs n'ont pas besoin d'être fermées dans

la structure de départ et, par conséquent, il n'est pas nécessaire de trouver les bonnes conformations internes des nucléotides afin d'ajuster un ensemble de paramètres hélicoïdaux choisi avant de commencer la minimisation d'énergie. Cette approche permet la recherche de structures d'acides nucléiques inhabituelles et irrégulières pour lesquelles aucune information conformationnelle n'existe à priori. D'autre part, les jonctions entre les nucléotides peuvent s'ouvrir lors de l'optimisation, ce qui permet des transitions entre différents états conformationnels, sans qu'il y ait de barrières énergétiques. Enfin, il n'est pas nécessaire de développer et de résoudre des équations décrivant la fermeture du système.

JUMNA optimise l'énergie du système avec un algorithme de minimisation de type gradient conjugué quasi-Newtonien (Harwell VA13A). Contrairement à l'algorithme du gradient conjugué simple qui n'utilise que l'information de la pente, la méthode quasi-Newtonienne utilise l'information de la pente et la dérivée seconde de l'énergie. La méthode quasi-Newtonienne nécessite un calcul analytique des dérivées premières de l'énergie de conformation par rapport à toutes les variables indépendantes. Ces dérivées sont obtenues en calculant les forces qui s'appliquent sur les atomes des différents nucléotides. Sous l'action de ces forces, les différents nucléotides se déplacent les uns par rapport aux autres par des mouvements de translations et de rotations. La dérivée par rapport à une variable de rotation donnée peut alors être obtenue en calculant le produit vectoriel des forces atomiques sur les atomes qui se déplacent avec les vecteurs reliant ces atomes à un point situé sur l'axe de rotation et en additionnant les composantes de ces produits dans la direction de cet axe. Dans le cas des variables de translation, le calcul des dérivées est simplifié puisqu'il suffit de sommer les forces atomiques des atomes en mouvement dans la direction de la translation. Il faut tout de même noter que cette procédure est un peu plus compliquée qu'il n'y paraît. En effet, modifier une variable hélicoïdale modifie la conformation du brin phosphodiester pour au moins une jonction entre deux nucléotides. Étant donné que les variables dépendantes contribuent également à l'énergie conformationnelle du système, leurs dérivées doivent également être calculées et additionnées à celles obtenues par les variables indépendantes. JUMNA intègre désormais le champ de force AMBER99 BSC0 qui est le champ de force le plus utilisé dans la modélisation, minimisation et simulation des acides nucléiques.

Dans JUMNA, les effets du solvant sont traités par un modèle implicite. Le programme utilise une fonction sigmoïdale ε(R) où l'interaction électrostatique entre deux charges est diminuée dans un solvant polaire (334).

$$\varepsilon(R) = D - \frac{D-1}{2[(RS)^2 + 2RS + 2]\exp(-RS)}$$

Où D=78 (constante diélectrique du solvant) et S=0,356. La gestion des effets du solvant est accompagnée d'une réduction de la charge nette des phosphates de -0,5e afin de simuler la présence de contre-ions. À noter que l'utilisation de la fonction ε(R) et de ces charges n'influence que très peu la conformation de l'ADN.

B. Préparation des données protéiques : PCHEM

PCHEM est un programme qui permet de préparer les données topologiques spécifiques aux protéines en définissant la connectivité entre les atomes, le type de chaque atome, leurs charges ainsi que leurs coordonnées. PCHEM définit également les variables internes du système. Le programme offre à l'utilisateur la possibilité de « geler » certaines variables internes dans le but par exemple de n'autoriser que les mouvements de chaînes latérales et de garder la chaîne principale rigide.

C. Enfilage moléculaire : ADAPT

1. Enfilage des séquences d'ADN

ADAPT est une méthode d'enfilage moléculaire développée dans l'équipe de Richard Lavery qui permet d'étudier l'impact de mutations sur la reconnaissance ADN-protéine (150). À partir d'une structure ADN-protéine, ADAPT va permettre d'enfiler toutes les séquences nucléiques possibles dans le site de liaison et calculer des énergies d'interaction et de déformation. Les sites de liaison impliquent entre 5 et 25 paires de bases soit de 4^5 à 4^{25} séquences possibles. Il est possible d'enfiler 4^{25} séquences dans ADAPT, mais cela demande une puissance et des temps de calcul importants. Afin de réduire la complexité de ces calculs et d'augmenter la vitesse de calculs, ADAPT découpe les calculs en fragments de N paires de bases chevauchants. Un ADN de M paires de bases est alors découpé en M-N+1 fragments. Sachant que chaque fragment peut contenir 4^N séquences, toutes les séquences peuvent être reconstituées en (M-N+1).4^N fragments. Par

exemple, si on découpe un ADN de 10 paires de bases en fragments de 5 paires de bases chevauchants, on peut reconstituer l'énergie des 4^{10} séquences possibles en calculant seulement 6144 séquences.

2. Protocole général

Déterminer la spécificité d'interaction ADN-protéine se réalise en plusieurs étapes. La première étape consiste à construire le complexe ADN-protéine à partir d'une structure cristallographique, RMN ou d'une structure provenant de dynamique moléculaire. Dans cette étape il faut aussi construire un ADN canonique ayant les mêmes séquences que celle du complexe qui permettra d'obtenir l'énergie non seulement d'interaction ADN-protéine, mais aussi de déformation de l'ADN lors de la dernière étape. Dans la structure étudiée, on enfile toutes les séquences possibles dans les différents fragments N puis on réalise une minimisation d'énergie et on calcule l'énergie. L'énergie des 4^{10} séquences du complexe et de l'ADN nu est obtenue en sommant les énergies de liaisons en utilisant un système de pondération qui tient compte des interactions base-acides aminés présents dans les segments chevauchant. En soustrayant l'énergie de la séquence de l'ADN nu à l'énergie du complexe, on obtient l'énergie de liaison du complexe. Enfin les séquences ayant une meilleure énergie de liaison (définies par un rayon de coupure) sont combinées afin de générer une matrice de poids qui sera visualisée par un logo.

3. Protocole appliqué

Le protocole ADAPT a été appliqué sur des structures provenant de dynamique moléculaire. Pour rétablir la planéité des bases, une minimisation de 1500 pas (500 en gradient simple et 1000 en gradient conjugué) grâce au programme AMBER et au champ de force PARM99 + BSC0 a été réalisée en amont du protocole ADAPT. L'enfilage moléculaire a été effectué sur des fragments de 4 ou 5 paires de bases de long. La minimisation a été réalisée avec JUMNA jusqu'à convergence (soit 2000 pas de minimisation en moyenne). Durant la minimisation les nucléotides du fragment d'ADN étaient flexibles à contrario des nucléotides à l'extérieur du fragment qui sont figés. Les chaînes latérales protéiques ont été rendues flexibles si elles étaient à une distance de maximum 20 Å des bases nucléiques du fragment. Les énergies obtenues ont ensuite été

converties en matrice de poids (PWM) et visualisées avec le logiciel Weblogo (283). Initialement le protocole ADAPT a été développé pour analyser la sélectivité de séquences à partir d'une seule structure. Dans le cadre de cette thèse, ADAPT a été étendu afin d'analyser la sélectivité provenant d'un ensemble de structures issues de la simulation.

IV. Outils de visualisation des structures

La visualisation des trajectoires de dynamique moléculaire a été réalisée avec l'outil VMD (335) et les structures de complexes ont été observées grâce à CHIMERA (336) et PYMOL (337).

V. Bases de données expérimentales

Les bases de données footprintDB (338), Jaspar (290) et TRANSFAC (339) ont été utilisées pour récupérer les matrices de poids (PWM) permettant de générer les logos de spécificité de séquences pour les différents complexes étudiés ici.

Chapitre IV. Étude par dynamique moléculaire atomique du processus de reconnaissance de quatre facteurs de transcriptions

Introduction et motivations

Ces dernières années, de plus en plus de complexes ADN-protéine sont venus enrichir les bases de données structurales. Bien que l'analyse de ces structures ait amélioré notre compréhension des interactions ADN-protéine, celle-ci reste incomplète puisqu'elles nous offrent une vision statique de ces interactions. Plusieurs groupes ont étudié expérimentalement le comportement dynamique des interactions ADN-protéine afin notamment d'expliquer les mécanismes qui permettent aux protéines de trouver et de se lier à leur cible et de comprendre les mécanismes qui leur permettent de différencier un site non spécifique d'un site spécifique. Les expériences RMN et les approches de résonnance paramagnétique ont été utilisées pour mieux caractériser les mécanismes de liaison non spécifique, et comprendre comment la diffusion d'une protéine le long de l'ADN, grâce aux ponts salins, peut induire la formation de contacts spécifiques avec les bases azotées (280, 340–343). Ces mécanismes ont également fait l'objet de plusieurs études théoriques (344, 345) et computationnelles (173, 175, 346–348) qui permettent de comprendre comment ces mécanismes contrôlent l'expression génique (5, 6, 349). La dynamique des interfaces ne se limite pas uniquement aux interactions non spécifiques et aux mécanismes de recherche du bon site de liaison, mais elle peut être importante également dans les interactions ADN-protéine spécifiques. Récemment, une étude en dynamique moléculaire sur les protéines TRF1 et TRF2, a montré que la dynamique de certaines chaînes latérales peut contribuer à la reconnaissance de plusieurs paires de bases, permettant ainsi d'expliquer la position alternative de certains résidus dans les structures expérimentales (350).

Dans le but de mieux comprendre les interactions ADN-protéine, nous avons utilisé la méthode de dynamique moléculaire couplée avec une méthode d'enfilage moléculaire ADAPT afin d'étudier l'aspect dynamique des interfaces et son impact sur les mécanismes de la reconnaissance spécifique ADN-protéine.

I. Systèmes étudiés

Durant cette thèse, quatre complexes ADN-protéine (TBP, SKN-1, SRY et le répresseur P22 c2) ont été étudiés par simulation en dynamique moléculaire. Le choix des complexes a été effectué selon différents critères. Le premier critère est la présence

de données expérimentales qui caractérisent le site de reconnaissance. Pour chacun des complexes, la séquence du site de liaison de la protéine est clairement définie. Enfin parmi les autres critères, chaque système doit posséder un motif d'interaction unique.

Le complexe TBP/ADN a été sélectionné car ce complexe qui initie le début de la transcription chez les eucaryotes a été très étudié. La protéine TBP possède une forte affinité (Kd ~3,5 nM (351)) pour la séquence consensus TATAWAAR (où W désigne les nucléotides A ou T et R désigne A ou G). Dans ce complexe, on sait également que la structure de l'ADN est l'élément principal de la reconnaissance et que l'interaction protéine-base de l'ADN (reconnaissance directe) ne contribue que peu à la reconnaissance. La protéine TBP se lie dans le petit sillon de l'ADN grâce à un large feuillet β, produisant une importante déformation, notamment un élargissement du petit sillon et un déroulement de la double hélice, et provoquant une importante courbure concave.

Le complexe SRY/ADN (« sex-determining Y protein ») est également une protéine responsable de la différenciation des gonades chez l'homme qui se lie dans le petit sillon de l'ADN à l'aide d'hélices α et de deux queues flexibles chargées positivement. L'interaction de SRY avec l'ADN provoque également une déformation de la double hélice (élargissement du petit sillon, diminution du twist), moins prononcée que dans le complexe TBP/ADN. Le complexe ADN-SRY a été déjà analysé au sein du laboratoire et nous possédons déjà certaines informations concernant la contribution des mécanismes de reconnaissance (347). Lors de ce précédent travail, l'impact d'éventuels changements conformationnels au niveau de l'interface n'a pas été étudié et fait l'objet d'investigation dans cette thèse.

Le complexe SKN-1/ADN est un complexe impliqué dans la différenciation des blastomères abdominaux de la nématode. Il a été choisi, car contrairement aux complexes TBP et SRY, cette protéine se lie à l'ADN dans le grand sillon grâce à une longue hélice α, homologue à celles rencontrées dans la famille protéique des leucines zipper. Une queue N-terminal cationique se lie au niveau du petit sillon en amont du site de reconnaissance sur des séquences riches en nucléotides A/T. Son rôle présumé est d'augmenter l'affinité du complexe. La protéine SKN-1 déforme peu l'ADN et la reconnaissance s'effectuerait principalement grâce au mécanisme de lecture directe. L'étude de la dynamique de l'interface de ce complexe est donc intéressante, car elle permettra d'étudier la relation entre la dynamique des résidus et la spécificité d'interaction.

Enfin le répresseur P22 c2 est impliqué dans le cycle lysogénique du bactériophage P22. Ce complexe a été sélectionné, car c'est le seul des quatre complexes qui est constitué d'un homodimère qui se lie à un long fragment d'ADN de 20 paires de bases grâce à deux motifs hélix-turn-hélix (un par monomère) dans le grand sillon de l'ADN. Chaque monomère se lie au niveau d'un demi site séparé par un tour d'hélice. La liaison du répresseur avec la molécule d'ADN ne produit qu'une faible déformation de cette dernière (légère augmentation du Twist et diminution de la largeur du petit sillon entre les deux sites de liaison, ainsi qu'une faible courbure convexe –vers le site de fixation de la protéine). À l'interface des deux monomères et de l'ADN, le complexe forme un canal électronégatif que des cations pourraient traverser, ce qui n'a pas été encore été étudié. P22 est un cas intéressant, car l'analyse de ce complexe peut nous donner des indications sur la reconnaissance d'un dimère. Chaque monomère reconnaît-il le demi-site de reconnaissance de manière synchrone ou asynchrone ?

Outre les différences mentionnées dans les paragraphes précédents, ces protéines ont été choisies en raison des mécanismes de reconnaissance qui prédominent. En effet, le nombre d'interactions directes entre les chaînes latérales des acides aminés et bases de l'ADN (liaisons hydrogène et interactions hydrophobes) diffère entre chaque complexe, modifiant probablement le rapport entre reconnaissance directe et indirecte. Dans le complexe TBP, il y a trois contacts polaires directs, quatre pour SRY, trois pour SKN-1 et enfin un par monomère dans le complexe ADN-P22 c2. La différence du nombre de contacts ainsi que le niveau de déformation de l'ADN suggère que l'équilibre entre le mécanisme de reconnaissance directe et le mécanisme de reconnaissance indirecte va être différent d'un complexe à l'autre.

II. Le complexe TBP

A. Rôle biologique

L'initiation de la transcription par l'ARN polymérase II (ARN pol II) nécessite la présence de plus de 70 polypeptides. La protéine qui coordonne l'activité de l'ensemble de ces polypeptides est le facteur de transcription IID (TFIID) qui va orienter la polymérase et les autres cofacteurs correctement. TFIID interagit avec la protéine TBP qui se fixe au niveau de la région promotrice nommée boîte TATA (car la séquence de ce promoteur est riche en nucléotides T/A). Cette séquence est localisée typiquement aux

environs de 30 paires de bases en amont du site du début de la transcription des génomes eucaryotes. La formation du complexe TFIID/TBP permet de réguler le recrutement des autres co-activateurs de la transcription notamment TFIIA, TFIIB et TFIIF formant ainsi le complexe de préinitiation de la transcription (352–354). La protéine TBP présente une particularité au niveau de l'extrémité N-terminale. En effet dans cette région, l'on observe une longue succession de résidus glutamine (entre 25 et 42) permettant de réguler la formation du complexe et donc de la transcription. L'élongation de cette chaîne de glutamine (plus de 45 glutamines) est associée à une maladie neurodégénérative nommée ataxie spinocérébelleuse de type 17. Cette maladie est associée à un dysfonctionnement de la transcription dont les mécanismes moléculaires sont encore inconnus (355).

B. Propriétés structurales

Le complexe TBP a été résolu par cristallographie avec une résolution de 1,9 Å et est disponible sur la PDB sous le code 1CDW. La structure cristallographique de la protéine contient 179 résidus. La protéine se compose de deux domaines quasi symétriques (environ 40 % de similarité et 30 % d'identité de séquence) très conservés dans tous les organismes eucaryotes alors que la partie N-terminale (qui n'est pas présente dans le cristal et qui représente ~150 résidus) est, elle, très variable. Cette protéine reconnaît spécifiquement une séquence riche en A/T de 8 paires de bases de long (boîte TATA). Lors de l'interaction entre les deux partenaires, la protéine adopte une forme dite de « selle », car elle recouvre l'ADN comme le ferait une selle sur un cheval. Chaque domaine est composé de cinq brins β formant un feuillet β antiparallèle et de deux hélices α (cf. Figure 34). L'alignement des deux domaines révèle une topologie identique puisque le RMSD (**R**oot **M**ean **S**quare **D**eviation) calculé sur la position des carbones α est de 0,92 Å. La présence de la protéine modifie de manière importante la structure de l'ADN notamment en diminuant d'environ 117° le Twist total pour l'ensemble des bases du site de reconnaissance TATAAAAG et courbe l'ADN de 66° dans la direction opposée à la protéine.

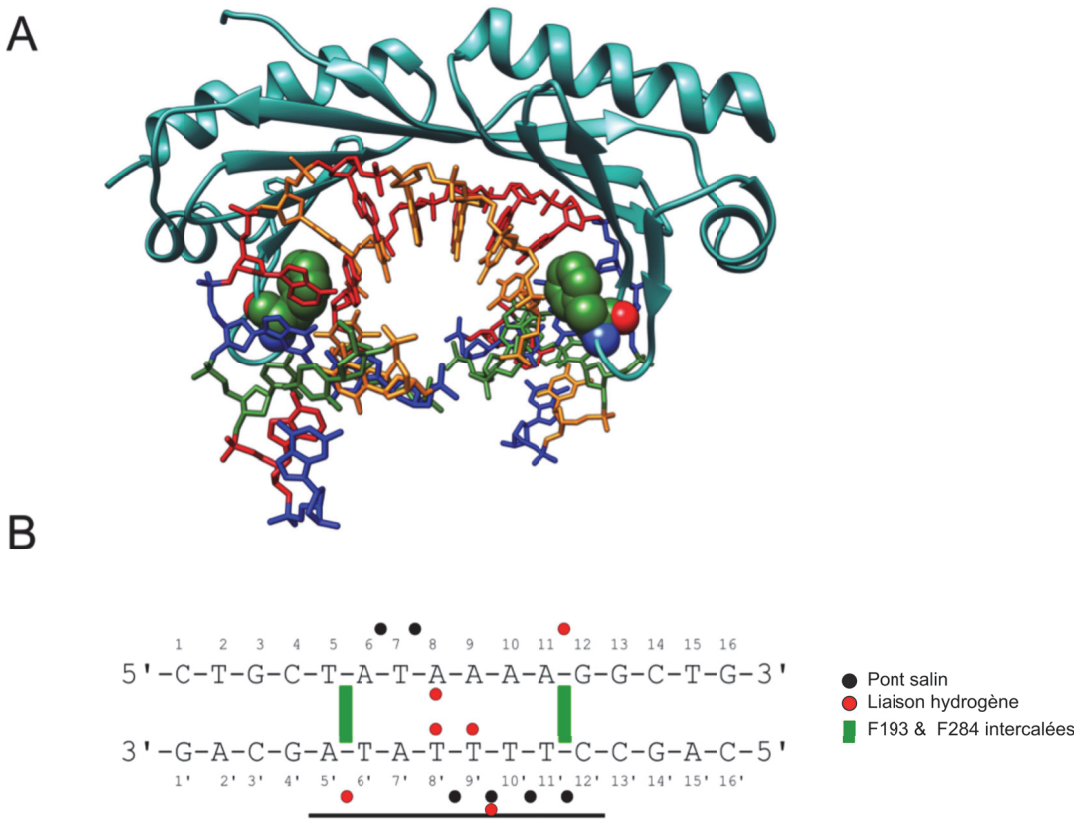


Figure 34 : A) Structure du complexe humain ADN-TBP. Les phénylalanines 193 et 284 partiellement intercalées au niveau des pas dinucléotidiques T5pA6 et A11pG12 sont représentées en sphère. L'ADN est coloré selon la séquence (A en rouge, T en orange, G en bleu et C en vert). B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 16 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-16' dans le sens 3'-5'

Au niveau de la surface d'interaction qui, rappelons-le, se situe au niveau du petit sillon de l'ADN, ce sont les interactions de van der Waals qui prédominent. En effet sur les 3100 Å^2 de surface enfouie au niveau de l'interface, environ 2300 Å^2 est constitué d'interaction hydrophobe (79 % de la surface protéique et 70 % de la surface nucléique). Au niveau des positions T5pA6 et A11pG12, deux phénylalanines Phe284 et Phe193 sont intercalées (cf. Figure 34) et provoquent une rupture dans l'empilement (« kink ») de ces paires de bases, créant un angle de Roll de 52,5 et 39,1° respectivement. Les phénylalanines 210 et 301 stabilisent la formation de ces cassures en formant de nombreuses interactions de van der Waals avec le désoxyribose des groupes A11 et T6'.

Malgré cette importante surface d'interaction, TBP n'effectue que peu de liaisons hydrogène avec les bases azotées de l'ADN. Dans la structure cristallographique, seule la base A8 du brin Watson et les bases T8' et T9' du brin Crick (cf. Figure 34) forment des interactions avec l'atome OG1 de la thréonine 309 et l'atome ND2 du résidu asparagine 163 respectivement (ND2 formant une double liaison hydrogène avec

l'atome O2 des bases T8' et T9'). Au niveau du brin phosphodiester, quatre arginines (Arg192, Arg199, Arg204 et Arg290) forment des ponts salins. Arg192 et Arg290 forment un double pont salin avec les nucléotides T10'(O1P)/T11'(O3') et T7(O1P)/A8(O1P) respectivement. Les arginines 199 et 204 forment des ponts salins avec les groupements phosphates du nucléotide T9' et T8' respectivement. La présence de ces ponts salins est probablement requise afin de neutraliser la charge négative des groupements phosphates et ainsi permettre de diminuer la répulsion entre groupements phosphates lorsque la distance entre eux diminue suite à l'importante courbure que provoque la protéine. À noter également, la présence de liaisons hydrogène entre les résidus T206, S212 et S303 avec les nucléotides T9'(O1P), G12(O1P) et A5'(O1P).

C. Site de reconnaissance de la protéine TBP

TBP se lie sur une séquence riche en nucléotides A/T, nommée la boîte TATA, localisée environ 30 paires de bases avant le site d'initiation de la transcription. Chez de nombreux organismes, le site de reconnaissance contient un motif 5'-TATA-3' très conservé (cf. Figure 35). On observe ensuite de petites variations de spécificité notamment au niveau des positions 5 et 7 où l'on retrouve une préférence soit pour une thymine soit pour une adénine selon les organismes, ou une préférence pour une guanine ou adénine pour la position 8. Il en résulte alors une séquence consensus de huit paires de bases dont la séquence est TATAWAWR (où W correspond aux bases A ou T et R à A ou G selon la nomenclature IUPAC (291)). Cette séquence est très conservée au sein des différentes espèces puisque l'interaction de TBP avec l'ADN engendre une importante déformation de la double hélice, rendue possible par la flexibilité accrue des séquences riches en A/T par rapport aux séquences riches en G/C.

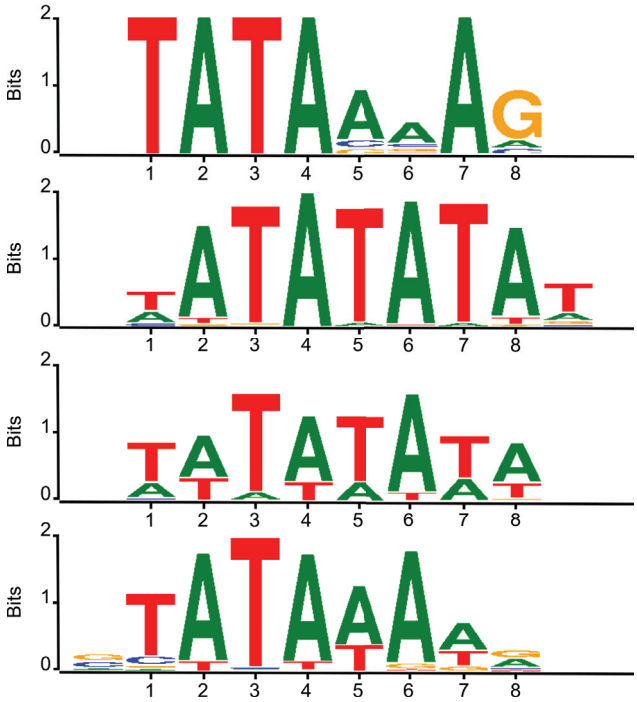


Figure 35 : Site de reconnaissance du facteur de transcription TBP pour plusieurs organismes eucaryotes issue de la base de données footprintDB (338). Avec de haut en bas *Arabidopsis thaliana (88)*, *Saccharomyces cerevisiae (356)*, *Mus musculus* (292) et *Homo sapiens (339)*.

D. Analyse de la dynamique moléculaire

Le complexe TBP a été analysé après 500 ns de simulation selon les paramètres définis dans le chapitre Méthodologie. Un ADN non lié avec la même séquence que celle présente dans le cristal a également été simulé afin d'observer les changements de structure, les modifications d'environnement, et plus particulièrement, le déplacement des cations lorsque la protéine est liée à l'ADN.

1. Structure de l'ADN

L'analyse de la trajectoire et la mesure d'un RMSD (calculé sur les atomes lourds de la protéine et de l'ADN) inférieure à 2 Å révèlent que le complexe a gardé une structure très proche de celle présente dans le cristal 1CDW. Au cours de la dynamique, l'ADN présente toujours une forte courbure d'environ 60° (contre 66° dans le cristal) dans la direction opposée à la protéine. Sur le fragment d'ADN non lié, on observe également une légère courbure intrinsèque à la séquence de 24° et dont l'axe hélicoïdal est orienté dans la même direction que la structure cristallographique. L'analyse comparative entre le complexe et l'ADN libre montre que plusieurs paramètres hélicoïdaux sont altérés. Ces modifications structurales sont plus importantes entre les positions 5 et 12 qui

correspondent au site de fixation de la protéine (cf. Figure 36). L'interaction de TBP au niveau du petit sillon provoque un élargissement de celui-ci de plus de 4 Å entre les positions 5 et 12. Cet élargissement du petit sillon est couplé avec un grand sillon qui devient alors plus étroit environ 12 Å pour l'ADN libre contre 9 Å en moyenne pour le complexe (à noter tout de même que le grand sillon dans le complexe peut être extrêmement étroit d'environ 5 Å entre les positions 7 et 9). La présence de la protéine influence également le niveau d'enroulement de l'ADN. L'analyse du paramètre hélicoïdal Twist indique que l'interaction provoque une diminution globale du Twist le long du site de liaison de 85° et donc un déroulement de la double hélice. Cette diminution du Twist est particulièrement importante au niveau du pas A8pA9 (centre du site de liaison).

La présence des phénylalanines intercalées aux positions T5pA6 et A11pG12 modifie de manière significative les paramètres hélicoïdaux inter-paires de bases, en particulier les paramètres Rise et Roll. Le paramètre Rise augmente respectivement de 1 et 2 Å pour les paires de bases T5pA6 et A11pG12. En dehors de ces deux positions la protéine n'affecte que peu le Rise des autres pas dinucléotidiques à l'exception là encore du pas central A8pA9 où l'on observe une diminution du Rise de 0,2 Å. L'intercalation des phénylalanines 193 et 284 provoque également une augmentation globale de l'angle de Roll de 148° le long de la séquence du site de fixation. Cette augmentation est cependant plus importante au niveau des sites d'intercalation, avec une augmentation de l'angle de Roll de 38,5° et 39,1° alors qu'entre ces positions, l'angle de Roll est augmenté de 14,5° en moyenne.

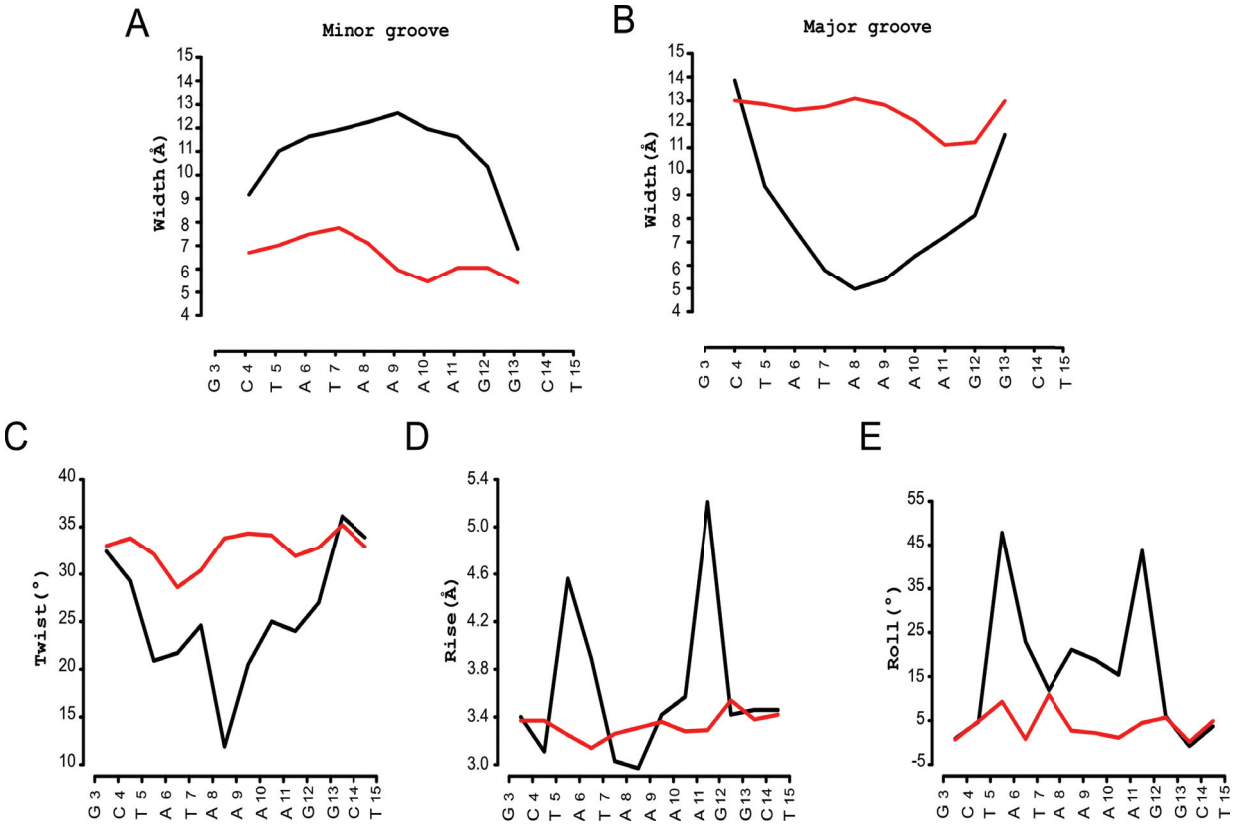


Figure 36 : Paramètres structuraux du complexe TBP-ADN (lignes noires) et de l'ADN libre (lignes rouges). A et B représentent la largeur du petit et du grand sillon de l'ADN (Å). C, D et E représente les valeurs moyennes des paramètres inter paires de bases Twist (°), Rise (Å) et Roll (°). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

2. Dynamique du complexe et des interactions ADN-TBP

Lorsqu'on s'intéresse à la dynamique du complexe, on s'aperçoit que les nombreuses interactions (ponts salins et liaisons hydrogène) qui ont lieu entre la protéine et le brin phosphodiester de l'ADN stabilisent et diminuent la mobilité des groupements phosphates de près de 25 % dans le complexe par rapport aux valeurs de l'ADN libre (RMSF ~1,75 Å) (cf. Figure 37A). Malgré l'importante surface d'interaction (~3000 Å^2), la dynamique moléculaire confirme le faible nombre de liaisons hydrogène observées expérimentalement (cf. Figure 37B) entre TBP et les bases nucléiques. Cinq nouvelles liaisons hydrogène absentes dans la structure cristallographique apparaissent lors de la dynamique moléculaire entre Thr218(OG1) et T10'(O2)/T9'(O2) (temps de présence de 12 % et 7 %) et Asn253(ND2) et A8(N3)/A9(N3) (temps de présence de 60 % et 78 % respectivement). On voit également apparaître de nouvelles liaisons hydrogène entre Ser257(OG) et A7'(O3') alors que la liaison hydrogène Thr218(OG1)-T9'5O4'), présente dans la structure cristallographique, disparaît complètement lors de la

dynamique. À cela viennent s'ajouter sept ponts salins (trois avec le brin Watson et quatre avec le brin Crick) (cf. Tableau 2). La plupart des interactions observées expérimentalement sont retrouvées en dynamique moléculaire. Cependant les interactions ne sont pas statiques et on peut observer la formation de nouveaux ponts salins ou liaisons hydrogène entre les groupements fonctionnels d'un acide aminé et différents atomes d'un même nucléotide. Les interactions électrostatiques se forment et se rompent rapidement avec une persistance d'environ 100 ps que ce soit pour les ponts salins ou pour les liaisons hydrogène. Cependant, certaines fois la durée de vie d'une liaison hydrogène est plus stable et peut persister plusieurs nanosecondes comme c'est le cas pour les serines 212 et 303 (cf. Tableau 2). Ces données *in silico*, sont en adéquation avec de récentes données expérimentales issues de RMN montrant que les ponts salins (en particulier ceux entre lysine et groupement phosphate) se forment et se rompent dans ces échelles de temps (357, 358).

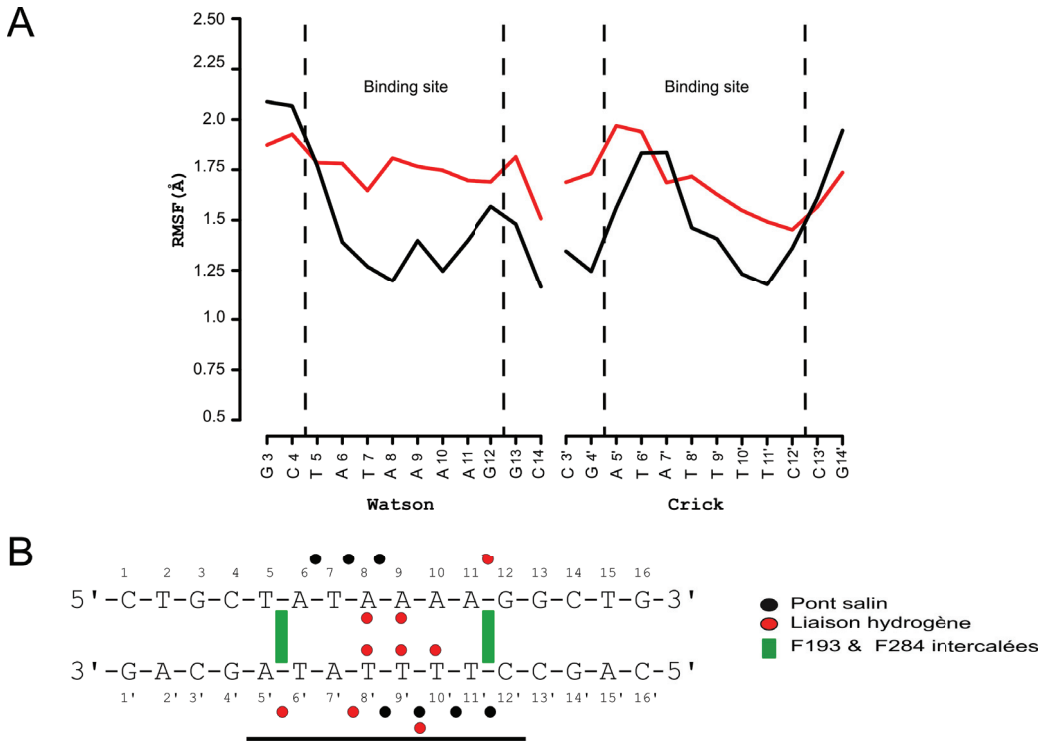


Figure 37 : Information sur la dynamique du complexe ADN-TBP. A) Moyenne du RMSF (Root Mean Square Fluctuation) des atomes de phosphore pour les brins Watson & Crick Crick pour le complexe (lignes noires) et de l'ADN libre (lignes rouges. Les lignes verticales en pointillés délimitent le site de reconnaissance spécifique. B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 16 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-16' dans le sens 3'-5'.

Tableau 2 : Temps de présence (% pres.) et temps de vie moyen (Lifetime en picoseconde) des ponts salins et liaisons hydrogène entre TBP et l'ADN (bases + brins phosphodiester) durant les 500 ns de simulation. La ligne noire marque la séparation entre ponts salins (au-dessus) et liaisons hydrogène (en dessous). Les interactions notées en vert sont présentes uniquement dans la structure cristallographique, celles en noir sont présentes dans la structure cristallographique et dans la simulation et celles en rouge sont présentes uniquement dans la simulation de dynamique moléculaire.

Protein	Backbone	%pres.	Lifetime	Protein	Base	%pres.	Lifetime
R192(NH1)	T10'(O1P)	54	100	N163(ND2)	T8'(O2)	94	1150
R192(NH2)	T10'(O1P)	48	120	N163(ND2)	T9'(O2)	77	125
R192(NH2)	T10'(O2P)	30	80	T218(OG1)	T10'(O2)	12	25
R192(NH2)	T11'(O3')	13	20	T218(OG1)	T9'(O2)	7	20
R192(NH2)	T11'(O1P)	16	165	N253(ND2)	A8(N3)	60	65
R192(NE)	T11'(O1P)	24	145	N253(ND2)	A9(N3)	78	125
R199(NH2)	T9'(O1P)	80	150	T309(OG1)	A8(N3)	56	50
R199(NE)	T9'(O1P)	74	160				
R199(NH2)	T9'(O2P)	18	25				
R199(NH2)	T10'(O3')	25	25				
R204(NE)	T8'(O1P)	13	55				
R204(NH2)	T8'(O1P)	21	155				
R204(NH1)	T8'(O1P)	18	120				
R204(NH2)	T8'(O2P)	4	24				
R290(NH2)	T7(O3')	33	30				
R290(NH2)	T7(O1P)	4	95				
R290(NH1)	A8(O1P)	92	1140				
R290(NH2)	A8(O1P)	58	200				
R295(NH1)	A9(O1P)	11	100				
R295(NH2)	A9(O1P)	8	100				
T206(OG1)	T9'(O1P)	78	1300				
S212(OG)	G12(O1P)	96	3950				
T218(OG1)	T9'(O4')	1	-				
S257(OG)	A7'(O3')	14	45				
S303(OG)	A5'(O1P)	98	11 450				

3. *L'environnement ionique*

La fixation de la protéine TBP sur le site de reconnaissance engendre un déplacement des ions présents dans le petit sillon de l'ADN libre (cf. Figure 38). Entre les positions T5 et G12, l'interaction ADN-protéine ne permet pas que des ions K^+ soient présents au niveau du petit sillon (molarité de 0). La présence de Glu282 en amont du site de fixation (position G3pC4) semble favoriser la présence d'ions dans le petit sillon, probablement afin de diminuer la répulsion électrostatique entre le glutamate et l'ADN. Au niveau du grand sillon devenu plus étroit lors de la fixation de la protéine, la concentration en ions potassium est augmentée tout le long du site de liaison. Dans le grand sillon de l'ADN, les bases G/C sont plus électronégatives que les bases A/T (et inversement dans le petit sillon). Le sillon étant plus étroit lorsque la protéine est présente, il n'est pas surprenant de voir la concentration ionique à la position G12pG13

augmenter (molarité de 1,6 M pour l'oligomère libre contre 7,5 M dans le complexe). Dans 75 % de la trajectoire, un ion potassium est présent au niveau de ce site dans le complexe, alors qu'il est présent uniquement 48 % du temps pour l'ADN isolé. De plus, la position 12-13 est occupée par deux ions pendant près de 15 % du temps (contre 5 % pour l'ADN libre). De part et d'autre du site de liaison spécifique, la présence de TBP ne modifie pas l'environnement ionique.

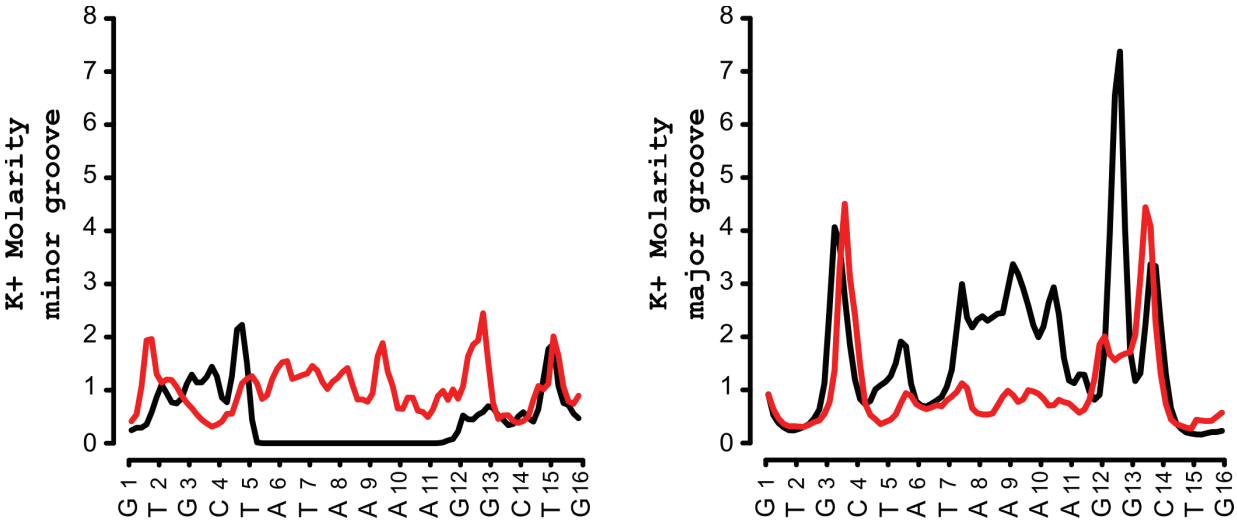


Figure 38 : Concentration ionique de potassium dans le petit et grand sillon de l'ADN pour le complexe ADN-TBP (ligne noire) et pour l'ADN libre (ligne rouge). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

4. Clustering et analyses de la spécificité de séquence

(a) Analyse de la spécificité de séquence

Le protocole d'analyse permettant le regroupement des structures ayant les mêmes interfaces ne permet pas d'obtenir des sous-états liés à la dynamique des acides aminés (cf. Figure 39). L'interface ADN-TBP semble très stable durant les 0,5 μs de simulation malgré le fait que les ponts salins et liaisons hydrogène se forment et se rompent rapidement avec le même nucléotide. La matrice de distance entre les acides aminés et bases de l'ADN est uniforme et ne permet pas d'établir des groupes (cf. Figure 39).

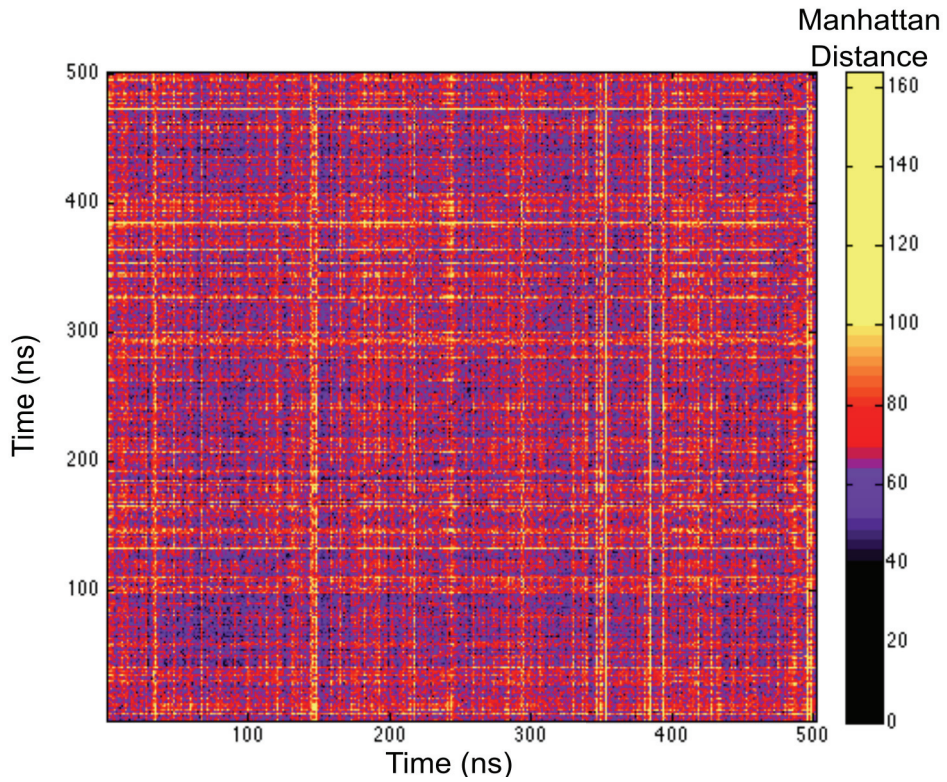


Figure 39 : Matrice de distance entre les différentes structures du complexe TBP provenant de la dynamique moléculaire.

Un jeu de données de 10 structures prises uniformément le long de la trajectoire a été créé et analysé selon la méthode d'enfilage moléculaire ADAPT (cf. chapitre Méthodologie pour plus de détails). L'enfilage moléculaire a été effectué sur chaque structure et un logo moyen a été créé à partir des données individuelles de chaque structure. Le logo moyen (MD Consensus) est en accord avec la séquence consensus expérimentale pour la TBP chez les vertébrés (cadran JASPAR) (cf. Figure 40). On peut alors mesurer l'information contenue (« Information content » ou IC) dans le logo qui correspond à la somme de la taille de chaque lettre pour chacune des positions et une corrélation (calculée à partir des matrices de poids PWM) afin de comparer la sélectivité de reconnaissance le long de la séquence consensus. Le signal de reconnaissance des bases le long du site de liaison de TBP est quasi identique entre la dynamique et la séquence expérimentale avec un IC de 10,1 et 9,3 respectivement. On retrouve également une forte corrélation entre les différentes positions avec localement des corrélations de 1/1/1/0,96/0,28/0,99/0,96 et 0,97 le long du site de liaisons pour les positions 5 à 12 et une corrélation globale entre les logos de 0,87. La méthode ADAPT est capable de décomposer l'énergie totale de la formation du complexe en énergie de déformation

(mécanisme de reconnaissance indirecte) et en énergie d'interaction (mécanisme de reconnaissance directe). Les logos issus de chacun de ces mécanismes sont présentés dans les cadrans *Deformation* et *Interaction* de la Figure 40. Comme attendu, la reconnaissance indirecte liée à la déformation de l'ADN joue un rôle essentiel dans la formation du complexe notamment pour les positions 5 à 8 qui correspondent à la séquence 5'-TATA-3'. La reconnaissance directe liée à l'interaction spécifique entre les bases nucléiques et les acides aminés reste cependant primordiale en particulier au niveau des positions centrales du site de reconnaissance. C'est au niveau de ces positions (A8-A10) que la majorité des liaisons hydrogène ADN-protéine sont présentes (cf. Figure 37 et Tableau 2).

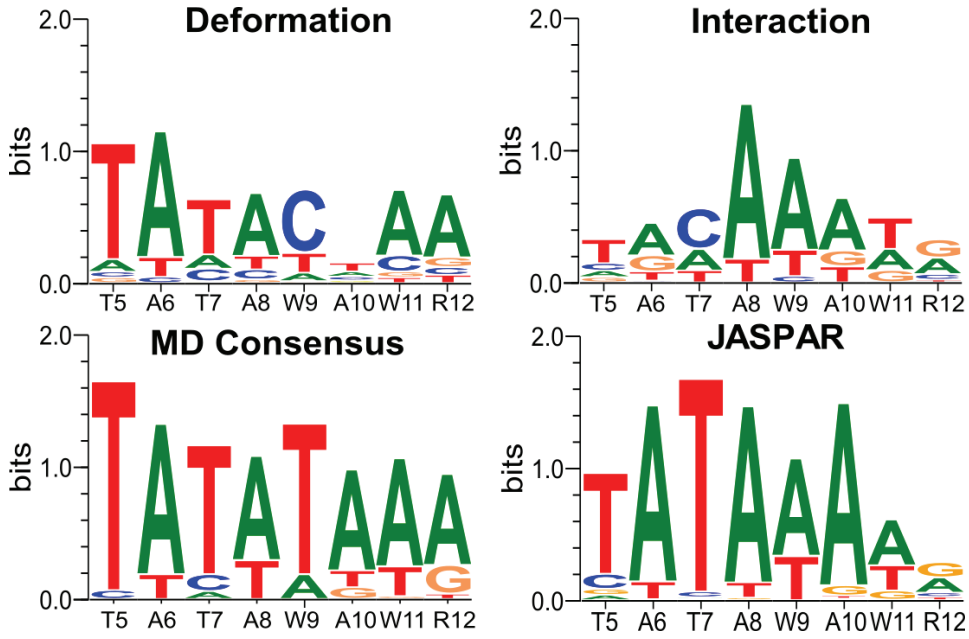


Figure 40 : Logos issus des matrices de poids (PWM) pour le complexe ADN-TBP lors de l'analyse de la trajectoire de dynamique moléculaire. La séquence en abscisse en dessous de chaque logo correspond à la séquence consensus expérimentale (W ≡ A/T, R ≡ A/G).

En conclusion, l'interface d'interaction entre TBP et l'ADN semble très stable et peu dynamique dans des temps de simulation de l'ordre de la microseconde. Au cours de la simulation, nous avons pu constater que les interactions spécifiques entre la protéine et l'ADN se forme et se rompe régulièrement (persistance de l'ordre de 0,1 ns), mais que cela n'affecte en rien le mécanisme de reconnaissance qui est principalement indirecte aux extrémités 5' et 3' du site liaison et directe au niveau des positions centrales 8-10.

III. Le complexe SRY

A. Rôle biologique

La protéine SRY (Sex-determining Y protein) est un facteur de transcription de la famille des SOX, présent chez les organismes ayant un système XY de détermination sexuelle. Cette protéine, codée par un gène du même nom présent sur le chromosome Y, contrôle la différentiation des gonades en testicules lors du développement du mâle. SRY est nécessaire et initie à elle seule la détermination du sexe mâle, en orientant le développement des cellules précurseurs en cellule de Sertoli au lieu des cellules de Granulosa présentes chez le sexe féminin. Il est à noter que la présence de ce gène est indispensable à la différenciation. En effet, il a été observé que des individus pouvaient avoir un phénotype sexuel ne correspondant pas à leur caryotype. Par exemple, on recense certaines femmes avec le caryotype XY et certains hommes avec XX. L'analyse de leur caryotype révèle que les hommes XX possèdent un gène SRY sur l'un de leurs chromosomes X et que les femmes XY possèdent un gène SRY muté le rendant non fonctionnel.

B. Propriétés structurales

La structure 3D du complexe SRY provient de données RMN et est disponible sur la PDB avec le code 1J46 (cf. Figure 41A). La structure SRY se compose de 85 acides aminés (domaine de liaison à l'ADN) lié à un oligomère de 14 paires de bases. La protéine SRY se lie à l'ADN au niveau du petit sillon via une hélice α dans la partie 3' du site liaison et une queue C-terminale flexible chargée positivement contenant quatre lysines (Lys73, Lys79, Lys81 et Lys85) et trois arginines (Arg75, Arg77 et Arg78) proches de la région 5' du site de reconnaissance. L'interaction de la protéine SRY provoque une déformation importante de la molécule d'ADN, notamment en engendrant une courbure de 43° et une modification localisée de certains paramètres (par exemple, une diminution du Twist entre les positions T8 et C11) et un élargissement du petit sillon. Dans le complexe, la surface d'interaction enfouie est d'environ 3000 Å. On compte six liaisons hydrogène directes au niveau de l'interface : entre le groupe hydroxyle de Tyr74 et A6(N3), entre le groupement guanidium de Arg7 et T6'(O2), entre Asn10(ND2) et G7'(N3), et enfin entre Asn32(ND2), Ser33 et 36(OG) et les atomes A10(N3), G11'(N3) et T10'(O2) respectivement. À cela il faut ajouter trois ponts salins formés par Lys44, Lys51 et Lys81

et 9 liaisons hydrogène entre acides aminés polaires non chargés et les brins de l'ADN. La Figure 41B représente schématiquement la position de toutes ces interactions électrostatiques le long de la séquence d'ADN.

La courbure de l'ADN est plus importante au niveau des positions 8 et 9 et est orientée dans la direction opposée à la protéine. À proximité de ces deux positions, trois résidus Met9, Phe12 et Trp43 consolident l'environnement hydrophobe et stabilisent l'isoleucine 13 qui est pseudo-intercalée entre les bases A8 et A9 et qui provoque une cassure importante avec un angle de Roll de 20° environ.

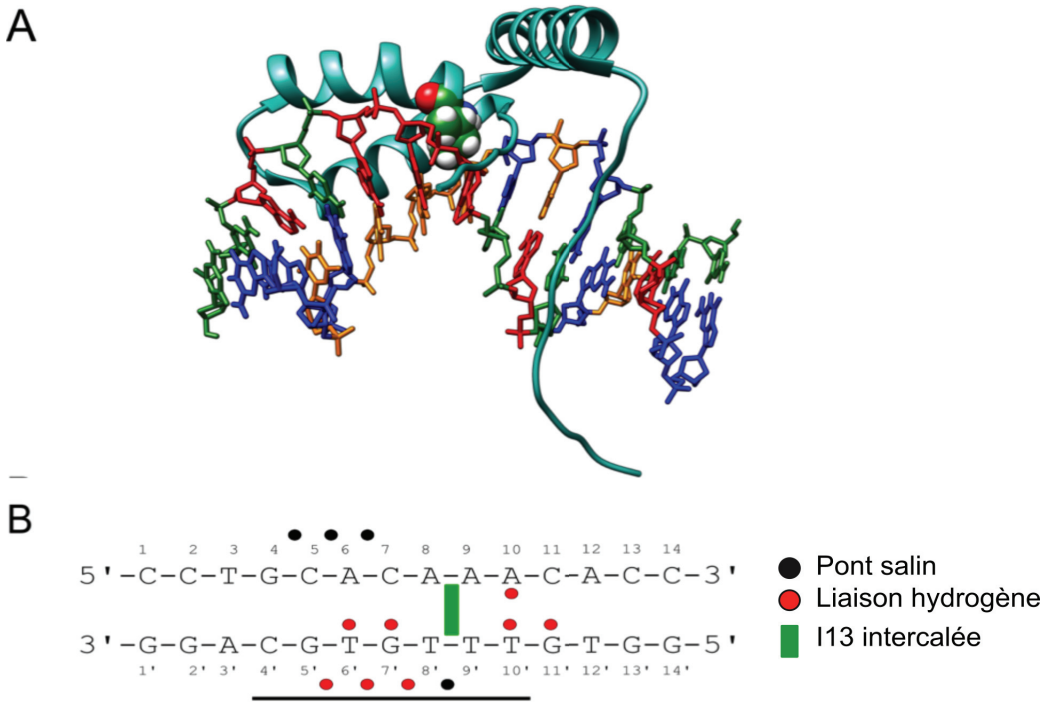


Figure 41 : A) Structure du complexe humain ADN-SRY. L'isoleucine 13 partiellement intercalée au niveau du pas dinucléotidique A8pA9 est représentée en sphère. L'ADN est coloré selon la séquence (A en rouge, T en orange, G en bleu et C en vert). B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 14 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-14' dans le sens 3'-5'.

C. Site de reconnaissance de la protéine SRY

Les protéines SOX/SRY se lient spécifiquement à la même séquence consensus WAACAAW (359, 360) (position 3 à 9 sur la Figure 42). Les séquences riches en nucléotides A/T sont plus flexibles et peuvent explorer davantage de conformations (sillon plus large, angle de courbure plus important) favorisant ainsi la formation du complexe avec SRY.

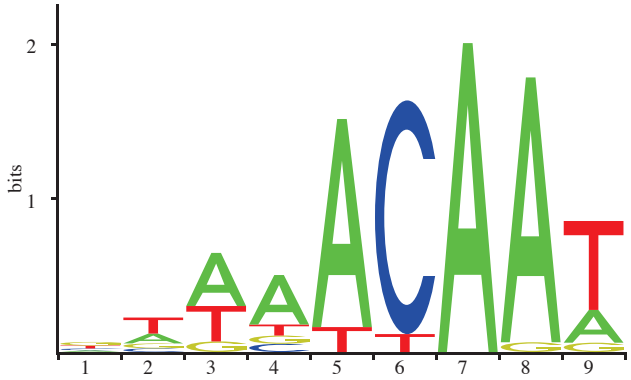


Figure 42 : Séquence consensus de la protéine SRY humaine obtenue par des expériences SELEX (360) .

D. Analyse de la dynamique moléculaire

1. Structure de l'ADN

Contrairement au complexe TBP précédent, le complexe ADN-SRY est plus flexible et dynamique notamment au niveau des résidus N et C-terminaux (constitué des résidus 1 à 9 et 76 à 85 respectivement) qui ne sont pas structurés en structures secondaires régulières (cf. Figure 43). Le RMSD moyen de ces queues terminales est d'environ 5 Å pour la queue N-terminale et de 10 Å pour la queue C-terminale.

Comme TBP, la présence de la protéine SRY modifie de manière importante l'axe hélicoïdal de l'ADN en induisant une forte courbure concave de 61° en moyenne lors de la dynamique moléculaire (contre 43° dans la structure RMN et 20° dans la dynamique de l'ADN libre).

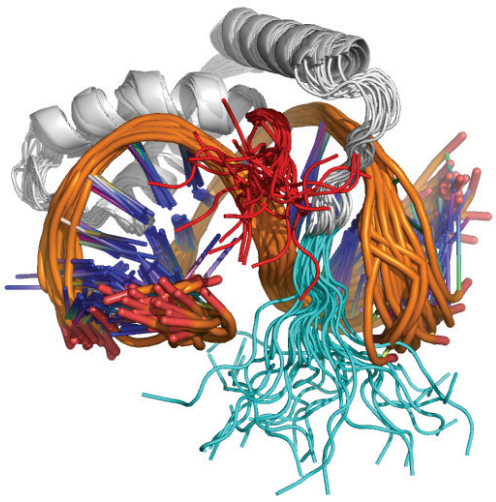


Figure 43 : Alignement de 28 structures du complexe SRY prises uniformément le long de la trajectoire de dynamique moléculaire. Les régions N et C-terminales sont colorées en rouge et cyan et le reste de la protéine en gris.

L'insertion de l'hélice α de SRY dans le petit sillon de la partie 3' du site de fixation provoque un élargissement de ce dernier de 6 Å, alors que le grand sillon devient plus étroit (cf. Figure 44). Ce changement dans la largeur des sillons s'accompagne d'une diminution locale du Twist entre les positions A8 et C11 de 41°.

La pseudo-intercalation de l'isoleucine 13 induit une augmentation du Rise de 1,8 Å au niveau du pas dinucléotidique A8pA9. La présence de Ile13 déstabilise l'empilement des bases A8pA9 en augmentant positivement l'angle de Roll entre les deux paires de bases (45° contre 4° pour l'ADN libre). Une augmentation moins importante de l'angle de Roll se produit entre les bases A9p10 et A10pC11 avec une augmentation d'environ 9° par rapport à l'oligomère libre.

Les paramètres décrivant la conformation des brins phosphodiester semblent peu affectés par la présence de la protéine. La conformation des désoxyriboses au sein du complexe est sensiblement identique à celle observée dans l'ADN libre à l'exception des désoxyriboses 8 du brin Watson et 9 du brin Crick qui adopte plus fréquemment une conformation C3'-endo (47 % et 31 % du temps dans le complexe contre 2 et 3 % dans l'ADN libre). La distribution des angles epsilon/zêta (décrivant les conformations BI/BII) est globalement la même le long du site de liaisons à quelques exceptions près. La protéine contraint la position 5 des deux brins à adopter une conformation BII (60 % du temps alors que dans la simulation d'ADN libre la position 5 est 15 % du temps dans la conformation BII). La base C11 sur le brin Watson adopte également une conformation BII (88 % du temps contre 5 % du temps pour l'ADN libre). La distribution des angles alpha/gamma reste inchangée dans le complexe.

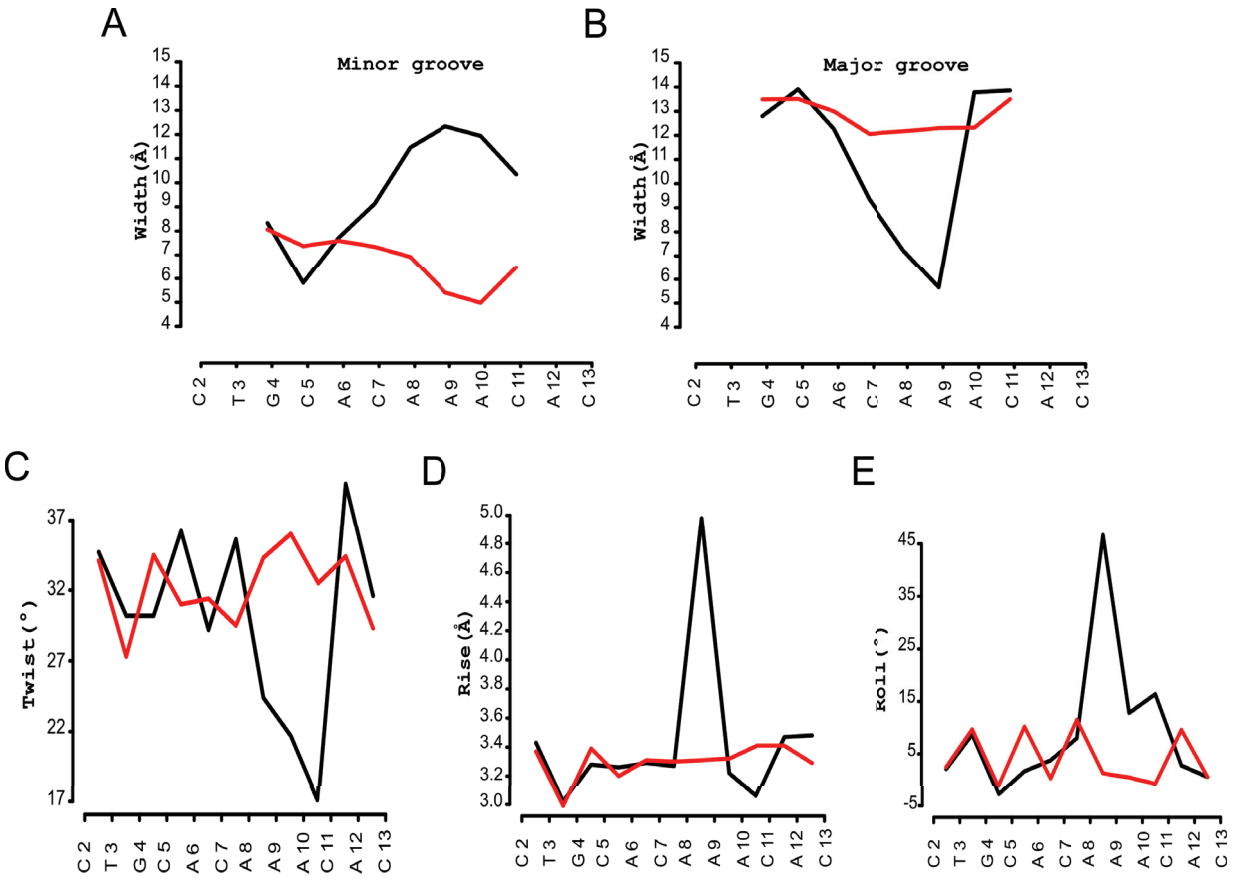


Figure 44 : Paramètres structuraux du complexe SRY-ADN (lignes noires) et de l'ADN libre (lignes rouges). A et B représentent la largeur du petit et du grand sillon de l'ADN (Å). C, D et E représente les valeurs moyennes des paramètres inter paires de bases Twist (°), Rise (Å) et Roll (°). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

2. Dynamique du complexe et des interactions ADN-SRY

La formation de ponts salins et de liaisons hydrogène (entre le brin phosphodiester de l'ADN et la protéine le long du site de fixation) diminue la dynamique du brin comme le laisse suggérer la baisse de 0,5 Å en moyenne du RMSF des atomes de phosphore par rapport à l'ADN libre (cf. Figure 45A). Au cours de la dynamique, huit groupements phosphates du brin Watson et cinq du brin Crick sont impliqués dans la formation de ponts salins (cf. Figure 45B). On compte également la présence de sept liaisons hydrogène avec les groupements phosphates et ce, principalement au niveau du site de reconnaissance. Ces nombreuses interactions permettent de stabiliser les brins phosphodiester et donc l'interaction ADN-protéine. Les ponts salins formés avec le brin Watson sont plus stables avec des temps de présence de 73 à 100 % du temps (en faisant abstraction du changement des atomes en interaction pour un même couple acide aminé/groupement phosphate) alors que les ponts salins formés avec le brin Crick sont

plus labiles (entre 9 et 45 % de présence). Malgré la persistance de certains ponts salins lors de la simulation, le temps de vie moyen de ces interactions est de l'ordre de la centaine de picosecondes, comme observé expérimentalement (357). Lors de la dynamique, SRY forme un nombre important de contacts directs (par des liaisons hydrogène) avec les bases azotées. On en compte plus de 12, dont 10 se situent au niveau du site de reconnaissance (cf. Figure 45B). Le temps de vie typique d'une liaison hydrogène est de l'ordre de 100 ps à l'exception des liaisons hydrogène formées entre les bases C11(O2), T6'(O2), T8'(O2), et T10'(O2) avec les acides aminés Asn32(ND2), Asn10(ND2), Arg7(NH1) et Ser36(OG) dont le temps de vie atteint plusieurs nanosecondes (cf. Tableau 3). Comme observé précédemment dans le complexe TBP, la dynamique de l'interface permet l'apparition de nouveaux ponts salins et de nouvelles liaisons hydrogène (en rouge dans le Tableau 3) au niveau du site de reconnaissance ou des bases adjacentes. Plusieurs interactions initialement présentes dans la structure RMN et notamment des liaisons hydrogène impliquant les résidus Asn10, Asn32, Gln62, Lys79 et Ser33 ne sont pas retrouvées lors de la dynamique moléculaire. Certaines de ces chaînes latérales se sont repositionnées pour interagir non plus avec le brin phosphodiester comme c'est le cas des asparagines 10 et 32 dans la structure RMN, mais pour former des interactions avec les bases azotées G7'(N3)/T8'(O2) et A10(N3)/C11(O2) dans la simulation (cf. Tableau 3).

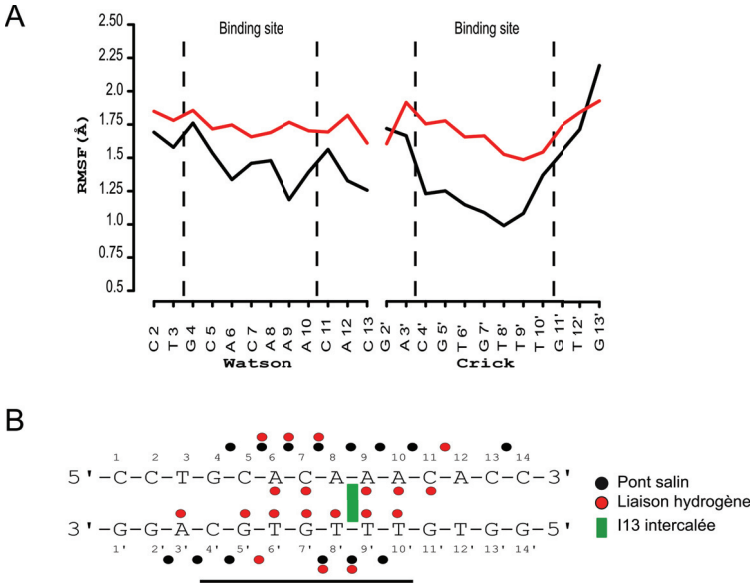


Figure 45 : Information sur la structure du complexe ADN-SRY. A) Moyenne du RMSF des atomes de phosphore pour les brins Watson & Crick Crick pour le complexe (lignes noires) et de l'ADN libre (lignes rouges. Les lignes verticales en pointillés délimitent le site de reconnaissance spécifique. B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 14 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-14' dans le sens 3'-5'.

Tableau 3 : Temps de présence (% pres.) et temps de vie moyen (Lifetime en picoseconde) des ponts salins et liaisons hydrogène entre SRY, l'ADN (bases + brins phosphodiester) durant les 500 ns de simulation. La ligne noire marque la séparation entre ponts salins (au-dessus) et liaisons hydrogène (en dessous). Les interactions notées en vert sont présentes uniquement dans la structure expérimentale, celles en noire sont présentes dans la structure RMN et dans la simulation et celles en rouge sont présentes uniquement dans la simulation de dynamique moléculaire.

Protein	Backbone	%pres.	Lifetime	Protein	Base	%pres.	Lifetime
R4(NH2)	A8(O1P)	21	135	R7(NH1)	C7(O2)	92	780
R4(NH1)	A8(O1P)	32	185	R7(NH2)	T6'(O2)	58	50
R4(NH2)	A8(O2P)	20	75	R7(NH1)	T6'(O2)	97	3590
R4(NH1)	A8(O2P)	14	75	N10(ND2)	T8'(O2)	97	27940
K6(NZ)	A9(O2P)	56	45	N10(ND2)	G7'(N3)	81	175
K6(NZ)	A9(O1P)	27	135	S16(OG)	A9(N3)	8	165
R17(NE)	A10(O1P)	53	55	R20(NH2)	T9'(O2)	13	100
R17(NH2)	A10(O1P)	71	205	N32(ND2)	A10(N3)	65	60
R17(NE)	A9(O3')	46	45	N32(ND2)	C11(O2)	94	1240
R21(NE)	C11(O1P)	64	165	S33(OG)	G11'(N3)	<1	-
R21(NE)	C11(O2P)	14	20	S36(OG)	T10'(O2)	100	128 735
R21(NH2)	C11(O2P)	65	375	Y74(OH)	A6(N3)	54	490
R21(NH2)	C11(O1P)	20	120	Y74(OH)	G5'(N3)	26	60
R31(NH2)	C14(O1P)	7	730	R78(NH1)	A3'(N3)	27	185
K37(NZ)	T9'(O1P)	45	75	R78(NH2)	A3'(N3)	14	60
K44(NZ)	T8'(O1P)	22	75				
K51(NZ)	G7'(O1P)	<1	-				
R66(NH2)	G7'(O1P)	9	360				
K73(NZ)	C4'(O1P)	28	135				
R75(NE)	A3'(O1P)	10	125				
R75(NH1)	A3'(O1P)	20	310				
R75(NH2)	A3'(O2P)	11	315				
R77(NE)	C7(O1P)	54	220				
R77(NE)	C7(O2P)	20	50				
R77(NH2)	C7(O2P)	44	175				
R77(NH2)	C7(O1P)	16	185				
R78(NH1)	G2'(O1P)	8	95				
K79(NZ)	A6(O2P)	21	70				
K79(NZ)	A6(O1P)	7	45				
K81(NZ)	C5(O1P)	12	8				
R7(N)	A8(O1P)	33	180				
R7(N)	A8(O5')	16	250				
R7(NH2)	G5'(O4')	11	35				
N10(ND2)	T8'(O4')	<1	-				
N32(N)	A12(O4')	73	170				
R77(N)	A6(O3')	76	150				
R77(N)	C7(O1P)	37	50				
N32(ND2)	C11(O4')	<1	-				
W43(NE1)	G7'(O1P)	88	290				
W43(NE1)	T8'(O3')	17	15				
Q62(NE2)	T6'(O3')	<1	-				
Q62(NE2)	G5'(O1P)	14	55				
K79(N)	A6(O1P)	<1	-				

3. L'environnement ionique

La présence de la protéine au niveau du petit sillon de l'ADN, réorganise la distribution des ions autour de l'ADN (cf. Figure 46). La présence de SRY élimine complètement la présence d'ions K^+ dans le petit sillon de l'ADN au niveau du site de reconnaissance (position 4 à 10), mais également au niveau des bases adjacentes. La distribution des ions dans le grand sillon de l'ADN semble moins affectée par la présence de la protéine. Cependant, la présence des ions au niveau des positions G4pC5 est fortement diminuée dans le complexe avec une molarité 2,5 fois moins importante que celle de l'ADN isolé (molarité de 1,65 M et 4,1 M respectivement). La présence de l'hélice α dans le petit sillon dans la région 3' du site de reconnaissance induit également une augmentation de la concentration en potassium dans le grand sillon de presque 100 % alors qu'aucune interaction n'a lieu dans ce sillon. L'augmentation de la concentration ionique dans le sillon opposé à celui qui interagit avec la protéine a déjà été observée plus clairement dans le complexe TBP développé page 102, Figure 38.

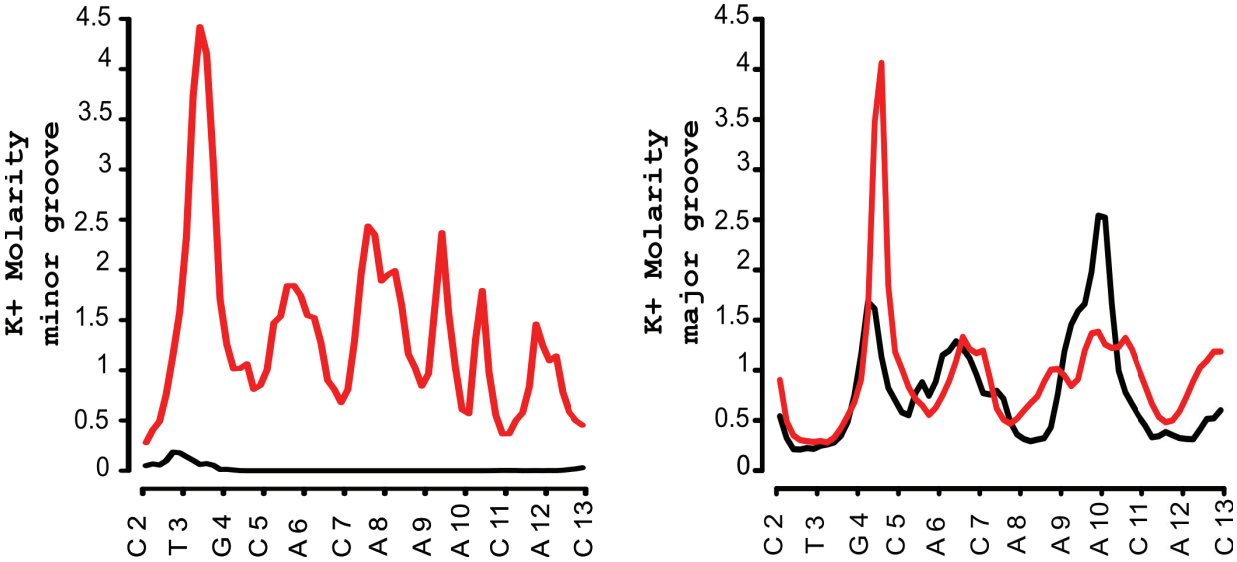


Figure 46 : Concentration ionique de potassium dans le petit et grand sillon de l'ADN pour le complexe ADN-SRY (ligne noire) et pour l'ADN libre (ligne rouge). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

4. Clustering et analyses de la spécificité de séquence

L'interface d'interaction du complexe ADN-SRY a été analysée selon le protocole développé page 83. L'analyse de la matrice d'interaction calculée pour l'ensemble des contacts acide aminé-base indique clairement la présence de plusieurs sous-états

conformationnels de la protéine SRY (cf. Figure 47). Afin de comprendre quels sont les résidus responsables de la création de ces sous-états, une matrice de contacts entre chaque acide aminé et les bases de l'ADN a été réalisée. Parmi les 85 matrices obtenues, il en ressort que les résidus tyrosine 74 (Y74) et arginine 78 (R78) sont les principaux contributeurs de la matrice globale (cf. Figure 47A). Ces deux résidus sont présents dans la queue C-terminal et interagissent avec la partie 5' du site d'interaction. Pour simplifier les analyses, l'analyse de la spécificité de reconnaissance sera scindée en deux sections, l'une sera concentrée sur la sélectivité de séquence de Tyr74 et la seconde sur Arg78.

L'analyse des structures révèle que la chaîne latérale du résidu Tyr74 est capable d'adopter trois conformations distinctes (cf. Figure 47B). La première conformation est présente 54 % du temps : Y74 effectue une liaison hydrogène avec l'atome N3 de la base azotée A6 (Cluster 1). Dans le second cluster (Cluster 2) (26 % de la trajectoire) Tyr74 est également donneur de liaison hydrogène, mais cette fois-ci avec la base du brin Crick G5'(N3). Enfin la dernière conformation (Cluster 3) est observée pendant environ 20 % du temps : la tyrosine 74 est impliquée dans une interaction bidentée en tant que donneur avec la base A6(N3) et accepteur de liaison hydrogène avec la G5'(N2).

Le résidu Arg78 forme différentes interactions au cours de la simulation. Ce résidu, positionné dans la queue C-terminale, possède une chaîne latérale très dynamique. L'inspection visuelle des structures révèle l'existence de plusieurs conformations dominées par trois sous états. Dans sa première conformation (environ 8 % du temps), Arg78 forme un pont salin avec les atomes du groupement phosphate de la base G2' (cluster A). En fin de simulation, Arg78 forme une liaison hydrogène avec la base azotée A3'(N3) (environ 25% du temps) (cluster B). Enfin dans le dernier groupe de structures, Arg78 n'interagit ni avec les bases l'ADN ni avec le brin phosphodiester (environ 20% du temps) (cluster C).

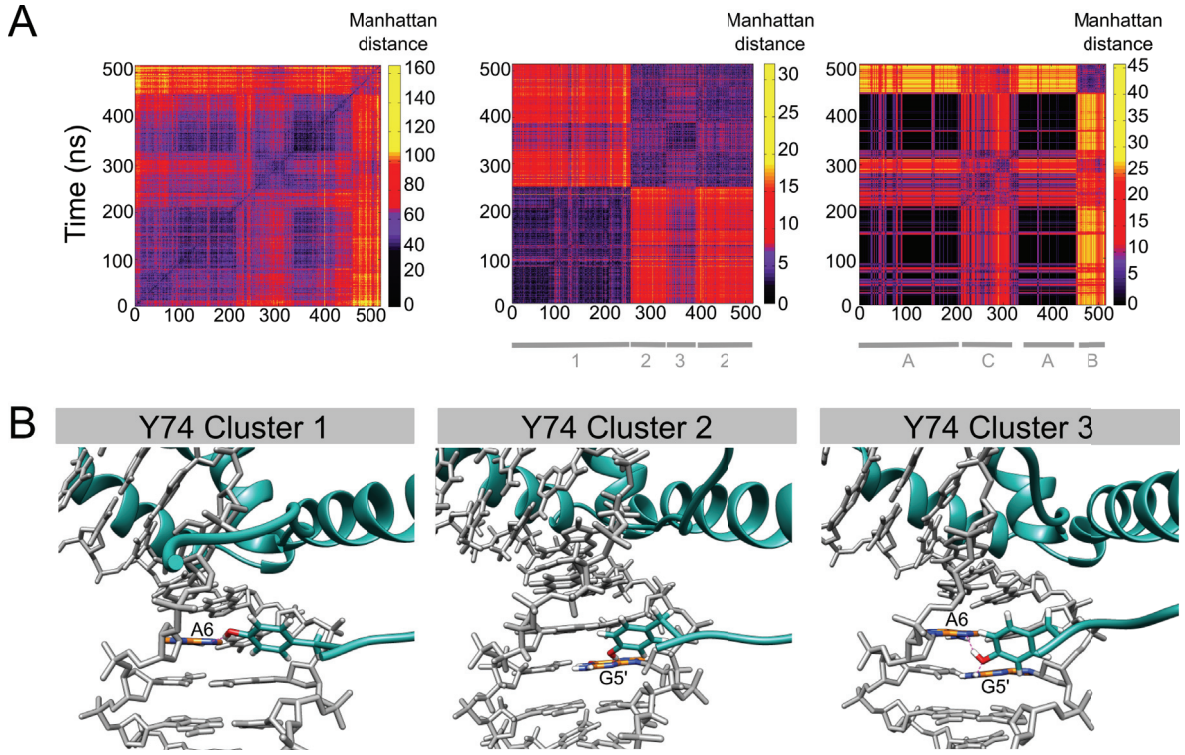


Figure 47 : Classification des structures du complexe ADN-SRY lors des 500 ns simulations. A) Matrices des distances de Manhattan pour tous les contacts ADN-protéine (à gauche), pour la tyrosine 74 (au milieu) et pour l'arginine 78 (à droite). L'échelle indique les distances croissantes (sens noir → jaune). Les barres grises horizontales indiquent les différents groupes le long de la matrice de contacts. B) Conformations alternatives observées pour la tyrosine 74 : liaison hydrogène avec A6(N3) (à gauche), G5'(N3) (au centre) et formant une interaction bidentée avec A6(N3) et G5'(N2) (à droite).

Un total de 21 structures, réparti en 6 structures pour le cluster 1, 10 pour le cluster 2 et 5 pour le cluster 3 a été utilisé pour effectuer l'analyse de la spécificité de séquence par la méthode d'enfilage ADAPT de Tyr74. L'analyse des logos issus de la méthode ADAPT révèle qu'un changement de conformation de la chaîne latérale Tyr74 n'altère que peu la sélectivité de séquence et qu'une préférence pour la base adénine en sixième position est privilégiée dans les trois sous états (cf. Figure 48). Cependant, on observe une sélectivité pour une thymine en position 4 (extrémité 5' du site de liaison) lorsque Tyr74 effectue une interaction avec la base G5' du brin Crick adjacente (cluster 2 et 3).

Le résidu Arg78 ne forme aucune interaction avec les bases du site de reconnaissance (position 4 à 10), mais avec les nucléotides flanquant la région 5'. L'analyse de la spécificité par enfilage moléculaire ne révèle aucune différence significative si ce n'est une perte de reconnaissance du nucléotide C en position 7 lorsque l'arginine ne forme aucune interaction avec l'ADN.

Les résidus Tyr74 et Arg78 ont un impact mineur sur la spécificité de reconnaissance. Le logo moyen (MD Consensus) est en bonne adéquation avec le logo

expérimental humain issu de JASPAR avec un IC de 6,0 pour le logo ADAPT contre 8,7 pour JASPAR. Localement, on observe une corrélation de Pearson de 0,48/0,44/1/0/1/0,91/0,99 pour les positions 4 à 10, et une corrélation globale de 0,69. La différence majeure entre les deux logos est le changement de sélectivité au niveau de la position centrale C7 du site de reconnaissance. Dans les données expérimentales, on observe une sélectivité accrue pour la base C et une sélectivité réduite pour la thymine alors qu'*in silico*, on observe une sélectivité accrue pour une thymine et une faible sélectivité pour la cytosine. Dans l'environnement de la paire de bases C7-G7', on retrouve une asparagine (N10) qui effectue une liaison hydrogène avec l'atome N3 de la base G7'. La conformation de la chaîne latérale l'asparagine 10 dans la structure RMN ne permet pas à l'atome OD1 d'accepter une liaison hydrogène provenant de l'atome G7'(N2). Une sélectivité pour une thymine est alors possible puisqu'elle autorise une liaison hydrogène Asn10(ND2)-T(O2) et évite une répulsion électrostatique entre Asn10(ND2) et G7'(N2) présents dans la structure RMN. Il est toutefois important de souligner la présence de deux logos expérimentaux pour un homologue de la protéine SRY chez la souris. Cet homologue possède une similarité de séquence de 86 % et une identité de séquence de 70 % avec la protéine humaine. En se basant sur l'alignement des séquences (aucune structure de l'homologue de la souris n'est disponible), le site d'interaction est virtuellement identique à 83 % (les résidus Arg31 et Ser33 de l'homme sont remplacés par une asparagine et une thréonine respectivement) et similaire à 100 %. Cependant, les résidus Arg31 et Ser33 ne sont pas en interaction avec les bases qui forme le site consensus mais avec les bases adjacentes en position 3'. Un de ces logos présente une préférence pour une thymine en position 7 (cf. Figure 48). Ce logo (JASPAR Mouse 2) présente une meilleure corrélation globale de 0,73 avec notre logo de modélisation.

On notera également que lors de la simulation, la séquence du site de fixation contient une double mutation A4A5 par G4C5 et que la séquence consensus qui dérive de l'analyse d'enfilage moléculaire ne présente pas de spécificité pour ces bases et une faible préférence pour les nucléotides A/T observés expérimentalement. Cela signifie que l'optimisation du positionnement des chaînes latérales lors de l'étape de minimisation du complexe est capable d'adapter la surface d'interaction et n'est pas biaisée par la séquence utilisée lors de la simulation.

Lorsque l'on décompose le logo consensus selon les mécanismes de reconnaissance directe et indirecte, on constate que la reconnaissance directe prédomine

dans la région 5’ du site de fixation, site qui, rappelons-le, forme des interactions avec la queue C-terminal. La reconnaissance indirecte est plus importante dans la région 3’, là où l’hélice α déforme le petit sillon.

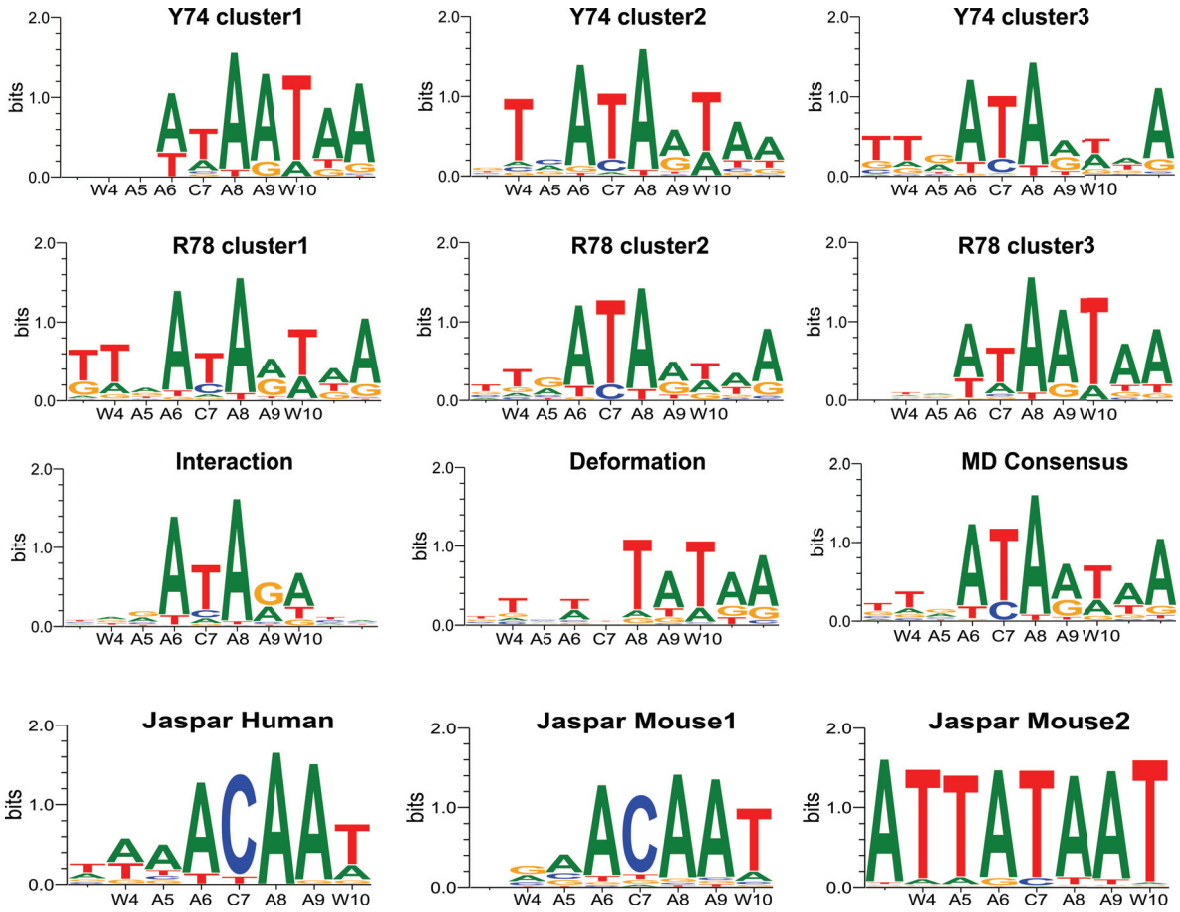


Figure 48 : Logo de spécificité de séquence obtenue lors de l’analyse de la dynamique moléculaire de SRY La décomposition des mécanismes d’interaction en reconnaissance directe liés à l’interaction ADN-protéine, indirecte avec la déformation de l’ADN et le logo moyen sont également présentés dans les panneaux Interaction, Deformation et MD consensus respectivement. Les logos obtenus à partir de données expérimentales provenant de la base de données JASPAR sont également représentés sur le panneau du bas avec à gauche le logo humain et son homologue chez la souris (panneau du centre et à droite). Sous chaque logo, la séquence consensus est exprimée le long de l’axe des abscisses (W ≡ A/T).

En conclusion, lorsque l’on décompose le mécanisme de reconnaissance en termes de reconnaissance directe et indirecte, on observe que les interactions entre la queue C-terminale de SRY et l’ADN déterminent la reconnaissance directe de la partie 5’ et centrale du site de liaison. La reconnaissance indirecte (liée à la déformation de l’ADN par l’interaction de l’hélice α dans le petit sillon) joue un rôle important dans la reconnaissance de la région 3’ du site reconnaissance. La formation de sous-états conformationnels liés à la dynamique de la queue C-terminale ne semble pas modifier la reconnaissance des bases par la protéine.

IV. Le complexe SKN-1

A. Rôle biologique

La protéine skinhead-1 (SKN-1) est un facteur de transcription présent chez le nématode Caenorhabditis elegans. SKN-1 joue un rôle essentiel dans la différenciation des blastomères abdominaux durant la phase embryonnaire du nématode, puis dans le développement des intestins chez le nématode adulte (361, 362). Ce facteur de transcription est également impliqué dans la régulation du gène Stl-1 en réponse à un stress oxydatif ou une anoxie (363). SKN-1 régule aussi le gène Gsc-1 et plusieurs gluthathione-S-transférases en réponse à un stress oxydatif généré lors d'infection bactérienne (364). Cette protéine est homologue aux protéines Nrf humaines qui sont également impliquées dans les mécanismes de réponse à un stress.

B. Propriétés structurales

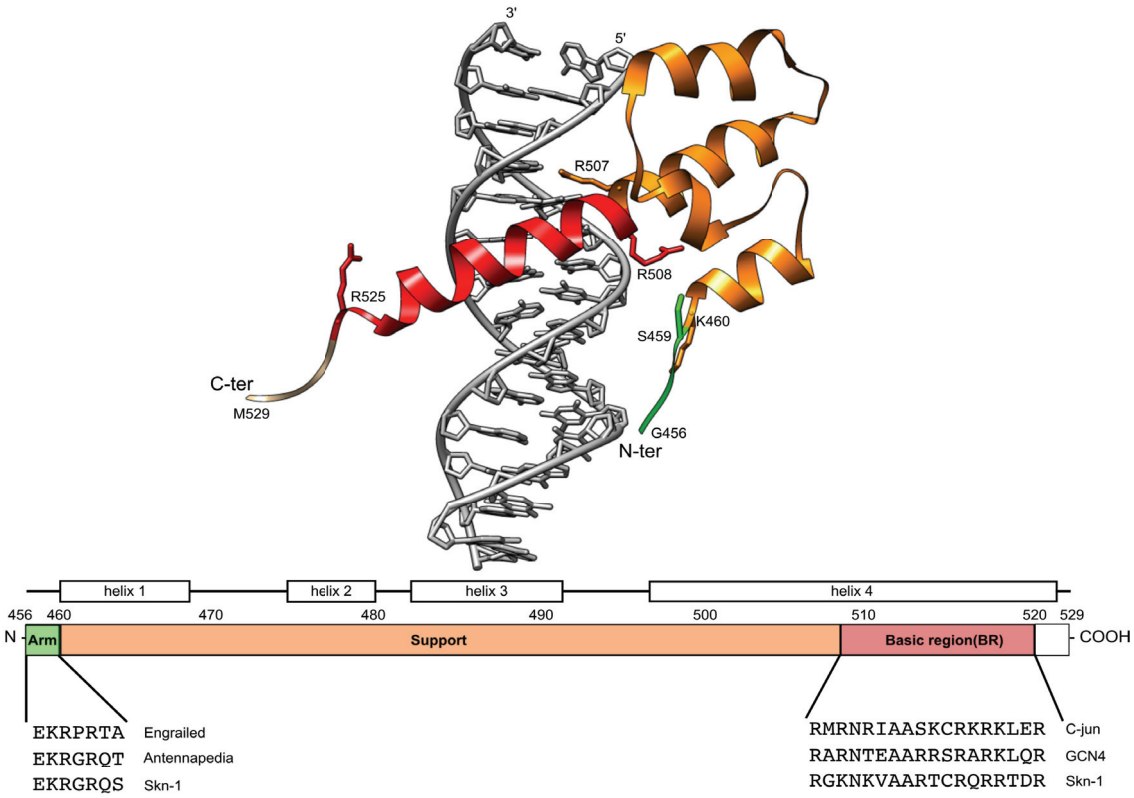
La structure du domaine de liaison à l'ADN de SKN-1 a été obtenue par cristallographie aux rayons X avec une résolution de 2,50 Å et est disponible dans la PDB sous le code 1SKN. Le domaine SKN-1, constitué de 72 acides aminés, se compose de quatre hélices α qui forment trois régions : le support, la région basique et le bras N-terminal (cf. Figure 49A). La protéine possède une hélice basique enfouie dans le grand sillon de l'ADN analogue à celle observée dans la famille des leucine zipper (par exemple c-Jun et GCN4) (cf. Figure 49), mais ne possède pas la région permettant une dimérisation de la protéine. La séquence d'ADN GTCAT, reconnue par SKN-1, correspond au demi-site de reconnaissance du dimer GCN4. Les résidus Asn 511, Ala 514, Ala 515 et Arg 519 de la protéine SKN-1 forment des interactions analogues à celles effectuées par les résidus Asn 235, Ala 238, Ala 239 et Arg 243 de la protéine GCN4. SKN-1 contient également un bras N-terminal basique de 5 résidus (EKRGR) dont la séquence est similaire à celle des protéines possédant un homéodomaine qui se lient spécifiquement au petit sillon d'ADN dont les séquences sont riches en nucléotides A/T (365).

SKN-1 effectue des contacts directs avec les bases de l'ADN (cf. Figure 49 B). L'arginine 519 forme une double liaison hydrogène avec les atomes N7 et O6 de la base G8. Le groupement méthyle de l'alanine 515 interagit par contact de van der Waals avec la thymine 9 et enfin l'asparagine 511 forme une double liaison hydrogène avec

C10(N4) et T11'(O4). Au niveau du cristal, il n'y a pas de contacts directs entre les acides aminés de la protéine et la paire de bases en position 12 malgré qu'elle fasse partie du site de reconnaissance.

Le complexe est également stabilisé par la présence de plusieurs ponts salins formés essentiellement par les résidus de l'hélice de reconnaissance. Dans la structure cristallographique, tous les ponts salins sont effectués par des arginines (R503, Arg506, Arg507, Arg508, Arg516, Arg521 et Arg522) avec les groupements phosphates G8, G10', T11' (du site de reconnaissance) et les groupements phosphates C13 et G14' qui sont à l'extérieur du site de liaison.

A



B

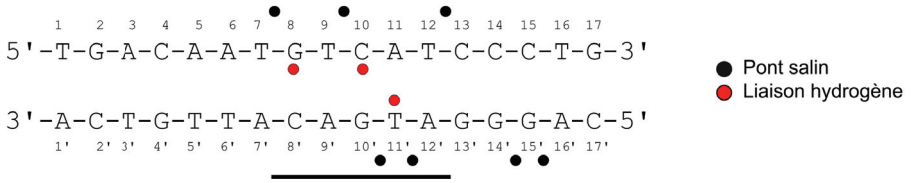


Figure 49 : Structure et interactions du complexe ADN-SKN-1. A) Structure 3D de SKN-1 en complexe avec l'ADN. Les résidus représentés en « stick » définissent les limites du bras N-terminal (en vert), de la région support (en orange) et de la région basique BR de l'hélice de reconnaissance (en rouge). Ces mêmes régions sont représentées sur la séquence primaire située en dessous. Sur la séquence primaire, est également représentée l'alignement de séquence du bras N-terminal de SKN-1 avec ses homologues chez *Drosophilia engrailed* (366) et *Drosophilia antennapedia* (367) ainsi que l'alignement de la partie basique de l'hélice de reconnaissance avec ses homologues c-jun (368) et GCN4 (369). B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 17 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-17' dans le sens 3'-5'.

C. Site de reconnaissance de la protéine SKN-1

Des études génomiques de régulation positive et de régulation négative des gènes par la protéine SKN-1 réalisées suite à un stress oxydatif (présence d'arsenites ou de tert-Butyl hydroperoxyde) (370, 371) montrent que la protéine se lie à une séquence d'ADN ayant le motif RTCAT (370–372) (où R désigne les bases A ou G d'après la nomenclature IUPAC (291)). Cependant SKN-1 ne reconnaît pas strictement la séquence RTCAT, mais présente certaines modulations de spécificité (cf. Figure 50 A-D). Au niveau de la position 8 par exemple, seule l'expérience de régulation positive de Staab (cf. Figure 50C) (371) retrouve une sélectivité pour G ou A alors que les autres résultats indiquent une préférence pour G uniquement (cf. Figure 50B), ou A uniquement (cf. Figure 50D) et même T/G/A dans l'expérience de régulation positive de Oliveira (cf. Figure 50A) (370). Notons également la position 12 où la sélectivité pour la base T n'est retrouvée que dans les gènes activés positivement par SKN-1. La sélectivité du site riche en A/T entre les positions 5 et 7 est également variable dans les expériences de régulation du groupe Oliveira (cf. Figure 50B et D) où les bases C et G sont préférées pour les position 5 et 6.

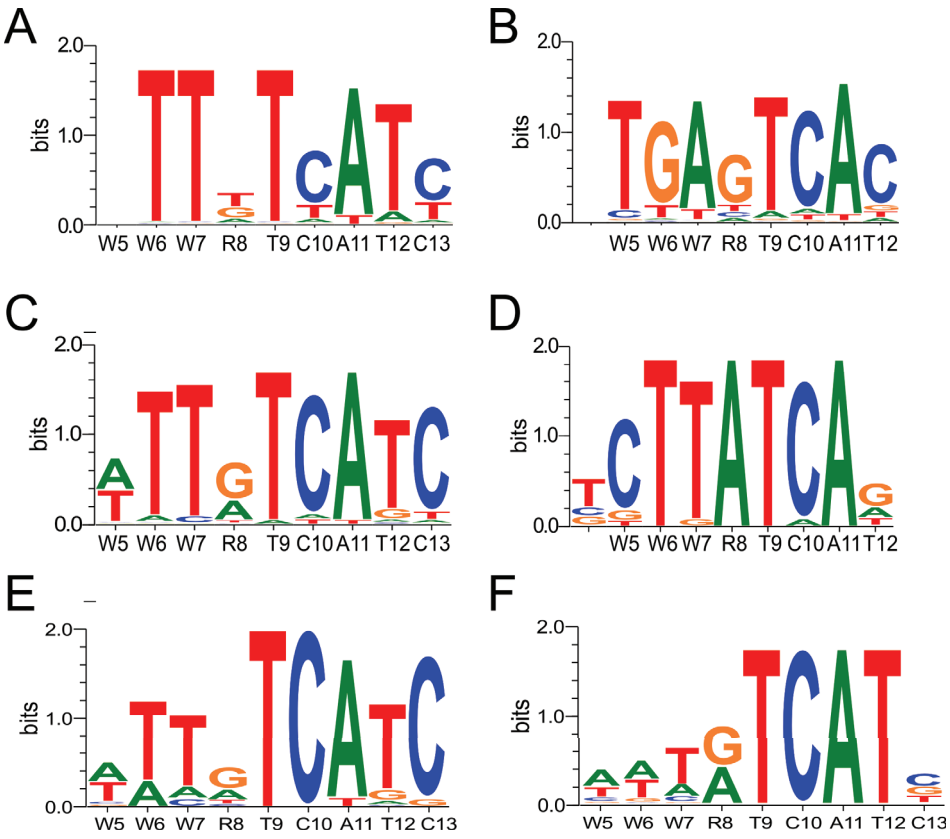


Figure 50 : Motifs de reconnaissance obtenus après des expériences de régulation positive (A,C) et de régulation négative (B,D) pour la protéine SKN-1 réalisées par le groupe de Oliveira (370) (A,B) et le groupe de Staab (371) (C,D). Les logos consensus disponibles sur les bases de données Jaspar et Transfac sont présentés en E et F respectivement. Sous chaque logo, la séquence consensus est exprimée le long de l'axe des abscisses (W ≡ A/T et R ≡ A/G).

En 1999, Blackwell et ses collaborateurs (373) ont démontré que le bras N-terminal de SKN-1 était responsable de la reconnaissance d'une séquence riche en AT à l'extrémité 5' du motif RTCAT. Chacune des trois paires de bases présentes en amont du motif RTCAT est essentielle et la substitution de l'une d'entre elle par un nucléotide G ou C diminue significativement l'affinité d'interaction ADN-SKN-1 (373). Les expériences de mutagenèse dirigée montrent que chaque résidu présent dans le bras N-terminal contribue à l'affinité et à la spécificité d'interaction. En particulier les mutations de la glycine 456 et de l'arginine 457 diminuent l'énergie d'affinité et la spécificité pour la séquence riche en A/T au profit d'une séquence riche en G/C (373). On notera que la spécificité de séquence pour la région riche en A/T, notée par les positions W5 à W7 sur la Figure 50 (où W représente les nucléotides A ou T selon la nomenclature IUPAC), est cohérente avec les expériences de Blackwell, mais présente une modulation de la spécificité comme observée dans le motif RTCAT. Deux logos ont été déposés sur les bases de données JASPAR (290) et TRANSFAC (289), obtenus grâce à des expériences de ChIP-seq et de SELEX (cf. Figure 50 E et F). Comparées aux logos présentés Figure 50, les données issues de ChIP-seq présentent une faible sélectivité pour les bases A/T au niveau des positions 5 à 7.

D. Analyse de la dynamique moléculaire

1. Structure de l'ADN

La liaison de la protéine skinhead-1 altère légèrement la structure de la molécule d'ADN (cf. Figure 51). L'insertion de la longue hélice dans le grand sillon (cf. Figure 49A) provoque une faible augmentation de la largeur du grand sillon de 1 à 1,5 Å le long du site de reconnaissance (position 8 à 13), et induit un rétrécissement dans le sillon opposé. La présence de la queue N-terminal dans le petit sillon (position A5-T7) provoque également une diminution de sa largeur de 1 Å. L'interaction entre les acides aminés et les bases de l'ADN provoque une augmentation du Twist de 2° en moyenne. Cette augmentation est cependant plus importante au niveau de la position C10-A11 où une l'asparagine 511 contraint les bases de l'ADN pour effectuer une double liaison hydrogène.

Le paramètre Rise est peu affecté le long du site de reconnaissance (position G8-C13) si ce n'est une légère diminution de 0,22 Å au niveau de la paire de base T7pG8 où l'arginine 519 est positionnée pour effectuer des liaisons hydrogène avec la base G8. Enfin

l'on constate une petite diminution de 1,7° de l'angle de Roll au niveau de la position 5' du site de fixation, probablement liée à l'insertion du bras N-terminal dans le petit sillon de l'ADN au niveau de la séquence 5'-AAT-3'.

Enfin la présence de la protéine SKN-1, n'altère pas les paramètres des brins phosphodiester de l'oligomère. La conformation du désoxyribose, ainsi que la distribution des angles epsilon/zêta (conformation BI/BII) et alpha/bêta (conformation g-/g+ pour un ADN B) est sensiblement identique entre l'ADN du complexe et l'ADN isolé.

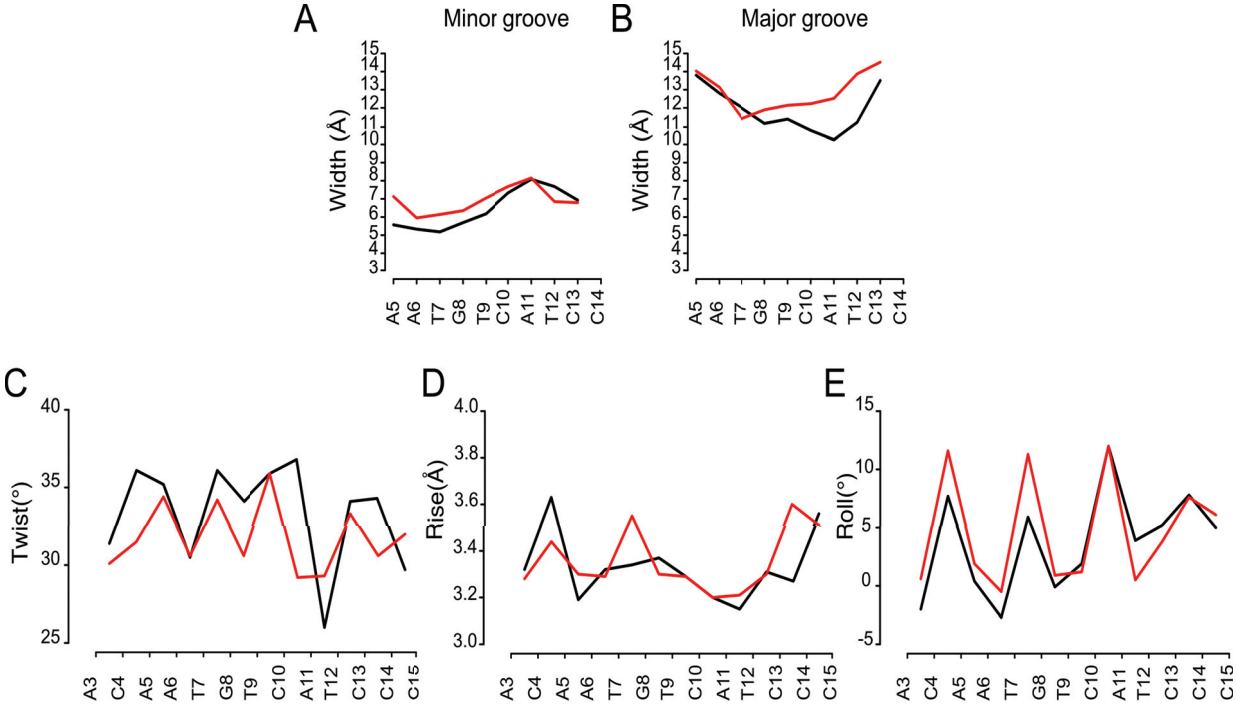


Figure 51 : Paramètres structuraux du complexe SKN-ADN (lignes noires) et de l'ADN libre (lignes rouges). A et B représentent la largeur du petit et du grand sillon de l'ADN (Å). C, D et E représentent les valeurs moyennes des paramètres inter-paires de bases Twist (°), Rise (Å) et Roll (°). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

2. Dynamique du complexe et des interactions ADN-SKN1

L'analyse de la structure cristallographique du complexe SKN-1 révèle la présence de sept ponts salins et trois liaisons hydrogène qui stabilisent l'interaction ADN-protéine (cf. Figure 49B), essentiellement au niveau du site d'interaction. Au cours de la dynamique moléculaire, le pouvoir stabilisant de ces ponts salins se confirme par une diminution de la mobilité des groupements phosphates des positions au niveau du site de reconnaissance (position 8 à 12), mais également sur les bases adjacentes C13-C15 où l'on observe la présence de trois ponts salins entre les chaînes latérales des arginines 503, 506 et lysine 510 avec le brin Crick (cf. Figure 52 et Tableau 4). Durant la dynamique, on

voit également apparaître des interactions entre la protéine et les bases de l'ADN qui n'étaient pas présentes dans la structure cristallographique. Au cours de la simulation, le bras N-terminal de la protéine s'introduit dans le petit sillon de l'ADN. L'enfouissement de ce bras N-terminal induit la formation de liaisons hydrogène entre l'arginine 457 et les bases T5'(O2) et T6'(O2) de l'ADN. Dans la structure, l'arginine 507 forme un pont salin avec la base C10. Cependant au cours de la simulation, l'arginine 507 établit de manière préférentielle des liaisons hydrogène avec les bases T12(O4) et G13'(O6 et N7).

La plupart des ponts salins observés par dynamique moléculaire impliquent les mêmes paires acide aminé-nucléotide que ceux observés expérimentalement. Cependant l'on observe quand même la formation de nouveaux ponts salins impliquant les résidus R457 et K460 du bras N-terminal. On notera que les ponts salins présents dans la structure cristallographique sont ceux qui persiste le plus longtemps lors de la dynamique. Le temps de vie d'une liaison hydrogène ou d'un pont salin est, comme les cas TBP et SRY présentés précédemment, de l'ordre de la centaine de picosecondes.

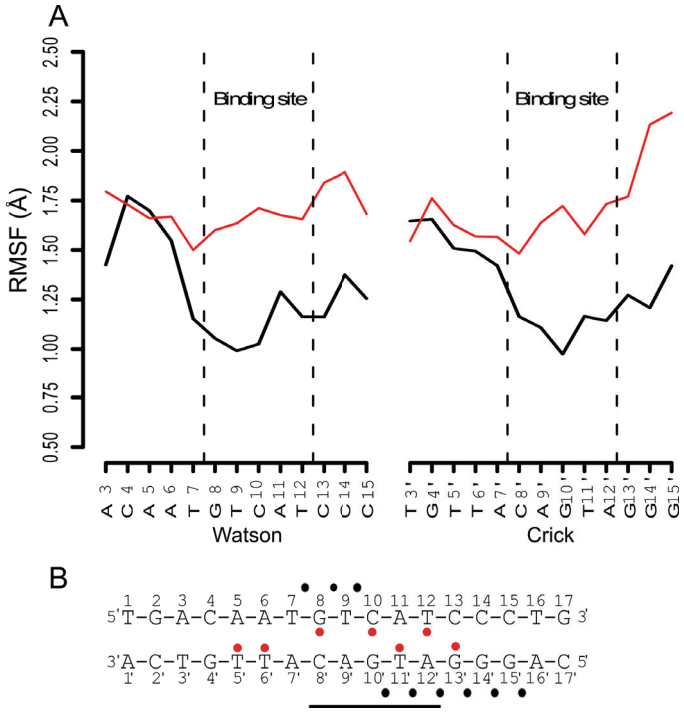


Figure 52 : Information sur la structure du complexe ADN-SKN-1. A) Moyenne du RMSF des atomes de phosphore pour les brins Watson & Crick pour le complexe (lignes noires) et de l'ADN libre (lignes rouges). Les lignes verticales en pointillés délimitent le site de reconnaissance spécifique. B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 17 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-17' dans le sens 3'-5'.

Tableau 4 : Temps de présence (% pres.) et de vie (lifetime en picoseconde) des ponts salins et liaisons hydrogène dans le complexe SKN (bases + brins phosphodiester) durant les 500 ns de simulation. Les interactions notées en noires sont présentes dans la structure cristallographique et dans la simulation et celles en rouges sont présentes uniquement dans la simulation de dynamique moléculaire.

Protein	Backbone	% pres.	lifetime	Protein	Base	% pres.	lifetime
ARG 503(NE)	G 15'(O2P)	52	21	ARG 457(NE)	T 5'(O2)	15	24
ARG 503(NE)	G 15'(O1P)	38	13	ARG 457(NH1)	T 5'(O2)	12	34
ARG 503(NH2)	G 15'(O2P)	20	25	ARG 457(NH2)	T 5'(O2)	19	40
ARG 503(NH2)	G 15'(O1P)	49	60	ARG 457(NH1)	T 6'(O2)	16	13
ARG 503(NH1)	A 16'(O2P)	13	10	ARG 507(NH2)	T 12(O4)	24	29
ARG 503(NH2)	A 16'(O2P)	19	18	ARG 507(NH1)	G 13'(N7)	25	34
ARG 503(NH2)	A 16'(O1P)	15	22	ARG 507(NH2)	G 13'(O6)	72	145
ARG 506(NE)	G 14'(O2P)	17	5	ASN 511(ND2)	T 11'(O4)	92	113
ARG 506(NE)	G 14'(O1P)	96	268	ASN 511(OD1)	C 10(N4)	97	2357
ARG 506(NH2)	G 14'(O2P)	98	695	ARG 519(NH1)	G 8(N7)	78	87
ARG 507(NH2)	G 15'(O1P)	3	83	ARG 519(NH2)	G 8(O6)	79	169
ARG 507(NE)	G 15'(O2P)	2	6				
ARG 507(NH2)	A 11(O1P)	3	66				
ARG 507(NH1)	C 10(O1P)	9	7				
ARG 508(NE)	T 9(O2P)	20	10				
ARG 508(NH2)	T 9(O2P)	37	9068				
ARG 508(NE)	T 9(O5')	11	6				
ARG 508(NH1)	C 10(O1P)	50	91				
LYS 510(NZ)	G 13'(O1P)	21	16				
ARG 516(NE)	G 8(O1P)	42	68				
ARG 516(NH2)	G 8(O2P)	30	22				
ARG 516(NH2)	G 8(O1P)	37	35				
ARG 519(NH2)	T 7(O1P)	5	19				
ARG 519(NH1)	T 7(O1P)	4	309				
ARG 521(NE)	T 11'(O1P)	44	14				
ARG 521(NH2)	T 11'(O1P)	89	267				
ARG 521(NH1)	A 12'(O1P)	59	44				
ARG 521(NH2)	A 12'(O1P)	15	6				
ARG 521(NH2)	A 12'(O5')	16	9				
ARG 522(NE)	G 10'(O1P)	28	24				
ARG 522(NH2)	G 10'(O1P)	42	90				
ARG 525(NH1)	T 11'(O2P)	12	91				
ARG 457(N)	G 8(O2P)	28	46				
LYS 460(N)	T 9(O2P)	37	50				

3. *L'environnement ionique*

La présence de SKN-1 modifie la distribution ionique le long du site de liaison à l'ADN (cf. Figure 53). L'insertion de l'hélice α dans le grand sillon de l'ADN entre les

positions A6-C13 déplace entièrement la population d'ions à l'extérieur du sillon alors que l'on observe une concentration de 1 à 2 M pour l'oligomère isolé. Ce déplacement d'ions dans le grand sillon provoque, en contrepartie, une augmentation de la concentration dans le petit sillon pour ces mêmes positions. On note la présence d'un ion localisé au niveau des bases G8-T9 qui est absent dans la simulation de l'ADN libre. La présence du bras N-terminal dans le petit sillon de l'ADN provoque également une diminution de la population des cations entre les positions A5 et T7.

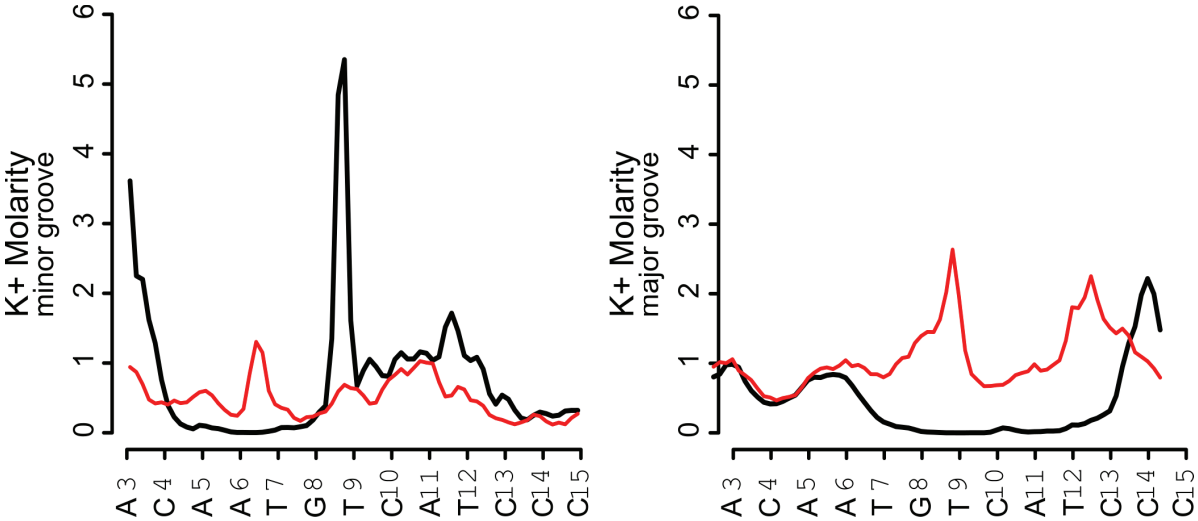


Figure 53 : Concentration ionique de potassium dans le petit et grand sillon de l'ADN pour le complexe ADN-SRY (ligne noire) et pour l'ADN libre (ligne rouge). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

4. Clustering et analyses de la spécificité de séquence

L'analyse de la dynamique de l'interface grâce aux matrices de distance révèle la présence de sous-états conformationnels (cf. Figure 54). L'analyse de la matrice des distances indique la présence de quatre sous états nommés CL1, CL2, CL3 et CL4 qui n'apparaissent pas avec la même fréquence durant les 500 ns de simulation. Le groupe CL1, qui possède une structure et une interface d'interaction proche de la structure cristallographique, disparaît après seulement 5 ns de simulation et ne réapparait qu'entre 300 et 400 ns de simulation. Le second groupe CL2, qui correspond au groupe le plus observé, est présent 60 % de la simulation. Le troisième groupe de conformation (CL3) n'apparaît que brièvement au début de la simulation entre 70 et 100 ns. Et enfin le dernier groupe (CL4) n'est présent qu'en milieu de simulation entre 200 et 300 ns.

L'analyse de ces différents groupes révèle que la majorité des fluctuations observées dans la matrice de contacts est liée à des changements d'interaction entre les résidus R507 et R519 et l'ADN (cf. Figure 55). Dans le groupe CL1, l'arginine 507 forme des interactions principalement avec le groupement phosphate C10 et plus rarement avec les groupements phosphates des bases A11 et G15' alors que dans les trois autres groupes, R507 est enfouie dans le grand sillon et forme une liaison hydrogène bidentée avec les atomes O6 et N7 de la base G13'. Parfois, R507 forme une triple liaison hydrogène en effectuant une interaction avec l'atome O4 de la thymine 12. Concernant, l'arginine 519, on observe deux conformations. Dans les groupes CL1, CL2 et CL4, R519 effectue une interaction bidentée avec les atomes O6 et N7 du nucléotide G8 alors que dans le groupe CL3, R519 forme un pont salin avec les atomes du groupement phosphate de la thymine T7.

Les groupes CL2 et CL4 sont très similaires, mais apparaissent comme des sous-états différents dans la matrice des contacts. L'analyse des structures indique que les deux groupes se différencient au niveau du bras N-terminal, qui dans le cas de CL2, est enfoui dans le petit sillon et interagit avec les bases azotées de l'ADN, alors que pour CL4, ce même bras terminal est en interaction avec le brin de l'ADN.

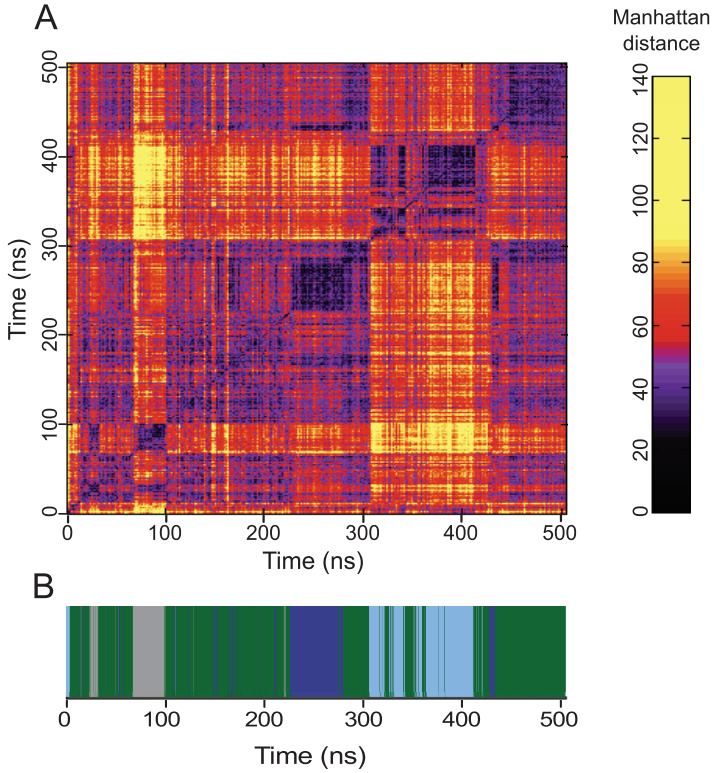
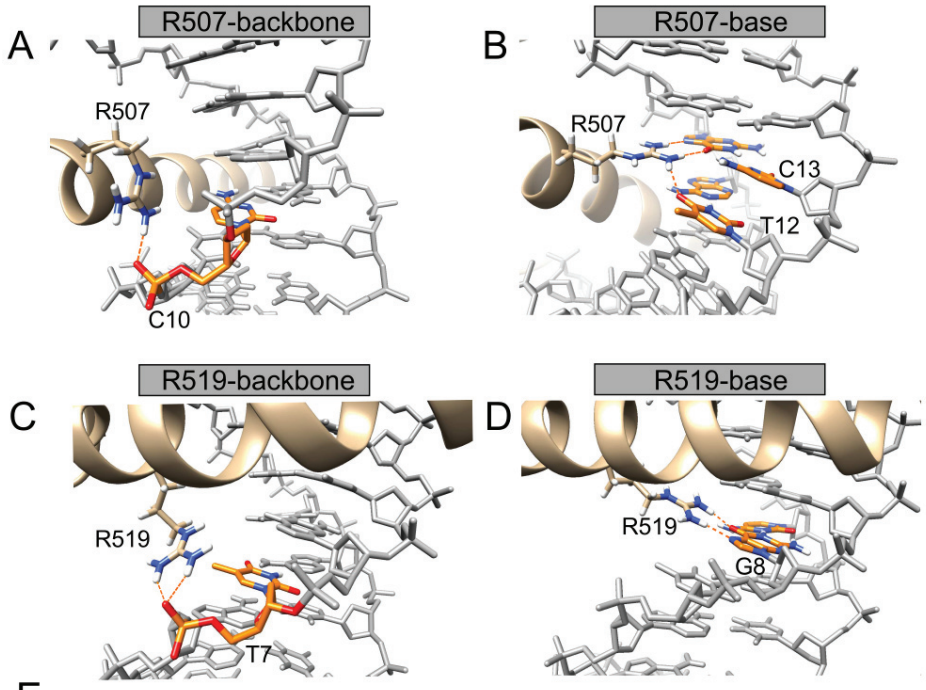


Figure 54 : Classification des structures du complexe ADN-SKN-1 durant les 500 ns de simulation. A) Matrice de distances de Manhattan pour tous les contacts ADN-protéine. L'échelle indique les distances croissantes (sens noir → jaune). B) La matrice de distances a été utilisée pour former quatre groupes de conformation distincte représentés le long de la trajectoire par les couleurs cyan (CL1), vert (CL2), gris (CL3) et bleu foncé (CL4).



E

Cluster	Color	R507 base	R507 backbone	R519 base	R519 backbone	% occurrence
Cl1		-	X	X	-	17
Cl2		X	-	X	-	60
Cl3		X	-	-	X	9
Cl4		X	-	X	-	14

Figure 55 : Conformations alternatives des résidus arginines 507 (A et B) et 519 (C et D). E) Le tableau récapitule la conformation des résidus R507/R519 dans chaque groupe. La fréquence de chaque groupe durant la dynamique moléculaire est également indiquée.

Le protocole d'enfilage moléculaire ADAPT, permettant de tester l'impact des changements de conformations des résidus Arg507 et Arg519, a été appliqué pour chacun des groupes. L'analyse des logos de spécificité de séquence indiqués sur la Figure 56 montre que les changements de conformations des deux arginines modifient de manière significative la reconnaissance entre SKN-1 et les bases de l'ADN. Les groupes CL2 et CL4 ne présentent qu'une différence structurale au niveau des interactions formées par le bras N-terminal. Les différentes conformations du bras N-terminal au niveau des positions A5 et T7 n'affecte pas la spécificité de reconnaissance (cf. Figure S1). C'est pourquoi, pour la suite des résultats, les deux groupes ont été fusionnés en CL2/4. Si on se concentre sur les positions 8, 12-13 qui sont les bases en interactions avec les résidus Arg519 et Arg507 respectivement, il est facile d'interpréter les résultats. La présence de liaisons hydrogène entre Arg519 et la position 8 favorise la base G à cette position (cf. Figure 56 CL1 et CL2/4). L'enfouissement de Arg507 dans le grand sillon laisse apparaître une sélectivité accrue pour les bases TC en position 12 et 13 (cf. Figure

56 CL2/4 et CL3). On constate également que la présence de Arg507 dans le sillon modifie la spécificité de reconnaissance au niveau des positions 10 et 11 et laisse apparaître un motif CATC dans les groupes CL2/4 et CL3. Pour CL1, où Arg507 n'est pas dans le grand sillon, on observe une sélectivité diminuée pour les bases CAT et une perte de la reconnaissance pour la base C13.

On remarque également que la queue N-terminal ne semble pas avoir d'impact sur la reconnaissance du site riche en A/T qui est localisé à l'extrémité 5' du site de liaison (position 5 à 7), (cf. Figure 56 et Figure S1) signifiant que la sélectivité est essentiellement électrostatique et qu'elle ne requiert pas la présence de l'arginine dans le petit sillon.

Le logo du groupe CL3 est en parfait accord avec les données expérimentales et principalement avec les données provenant de JASPAR (corrélation de 0,82). Les groupes CL1 et CL2/4 montrent une corrélation plus modérée de 0,5 et 0,52 respectivement. Si l'on regarde la corrélation au niveau de chacune des positions, on obtient une information complémentaire. En effet au niveau de la position 8, l'on observe une corrélation de 0,89, 0,95 et 0,29 entre les groupes CL1, CL2/4 et CL3 et les données de JASPAR signifiant que seuls les groupes où R519 est enfoui dans l'ADN (CL1 et CL2/4) sont capables de reproduire les données expérimentales. De même, si l'on observe les positions 12 et 13, seuls les groupes CL2/4 et CL3, où l'arginine 507 est en interaction avec les bases du grand sillon de l'ADN, sont capables de reproduire les données expérimentales avec des corrélations de (0,99, 1) et (0,97, 1) contre (0,84, -0,50) pour le groupe CL1 où Arg507 n'est pas dans le sillon.

Nous terminerons cette partie en indiquant que pour ce complexe, l'interaction spécifique entre les acides aminés et les bases de l'ADN est le mécanisme de reconnaissance principale de ce complexe, même s'il faut noter que la déformation de l'ADN permet la reconnaissance des bases T en positions 10 et 13 (cf. Figure S2).

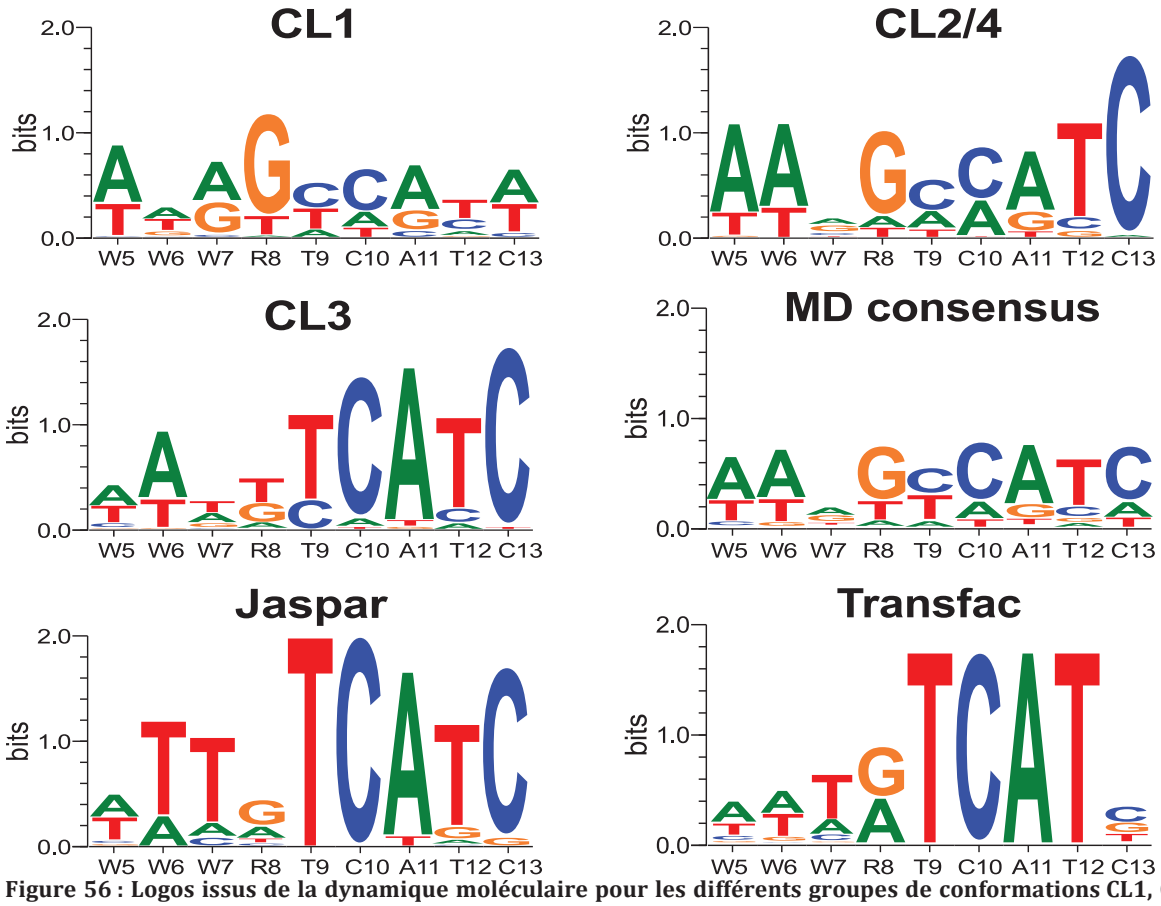


Figure 56 : Logos issus de la dynamique moléculaire pour les différents groupes de conformations CL1, CL2/4, CL3 et du logo moyen (MD consensus). Les logos expérimentaux de JASPAR et Transfac sont également présents. Sous chaque logo, la séquence consensus est exprimée le long de l'axe des abscisses (W ≡ A/T et R ≡ A/G).

En conclusion, on notera que l'interface du complexe ADN-SKN-1 est dynamique et que deux arginines jouent un rôle important. Les deux arginines 507 et 519 oscillent entre deux conformations afin de former soit des liaisons hydrogène avec les bases de l'ADN soit des ponts salins avec le brin phosphodiester. Dans cette étude, nous avons pu constater que la perte des liaisons hydrogène avec les bases de l'ADN provoque une perte significative de sélectivité laissant suggérer que la séquence consensus observée expérimentalement peut être perçu comme la moyenne de plusieurs sous-états qui reconnaissent différentes parties du site de liaison.

V. Le complexe P22 c2

A. Rôle biologique

Le répresseur P22 c2 est un homodimère protéique présent chez *Enterobacteria phage P22* qui va permettre au bactériophage de rester inactif dans le génome de l'hôte qu'il aura infecté. Pour rester inactif, le phage P22 produit la protéine répresseur se liant au niveau de sites opérateurs qui sont localisés de chaque côté du gène répresseur, afin d'inhiber l'expression des autres gènes du phage. L'opérateur qui se situe au début du génome du bactériophage est appelé O_L et celui en fin de génome O_R. Pour chacun de ces opérateurs, il existe trois opérateurs naturels O_L1, O_L2, O_L3, O_R1, O_R2 et O_R3 pour lequel le répresseur P22 c2 se lie avec plus ou moins d'affinité.

B. Propriétés structurales

La structure cristallographique du complexe P22 c2 est disponible via le code PDB 2R1J. P22 est une protéine globulaire qui se lie spécifiquement dans le grand sillon de l'ADN à l'aide d'un motif helix-turn-helix. La protéine P22 c2 provoque de faibles modifications structurales de la molécule d'ADN. Dans la structure cristallographique, l'ADN est courbé d'environ 20° et les sillons de l'ADN sont plus étroits que dans un ADN B canonique ce qui n'est pas lié à la courbure de l'ADN (puisque généralement la formation d'une courbure à un effet opposé sur les deux sillons), mais à une augmentation du Twist le long du site de liaison. L'interface formée par la protéine et l'ADN crée un canal au niveau des positions A9-T12 un canal. Ce canal est électronégatif de par la présence de la séquence ATAT (côté petit sillon) et de quatre acides glutamiques (deux par monomère), peut en théorie être occupé par des molécules d'eau ou des cations monovalents. Toutefois, la fonction de ce canal dans le mécanisme de reconnaissance est encore inconnue.

Chaque monomère de P22 c2 n'effectue qu'un seul contact direct avec les bases de l'ADN. Cette interaction est effectuée par les glutamines 37 qui forment une liaison hydrogène avec l'atome O4 des bases C8' et C13. Chaque monomère effectue également plusieurs interactions avec les bases par des contacts médiés par des molécules d'eau comme c'est le cas des résidus N32, Q37 et E42 qui peuvent ainsi former des interactions avec les bases des positions T4-G8 et T17'-G13'.

La valine 33 est insérée dans une cavité formée par les groupements méthyles des quatre thymines du fragment 5'-TTAA-3' (cf. Figure 57) au niveau du grand sillon. L'interaction de cette valine semble importante dans le positionnement des monomères et confère une stabilité et une spécificité lors de la reconnaissance (206). Cette cavité formée par ces groupements méthyles ne semble pas être induite par la fixation de la protéine puisqu'une cavité de même dimension est observée sur des fragments d'ADN non liés présentant le même motif 5'-TTAA-3' (374). Le rapprochement des groupes méthyle est facilité par des propriétés intrinsèques à cette séquence, notamment par une augmentation de l'angle de Roll au niveau du pas TA et des valeurs de propeller-twist négatives.

Chaque monomère forme également trois ponts salins et huit liaisons hydrogène avec les groupements phosphates de l'ADN. Les arginines 11, 14 et 40 forment des ponts salins avec les nucléotides 5'-TTT-3' de chaque brin. À cela viennent s'ajouter trois liaisons hydrogène formées par les résidus Ser31, Ser36 et Trp38 de l'hélice de reconnaissance et les phosphates de T7', T4 et C8' (correspondance T14, T17' et C13 pour le second monomère) de l'ADN permettant ainsi un bon positionnement de l'hélice. Enfin les résidus Gln21, Asn6 et Asn49 forment également des liaisons hydrogène avec T3(O2P)

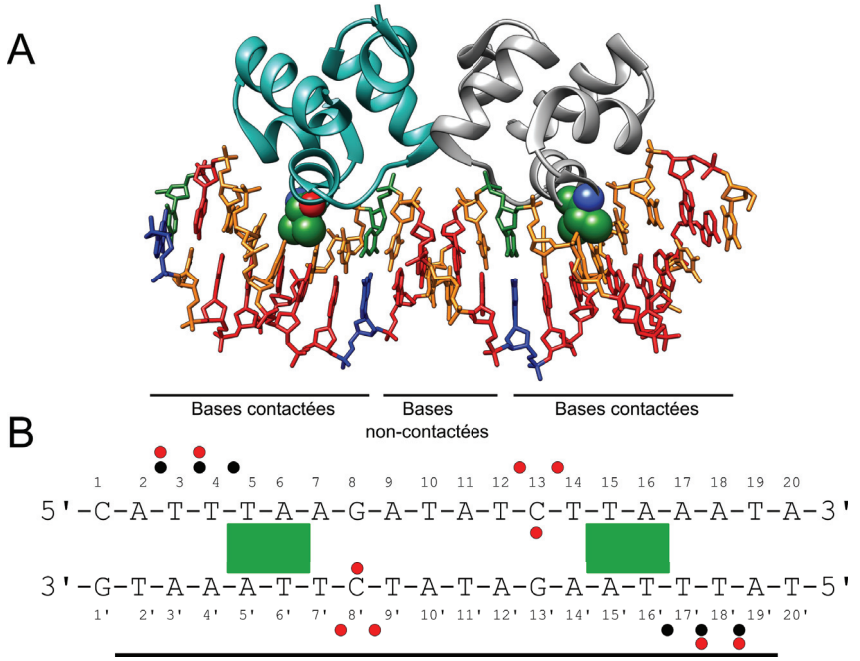


Figure 57 : A) Structure du complexe ADN-répresseur P22 c2 (les deux monomères sont représentés en gris et cyan). La valine 33 de chaque monomère qui interagit avec les groupements méthyles des paires de bases T4-A7 et T14-A17 est représentée en sphère. L'ADN est coloré selon la séquence (A en rouge, T en orange, G en bleu et C en vert). B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 20 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-20' dans le sens 3'-5'. Les ponts salins, liaisons hydrogène ainsi que la position des valines 33 observées dans la structure cristallographique sont indiqués par des cercles noirs, rouges et par des rectangles verts respectivement.

et C8'(O1P/O5') (équivalent T18' et C13 pour le second monomère) qui vont renforcer la stabilité et l'ancrage de l'hélice de reconnaissance dans le grand sillon de l'ADN.

C. Site de reconnaissance de la protéine P22 c2

Le dimère P22 c2 reconnaît un fragment d'ADN de 20 paires de bases. Chaque monomère reconnaît un demi-site d'interaction séparé par un tour d'hélice. P22 se lie à six opérateurs naturels O_L1, O_L2, O_L3, O_R1, O_R2 et O_R3 dont les séquences sont données dans la Figure 58. L'alignement des différents sites révèle une séquence consensus **A**N**T**N**AAG**NNNN**CTT**N**A**N**T** (où N correspond à l'une des quatre bases A, T, C ou G). Les quatre bases centrales du site de liaison sont variables d'un opérateur à l'autre et permettraient de modifier l'affinité de liaison de la protéine (206, 375) (cf. Figure 58). En effet, au niveau de ces positions, l'ADN adopte une conformation appelée B'. Contrairement à la conformation B, un ADN ayant une conformation B' présente un petit sillon plus étroit, un Twist plus important et la présence de molécules d'eau très ordonnées. Une séquence comportant une succession d'adénines (« A-tract ») ou une séquence riche en A/T sera favorable à la formation de l'état B' alors que des séquences contenant des nucléotides G/C ne permettent pas d'obtenir l'état B' (376). Le répresseur ne formant aucune interaction directe avec ces quatre bases centrales, il est probable que P22 discrimine ces séquences très similaires par un mécanisme de reconnaissance indirecte engendré par la déformation de ces bases.

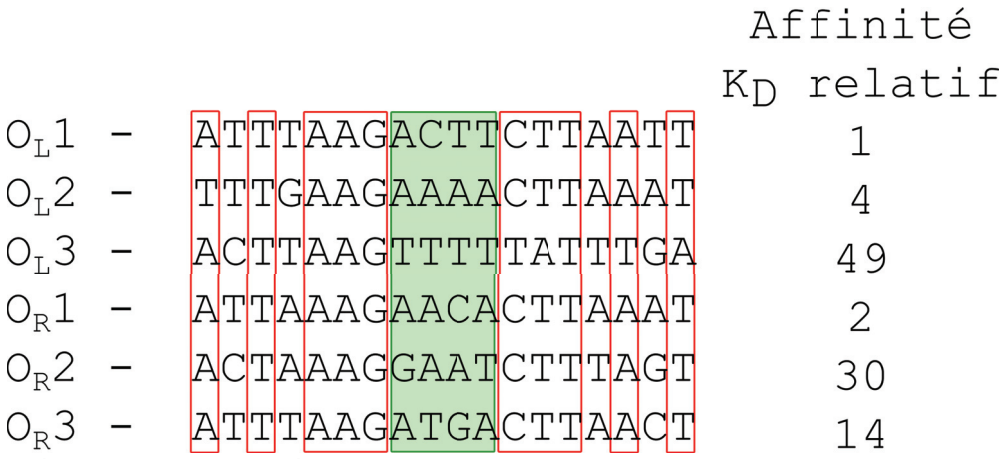


Figure 58 : Opérateurs naturels du répresseur P22 c2. Les cadres rouges indiquent les positions conservées entre les six opérateurs. Le cadre vert représente la région où la protéine ne forme aucun contact direct avec les bases de l'ADN et qui est très variable entre les opérateurs. Les valeurs du K_D sont normalisées par rapport au K_D de l'opérateur O_L1 (ou 1 = 5 x 10^{-8} M) (377).

D. Analyse de la dynamique moléculaire

1. Structure de l'ADN

La protéine P22 c2 induit de faibles déformations lors de la formation du complexe. En moyenne, sur 1 μs de simulation effectuée (compte tenu de la taille importante du système), l'ADN est courbé dans la direction de la protéine d'environ 23°. Une courbure de même amplitude et dans la même direction est observée dans la structure cristallographique (code PDB : 2R1J). On observe également une courbure de 20° dans les simulations d'ADN libre. Cependant, dans le complexe, la courbure est générée au centre de l'oligomère (position 9 à 11) alors que pour l'ADN libre, le centre de l'oligomère présente un axe hélicoïdal plutôt linéaire et des extrémités 5' et 3' courbées.

Lors de la dynamique moléculaire, le grand et le petit sillon de l'ADN du complexe deviennent plus étroits (cf. Figure 59A et B). Généralement, la modification de la largeur d'un sillon induit par une protéine produit l'effet opposé sur l'autre sillon. En d'autres mots, si, par exemple, la protéine se fixe dans le petit sillon de l'ADN et induit un élargissement de celui-ci, le grand sillon sera lui rétréci. Or ce n'est pas le cas dans ce complexe, ce qui signifie que la diminution de la taille des sillons de l'ADN n'est pas induite directement par l'insertion des hélices α de l'homodimère P22 c2, mais qu'un autre facteur en est responsable. Le paramètre hélicoïdal twist, définit le niveau d'enroulement de la double hélice. Ainsi si l'on augmente ce dernier, la double hélice va s'enrouler davantage provoquant une diminution de la largeur des sillons alors qu'une baisse du Twist provoquera un élargissement des sillons de l'ADN. Dans notre complexe ADN-P22 c2, on observe une augmentation du Twist de 40° le long du site de fixation de la protéine par rapport à l'oligomère isolé (cf. Figure 59C). Cette augmentation de Twist est notable au niveau du fragment T5-A9 et T12-A17 et au niveau du pas central T10pA11 où le Twist est augmenté de 8° alors que les paires de bases adjacentes A9pT10 et A11pT12 ne sont pas affectées.

Le répresseur P22 c2 n'affecte pas le paramètre Rise à l'exception du pas T5pA6 et T15pA16 où l'on observe une augmentation de 0,3 Å suite à l'interaction entre ces paires de bases et les valines 33 de chaque monomère du répresseur. L'angle de Roll de 11° au niveau de ces positions n'est pas induit par la présence de ces valines puisque l'on retrouve ces mêmes caractéristiques pour l'ADN libre. La formation de ces angles de Roll

positifs est intrinsèque à la séquence TpA. La présence de P22 engendre une diminution de l'angle de Roll entre les positions G8-C13 qui correspondent à la partie du site de fixation où aucune base azotée de l'ADN ne forme d'interaction avec le dimère. La paire centrale T10pA11 est encore une fois très affectée puisque l'angle de Roll devient négatif -5,4° alors qu'il est de 7,4 dans l'ADN libre.

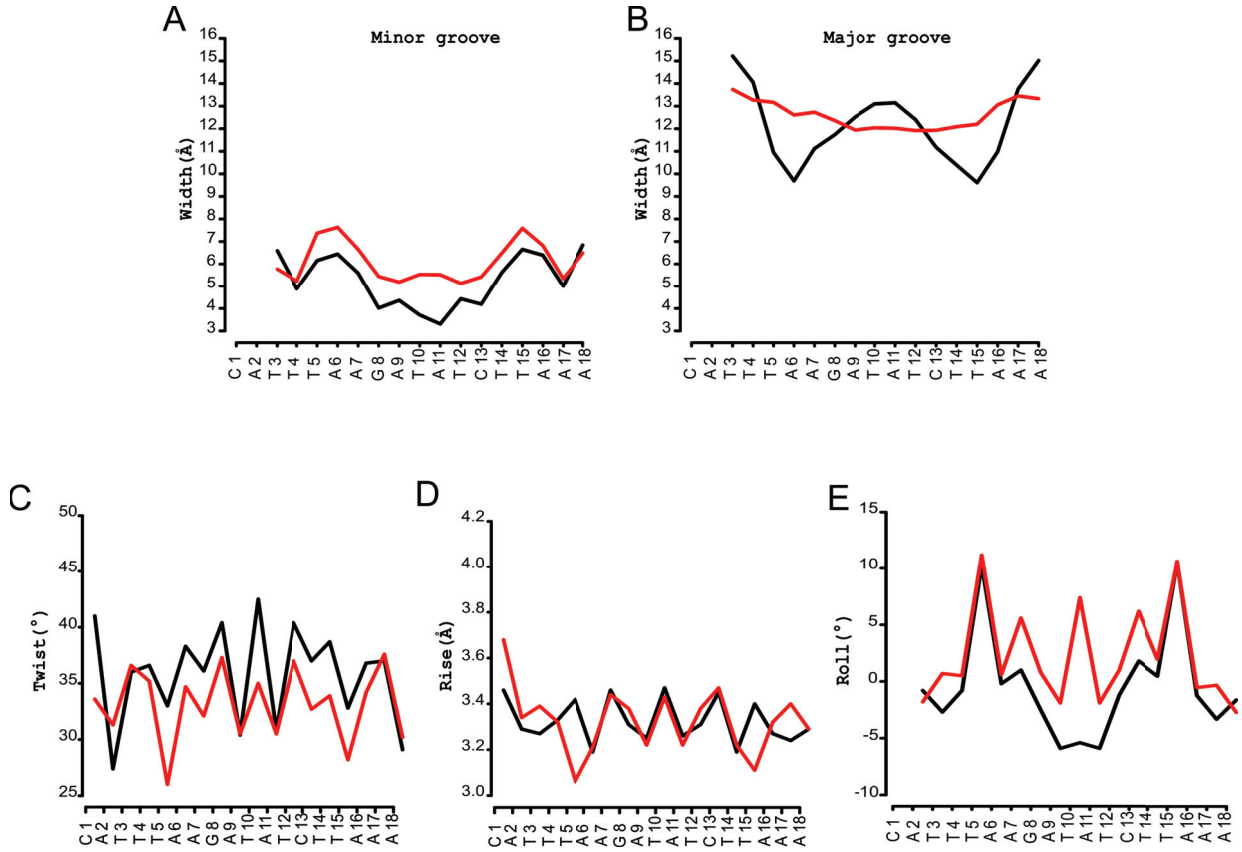


Figure 59 : Paramètres structuraux du complexe P22-ADN (lignes noires) et de l'ADN libre (lignes rouges). A et B représentent la largeur du petit et du grand sillon de l'ADN (Å). C, D et E représentent les valeurs moyennes des paramètres inter paires de bases Twist (°), Rise (Å) et Roll (°). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'.

Enfin l'interaction entre P22-ADN n'engendre que peu de modifications des paramètres du brin phosphodiester. On observe uniquement un changement dans la distribution des angles epsilon/zêta au niveau des positions T5-A6, T15-A16 qui interagissent avec la valine 33 ou le pourcentage de BII est de 4, 18, et 14 % contre 46, 45 32 et 38 % dans l'ADN libre. Le répresseur provoque également une augmentation de la conformation BII du brin Crick au niveau du pas central A10'-T11' (conformation BII présente 20 % du temps contre 2 % dans l'oligomère non lié).

2. Dynamique du complexe et des interactions ADN-P22 c2

Comme pour les trois complexes précédents, la présence de la protéine réduit la mobilité des phosphates. Cependant, on constate que cette diminution de la mobilité est uniforme sur le brin Crick alors que sur le brin Watson, notamment au niveau des positions A6-G8 et A16-A18, la présence du répresseur ne modifie pas la mobilité des groupements phosphates (cf. Figure 60A). Les informations structurales ne permettent pas pour le moment d'expliquer ce phénomène. Le groupe de Modesto Orozco à Barcelone a effectué une simulation du répresseur P22 c2(code PDB : 3JXB) durant 2 μs à l'aide du nouveau champ de force BSC1 (378). L'analyse du RMSF des phosphates de cette simulation montre une diminution de la mobilité des phosphates uniforme, quel que soit le brin. L'analyse des paramètres hélicoïdaux de ce nouveau champ de forces est néanmoins en adéquation avec les résultats obtenus lors de nos simulations avec le champ de force BSC0.

Lors de la dynamique, de nouveaux ponts salins et de nouvelles liaisons hydrogène se forment (cf. Tableau 5). Le plus souvent, les ponts salins et liaisons hydrogène se forment et se rompent entre plusieurs atomes d'un même couple acide aminé-nucléotide. Les ponts salins sont essentiellement localisés dans la région 5' du site de liaisons alors que les liaisons hydrogène sont plus présentes au niveau des bases adjacentes T10-A11. Les ponts salins les plus stables impliquent les arginines 14 et 20 qui interagissent entre 73 et 97 % du temps avec les nucléotides T18'-A19' et T3-A2. Les interactions effectuées par chacun des monomères du répresseur sont assez identiques en termes de temps de présence et de temps de vie des interactions. Comme pour les systèmes précédents, le temps de vie moyen d'une liaison hydrogène ou d'un pont salin est de l'ordre d'une centaine de picosecondes. Les résidus Gln21, Ser31 et Ser36 dérogent pourtant à cette règle en effectuant des liaisons hydrogène dont le temps de vie peut atteindre plusieurs nanosecondes voire centaines de nanosecondes (exemple Ser31 avec un temps de vie d'environ 500 ns).

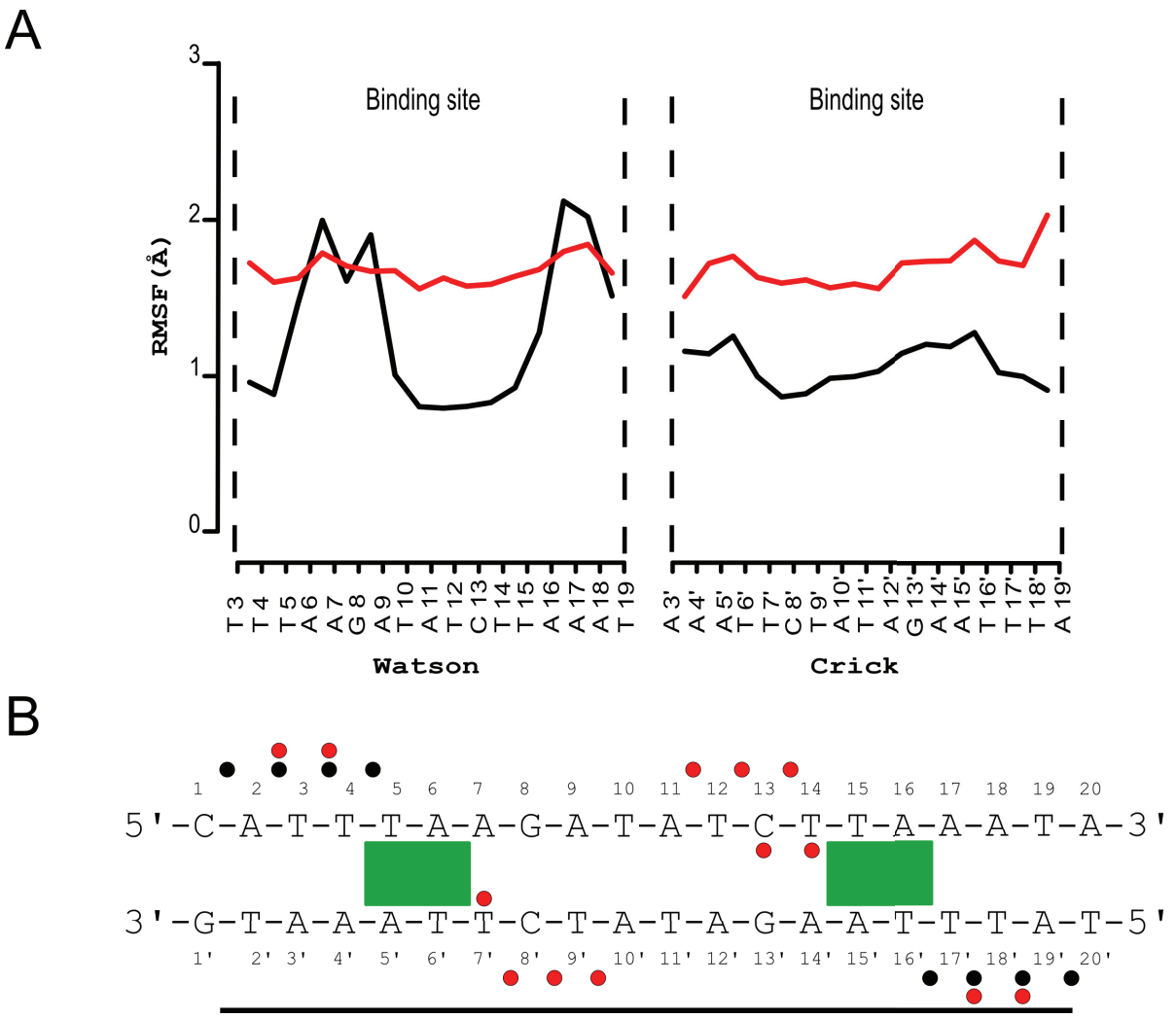


Figure 60 : Information sur la structure du complexe ADN-P22 c2. A) Moyenne du RMSF des atomes de phosphore pour les brins Watson & Crick Crick pour le complexe (lignes noires) et de l'ADN libre (lignes rouges. Les lignes verticales en pointillés délimitent le site de reconnaissance. B) Le site de fixation de la protéine est indiqué par la barre noire horizontale. Le brin « Watson » est numéroté de 1 à 20 dans le sens 5'-3' et le brin « Crick » est numéroté 1'-20' dans le sens 3'-5'. Les ponts salins, les liaisons hydrogène ainsi que la position des valines 33 de chaque monomère observées lors de la dynamique moléculaire sont indiqués par des cercles noirs, rouges et par les rectangles verts respectivement.

Tableau 5 : Temps de présence (% pres.) et temps de vie moyen (Lifetime en picoseconde) des ponts salins et liaisons hydrogène entre le brin phosphodiester (A) et les bases (B) de l'ADN et chaque monomère (M1 et M2) du répresseur (durant les 1µs de simulation. La ligne noire marque la séparation entre ponts salins (au-dessus) et liaisons hydrogène (en dessous). Les interactions notées en vertes sont présentes uniquement dans la structure expérimentale, celles en noires sont présentes dans la structure cristallographique et dans la simulation et celles en rouges sont présentes uniquement dans la simulation de dynamique moléculaire.

A

M1	Backbone	%pres.	Lifetime	M2	Backbone	%pres.	Lifetime
R11(NH1)	T17'(O1P)	24	70	R11(NH1)	T4(O1P)	12	80
R11(NH2)	T17'(O1P)	32	185	R11(NH2)	T4(O1P)	16	160
R11(NH2)	T17'(O2P)	14	40	R11(NH2)	T4(O2P)	8	35
R14(NH1)	T18'(O1P)	89	85	R14(NH1)	T3(O1P)	92	610
R14(NH2)	T18'(O1P)	61	80	R14(NH2)	T3(O1P)	67	95
R20(NE)	A19'(O1P)	14	55	R20(NE)	A2(O1P)	29	135
R20(NE)	A19'(O5')	10	40	R20(NE)	A2(O5')	5	20
R20(NH1)	A19'(O2P)	29	60	R20(NH1)	A2(O2P)	27	65
R20(NH1)	A19'(O1P)	18	50	R20(NH1)	A2(O1P)	17	45
R20(NH2)	A19'(O1P)	19	95	R20(NH2)	A2(O1P)	26	25
R40(NE)	T17'(O2P)	24	25	R40(NE)	T4(O2P)	30	25
R40(NH1)	T16'(O2P)	13	80	R40(NH1)	T5(O2P)	6	55
Q21(N)	T18'(O2P)	95	35 535	Q21(N)	T3(O2P)	100	67 910
Q21(NE2)	T17'(O2P)	98	249 260	Q21(NE2)	T4(O2P)	100	5020
S31(N)	T14(O2P)	100	166 170	S31(N)	T7'(O2P)	100	83 080
S31(OG)	T14(O2P)	100	498 530	S31(OG)	T7'(O2P)	100	498 530
S36(OG)	T17'(O2P)	99	9575	S36(OG)	T4(O2P)	99	8120
Q37(NE2)	C13(O2P)	11	75	Q37(NE2)	C8'(O2P)	3	60
W38(NE1)	C13(O2P)	95	705	W38(NE1)	C8'(O2P)	94	860
W38(NE1)	C13(O1P)	15	25	W38(NE1)	C8'(O1P)	17	25
T43(OG1)	C13(O2P)	85	305	T43(OG1)	C8'(O2P)	81	290
N46(ND2)	T9'(O1P)	42	165	N46(ND2)	T12(O1P)	15	195
N46(N)	C13(O1P)	2	140	N46(N)	C8'(O1P)	1	2725
N49(ND2)	C13(O1P)	3	137	N49(ND2)	C8'(O1P)	<1	-
N49(ND2)	C13(O5')	<1	-	N49(ND2)	C8'(O5')	<1	-

B

M1	Base	%pres.	Lifetime	M2	Base	%pres.	Lifetime
Q37(NE2)	C13(O4)	3	8	Q37(NE2)	C8'(O4)	<1	-
Q37(NE2)	T14(O4)	3	125	Q37(NE2)	T7'(O4)	1	160

3. L'environnement ionique

L'interaction du répresseur modifie la distribution des ions le long du fragment d'ADN. Le changement le plus important se produit au niveau du petit sillon au niveau du fragment central ATAT (cf. Figure 61A). On observe une molarité de 15 M, ce qui correspond à la présence d'un ion K^+ durant 75 % du temps. Au niveau de cette position centrale, on observe une organisation des ions potassium en deux couches au niveau du petit sillon. Cette organisation permet de réduire d'une part la répulsion entre les

phosphates et d'autre part de réduire la répulsion entre les acides glutamiques 44, 48, et les groupements phosphates de l'ADN (375, 379).

L'interface entre le dimère P22 c2 et le fragment central T10-T12, forme un petit canal électronégatif dans lequel des molécules peuvent transiter (206). Durant la dynamique moléculaire, il arrive que certains ions K^+ traversent également ce canal ionique. Le passage des ions potassium s'effectue dans la première couche ionique, c'est-à-dire à l'intérieur du sillon et non dans la seconde couche (entre les phosphates et glutamates 44 et 48, voir Figure 61B. Cependant le passage d'un cation dans le canal reste un événement rare et rapide (quelques dizaines de picosecondes pour traverser le canal) qui n'a été observé que 5 fois lors des 500 dernières nanosecondes de simulations. Le mécanisme et l'importance de ce canal ionique restent encore inconnus et son implication dans le mécanisme de reconnaissance encore incertaine.

Dans le grand sillon de l'ADN où la protéine se lie, la distribution ionique de 2,5 M au niveau des positions G8 et C13 est la même lorsque la protéine est présente ou non (cf. Figure 61A). Sur ces deux positions, un ion K^+ persiste près de 70 % du temps. Expérimentalement, un ion Tl^+ lié à la base G8 a déjà été observé, le second site (C13) étant occupé une lysine (379). Le rôle de ce cation dans la reconnaissance reste encore à définir, joue-t-il un rôle dans la déformation de l'ADN et donc dans la reconnaissance indirecte ; permet-il de réduire l'électrostatique défavorable entre l'ADN et le glutamate 42 ? Toutes ces questions restent actuellement sans réponse.

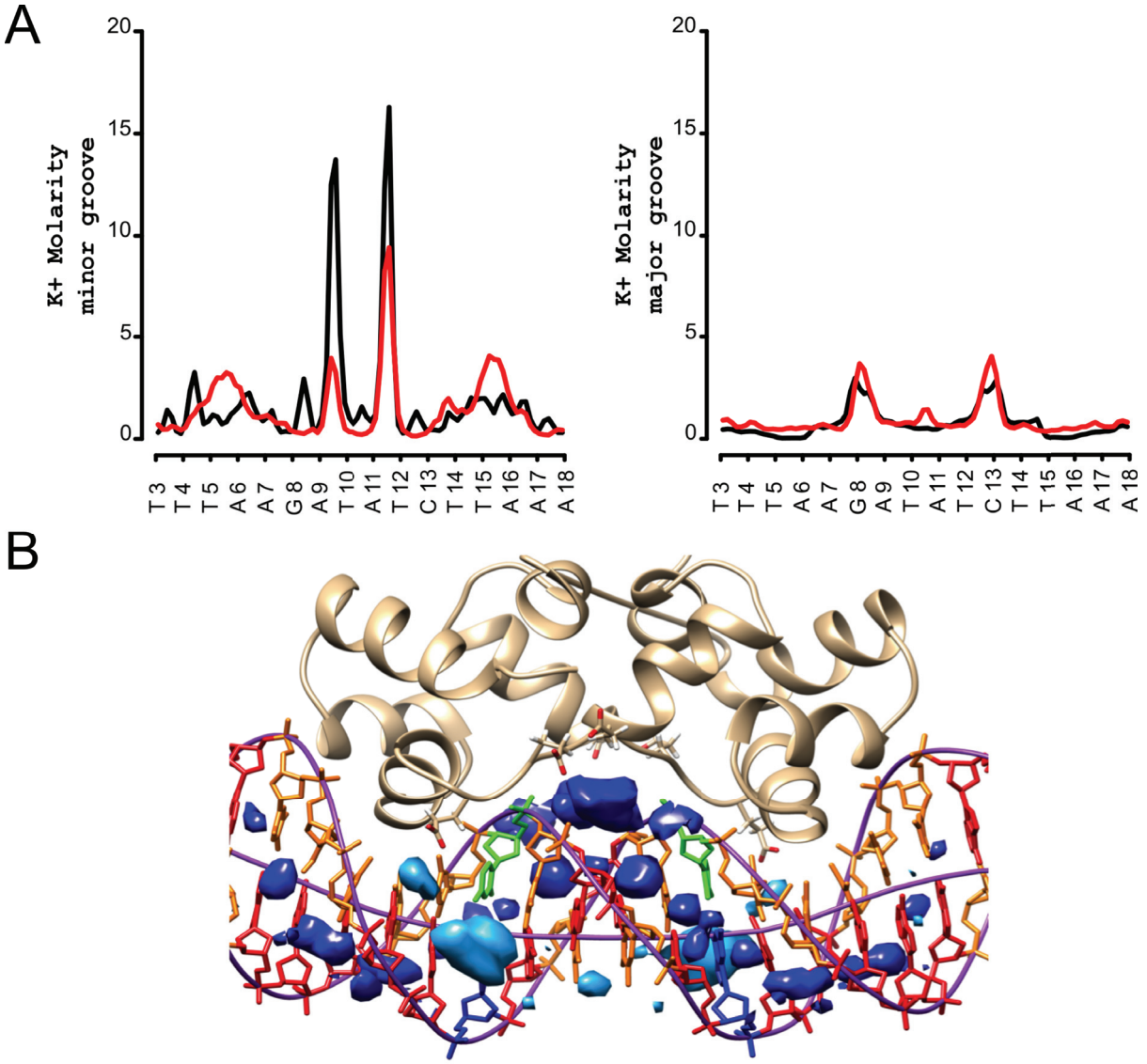


Figure 61 : Distribution 1D et 3D des ions K^+ le long du site de liaison. A) Concentration ionique de potassium dans le petit et grand sillon de l'ADN pour le complexe ADN-SRY (ligne noire) et pour l'ADN libre (ligne rouge). Les séquences sont indiquées pour le brin Watson dans le sens 5'-3'. B) Distribution moyenne des ions K^+ dans le petit (bleu foncé) et grand sillon (cyan) de l'ADN pour une densité de surface 4M. L'ADN est coloré selon la séquence (A en rouge, T en orange, G en bleu et C en vert). L'axe hélicoïdal provenant de Curves+ (324) est indiqué en violet.

4. Clustering et analyses de la spécificité de séquence

Bien que la symétrie entre chaque demi-site de liaison et les monomères soit conservée durant les 1µs de simulation en termes de surface d'interaction (613± 33 Å^2 et 588±55 Å^2 respectivement), de temps de présence et de vie des ponts salins et liaisons hydrogène, des changements de conformations se produisent de manière indépendante dans chaque site. La matrice des contacts acides aminés-base azotée de la Figure 62A met en évidence ces changements indépendants entre monomères. L'analyse des contacts par acide aminé a permis d'identifier le résidu Gln37 (résidu se trouvant dans l'hélice α qui

interagit avec le site de liaison) comme étant le principal responsable des fluctuations observées dans les matrices. Dans la structure cristallographique, le résidu glutamine 37 de chaque monomère forme une liaison hydrogène avec la base T7'(O4) et T14(O4). Lors de la dynamique, cette interaction ne se produit que rarement (3 % du temps). Le reste du temps, le résidu Q37 forme des interactions avec le groupement phosphate C8'pT10' ou C13pT12 (11 % du temps) (Figure 62 B panneau de droite) ou ne forme aucune interaction avec la double hélice (Figure 62 B panneau de gauche). En considérant les deux états (lié au brin et non lié) pour chaque monomère, il en résulte la possibilité de former quatre groupes ; I) Q37 du monomère 1 (M1) liée au brin alors que Gln37 du monomère 2 (M2) ne forme pas d'interaction ; II) Q37 de M1 n'est pas liée et Gln37 de M2 est liée au brin ; III) les deux Gln37 ne sont pas liées et IV) les deux glutamines sont liées au brin phosphodiester.

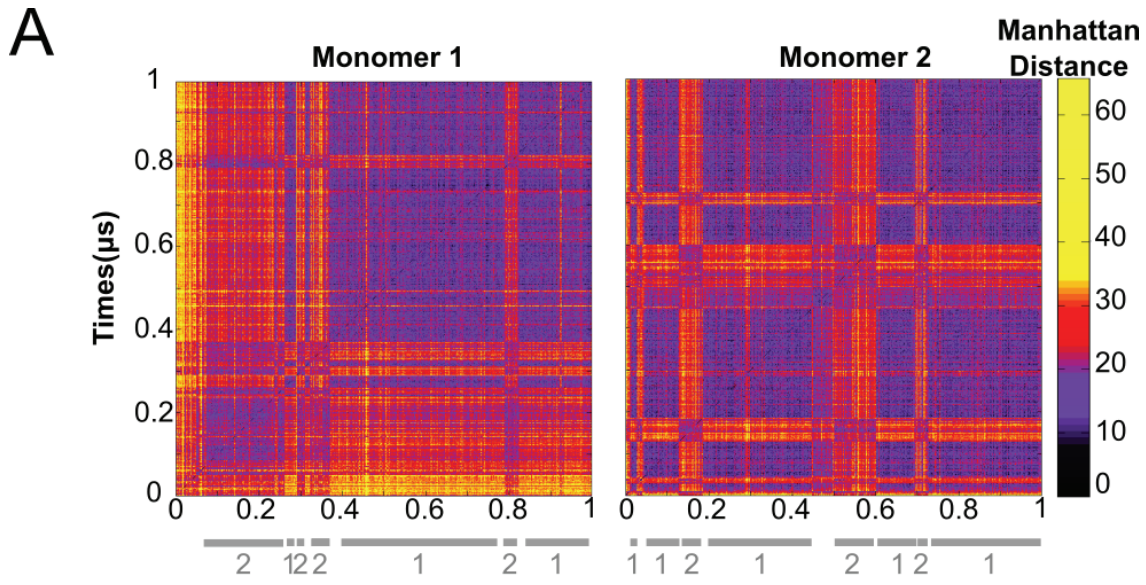
L'analyse par enfilage moléculaire sur les quatre groupes de conformations révèle que la reconnaissance du fragment 5'-TTAA-3' (position 4 à 7) du demi-site de reconnaissance ne se produit que lorsque le résidu Gln37 n'est pas lié au brin phosphodiester et qu'il ne forme aucune liaison hydrogène avec l'ADN (cf. M2 dans le groupe 1, M1 dans le groupe 2, et pour les monomères dans le groupe 3 de la Figure 63). L'IC moyen sur ces positions 4 à 7 lorsque Gln37 ne forme pas de liaisons hydrogène avec l'ADN est de 5 environ contre 2,5 d'IC lorsque Gln37 forment une interaction avec le brin. La perte de reconnaissance qui se produit lorsque Gln37 est en interaction avec le brin phosphodiester est liée à un déplacement global du monomère. À noter également que la réorientation de la chaîne latérale en direction du brin réduit les interactions hydrophobes puisque la distance entre les carbones β et δ de la glutamine et les groupements méthyles des thymines augmente.

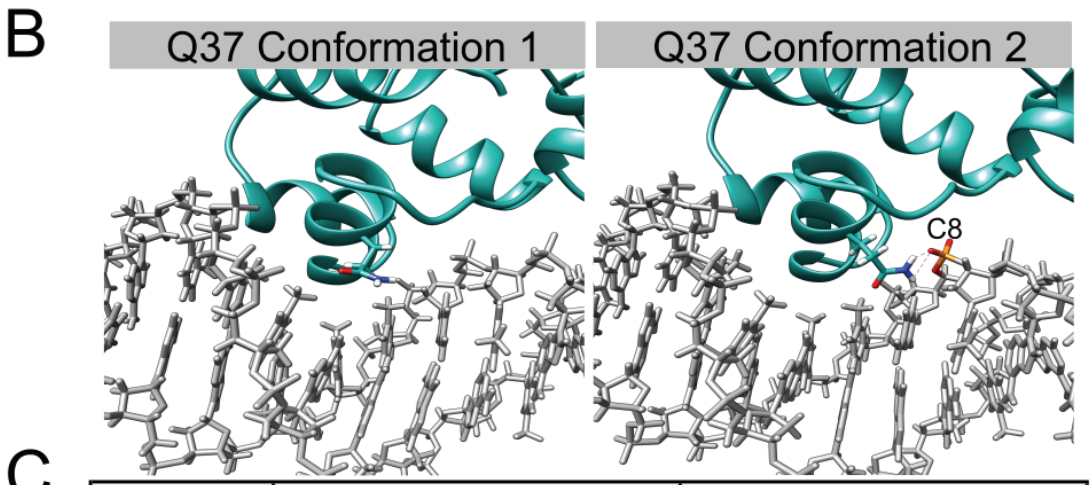
L'analyse du logo moyen (MD Consensus) de P22 c2 indique que la séquence 5'-TTAA-3' qui interagit également avec la valine 33 est assez bien détectée par la méthode ADAPT. On ne détecte pas de préférence G/C pour les positions 8 et 13 et aucune préférence pour les quatre bases centrales où aucune interaction directe acide aminé-base de l'ADN n'est observée.

La spécificité de reconnaissance décrite dans la littérature ne repose que sur l'alignement des six opérateurs naturels du répresseur. Actuellement il n'existe aucun logo pour P22 c2 qui aurait été généré à partir de données de puces à protéine ou données SELEX. La préférence de sélectivité pour une séquence riche en A/T pour la

région séparant chaque demi-site de fixation a été interprétée comme étant due à la reconnaissance indirecte résultante de la transition d'une structure d'ADN B en ADN B'. L'ADN B' se caractérise par un petit sillon plus étroit, une augmentation du twist, une organisation structurée des molécules d'eau très ordonnées ainsi que la présence de cations monovalents dans le petit sillon (379). Malgré que ces changements de conformations vers un état B' soient observés durant la simulation, aucune sélectivité de séquence n'apparaît. La présence de cations monovalents serait favorable d'un point de vue électrostatique à une sélectivité des paires A/T. Le rôle électrostatique de ces ions a été démontré dans des expériences où la mutation de Glu44 et Glu48 (résidus présents au niveau du canal ionique, cf. Figure 61B) par des résidus neutres provoque une baisse de la sélectivité ADN-protéine (375). La dynamique moléculaire effectuée lors de cette thèse confirme la présence d'ions dans le canal ionique avec deux sites particulièrement denses (cf. Figure 61B) qui favorisent certainement les paires de bases AT.

Malheureusement, la méthode ADAPT, ne permet pas d'effectuer des minimisations en solvant ou ions explicites. Les effets du solvant sont traités de manière implicite grâce à une fonction sigmoïde. Les effets explicites des ions et de certaines molécules d'eau qui semble être important dans la reconnaissance du site G8, C13 et de cette région centrale (206) ne sont alors pas pris en compte ce qui peut expliquer la perte de sélectivité observée dans notre étude.





C

Clusters	Q37 conformation M1	Q37 conformation M2
Cluster 1	Conformation 2	Conformation 1
Cluster 2	Conformation 1	Conformation 2
Cluster 3	Conformation 1	Conformation 1
Cluster 4	Conformation 2	Conformation 2

Figure 62 : Classification des structures du complexe ADN-P22 c2. A) Matrices des distances de Manhattan pour tous les contacts ADN-protéine du monomère 1 (à gauche) et du monomère 2 (à droite). L'échelle indique les distances croissantes (sens noir → jaune). Les barres grises horizontales indiquent les différentes conformations adoptées par le résidu Gln37 (Q37). B) Conformations alternatives observées pour le résidu Gln37 : positionné dans le grand sillon (à gauche) et lié au brin (à droite). Résumé des conformations adoptées par Gln37 pour chacun des clusters.

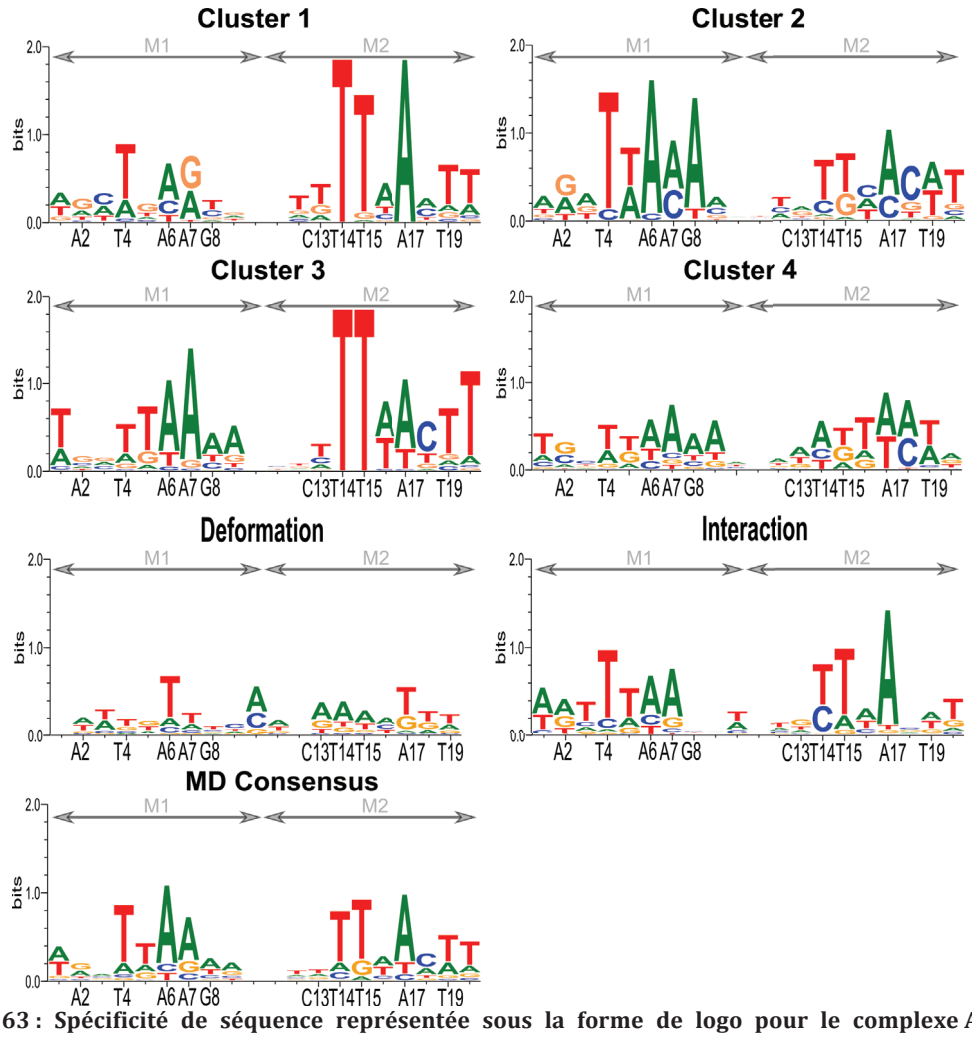


Figure 63 : Spécificité de séquence représentée sous la forme de logo pour le complexe ADN-P22 c2. La dynamique de la glutamine 37 (Q37) de chaque monomère (M1 et M2) génère la formation de quatre groupes : Q37/M1 liée au brin et Q37/M2 non liée (cluster1) ; Q37/M1 non liée au brin et Q37/M2 liée au brin (cluster2) ; Q37/M1 non liée au brin et Q37/M2 non liée au brin (cluster3) et Q37/M1 liée au brin et Q37/M2 liée au brin (cluster4). La décomposition des mécanismes d'interaction en reconnaissance indirecte (déformation de l'ADN), directe (interaction ADN-protéine) et le logo moyen sont également présentés dans les panneaux Deformation, Interaction et MD consensus respectivement. Les flèches grises délimitent le demi-site d'interaction de chaque monomère. Sous chaque logo, la séquence consensus est exprimée le long de l'axe des abscisses.

En conclusion, le répresseur P22 c2 forme une interaction stable durant la microseconde de simulation. Malgré une symétrie conservée entre les deux demi-sites de fixation, chaque monomère possède une dynamique complètement indépendante de l'autre monomère. Le résidu Gln37 de chaque monomère est particulièrement mobile et conduit à la formation de sous-états qui modulent la spécificité d'interaction entre la protéine et l'ADN. La spécificité de reconnaissance de la région centrale AATT, interprétée comme la formation d'un état B' de l'ADN, n'est pas retrouvé *in silico* par la méthode ADAPT qui ne peut reproduire explicitement le niveau d'organisation des molécules d'eau et la présence d'ions K^+ dans le petit sillon qui sont nécessaire et stabilisent la conformation B' de l'ADN.

VI. Conclusion

L'analyse des interfaces ADN-protéine grâce à la dynamique moléculaire est souvent longue et délicate puisqu'elles ne sont pas rigides. Nous avons pu constater que les liaisons de nature électrostatiques (ponts salins et liaisons hydrogène) se forment et se rompent pour former de nouvelles interactions qui n'étaient pas observées dans les différentes structures cristallographiques ou RMN. Ces interactions qui persistent en moyenne quelques centaines de picosecondes peuvent parfois être maintenu plusieurs nanosecondes, permettant d'une part de stabiliser le complexe en réduisant la mobilité du brin phosphodiester, et d'autre part de diminuer la répulsion électrostatique entre les groupements phosphates lorsque l'ADN est très déformé (courbure importante, sillon étroit). De récentes études en RMN et en dynamique moléculaire (358, 380, 381) ont démontrées l'aspect dynamique des ponts salins lysine-groupement phosphate pour lequel on observe des transitions entre deux sous-états : appariements directs (pont salin) ou appariements médiés par des molécules d'eau, dont les temps de persistances sont également de l'ordre de quelques centaines de picosecondes.

Une étude de 2016 effectuée par le groupe de Iwahara (382) réalisée sur le complexe Egr-1, a cependant montré que malgré une importante mobilité des résidus arginines et lysines, au niveau des groupements phosphates, les liaisons hydrogènes bidentées entre le groupement guanidium de l'arginine et la base guanine sont très stables et peu dynamiques. Parmi les complexes que nous avons étudiés, on observe ce comportement dans le cas de la protéine SKN-1 où les arginines R507 et R519 forment une interaction bidentée avec les bases G8 et G13. On observe également des interactions très stable impliquant la formation de liaison hydrogène entre les asparagines N10 de SRY et N511 de SKN-1 qui forment une double interaction avec deux bases adjacentes plus de 80% du temps.

A travers ces simulations, nous avons pu démontrer que la dynamique des interfaces ADN-protéine n'impliquait pas uniquement la formation et la rupture de liaisons hydrogène et de ponts salins avec le brin phosphodiester mais que des interactions protéine-bases azotées peuvent également se former et se défaire. Cependant, cette dynamique au niveau des bases de l'ADN sur la reconnaissance du site de liaison varie d'un complexe à l'autre : TBP ne semble pas affecté, alors que la dynamique du résidus Arg78 de la protéine SRY modifie modérément la séquence de

reconnaissance au niveau de l'extrémité 5' du site de liaison. La dynamique du résidus Gln37 de la protéine P22 c2 induit des changements important du site de reconnaissance 5'-TTAA-3' selon la conformation qu'il adopte. Et enfin pour le complexe SKN-1, la perte temporaire des interactions base-Arg507/Arg519 conduit à une perte de sélectivité des extrémités 5' et 3' respectivement.

La modulation de spécificité observée lors de notre étude laisse donc suggérer que la séquence consensus observée expérimentalement, peut, dans certains cas être une séquence moyenne produit par différents sous-états conformationnels, qui sont capables de reconnaître certaines parties du site de liaison. De plus, l'aspect dynamique des chaînes latérales peut jouer un rôle essentiel lors de la recherche du bon site de liaison, puisqu'elle peut permettre de faire des transitions entre des contacts spécifiques et non spécifiques et donc hypothétiquement réduire la pénalité entropique du système lors de la liaison ADN-protéine.

Chapitre V. Analyse de la spécificité de séquence par une méthode originale : Le Modèle de Protéine Modulable

I. Introduction générale et motivation

Nous avons vu précédemment que l'interface ADN-protéine était dynamique et qu'elle pouvait moduler la reconnaissance d'une ou plusieurs bases du site de liaison. On se propose dans ce nouveau chapitre de développer un outil original pour décomposer les interactions ADN-protéine afin d'identifier les propriétés de l'interface qui sont requises pour déterminer la reconnaissance spécifique du site de liaison.

Depuis la résolution des premiers complexes ADN-protéine, les chercheurs ont essayé de déterminer un ensemble de règles permettant de caractériser les interactions spécifiques qui ont lieu au sein des complexes. Ces règles ont été initialement basées sur la complémentarité entre donneurs et accepteurs de liaisons hydrogène entre les acides aminés et les bases de l'ADN (181–187). Néanmoins, dans certains complexes tels que le répresseur tryptophane, la protéine n'effectue aucune interaction directe avec les bases de l'ADN laissant suggérer que d'autres facteurs (déformation de l'ADN, interaction médiée par des molécules d'eau) sont responsables de la reconnaissance spécifique (80, 322).

En 1979, Mirzabekov et Rich ont proposé un mécanisme pour expliquer la déformation de l'ADN lors des interactions ADN-protéine (383). Leur modèle est basé sur des propriétés électrostatiques. Ils suggèrent que la neutralisation asymétrique des groupements phosphate d'une face de l'ADN par des protéines cationiques crée un déséquilibre électrostatique et induit une courbure vers la protéine. Ils proposent que ce mécanisme explique la stabilisation des nucléosomes, malgré la déformation majeure de l'ADN enroulé autour du coeur des histones (383).

En 1996, Elcock et McCammon ont abordé le rôle de l'électrostatique d 'une autre manière (384). Ils émettent l'hypothèse que l'interaction d'un corps ayant une faible constante diélectrique, tel qu'une protéine, est suffisante pour provoquer d'importants changements dans la structure de l'ADN. En effet, l'approche de la protéine engendre une désolvatation autour des groupements phosphates, créant un milieu à faible constant diélectrique. Ceci augment la répulsion entre les phosphates près de la protéine et peut induire la courbure de l'ADN, mais dans le sens opposé à celui imaginé par Mirzabekov et Rich. Pour vérifier cette hypothèse, Elcock et McCammon utilisent un modèle simple de protéine, composé de sept sphères vides non chargées se liant au petit sillon d'un ADN de seize paires de bases. Ce modèle simplifié de protéine est ensuite simulé dans un

environnement avec de l'eau et des ions implicites (forces calculées par l'équation de Poisson-Boltzmann). Lors de la simulation, le modèle de la protéine est approché de l'ADN du côté du petit sillon. Ceci provoque en effet un élargissement du petit sillon et une courbure de l'ADN dans la direction opposée à la protéine (384).

Ces deux études mettent en avant que la présence de la protéine peut avoir deux conséquences opposées sur la déformation de l'ADN. Dans un cas, la neutralisation des charges permet aux groupements phosphates de se rapprocher et donc de se courber vers la protéine alors que dans le second cas la déshydratation des groupements phosphates induit une répulsion qui permet à l'ADN de se courber dans la direction opposée à la protéine. Ce sont par conséquent, deux mécanismes susceptibles d'être retrouvés dans de vrais complexes protéine-ADN.

Nous avons élaboré un modèle plus réaliste de protéine afin d'étudier l'impact du volume de la protéine et de l'électrostatique sur la reconnaissance ADN-protéine. Ce modèle nous permettra également de vérifier les hypothèses quant à l'origine de la courbure de Rich (383) et de McCammon (384). Notre modèle de protéine modulable (MPM), est simulé en dynamique moléculaire en présence d'ADN dans un milieu contenant de l'eau et des ions explicites (cf. Chapitre 3 : protocole de simulation). Pour chaque complexe étudié dans ce chapitre, des modèles MPM avec différentes complexités ont été testés dans le but d'analyser : 1) Si le modèle MPM se lie à l'ADN de façon stable et est-elle capable de reproduire des propriétés physiques de l'ADN (déformations, courbure, etc.) ; 2) Quel niveau de complexité du modèle permet de retrouver la sélectivité de séquence de bases observée dans le complexe tout atome.

II. Construction des MPMs

A. Modèle gros grains des protéines

Afin de reproduire une protéine rigide, une modélisation par représentation gros grains a été utilisée dans cette thèse. En plus d'augmenter la vitesse de calcul, le modèle gros grains va nous permettre de décomposer les mécanismes de la reconnaissance en fonction de diverses propriétés comme le volume, la charge et la représentation de la protéine, mais surtout de déterminer l'implication de chacun de ces facteurs dans la reconnaissance. La modélisation gros grains choisie pour notre modèle est celle décrite par Zacharias pour le docking protéine-protéine (385). Dans cette représentation chaque

acide aminé est représenté par un pseudo-atome (PA) situé sur le carbone α et un ou deux pseudo-atomes situés sur la chaîne latérale (voir Figure 64).

1. Modèle avec ou sans charge

Nous avons élaboré un premier modèle très simple, le modèle Z (voir Base Commune + modèle noDA sur Figure 64), qui comporte uniquement des sphères gros grains sans charge. Il permet d'observer l'impact d'un volume vide sur les interactions avec l'ADN.

Dans un modèle plus complexe, la charge des PAs des résidus acido-basiques ont été initialement placées aux centres des PAs, déplaçant ainsi les charges d'environ 1,2 à 1,5 Å par rapport à leurs positions dans un modèle atomique. Ce déplacement des charges induit un affaiblissement significatif dans l'énergie d'interaction avec l'ADN. Pour remédier à ce problème, des modifications ont été apportées au modèle. Lors de la conversion du modèle atomique en modèle en gros grains, les charges ne sont plus délocalisées au centre du PA, mais sont maintenus au niveau des oxygènes des groupements carboxylate des chaînes latérales Asp et Glu (-0.5e par groupement carboxylate) et des azotes du groupement guanidium de l'arginine (+0,5e par azote) et du groupement amine de la lysine (+1.0e). Ce modèle sera nommé modèle Z*.

2. Modèle avec liaison hydrogène

Dans l'étape suivant, le modèle a été complexifié : modèle ZDA (voir Base Commune + modèle DA sur la Figure 64 où DA signifie **D**onneurs-**A**ccepteurs) en ajoutant des pseudo-atomes accepteurs et donneurs de liaisons hydrogène sans que ceux-ci ne soient chargés afin d'effectuer uniquement des contacts van der Waals plus précis que le modèle Z. Dans le modèle Z*DA* plus complexe, des charges partielles provenant du champ de force ff99SB ont été ajoutées sur les centres dénommés PA5, PA6 et PA7 afin de permettre la formation des liaisons hydrogène. Dans ce modèle (Z*DA*), une charge complémentaire a été ajoutée sur les pseudo-atome PA1 ou PA2 pour conserver la neutralité du résidu.

3. Modèle avec quelques résidus atomiques

Pour certains complexes, la modélisation gros grains de résidus comme les acides aminés aromatiques peut influencer grandement la stabilité de la structure. En effet, le

remplacement d'un cycle par une sphère modifie de manière importante le comportement de l'acide aminé, notamment si celui-ci est intercalé dans l'ADN. Dans des cas où certains résidus jouent un rôle clé, il est possible de les remplacer par une représentation tout atome afin de reproduire des comportements de façon plus précise. Ces résidus seront décrits au cas par cas et seront nommés : modèle R+X* (où R = Z, ZDA, Z* ou Z*DA*) dans lequel est ajouté une ou plusieurs chaînes latérales tout atome (X*) qui possèderont les charges formelles décrites dans le champ de force ff99SB.

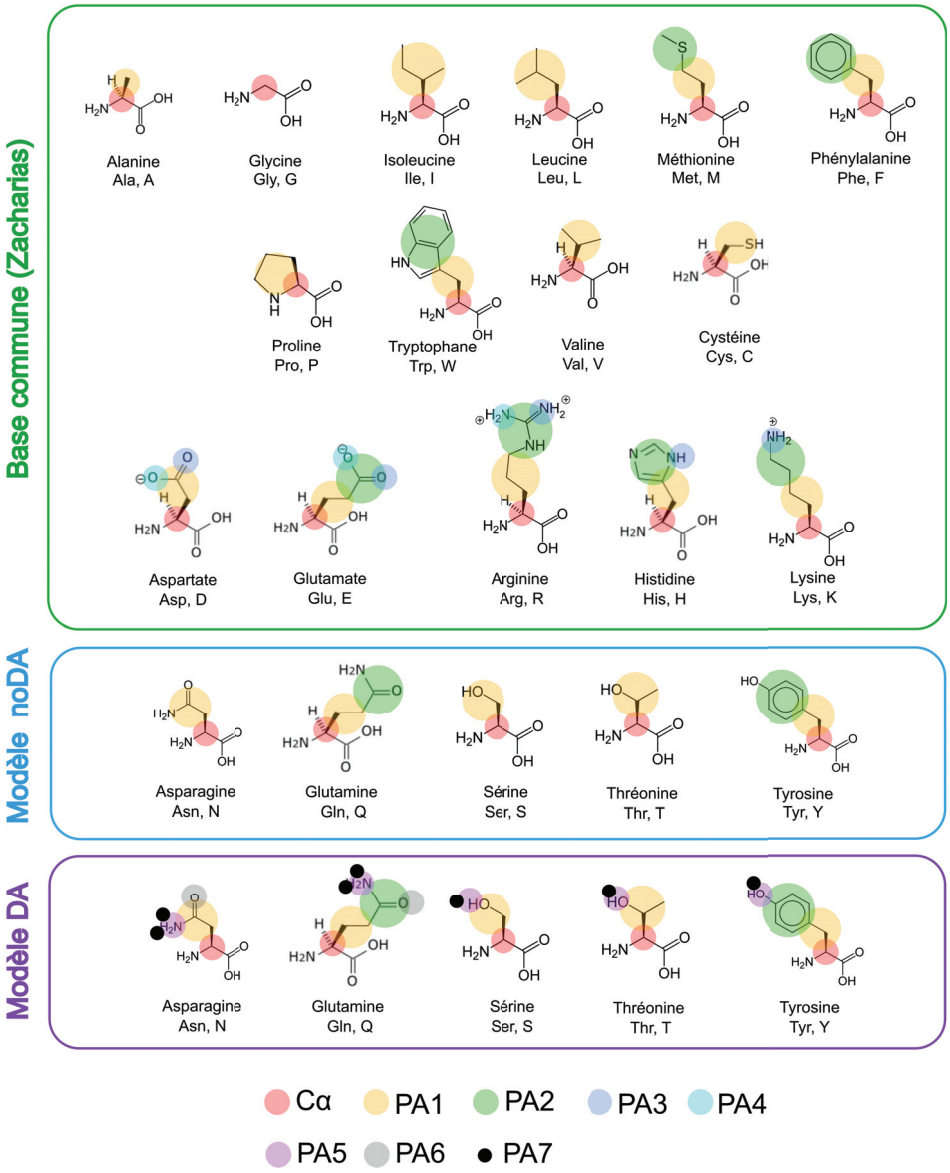


Figure 64 : Représentation gros grains des acides aminés et caractéristique des différents MPM. La nomenclature DA signifie Donneurs-Accepteurs.

B. Paramétrisation des pseudo-atomes

Dans un champ de force, il est nécessaire d'attribuer des rayons aux pseudo-atomes et une valeur de profondeur de puits de potentiel Lennard-Jones ε pour modéliser les phénomènes d'attractions et de répulsions entre deux pseudo-atomes.

1. Paramètres Lennard-Jones

(a) Rayons de Lennard-Jones

Les rayons de Lennard-Jones (LJ) ont été calculés à partir d'un jeu de 100 complexes ADN-protéine avec une résolution inférieure à 2,6 Å (cf. Tableau S1 en Annexe). Ces complexes regroupent des protéines de tailles différentes se liant dans le grand ou le petit sillon de l'ADN. L'ensemble des grandes familles structurales ADN-protéine est représenté. Pour déterminer les rayons LJ du modèle gros grains de la protéine, les PAs de la protéine et les atomes lourds de l'ADN présents à l'interface (séparés par moins de 6 Å) ont été sélectionnés. Pour chaque type de PA, seule l'interaction ayant la plus faible distance est gardée. Puis pour les PAs Cα, PA1 et PA2, on obtient une distribution des distances entre le PA et les atomes lourds de l'ADN. Le premier pic de la distribution correspond à la distance minimale pour laquelle l'énergie de vdW est égale à 0 et qui correspond à la valeur D_{min} (voir Figure 28). On calcule ensuite le rayon moyen d'un atome lourd de l'ADN $R_{ADN\ atome\ moy}$ présent à l'interface grâce à une moyenne pondérée par la fréquence d'apparition des atomes lourds de l'ADN à l'interface grâce à la relation suivante :

$$R_{ADN\ atome\ moy} = \sum (Px * Rx)$$

Où Px est la probabilité d'avoir un atome lourd en interaction avec le pseudo-atome (définit par le ratio $\frac{A_i}{A_{total}}$ avec A_i type de l'atome observé et A_{total} le nombre total d'atomes quelque soit leur type) et Rx le rayon LJ de l'atome lourd de l'ADN. Avec cette relation, le rayon moyen obtenu est de 1,84 Å. Dès lors, on peut obtenir le rayon du pseudo-atome R_{PA} par la relation :

$$R_{PA} = \left(2^{\frac{1}{6}} * D_{min}\right) - R_{atome\ moy}$$

Où D_{min} est la distance minimale à laquelle l'énergie de van der Waals entre le PA et l'atome moyen est nulle.

Pour réduire la complexité liée au nombre important de rayons LJ (20 Cα, 19 PA1 et 9 PA2), les distances ont été regroupées en six types de PAs : 1 rayon pour les Cα, 3 pour les PA1 et 2 les PA2. Puisque les PAs 3 à 7 sont une représentation des atomes N, O et H, nous avons attribués les rayons LJ présents dans le champ de force AMBER99 + BSC0 (cf. Tableau S2).

(b) Paramètre du puits de potentiel ε

Afin d'avoir un modèle le plus simple possible, un ε unique a été attribué à l'ensemble des pseudo-atomes Cα, PA1 et PA2. Ce paramètre ε a été déterminé de telle façon que l'énergie d'interaction van der Waals du modèle gros grains, soit la plus proche possible de celle du modèle tout atome. Cet ajustement a été réalisé sur plusieurs protéines de tailles différentes et appartenant à différentes familles structurales. La Figure 65 montre le résultat obtenu avec ε = 0.6 kcal.mol^{-1} après 1 ns de simulation de dynamique moléculaire pour le modèle tout atome et le modèle Z*DA*. Pour les PAs 3 à 7 le puits ε est celui des atomes N, O et H présents dans le champ de force AMBER99 + BSC0.

Le tableau récapitulatif des différents rayons LJ, valeurs de ε et charges est donné en annexe Tableau S2.

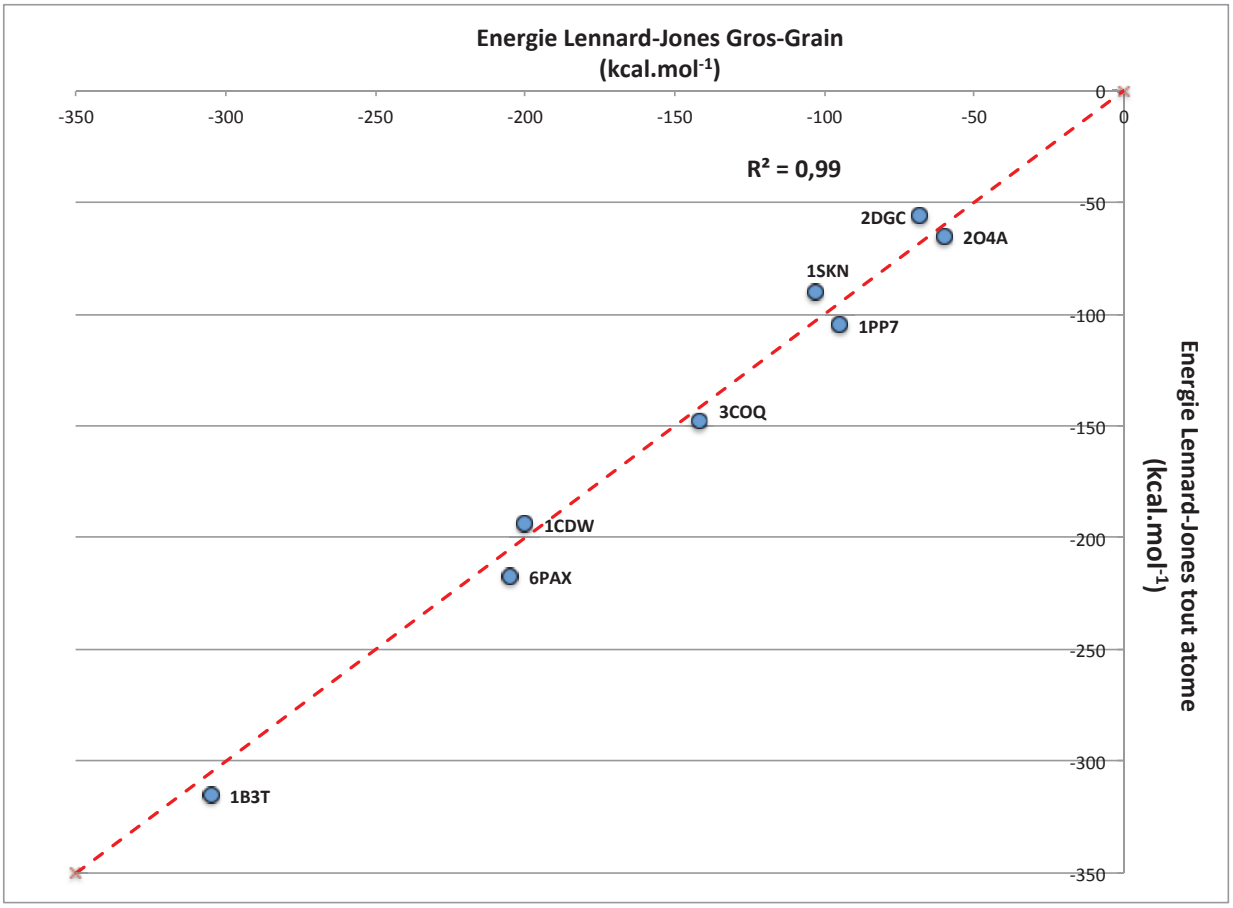


Figure 65 : Corrélation entre l'énergie de Lennard-Jones obtenue pour les modèles tout atome et gros-grain Z*DA* pour différents complexes ADN-protéine.

2. *Paramètres électrostatiques*

Dans le modèle Z*, tous les pseudo-atomes possèdent une charge de 0 à l'exception de l'arginine qui porte une charge de +0,5e sur PA3 et PA4, lysine porte une charge de +1e sur PA3 et deux charges de -0,5e sur PA3 et PA4 pour aspartate et glutamate. Le modèle Z*DA* est plus complexe, car l'ajout de charges sur les groupements polaires pour rendre possible les interactions hydrogène nécessite l'ajout de charges sur le PA précédent afin d'assurer l'électroneutralité de l'acide aminé. Pour les résidus non polaires, la charge partielle des PAs est de 0. Dans chacun des modèles Z* et Z*DA*, le premier résidu protéique porte une charge de +1 situé sur le PA Cα et le dernier résidu une charge de -1 également porté par le PA Cα.

C. Création et rigidification des MPMs

Au cours de cette étude, l'idée première est la décomposition des interactions protéine-ADN par l'utilisation d'un modèle le plus simplifié possible. Dans cette optique, nous avons choisi de figer les mouvements internes à la protéine en créant un modèle entièrement rigide. Dans un système composé de N particules, il faut 3N-6 contraintes pour bloquer tous les degrés de liberté internes. L'approche initiale consiste à établir un ensemble de trois contraintes uniques par pseudo-atome. Les trois pseudo-atomes sont choisis de manière aléatoire dans la structure afin de former des contraintes aussi bien courtes que longues. Pour obtenir une protéine rigide, le pseudo-atome i ne doit pas être dans le plan formé par les trois pseudo-atomes tirés aléatoirement et ne doivent pas être proches l'un de l'autre.

1. L'algorithme

Étape 1) Pour chaque PA (Cα et PA1 à PA7), l'on tire aléatoirement trois autres PAs. Si l'une des associations PA_i, PA_j existe déjà, on tire de nouveaux PAs, sinon l'on calcule les trois distances entre le PA_i et chacun des 3 autres PAs.

Étape 2) À partir des 4 PAs, dont le sommet est le PA_i, on crée un tétraèdre où la distance sommet-base (distance cercle rouge-vert) est de 1Å comme indiqué sur la Figure 66. La distance entre les PAs aléatoire (cercles bleus, verts) est variable.

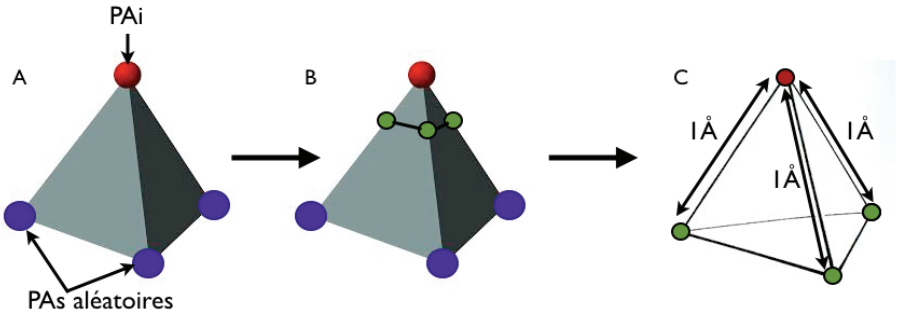


Figure 66 : Schéma indiquant la démarche de création du tétraèdre de 1 Å de côté à partir du PA d'intérêt (PA_i) et des 3 PAs pris aléatoirement dans la structure.

Étape 3) On calcule le volume de ce tétraèdre dont le volume optimal doit être de 0,118 $Å^3$, correspondant au volume d'un tétraèdre régulier de 1 Å de côté. Pour faire converger l'algorithme, on autorise une déviation autour de cette valeur optimale de 30 % (soit 0,118 $Å^3$ ± 0,035 $Å^3$), valeur pour laquelle l'algorithme est capable d'attribuer des contraintes pour les différentes structures étudiées lors de la thèse. Si ce critère n'est pas rempli, on retourne à l'Étape 1.

Étape 4) On ajoute une contrainte entre tous les PAs successifs de type Cα.

Les premiers tests de simulation de ces MPMs ont cependant révélé que cette approche n'est pas suffisante pour maintenir la structure rigide avec les contraintes quadratiques appliquées par le logiciel de simulation. Pour rigidifier un MPM, une nouvelle contrainte a été ajoutée entre le pseudo-atome i et le pseudo- atome i+1 qui s'avère être suffisante pour bloquer entièrement les mouvements de la protéine.

2. Paramètres appliqués aux MPMs

Une fois les contraintes du modèle définies, il est nécessaire de définir la constante de force qui va être appliquée pour maintenir la contrainte harmonique de distance entre les PAs. Pour obtenir un minimum de mouvement au sein de la protéine, une contrainte très forte de 1000 kcal.mol.$Å^{-2}$ est appliquée.

III. Déconvolution des interactions appliquées à différents complexes ADN-protéine

Le modèle de protéine modulable a été appliqué sur les quatre complexes TBP, SRY, SKN et P22 c2 qui ont fait l'objet d'analyse dans le chapitre 4 et pour lesquels l'on sait que la méthode ADAPT donne des résultats en adéquation avec les données expérimentales. La transformation de ces modèles atomiques en MPM peut nous permettre d'identifier des propriétés de l'interface qui sont nécessaires dans la détection des bases nucléiques lors de la reconnaissance. Les résultats avec les modèles simplifiés seront comparés aux résultats des modèles tout atome du précédent chapitre. Tous les MPMs ont été élaborés à partir du modèle atomique minimisé et équilibré en dynamique moléculaire pendant 1ns, permettant la relaxation du complexe. Les modèles gros grains sont ensuite simulés durant 500 ns dans un solvant explicite en présence de 150 mM de KCl.

A. Les MPMs de TBP

1. Rappel des propriétés structurales de l'ADN dans le complexe tout atome

La protéine TBP se lie spécifiquement à une séquence consensus de 8 paires de bases TATAWAAR à l'aide d'un large feuillet β inséré dans le petit sillon de l'ADN. Cette insertion de TBP provoque une importante déformation de l'ADN. Le petit sillon est élargi (environ 4 Å par rapport à un ADN isolé). La présence de la protéine induit une diminution de l'enroulement de l'ADN (diminution du twist de 85° le long du site de liaison). La protéine possède deux phénylalanines Phe193 et Phe284 intercalées (cf. Figure 34) aux positions T5pA6 et A11pG12 provoquant une rupture de l'empilement des bases (angle de roll positif de 52° et 39° respectivement) et une courbure de l'ADN de 60° dans la direction opposée à la protéine.

2. Modèles préliminaires

Quatre modèles de MPMs ont été construits et simulés en dynamique moléculaire pour la protéine TBP. Les quatre premiers modèles sont composés uniquement d'acides aminés gros-grains, et sont Z, ZDA, Z* et Z*DA* (voir tableau au sein de la Figure 64 pour un rappel de ces codes).

L'analyse des résultats préliminaires révèle que malgré l'absence de charges, les MPMs de TBP restent liées à l'ADN. Cependant, cela n'est pas suffisant pour reproduire la déformation de l'ADN, en particulier l'angle de roll positif entre les positions T5pA6 et A11pG12 (voir valeurs moyennes issues de la trajectoire de dynamique moléculaire sur la Figure 67). L'ajout de charges (modèle Z*, Z*DA*) ne permet pas de restituer cette propriété, suggèrent qu'elle est liée à l'intercalation des phénylalanines (données non présentées).

Pour mieux reproduire la déformation de l'ADN et donc améliorer la spécificité de reconnaissance avec notre MPM, les chaînes latérales des résidus Phe193, Phe284 intercalés ainsi que les résidus Phe 210 et Phe301 (qui renforcent l'intercalation des phénylalanines intercalées) ont été modélisés en tout atome. Cependant, ceci n'a pas permis de reproduire précisément les modifications structurales de l'ADN (données non présentées). Lors des recherches bibliographiques, il s'est avéré que deux résidus : la proline 285 et l'alanine 194 sont des résidus clés qui contrôlent l'orientation de la protéine TBP sur l'ADN (386). Ces deux résidus ont donc été représentés de manière atomique dans les modèles présentés ci-après.

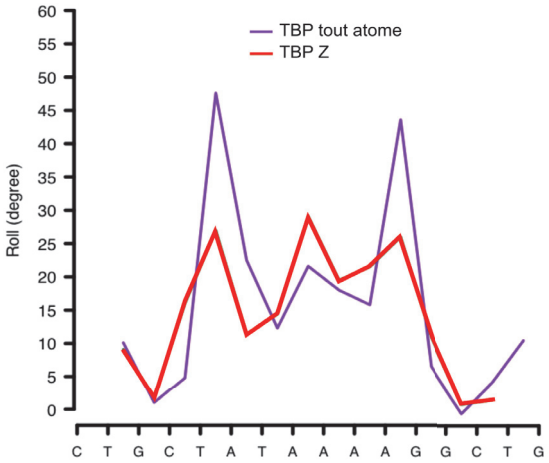


Figure 67 : Angle de Roll (en degré) pour le complexe TBP tout atome (en violet) et pour le MPM Z (en rouge). La séquence est indiquée dans le sens 5'-3' pour le brin Watson.

3. Influence du modèle sur la structure, la distribution des ions et la spécificité de reconnaissance

Le MPM de TBP est donc gros-grains à l'exception des résidus Phe193, Phe210, Phe 284, Phe 301, Ala194 et Pro285 qui possèdent une chaîne latérale atomique permettant de reproduire la déformation de l'ADN et l'orientation correcte de la protéine. Ces résidus tout atome seront présents dans les modèles Z+FAP*, ZDA+FAP*, Z*DA+FAP* et Z*DA*+FAP*.

(a) Modèle Z+FAP*

À l'issue des 500 ns de simulation, bien que toujours liée, le MPM altère la structure de l'ADN en rendant le petit sillon plus large d'environ 1 Å au niveau des positions T5pA6 et A11pG12 et en élargissant le grand sillon de 1,5 Å en moyenne le long du site de reconnaissance par rapport au complexe atomique.

Le modèle Z+FAP* induit des modifications structurales au niveau du pas A8pA9 où le Rise et l'angle de Roll augmentent de 1,8 Å et 10° respectivement par rapport à la structure atomique. De plus, ce modèle non chargée influence l'intercalation de la phénylalanine F193 au niveau de la paire de bases T5pA6, réduisant l'angle de Roll de plus de 20° par rapport à la structure atomique (cf. Figure 69).

Dans le petit sillon de l'ADN, la présence de la protéine gros grains ne permet pas à des ions de venir s'intercaler au niveau des bases le long du site de liaison entre les positions T4 et G12, comme montré dans la Figure 70. En revanche, dans le grand sillon, la concentration d'ions K^+ à tendance à augmenter en amont du site de reconnaissance (positions G3pC4), mais également au niveau de l'extrémité 5' (T5-A8) où la concentration est augmentée de 1,3 M par rapport au complexe tout atome.

L'inspection visuelle des logos de spécificité issus des matrices de poids générées par la méthode ADAPT sur les structures extraites des trajectoires de dynamique moléculaire, présentées dans la Figure 71, révèle que le modèle Z+FAP* maintient une reconnaissance spécifique entre les positions 9 et 12 identique à la reconnaissance observée en tout atome (corrélation de Pearson 0,95 par rapport au modèle atomique). L'intensité du signal de reconnaissance est similaire : IC de 3,8 pour le modèle Z+FAP (logo MD Consensus) contre 4,6 dans le modèle atomique (logo MD Consensus). En revanche, ce modèle ne permet pas une reconnaissance complète du motif TATA à l'extrémité 5' du site de reconnaissance, en particulier pour les deux bases centrales A6-pT7 (corrélation avec le modèle tout atome de 0,54).

(b) Modèle ZDA+FAP*

Ce modèle n'améliore ni la conformation liée de l'ADN (largeur des sillons, paramètres inter-paires de base, voir Figure 68 et Figure 69), ni la distribution des ions observés dans le modèle précédent (cf. Figure 70) et n'apporte aucune information complémentaire sur la spécificité de reconnaissance (cf. Figure 71).

(c) Modèle Z*+FAP*

Ce modèle permet de reproduire plus précisément la largeur des sillons puisque l'on a une variation de 0,1 Å au niveau du petit sillon et une différence de 0,5 Å au niveau du grand sillon entre ce modèle chargé et le complexe atomique. Les paramètres Twist, Rise et Roll sont également mieux reproduits dans Z*+FAP* que dans les modèles précédents, sans être identique avec le modèle tout atome (voir Figure 68 et Figure 69).

On voit également que la présence de charges sur la surface de la protéine rétablit la distribution des ions le long du grand sillon (cf. Figure 70).

L'ajout de charges sur les résidus acido-basiques n'améliore pas la reconnaissance (corrélation globale de 0,88 avec le modèle tout atome contre 0,87 pour le modèle Z) et l'intensité du signal de l'extrémité 3' (IC de 3,3 contre 3,75 dans les modèles sans charges Z et ZDA). Elles permettent cependant, de récupérer une préférence pour le motif TA avec une corrélation de 0,86 et 0,89 en position 5 et 6 avec le modèle atomique.

(d) Modèle Z*DA*+FAP*

La présence de charges partielles sur les donneurs et accepteurs de liaisons hydrogène permet de reproduire plus fidèlement l'influence de la protéine sur la conformation de l'ADN, notamment au niveau des bases 8 et 9 qui sont en contact avec les résidus Asn163 et Thr309 où les paramètres Twist, Rise et Roll observés pour ce modèle sont presque identiques à ceux observés dans la structure tout atome.

Par contre, ce modèle n'apporte aucune information complémentaire sur la distribution des ions le long de l'ADN par comparaison avec les modèles précédents.

Avec ce modèle, on retrouve une spécificité de reconnaissance pour le motif 5'-TATA-3' (voir les logos Deformation et MD Consensus) qui n'était pas présente dans les modèles précédents (corrélation de Pearson de 0,98 pour le motif TATA avec le modèle tout atome). Au niveau du motif TATA, on ne retrouve pas de liaisons hydrogène entre la protéine et les bases, mais des liaisons hydrogène entre la protéine et le brin phosphodiester (cf. structure cristallographique Figure 34B). Ces interactions améliorent la reconnaissance par un mécanisme indirecte (voir logo Deformation sur la Figure 71). La présence de liaisons hydrogène rétablit également la reconnaissance directe, principalement au niveau des bases 8 et 9 (voir logo Interaction du modèle Z*DA*+FAP* sur la Figure 71).

Quel que soit le modèle, il semble que le nucléotide A en position 10 ne soit pas spécifiquement reconnu par nos protéines gros grains (cf. Figure 71 MD Consensus).

Pour conclure, la présence de charges n'est pas essentielle dans la spécificité de reconnaissance du site 3' du site de reconnaissance. La seule présence du volume de la protéine dans cette région suffit à maintenir complexe stable et une sélectivité partielle puisque les modèles Z+FAP* et ZDA+FAP* ne suffisent pas à retrouver le motif TATA à l'extrémité 5'. Le simple ajout de charges sur les résidus Arg/Lys et Asp/Glu (modèle Z*DA+FAP*) n'est également pas suffisant pour retrouver ce motif TATA. L'ajout de donneurs/accepteurs de liaisons hydrogène en plus des charges sur les résidus acido-basiques (modèle Z*DA*+FAP*) permet de retrouver une préférence, bien que modérée pour le motif TATA.

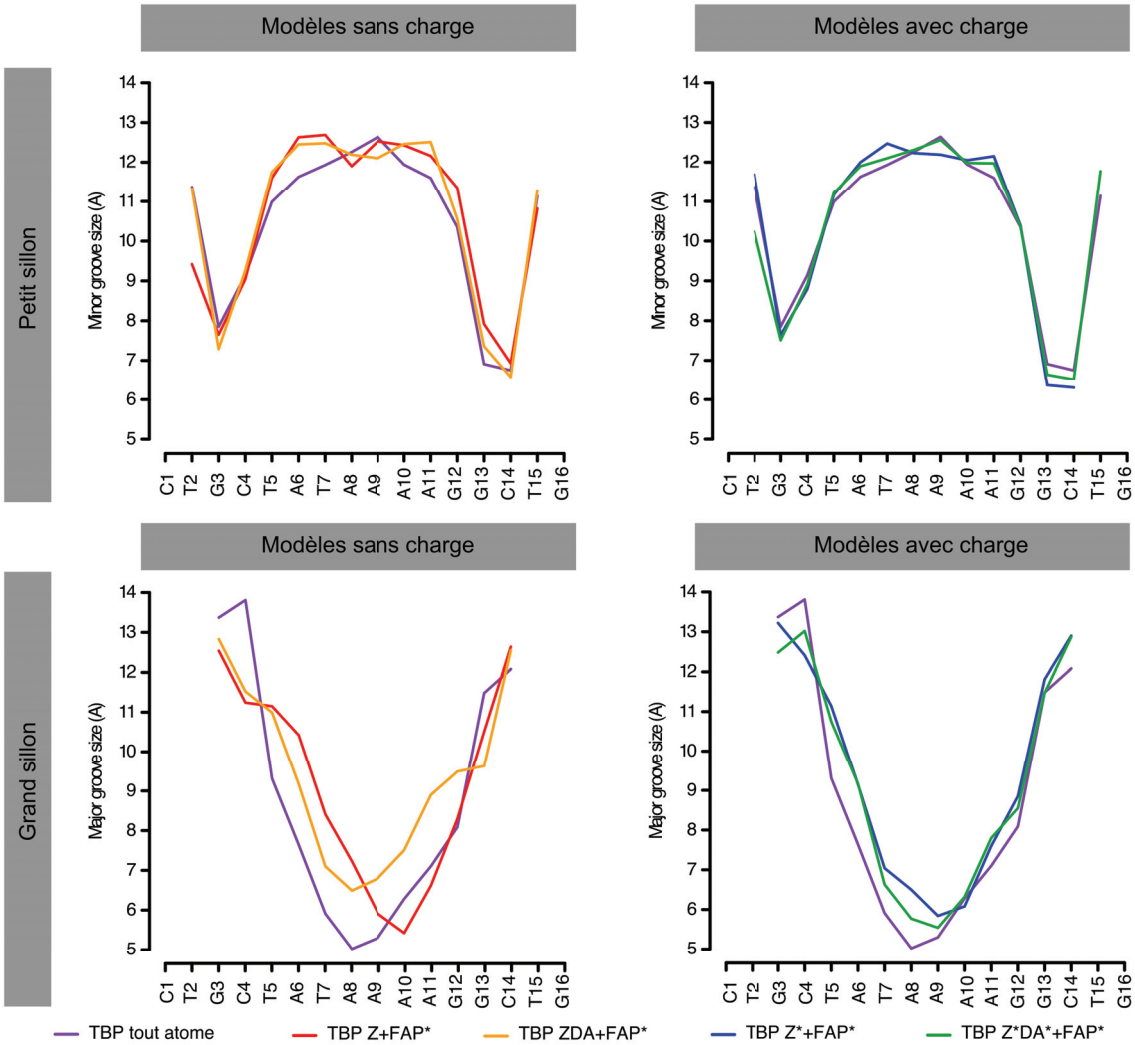


Figure 68 : Largeur du petit et du grand sillon de l'ADN pour différents MPMs et pour le complexe TBP tout atome. La séquence est donnée dans le sens 5'-3' pour le brin Watson.

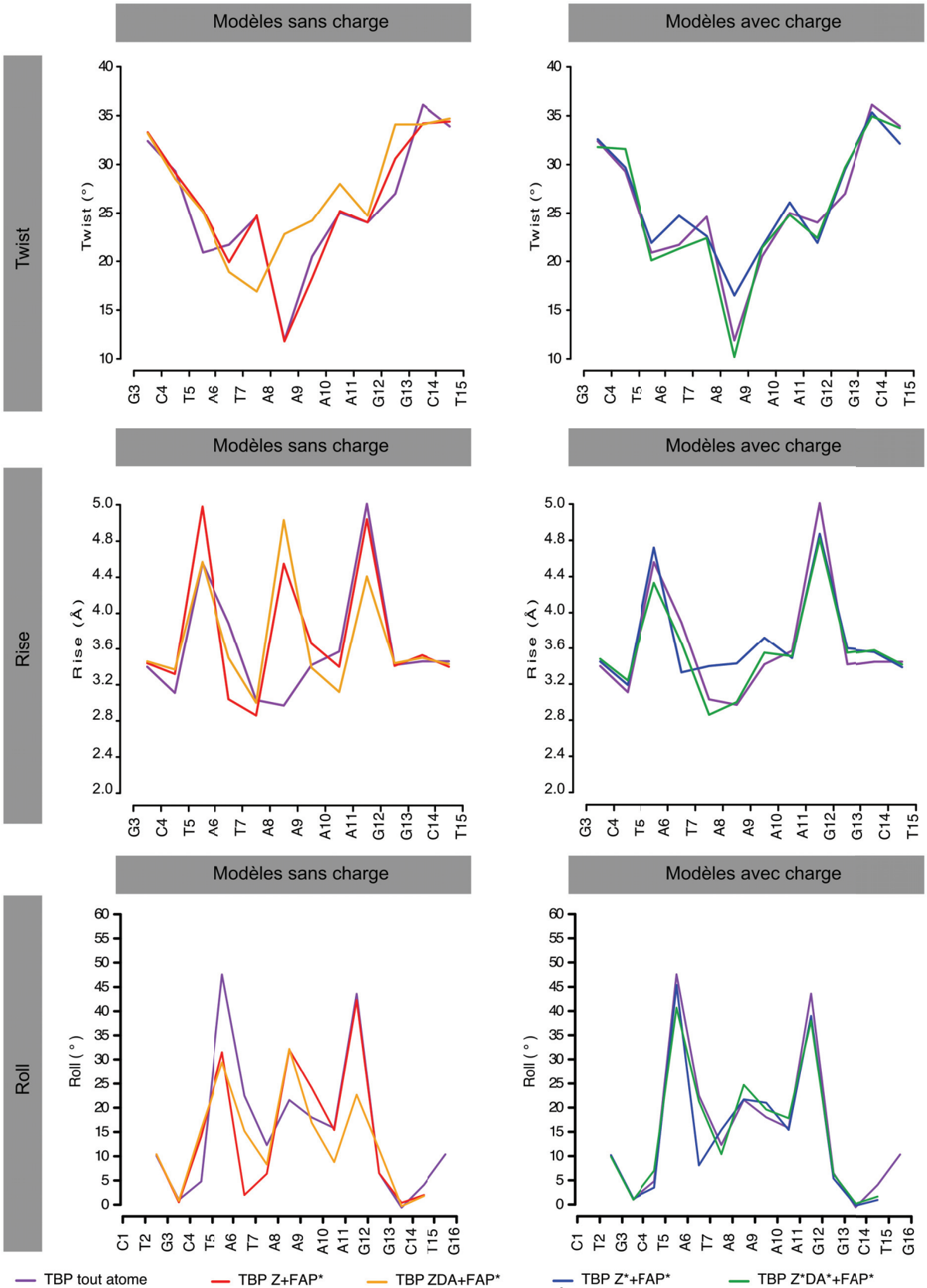


Figure 69 : Moyenne des paramètres hélicoïdaux Twist (°), Rise (Å) et Roll (°) le long du site de liaison pour les différents modèles de MPMs et pour le complexe TBP tout atome. Les séquences sont indiquées dans le sens 5'-3'.

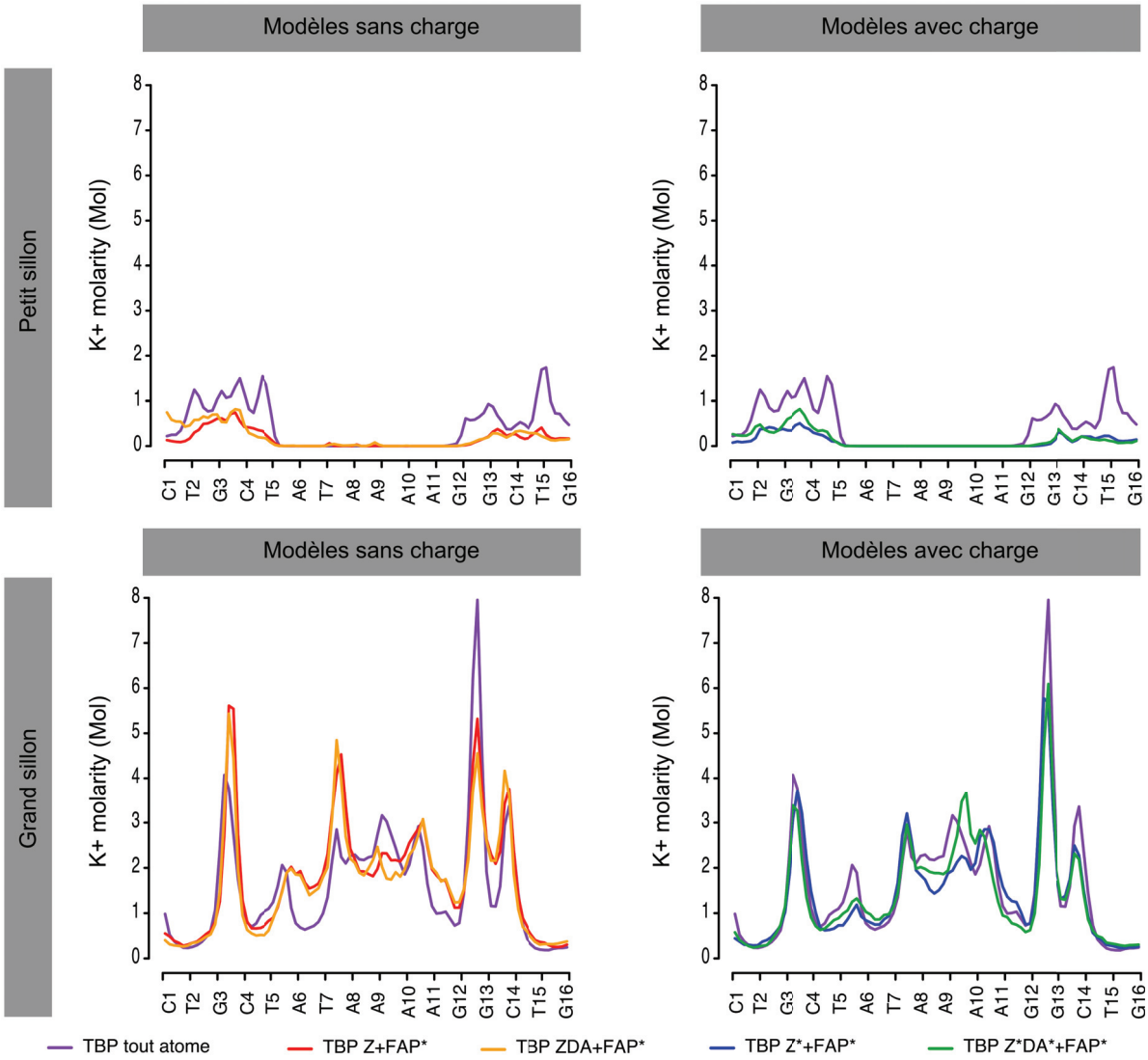


Figure 70 : Concentration en ions K^+ pour les différents MPMs de TBP le long des sillons de l'ADN.

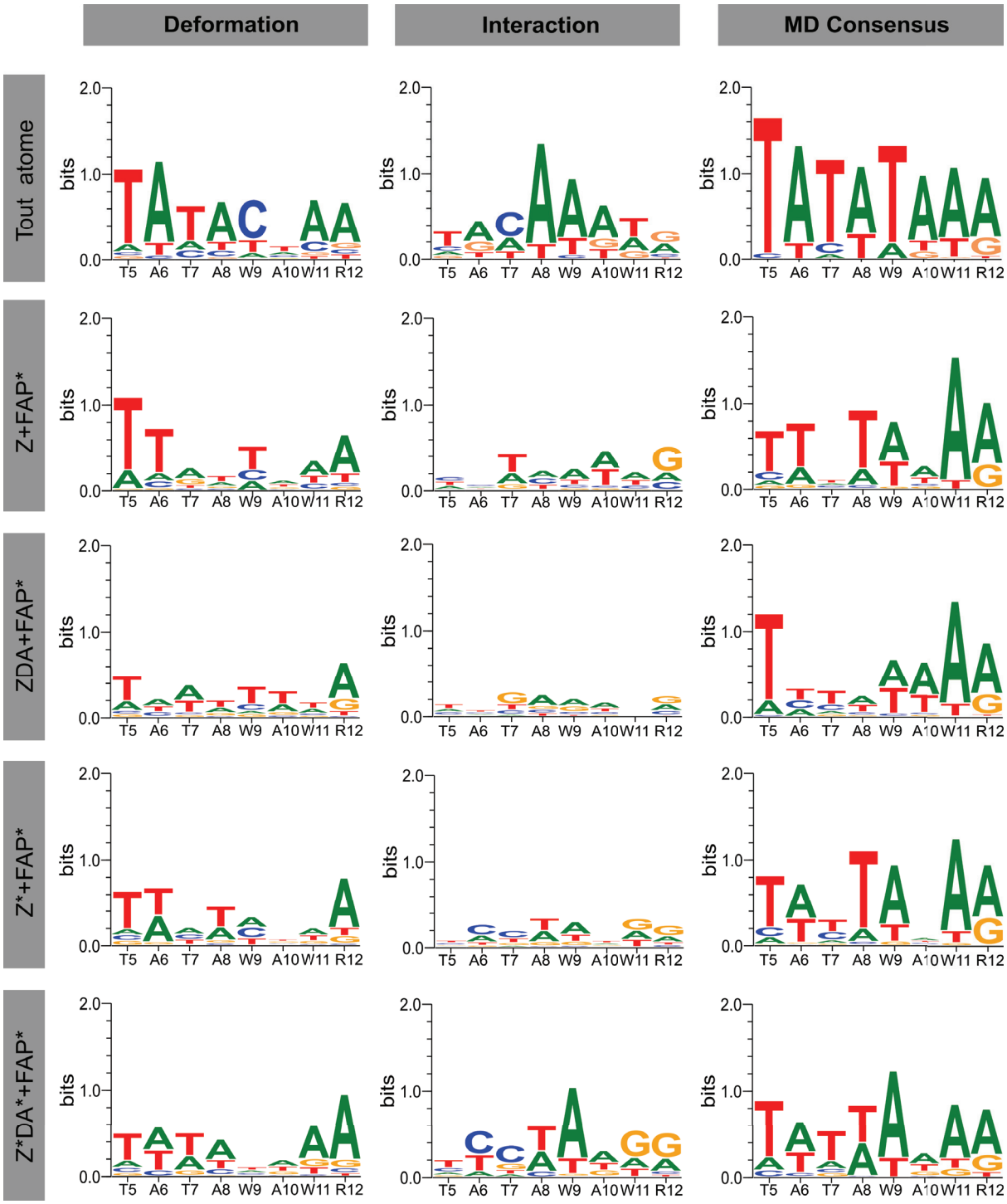


Figure 71 : Logos de spécificité de reconnaissance pour les différents modèles gros grains et pour le modèle TBP tout atome. Pour chaque cas, la contribution de la déformation (reconnaissance indirecte), de l'interaction ADN-protéine (reconnaissance directe) et la contribution totale (MD Consensus) sont représentées.

B. Les MPMs de SRY

1. Rappel des propriétés structurales de l'ADN dans le complexe tout atome

La protéine SRY se lie à un fragment d'ADN de 14 paires de bases possédant une séquence consensus WAACAAW. SRY se lie à l'ADN au niveau du petit sillon via une hélice α et une queue C-terminale flexible chargée positivement. L'insertion de l'hélice au niveau du petit sillon provoque une modification locale des paramètres hélicoïdaux, notamment en élargissant le petit sillon, en diminuant le Twist (avec une diminution de 41° entre les positions A8 et C11, voir Figure 44). La présence de l'isoleucine 13 déstabilise l'empilement des bases A8pA9 augmentant positivement l'angle de Roll de +40° et créant une courbure globale de l'axe hélicoïdale de 61° dans la direction opposée à la protéine. Enfin SRY, effectue de nombreuses interactions électrostatiques avec les brins et les bases de l'ADN, résumées page 106.

2. Influence du modèle sur la structure, la distribution des ions et la spécificité de reconnaissance

Cinq modèles de protéine modulable ont été construits pour la protéine SRY. Les quatre premiers modèles sont composés uniquement d'acides aminés gros-grains, et correspondent aux quatre modèles Z, ZDA, Z* et Z*DA* présentés précédemment page 149. Un dernier modèle de type Z*DA* a été créé, dans lequel les chaînes latérales des résidus Arg7 et Asn10 ont été modélisées en tout atome : modèle Z*DA*+RN*.

(a) Modèle Z

Malgré l'absence totale de charges sur le modèle Z, SRY maintient le complexe ADN-protéine, mais induit une modification différente de la conformation de l'ADN par rapport à ce qui est observé dans le complexe tout atome (cf. Figure 72 et Figure 73). Ce modèle ne permet pas de reproduire correctement la largeur du petit sillon, car l'on observe une augmentation de la largeur du petit sillon d'environ 1,5 Å le long du site de reconnaissance entre les positions G4 et A9 et une diminution de la largeur du grand sillon de 1 Å par rapport au complexe tout atome. En revanche, ce modèle affecte peu les

autres paramètres hélicoïdaux inter-paires de bases (voir Figure 73). On observe une diminution du Twist de 2° en moyenne le long du site de liaison Cette diminution du Twist est cependant plus importante entre les positions A9-A10 (environ 8°) où l'isoleucine 13 est intercalée. Dans ce complexe, l'isoleucine 13 a été modélisée selon le modèle gros-grains présenté page 151. Malgré cette simplification, ce modèle est capable de maintenir la déformation qu'induit l'intercalation de ce résidu, comme le suggèrent les paramètres Rise et Roll en position A9-A10 (cf. Figure 73).

Notons que, quel que soit le MPM considéré, on observe une augmentation de la largeur du petit sillon en amont du site de reconnaissance (position C2-T3) qui est en interaction avec la queue flexible C-terminale de la protéine. Dans nos modèles, cette queue C-terminale étant rigide, il n'est pas anormal d'observer une différence qui pourrait sans doute être corrigée en ajoutant de la flexibilité aux modèles.

Le modèle Z altère néanmoins l'environnement ionique du complexe dans les deux sillons de l'ADN. On observe une concentration ionique toujours plus importante dans les deux sillons par rapport au modèle atomique (cf. Figure 74). Contrairement au complexe tout atome pour lequel aucun ion n'est présent dans le petit sillon entre les positions G4 et C14, l'absence de charge sur l'arginine 7 de la protéine gros grains, qui est en contact avec les bases A6pC7, permet à un ion K^+ de venir se positionner entre la chaîne latérale et le brin phosphodiester. L'absence de charge induit aussi une augmentation de la concentration ionique de 4,1 M à l'extrémité 5' du site de fixation. Cette augmentation s'explique par le fait que l'extrémité 5' est en interaction avec la queue C-terminal de SRY qui contient trois lysines et une arginine. L'absence de ces charges positives dans l'environnement immédiat de l'ADN permet donc à des ions de s'intercaler. Ce modèle n'est donc pas suffisant pour reproduire l'environnement ionique du complexe ADN-SRY.

La Figure 75 nous montre que malgré une structure simplifiée et sans charges, le modèle Z maintient une certaine spécificité (voir logo MD Consensus de la Figure 75). La spécificité du complexe provient à la fois de la déformation de l'ADN pour les extrémités 5' et 3' du site de reconnaissance, et de l'interaction bases-protéine gros grains pour les bases centrales 6 et 7 où l'on retrouve une préférence pour les bases AC malgré que la contribution totale (déformation + interaction) conduise à une perte de la sélectivité pour la base C7 (voir logo Z Interaction *versus* logo MD Consensus sur la Figure 75). Le volume de la protéine induit un changement de spécificité pour la base A8 en T (cf. logo MD Consensus de la Figure 75). Globalement ce modèle permet de maintenir une

sélectivité (corrélation de Pearson de 0,73 par rapport au modèle tout atome) et un IC important du logo en dehors des positions 7 et 8 avec une valeur de 4,6 contre 5,4 dans le modèle atomique.

(b) Modèle ZDA

L'ajout des pseudo-atomes donneurs et accepteurs de liaisons maintient l'interface protéine-ADN, mais n'améliore pas la modélisation de la déformation de l'ADN (cf. Figure 72 et Figure 73).

Le profil ionique de ce second modèle non chargé est très similaire au modèle Z (cf. Figure 74). Cependant, l'on observe une différence significative au niveau de la position T3 du petit sillon et A10 du grand sillon où les concentrations en potassium sont multipliées par 2,5 par rapport au modèle tout atome. Ce modèle n'est toujours pas suffisant pour reproduire la distribution des ions observée dans le complexe tout atome.

La spécificité de reconnaissance est diminuée lorsque l'on ajoute les atomes donneurs et accepteurs de liaisons hydrogène sans leurs charges : corrélation de Pearson de 0,33 et IC de 4 dans le modèle ZDA (logo MD Consensus Figure 75) contre corrélation de 0,73 et IC de 4,6 dans le modèle Z avec une perte de sélectivité notable au niveau de la base A6 (voir Figure 75, logo Interaction et MD Consensus).

(c) Modèle Z*

La présence de charges sur les résidus acido-basiques reproduit avec plus de précision la conformation de l'ADN en particulier la largeur des sillons qui s'accorde parfaitement avec le résultat tout atome (cf. courbe bleue Figure 72 et Figure 73). Ce modèle de SRY, permet également de reproduire les paramètres inter-paires de bases Twist, Rise et Roll dans la partie 5' du site de reconnaissance entre les positions G4 et C7, voir Figure 73.

Lorsque l'on ajoute des charges sur les résidus acido-basiques on retrouve une distribution des ions dans le grand et le petit sillon de l'ADN similaire à celle observée pour le complexe tout atome, à l'exception d'un ion K^+ en position 12 qui n'est pas présent dans le complexe tout atome et qui est sans doute lié à un repositionnement de l'arginine 31 dans le complexe tout atome qui empêche l'ion de venir s'intercaler.

L'ajout de charges sur les résidus acido-basiques induit une perte de spécificité en position 7, mais accroît la spécificité de reconnaissance au niveau des positions 8 à 10 (logo MD Consensus de la Figure 75) où l'on observe une forte préférence pour les bases

WAT (où W =A/T) avec un IC de 5,2, ce qui est supérieur à ce qui est observé dans le complexe tout atome ou l'IC est de 2,8. La présence de charge sur le modèle Z* permet également de retrouver une préférence accrue pour le nucléotide A en position 8 qui n'était pas retrouvée par les précédents modèles sans charges (voir Figure 75, logo MD Consensus). Pour les positions 3 et 4, on remarque que les charges améliorent la préférence pour les bases T et A alors que celles-ci étaient peu sélectionnées dans les modèles précédents (IC de 2,6 dans le modèle Z* contre 0,9 en moyenne dans les modèles Z et ZDA). On remarque également que dans le modèle tout atome, les positions 3 et 4 ne sont pas fortement reconnues (IC de 0,7). Cette différence avec les MPMs peut être due aux multiples conformations adoptées par la queue flexible C-terminale dans le complexe tout atome et dont la flexibilité n'est pas modélisée dans nos protéines simplifiées, car rappelons-le, le logo moyen provient de plusieurs structures.

Ce modèle permet donc de déterminer que la charge des résidus acido-basiques est nécessaire pour retrouver une spécificité de reconnaissance des bases en positions 3, 4 et 8

(d) Modèle Z*DA*

La présence de liaisons hydrogène au niveau de l'interface SRY-ADN n'est pas fondamentale dans le maintien de l'interaction du complexe, et n'apporte pas d'amélioration des paramètres hélicoïdaux par rapport au modèle Z* (voir Figure 72 et Figure 73) et ne modifie pas significativement la distribution des ions autour du complexe.

La formation de liaisons hydrogène n'altère pas la reconnaissance le long du site de liaison à l'exception de la position 7 pour laquelle on observe une sélectivité pour les nucléotides A ou G (voir logo MD Consensus de la Figure 75). Ce résultat est néanmoins différent de la sélectivité obtenue avec le modèle tout atome (T ou C en position 7) indiquant que notre modèle nécessite un niveau supplémentaire de décomposition, notamment par l'ajout de chaînes latérales atomiques.

(e) Modèle Z*DA*+RN*

Jusqu'à présent l'analyse de spécificité par la méthode des MPMs entièrement gros-grains ne nous permettait pas d'obtenir une sélectivité pour les bases C ou T en position 7 comme observée dans la structure tout atome du complexe. Nous avons donc choisi d'ajouter deux résidus en représentation atomique : Arg7 et Asn10. Lors de la

simulation du complexe tout atome, ces deux résidus ont rapidement adopté une conformation différente de celle de la structure RMN.

L'ajout de chaînes latérales tout atome pour les résidus Arg7 et Asn10 n'apporte aucune information complémentaire sur la conformation de l'ADN et la distribution des ions le long de l'ADN.

La Figure 75 révèle que l'ajout de flexibilité pour ces deux chaînes latérales permet de restituer une préférence pour les nucléotides T ou C en position 7 et permet de mieux reproduire la sélectivité du site de reconnaissance : corrélation de Pearson globale de 0,99 par rapport au modèle tout atome contre 0,65 en moyenne pour les autres modèles.

En conclusion, la présence de charges n'est pas indispensable à la protéine SRY pour maintenir une interaction stable avec l'ADN, mais n'est pas suffisante pour reproduire la déformation de l'ADN et notamment la largeur des sillons. Cette propriété est restaurée lorsque l'on ajoute des charges sur les résidus acido-basiques. La présence de charges permet également de restituer l'environnement ionique le long du fragment d'ADN. En termes de reconnaissance, les charges permettent une reconnaissance spécifique des bases centrales 7 et 8 du site consensus alors que la déformation de l'ADN suffit à expliquer la majeure partie de la reconnaissance (voir logo MD Consensus pour les modèles sans charges dans la Figure 75). Finalement, une modélisation atomique flexible des chaînes latérales des acides aminés Arg7 et Asn10 semble importante pour la reconnaissance de la base C7.

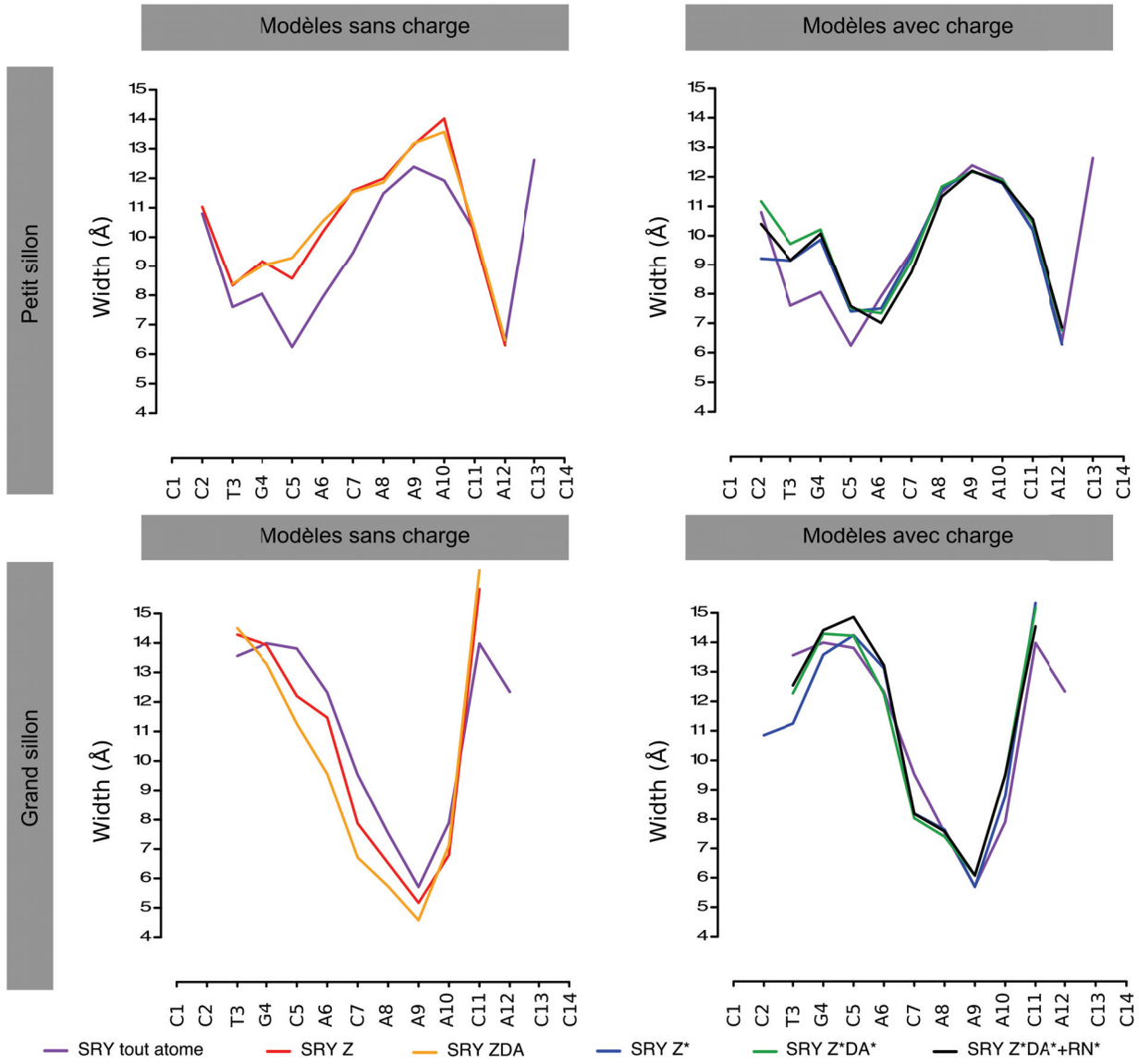


Figure 72 : Largeur du petit et du grand sillon de l'ADN pour différents MPMs et pour le complexe SRY tout atome. La séquence est donnée dans le sens 5'-3' pour le brin Watson.

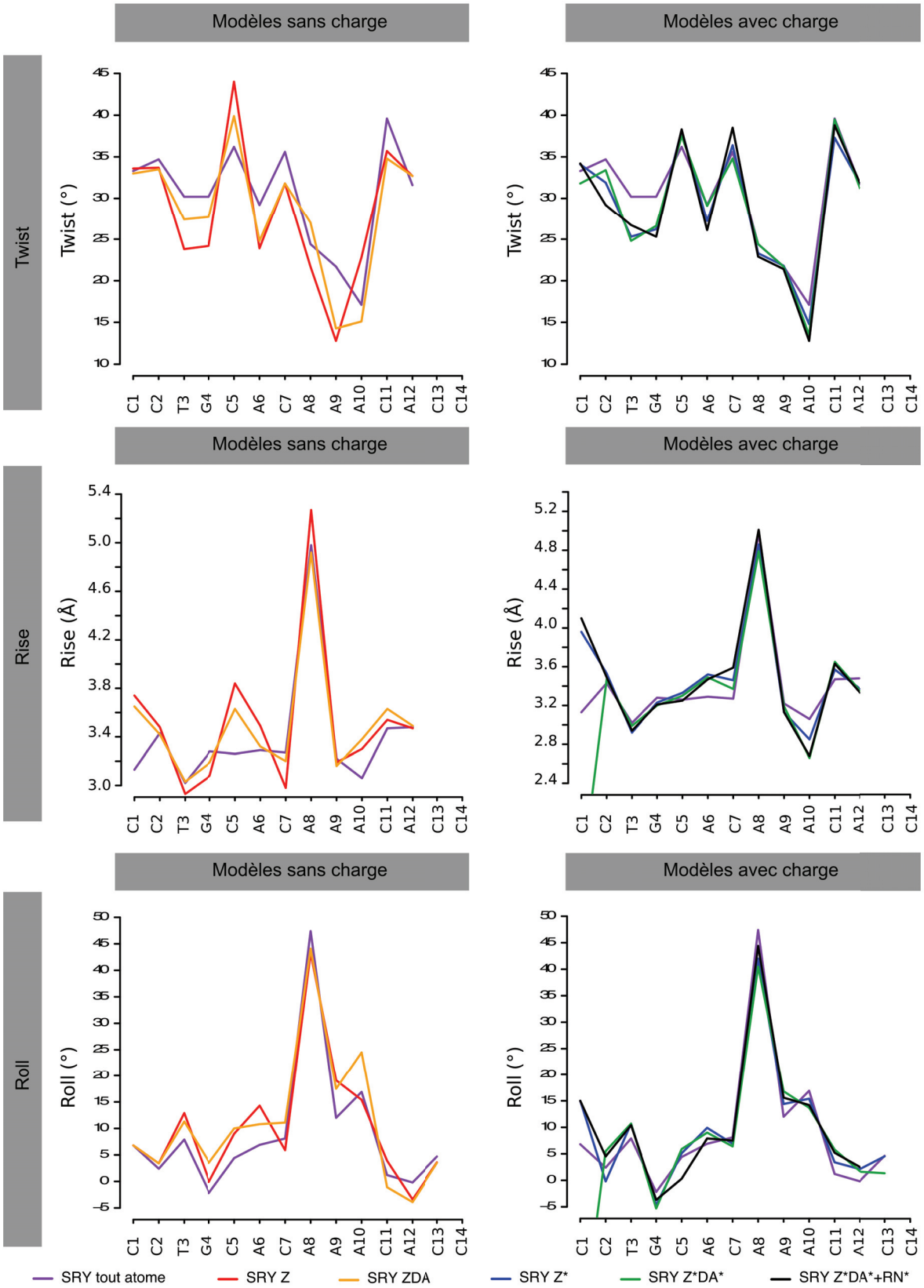


Figure 73 : Moyenne des paramètres hélicoïdaux Twist (°), Rise (Å) et Roll (°) le long du site de liaison pour les différents MPMs et pour le complexe SRY tout atome. Les séquences sont indiquées dans le sens 5'-3'.

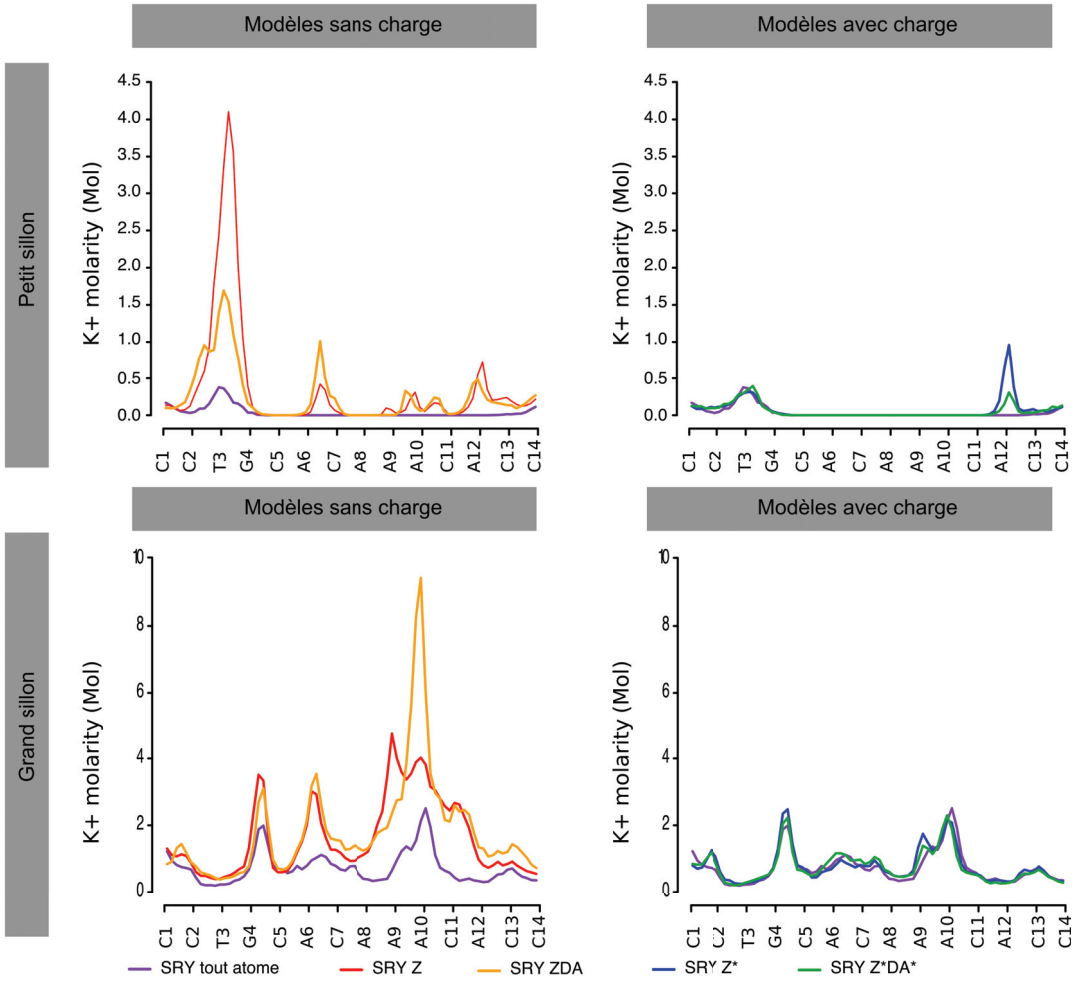


Figure 74 : Concentration en ions K^+ pour les différents modèles de SRY le long des sillons de l'ADN.

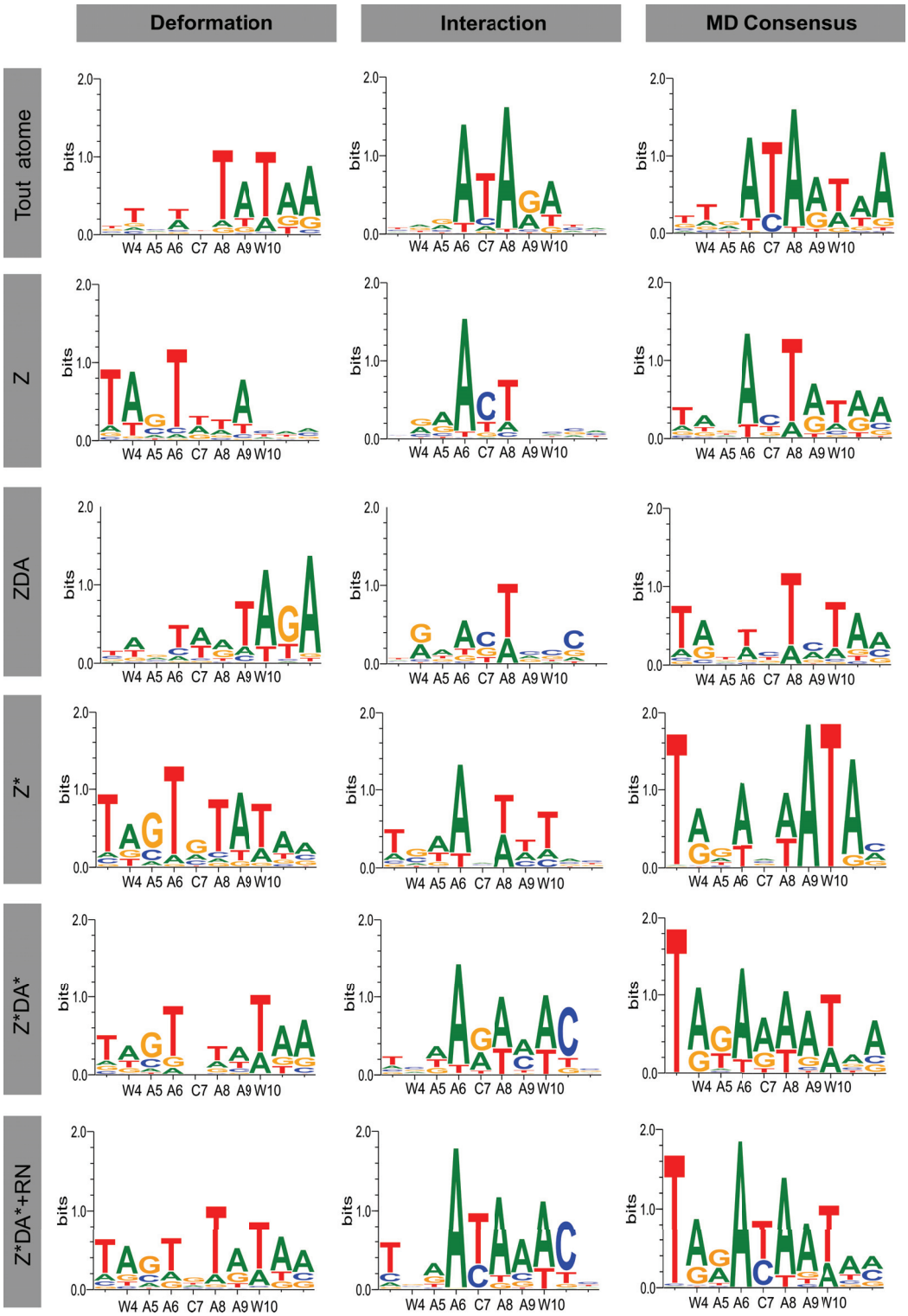


Figure 75 : Logos de spécificité de reconnaissance pour les différents modèles gros grains et pour le modèle SRY tout atome. Pour chaque cas, la contribution de la déformation (reconnaissance indirecte), de l'interaction ADN-protéine (reconnaissance directe) et la contribution totale (MD Consensus) sont représentées.

C. Les MPMs de SKN-1

1. Rappel des propriétés structurales de l'ADN dans le complexe tout atome

La protéine SKN-1 se lie à un fragment d'ADN de 17 paires de bases possédant la séquence GTCAT reconnue par une hélice basique analogue à l'hélice de la famille des leucine zipper. SKN-1 possède un bras N-terminal qui reconnaît des séquences riches en A/T. L'insertion de l'hélice de SKN-1 dans le grand sillon de l'ADN ne provoque que peu de déformation de l'ADN (voir Figure 51) et permet à la protéine d'effectuer quatre contacts directs avec les bases du site de reconnaissance et deux avec les bases flanquantes en position 5' (cf. Figure 52B).

2. Influence du modèle sur la structure, la distribution des ions et la spécificité de reconnaissance

Pour le complexe SKN-1, les quatre modèles gros-grains présentés page 149 ont été utilisés : Z, ZDA, Z* et Z*DA*. Pour ce complexe, aucune chaîne latérale tout atome n'a été ajoutée à la protéine gros grains de SKN-1.

(a) Modèle Z

Ce modèle de SKN-1 provoque une dissociation du complexe après quelques nanosecondes de simulations, permettant à l'ADN de retrouver la même conformation que l'ADN libre (données non présentées).

(b) Modèle ZDA

Comme pour le modèle précédent, le modèle ZDA ne permet pas de maintenir une interaction entre l'ADN et la protéine (données non présentées).

(c) Modèle Z*

L'approximation de ce modèle affecte légèrement la structure de l'ADN le long du site de fixation. La Figure 76 montre que l'insertion de la queue N-terminale des modèles chargés réduit la largeur du petit sillon de l'ADN par rapport au complexe tout atome. On observe une diminution de la largeur du petit sillon de 0,5 Å en moyenne entre les positions A5 et T7, et une augmentation de 1,5 Å au niveau de la position T9. La présence

de la protéine simplifiée avec des charges dans le grand sillon de l'ADN induit une légère diminution de sa largeur le long du site de fixation (environ 0,15 Å).

Les paramètres inter-paires de bases Twist, Rise et Roll semblent également peu affectés par le modèle Z*, à l'exception de la région riche en A/T en amont du site de reconnaissance (cf. Figure 77). Cette région est en interaction avec le bras N-terminal de la protéine SKN-1 qui est relativement dynamique et mobile dans la simulation tout atome. Nos protéines gros grains étant rigides, on oblige l'ADN à s'adapter à la position choisie pour la queue N-terminale de la protéine.

La présence de charges sur ce modèle, provoque un changement dans la distribution des ions le long du site de fixation par rapport à la simulation d'ADN libre. Au niveau du petit sillon, on observe une distribution des ions similaire entre le modèle atomique et la protéine gros grain Z* au niveau du site de reconnaissance à l'exception de la position G8pT9 pour laquelle on observe une diminution plus importante (cf. Figure 78). Dans le grand sillon de l'ADN, au niveau du site de reconnaissance aucun ion potassium n'est présent. La Figure 78 montre une diminution de la concentration ionique des ions K^+ d'un facteur 2 entre les positions C14 et C15 correspondant à la présence d'un ion 27 % du temps pour les MPMs Z* contre 78 % pour le complexe tout atome.

La séquence consensus de SKN-1 est RTCATC (où R = A/G). Le modèle Z* permet d'obtenir une spécificité de reconnaissance pour les bases G8, T12 et C13 (cf. Figure 79 ligne Z*, MD Consensus) similaire à ce qui est observé dans le modèle atomique (corrélation de 0,99, 0,99 et 1 respectivement). Cette reconnaissance est induite par une reconnaissance directe entre la protéine gros grains et les bases de l'ADN. La présence de charges favorise également une faible préférence pour les bases T ou C en position 9 (voir logos Interaction et MD Consensus du modèle Z* de la Figure 79). Les bases centrales C10pA11 ne sont spécifiquement sélectionnées par le modèle Z* (voir Figure 79, logos Interaction/MD Consensus). On observe même une sélectivité pour le nucléotide A sur ces deux positions.

(d) Modèle Z*DA*

L'ajout de donneurs-accepteurs de liaisons hydrogène dans ce modèle n'améliore pas la conformation liée de l'ADN (voir Figure 76 et Figure 77). On remarque cependant que la largeur du grand sillon au niveau du pas dinucléotidique C10pA11 augmente de 0,65 Å par rapport à ce qui est observé dans le modèle tout atome et le modèle Z*

(présence d'une asparagine (Asn511) capable d'effectuer une double liaison hydrogène avec les atomes C10(N4) et T11'(O4)).

La présence de charges formelles sur les résidus polaires non chargés ne modifie pas la distribution des ions le long de l'ADN observé dans le modèle Z*.

Comme pour le modèle Z*, on observe une préférence pour les bases G et TC en position 8, 12 et 13. De plus, l'ajout de charges partielles sur les résidus polaires permet de retrouver visuellement une sélectivité pour les bases C10 et A11 (corrélation de 0,89 et 0,90 par rapport au modèle atomique) et une intensité du signal plus importante (IC de 2,95 dans le modèle Z*DA* contre 1,15 dans le modèle Z*), qui n'étaient pas sélectionnées dans le modèle précédent. La présence de donneurs/accepteurs de liaisons hydrogène renforce la préférence pour la base T en position 9 (IC de 0,87 contre 0,32 dans le modèle Z*).

La présence de charges partielles sur les donneurs/accepteurs de liaisons hydrogène sur SKN-1 ne modifie pas la spécificité de reconnaissance pour la séquence riche en A/T (corrélation de Pearson de 0,54 par rapport au modèle tout atome et IC de 1,7 contre corrélation de 0,55 et IC de 1,7 pour le modèle Z*) qui se situe en amont du site de reconnaissance (cf. Figure 79) ce qui indique que cette reconnaissance est principalement due à la présence de résidus basiques.

En conclusion, SKN-1 est un bon exemple où la décomposition des interactions permet d'identifier les facteurs clés de la reconnaissance. L'analyse des MPMs révèle que la présence des charges sur la surface de SKN-1 est nécessaire pour maintenir l'intégrité du complexe, reproduire la distribution des ions le long de l'ADN et que la spécificité de reconnaissance est effectuée par les résidus acido-basiques pour les positions 8, 12 et 13 alors que ce sont les liaisons hydrogène à l'interface qui sont responsables de la reconnaissance des bases en position 9, 10 et 11.

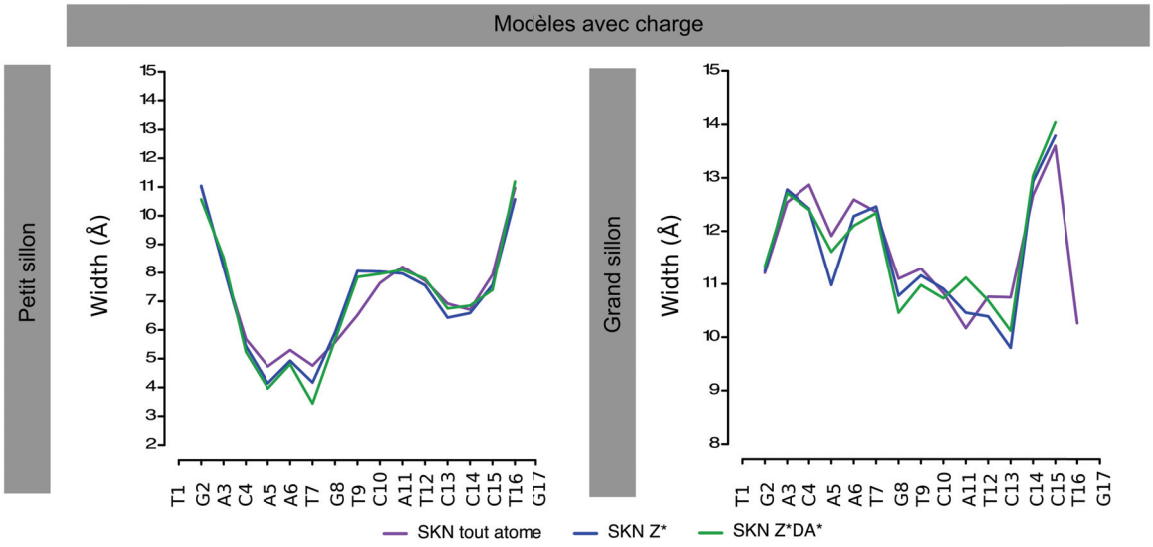


Figure 76 : Largeur du petit et du grand sillon de l'ADN pour les différents MPMs et pour le complexe SKN tout atome. La séquence est donnée dans le sens 5'-3' pour le brin Watson.

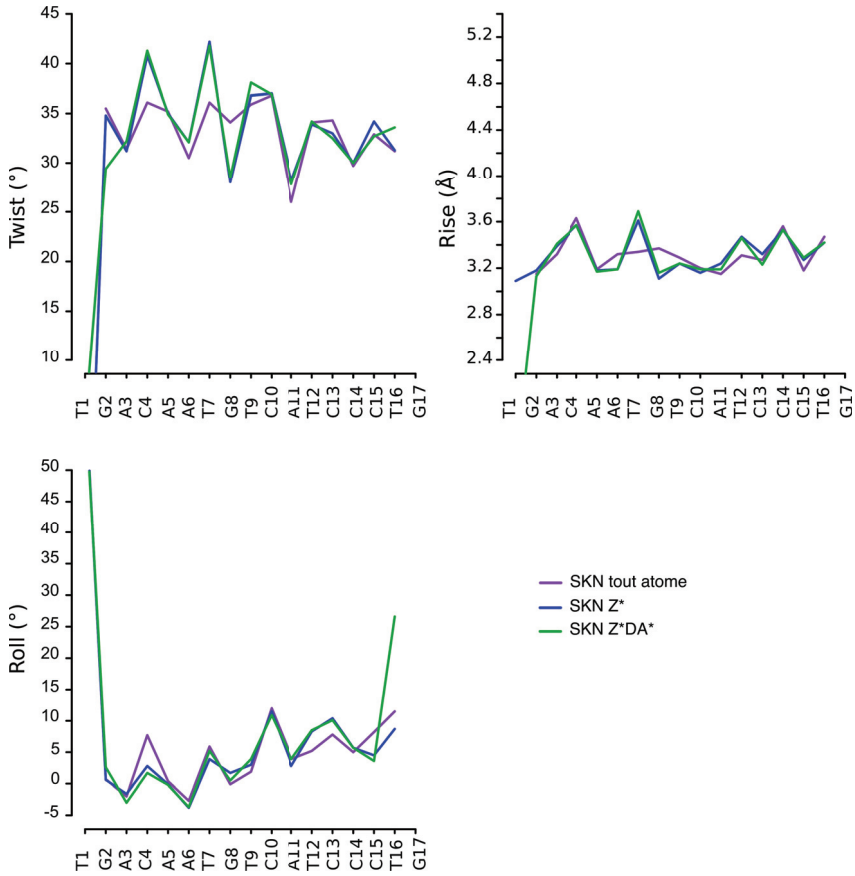


Figure 77 : Moyenne des paramètres hélicoïdaux Twist (°), Rise (Å) et Roll (°) le long du site de liaison pour les différents MPMs et pour le complexe SKN-1 tout atome. Les séquences sont indiquées dans le sens 5'-3'.

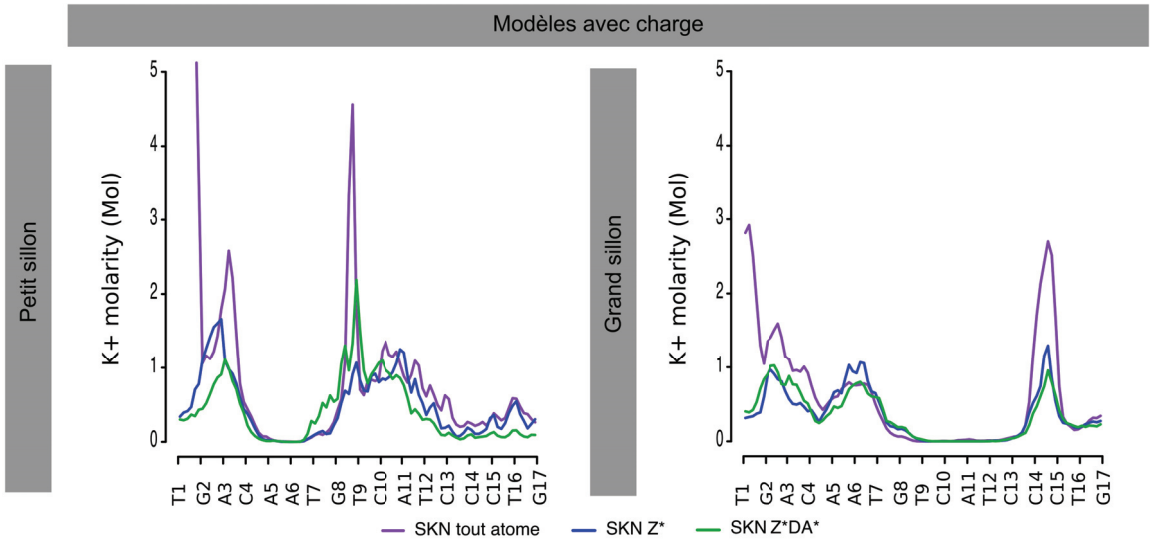


Figure 78 : Concentration en ions K^+ pour les différents modèles de SKN-1 le long des sillons de l'ADN.

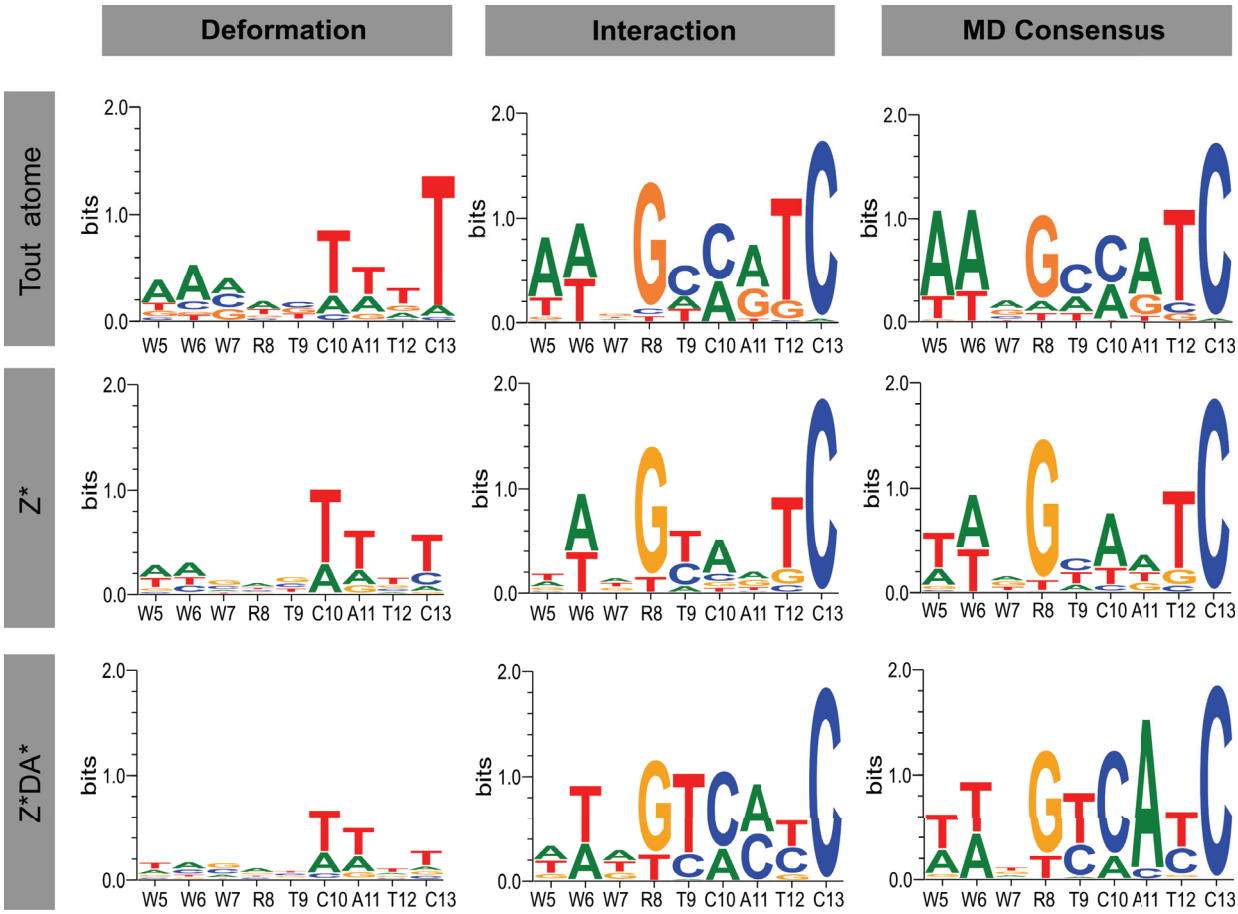


Figure 79 : Logos de spécificité de reconnaissance pour les différents modèles gros grains et pour le complexe SKN-1 tout atome. Pour chaque cas, la contribution de la déformation (reconnaissance indirecte), de l'interaction ADN-protéine (reconnaissance directe) et la contribution totale (MD Consensus) sont représentées.

D. Les MPMS de P22 c2

1. Rappel des propriétés structurales de l'ADN dans le complexe tout atome

Le répresseur P22 c2 se lie spécifiquement avec un ADN de 20 paires de bases dont la séquence consensus est **A**N**T**N**AAG**NNNN**CTT**N**A**N**T** (où N = A, T, C ou G). Chaque monomère se lie dans le grand sillon de l'ADN à un demi-site de reconnaissance de l'ADN grâce à une hélice α. Les monomères de P22 c2 n'effectuent qu'une seule liaison hydrogène avec les bases C8' ou C13'. La grande majorité des contacts sont effectués avec le brin de l'ADN, où l'on observe 8 ponts salins et 10 liaisons hydrogène (voir Figure 60B). Chaque monomère possède une valine (Val33) qui est enfouie dans une cavité formée par les quatre groupements méthyles au sein du grand sillon de la séquence 5'-TTAA-3'. La présence de la protéine induit une légère déformation de la molécule d'ADN (augmentation du twist, diminution de la largeur des deux sillons et une courbure de 23° par rapport à un ADN libre). On observe également une conformation B' de l'ADN au niveau des pas centraux qui séparent les demi-site de reconnaissance.

2. Influence du modèle sur la structure, la distribution des ions et la spécificité de reconnaissance

Comme pour les complexes précédents, les quatre modèles ZDA, ZDA, Z*et Z*DA* ont été utilisés pour simuler le répresseur P22 c2. Avec l'expérience acquise sur les complexes précédents et l'importance de la structure de la structure de certaines chaînes latérales pour reproduire des conformations de l'ADN, les chaînes latérales des Val33 ont été modélisées en tout atome plutôt qu'en gros-grains menant aux modèles Z+V*, ZDA+V*, Z*DA+V* et Z*DA*+V*.

(a) Modèle Z+V*, ZDA+V*

Lors de la dynamique moléculaire, les modèles simplifiés et sans charges Z+V* et ZDA+V* induisent la dissociation du complexe.

(b) Z*DA+V*

L'ajout de charges sur les résidus acido-basiques n'est également pas suffisant pour maintenir un complexe stable. Après une dizaine de nanosecondes de simulation, le complexe se dissocie.

(c) Z*DA*+V*

Dans la structure cristallographique, chaque monomère est stabilisé par la présence de trois ponts salins et deux liaisons hydrogène à l'extrémité 5' de chaque brin (cf. Figure 57B). Nous avons vu précédemment que la présence de charges nettes sur les résidus acido-basiques du modèle Z*DA+V* ne permet pas au répresseur de rester lié à l'ADN. L'interface formée entre les deux monomères et le tétranucléotide central 5'-ATAT-3' crée un canal ionique électronégatif. Au niveau de ce tétranucléotide, l'ADN adopte une conformation B' dont le petit sillon est très étroit. Dans le chapitre précédent, nous avons observé la présence de six liaisons hydrogène entre les monomères et les brins de l'ADN (trois par monomère) qui pourraient stabiliser l'interaction du complexe en diminuant la répulsion entre les groupements phosphates induite par l'étroitesse du petit sillon. Cette hypothèse se confirme dans ce modèle de Z*DA*+V* qui permet la formation de liaisons hydrogène. En effet, pour ce modèle, l'interaction ADN-protéine est maintenue tout au long de la simulation. La présence des donneurs-accepteurs de liaison Ser31-T14/T7' et Trp38-C13/C8' semble donc essentielle.

Cependant la présence de ces liaisons hydrogène n'est pas suffisante pour maintenir la conformation de l'ADN. En effet, la largeur des sillons et les paramètres hélicoïdaux sont grandement affectés. Comme le montre la Figure 80, la présence de la protéine gros grains dans le grand sillon élargit ce dernier d'environ 2 Å au niveau des positions 4 à 7 et 14 à 17 par rapport au complexe tout atome, correspondant à la poche dans laquelle sont enfouies les valines 33. Cette augmentation de largeur des sillons est liée à une diminution du Twist (cf. Figure 81). Au niveau des bases centrales 5'-ATAT-3' non contactées par le modèle gros grains de la protéine, on observe une diminution de la largeur du petit sillon qui atteint 1,3 Å de large au niveau de la position T11pA 12 dans la simulation Z*DA*+V* contre 3,3 Å pour la simulation tout atome. Ceci s'explique par une augmentation du Twist pour ces mêmes positions. Ces modifications de largeur de sillon pourraient potentiellement être corrigées, en intégrant des chaînes latérales atomiques pour les résidus Thr43 et Asn46. Ces résidus forment des liaisons hydrogène avec les

groupements phosphates des nucléotides C13/C8' et T9'/T12 (cf. Tableau 5 page 137) dans la simulation du modèle atomique.

Enfin l'on observe une modification de l'angle de Roll entre les paires A7pG8 et C13pT14. L'angle de Roll à ces positions correspond à celui observé dans la simulation d'ADN libre (environ 6,5°) alors que dans la simulation du complexe atomique, l'angle de Roll au niveau de ces positions est diminué de 5°, ce qui signifie que certaines chaînes latérales du modèle Z*DA*+V* ne contraignent pas l'ADN en position 8.

La présence de la protéine Z*DA*+V* reproduit la distribution des ions K^+ observés dans le complexe atomique (cf. Figure 82). La présence des ions au niveau des bases G8, C13 et des pas dinucléotidiques A9pT10 et A11pT12 a été discutée précédemment page 137.

L'analyse des composantes de la spécificité indique que la reconnaissance s'effectue principalement par l'interaction directe entre l'ADN et P22 c2 (voir Figure 83). Il est intéressant de noter que dans le modèle P22 Z*DA*+V*, on observe une différence de spécificité entre les monomères 1 et 2 comme dans la trajectoire tout atome. Le monomère 2 sélectionne spécifiquement le nucléotide T en position 14 et 15 alors que le monomère 1 ne reconnaît pas spécifiquement ces mêmes nucléotides en position 6 et 7 (cf. logos Interaction/MD Consensus Figure 83). Cette différence est due à un changement de conformation du résidu Gln37. Dans la structure qui a permis de modéliser le MPM Z*DA*+V* de P22 c2, la glutamine 37 du monomère 1 ne forme pas d'interaction avec le nucléotide C8' alors que celle du monomère 2 forme une liaison hydrogène avec le nucléotide C13. Bien que simplifié, notre modèle Z*DA*+V* est capable de reproduire la différence de sélectivité observée dans le complexe tout atome pour le fragment AAG et CTT selon la conformation adoptée par Gln37. Comme pour le complexe tout atome, la reconnaissance indirecte de la séquence riche en A/T pour la région séparant chaque demi-site de fixation qui est interprétée comme la transition de l'ADN B en ADN B' n'est pas retrouvé. La présence de cations monovalents serait un facteur électrostatique favorable à une sélectivité des paires A/T. Cependant la méthode ADAPT ne permet pas de traiter le solvant ou les ions explicitement dans le protocole d'enfilage moléculaire, ce qui peut expliquer la perte de spécificité de cette région pour les modèles gros grains ou atomiques.

En conclusion, notre modèle révèle que la présence de liaisons hydrogène dans le complexe ADN-P22 c2 est indispensable pour maintenir une interaction entre le répresseur et l'ADN. Malgré une conformation de l'ADN qui n'est pas parfaitement reproduite par le modèle Z*DA*+V*, la distribution des ions K^+ est bien reproduite par rapport au complexe tout atome et la spécificité de reconnaissance n'est pas affectée par ces changements de conformation de l'ADN, puisque la reconnaissance provient principalement de l'interaction directe entre les bases de l'ADN et la protéine gros grains. L'absence de traitement explicite du solvant et des ions dans la méthode ADAPT conduit probablement à la perte de reconnaissance de la région séparant les deux demi-sites de reconnaissance et s'applique aux modèles atomiques et gros grains. Bien que simplifié ce modèle est capable de définir une sélectivité de séquence liée à différente conformation de chaînes latérales.

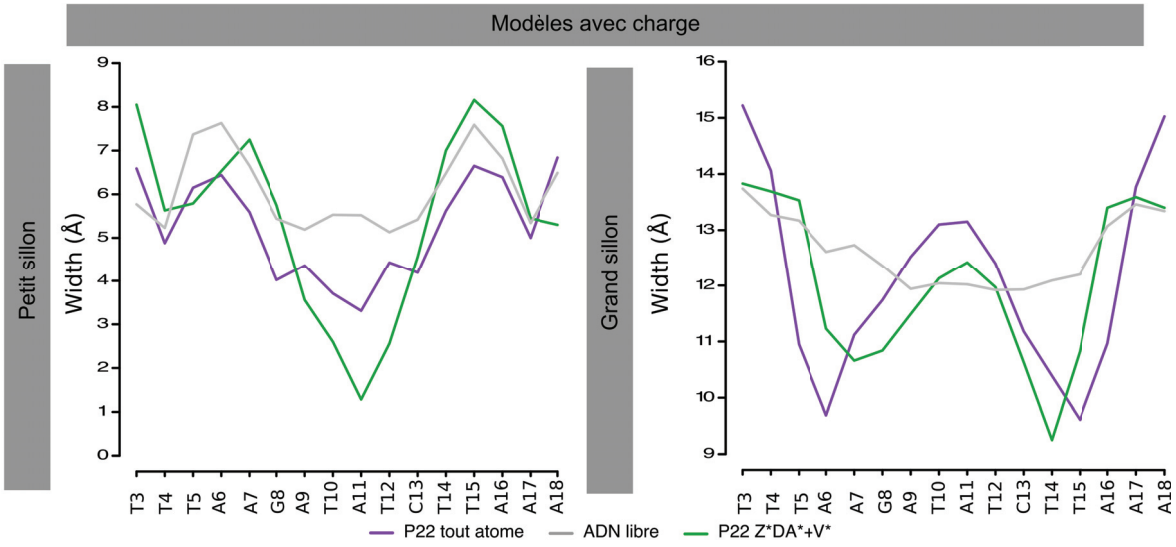


Figure 80 : Largeur du petit et du grand sillon de l'ADN pour les différents MPMs et pour le complexe ADN-P22 c2 tout atome. La séquence est donnée dans le sens 5'-3' pour le brin Watson.

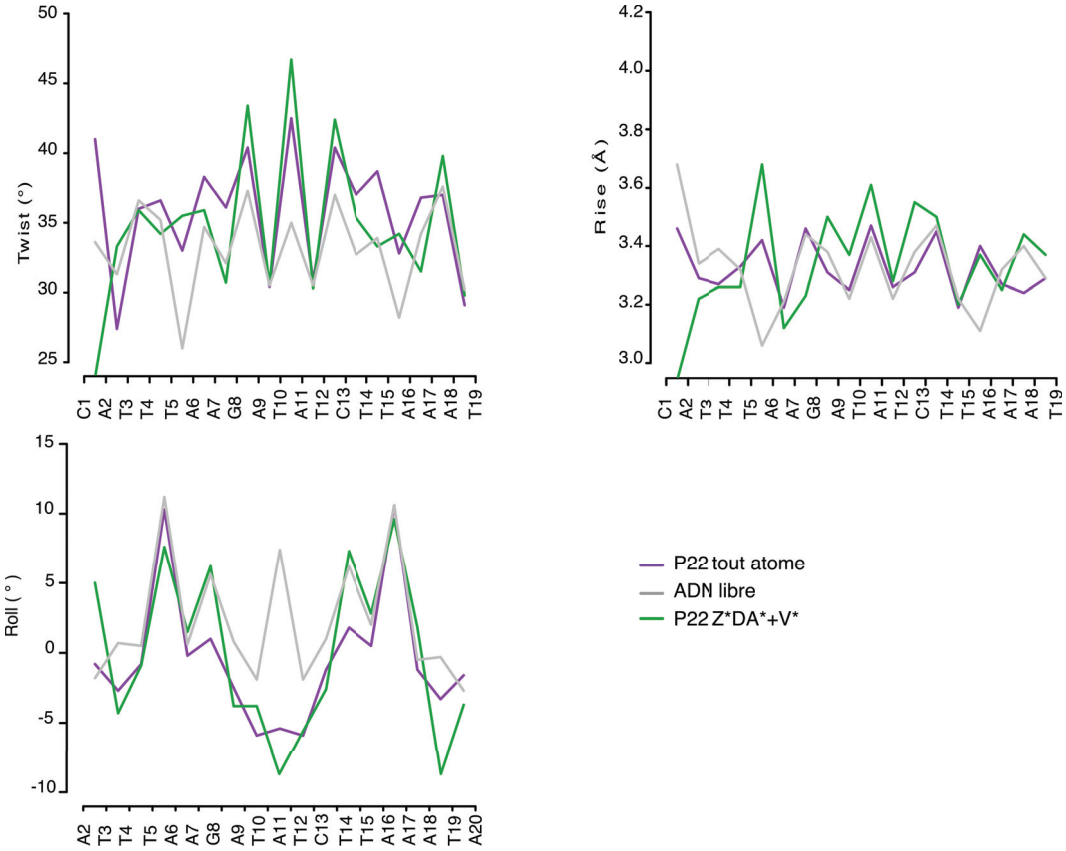


Figure 81 : Moyenne des paramètres hélicoïdaux Twist (°), Rise (Å) et Roll (°) le long du site de liaison pour les différents MPMs et pour le complexe ADN-P22 c2 tout atome. Les séquences sont indiquées dans le sens 5'-3'.

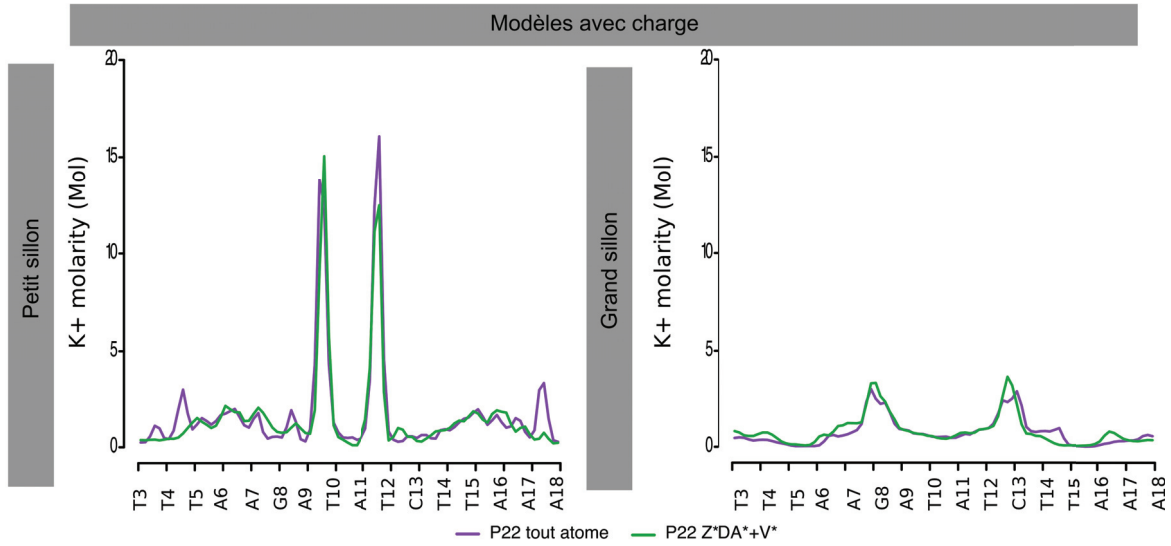


Figure 82 : Concentration en ions K^+ pour les différents modèles de P22 c2 le long des sillons de l'ADN.

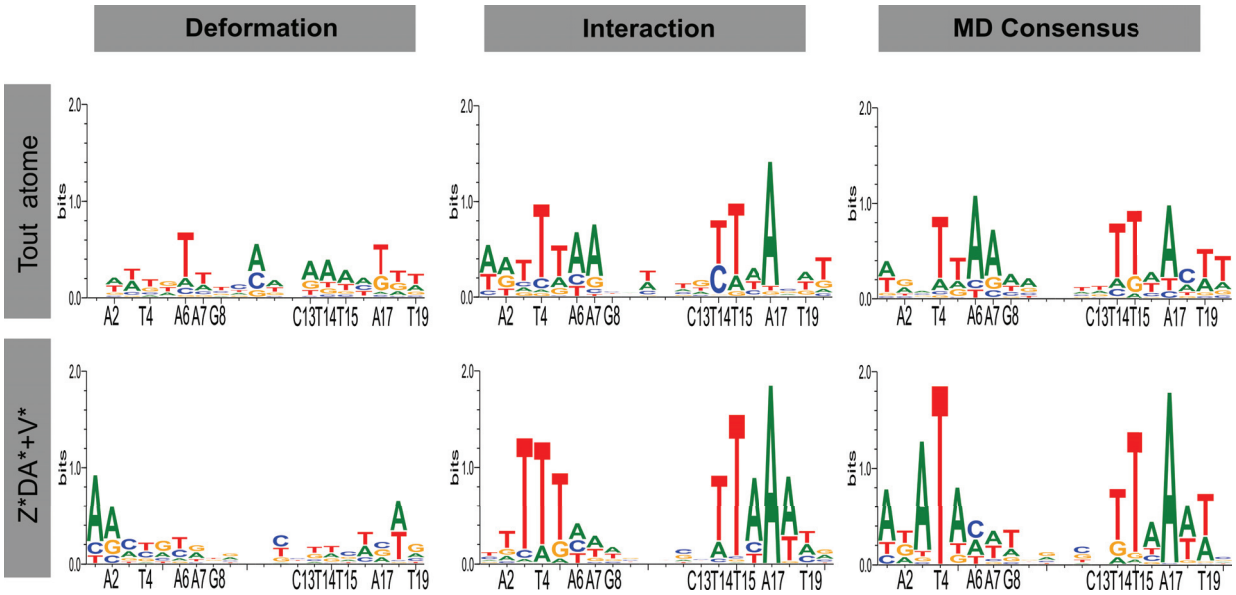


Figure 83 : Logos de spécificité de reconnaissance pour le modèle Z*DA* et pour le complexe P22 tout atome. Pour chaque cas, la contribution de la déformation (reconnaissance indirecte), de l'interaction ADN-protéine (reconnaissance directe) et la contribution totale (MD Consensus) sont représentées.

IV. Conclusion

Le modèle gros grains de protéine qui a été développé au cours de cette thèse est un modèle qui permet de représenter de manière simplifiée une protéine pour décomposer les interactions ADN-protéine et d'identifier : 1) les facteurs qui maintiennent la stabilité et la conformation de l'ADN ; 2) les facteurs responsables de la spécificité de reconnaissance ADN-protéine.

L'analyse des MPMs nous permet de mettre en évidence que parmi les quatre complexes étudiés, les protéines se liant dans le grand sillon de l'ADN (SKN-1 et P22 c2) nécessitent la présence de charges pour maintenir l'interaction. Dans le cas de P22 c2, cela va encore plus loin, puisqu'il est nécessaire de modéliser les liaisons hydrogène.

Dans le cas des protéines TBP et SRY, qui se lient dans le petit sillon, les modèles simplifiés et sans charges ne modifient pas la stabilité du complexe. Le volume de la protéine suffit à maintenir une interaction et une certaine spécificité d'interaction. Dans le complexe TBP, le volume de la protéine suffit à maintenir une reconnaissance spécifique au niveau de l'extrémité 3' du site de liaison. La déformation de l'ADN est le facteur dominant de la reconnaissance du complexe ADN-TBP. Notons également que la présence d'interactions électrostatique entre TBP et les brins phosphodiester de l'ADN dans la région 5' est nécessaire pour obtenir une déformation optimale de l'ADN permettant de reconnaître le motif TATA (lecture indirecte).

Pour le complexe SRY, la reconnaissance s'effectue partiellement par le volume de la protéine. La présence de charges améliore la reconnaissance des bases notamment dans la région 5' du site de liaison. En revanche, la présence de liaisons hydrogène a mis en évidence un mauvais positionnement de chaînes latérales dans la structure conduisant à une perte de sélectivité pour la base C7. Cette sélectivité a été retrouvée après insertion de chaînes latérales atomiques permettant leur bon repositionnement. L'absence de flexibilité de la queue C-terminale provoque un changement de conformation de l'ADN en amont du site de reconnaissance.

Dans le complexe SKN-1, la déconvolution des interactions nous permet d'affirmer que la reconnaissance du site s'effectue par les résidus chargés positivement pour les positions 8, 12 et 13 du site de liaison alors que la reconnaissance des positions 10 et 11 est effectuée par la présence de liaisons hydrogène assurées par le résidu Asn511.

Les protéines SRY et SKN-1 possèdent une queue terminale flexible que nous avons rendu rigide par la modélisation gros grains. Cependant, plusieurs études on démontrer que limiter la flexibilité des queues terminales réduit les changements de conformations accessibles par l'ADN, l'affinité de liaison et également la diffusion des protéine dans la recherche du bon site de liaison (387, 388). L'analyse des simulations gros grains de ces deux complexes confirme que les paramètres hélicoïdaux des bases en contacts avec ces queues terminales sont partiellement altérés par rapport aux complexes tout atome, mais que la reconnaissance spécifique est similaire malgré la perte de flexibilité.

Pour P22 c2, la reconnaissance s'effectue principalement par l'interaction directe entre la protéine et l'ADN. Le modèle MPM Z*DA*+V* permet dans cette étude de mettre en évidence la différence de sélectivité qui provient d'un changement de conformation de la chaîne latérale du résidu Gln37 malgré la simplification de représentation.

À travers ces exemples, nous avons démontré que la reconnaissance au sein des complexes ADN-protéine ne suit pas une loi universelle. Chaque protéine semble adopter un mécanisme de reconnaissance qui lui est propre et qui nécessite une analyse au cas par cas. La décomposition des interactions et des mécanismes de la reconnaissance peuvent être étudiés grâce à des modèles gros grains simples comme celui développé au cours de cette thèse, et qui vont permettent une analyse plus rapide et systématique des complexes ADN-protéine. Les protéines qui se lient dans le grand sillon ont un mécanisme d'interaction qui repose sur la présence de ponts salins et de liaisons hydrogène

correspondant au mécanisme décrit par Mirabekov et Rich (383), contrairement aux protéines qui se lient dans le petit sillon de l'ADN et pour lesquelles l'absence de charges provoque une courbure dans la direction opposé à la protéine comme suggéré par Elcock et McCammon (384).

Conclusions générales & perspectives

Depuis la résolution des premières structures ADN-protéine, les chercheurs ont essayé de déterminer un code universel qui permettrait de décrire les mécanismes de la reconnaissance spécifique du site de reconnaissance sans y parvenir. Jusqu'à présent, on ne possède qu'une information fragmentée des mécanismes de la reconnaissance provenant de structures principalement « statiques » venant des études par RMN ou par cristallographie aux rayons X. Le premier objectif de cette thèse était d'analyser les propriétés dynamiques des interfaces ADN-protéine à l'aide de dynamique moléculaire tout atome et son impact sur la reconnaissance ADN-protéine. Le second objectif était d'identifier et isoler les différents paramètres permettant de moduler la spécificité de reconnaissance à travers les déformations de l'ADN et les interactions à l'interface ADN-protéine grâce au développement d'un modèle gros grains.

Nous avons mis au point un protocole qui permet d'étudier la dynamique des complexes ADN-protéine et la spécificité de reconnaissance couplant dynamique et enfilage moléculaire. Nous avons découvert que les interactions entre l'ADN (brin phosphodiester ou bases azotées) et la protéine sont dynamiques. Elle se forment et se rompent régulièrement avec des temps de vie de l'ordre de la centaine de picosecondes mais peuvent dans certains cas perdurer plusieurs nanosecondes. À partir du protocole d'analyse mis en place, nous avons pu mettre en évidence que la dynamique de certaines interactions acides aminé-base peut moduler la spécificité de reconnaissance d'une ou plusieurs bases le long du site de reconnaissance.

Malgré que les ressources computationnelles permettent des simulations tout atome de plus en plus longues, la description de certains processus biologiques reste inaccessible car les échelles de temps simulées sont trop courtes. C'est pourquoi, plusieurs groupes ont développé des modèles gros grains. Ces modèles réduisent les temps de calculs et permettent d'effectuer des simulations pouvant atteindre plusieurs millisecondes dans des temps acceptables. Actuellement, les modèles gros grains qui reproduisent des comportements proches des modèles tout atomes (interactions, structures secondaires, flexibilité) nécessitent une paramétrisation fine et complexe. Nous avons donc développé un modèle gros grains simple, nécessitant peu de paramètres mais adaptable dans le but de mieux comprendre les mécanismes en jeu lors des interactions protéine-ADN. Ce modèle peut être complexifié de différents façons, notamment en ajoutant des charges et des chaînes latérales atomiques. La complexification progressive du modèle permet :1) d'identifier les facteurs responsables

de la liaison ADN-protéine ; 2) d'isoler les paramètres permettant de reproduire la forme liée de l'ADN et 3) de déterminer les facteurs responsables de la spécificité de reconnaissance et ce pour plusieurs familles de complexes.

Décomposer les interactions nous a permis de mettre en évidence un mécanisme d'interaction différent entre les protéines qui se lient dans le petit sillon, de celles se liant dans le grand sillon. La présence de ponts salins ou de liaisons hydrogène semble essentielle pour maintenir une interaction dans les complexes où la protéine se lie dans le grand sillon. La reconnaissance spécifique de ces protéines s'effectue principalement par la présence des résidus polaires chargés ou neutres le long du site de reconnaissance. Pour les protéines se liant dans le petit sillon que nous avons étudié, une interaction et une reconnaissance du site de liaison partielle est possible même en absence de charges. L'ajout de charges sur les protéines positionnées dans le petit sillon permet de compléter la spécificité pour certaines bases du site. Cette observation nécessite cependant une analyse plus systématique des complexes ADN-protéine qui permettrait de valider cette différence entre les deux types liaisons à l'ADN.

Dans cette thèse, nous avons mis en avant l'aspect dynamique des interfaces ADN-protéine. Il serait intéressant de transférer cette propriété aux modèles de protéine gros grains dans l'optique d'étudier les effets de différentes conformations des complexes sur la décomposition des mécanismes d'interaction et de reconnaissance. Deux approches peuvent être envisagées pour reproduire l'aspect dynamique des interfaces ADN-protéine avec nos modèles. La plus simple à développer consiste à utiliser des structures représentatives provenant des analyses de clustering. La seconde approche nécessite une refonte du modèle afin d'intégrer des paramètres physiques (angle de torsions, dièdres, etc.) qui permettront aux pseudo-atomes des acides aminés de se déplacer.

Dans le cas du répresseur P22 c2, nous avons constaté que certaines bases du site de reconnaissance qui ne sont pas en contact direct avec la protéine ne sont pas spécifiquement reconnues par la méthode ADAPT. L'absence de molécules d'eau et d'ions explicites dans notre approche actuelle ne nous donne qu'une vue fragmentée de tous les composants qui interviennent dans la reconnaissance ADN-protéine. L'ajout de solvant explicite reste actuellement un réel challenge méthodologique et computationnel qui devrait néanmoins conduire à une compréhension plus fine et complète de la reconnaissance indirecte.

ANNEXES

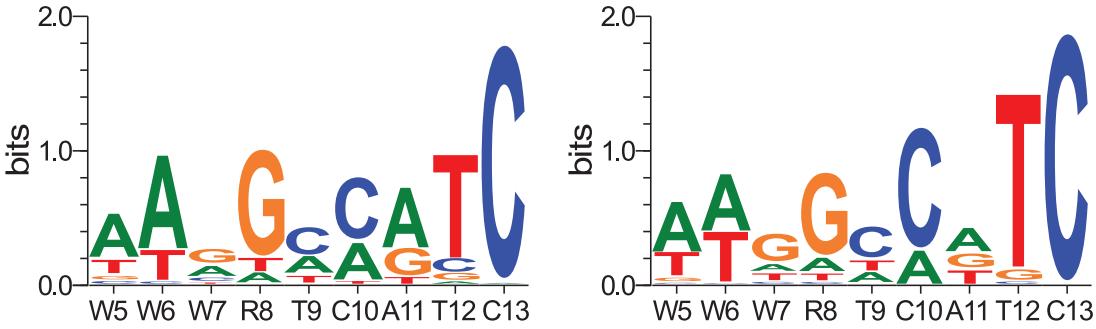


Figure S1 : Logo de séquence pour les groupes de conformations CL2 et CL4 pour le complexe ADN-SKN-1. W correspond aux nucléotides A/T alors que R correspond aux nucléotides A/G (nomenclature IUPAC).

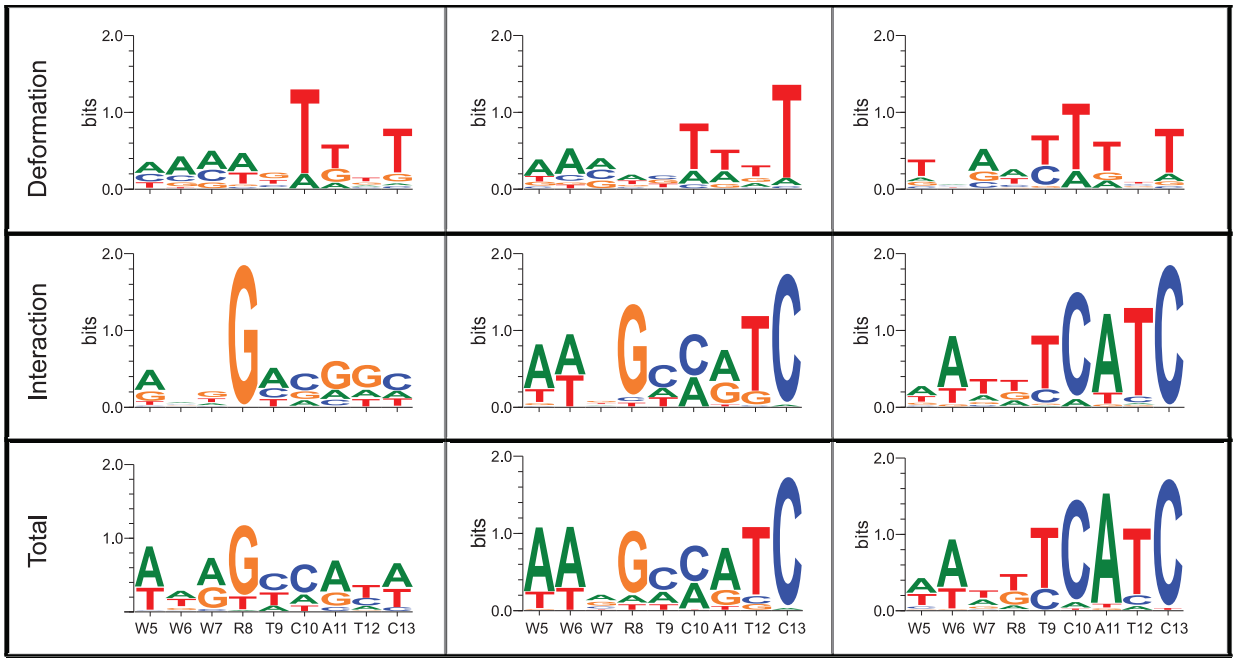


Figure S2 : Contributions de la déformation de l'ADN (panneau du haut), des interactions ADN-protéine (au milieu) et logo déformation+interaction (en bas) pour les groupes CL1, CL2/4 et CL3 (de gauche à droite) pour le complexe ADN-SKN-1.

Tableau S1 : Liste des structures PDB utilisées pour déterminer les rayons de Lennard-Jones des pseudo-atomes des modèles de protéines modulables.

PDB	Sillon de liaison	PDB	Sillon de liaison	PDB	Sillon de liaison
4gzn	Major	3ikt	Major	3mln	Major/Minor
1a1i	Major	3jxc	Major	2x6v	Major/Minor
1ubd	Major	3qws	Major	1h9d	Major/Minor
3vd6	Major/Minor	1k79	Major	3ted	Major
2i13	Major	2or1	Major	3v6t	Major
3uk3	Major	1pdn	Major/Minor	3pvv	Major/Minor
2wbs	Major	3ere	Major	4a04	Major/Minor
1ga5	Major/Minor	9ant	Major	1j1v	Major
1zme	Major/Minor	1e3o	Major	1h6f	Major/Minor
1hcq	Major	3osg	Major	3l2c	Major
1dsz	Major/Minor	4izz	Major	1nkp	Major
3g9m	Major	3zp5	Major	2c9l	Major
2han	Major	1cf7	Major	4ati	Major
2a66	Major	3vok	Major	2wt7	Major
2c7a	Major	1yo5	Major	1nlw	Major
4f6m	Major/Minor	3lnq	Major/Minor	1xbr	Major/Minor
2gli	Major	1pp7	Major/Minor	2xsd	Major
3coq	Major	3jtg	Major	1jj4	Major
1r0n	Major	1ais	Major/Minor	2o4a	Major
1ig7	Major	1qna	Minor	1b3t	Major/Minor
4aij	Major	3bs1	Major/Minor	3iag	Major/Minor
2d5v	Major	1cdw	Minor	2bop	Major
3jso	Major/Minor	3brd	Major/Minor	2hdd	Major
1bl0	Major	1ckt	Minor	4hqe	Major
3fdq	Major/Minor	1j3e	Major	1wet	Major/Minor
3g73	Major	3tq6	Minor	3o9x	Major/Minor
1puf	Major/Minor	3u2b	Minor	1lmb	Major
3s8q	Major	1mnn	Major/Minor	3cro	Major
4hf1	Major	1vtn	Major	1skn	Major
1bc8	Major	1jnm	Major	4ix7	Major/Minor
3sjm	Major	1gu4	Major	1qpi	Major
4i2o	Major	4h10	Major	1yrn	Major
1tro	Major	2dgc	Major		
1pue	Major	6pax	Major/Minor		

Tableau S2 : Paramètres pour les différents pseudo-atomes des MPMs Z*DA*.

Acide aminé	Pseudo-atome	Rayons VdW	ε	Charges
Ala	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
Arg	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
	PA3	1,820	0,6000	0,5000
	PA4	1,820	0,6000	0,5000
Asn	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,6730
	PA5	1,820	0,1700	-0,9191
	PA7	0,600	0,0157	0,4196
	PA7	0,600	0,0157	0,4196
	PA6	1,660	0,2100	-0,5931
Asp	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA3	1,660	0,6000	-0,5000
	PA4	1,660	0,6000	-0,5000
Cys	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
Gln	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,6991
	PA5	1,820	0,1700	-0,9407
	PA7	0,600	0,0157	0,4251
	PA7	0,600	0,0157	0,4251
	PA6	1,660	0,2100	-0,6086
Glu	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
	PA3	1,660	0,6000	-0,5000
	PA4	1,660	0,6000	-0,5000
Gly	Cα	2,280	0,6000	0,0000
His	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
	PA3	1,820	0,6000	0,0000
Ile	Cα	2,280	0,6000	0,0000
	PA1	2,900	0,6000	0,0000
Leu	Cα	2,280	0,6000	0,0000
	PA1	2,900	0,6000	0,0000

(Suite du tableau S2)

Acide aminé	Pseudo-atome	Rayons VdW	ε	Charges
Lys	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
	PA3	1,820	0,6000	1,0000
Met	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
Phe	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,170	0,6000	0,0000
Pro	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
Ser	Cα	2,280	0,6000	0,0000
	PA1	1,600	0,6000	0,2271
	PA5	1,721	0,2104	-0,6546
	PA7	0,000	0,0000	0,4275
Thr	Cα	2,280	0,6000	0,0000
	PA1	1,600	0,6000	0,2659
	PA5	1,721	0,2104	-0,6761
	PA7	0,000	0,0000	0,4102
Trp	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,900	0,6000	0,0000
Tyr	Cα	2,280	0,6000	0,0000
	PA1	2,320	0,6000	0,0000
	PA2	2,900	0,6000	0,1587
	PA5	1,721	0,2104	-0,5579
	PA7	0,000	0,0000	0,3992
Val	Cα	2,280	0,6000	0,0000
	PA1	2,900	0,6000	0,0000

Tableau S3 : Corrélation de Pearson par position le long du site de reconnaissance entre les logos des différents modèles MPMs et le logo du complexe atomique TBP.

	Séquence Consensus							
Modèle	T	A	T	A	W	A	A	R
Z+FAP*	0,99	0,18	0,80	0,29	0,43	0,93	0,98	0,99
ZDA+FAP*	0,98	0,16	0,97	0,93	0,66	0,90	0,99	0,99
Z*+FAP*	0,99	0,86	0,89	0,18	0,08	0,87	0,98	0,97
Z*DA*+FAP*	0,98	0,87	0,98	0,77	0,04	0,87	0,98	0,99

Tableau S4 : Corrélation de Pearson par position le long du site de reconnaissance entre les logos des différents modèles MPMs et le logo du complexe atomique SRY.

	Séquence Consensus									
Modèle		W	A	A	C	A	A	W		
Z	0,92	0,13	0,73	0,99	0,02	-0,13	0,99	0,93	0,72	0,97
ZDA	0,91	-0,4	-0,39	0,45	0,22	-0,05	0,12	0,94	0,89	0,87
Z*	0,94	-0,33	0,87	0,97	-0,99	0,82	0,84	0,87	0,84	0,37
Z*DA*	0,94	-0,25	0,49	1,00	-0,35	0,92	0,91	0,99	0,85	0,96
Z*DA*+RN*	0,93	-0,25	1,00	0,99	0,78	1,00	0,87	0,99	0,94	0,61

Tableau S5 : Corrélation de Pearson par position le long du site de reconnaissance entre les logos des différents modèles MPMs et le logo du complexe atomique SKN-1.

	Séquence Consensus								
Modèle	W	W	W	R	T	C	A	T	C
Z*	0,33	0,91	0,55	0,99	0,73	0,29	0,88	0,99	1,00
Z*DA*	0,54	0,79	-0,16	0,98	0,29	0,89	0,90	0,80	1,00

Tableau S6 : Corrélation de Pearson par position le long du site de reconnaissance entre le logo Z*DA*+V* et le logo du complexe atomique P22 c2. Pour plus de visibilité chaque demi site de reconnaissance est indiqué séparément.

	Séquence Consensus										
Modèle			A		T		A	A	G		
	Monomère 1	0,72	0,42	0,84	0,99	0,00	0,50	0,45	0,21	-0,75	0,65
	Séquence Consensus										
	Monomère 2		T		A		T	T	C		
Z*DA*+V*		-0,58	0,99	-0,29	0,99	0,74	0,98	0,91	-0,70	-0,47	0,52

Références Bibliographiques

1. Avery,O.T., MacLeod,C.M. and McCarty,M. (1944) STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *J. Exp. Med.*, **79**, 137–158.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
4. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
5. Sela,I. and Lukatsky,D.B. (2011) DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.*, **101**, 160–166.
6. Afek,A. and Lukatsky,D.B. (2013) Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophys. J.*, **105**, 1653–1660.
7. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
8. Chargaff,E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **6**, 201–209.
9. Hoogsteen,K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, **16**, 907–916.
10. Stofer,E., Chipot,C. and Lavery,R. (1999) Free Energy Calculations of Watson−Crick Base Pairing in Aqueous Solution. *J. Am. Chem. Soc.*, **121**, 9503–9508.
11. Hunter,C.A. (1993) Sequence-dependent DNA structure. The role of base stacking interactions. *J. Mol. Biol.*, **230**, 1025–1054.
12. Sponer,J., Gabb,H.A., Leszczynski,J. and Hobza,P. (1997) Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys. J.*, **73**, 76–87.
13. Bonner,G. and Klibanov,A.M. (2000) Structural stability of DNA in nonaqueous solvents. *Biotechnol. Bioeng.*, **68**, 339–344.
14. Hillen,W., Goodman,T.C. and Wells,R.D. (1981) Salt dependence and thermodynamic interpretation of the thermal denaturation of small DNA restriction fragments. *Nucleic Acids Res.*, **9**, 415–436.
15. Nikolova,E.N., Kim,E., Wise,A.A., O'Brien,P.J., Andricioaei,I. and Al-Hashimi,H.M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, **470**, 498–502.
16. Bommarito,S., Peyret,N. and Jr,J.S. (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.*, **28**, 1929–1934.
17. Petersheim,M. and Turner,D.H. (1983) Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry (Mosc.)*, **22**, 256–263.
18. Yakovchuk,P., Protozanova,E. and Frank-Kamenetskii,M.D. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids*

Res., **34**, 564–574.
19. Manning,G.S. (1977) Limiting laws and counterion condensation in polyelectrolyte solutions. IV. The approach to the limit and the extraordinary stability of the charge fraction. *Biophys. Chem.*, **7**, 95–102.
20. Manning,G.S. (1978) The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys.*, **11**, 179–246.
21. Conner,B.N., Yoon,C., Dickerson,J.L. and Dickerson,R.E. (1984) Helix geometry and hydration in an A-DNA tetramer: IC-C-G-G. *J. Mol. Biol.*, **174**, 663–695.
22. Beckers,M.L.M. and Buydens,L.M.C. (1998) Multivariate analysis of a data matrix containing A-DNA and B-DNA dinucleoside monophosphate steps: Multidimensional Ramachandran plots for nucleic acids. *J. Comput. Chem.*, **19**, 695–715.
23. Srinivasan,A.R. and Olson,W.K. (1987) Nucleic acid model building: the multiple backbone solutions associated with a given base morphology. *J. Biomol. Struct. Dyn.*, **4**, 895–938.
24. Castagné,C., Murphy,E.C., Gronenborn,A.M. and Delepierre,M. (2000) 31P NMR analysis of the DNA conformation induced by protein binding SRY/DNA complexes. *Eur. J. Biochem. FEBS*, **267**, 1223–1229.
25. Madhumalar,A. and Bansal,M. (2005) Sequence Preference for BI/BII Conformations in DNA: MD and Crystal Structure Data Analysis. *J. Biomol. Struct. Dyn.*, **23**, 13–27.
26. Derreumaux,S., Chaoui,M., Tevanian,G. and Fermandjian,S. (2001) Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res.*, **29**, 2314–2326.
27. Reddy,S.Y., Obika,S. and Bruice,T.C. (2003) Conformations and dynamics of Ets-1 ETS domain–DNA complexes. *Proc. Natl. Acad. Sci.*, **100**, 15475–15480.
28. 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.
29. Wellenzohn,B., Flader,W., Winger,R.H., Hallbrucker,A., Mayer,E. and Liedl,K.R. (2001) Complex of B-DNA with polyamides freezes DNA backbone flexibility. *J. Am. Chem. Soc.*, **123**, 5044–5049.
30. Westhof,E. and Sundaralingam,M. (1983) A method for the analysis of puckering disorder in five-membered rings: the relative mobilities of furanose and proline rings and their effects on polynucleotide and polypeptide backbone flexibility. *J. Am. Chem. Soc.*, **105**, 970–976.
31. 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.
32. Lu,X.-J. and Olson,W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.
33. Arnott,S. and Hukins,D.W.L. (1972) Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509.
34. Arnott,S. and Hukins,D.W.L. (1973) Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *J. Mol.*

Biol., **81**, 93–105.
35. Wang,A.H., Quigley,G.J., Kolpak,F.J., Crawford,J.L., van Boom,J.H., van der Marel,G. and Rich,A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
36. Leslie,A.G.W., Arnott,S., Chandrasekaran,R. and Ratliff,R.L. (1980) Polymorphism of DNA double helices. *J. Mol. Biol.*, **143**, 49–72.
37. Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
38. Lucas,A.A. (2008) A-DNA and B-DNA: Comparing Their Historical X-ray Fiber Diffraction Images. *J. Chem. Educ.*, **85**, 737.
39. Jiang,H., Zacharias,W. and Amirhaeri,S. (1991) Potassium permanganate as a in situ probe for B-Z and Z-Z junctions. *Nucleic Acids Res.*, **19**, 6943–6948.
40. Paleček,E., Rašovská,E. and Boublíková,P. (1988) Probing of DNA polymorphic structure in the cell with osmium tetroxide. *Biochem. Biophys. Res. Commun.*, **150**, 731–738.
41. Chen,Y.Z. and Prohofsky,E.W. (1993) Salt dependent premelting base pair opening probabilities of B and Z DNA Poly [d(G-C)] and significance for the B-Z transition. *Biophys. J.*, **64**, 1394–1397.
42. Adams,R.L.P. and Burdon,R.H. (1985) Molecular Biology of DNA Methylation Springer New York, New York, NY.
43. Ha,S.C., Kim,D., Hwang,H.-Y., Rich,A., Kim,Y.-G. and Kim,K.K. (2008) The crystal structure of the second Z-DNA binding domain of human DAI (ZBP1) in complex with Z-DNA reveals an unusual binding mode to Z-DNA. *Proc. Natl. Acad. Sci.*, **105**, 20671–20676.
44. A Rich, A Nordheim and Wang, and A.H.J. (1984) The Chemistry and Biology of Left-Handed Z-DNA. *Annu. Rev. Biochem.*, **53**, 791–846.
45. Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
46. Pauling,L., Corey,R.B. and Branson,H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **37**, 205–211.
47. Honig,B. (1999) Protein folding: from the levinthal paradox to structure prediction. *J. Mol. Biol.*, **293**, 283–293.
48. Anderson,D.E., Becktel,W.J. and Dahlquist,F.W. (1990) pH-Induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry (Mosc.)*, **29**, 2403–2408.
49. Bosshard,H.R., Marti,D.N. and Jelesarov,I. (2004) Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J. Mol. Recognit. JMR*, **17**, 1–16.
50. Phelan,P., Gorfe,A.A., Jelesarov,I., Marti,D.N., Warwicker,J. and Bosshard,H.R. (2002) Salt Bridges Destabilize a Leucine Zipper Designed for Maximized Ion Pairing between Helices. *Biochemistry (Mosc.)*, **41**, 2998–3008.
51. Marti,D.N. and Bosshard,H.R. (2003) Electrostatic interactions in leucine zippers: thermodynamic analysis of the contributions of Glu and His residues and the effect of

mutating salt bridges. *J. Mol. Biol.*, **330**, 621–637.
52. Czaplewski,C., Ołdziej,S., Liwo,A. and Scheraga,H.A. (2004) Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds. *Protein Eng. Des. Sel.*, **17**, 29–36.
53. 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.
54. Olins,D.E. and Olins,A.L. (2003) Chromatin history: our view from the bridge. *Nat. Rev. Mol. Cell Biol.*, **4**, 809–814.
55. Schoeffler,A.J. and Berger,J.M. (2005) Recent advances in understanding structure–function relationships in the type II topoisomerase mechanism. *Biochem. Soc. Trans.*, **33**, 1465–1470.
56. Champoux,J.J. (2001) DNA Topoisomerases: Structure, Function, and Mechanism. *Annu. Rev. Biochem.*, **70**, 369–413.
57. Wang,J.C. (2002) Cellular roles of DNA topoisomerases: a molecular perspective. *Nat. Rev. Mol. Cell Biol.*, **3**, 430–440.
58. Tuteja,N. and Tuteja,R. (2004) Unraveling DNA helicases. *Eur. J. Biochem.*, **271**, 1849–1863.
59. Joyce,C.M. and Steitz,T.A. (1995) Polymerase structures and function: variations on a theme? *J. Bacteriol.*, **177**, 6321–6329.
60. McClarin,J.A., Frederick,C.A., Wang,B.C., Greene,P., Boyer,H.W., Grable,J. and Rosenberg,J.M. (1986) Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. *Science*, **234**, 1526–1541.
61. Pingoud,A. and Jeltsch,A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res.*, **29**, 3705–3727.
62. Newman,M., Strzelecka,T., Dorner,L.F., Schildkraut,I. and Aggarwal,A.K. (1995) Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science*, **269**, 656–663.
63. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
64. de Boer,C.G. and Hughes,T.R. (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.*, **40**, D169–D179.
65. Kristiansson,E., Thorsen,M., Tamás,M.J. and Nerman,O. (2009) Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements. *Mol. Biol. Evol.*, **26**, 1299–1307.
66. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
67. Liu,Y. and Ringnér,M. (2007) Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol.*, **8**, R77.
68. García,R., Bermejo,C., Grau,C., Pérez,R., Rodríguez-Peña,J.M., Francois,J., Nombela,C. and Arroyo,J. (2004) The Global Transcriptional Response to Transient Cell Wall Damage in Saccharomyces cerevisiae and Its Regulation by the Cell Integrity Signaling Pathway. *J.*

Biol. Chem., **279**, 15183–15195.
69. Crossley,M. and Brownlee,G.G. (1990) Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature*, **345**, 444–446.
70. Martin,D.I., Tsai,S.F. and Orkin,S.H. (1989) Increased gamma-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature*, **338**, 435–438.
71. Faustino,P., Lavinha,J., Marini,M.G. and Moi,P. (1996) beta-Thalassemia mutation at -90C-->T impairs the interaction of the proximal CACCC box with both erythroid and nonerythroid factors. *Blood*, **88**, 3248–3249.
72. Lalioti,M.D., Scott,H.S., Buresi,C., Rossier,C., Bottani,A., Morris,M.A., Malafosse,A. and Antonarakis,S.E. (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, **386**, 847–851.
73. Wilson,A.G., Symons,J.A., McDowell,T.L., McDevitt,H.O. and Duff,G.W. (1997) Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 3195–3199.
74. Lewis,M., Chang,G., Horton,N.C., Kercher,M.A., Pace,H.C., Schumacher,M.A., Brennan,R.G. and Lu,P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, **271**, 1247–1254.
75. Schumacher,M.A., Choi,K.Y., Zalkin,H. and Brennan,R.G. (1994) Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science*, **266**, 763–770.
76. Harrison,S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature*, **353**, 715–719.
77. Malcolm James,D. (1995) DNA-protein : structural interactions. *Trove*.
78. Harrison,S.C. and Aggarwal,A.K. (1990) DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.*, **59**, 933–969.
79. Brennan,R.G. and Matthews,B.W. (1989) The helix-turn-helix DNA binding motif. *J. Biol. Chem.*, **264**, 1903–1906.
80. Otwinowski,Z., Schevitz,R.W., Zhang,R.-G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp represser/operator complex at atomic resolution. *Nature*, **335**, 321–329.
81. Suzuki,M., Yagi,N. and Gerstein,M. (1995) DNA recognition and superstructure formation by helix-turn-helix proteins. *Protein Eng.*, **8**, 329–338.
82. Gajiwala,K.S. and Burley,S.K. (2000) Winged helix proteins. *Curr. Opin. Struct. Biol.*, **10**, 110–116.
83. Clark,K.L., Halay,E.D., Lai,E. and Burley,S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, **364**, 412–420.
84. Kodandapani,R., Pio,F., Ni,C.Z., Piccialli,G., Klemsz,M., McKercher,S., Maki,R.A. and Ely,K.R. (1996) A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. *Nature*, **380**, 456–460.
85. Ellenberger,T.E., Brandl,C.J., Struhl,K. and Harrison,S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.

86. Murre,C., Bain,G., van Dijk,M.A., Engel,I., Furnari,B.A., Massari,M.E., Matthews,J.R., Quong,M.W., Rivera,R.R. and Stuiver,M.H. (1994) Structure and function of helix-loop-helix proteins. *Biochim. Biophys. Acta*, **1218**, 129–135.
87. Kim,Y., Geiger,J.H., Hahn,S. and Sigler,P.B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**, 512–520.
88. Kim,J.L., Nikolov,D.B. and Burley,S.K. (1993) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**, 520–527.
89. Kim,S.H. (1992) Beta ribbon: a new DNA recognition motif. *Science*, **255**, 1217–1218.
90. Phillips,S.E. (1994) The beta-ribbon DNA recognition motif. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 671–701.
91. Cho,Y., Gorina,S., Jeffrey,P.D. and Pavletich,N.P. (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science*, **265**, 346–355.
92. Ghosh,G., van Duyne,G., Ghosh,S. and Sigler,P.B. (1995) Structure of NF-kappa B p50 homodimer bound to a kappa B site. *Nature*, **373**, 303–310.
93. Hanas,J.S., Hazuda,D.J., Bogenhagen,D.F., Wu,F.Y. and Wu,C.W. (1983) Xenopus transcription factor A requires zinc for binding to the 5 S RNA gene. *J. Biol. Chem.*, **258**, 14120–14125.
94. Berg,J.M. (1990) Zinc fingers and other metal-binding domains. Elements for interactions between macromolecules. *J. Biol. Chem.*, **265**, 6513–6516.
95. Marmorstein,R., Carey,M., Ptashne,M. and Harrison,S.C. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.
96. Schreiter,E.R. and Drennan,C.L. (2007) Ribbon–helix–helix transcription factors: variations on a theme. *Nat. Rev. Microbiol.*, **5**, 710–720.
97. Rhee,S., Martin,R.G., Rosner,J.L. and Davies,D.R. (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 10413–10418.
98. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
99. Afek,A., Schipper,J.L., Horton,J., Gordân,R. and Lukatsky,D.B. (2014) Protein–DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci.*, **111**, 17140–17145.
100. Mordelet,F., Horton,J., Hartemink,A.J., Engelhardt,B.E. and Gordân,R. (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinforma. Oxf. Engl.*, **29**, i117–125.
101. Djordjevic,M. (2007) SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomol. Eng.*, **24**, 179–189.
102. Maerkl,S.J. and Quake,S.R. (2009) Experimental determination of the evolvability of a transcription factor. *Proc. Natl. Acad. Sci.*, **106**, 18650–18655.
103. Geertz,M. and Maerkl,S.J. (2010) Experimental strategies for studying transcription factor–DNA binding specificities. *Brief. Funct. Genomics*, **9**, 362–373.
104. Fried,M. and Crothers,D.M. (1981) Equilibria and kinetics of lac repressor-operator

interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.*, **9**, 6505–6525.
105. Garner,M.M. and Revzin,A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, **9**, 3047–3060.
106. Horak,C.E. and Snyder,M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.*, **350**, 469–483.
107. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
108. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
109. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.
110. Warren,C.L., Kratochvil,N.C.S., Hauschild,K.E., Foister,S., Brezinski,M.L., Dervan,P.B., Phillips,G.N. and Ansari,A.Z. (2006) Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 867–872.
111. Bulyk,M.L. (2007) Protein binding microarrays for the characterization of DNA-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, **104**, 65–85.
112. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
113. Andrilenas,K.K., Penvose,A. and Siggers,T. (2014) Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes. *Brief. Funct. Genomics*, 10.1093/bfgp/elu046.
114. Gilmour,D.S. and Lis,J.T. (1985) In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Mol. Cell. Biol.*, **5**, 2009–2018.
115. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E., *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
116. Hanlon,S.E. and Lieb,J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.*, **14**, 697–705.
117. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
118. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E., *et al.* (2010) Variation in Transcription Factor Binding Among Humans. *Science*, **328**, 232–235.
119. Liu,X., Noll,D.M., Lieb,J.D. and Clarke,N.D. (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.*, **15**, 421–427.
120. Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
121. Brenowitz,M., Senear,D.F. and Kingston,R.E. (2001) DNase I footprint analysis of

protein-DNA binding. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al*, **Chapter 12**, Unit 12.4.
122. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
123. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell*, **152**, 327–339.
124. Ogawa,N. and Biggin,M.D. (2012) High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Methods Mol. Biol. Clifton NJ*, **786**, 51–63.
125. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
126. Angarica,V.E., Pérez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.
127. Mandel-Gutfreund,Y., Baron,A. and Margalit,H. (2001) A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*
128. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
129. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci. Publ. Protein Soc.*, **11**, 2714–2726.
130. Zhang,C., Liu,S., Zhu,Q. and Zhou,Y. (2005) A Knowledge-Based Energy Function for Protein–Ligand, Protein–Protein, and Protein–DNA Complexes. *J. Med. Chem.*, **48**, 2325–2335.
131. Liu,Z., Mao,F., Guo,J., Yan,B., Wang,P., Qu,Y. and Xu,Y. (2005) Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Res.*, **33**, 546–558.
132. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein–DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
133. Kussell,E., Shimada,J. and Shakhnovich,E.I. (2002) A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci.*, **99**, 5343–5348.
134. Chen,W.W. and Shakhnovich,E.I. (2005) Lessons from the design of a novel atomic potential for protein folding. *Protein Sci. Publ. Protein Soc.*, **14**, 1741–1752.
135. Hubner,I.A., Deeds,E.J. and Shakhnovich,E.I. (2005) High-resolution protein folding with a transferable potential. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 18914–18919.
136. AlQuraishi,M. and McAdams,H.H. (2011) Direct inference of protein–DNA interactions using compressed sensing methods. *Proc. Natl. Acad. Sci.*, **108**, 14819–14824.
137. Alibés,A., Nadra,A.D., De Masi,F., Bulyk,M.L., Serrano,L. and Stricher,F. (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein–DNA interactions: the Pax6 example. *Nucleic Acids Res.*, **38**, 7422–7431.

138. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
139. Rahi,S.J., Virnau,P., Mirny,L.A. and Kardar,M. (2008) Predicting transcription factor specificity with all-atom models. *Nucleic Acids Res.*, **36**, 6209–6217.
140. Moroni,E., Caselle,M. and Fogolari,F. (2007) Identification of DNA-binding protein target sequences by physical effective energy functions: free energy analysis of lambda repressor-DNA complexes. *BMC Struct. Biol.*, **7**, 61.
141. Liu,L.A. and Bader,J.S. (2007) Ab initio prediction of transcription factor binding sites. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*
142. Beierlein,F.R., Kneale,G.G. and Clark,T. (2011) Predicting the Effects of Basepair Mutations in DNA-Protein Complexes by Thermodynamic Integration. *Biophys. J.*, **101**, 1130–1138.
143. Seeliger,D., Buelens,F.P., Goette,M., de Groot,B.L. and Grubmüller,H. (2011) Towards computational specificity screening of DNA-binding proteins. *Nucleic Acids Res.*, **39**, 8281–8290.
144. Boresch,S. and Karplus,M. (1995) The meaning of component analysis: decomposition of the free energy in terms of specific interactions. *J. Mol. Biol.*, **254**, 801–807.
145. Temiz,N.A. and Camacho,C.J. (2009) Experimentally based contact energies decode interactions responsible for protein–DNA affinity and the role of molecular waters at the binding interface. *Nucleic Acids Res.*, **37**, 4076–4088.
146. 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.*, **95**, 11163–11168.
147. Angarica,V.E., Pérez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.
148. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
149. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
150. Deremble,C., Lavery,R. and Zakrzewska,K. (2008) Protein–DNA recognition: Breaking the combinatorial barrier. *Comput. Phys. Commun.*, **179**, 112–119.
151. Paillard,G. and Lavery,R. (2004) Analyzing Protein-DNA Recognition Mechanisms. *Structure*, **12**, 113–122.
152. Paillard,G., Deremble,C. and Lavery,R. (2004) Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.*, **32**, 6673–6682.
153. Zakrzewska,K., Bouvier,B., Michon,A., Blanchet,C. and Lavery,R. (2009) Protein-DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. *Phys. Chem. Chem. Phys. PCCP*, **11**, 10712–10721.
154. Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.*, **344**, 59–70.

155. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
156. Endres,R.G., Schulthess,T.C. and Wingreen,N.S. (2004) Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, **57**, 262–268.
157. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
158. Berg,O.G. and Hippel,P.H. von (1985) Diffusion-Controlled Macromolecular Interactions. *Annu. Rev. Biophys. Biophys. Chem.*, **14**, 131–158.
159. Nobbs,T.J., Szczelkun,M.D., Wentzell,L.M. and Halford,S.E. (1998) DNA excision by the SfiI restriction endonuclease1. *J. Mol. Biol.*, **281**, 419–432.
160. Gottlieb,P.A., Wu,S., Zhang,X., Tecklenburg,M., Kuempel,P. and Hill,T.M. (1992) Equilibrium, kinetic, and footprinting studies of the Tus-Ter protein-DNA interaction. *J. Biol. Chem.*, **267**, 7434–7443.
161. Hoopes,B.C., LeBlanc,J.F. and Hawley,D.K. (1992) Kinetic analysis of yeast TFIID-TATA box complex formation suggests a multi-step pathway. *J. Biol. Chem.*, **267**, 11539–11547.
162. Wallis,R., Leung,K.Y., Pommer,A.J., Videler,H., Moore,G.R., James,R. and Kleanthous,C. (1995) Protein-protein interactions in colicin E9 DNase-immunity protein complexes. 2. Cognate and noncognate interactions that span the millimolar to femtomolar affinity range. *Biochemistry (Mosc.)*, **34**, 13751–13759.
163. Dhavan,G.M., Crothers,D.M., Chance,M.R. and Brenowitz,M. (2002) Concerted binding and bending of DNA by Escherichia coli integration host factor. *J. Mol. Biol.*, **315**, 1027–1037.
164. Bagchi,B., Blainey,P.C. and Xie,X.S. (2008) Diffusion Constant of a Nonspecifically Bound Protein Undergoing Curvilinear Motion along DNA. *J. Phys. Chem. B*, **112**, 6282–6284.
165. Blainey,P.C., Luo,G., Kou,S.C., Mangel,W.F., Verdine,G.L., Bagchi,B. and Xie,X.S. (2009) Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.*, **16**, 1224–1229.
166. Riggs,A.D., Bourgeois,S. and Cohn,M. (1970) The lac repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.*, **53**, 401–417.
167. Mirny,L., Slutsky,M., Wunderlich,Z., Tafvizi,A., Leith,J. and Kosmrlj,A. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. Math. Theor.*, **42**, 434013.
168. Tafvizi,A., Mirny,L.A. and van Oijen,A.M. (2011) Dancing on DNA: kinetic aspects of search processes on DNA. *Chemphyschem Eur. J. Chem. Phys. Phys. Chem.*, **12**, 1481–1489.
169. Elf,J., Li,G.-W. and Xie,X.S. (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, **316**, 1191–1194.
170. Hammar,P., Leroy,P., Mahmutovic,A., Marklund,E.G., Berg,O.G. and Elf,J. (2012) The lac repressor displays facilitated diffusion in living cells. *Science*, **336**, 1595–1598.
171. Gorman,J., Chowdhury,A., Surtees,J.A., Shimada,J., Reichman,D.R., Alani,E. and Greene,E.C. (2007) Dynamic Basis for One-Dimensional DNA Scanning by the Mismatch

Repair Complex Msh2-Msh6. *Mol. Cell*, **28**, 359–370.
172. Tafvizi,A., Huang,F., Leith,J.S., Fersht,A.R., Mirny,L.A. and van Oijen,A.M. (2008) Tumor Suppressor p53 Slides on DNA with Low Friction and High Stability. *Biophys. J.*, **95**, L01–L03.
173. Ando,T. and Skolnick,J. (2014) Sliding of Proteins Non-specifically Bound to DNA: Brownian Dynamics Studies with Coarse-Grained Protein and DNA Models. *PLoS Comput Biol*, **10**, e1003990.
174. Givaty,O. and Levy,Y. (2009) Protein sliding along DNA: dynamics and structural characterization. *J. Mol. Biol.*, **385**, 1087–1097.
175. Furini,S., Domene,C. and Cavalcanti,S. (2010) Insights into the Sliding Movement of the Lac Repressor Nonspecifically Bound to DNA. *J. Phys. Chem. B*, **114**, 2238–2245.
176. Bhattacherjee,A. and Levy,Y. (2014) Search by proteins for their DNA target site: 1. The effect of DNA conformation on protein sliding. *Nucleic Acids Res.*, 10.1093/nar/gku932.
177. Halford,S.E. (2009) An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem. Soc. Trans.*, **37**, 343–348.
178. Bhattacherjee,A. and Levy,Y. (2014) Search by proteins for their DNA target site: 2. The effect of DNA conformation on the dynamics of multidomain proteins. *Nucleic Acids Res.*, **42**, 12415–12424.
179. Broek,B. van den, Lomholt,M.A., Kalisch,S.-M.J., Metzler,R. and Wuite,G.J.L. (2008) How DNA coiling enhances target localization by proteins. *Proc. Natl. Acad. Sci.*, **105**, 15738–15742.
180. Janin,J., Rodier,F., Chakrabarti,P. and Bahadur,R.P. (2007) Macromolecular recognition in the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **63**, 1–8.
181. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci.*, **73**, 804–808.
182. Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
183. Matthews,B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
184. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
185. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
186. Lejeune,D., Delsaux,N., Charloteaux,B., Thomas,A. and Brasseur,R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, **61**, 258–271.
187. Coulocheri,S.A., Pigis,D.G., Papavassiliou,K.A. and Papavassiliou,A.G. (2007) Hydrogen bonds in protein–DNA complexes: Where geometry meets plasticity. *Biochimie*, **89**, 1291–1303.
188. Weber,I.T. and Steitz,T.A. (1984) Model of specific complex between catabolite gene

activator protein and B-DNA suggested by electrostatic complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, **81**, 3973–3977.
189. König,P., Giraldo,R., Chapman,L. and Rhodes,D. (1996) The Crystal Structure of the DNA-Binding Domain of Yeast RAP1 in Complex with Telomeric DNA. *Cell*, **85**, 125–136.
190. Brownlie,P., Ceska,T.A., Lamers,M., Romier,C., Stier,G., Teo,H. and Suck,D. (1997) The crystal structure of an intact human Max–DNA complex: new insights into mechanisms of transcriptional control. *Structure*, **5**, 509–520.
191. Hong,M. and Marmorstein,R. (2008) Chapter 3. In *Chapter 3:Structural Basis for Sequence-specific DNA Recognition by Transcription Factors and their Complexes*.
192. Omichinski,J.G., Clore,G.M., Schaad,O., Felsenfeld,G., Trainor,C., Appella,E., Stahl,S.J. and Gronenborn,A.M. (1993) NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science*, **261**, 438–446.
193. Bates,D.L., Chen,Y., Kim,G., Guo,L. and Chen,L. (2008) Crystal Structures of Multiple GATA Zinc Fingers Bound to DNA Reveal New Insights into DNA Recognition and Self-Association by GATA. *J. Mol. Biol.*, **381**, 1292–1306.
194. Yamasaki,K., Kigawa,T., Watanabe,S., Inoue,M., Yamasaki,T., Seki,M., Shinozaki,K. and Yokoyama,S. (2012) Structural basis for sequence-specific DNA recognition by an Arabidopsis WRKY transcription factor. *J. Biol. Chem.*, **287**, 7683–7691.
195. Carole A. Bewley, Angela M. Gronenborn and Clore, and G.M. (1998) MINOR GROOVE-BINDING ARCHITECTURAL PROTEINS: Structure, Function, and DNA Recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 105–131.
196. Reddy,C.K., Das,A. and Jayaram,B. (2001) Do water molecules mediate protein-DNA recognition?1. *J. Mol. Biol.*, **314**, 619–632.
197. Tainer,J.A. and Cunningham,R.P. (1993) Molecular recognition in DNA-binding proteins and enzymes. *Curr. Opin. Biotechnol.*, **4**, 474–483.
198. Wilson,D.S., Guenther,B., Desplan,C. and Kuriyan,J. (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.
199. Chiu,T.K., Sohn,C., Dickerson,R.E. and Johnson,R.C. (2002) Testing water-mediated DNA recognition by the Hin recombinase. *EMBO J.*, **21**, 801–814.
200. Woda,J., Schneider,B., Patel,K., Mistry,K. and Berman,H.M. (1998) An analysis of the relationship between hydration and protein-DNA interactions. *Biophys. J.*, **75**, 2170–2177.
201. Harris,L.F., Sullivan,M.R. and Popken-Harris,P.D. (1999) Molecular dynamics simulation in solvent of the bacteriophage 434 cI repressor protein DNA binding domain amino acids (R1-69) in complex with its cognate operator (OR1) DNA sequence. *J. Biomol. Struct. Dyn.*, **17**, 1–17.
202. Marco,E., García-Nieto,R. and Gago,F. (2003) Assessment by molecular dynamics simulations of the structural determinants of DNA-binding specificity for transcription factor Sp1. *J. Mol. Biol.*, **328**, 9–32.
203. Sen,S. and Nilsson,L. (1999) Structure, interaction, dynamics and solvent effects on the DNA-EcoRI complex in aqueous solution from molecular dynamics simulation. *Biophys. J.*, **77**, 1782–1800.
204. Suenaga,A., Yatsu,C., Komeiji,Y., Uebayasi,M., Meguro,T. and Yamato,I. (2000)

Molecular dynamics simulation of trp-repressor/operator complex: analysis of hydrogen bond patterns of protein–DNA interaction. *J. Mol. Struct.*, **526**, 209–218.
205. Tsui,V., Radhakrishnan,I., Wright,P.E. and Case,D.A. (2000) NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex. *J. Mol. Biol.*, **302**, 1101–1117.
206. Watkins,D., Hsiao,C., Woods,K.K., Koudelka,G.B. and Williams,L.D. (2008) P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. *Biochemistry (Mosc.)*, **47**, 2325–2338.
207. Murphy,E.C., Zhurkin,V.B., Louis,J.M., Cornilescu,G. and Clore,G.M. (2001) Structural Basis for SRY-dependent 46-X,Y Sex Reversal: Modulation of DNA Bending by a Naturally Occurring Point Mutation. *J. Mol. Biol.*, **312**, 481–499.
208. Cysewski,P. (2008) A post-SCF complete basis set study on the recognition patterns of uracil and cytosine by aromatic and π–aromatic stacking interactions with amino acid residues. *Phys. Chem. Chem. Phys.*, **10**, 2636–2645.
209. Rutledge,L.R., Durst,H.F. and Wetmore,S.D. (2009) Evidence for Stabilization of DNA/RNA–Protein Complexes Arising from Nucleobase–Amino Acid Stacking and T-Shaped Interactions. *J. Chem. Theory Comput.*, **5**, 1400–1410.
210. Rutledge,L.R., Campbell-Verduyn,L.S. and Wetmore,S.D. (2007) Characterization of the stacking interactions between DNA or RNA nucleobases and the aromatic amino acids. *Chem. Phys. Lett.*, **444**, 167–175.
211. Cauët,E., Rooman,M., Wintjens,R., Liévin,J. and Biot,C. (2005) Histidine-Aromatic Interactions in Proteins and Protein-Ligand Complexes: Quantum Chemical Study of X-ray and Model Structures. *J. Chem. Theory Comput.*, **1**, 472–483.
212. Wilson,K.A., Kellie,J.L. and Wetmore,S.D. (2014) DNA–protein π-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res.*, 10.1093/nar/gku269.
213. Rutledge,L.R. and Wetmore,S.D. (2011) Modeling the Chemical Step Utilized by Human Alkyladenine DNA Glycosylase: A Concerted Mechanism Aids in Selectively Excising Damaged Purines. *J. Am. Chem. Soc.*, **133**, 16258–16269.
214. Przybylski,J.L. and Wetmore,S.D. (2011) A QM/QM investigation of the hUNG2 reaction surface: the untold tale of a catalytic residue. *Biochemistry (Mosc.)*, **50**, 4218–4227.
215. Bruner,S.D., Norman,D.P.G. and Verdine,G.L. (2000) Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature*, **403**, 859–866.
216. Drew,H.R. and Travers,A.A. (1985) Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities. *Nucleic Acids Res.*, **13**, 4445–4467.
217. Hagerman,P.J. (1988) Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 265–286.
218. Patel,D.J., Kozlowski,S.A., Ikuta,S., Itakura,K., Bhatt,R. and Hare,D.R. (1983) NMR Studies of DNA Conformation and Dynamics in Solution. *Cold Spring Harb. Symp. Quant. Biol.*, **47**, 197–206.
219. Leroy,J.L., Kochoyan,M., Huynh-Dinh,T. and Guéron,M. (1988) Characterization of

base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *J. Mol. Biol.*, **200**, 223–238.
220. Laundon,C.H. and Griffith,J.D. (1987) Cationic metals promote sequence-directed DNA bending. *Biochemistry (Mosc.)*, **26**, 3759–3762.
221. Diekmann,S. and Wang,J.C. (1985) On the sequence determinants and flexibility of the kinetoplast DNA fragment with abnormal gel electrophoretic mobilities. *J. Mol. Biol.*, **186**, 1–11.
222. Li,T., Jin,Y., Vershon,A.K. and Wolberger,C. (1998) Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.*, **26**, 5707–5718.
223. Strauss,J.K., Roberts,C., Nelson,M.G., Switzer,C. and Maher,L.J. (1996) DNA bending by hexamethylene-tethered ammonium ions. *Proc. Natl. Acad. Sci.*, **93**, 9515–9520.
224. Dlakic,M. and Harrington,R.E. (1995) Bending and torsional flexibility of G/C-rich sequences as determined by cyclization assays. *J. Biol. Chem.*, **270**, 29945–29952.
225. Brukner,I., Susic,S., Dlakic,M., Savic,A. and Pongor,S. (1994) Physiological concentration of magnesium ions induces a strong macroscopic curvature in GGGCCC-containing DNA. *J. Mol. Biol.*, **236**, 26–32.
226. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Struct. Lond. Engl. 1993*, **13**, 1499–1509.
227. Hegde,R.S. (2002) The papillomavirus E2 proteins: structure, function, and biology. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 343–360.
228. Hines,C.S., Meghoo,C., Shetty,S., Biburger,M., Brenowitz,M. and Hegde,R.S. (1998) DNA structure and flexibility in the sequence-specific binding of papillomavirus E2 proteins. *J. Mol. Biol.*, **276**, 809–818.
229. Kim,S.S., Tam,J.K., Wang,A.F. and Hegde,R.S. (2000) The structural basis of DNA target discrimination by papillomavirus E2 proteins. *J. Biol. Chem.*, **275**, 31245–31254.
230. Strauss,J.K., Prakash,T.P., Roberts,C., Switzer,C. and Maher,L.J. (1996) DNA bending by a phantom protein. *Chem. Biol.*, **3**, 671–678.
231. Maher III,L.J. (1998) Mechanisms of DNA bending. *Curr. Opin. Chem. Biol.*, **2**, 688–694.
232. Strauss,J.K. and Maher,L.J.,3rd (1994) DNA bending by asymmetric phosphate neutralization. *Science*, **266**, 1829–1834.
233. Gurlie,R. and Zakrzewska,K. (1998) DNA curvature and phosphate neutralization: an important aspect of specific protein binding. *J. Biomol. Struct. Dyn.*, **16**, 605–618.
234. Strauss-Soukup,J.K. and Maher,L.J. (1997) Role of Asymmetric Phosphate Neutralization in DNA Bending by PU.1. *J. Biol. Chem.*, **272**, 31570–31575.
235. Gurlie,R. and Zakrzewska,K. (2001) Protein-induced DNA bending: the role of phosphate neutralisation. *Theor. Chem. Acc.*, **106**, 83–90.
236. Nikolov,D.B. and Burley,S.K. (1994) 2.1 Å resolution refined structure of a TATA box-binding protein (TBP). *Nat. Struct. Mol. Biol.*, **1**, 621–637.
237. Lu,X.J., Shakked,Z. and Olson,W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
238. Lu,X.J., Shakked,Z. and Olson,W.K. (2000) A-form conformational motifs in ligand-

bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
239. Tolstorukov,M.Y., Jernigan,R.L. and Zhurkin,V.B. (2004) Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. *J. Mol. Biol.*, **337**, 65–76.
240. Locasale,J.W., Napoli,A.A., Chen,S., Berman,H.M. and Lawson,C.L. (2009) Signatures of protein-DNA recognition in free DNA binding sites. *J. Mol. Biol.*, **386**, 1054–1065.
241. Guzikevich-Guerstein,G. and Shakked,Z. (1996) A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nat. Struct. Biol.*, **3**, 32–37.
242. Elrod-Erickson,M., Benson,T.E. and Pabo,C.O. (1998) High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Struct. Lond. Engl. 1993*, **6**, 451–464.
243. Nekludova,L. and Pabo,C.O. (1994) Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 6948–6952.
244. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.
245. Choo,Y. and Klug,A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.
246. Barraud,P. and Allain,F.H.-T. (2012) ADAR Proteins: Double-stranded RNA and Z-DNA Binding Domains. *Curr. Top. Microbiol. Immunol.*, **353**, 35–60.
247. Schwartz,T., Behlke,J., Lowenhaupt,K., Heinemann,U. and Rich,A. (2001) Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat. Struct. Biol.*, **8**, 761–765.
248. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
249. Rocchia,W., Sridharan,S., Nicholls,A., Alexov,E., Chiabrera,A. and Honig,B. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.*, **23**, 128–137.
250. Li,L., Li,C., Sarkar,S., Zhang,J., Witham,S., Zhang,Z., Wang,L., Smith,N., Petukh,M. and Alexov,E. (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys.*, **5**, 9.
251. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, 10.1093/nar/gku855.
252. Beveridge,D.L., Barreiro,G., Byun,K.S., Case,D.A., Cheatham,T.E., Dixit,S.B., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.*, **87**, 3799–3813.
253. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E., Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H., *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**,

3721–3740.
254. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
255. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
256. Maehigashi,T., Hsiao,C., Woods,K.K., Moulaei,T., Hud,N.V. and Williams,L.D. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–3722.
257. Lankas,F., Sponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–2883.
258. Winger,R.H., Liedl,K.R., Rüdisser,S., Pichler,A., Hallbrucker,A. and Mayer,E. (1998) B-DNA's BI → BII Conformer Substate Dynamics Is Coupled with Water Migration. *J. Phys. Chem. B*, **102**, 8934–8940.
259. Pichler,A., Rüdisser,S., Winger,R.H., Liedl,K.R., Hallbrucker,A. and Mayer,E. (2000) The role of water in B-DNAs BI to BII conformer substates interconversion: a combined study by calorimetry, FT-IR spectroscopy and computer simulation. *Chem. Phys.*, **258**, 391–404.
260. Oguey,C., Foloppe,N. and Hartmann,B. (2010) Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions. *PLoS ONE*, **5**.
261. Pichler,A., Hallbrucker,A., Winger,R.H., Liedl,K.R. and Mayer,E. (2000) B-DNA's BII Conformer Substate Population Increases with Decreasing Water Activity. 2. A Fourier Transform Infrared Spectroscopic Study of Nonoriented d(CGCGAATTCGCG)2. *J. Phys. Chem. B*, **104**, 11354–11359.
262. Bandyopadhyay,D. and Bhattacharyya,D. (2000) Effect of neighboring bases on base-pair stacking orientation: a molecular dynamics study. *J. Biomol. Struct. Dyn.*, **18**, 29–43.
263. Pastor,N., MacKerell,A.D. and Weinstein,H. (1999) TIT for TAT: the properties of inosine and adenosine in TATA box DNA. *J. Biomol. Struct. Dyn.*, **16**, 787–810.
264. Derreumaux,S. and Fermandjian,S. (2000) Bending and adaptability to proteins of the cAMP DNA-responsive element: molecular dynamics contrasted with NMR. *Biophys. J.*, **79**, 656–669.
265. Norberg,J. and Nilsson,L. (2000) On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.*, **79**, 1537–1553.
266. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional Specificity of a Hox Protein Mediated by the Recognition of Minor Groove Structure. *Cell*, **131**, 530–543.
267. Rohs,R., West,S.M., Liu,P. and Honig,B. (2009) Nuance in the Double-Helix and its Role in Protein-DNA Recognition. *Curr. Opin. Struct. Biol.*, **19**, 171–177.
268. Mauro,S.A., Pawlowski,D. and Koudelka,G.B. (2003) The Role of the Minor Groove Substituents in Indirect Readout of DNA Sequence by 434 Repressor. *J. Biol. Chem.*, **278**, 12955–12960.
269. Mann,R.S., Lelli,K.M. and Joshi,R. (2009) Hox specificity unique roles for cofactors

and collaborators. *Curr. Top. Dev. Biol.*, **88**, 63–101.
270. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
271. Churchill,M.E. and Travers,A.A. (1991) Protein motifs that recognize structural features of DNA. *Trends Biochem. Sci.*, **16**, 92–97.
272. Shen,A., Higgins,D.E. and Panne,D. (2009) Recognition of AT-rich DNA binding sites by the MogR repressor. *Struct. Lond. Engl. 1993*, **17**, 769–777.
273. Kitayner,M., Rozenberg,H., Rohs,R., Suad,O., Rabinovich,D., Honig,B. and Shakked,Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
274. Gajiwala,K.S., Chen,H., Cornille,F., Roques,B.P., Reith,W., Mach,B. and Burley,S.K. (2000) Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature*, **403**, 916–921.
275. Lane,W.J. and Darst,S.A. (2006) The structural basis for promoter -35 element recognition by the group IV sigma factors. *PLoS Biol.*, **4**, e269.
276. Yang,W. and Steitz,T.A. (1995) Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. *Cell*, **82**, 193–207.
277. Fairall,L., Schwabe,J.W., Chapman,L., Finch,J.T. and Rhodes,D. (1993) The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature*, **366**, 483–487.
278. Horton,N.C., Dorner,L.F. and Perona,J.J. (2002) Sequence selectivity and degeneracy of a restriction endonuclease mediated by DNA intercalation. *Nat. Struct. Biol.*, **9**, 42–47.
279. Sierk,M.L., Zhao,Q. and Rastinejad,F. (2001) DNA deformability as a recognition feature in the reverb response element. *Biochemistry (Mosc.)*, **40**, 12833–12843.
280. Kalodimos,C.G., Biris,N., Bonvin,A.M.J.J., Levandoski,M.M., Guennuegues,M., Boelens,R. and Kaptein,R. (2004) Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science*, **305**, 386–389.
281. Reményi,A., Lins,K., Nissen,L.J., Reinbold,R., Schöler,H.R. and Wilmanns,M. (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.*, **17**, 2048–2059.
282. Palasingam,P., Jauch,R., Ng,C.K.L. and Kolatkar,P.R. (2009) The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *J. Mol. Biol.*, **388**, 619–630.
283. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res.*, **14**, 1188–1190.
284. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
285. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
286. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinforma. Oxf. Engl.*, **16**, 16–23.
287. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling Dependencies in protein-DNA Binding Sites. In *Proceedings of the Seventh Annual International Conference*

on Research in Computational Molecular Biology, RECOMB '03. ACM, New York, NY, USA, pp. 28–37.
288. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinforma. Oxf. Engl.*, **23**, 933–941.
289. Matys,V., Fricke,E., Geffers,R., Gössling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V., *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
290. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., Chou,A., Ienasescu,H., *et al.* (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 10.1093/nar/gkt997.
291. Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.
292. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X., *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
293. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.-B., Reynolds,D.B., Yoo,J., *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
294. Zhu,C., Byers,K.J.R.P., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M., *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
295. Mordelet,F., Horton,J., Hartemink,A.J., Engelhardt,B.E. and Gordân,R. (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinforma. Oxf. Engl.*, **29**, i117–125.
296. De Jong,A.T. (2013) Effect of Flanking Bases on the DNA Specificity of EmBP-1. *Biochemistry (Mosc.)*, **52**, 786–794.
297. Liberzon,A., Ridner,G. and Walker,M.D. (2004) Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. *Nucleic Acids Res.*, **32**, 54–64.
298. Stringham,J.L., Brown,A.S., Drewell,R.A. and Dresch,J.M. (2013) Flanking sequence context-dependent transcription factor binding in early Drosophila development. *BMC Bioinformatics*, **14**, 298.
299. CASE,D.A., CHEATHAM,T.E., DARDEN,T., GOHLKE,H., LUO,R., MERZ,K.M., ONUFRIEV,A., SIMMERLING,C., WANG,B. and WOODS,R.J. (2005) The Amber Biomolecular Simulation Programs. *J. Comput. Chem.*, **26**, 1668–1688.
300. Arfken,G.B. and Weber,H.J. (2005) Mathematical Methods for Physicists, 6th Edition 6th edition. Academic Press, Boston.
301. Hestenes,M.R. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear systems National Bureau of Standards.
302. Verlet,L. (1967) Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, **159**, 98–103.
303. Ryckaert,J.-P., Ciccotti,G. and Berendsen,H.J.. (1977) Numerical integration of the

cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
304. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
305. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
306. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
307. Mahoney,M.W. and Jorgensen,W.L. (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, **112**, 8910–8922.
308. Dans,P.D., Danilāne,L., Ivani,I., Dršata,T., Lankaš,F., Hospital,A., Walther,J., Pujagut,R.I., Battistini,F., Gelpí,J.L., *et al.* (2016) Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, 10.1093/nar/gkw264.
309. Dang,L.X. (1995) Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
310. Dick,D.A. (1978) The distribution of sodium, potassium and chloride in the nucleus and cytoplasm of Bufo bufo oocytes measured by electron microprobe analysis. *J. Physiol.*, **284**, 37–53.
311. Lang,F. (2007) Mechanisms and significance of cell volume regulation. *J. Am. Coll. Nutr.*, **26**, 613S–623S.
312. Mezei,M. and Beveridge,D.L. (1981) Monte Carlo studies of the structure of dilute aqueous sclutions of Li+, Na+, K+, F–, and Cl–. *J. Chem. Phys.*, **74**, 6902–6910.
313. Várnai,P. and Zakrzewska,K. (2004) DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.*, **32**, 4269–4280.
314. Cheng,Y., Korolev,N. and Nordenskiöld,L. (2006) Similarities and differences in interaction of K+ and Na+ with condensed ordered DNA. A molecular dynamics computer simulation study. *Nucleic Acids Res.*, **34**, 686–696.
315. Dixit,S.B., Mezei,M. and Beveridge,D.L. (2012) Studies of base pair sequence effects on DNA solvation based on all-atom molecular dynamics simulations. *J. Biosci.*, **37**, 399–421.
316. 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.
317. Darden,T., Perera,L., Li,L. and Pedersen,L. (1999) New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, **7**, R55–R60.
318. Toukmaji,A., Sagui,C., Board,J. and Darden,T. (2000) Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.*, **113**, 10913–10927.
319. Cheatham,T.E., Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct.*

Dyn., **16**, 845–862.
320. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
321. Case,D.A., Berryman,J., Betz,R.., Cerutti,D., Cheatham III,T., Darden,T., Duke,R. and Glese,T. (2015) AMBER 2015.
322. Bonvin,A.M.J.J., Sunnerhagen,M., Otting,G. and van Gunsteren,W.F. (1998) Water molecules in DNA recognition II: a molecular dynamics view of the structure and hydration of the trp operator1. *J. Mol. Biol.*, **282**, 859–873.
323. 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.
324. Blanchet,C., Pasi,M., Zakrzewska,K. and Lavery,R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.*, **39**, W68–W73.
325. McLachlan,A.D. (1979) Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.*, **128**, 49–79.
326. R Development Core Team (2009) {R: A language and environment for statistical computing} R Foundation for Statistical Computing, Vienna, Austria.
327. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
328. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, 10.1093/nar/gkv080.
329. Krause,E.F. (1987) Taxicab Geometry: Adventure in Non-Euclidean Geometry (Paperback) par Eugene F. Krause: Dover Publications Inc., United States 9780486252025 Paperback, New edition. - The Book Depository US.
330. Ward,J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*, **58**, 236–244.
331. L. Kaufman,P.J.R. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. *Wiley N. Y. ISBN 0-471-87876-6*, 10.2307/2532178.
332. Murtagh,F. and Legendre,P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.*, **31**, 274–295.
333. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (junction minimisation of nucleic acids). *Comput. Phys. Commun.*, **91**, 135–158.
334. Hingerty,B.E., Ritchie,R.H., Ferrell,T.L. and Turner,J.E. (1985) Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers*, **24**, 427–439.
335. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
336. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
337. DeLano,W.L. (2002) The PyMOL Molecular Graphics System (2002) DeLano

Scientific, Palo Alto, CA, USA. http://www.pymol.org. *ResearchGate*.
338. Sebastian,A. and Contreras-Moreira,B. (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinforma. Oxf. Engl.*, **30**, 258–265.
339. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K., *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–110.
340. Kalodimos,C.G., Boelens,R. and Kaptein,R. (2004) Toward an Integrated Model of Protein–DNA Recognition as Inferred from NMR Studies on the Lac Repressor System. *Chem. Rev.*, **104**, 3567–3586.
341. Iwahara,J., Zweckstetter,M. and Clore,G.M. (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proc. Natl. Acad. Sci.*, **103**, 15062–15067.
342. Iwahara,J. and Clore,G.M. (2006) Direct Observation of Enhanced Translocation of a Homeodomain between DNA Cognate Sites by NMR Exchange Spectroscopy. *J. Am. Chem. Soc.*, **128**, 404–405.
343. Clore,G.M., Tang,C. and Iwahara,J. (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr. Opin. Struct. Biol.*, **17**, 603–616.
344. Mirny,L., Slutsky,M., Wunderlich,Z., Tafvizi,A., Leith,J. and Kosmrlj,A. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. Math. Theor.*, **42**, 434013.
345. Kolomeisky,A.B. (2011) Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys. Chem. Chem. Phys. PCCP*, **13**, 2088–2095.
346. Temiz,A.N., Benos,P.V. and Camacho,C.J. (2010) Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains. *Nucleic Acids Res.*, **38**, 2134–2144.
347. Bouvier,B., Zakrzewska,K. and Lavery,R. (2011) Protein–DNA Recognition Triggered by a DNA Conformational Switch. *Angew. Chem. Int. Ed.*, **50**, 6516–6518.
348. Chen,C. and Pettitt,B.M. (2011) The Binding Process of a Nonspecific Enzyme with DNA. *Biophys. J.*, **101**, 1139–1147.
349. Afek,A. and Lukatsky,D.B. (2013) Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. *Biophys. J.*, **104**, 1107–1115.
350. Garton,M. and Laughton,C. (2013) A comprehensive model for the recognition of human telomeres by TRF1. *J. Mol. Biol.*, **425**, 2910–2921.
351. Hahn,S., Buratowski,S., Sharp,P.A. and Guarente,L. (1989) Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **86**, 5718–5722.
352. Lee,T.I. and Young, and R.A. (2000) Transcription of Eukaryotic Protein-Coding Genes. *Annu. Rev. Genet.*, **34**, 77–137.
353. Burley,S.K. and Roeder,R.G. (1996) Biochemistry and structural biology of

transcription factor IID (TFIID). *Annu. Rev. Biochem.*, **65**, 769–799.
354. Hernandez,N. (1993) TBP, a universal eukaryotic transcription factor? *Genes Dev.*, **7**, 1291–1308.
355. Nakamura,K., Jeong,S.Y., Uchihara,T., Anno,M., Nagashima,K., Nagashima,T., Ikeda,S., Tsuji,S. and Kanazawa,I. (2001) SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.*, **10**, 1441–1448.
356. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–82.
357. Anderson,K.M., Esadze,A., Manoharan,M., Brüschweiler,R., Gorenstein,D.G. and Iwahara,J. (2013) Direct observation of the ion-pair dynamics at a protein-DNA interface by NMR spectroscopy. *J. Am. Chem. Soc.*, **135**, 3613–3619.
358. Chen,C., Esadze,A., Zandarashvili,L., Nguyen,D., Montgomery Pettitt,B. and Iwahara,J. (2015) Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein-DNA Complexes. *J. Phys. Chem. Lett.*, **6**, 2733–2737.
359. Kamachi,Y., Cheah,K.S. and Kondoh,H. (1999) Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9. *Mol. Cell. Biol.*, **19**, 107–120.
360. Harley,V.R., Lovell-Badge,R. and Goodfellow,P.N. (1994) Definition of a consensus DNA binding site for SRY. *Nucleic Acids Res.*, **22**, 1500–1501.
361. Bowerman,B., Eaton,B.A. and Priess,J.R. (1992) skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell*, **68**, 1061–1075.
362. Bowerman,B., Draper,B.W., Mello,C.C. and Priess,J.R. (1993) The maternal gene skn-1 encodes a protein that is distributed unequally in early C. elegans embryos. *Cell*, **74**, 443–452.
363. Ghose,P., Park,E.C., Tabakin,A., Salazar-Vasquez,N. and Rongo,C. (2013) Anoxia-Reoxygenation Regulates Mitochondrial Dynamics through the Hypoxia Response Pathway, SKN-1/Nrf, and Stomatin-Like Protein STL-1/SLP-2. *PLoS Genet.*, **9**.
364. Hoeven,R. van der, McCallum,K.C., Cruz,M.R. and Garsin,D.A. (2011) Ce-Duox1/BLI-3 generated reactive oxygen species trigger protective SKN-1 activity via p38 MAPK signaling during infection in C. elegans. *PLoS Pathog.*, **7**, e1002453.
365. Rupert,P.B., Daughdrill,G.W., Bowerman,B. and Matthews,B.W. (1998) A new DNA-binding motif in the Skn-1 binding domain-DNA complex. *Nat. Struct. Biol.*, **5**, 484–491.
366. Fjose,A., McGinnis,W.J. and Gehring,W.J. (1985) Isolation of a homoeo box-containing gene from the engrailed region of Drosophila and the spatial distribution of its transcripts. *Nature*, **313**, 284–289.
367. Scott,M.P., Weiner,A.J., Hazelrigg,T.I., Polisky,B.A., Pirrotta,V., Scalenghe,F. and Kaufman,T.C. (1983) The molecular organization of the Antennapedia locus of Drosophila. *Cell*, **35**, 763–776.
368. Bohmann,D., Bos,T.J., Admon,A., Nishimura,T., Vogt,P.K. and Tjian,R. (1987) Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP-1. *Science*, **238**, 1386–1392.

369. Hinnebusch,A.G. (1984) Evidence for translational regulation of the activator of general amino acid control in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, **81**, 6442–6446.
370. Oliveira,R.P., Abate,J.P., Dilks,K., Landis,J., Ashraf,J., Murphy,C.T. and Blackwell,T.K. (2009) Condition-adapted stress and longevity gene regulation by Caenorhabditis elegans SKN-1/Nrf. *Aging Cell*, **8**, 524–541.
371. Staab,T.A., Griffen,T.C., Corcoran,C., Evgrafov,O., Knowles,J.A. and Sieburth,D. (2013) The conserved SKN-1/Nrf2 stress response pathway regulates synaptic function in Caenorhabditis elegans. *PLoS Genet.*, **9**, e1003354.
372. Blackwell,T.K., Bowerman,B., Priess,J.R. and Weintraub,H. (1994) Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. *Science*, **266**, 621–628.
373. Kophengnavong,T., Carroll,A.S. and Blackwell,T.K. (1999) The SKN-1 amino-terminal arm is a DNA specificity segment. *Mol. Cell. Biol.*, **19**, 3039–3050.
374. Mack,D.R., Chiu,T.K. and Dickerson,R.E. (2001) Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.*, **312**, 1037–1049.
375. Harris,L.-A., Watkins,D., Williams,L.D. and Koudelka,G.B. (2013) Indirect Readout of DNA Sequence by P22 Repressor: Roles of DNA and Protein Functional Groups in Modulating DNA Conformation. *J. Mol. Biol.*, **425**, 133–143.
376. Watkins,D., Mohan,S., Koudelka,G.B. and Williams,L.D. (2010) Sequence Recognition of DNA by Protein-Induced Conformational Transitions. *J. Mol. Biol.*, **396**, 1145–1164.
377. Harris,L.-A., Williams,L.D. and Koudelka,G.B. (2014) Specific minor groove solvation is a crucial determinant of DNA binding site recognition. *Nucleic Acids Res.*, 10.1093/nar/gku1259.
378. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
379. Watkins,D., Mohan,S., Koudelka,G.B. and Williams,L.D. (2010) Sequence Recognition of DNA by Protein-Induced Conformational Transitions. *J. Mol. Biol.*, **396**, 1145–1164.
380. Iwahara,J., Esadze,A. and Zandarashvili,L. (2015) Physicochemical Properties of Ion Pairs of Biological Macromolecules. *Biomolecules*, **5**, 2435–2463.
381. Anderson,K.M., Nguyen,D., Esadze,A., Zandrashvili,L., Gorenstein,D.G. and Iwahara,J. (2015) A chemical approach for site-specific identification of NMR signals from protein side-chain NH_3^+ groups forming intermolecular ion pairs in protein-nucleic acid complexes. *J. Biomol. NMR*, **62**, 1–5.
382. Esadze,A., Chen,C., Zandarashvili,L., Roy,S., Pettitt,B.M. and Iwahara,J. (2016) Changes in conformational dynamics of basic side chains upon protein–DNA association. *Nucleic Acids Res.*, 10.1093/nar/gkw531.
383. Mirzabekov,A.D. and Rich,A. (1979) Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending. *Proc. Natl. Acad. Sci. U. S. A.*, **76**, 1118–1121.
384. Elcock,A.H. and McCammon,J.A. (1996) The Low Dielectric Interior of Proteins is Sufficient To Cause Major Structural Changes in DNA on Association. *J. Am. Chem. Soc.*,

118, 3787–3788.
385. Zacharias,M. (2003) Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci. Publ. Protein Soc.*, **12**, 1271–1282.
386. Spencer,J.V. and Arndt,K.M. (2002) A TATA Binding Protein Mutant with Increased Affinity for DNA Directs Transcription from a Reversed TATA Sequence In Vivo. *Mol. Cell. Biol.*, **22**, 8744–8755.
387. Marcovitz,A. and Levy,Y. (2011) Frustration in protein–DNA binding influences conformational switching and target search kinetics. *Proc. Natl. Acad. Sci.*, **108**, 17957–17962.
388. Vuzman,D. and Levy,Y. (2012) Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.*, **8**, 47–57.

Article 1

Nucleic Acids Research Advance Access published December 31, 2015

doi: 10.1093/nar/gkv1511

Dynamics and recognition within a protein–DNA complex: a molecular dynamics study of the SKN-1/DNA interaction

Loïc Etheve, Juliette Martin and Richard Lavery*

BMSSI UMR 5086 CNRS / Univ. Lyon I, Institut de Biologie et Chimie des Protéines, 7 passage du Vercors, Lyon 69367, France

Received September 23, 2015; Revised November 30, 2015; Accepted December 15, 2015

ABSTRACT

Molecular dynamics simulations of the *Caenorhabditis elegans* transcription factor SKN-1 bound to its cognate DNA site show that the protein–DNA interface undergoes significant dynamics on the microsecond timescale. A detailed analysis of the simulation shows that movements of two key arginine side chains between the major groove and the backbone of DNA generate distinct conformational substates that each recognize only part of the consensus binding sequence of SKN-1, while the experimentally observed binding specificity results from a time-averaged view of the dynamic recognition occurring within this complex.

INTRODUCTION

In recent years, a large structural database of protein–DNA complexes has been established, mainly through the contribution of X-ray crystallography. Although this information has undoubtedly been invaluable in understanding many aspects of protein–DNA interactions, it is true that it gives a rather static view of such complexes. The possible role of the dynamics of protein–DNA interfaces has nevertheless been a subject of interest for many years. A significant number of experimental studies have notably aimed at understanding how proteins approach and bind to their DNA targets and how they distinguish non-specific from cognate sites. Both, nuclear magnetic resonance (NMR) and paramagnetic resonance approaches have been used to better characterize non-specific protein binding and to analyze how such largely electrostatic interactions (reliant on arginine or lysine salt bridges with DNA phosphate groups) enable enhanced diffusion along DNA and can be subsequently transformed into specific binding, at least partially through the establishment of direct contacts with the nucleic acid bases (1–5). These mechanisms have also been the subject of a large number of theoretical (6–9) and molecular simulation studies (10–15) at various levels of detail, providing models of recognition mechanisms and suggesting how these mechanisms finally control the kinetics of gene expression at the cellular level (16–18).

The role of dynamics is however not limited to non-specific complexes and search mechanisms. Dynamics can also be important for specific protein–DNA complexes. Flexible, positively charged protein tails are a feature of many transcription factors. These tails, and also flexible linkers between DNA binding domains, can assist binding and can serve to fine tune specificity (19,20). Novel NMR studies using ^{15}N relaxation times and ^{15}N-^{31}P scalar coupling have also shown that lysine-phosphate salt bridges within specific complexes are themselves dynamic and direct interactions are regularly broken and remade (21–23), in line with earlier studies of salt bridges within proteins (24). This finding has recently been supported by all-atom molecular dynamics (MD) studies of homeodomain and Zn-finger complexes with DNA (25). Another aspect of protein–DNA interface dynamics is illustrated in a recent MD study of telomere repeat binding factors (TRF1 and TRF2), where the dynamics of individual amino acids chains suggested that they could contribute to the recognition of more than one base pair, helping to resolve conflicting experimental data (26).

As part of our ongoing attempt to better understand protein–DNA interactions using computer simulation techniques, we decided to couple long MD simulations with a time-dependent analysis of sequence selectivity using a sequence threading technique (ADAPT) that we have developed (27–30). ADAPT enables us to calculate and rank the binding energy of all possible DNA sequences within a protein–DNA complex (energy minimizing the interface structure for every sequence) and thus to obtain a computational position weight matrix (PWM). We already used this approach to study the appearance of base sequence selectivity during the approach of the mammalian transcription factor SRY to its DNA target (12). SRY, which controls the development of the male phenotype, is a member of the SOX (SRY-type HMG Box) family (31). By binding to the DNA minor groove, this protein creates significant DNA

*To whom correspondence should be addressed. Tel: +33 0 4 72 72 26 37; Fax: +33 0 4 72 72 26 04; Email: richard.lavery@ibcp.fr

deformation (32). We were able to show that this deformation indeed plays a major role in the resulting binding selectivity and that SRY therefore relies on a so-called indirect recognition mechanism.

Here, we chose to study a very different protein, the transcription factor SKN-1. SKN-1 is a *Caenorhabditis elegans* transcription factor involved in early embryonic development, oxidative stress resistance and aging (33,34). It is homologous to the human Nrf proteins that are also involved in stress response. Although it contains a basic C-terminal helix bound in the major groove of DNA analogous to the bZIP transcription factors (e.g. c-Jun and GCN4), it lacks a leucine zipper and does not dimerize. It also contains a basic N-terminal tail similar to those of the homeodomain proteins (35) that is responsible for high-affinity binding to AT-rich sequences at the 5′-end of the binding site (36). Its consensus binding site involves five base pairs RTCAT (where R ≡ A/G) (37). Genomic studies of genes up- or down-regulated by SKN-1 are consistent with this consensus, but show some modulations in specificity within the consensus site (38,39). The crystal structure of the 84 residue C-terminal DNA binding domain complexed with a cognate DNA oligomer (35) shows that this transcription factor induces only moderate DNA deformation and is consequently expected to recognize its binding site via a direct mechanism involving specific amino-acid base contacts.

In line with the NMR and simulations studies cited above (21,22,25), the 0.5 μs MD simulation of the SKN-1/DNA complex we have carried out shows significant dynamics at the protein–DNA interface. Most interestingly, this involves the breakage of backbone salt bridges and formation of base contacts, recalling the mechanisms described for the passage between non-specific and specific complexes (1,2,11), but here occurring within an existing specific complex.

By coupling our MD simulation with ADAPT sequence threading we have been able to establish that the observed interface dynamics indeed affects sequence selectivity. This suggests that the protein–DNA interfaces of specifically bound transcription factors may be considerably more dynamic than previously expected and, moreover, that an observed binding specificity may, at least in some cases, be the time-averaged result of a number of different sub-states where only parts of the overall cognate sequence are actually recognized.

MATERIALS AND METHODS

MD simulations

The structure of the SKN-1/DNA complex (PDB code 1SKN) was taken from the X-ray study of Rupert *et al.* (35). The single-stranded ends of the DNA oligomer were completed with complementary nucleotides to form a 17-mer (see Figure 1A and B). Hydrogen atoms were added to both the DNA and the protein and the complex was solvated with SPC/E water molecules (40) within a truncated octahedral box, ensuring a solvent shell of at least 10 Å around the solute. The solute was neutralized with 32 potassium ions and then sufficient K^+/Cl^- ion pairs were added to reach a concentration of 150 mM. The ions were initially placed at

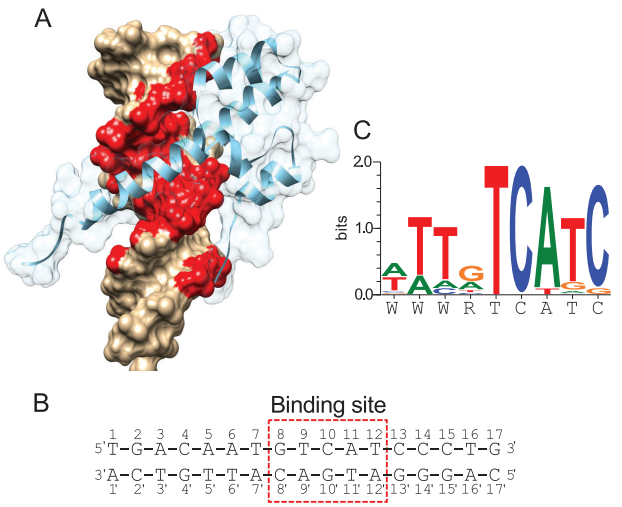


Figure 1. (**A**) Structure of SKN-1 protein–DNA complex (35). SKN-1 is shown in blue indicating its secondary structure and its surface envelope. The DNA oligomer is shown as a brown surface envelope with the protein-binding surface indicated in red. (**B**) DNA sequence used for the MD simulations, with the principal protein-binding site delimited by the red dashed box. Note that the first 'Watson' strand of the oligomer is numbered 1–17 in the 5′-3′ sense. Each complementary nucleotide in the 'Crick' strand has an identical number with a quote. (**C**) Experimental PWM for SKN-1 (W ≡A/T, R ≡A/G) from the JASPAR database (63).

random, but at least 5 Å from DNA and 3.5 Å from one another. The resulting system contained roughly 10 400 water molecules and 34 000 atoms in total.

MD simulations were performed with the AMBER 12 suite of programs (41,42) using PARM99 parameters (43) and the bsc0 modifications (44) for the solute and Dang parameters (45) for the surrounding ions. Simulations employed periodic boundary conditions and electrostatic interactions were treated using the particle-mesh Ewald algorithm (46,47) with a real space cutoff of 9 Å. Lennard-Jones interactions were truncated at 9 Å. A pair list was built with a buffer region and a list update was triggered whenever a particle moved by more than 0.5 Å with respect to the previous update.

The system was initially subjected to energy minimization with harmonic restraints of 25 kcal mol^{-1} $Å^{-2}$ on the solute atoms. The system was then heated to 300 K at constant volume during 100 ps. Constraints were then relaxed from 5 to 1 kcal mol^{-1} $Å^{-2}$ during a series of 1000 steps of energy minimization (500 steps of steepest descent and 500 steps of conjugate gradient) followed by 50 ps of equilibration with restraints of 0.5 kcal mol^{-1} $Å^{-2.}$ and 50 ps without solute restraints. The 500 ns production simulations were carried out at constant temperature (300 K) and pressure (1 bar) with a 2 fs time step. During these simulations pressure and temperature were maintained using the Berendsen algorithm (48) with a coupling constant of 5 ps and SHAKE constraints were applied to all bonds involving hydrogens (49). Conformational snapshots were saved for further analysis every ps. For comparison purposes, the 17-mer DNA oligomer was also simulated alone using an identical protocol, creating a second 500 ns trajectory.

Conformational and environmental analysis

Average DNA conformation, DNA conformational fluctuations and ion distributions around the SKN-1/DNA complex during the MD simulations were analyzed with the Curves+ program (50) and the Canal and Canion utilities (https://bisi.ibcp.fr/tools/curves_plus/). Using the recently developed ion analysis approach, based on describing ion positions with respect to the DNA helical axis, it was notably possible to calculate average ion molarities within the DNA grooves (51,52). As in our earlier work, the groove limit was set at a radius of 10.25 Å from the DNA helical axis (the average radial position of the backbone phosphorus atoms), while the angular limits were determined by the average position of the sugar C1' atoms. Lastly, hydrogen bond and salt bridges were analyzed using AMBER Tools (53) applying a distance cut-off of ≤ 3.5 Å between the relevant heavy atoms and an angle cut-off of $\geq 135^{\circ}$ at the intervening hydrogen atom.

Clustering the MD trajectory

In order to identify conformational clusters within the MD trajectory, we began by extracting snapshots every 200 ps. Since we were principally interested in the evolution of the protein–DNA binding specificity, we characterized each snapshot by counting the number of contacts between the protein and the DNA bases. Each contact between heavy atoms scored 1 for distances r_{ij} below 4 Å (using shorter distances would result in many transient 'breaks' that add noise to the analysis). In order to further increase the robustness, we used a buffer zone from 4 Å to 5 Å over which the score was modulated with a sigmoidal function $s(i,j)$ of the distance r_{ij} between the atoms i and j:

$$s(i,j) = \frac{1}{1 + e^{10*(r_{ij}-4.5)}}$$

This analysis yielded a 74 (amino acid) by 34 (DNA base) matrix for each snapshot. The overall distance $\mathbf{d(x,y)}$ between any two such matrices $\mathbf{x}$ and $\mathbf{y}$ was then calculated using the Manhattan algorithm (54).

$$\mathbf{d}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} |\mathbf{x}_i - \mathbf{y}_i|$$

Next, the Ward agglomerative hierarchical clustering method (55–57) was used to classify the different snapshots into groups by minimizing the variance within each cluster and increasing the weighted squared distance between cluster centers. The distance matrix and cluster representation were carried out using the R software package (58).

Binding specificity analysis

SKN-1 binding specificity was determined for any chosen snapshot from the MD trajectory (after a brief Cartesian coordinate energy minimization to remove bond length and base plane deformations) using the so-called ADAPT approach (28,29) implemented within the JUMNA program (59). This consists of calculating the complexation energy of the SKN-1/DNA complex for all possible DNA base sequences and deriving a PWM. In order to do this, it is necessary to thread all possible base sequences into the DNA oligomer within the complex, adapting the protein–DNA interface in each case using internal coordinate energy minimization. This was performed with the same AMBER parameterization used for the MD simulations, but replacing the explicit solvent and ion shell with a simple continuum model using a sigmoidal distance-dependent dielectric function and reduced phosphate charges (29). In parallel, an identical base sequence is threaded into the average conformation of the isolated DNA oligomer and energy minimization is again performed. Finally, another energy minimization is performed for the isolated protein (with flexibility limited to the side chains included within the interface cutoff distance, see below). Subtracting the isolated DNA oligomer and protein energies from the SKN-1/DNA complex energy yields the complex formation energy, which can be further analyzed in terms of two components: the DNA deformation energy and the protein–DNA interaction energy. In the present case, the nine central base pairs of the DNA oligomer were scanned, corresponding to the SKN-1 cognate site flanked by two extra base pairs on either end, leading to $4^9 = 262{,}144$ possible sequences. ADAPT calculations were accelerated using a divide-and-conquer technique, breaking each sequence into overlapping 4 bp fragments and thus reducing the total number of calculations to $6 \times 4^4 = 1024$ (for the complex and for the isolated DNA oligomer), without significant loss of accuracy (29). Protein flexibility was also limited to side chains within 20 Å of the protein–DNA interface. The energies resulting from this analysis were converted into PWMs using the WebLogo software (60). Finally, by analyzing the binding specificity derived from the sequence-dependent DNA deformation energy, or from the sequence-dependent protein–DNA interaction energy we could also analyze specificity in terms of its indirect and direct components.

We remark that for the purpose of this study we extended the utility programs associated with ADAPT to be able to derive a single PWM from a number of MD snapshots. In this case, ADAPT calculations were based on sequence-dependent energy differences with respect to the minimum energy for each snapshot, enabling us to overcome sequence-independent energy changes mainly caused by the necessary simplification of the electrostatic calculations (which rely on a rudimentary implicit solvent representation). Using this approach it was possible to describe the sequence selectivity of each of the conformational substates detected by the cluster analysis and to compare this to the consensus selectivity for the entire MD simulation, or to experimental binding data.

RESULTS AND DISCUSSION

We begin by considering the general impact of SKN-1 binding on DNA structure and dynamics. As shown in Figure 1A (see also Supplementary Figure S1) the protein inserts its long C-terminal α-helix in the major groove of the DNA binding site, while its N-terminal arm binds to the adjacent minor groove. In addition to amino acid side chain contacts with the DNA bases, the crystal structure of the

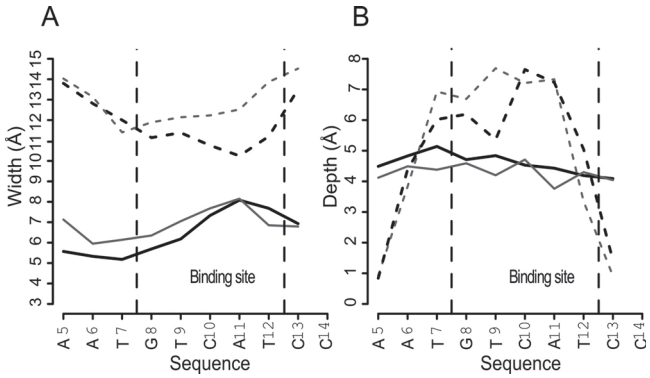


Figure 2. DNA groove dimensions (Å), width (**A**) and depth (**B**), within the isolated DNA oligomer (gray lines) and within the SKN-1/DNA complex (thick black lines). Major groove dimensions are indicated with dashed lines and minor groove dimensions with solid lines. Vertical dashed lines indicate the protein-binding site.

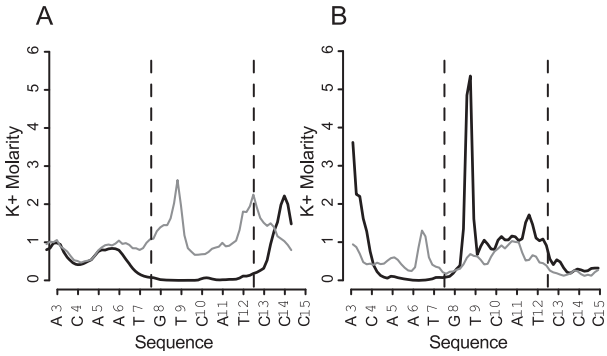


Figure 3. Potassium ion molarity: (**A**) inside the major groove and (**B**) inside the minor groove for the isolated DNA oligomer (gray lines) and for the DNA/SKN-1 complex (thick black lines). The sequences of both strand are shown in the 5′-3′ direction. Vertical dashed lines indicate the protein-binding site.

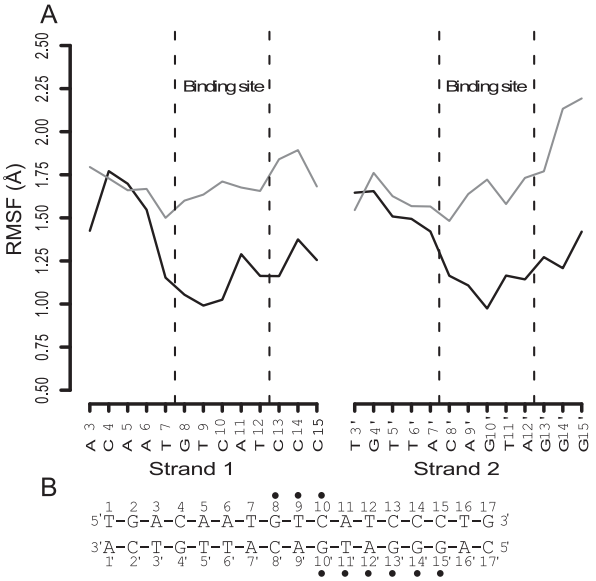


Figure 4. (**A**) Root mean square fluctuation (Å) of phosphorus atoms within the isolated DNA oligomer (gray lines) and within the DNA/SKN-1 complex (thick black lines). The sequences of both strand are shown in the 5′-3′ direction. Vertical dashed lines indicate the protein-binding site. (**B**) Black circles show the position of salt bridges within the DNA/SKN-1 complex.

complex is stabilized by seven salt bridges involving seven arginines (R503, R506, R507, R508, R516, R521, R522). Of these residues, four (R503, R506, R507, R508) belong to the central support region (see Supplementary Figure S1) and three (R516, R521, R522) are located in the C-terminal helix of the protein. These interactions link the protein with the phosphate groups at positions G8 and C10 in the Watson strand and positions G10', T11', G14' and G15' in the Crick strand.

Comparing the average structures derived from the MD simulations of the SKN-1/DNA complex and of DNA alone, we see that protein binding has relatively little structural impact. There are no major changes in helical parameters or backbone parameters, although the average twist along the binding site increases by 2° in the presence of the protein. We also observe slight bending of the DNA toward the protein (6.5° versus 2.5° in the isolated DNA oligomer), but this value is less than that seen in the crystal structure (22°). These changes are coupled to a change in groove geometry, as shown in Figure 2. Insertion of the C-terminal α-helix in the major groove leads to a decrease in width of roughly 2 Å at positions C10-C13 and a localized decrease in depth at position T9. The binding of the N-terminal tail has a smaller effect on the minor groove (positions 5–7), where we see a narrowing of roughly 1 Å coupled with a small increase in depth.

Before passing to an analysis of the dynamics of the SKN-1/DNA complex, we lastly consider the effect of protein binding on the ionic environment of DNA. As shown in Figure 3, protein binding, not surprisingly, almost completely removes potassium cations from the major groove between positions T6 and C13, whereas we observe roughly 1–2 M potassium in this region for isolated DNA. In compensation, the K^+ molarity increases in the minor groove of the binding site, notably with a strongly localized ion site at the step G8-T9 that is absent in the isolated DNA oligomer. Secondary increases in potassium molarity are also seen at A11-T12 in the minor groove and at C13-C14 in the major groove.

We now turn to the dynamics of the SKN-1/DNA complex. The first observation is that DNA backbone dynamics decrease in the presence of the protein. This is illustrated in Figure 4 using the root mean square fluctuations of the phosphate atoms. We recall that these values were obtained by analyzing the position of the phosphorus atoms within each MD snapshot using curvilinear helicoidal coordinates with respect to the instantaneous helical axis, and then replotting them in Cartesian space with respect to the helical axis of the average DNA structure (52). This has the effect of removing any fluctuations due to DNA bending, stretching or twisting and gives an accurate view of phosphorus atom mobility. The protein clearly reduces the mobility of the phosphate groups within the binding site and the effect is particularly strong for the phosphates involved in salt bridges with SKN-1 (see Figure 4B).

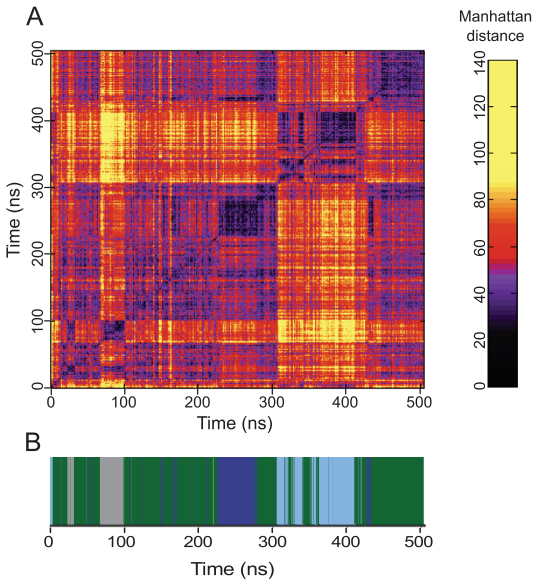
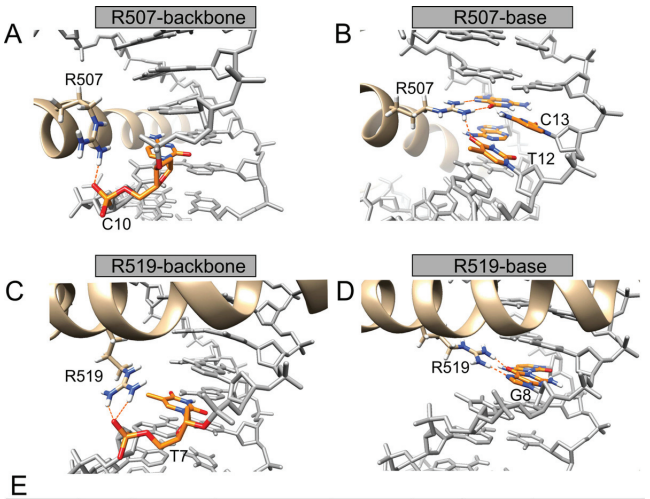


Figure 5. Clustering snapshots from the 500 ns MD trajectory of the DNA/SKN-1 complex: (**A**) Manhattan distance matrix. The vertical black to yellow scale represents increasing distances. (**B**) Clustering using the distance matrix leads to four distinct clusters whose appearance during the trajectory is indicated by the colors cyan (CL1), green (CL2), gray (CL3) and dark blue (CL4).



E

Cluster	Color	R507 base	R507 backbone	R519 base	R519 backbone	% occurrence
Cl1		-	X	X	-	17
Cl2		X	-	X	-	60
Cl3		X	-	-	X	9
Cl4		X	-	X	-	14

Figure 6. Alternative orientations observed for arginines 507 (**A**, **B**) and 519 (**C**, **D**). Orange dashed lines indicate hydrogen bonds between these arginines and DNA. The table (**E**) shows the link between the clusters observed during the MD trajectory and the R507/R519 orientations in addition to the percentage occurrence of each cluster during the trajectory.

In contrast to this apparent rigidification, we see significant dynamics at the protein–DNA interface. Note that Figure 4B indicates nine salt bridges, in contrast to the seven seen in the crystal structure. This change is indicative of what occurs during the MD simulation where we see many intermittent protein–DNA contacts. Most of these are alternative interactions involving the same side chains that form salt bridges in the crystal structure, although some are completely new, notably involving Arg 457 and Lys 460 within the N-terminal tail. Table 1 shows contacts seen in both the crystal structure and the MD simulation in black, while those appearing only in the simulation are shown in bold/red. From these results, we can see that most interactions are only present for a fraction of the 0.5 μs trajectory, although those observed in the crystal structure are generally the longest lived. It also shows that interactions between given side chains and nucleotides often involve different sets of atoms, in some cases simultaneously, creating bidentate interactions.

On the basis of this finding, we decided to see if the interface dynamics were random or reflected the existence of conformational sub-states. As described in the methodology section we carried out this analysis by building a contact matrix between protein side chains and DNA bases for snapshots every 200 ps along the trajectory, leading to a total of 2500 matrices. Measuring the Manhattan distances between these matrices created a new distance matrix 2500 × 2500 that could then be analyzed to detect conformational clusters. The results shown in Figure 5 indicate that the MD trajectory is in fact composed of four distinct conformational clusters.

The initial cluster, CL1 (cyan) is closest to the X-ray conformation of the complex. It is lost after only 5 ns, but then reappears intermittently during the last third of the trajectory and finally constitutes 17% of the trajectory. The second cluster to appear, CL2 (green) is the most common and reappears throughout the simulation representing in total 60% of the trajectory. A third cluster, CL3 (gray) appears around 70 ns, but only makes up 9% of the trajectory and is not seen after the first 100 ns. The final cluster, CL4 (dark blue) appears in the middle of the simulation and again briefly toward the end, making up 14% of the total.

By extracting snapshots belonging to each of the four clusters we can analyze their structural characteristics. The first observation is that the CL2 (green) and CL4 (dark blue) clusters are very similar to one another, differing only by the position of the N-terminal arm, which interacts with the bases in the DNA minor groove in the more common CL2 (green) cluster (without affecting the groove geometry), but with the DNA backbone in the CL4 (dark blue) cluster. We will consequently temporarily group these two clusters together (and denominate them as CL2/4). The main feature distinguishing the remaining clusters turns out to be to the position of the side chains of two arginines: R507 and R519. In CL1, R507 lies close to the DNA backbone, intermittently forming a salt bridge with the phosphate of C10 or, more rarely, those of A11 and G15'. In contrast, in CL2/4 and CL3 it binds in the DNA major groove forming a bidentate interaction with O6 and N7 of G13' (as seen in other protein–DNA complexes (61,62)) and, intermittently, to O4 of T12. Similarly, in CL1 and CL2/4, R519 also forms a bidentate interaction with O6 and N7 of G8, whereas in CL3 it is close to the backbone, intermittently forming a salt bridge with the phosphate of T7. The alternate conformations of R507 and R519 are illustrated in Figure 6. As summarized in Figure 6E, the combination of these two side

Table 1. SKN-1 salt bridges with the DNA backbone (columns 1–3) and hydrogen bonds with the DNA bases (columns 4–6) are highly dynamic

Protein	DNA Backbone	Percentage	Protein	DNA base	Percentage
ARG 457(N)	**G 8(O2P)**	**28**	**ARG 457(NE)**	**T 5'(O2)**	**15**
LYS 460(N)	**T 9(O2P)**	**37**	**ARG 457(NH1)**	**T 5'(O2)**	**12**
ARG 503(NE)	G 15'(O2P)	52	**ARG 457(NH2)**	**T 5'(O2)**	**19**
ARG 503(NE)	**G 15'(O1P)**	**38**	**ARG 457(NH1)**	**T 6'(O2)**	**16**
ARG 503(NH2)	**G 15'(O2P)**	**20**	**ARG 507(NH2)**	**T 12(O4)**	**24**
ARG 503(NH2)	**G 15'(O1P)**	**49**	**ARG 507(NH1)**	**G 13'(N7)**	**25**
ARG 503(NH1)	**A 16'(O2P)**	**13**	**ARG 507(NH2)**	**G 13'(O6)**	**72**
ARG 503(NH2)	**A 16'(O2P)**	**19**	ASN 511(ND2)	T 11'(O4)	92
ARG 503(NH2)	**A 16'(O1P)**	**15**	ASN 511(OD1)	C 10(N4)	97
ARG 506(NE)	G 14'(O2P)	17	ARG 519(NH1)	G 8(N7)	78
ARG 506(NE)	G 14'(O1P)	96	ARG 519(NH2)	G 8(O6)	79
ARG 506(NH2)	G 14'(O2P)	98			
ARG 507(NH2)	**G 15'(O1P)**	**3**			
ARG 507(NE)	**G 15'(O2P)**	**2**			
ARG 507(NH2)	**A 11(O1P)**	**3**			
ARG 507(NH1)	C 10(O1P)	9			
ARG 508(NE)	**T 9(O2P)**	**20**			
ARG 508(NH2)	**T 9(O2P)**	**37**			
ARG 508(NE)	**T 9(O5')**	**11**			
ARG 508(NH1)	C 10(O1P)	50			
LYS 510(NZ)	**G 13'(O1P)**	**21**			
ARG 516(NE)	**G 8(O1P)**	**42**			
ARG 516(NH2)	G 8(O2P)	30			
ARG 516(NH2)	G 8(O1P)	37			
ARG 519(NH2)	**T 7(O1P)**	**5**			
ARG 519(NH1)	**T 7(O1P)**	**4**			
ARG 521(NE)	T 11'(O1P)	44			
ARG 521(NH2)	T 11'(O1P)	89			
ARG 521(NH1)	**A 12'(O1P)**	**59**			
ARG 521(NH2)	**A 12'(O1P)**	**15**			
ARG 521(NH2)	**A 12'(O5')**	**16**			
ARG 522(NE)	**G 10'(O1P)**	**28**			
ARG 522(NH2)	G 10'(O1P)	42			
ARG 525(NH1)	**T 11'(O2P)**	**12**			

The percentage time each interaction was observed during the 0.5 μs MD trajectory is given in columns 3 and 6. Interactions shown in bold/red are only observed in the MD trajectory, while those in black are seen in the crystal structure (35) and in the MD trajectory.

chain flips gives rise to three conformational sub-states that distinguish the clusters CL1, CL2/4 and CL3.

The dynamical behavior of R507 and R519 are illustrated by the time series of side chain-DNA backbone/base distances in Supplementary Figure S2, which, for reference, also shows the distance fluctuations for the R506 salt bridge with the phosphate of G14'. While the significant perturbations of the R506 interaction occur only occasionally, R507 and R519 show complex fluctuations whether they are interacting with DNA bases or DNA phosphates. Analyzing snapshots every picosecond along the MD trajectory, with distance and angle cutoffs of 3.5 Å and 135°, respectively, leads to lifetimes of less than 30 ps for either base or phosphate interactions. However, ignoring breaks that last no longer than 1 ps typically increases the lifetimes to 100–400 ps. By comparison, the R506 salt bridge has lifetimes of roughly 100 ps or 1800 ps, depending on whether 1 ps breaks are taken into account or ignored.

By applying our sequence-threading approach ADAPT to multiple snapshots belonging to each cluster (7, 12, 10 and 2 for CL1, CL2, CL3 and CL4, respectively), we were able test whether the very localized changes in the two key arginines have any significant impact on how SKN-1 is recognizing the DNA base sequence. The results are shown in Figure 7, where CL2 and CL4 have again been grouped together since they yield identical PWMs. If we concentrate on the bases at positions 8, 12 and 13, the results are relatively easy to interpret. When R519 interacts with position 8 in CL1, a 'G' appears strongly at this position in the PWM. Similarly when R507 interacts with positions 12 and 13 in

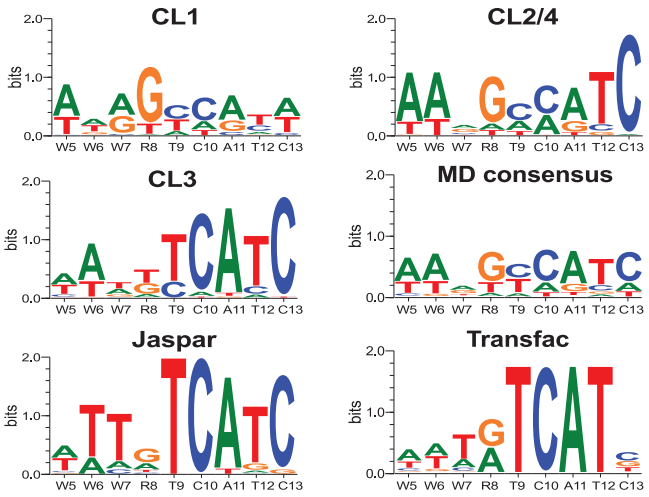


Figure 7. SKN-1 PWMs resulting from the analysis of snapshots belonging to each of the three distinct clusters and also a consensus PWM using the snapshots from the entire MD trajectory. These results can be compared to the experimental results from the JASPAR (63) and TRANSFAC databases (64) (W ≡A/T, R ≡ A/G).

CL2/4 and CL3, a clear 'TC' appears at these positions. Finally, when both arginines bind within the major groove in CL2/4, both a 'G' at position 8 and a 'TC' at positions 12 and 13 dominate. However, we can also see that the R507 groove interaction also impacts positions 10 and 11 at the 3′-end of the binding site and leads to the appearance of the CATC motif in both CL2/4 and CL3. As expected the majority of the recognition in each cluster comes from direct protein–DNA contacts. Although some base pairs show selectivity due to DNA deformation (notably for T at positions 10 and 13, see Supplementary Figure S3), protein–DNA interaction is clearly the dominant factor in the overall PWM.

We remark that the movement of the N-terminal tail does not appear to have any significant impact on the PWM since the A/T-rich preference seen at the 5′-end of the SKN-1 binding site, corresponding to the location of the N-terminal tail, is virtually unchanged whether the tail lies within the minor groove (CL2 and CL3), or closer to the DNA backbones (CL1 and CL4). Supplementary Figure S4 shows one such comparison for the clusters CL2 and CL4. We conclude that its role is largely electrostatic (its cationic residues favoring the more negative minor groove potentials generated by AT base pairs) and does not require binding to a specific base site.

We can make this analysis of selectivity more quantitative by calculating Pearson correlation coefficients (PCCs) between the PWMs of the various clusters and the experimental results. We limit our analysis to the PWM for SKN-1 from the JASPAR database (63), but remark that very similar results are obtained with the equivalent data in TRANSFAC (64). The overall correlation between CL1, CL2/4 and CL3 PWMs with the JASPAR data is 0.50, 0.52 and 0.82, respectively. Thus CL3 is closest to the experimental data (which can be seen visually in Figure 7). However, if we now look at the correlations at each position within the binding site, another picture emerges. At position 8, the correlations for CL1, CL2/4 and CL3 become 0.89, 0.95 and 0.29, respectively. Thus, only CL1 and CL2/4 (where R519 is bound in the DNA groove) reproduce the experimental result. In contrast, at positions 12 and 13, the correlations for CL1, CL2/4 and CL3 change again to (0.84, -0.50), (0.99, 1.0) and (0.97, 1.0) and thus only CL2/4 and CL3 (where R507 is bound in the DNA groove) fit the experiments. This confirms the notion that each conformational sub-state is recognizing only part of the binding site. In addition, we can note that these partial recognition events are not fully compatible with one another since the consensus correlation between the simulation (using all the snapshots extracted from the MD run) and the JASPAR PWM is only 0.57. This loss of selectivity can also be quantified by calculating the total information content of the various PWMs (65), which yields 6.2, 9.0 and 9.5 for CL1, CL2/4 and CL3, respectively, but only 5.3 for the MD consensus. In contrast, if we model recognition events occurring separately in different regions of the binding site, by combining columns 1–4 from the PWM of CL1 with columns 5–9 from the PWM of CL3, the total information content becomes 10.5, close to that of the experimental JASPAR logo (11.6).

CONCLUSIONS

This computational study of the transcription factor SKN-1 bound to its cognate DNA site shows that the protein–DNA interface is dynamic and, notably, that two arginine side chains oscillate between the formation of direct interactions with DNA bases and interactions with the DNA backbone. The cationic N-terminal arm of SKN-1 undergoes similar oscillations. This dynamics is analogous to what has been seen at protein-protein interfaces (66,67) and is compatible with recent NMR studies and simulation studies showing that protein–DNA salt bridges are broken on sub-nanosecond timescales (21,25). In our case, the temporary loss of protein-base interactions significantly alters sequence selectivity and suggests that the observed consensus binding sequence of the transcription factor exists as the time-averaged ensemble of a number of distinct conformational sub-states that each recognize different parts of the binding site. As other authors have already pointed out, the dynamic nature of the protein–DNA interface may aid binding both by making the transition between non-specific and specific sites easier and by reducing the entropic penalty for binding. From a computational point of view the 0.5 μs simulations carried out here led to the detection of four distinct sub-states, but we cannot exclude that this number would grow with longer simulations, or that the the relative sub-state populations could evolve. We conclude that understanding protein–DNA recognition mechanisms using molecular simulations, at least in some cases, may very well require trajectories on the microsecond scale.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to acknowledge GENCI for a generous allocation of supercomputer resources at the CINES center in Montpellier.

FUNDING

This work is funded by the ANR project CHROME ANR-12-BSV5-0017-01 and GENCI for a generous allocation of supercomputer resources at the CINES center in Montpellier. L.E. was funded by a doctoral grant from Rhône-Alpes ARC 1 Santé. Funding for open access charge: ANR project CHROME ANR-12-BSV5-0017-01.
Conflict of interest statement. None declared.

REFERENCES

1. Kalodimos,C.G., Biris,N., Bonvin,A.M., Levandoski,M.M., Guennuegues,M., Boelens,R. and Kaptein,R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386–389.
2. Kalodimos,C.G., Boelens,R. and Kaptein,R. (2004) Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. *Chem. Rev.*, **104**, 3567–3586.
3. Iwahara,J., Zweckstetter,M. and Clore,G.M. (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10562–10567.

4. Iwahara,J. and Clore,G.M. (2006) Direct observation of enhanced translocation of a homeodomain between DNA cognate sites by NMR exchange spectroscopy. *J. Am. Chem. Soc.*, **128**, 404–405.
5. Clore,G.M., Tang,C. and Iwahara,J. (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr. Opin. Struct. Biol.*, **17**, 603–616.
6. Mirny,L., Slutsky,M., Wunderlich,Z., Tafvizi,A., Leith,J. and Kosmrlj,A. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A Math. Theor.*, **42**, 434013.
7. Das,R.K. and Kolomeisky,A.B. (2010) Facilitated search of proteins on DNA: correlations are important. *Phys. Chem. Chem. Phys.*, **12**, 2999–3004.
8. Koslover,E.F., Díaz de la Rosa,M.A. and Spakowitz,A.J. (2011) Theoretical and computational modeling of target-site search kinetics in vitro and in vivo. *Biophys. J.*, **101**, 856–865.
9. Kolomeisky,A.B. (2011) Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.*, **13**, 2088–2095.
10. Furini,S., Domene,C. and Cavalcanti,S. (2010) Insights into the sliding movement of the lac repressor nonspecifically bound to DNA. *J. Phys. Chem. B*, **114**, 2238–2245.
11. Temiz,A.N., Benos,P.V. and Camacho,C.J. (2010) Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains. *Nucleic Acids Res.*, **38**, 2134–2144.
12. Bouvier,B., Zakrzewska,K. and Lavery,R. (2011) Protein-DNA recognition triggered by a DNA conformational switch. *Angew. Chem. Int. Ed. Engl.*, **50**, 6516–6518.
13. Chen,C. and Pettitt,B.M. (2011) The binding process of a nonspecific enzyme with DNA. *Biophys. J.*, **101**, 1139–1147.
14. Furini,S., Barbini,P. and Domene,C. (2013) DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res.*, **41**, 3963–3972.
15. Ando,T. and Skolnick,J. (2014) Sliding of proteins non-specifically bound to DNA: Brownian dynamics studies with coarse-grained protein and DNA models. *PLoS Comput. Biol.*, **10**, e1003990.
16. Sela,I. and Lukatsky,D. (2011) DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.*, **101**, 160–166.
17. Afek,A. and Lukatsky,D.B. (2013) Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophys. J.*, **105**, 1653–1660.
18. Afek,A. and Lukatsky,D. (2013) Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. *Biophys. J.*, **104**, 1107–1115.
19. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233.
20. Fuxreiter,M., Simon,I. and Bondos,S. (2011) Dynamic protein-DNA recognition: beyond what can be seen. *Trends Biochem. Sci.*, **36**, 415–423.
21. Anderson,K.M., Esadze,A., Manoharan,M., Bru schweiler,R., Gorenstein,D.G. and Iwahara,J. (2013) Direct observation of the ion-pair dynamics at a protein–DNA interface by NMR spectroscopy. *J. Am. Chem. Soc.*, **135**, 3613–3619.
22. Zandarashvili,L., Esadze,A. and Iwahara,J. (2013) NMR studies on the dynamics of hydrogen bonds and ion pairs involving lysine side chains of proteins. *Adv. Protein Chem. Struct. Biol.*, **93**, 37–80.
23. Zandarashvili,L. and Iwahara,J. (2014) Temperature dependence of internal motions of protein side-chain NH3+ groups: insight into energy barriers for transient breakage of hydrogen bonds. *Biochemistry*, **54**, 538–545.
24. Esadze,A., Li,D.W., Wang,T., Brüschweiler,R. and Iwahara,J. (2011) Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear 1H−15N NMR spectroscopy. *J. Am. Chem. Soc.*, **133**, 909–919.
25. Chen,C., Esadze,A., Zandarashvili,L., Nguyen,D., Pettitt,B.M. and Iwahara,J. (2015) Dynamic equilibria of short-range electrostatic interactions at molecular interfaces of protein–DNA complexes. *J. Phys. Chem. Lett.*, **6**, 2733–2737.
26. Garton,M. and Laughton,C. (2013) A comprehensive model for the recognition of human telomeres by TRF1. *J. Mol. Biol.*, **425**, 2910–2921.
27. Lafontaine,I. and Lavery,R. (2001) High-speed molecular mechanics searches for optimal DNA interaction sites. *Comb. Chem. High Throughput Screen.*, **4**, 707–717.
28. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
29. Deremble,C., Lavery,R. and Zakrzewska,K. (2008) Protein-DNA recognition: Breaking the combinatorial barrier. *Comput. Phys. Commun.*, **179**, 112–119.
30. Zakrzewska,K., Bouvier,B., Michon,A., Blanchet,C. and Lavery,R. (2009) Protein–DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. *Phys. Chem. Chem. Phys.*, **11**, 10712–10721.
31. Berta,P., Hawkins,J.B., Sinclair,A.H., Taylor,A., Griffiths,B.L., Goodfellow,P.N. and Fellous,M. (1990) Genetic evidence equating SRY and the testis-determining factor. *Nature*, **348**, 448–450.
32. Murphy,E.C., Zhurkin,V.B., Louis,J.M., Cornilescu,G. and Clore,G.M. (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.*, **312**, 481–499.
33. Bowerman,B., Eaton,B.A. and Priess,J.R. (1992) skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell*, **68**, 1061–1075.
34. Bowerman,B., Draper,B.W., Mello,C.C. and Priess,J.R. (1993) The maternal gene skn-1 encodes a protein that is distributed unequally in early C. elegans embryos. *Cell*, **74**, 443–452.
35. Rupert,P.B., Daughdrill,G.W., Bowerman,B. and Matthews,B.W. (1998) A new DNA-binding motif in the Skn-1 binding domain–DNA complex. *Nat. Struct. Biol.*, **5**, 484–491.
36. Kophengnavong,T., Carroll,A.S. and Blackwell,T.K. (1999) The SKN-1 amino-terminal arm is a DNA specificity segment. *Mol. Cell. Biol.*, **19**, 3039–3050.
37. Blackwell,T.K., Bowerman,B. and Weintraub,H. (1994) Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. *Science*, **266**, 621–628.
38. Oliveira,R.P., Abate,J.P., Dilks,K., Landis,J., Ashraf,J., Murphy,C.T. and Blackwell,T.K. (2009) Condition-adapted stress and longevity gene regulation by Caenorhabditis elegans SKN-1/Nrf. *Aging Cell*, **8**, 524–541.
39. Staab,T.A., Griffen,T.C., Corcoran,C., Evgrafov,O., Knowles,J.A. and Sieburth,D. (2013) The conserved SKN-1/Nrf2 stress response pathway regulates synaptic function in Caenorhabditis elegans. *PLoS Genet.*, **9**, e1003354.
40. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
41. Pearlman,D.A., Case,D.A., Caldwell,J.W., Ross,W.S., Cheatham,T.E., DeBolt,S., Ferguson,D., Seibel,G. and Kollman,P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.*, **91**, 1–41.
42. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
43. Cheatham,T.E. 3rd, Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
44. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
45. Dang,L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous-solutions of 18-crown-6 ether - A molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
46. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
47. Darden,T., Perera,L., Li,L. and Pedersen,L. (1999) New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, **7**, R55–R60.

48. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
49. Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.*, **23**, 327–341.
50. 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.
51. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
52. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2413–2423.
53. Case,D.A., Berryman,J., Betz,R.M., Cerutti,D., Cheatham,T. III, Darden,T., Duke,R., Glese,T., Gohlke,H. *et al.* (2015) *AMBER 2015*.
54. Krause,E.F. (1987) *Taxicab geometry: an adventure in non-Euclidean geometry* . Courier Corporation, Dover, London .
55. Ward,J.H. Jr (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
56. Kaufman,L. and Rousseeuw,P.J. (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, New York .
57. Murtagh,F. and Legendre,P. (2014) Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *J. Classif.*, **31**, 274–295.
58. R Development Core Team (2009) *R: A language and environment for statistical computing*.
59. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (Junction Minimization of Nucleic-Acids). *Comput. Phys. Commun.*, **91**, 135–158.
60. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
61. McClarin,J.A., Frederick,C.A., Wang,B.-C., Greene,P., Boyer,H.W., Grable,J. and Rosenberg,J.M. (1986) Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. *Science*, **234**, 1526–1541.
62. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp represser/operator complex at atomic resolution. *Nature*, **335**, 321–329.
63. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A. and Ienasescu,H. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142-D147.
64. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
65. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
66. Lee,H.J., Hota,P.K., Chugha,P., Guo,H., Miao,H., Zhang,L., Kim,S.-J., Stetzik,L., Wang,B.-C. and Buck,M. (2012) NMR structure of a heterodimeric SAM: SAM complex: characterization and manipulation of EphA2 binding reveal new cellular functions of SHIP2. *Structure*, **20**, 41–55.
67. Zhang,L. and Buck,M. (2013) Molecular simulations of a dynamic protein complex: role of salt-bridges and polar interactions in configurational transitions. *Biophys. J.*, **105**, 2412–2417.

Supplementary data

Dynamics and recognition within a protein-DNA complex: a molecular dynamics study of the SKN-1/DNA interaction

L. Etheve, J. Martin and R. Lavery

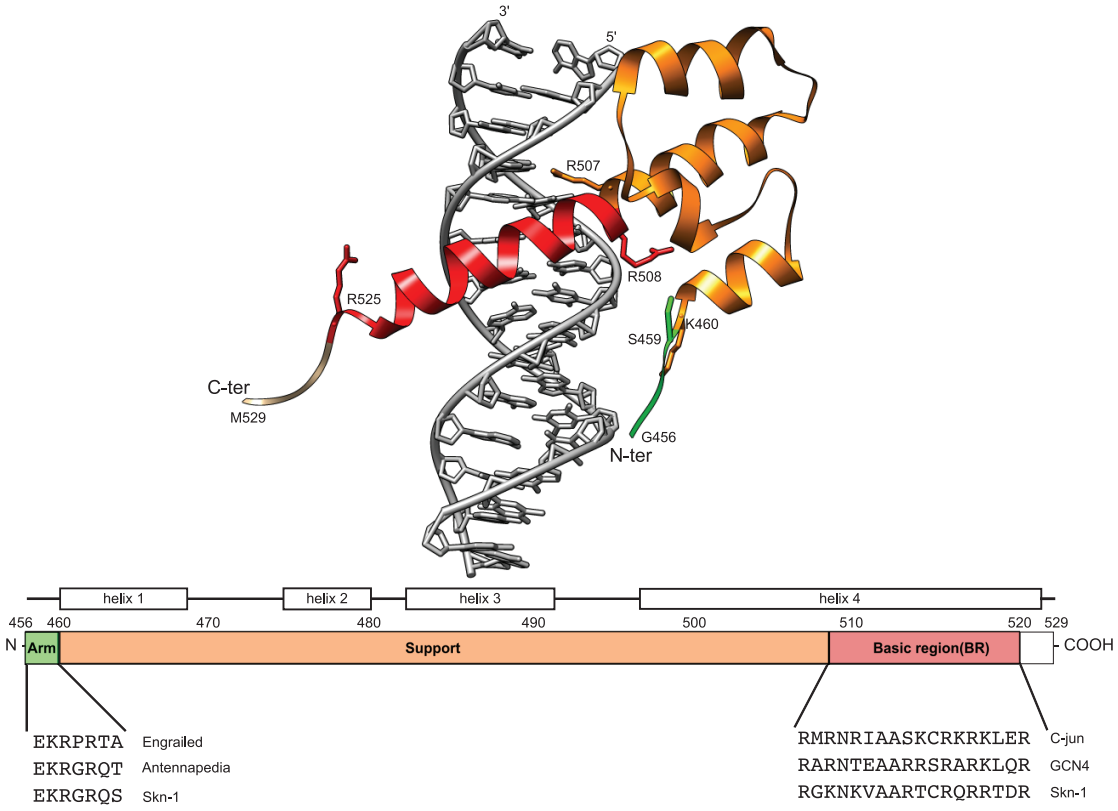


Figure S1. SKN-1 structure. The upper image shows the 3D structure of SKN-1 (1). Labeled residues define the limits of the N-terminal arm (green), the support region (orange) and the basic region (BR) helix (red). The location of these regions in the primary sequence are shown in the lower figure, along with alignments of the N-terminal arm of SKN-1 with the *Drosophilia engrailed* region homeobox (2) and *Drosophilia antennapedia* locus (3) and of the BR region with the equivalent regions of *Human c-jun (4)* and *Yeast GCN4 (5)*.

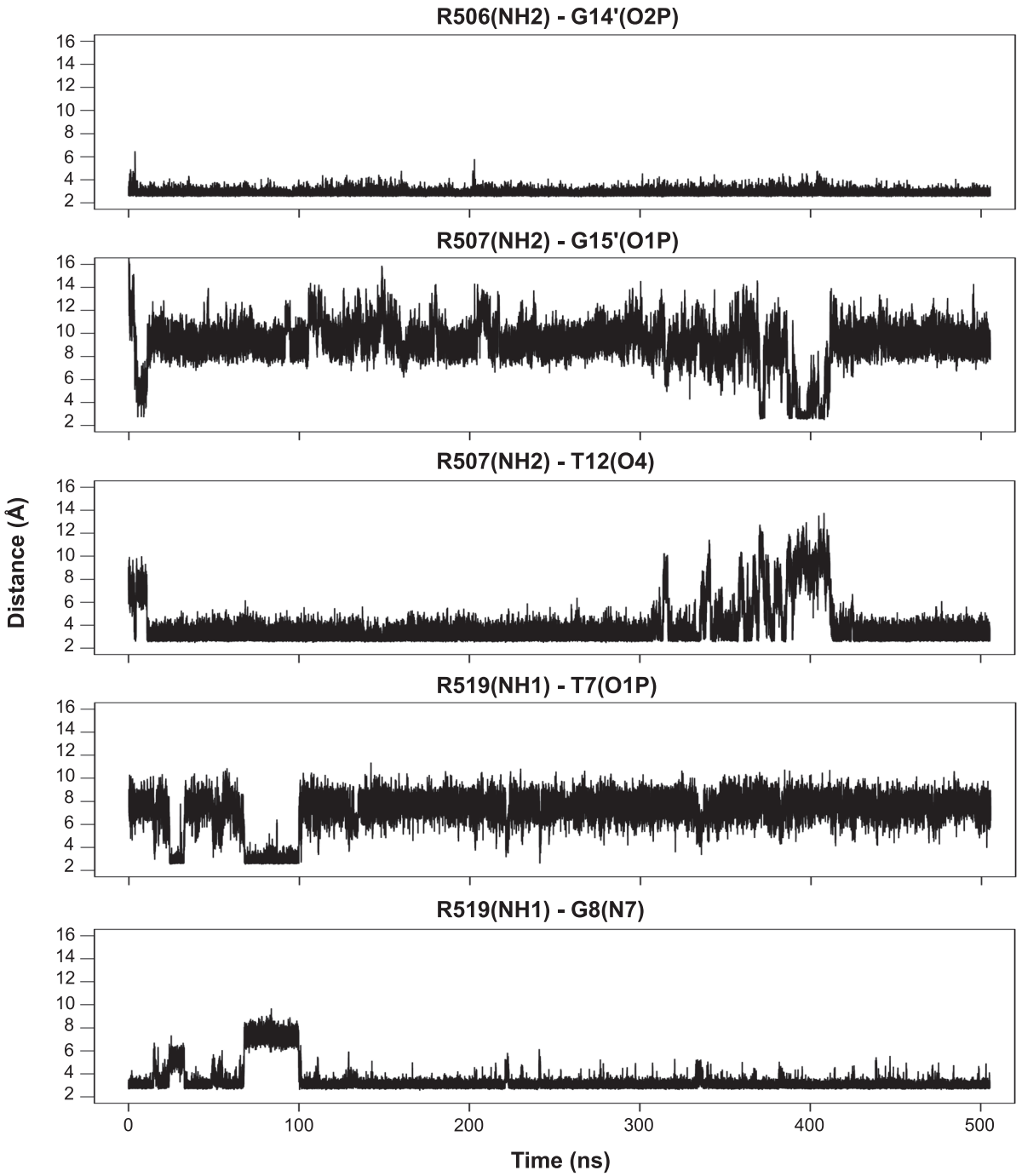


Figure S2. Time series of selected DNA backbone or base interactions formed by arginine side chains. The topmost plot shows the Arg 506-backbone interaction that persists virtually throughout the 500 ns trajectory. The two central plots show the alternating backbone/base interactions of Arg 507 and the two bottom plots shows the alternating backbone/base interactions of Arg 519.

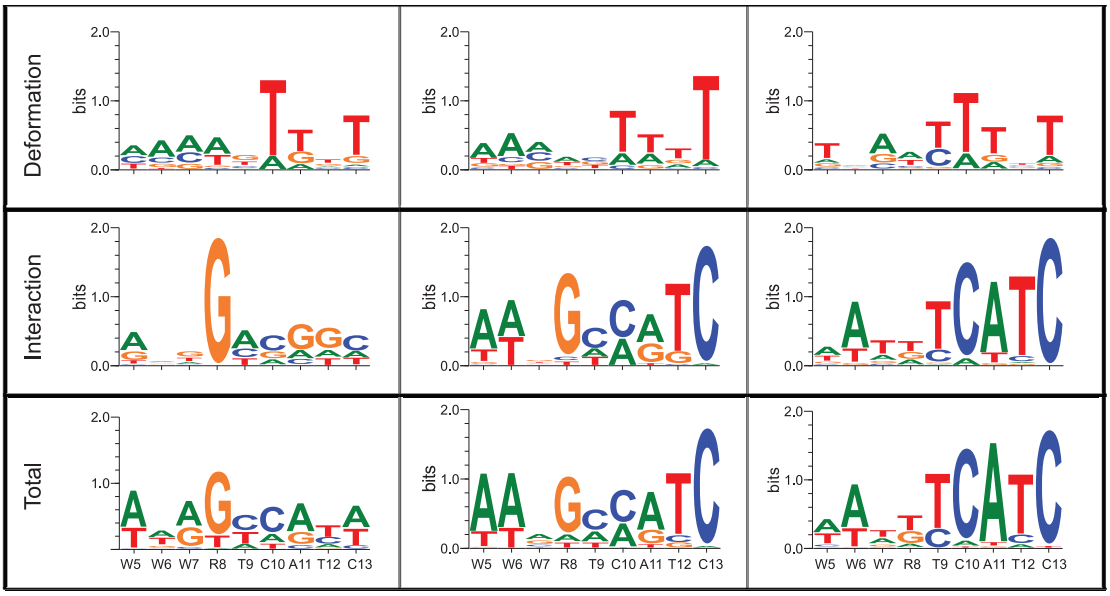


Figure S3. Contribution of DNA deformation (top) and protein-DNA interaction (center) to the overall PWMs for the conformational clusters CL1 (left), CL2/4 (center) and CL3 (right)

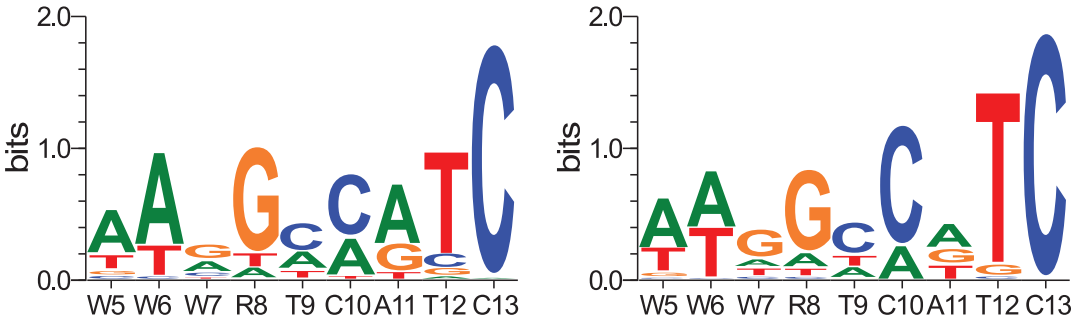


Figure S4. PWMs for the conformational clusters CL2 (left) and CL4 (right)

References

1. Rupert, P.B., Daughdrill, G.W., Bowerman, B. and Matthews, B.W. (1998) A new DNA-binding motif in the Skn-1 binding domain–DNA complex. *Nat Struct Biol,* **5**, 484-491.
2. Fjose, A., McGinnis, W.J. and Gehring, W.J. (1985) Isolation of a homoeo box-containing gene from the engrailed region of Drosophila and the spatial distribution of its transcripts. .
3. Scott, M.P., Weiner, A.J., Hazelrigg, T.I., Polisky, B.A., Pirrotta, V., Scalenghe, F. and Kaufman, T.C. (1983) The molecular organization of the Antennapedia locus of Drosophila. *Cell,* **35**, 763-776.
4. Bohmann, D., Bos, T.J., Admon, A., Nishimura, T., Vogt, P.K. and Tjian, R. (1987) Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP-1. *Science,* **238**, 1386-1392.
5. Hinnebusch, A.G. (1984) Evidence for translational regulation of the activator of general amino acid control in yeast. *Proc Natl Acad Sci U S A,* **81**, 6442-6446.

Article 2

Nucleic Acids Research Advance Access published September 21, 2016

doi: 10.1093/nar/gkw841

Protein–DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors

Loïc Etheve, Juliette Martin and Richard Lavery*

MMSB UMR 5086 CNRS/University of Lyon I, Institut de Biologie et Chimie des Protéines, 7 passage du Vercors, Lyon 69367, France

Received July 7, 2016; Revised September 12, 2016; Accepted September 13, 2016

ABSTRACT

We have studied the dynamics of three transcription factor–DNA complexes using all-atom, microsecond-scale MD simulations. In each case, the salt bridges and hydrogen bond interactions formed at the protein–DNA interface are found to be dynamic, with lifetimes typically in the range of tens to hundreds of picoseconds, although some interactions, notably those involving specific binding to DNA bases, can be a hundred times longer lived. Depending on the complex studied, this dynamics may or may not lead to the existence of distinct conformational substates. Using a sequence threading technique, it has been possible to determine whether DNA sequence recognition is sensitive or not to such conformational changes, and, in one case, to show that recognition appears to be locally dependent on protein-mediated cation distributions.

INTRODUCTION

We recently carried out a molecular dynamics study of the interface dynamics of the complex between SKN-1, a transcription factor and its DNA cognate binding site (1). We found that arginine-phosphate salt bridges broke and reformed regularly with lifetimes of the order of hundreds of picoseconds. This result was in line with recent nuclear magnetic resonance (NMR) experiments (2–4), coupled with computational studies (5), showing that lysine-phosphate salt bridges were also dynamic within protein–DNA complexes. However, in the case of our work, we found that some arginine side chains could oscillate between backbone and base binding sites. By identifying the distinct conformational substates associated with these movements, and using a sequence threading technique to analyze binding selectivity, we found that different arginine-linked substates could explain different parts of the experimentally observed consensus binding sequence. It thus appeared that recognition, at least with this particular transcription factor, was the result of a dynamic process.

In order to test whether this result can be generalized, we have now extended our study to three other transcription factor–DNA complexes involving both major and minor groove binding and different degrees of protein-induced DNA deformation. First, we chose the ubiquitous TATA-box binding protein (TBP) that, as part of the TFIID factor, initiates the assembly of the transcriptosome on core promoters. TBP binds in the minor groove of the double helix via an extended β-sheet, producing a large DNA deformation, opening the minor groove, unwinding the double helix, bending it away from the protein and creating kinks at either end of the binding site due to the partial intercalation of phenylalanine residues (6). For the second protein, we chose sex-determining Y protein (SRY) that again binds in the minor groove, but this time via an α-helix and a flexible cationic tail (7). SRY binding, that also includes the partial intercalation of an isoleucine residue, again deforms DNA, but less extensively than TBP. The third protein chosen was the P22 c2 repressor (8). P22 is a homodimer that binds at two major groove sites separated by one turn of the double helix. P22 binding produces limited DNA deformation, but includes the close packing of DNA methyl groups around a valine residue within each half-site.

In addition to the differences already mentioned, our chosen proteins differ in the extent of their direct, and presumably sequence-specific, contacts between amino acid side chains and DNA bases. There are relatively few such contacts with TBP, only one in each half-site of P22, but many with SRY. This suggests that the balance between so-called direct and indirect recognition will vary significantly for these three proteins.

We have carried out microsecond-scale simulations on each of these complexes in water at a physiologically reasonable salt concentration and also performed reference simulations on the corresponding, isolated DNA oligomers. The results show that most protein–DNA contacts fluctuate on a sub-nanosecond timescale. A subset of these contacts oscillate between different DNA target sites, and a further subset

*To whom correspondence should be addressed. Tel: +33 4 7272 2637; Fax: +33 4 7272 2604; Email: richard.lavery@ibcp.fr

are able to modulate the optimal DNA binding sequence of the protein.

While the sequence-threading technique we previously developed is an essential part of this study, for computational reasons it cannot treat explicit water molecules, or ions, at the protein–DNA interface (9,10). For the cases studied here this restriction actually helps in determining whether such 'environmental factors' indeed play an important role in the recognition mechanisms of the proteins we have studied.

MATERIALS AND METHODS

Starting conformations

The initial construction of our three chosen protein–DNA complexes was based on coordinates drawn from the Protein Data Bank (11): the crystal structure of human TBP at a resolution of 1.9 Å (1CDW; (6)), the NMR structure of human SRY (1J46; (7)) and the crystal structure of lambdoid bacteriophage P22 c2 repressor (P22) at a resolution of 1.53 Å (2R1J; (8)). The internal/helicoidal variable modeling program JUMNA (12) was used to construct complexes within oligomers containing the experimentally studied binding sequences, maintaining the conformation of the protein and of the protein–DNA interface. We used a 16-mer for TBP, a 14-mer for SRY and a 20-mer for the dimeric P22. Their exact sequences are shown and discussed in the results section.

MD simulations

The initial conformations of the protein/DNA complexes were solvated with SPC/E water molecules (13). Periodic boundary conditions were imposed using a truncated octahedral box, ensuring a solvent shell of at least 10 Å around the solute. The solute was neutralized with potassium ions and then sufficient K^+/Cl^- ion pairs were added to reach a concentration of 150 mM. The ions were initially placed at random, but at least 5 Å from DNA and 3.5 Å from one another. The resulting systems contained between 9800 and 11 200 water molecules, corresponding to a total of 33 456, 32 516 and 37 091 atoms for the TBP, SRY and P22 complexes respectively.

Molecular dynamics simulations were performed with the AMBER 12 suite of programs (14,15) using PARM99 parameters (16) and the bsc0 modifications (17) for the solute and Dang parameters (18) for the surrounding ions. Simulations employed periodic boundary conditions and electrostatic interactions were treated using the particle-mesh Ewald algorithm (19,20) with a real space cutoff of 9 Å. Lennard–Jones interactions were truncated at 9 Å. A pair list was built with a buffer region and a list update was triggered whenever a particle moved by more than 0.5 Å with respect to the previous update.

Each system was initially subjected to energy minimization with harmonic restraints of 25 kcal mol^{-1} $Å^{-2}$ on the solute atoms. The system was then heated to 300 K at constant volume during 100 ps. Constraints were then relaxed from 5 to 1 kcal mol^{-1} $Å^{-2}$ during a series of 1000 steps of energy minimization (500 steps of steepest descent and 500 steps of conjugate gradient) followed by 50 ps of equilibration with restraints of 0.5 mol^{-1} $Å^{-2}$ and 50 ps without solute restraints. The 500 ns production simulations (or 1 μs in the case of P22) were carried out at constant temperature (300 K) and pressure (1 bar) with a 2 fs time step. During these simulations pressure and temperature were maintained using the Berendsen algorithm (21) with a coupling constant of 5 ps and SHAKE constraints were applied to all bonds involving hydrogens (22). Conformational snapshots were saved for further analysis every ps. For comparison purposes, the isolated DNA oligomers from each complex were also simulated alone using an identical protocol, creating a second set of 500 ns trajectories.

Conformational and environmental analysis

Average DNA conformation, DNA conformational fluctuations and ion distributions around the protein/DNA complexes during the MD simulations were analyzed with the Curves+ program (23) and the Canal and Canion utilities (https://bisi.ibcp.fr/tools/curves_plus/). In addition to intra-bp, inter-bp and bp-axis parameters Curves+ can calculate groove geometries and the overall bend of a helical axis. Note that the values of axis bend presented here ignore the terminal base pairs of the oligomers since these often suffer from local deformations.

Using the recently developed Curves+ ion analysis approach, based on describing ion positions using curvilinear helicoidal coordinates with respect to the DNA helical axis, it was notably possible to calculate average ion molarities and ion populations within the DNA grooves (24,25). As in our earlier work, the groove limit was set at a radius of 10.25 Å from the DNA helical axis (the average radial position of the backbone phosphorus atoms), while the angular limits defining the major and minor grooves were determined by the average position of the sugar C1' atoms. Spatial ion densities, and all molecular graphics, were generated using Chimera (26,27).

Lastly, hydrogen bond and salt bridges were analyzed using AMBER Tools (28). We chose to limit our analysis to direct interactions by applying a distance cut-off of $\leq$3.5 Å between the relevant heavy atoms and an angle cut-off of $\geq$135° at the intervening hydrogen atom. These interactions are characterized by the percentage of the trajectory during which they are observed (% presence) and by their average lifetimes, which are calculated ignoring interruptions in the interaction that last less than 1 ps. As shown in Supplementary Figure S1, for the case of salt bridges, longer range interactions, notably in the range 3.5–6.0 Å (presumably involving a bridging water molecule (5)) exist and even more distant interactions (generally involving concurrent interactions with a neighboring nucleic acid residue) can also occur. It is however difficult to characterize these indirect interactions with a simple distance criteria and they have been excluded from the present analysis.

Clustering the MD trajectory

In order to identify conformational clusters within the MD trajectory, we began by extracting snapshots every 200 ps. Since we were principally interested in the evolution of

the protein–DNA binding specificity, we characterized each snapshot by counting the number of contacts between the protein and the DNA bases. Each contact between heavy atoms scored 1 for distances r_{ij} below 4 Å (using shorter distances would result in many transient 'breaks' that add noise to the analysis). In order to further increase the robustness, we used a buffer zone from 4 Å to 5 Å over which the score was modulated with a sigmoidal function $s(i,j)$ of the distance r_{ij} between the atoms i and j:

$$s(i,j) = \frac{1}{1 + e^{10*(r_{ij}-4.5)}}$$

This analysis yielded a rectangular *Na* amino acid by *Nb* base matrix for each snapshot. The overall distance d(x,y) between any two such matrices **x** and **y** was then calculated using the Manhattan algorithm (29).

$$d(x,y) = \sum_{k=1}^{Na}\sum_{l=1}^{Nb} |x_{kl} - y_{kl}|$$

Next, the Ward agglomerative hierarchical clustering method (30–32) was used to classify the different snapshots into groups by minimizing the variance within each cluster and increasing the weighted squared distance between cluster centers. The distance matrix and cluster representations were obtained using the R software package (33).

When an MD trajectory shows the existence of conformational substates, we create new clustered maps for each amino acid at the interface. These component maps indicate which residue, or residues, are responsible for the observed changes and, in the case that several residues are involved, indicate whether these residues act together or separately to create conformational substates.

Binding specificity analysis

Binding specificity was determined for any chosen snapshot from the MD trajectory (after a brief Cartesian coordinate energy minimization to remove bond length and base plane deformations) using the so-called ADAPT sequence threading approach (9,10) implemented within the JUMNA program (12). This consists of calculating the complex formation energy of the protein–DNA complex for all possible DNA base sequences and then deriving a position weight matrix (PWM) from the best binding sequences. In order to do this, it is necessary to thread all possible base sequences into the binding site of the DNA oligomer within the complex, adapting the protein–DNA interface in each case using internal coordinate energy minimization. Minimization was performed with the same AMBER parameterization used for the MD simulations, but replacing the explicit solvent and ion shell with a simple continuum model using a sigmoidal distance-dependent dielectric function and reduced phosphate charges (10). In parallel, an identical base sequence is threaded into the average conformation of the isolated DNA oligomer and energy minimization is again performed. Finally, another energy minimization is performed for the isolated protein (with flexibility limited to the side chains included within the interface cutoff distance, see below).

Subtracting the isolated DNA oligomer and protein energies from the protein–DNA complex energy yields the complex formation energy, which can be further analyzed in terms of two components: the DNA deformation energy and the protein–DNA interaction energy. In this work, we used ADAPT to scan 8, 10 and 20 bp belonging to the binding sites of the TBP, SRY and P22 complexes respectively (this implies analyzing binding for between 6.5×10^4 and 1.1×10^{12} potential base sequences). ADAPT calculations achieve this task by a divide-and-conquer technique, breaking each sequence down into overlapping 5 bp fragments and thus dramatically reducing the total number of calculations for the complex and for the isolated DNA oligomer, without significant loss of accuracy (10). Protein flexibility was also limited to side chains within 20 Å of the protein–DNA interface. The energies resulting from this analysis were converted into PWMs using the WebLogo software (34). Finally, by analyzing the binding specificity derived from the sequence-dependent DNA deformation energy, or from the sequence-dependent protein–DNA interaction energy we could also describe binding specificity in terms of its so-called indirect and direct components.

We remark that the utility programs associated with ADAPT have been extended to be able to derive a single PWM from a number of MD snapshots belonging to a given conformational substate (in the present work, between 5 and 10 snapshots per substate, depending on its overall duration). In this case, ADAPT calculations were based on sequence-dependent energy differences with respect to the minimum energy for each snapshot, enabling us to overcome sequence-independent energy changes mainly caused by the necessary simplification of the electrostatic calculations (which rely on a rudimentary implicit solvent representation). Using this approach it was possible to describe the sequence selectivity of each of the conformational substates detected by the cluster analysis and to compare this to the consensus selectivity for the entire MD simulation, or to experimental binding data.

RESULTS AND DISCUSSION

TATA-box binding protein (TBP)

We chose to study human TBP as a casebook example of a protein binding in the minor groove of DNA, producing significant DNA deformation (6). In this case, protein binding causes a wide opening of the minor groove, a strongly reduced twist and ∼60° bending away from the protein. TBP interacts with DNA via an extensive β-sheet covering 8-bp site (T5 → G12) within the 16 bp oligomer we studied. Despite this extensive contact surface, the MD simulations confirm that the protein establishes relatively few hydrogen bonds with the DNA bases, only two with the Watson strand and three with the Crick strand involving asparagine or threonine side chains binding to the bases A8, A9 and T8' → T10' (see Table 1). These are complemented by eight arginine-phosphate salt bridges involving seven phosphate groups, three in the Watson strand and four in the Crick strand, and three serine-phosphate hydrogen bonds (with G12, A5', A7', see Figure 1). For comparison, the contacts found in the crystal structure are shown in Supplementary Figure S2A. Note that, by convention, phosphate contacts

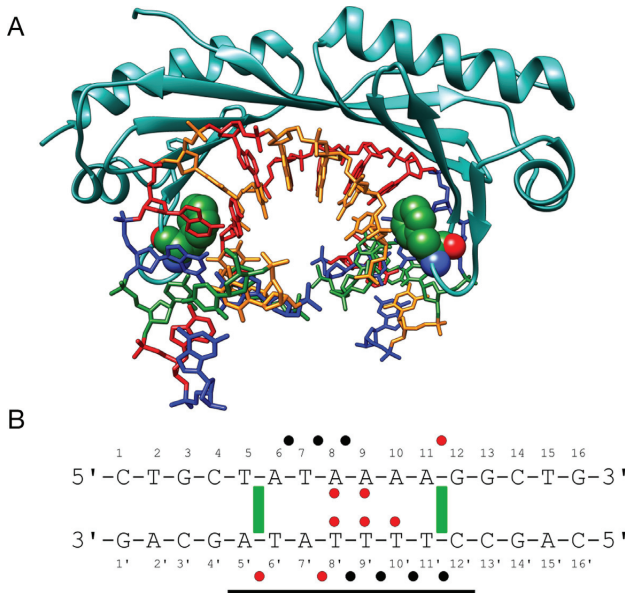


Figure 1. (**A**) Structure of the human TBP/DNA complex (6). Two phenylalanine residues (Phe 193 and Phe 284, green spheres) are partially intercalated at the T5pA6 and A11pG12 steps. DNA is colored according to sequence (A = red, T = orange, G = blue, C = green). (**B**) DNA sequence used for the MD simulations with the binding site delimited by a horizontal black line. The 'Watson' strand of the oligomer is numbered 1–16 in the 5′-3′ sense and the 'Crick' strand is numbered 1′-16′ in the 3′-5′ sense. Salt bridges, hydrogen bonds and important apolar interactions observed during the MD simulations are indicated by black dots, red dots and green rectangles respectively.

refer to the 5′-phosphate groups of the cited nucleotides. Table 1 shows that, as we found earlier for SKN-1 (1), while all but one of the interactions found experimentally are observed, the MD trajectory leads to new interactions, mainly linked to the dynamics of amino acid side chains that enables them to contact several neighboring phosphates or bases.

Given the paucity of amino acid-base interactions, TBP's mode of recognition is expected to involve a significant indirect component relying on the sequence-dependent nature of the induced DNA deformation. In addition to the large-scale bending and twisting components already mentioned, TBP binding also includes the partial intercalation of phenylalanine residues which lead to the formation of kinks at either end of the binding site at T5pA6 and A11pG12, locally increasing the rise and the positive roll. The consensus binding sequence of TBP is TATAWAWR (where W implies A/T and R implies A/G), although some dependence on the flanking base sequences has also been demonstrated (35). In line with a dominantly indirect recognition mechanism, it is also possible to favor TBP binding by appropriately pre-bending DNA, in order to widen the minor groove at the interaction site (36).

Analysis of our 0.5 μs molecular dynamics simulation shows that the TBP–DNA complex stays relatively close to the crystallographic structure (heavy atom root mean square difference (RMSD) $\leq$ 2Å). DNA remains strongly bent away from the protein by an average of 57° (compared to 66° in the crystal structure and to only 24° in the isolated DNA oligomer, which bends in the same direction as that observed in the complex). The total twist over the binding site is reduced by an average of 85° compared to the free oligomer. Supplementary Figure S3 summarizes the average conformation of the binding site in terms of twist, rise, roll and groove width.

From a dynamic point of view, the multiple salt bridges established between TBP and DNA lead to restricted phosphate mobility typically reducing values in the free oligomer (root mean square fluctuation (RMSF) ≈1.75 Å) by around 25% (Supplementary Figure S3). The salt bridges have a percentage presence ranging from 15 to 97% of the simulated trajectory (ignoring changes in the closest interacting atom pairs, see Supplementary Table S1). Those with A8 and T9' are the longest lived, while those with A9 and T10' are the shortest. As shown in Table 1, specific salt bridge interactions, and also specific hydrogen bonding across the protein–DNA interface have lifetimes that are typically around 100 ps, although some may persist for many nanoseconds.

As shown in Supplementary Figure S4, TBP binding modifies the ion distribution around DNA. The minor groove ion density is not surprisingly reduced to zero throughout the binding site. However, we also see changes in the narrow major groove, where there is an increase in ion density for the base pairs belonging to the binding site and also the appearance of a particularly strong ion density (5x that in the isolated oligomer) at G12pG13. For 75% of the trajectory there is a K^+ ion resident at this site and for 15% the site is occupied by two ions (the equivalent results for the isolated oligomer being 48 and 5%).

When we use the interface analysis protocol, previously developed for our study of the SKN-1/DNA complex, we do not see any of the substates related to the amino acid side chain dynamics that we observed with the former protein. The TBP binding interface turns out to be very stable. Although both salt bridges and hydrogen bonds to the bases both break repeatedly during the simulation (see the lifetimes listed in Table 1), they generally reform with the same nucleotides (although the donor and acceptor atoms may change, as shown in the table). The amino acid-base interaction matrix is very smooth and cannot be clustered (data not shown). Consequently, we can generate an overall PWM logo by studying a set of 10 snapshots drawn randomly from the trajectory.

Sequence-threading using ADAPT on each snapshot, followed by averaging, leads to the overall logo shown in Figure 2. This result is in good agreement with the experimental result from JASPAR (37), in terms of the base recognition along the binding site and in terms of the overall information content (10.1 for the MD snapshots versus 9.3 for JASPAR, with an overall correlation coefficient of 0.87). Dividing the ADAPT results into indirect (DNA deformation) and direct (protein–DNA interaction energy) components confirms that indirect recognition plays a major role in this complex (as we saw in earlier work based on a sequence threading analysis applied to the experimental structure of the complex (9,10)). However, the direct interactions remain critical in establishing the overall consensus, particularly toward the 3′-end of the binding site, where the majority of amino acid-base hydrogen bonds are indeed formed

Table 1. TBP interactions with the DNA backbone and bases showing percentage presence during the 0.5 μs MD trajectory and the average lifetime (ps)

Protein	Backbone	%pres.	Lifetime	Protein	Base	%pres.	Lifetime
R192(NH1)	T10'(O1P)	54	100	N163(ND2)	T8'(O2)	94	1150
R192(NH2)	T10'(O1P)	48	120	N163(ND2)	T9'(O2)	77	125
R192(NH2)	T10'(O2P)	30	80	T218(OG1)	T10'(O2)	12	25
R192(NH2)	T11'(O3')	13	20	T218(OG1)	T9'(O2)	7	20
R192(NH2)	T11'(O1P)	16	165	N253(ND2)	A8(N3)	60	65
R192(NE)	T11'(O1P)	24	145	N253(ND2)	A9(N3)	78	125
R199(NH2)	T9'(O1P)	80	150	T309(OG1)	A8(N3)	56	50
R199(NE)	T9'(O1P)	74	160				
R199(NH2)	T9'(O2P)	18	25				
R199(NH2)	T10'(O3')	25	25				
R204(NE)	T8'(O1P)	13	55				
R204(NH2)	T8'(O1P)	21	155				
R204(NH1)	T8'(O1P)	18	120				
R204(NH2)	T8'(O2P)	4	24				
R290(NH2)	T7(O3')	33	30				
R290(NH2)	T7(O1P)	4	95				
R290(NH1)	A8(O1P)	92	1140				
R290(NH2)	A8(O1P)	58	200				
R295(NH1)	A9(O1P)	11	100				
R295(NH2)	A9(O1P)	8	100				
T206(OG1)	T9'(O1P)	78	1300				
S212(OG)	G12(O1P)	96	3950				
T218(OG1)	T9'(O4')	1	-				
S257(OG)	A7'(O3')	14	45				
S303(OG)	A5'(O1P)	98	11 450				

Bold horizontal lines indicate the separation between salt bridges (above) and hydrogen bonds (below). Interactions in black are common to the experimental structure and the MD trajectory, those in red only occur in the MD trajectory and those in green only occur in the experimental structure.

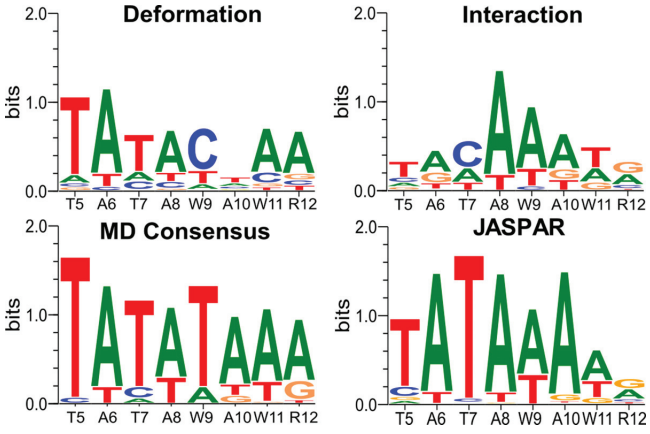


Figure 2. PWM logos for the TBP/DNA complex obtained from the analysis of the MD trajectory. Top left: DNA deformation energy (indirect recognition). Top right: DNA–protein interaction energy (direct recognition). Bottom left: overall recognition. Bottom right: experimental logo from the JASPAR database. Each panel also shows the experimental consensus along the abscissa (W ≡ A/T, R ≡ A/G).

(see Figure 1 and Table 1). In conclusion, TBP presents a much simpler case than our earlier study of the protein skinhead 1, SKN-1. Individual protein–DNA interactions regularly break and reform (typically on a 0.1 ns timescale), and sometimes oscillate between neighboring nucleotides, but these dynamics do not influence the recognition mechanism that can be understood using a single conformational state.

Sex-determining region Y protein (SRY)

SRY determines the male sex in humans and belongs to the Sry-related HMG box (SOX) gene family. It binds in the minor groove of DNA, via an α-helix at the 3′-end of the binding site and via a flexible cationic C-terminal tail (with four lysines and three arginines in proximity to DNA) at the 5′-end. It recognizes a 7 bp binding site with a weak consensus sequence WAACAAT. Our simulations were carried out using a 14 bp oligomer, with a centrally positioned site G4 → A10 (GCACAAA) based on the sequence used in the NMR structure determination (7) (see Figure 3). Note that the α-helix contains a conserved isoleucine that partially intercalates at the ApA step within the CAAA end of the binding site (numbered A8pA9 with the 14 bp DNA oligomer we studied). SRY makes extensive hydrogen bonds with base sites, five in the Watson strand and seven in the Crick strand, as well as numerous arginine-phosphate salt bridges, seven in the Watson strand and six in the Crick strand (see Figure 3 and Table 2. For comparison, Supplementary Figure S2B shows the experimentally observed contacts).

As for TBP, the minor groove binding of SRY distorts DNA. The double helix bends significantly away from the protein by an average of 61° during the simulations (43° in

Table 2. SRY interactions with the DNA backbone and bases showing percentage presence during the 0.5 μs MD trajectory and the average lifetime (ps)

Protein	Backbone	%pres.	Lifetime	Protein	Base	%pres.	Lifetime
R4(NH2)	A8(O1P)	21	135	R7(NH1)	C7(O2)	92	780
R4(NH1)	A8(O1P)	32	185	R7(NH2)	T6'(O2)	58	50
R4(NH2)	A8(O2P)	20	75	R7(NH1)	T6'(O2)	97	3590
R4(NH1)	A8(O2P)	14	75	N10(ND2)	T8'(O2)	97	27940
K6(NZ)	A9(O2P)	56	45	N10(ND2)	G7'(N3)	81	175
K6(NZ)	A9(O1P)	27	135	S16(OG)	A9(N3)	8	165
R17(NE)	A10(O1P)	53	55	R20(NH2)	T9'(O2)	13	100
R17(NH2)	A10(O1P)	71	205	N32(ND2)	A10(N3)	65	60
R17(NE)	A9(O3')	46	45	N32(ND2)	C11(O2)	94	1240
R21(NE)	C11(O1P)	64	165	S33(OG)	G11'(N3)	<1	-
R21(NE)	C11(O2P)	14	20	S36(OG)	T10'(O2)	100	128 735
R21(NH2)	C11(O2P)	65	375	Y74(OH)	A6(N3)	54	490
R21(NH2)	C11(O1P)	20	120	Y74(OH)	G5'(N3)	26	60
R31(NH2)	C14(O1P)	7	730	R78(NH1)	A3'(N3)	27	185
K37(NZ)	T9'(O1P)	45	75	R78(NH2)	A3'(N3)	14	60
K44(NZ)	T8'(O1P)	22	75				
K51(NZ)	G7'(O1P)	<1	-				
R66(NH2)	G7'(O1P)	9	360				
K73(NZ)	C4'(O1P)	28	135				
R75(NE)	A3'(O1P)	10	125				
R75(NH1)	A3'(O1P)	20	310				
R75(NH2)	A3'(O2P)	11	315				
R77(NE)	C7(O1P)	54	220				
R77(NE)	C7(O2P)	20	50				
R77(NH2)	C7(O2P)	44	175				
R77(NH2)	C7(O1P)	16	185				
R78(NH1)	G2'(O1P)	8	95				
K79(NZ)	A6(O2P)	21	70				
K79(NZ)	A6(O1P)	7	45				
K81(NZ)	C5(O1P)	12	8				
R7(N)	A8(O1P)	33	180				
R7(N)	A8(O5')	16	250				
R7(NH2)	G5'(O4')	11	35				
N10(ND2)	T8'(O4')	<1	-				
N32(N)	A12(O4')	73	170				
R77(N)	A6(O3')	76	150				
R77(N)	C7(O1P)	37	50				
N32(ND2)	C11(O4')	<1	-				
W43(NE1)	G7'(O1P)	88	290				
W43(NE1)	T8'(O3')	17	15				
Q62(NE2)	T6'(O3')	<1	-				
Q62(NE2)	G5'(O1P)	14	55				
K79(N)	A6(O1P)	<1	-				

Bold horizontal lines indicate the separation between salt bridges (above) and hydrogen bonds (below). Interactions in black are common to the experimental structure and the MD trajectory, those in red only occur in the MD trajectory and those in green only occur in the experimental structure.

the NMR structure and 20° in the isolated oligomer). The minor groove is widened by roughly 6 Å where the α-helix contacts DNA at the 3′-end of the binding site and is locally unwound by 41°. We also see an increased rise (5 Å) and positive roll (45°) at the isoleucine intercalation site. Supplementary Figure S5 summarizes the conformational characteristics of the SRY complex.

Also as noted for TBP, salt bridge formation reduces the dynamics of the phosphodiester backbones within the binding site as judged by the phosphate RMSF values which drop from an average of 1.75 Å to 1.25 Å (see Supplementary Figure S5). The salt bridges on the Watson strand generally have a longer percentage presence, and often multiple arginine or lysine interactions, compared to those of the

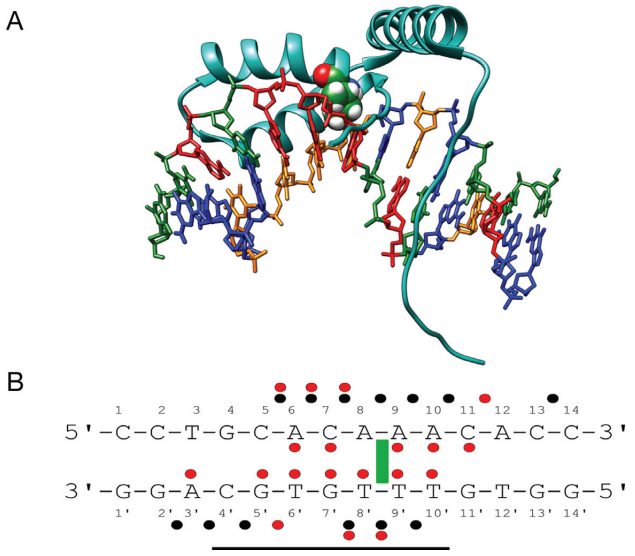


Figure 3. (**A**) Structure of the human SRY/DNA complex (7). Isoleucine 13 (green spheres) is partially intercalated at the A8pA9 step. DNA is colored according to sequence (A = red, T = orange, G = blue, C = green). (**B**) DNA sequence used for the MD simulations with the binding site delimited by a horizontal black line. The 'Watson' strand of the oligomer is numbered 1–14 in the 5′-3′ sense and the 'Crick' strand is numbered 1′-14′ in the 3′-5′ sense. Salt bridges, hydrogen bonds and important apolar interactions observed during the MD simulations are indicated by black dots, red dots and green rectangles respectively.

Crick strand (see Supplementary Table S1). The two outlying interactions (R31-C14 and R78-G2') are both present for less than 10% of the trajectory. Individual salt bridge and hydrogen bond interactions at the protein–DNA interface typically have lifetimes of the order of 100 ps, but several specific hydrogen bonds (notably those with C11, T6', T8' and T10') persist for many nanoseconds (see data in Table 2). As for TBP, the interface dynamics adds many contacts to those seen experimentally (red lines in Table 2) with a significant increase in the number of salt bridges and hydrogen bonds, where most of the amino acids involved are able to contact several nucleotides within (or adjacent to) the binding site.

The extensive SRY–DNA interface understandably restructures the counterion distribution around DNA, virtually eliminating K^+ ions from the minor groove. The major groove ions are less perturbed, although a strong binding site at G4pC5 is significantly reduced in the complex, while ion density at A9pA10 opposite the SRY α-helix (and the widened minor groove) increases (see Supplementary Figure S6).

We now consider the impact of SRY/DNA interface dynamics on recognition by first calculating the amino acid-base contact matrix for the trajectory. These results make it clear that SRY binding involves several distinct conformational substates. In order to understand which amino acids are playing a major role we calculated the contact matrices for each residue involved in the SRY/DNA interface. This analysis showed that two residues belonging to the flexible C-terminal tail, tyrosine 74 (Y74) and arginine 78 (R78), were the key players. Their individual contact matrices taken together explain the major variations seen in the overall interface matrix (see Figure 4).

We begin by considering Y74. This side chain can adopt three states: interacting as a hydrogen bond donor to A6(N3) (54% of the trajectory), as a donor to G5'(N3) (26%), or positioned to interact in a bidentate manner with A6(N3) and G5'(N2) (20%). Sequence threading shows that these conformational changes have a relatively small impact on recognition since an A in position 6 is favored whatever the state of Y74 (see Figure 5). However, a preference for T in position 4 (at the 5′-end of the binding site) only occurs when Y74 is interacting with the adjacent base at position 5. Similarly, T/A recognition in position 10 is diminished when Y74 is bound in a bidentate manner (although how these effects are coupled is not clear). For R78, we again find three substates: interacting with the backbone phosphate group of G2' (8% of the trajectory), interacting with A3'(N3) (27%) or not interacting directly with DNA (65%). Since the bases contacted by R78 flank the 5′-end of the SRY binding site, this side chain has little impact on the calculated consensus, although we note that the weak preference of C at position 7 disappears when R78 does not interact with DNA (data not shown).

Looking at the overall consensus derived from the trajectory in Figure 5 we see a reasonable agreement with the experimental result with the exception of the stronger experimental C preference at position 7 (information content 6.0 versus a JASPAR value of 8.7, with an overall correlation coefficient of 0.69). It is worth noting that two experimental logos are available for the highly homologous mouse SRY protein (86% homology, with a virtually identical DNA-binding interface based on sequence alignment) and one of these shows a dominant recognition of thymine at this position 7 as in our PWM (38). It is also interesting to note that although the simulations involved an oligomer containing G4-C5, the consensus derived by sequence threading shows no preference for these bases, and rather favors the experimental weak preference for A/T. This implies that the conformational optimization carried out for each overlapping fragment of the complex during threading is capable of correctly adapting the protein–DNA interface and is not biased by the DNA sequence used for the simulation.

Looking at the direct and indirect components of the MD-derived consensus shows, not unreasonably, that direct interactions dominate the recognition at the 5′-end, where the C-terminal tail binds. In contrast, indirect, deformation-related recognition, dominates where the α-helix deforms the minor groove at the 3′-end and both mechanisms play a role in the center of the binding site. In conclusion, while SRY binding does involve conformational substates, these play a relatively minor role in determining the base sequence recognized by the protein.

Bacteriophage P22 c2 repressor protein (P22)

P22 is a homodimer that is involved in controlling the lysogenic pathway of the lambdoid P22 bacteriophage. Each monomer binds to DNA via an α-helix within a major groove half-site, the two half-sites being separated by one turn of the DNA double helix (8). P22 binds to six naturally occurring operator sequences having an overall consensus

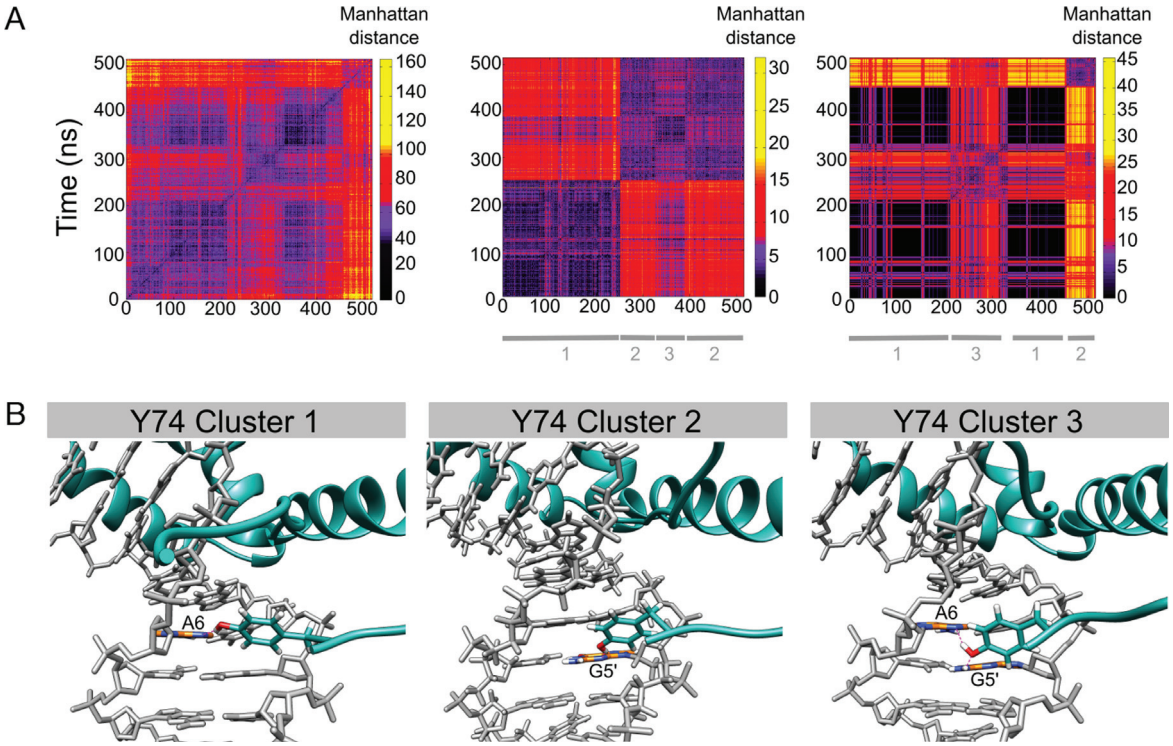


Figure 4. Clustering snapshots from the 500 ns MD trajectory of the SRY/DNA complex. (**A**) Manhattan distance matrix for all protein–DNA base contacts (left), for tyrosine 74 (center) and for arginine 78 (right). The vertical scale shows increasing distances (black → yellow). (**B**) Alternative orientations observed for tyrosine 74: bound to A6(N3), bound to G5'(N3), bidentate interactions with A6(N3) and G5'(N2).

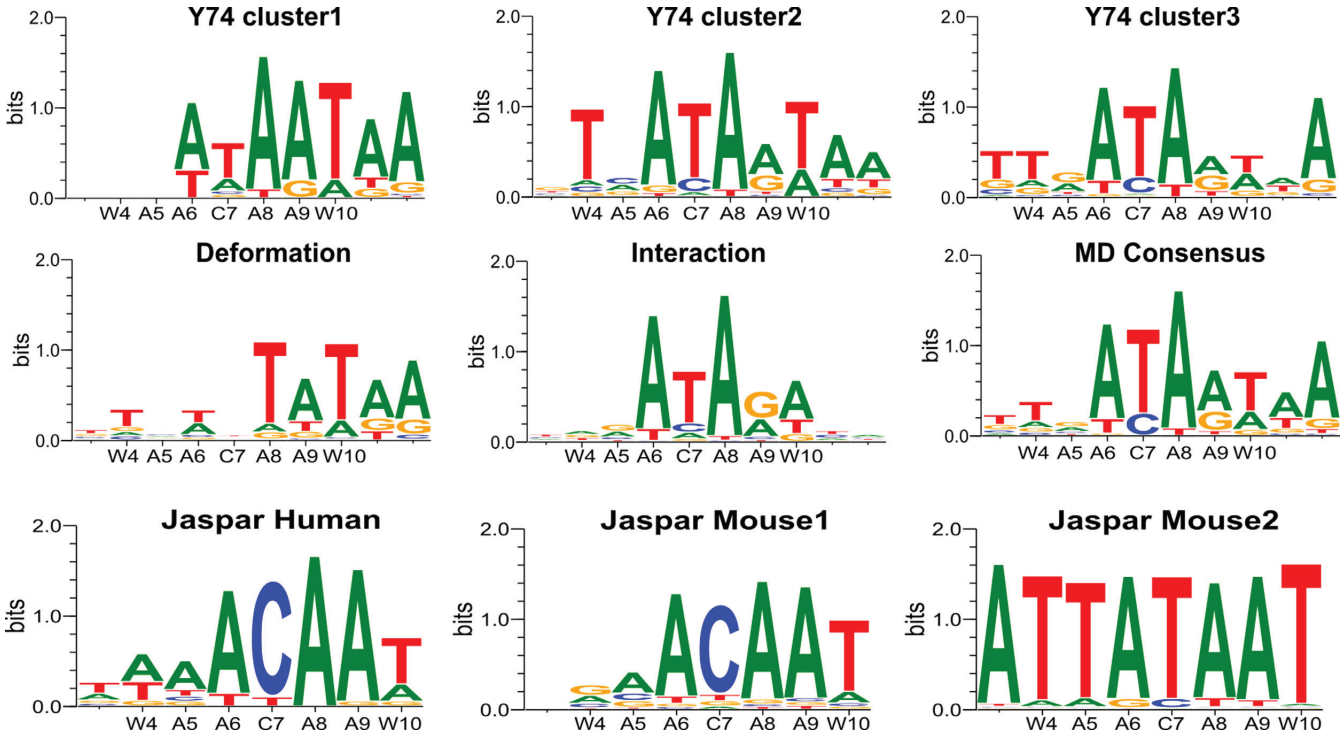


Figure 5. PWM logos for the SRY/DNA complex obtained from the analysis of MD trajectory. Tyrosine 74 dynamics generate three substates: binding to A6 (top left), binding to G5' (top center), bidentate binding to A6/G5' (top right). Components of recognition: indirect from DNA deformation energy (middle left), direct from DNA–protein interaction energy (middle center), overall (middle right). Experimental PWM logos from the JASPAR database: human SRY (bottom left), mouse SRY (bottom center and right). Each panel also shows the experimental consensus along the abscissa (W ≡ A/T, R ≡ A/G).

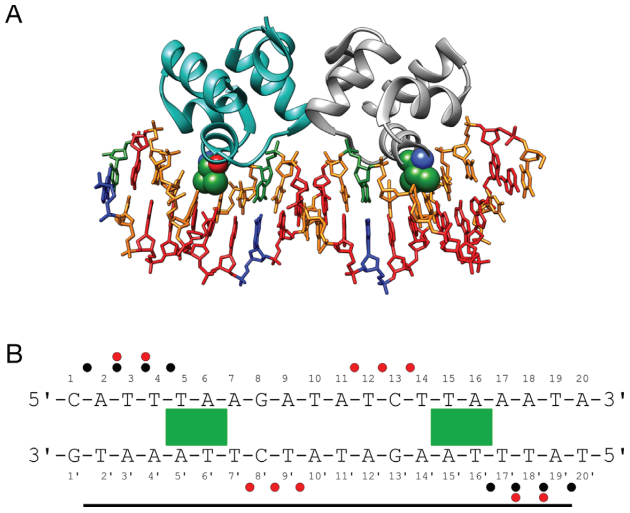


Figure 6. (**A**) Structure of the bacteriophage P22/DNA complex (the two monomers are shown as blue and gray ribbons) (8). Valine 33 from each monomer interacts with the thymine methyl groups of the T4-A7 and T14-A17 base pairs. DNA is colored according to sequence (A = red, T = orange, G = blue, C = green). (**B**) DNA sequence used for the MD simulations with the binding site delimited by a horizontal black line. The 'Watson' strand of the oligomer is numbered 1–20 in the 5′-3′ sense and the 'Crick' strand is numbered 1′-20′ in the 3′-5′ sense. Salt bridges, hydrogen bonds and important apolar interactions observed during the MD simulations are indicated by black dots, red dots and green rectangles respectively.

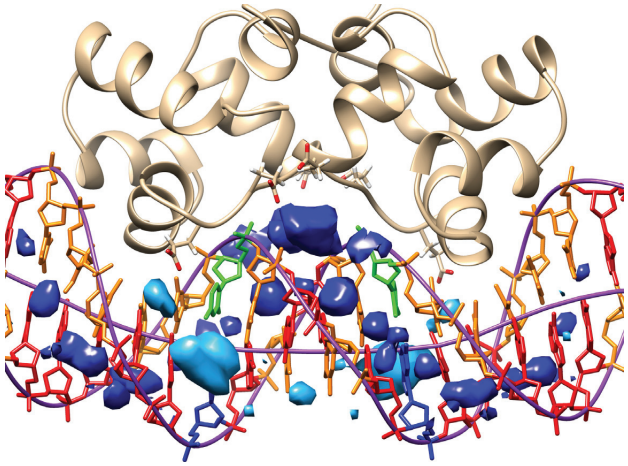


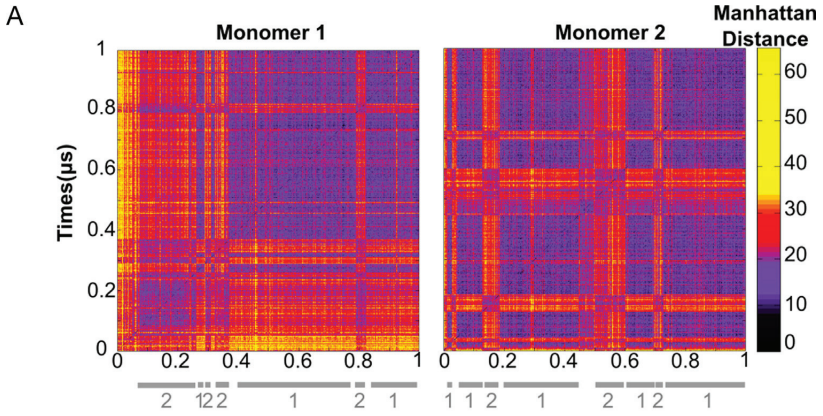
Figure 7. Average K^+ distribution in the minor (dark blue) and major (pale blue) grooves of DNA within the P22/DNA complex plotted as 4 M isodensity surfaces. Ions accumulate within the central minor groove (both near the bases and at the entrance to the groove) due to neighboring P22 glutamic acid residues. Strong major groove densities are also seen close to the G8 and C13 base pairs. Nucleotides are color-coded (A = red, T = orange, G = blue, C = green). The backbone pathway and the helical axis from Curves+ are shown in purple.

ATTTAAGATAT**CTTAAAT**, where the bases in bold font are highly conserved. Each α-helix carries a conserved valine residue in close contact with the bases of each half-site. In the crystal structure, the half sites have the sequence TTAAG and they are separated by a central 4-bp fragment ATAT. The minor groove of this fragment faces the protein but is not contacted by it, although four glutamic acid residues (E44 and E48 in each monomer) are close by. Our simulations involved a 20 bp DNA oligomer with the sequence shown in Figure 6. The two half sites are located at positions T4 → G8 and C13 → A17. The important valine residues (V33 in each monomer) contact the steps T5pA6 and T15pA16 and are each surrounded by the four thymine methyl groups of the TTAA segments. During the MD simulation, P22 forms four salt-bridges with each phosphodiester strand of the binding site (versus six in the crystal structure, see Supplementary Figure S2C). Apart from the valine contacts already mentioned, only fleeting contacts are seen with the bases within the binding site (see Table 3).

P22 causes relatively little DNA deformation upon binding. On average, during the 1 μs MD trajectory, DNA is bent by 23° toward the protein (as in the crystal structure), but this is only slightly more than the bend in the free oligomer. Both major and minor grooves are narrowed following protein binding, with the exception of a small broadening of the central major groove. This is not related to bending (which generally has opposite impacts on the major and minor grooves), but to over-twisting the double helix (the twist over the full binding site increasing by 40° compared to the isolated oligomer). This change involves the segments T5-A9 and T12-A17, plus the central T10pA11 step (which exhibits an 8° increase in twist, although the flanking ApT steps are unaffected). Rise is largely unaffected by P22, with the exception of small increases (0.3 Å, coupled with 10° of roll) at the TpA steps contacted by the Val33 residues. These conformational changes are summarized in Supplementary Figure S7.

As for the other cases studied here, protein binding reduces phosphate mobility by roughly 0.5 Å RMSF. However, while this effect is uniform on the Crick strand, the phosphates A6-G8 and A16-A18 on the Watson strand are not affected (see Supplementary Figure S7). The most stable salt bridges are those involving arginines 14 and 20 that are present between 73 and 97% of the trajectory. Those involving arginines 11 and 40 are considerably more labile (see Supplementary Table S1). Both salt bridge and hydrogen bond lifetimes are again of the order of 100 ps, but as already seen for SRY, several backbone hydrogen bonds are much longer lived. Also, as for the other proteins studied, many interactions fluctuate between neighboring backbone sites (see Table 3).

Although P22 binding influences the ion distribution around DNA, the changes in the major groove are relatively small and, surprisingly, the 2.5 M ion densities at G8 and C13, observed in the isolated oligomer, remain after P22 binding (with a K^+ ion resident for 70% of the trajectory) (see Supplementary Figure S8). Interestingly a bound cation was observed experimentally at one of these positions (the other being occupied by a lysine residue) (39). The most important change however occurs for the ApT steps in the central minor groove. Here, we observe a cation density of roughly 15 M with a corresponding probability of 75% for finding a K^+ ion in this zone (see Figure 7). As shown in figure, these ions undoubtedly help to offset the repulsion between the P22 glutamic acid groups and the DNA phosphates (39,40).



C

Clusters	Q37 conformation M1	Q37 conformation M2
Cluster 1	Conformation 2	Conformation 1
Cluster 2	Conformation 1	Conformation 2
Cluster 3	Conformation 1	Conformation 1
Cluster 4	Conformation 2	Conformation 2

Figure 8. Clustering snapshots from the 1 μs MD trajectory of the P22/DNA complex. (**A**) Manhattan distance matrix for all protein–NA base contacts involving each monomer. The vertical scale shows increasing distances (black → yellow). (**B**) Alternative orientations observed for glutamine 37 (Q37): positioned in the major groove (left), bound to the backbone (right). (**C**) summary of the position of the Q37 residues in each cluster.

During 1 μs trajectory, symmetry is largely conserved between the two half-sites in terms of their buried surface areas (613 ± 33 Å and 588 ± 55 Å respectively) and the percentage presence and lifetimes of the P22-DNA contacts. However, independent conformational fluctuations occur at each site. These can be seen in the amino acid side chain-DNA base contact maps shown in Figure 8. Carrying out the residue-by-residue analysis already described enabled us to identify glutamine 37 (Q37) within the interacting α-helix of each P22 monomer as responsible for the main fluctuations in the protein–DNA interface. The interaction of the Q37 residues of each monomer with T7' and T14 seen in the crystal structure, only occurs fleetingly during the MD trajectory (3%). For the rest of the time Q37 binds to the adjacent CpT phosphate group (11%), or has no direct interaction with DNA. Considering the backbone bound or unbound states of the two Gln37 residues leads to four possible substates (Figure 8). Strong recognition of the TTAA half-site motif only occurs when the corresponding Q37 is not bound to the DNA backbone (i.e. for M2 in cluster 1, for M1 in cluster 2 and for both monomers in cluster 3). The loss of recognition occurring during Q37-backbone binding appears to be due both to an overall displacement of the P22 monomer and to the reorientation of the Q37 side chain, reducing favorable apolar interactions with the proximal thymine methyl groups.

However, if we consider the overall MD consensus for P22 shown in Figure 9, we see that although the TTAA sequences interacting with the valine 33 residues are well detected, we see no G/C preference at positions 8 and 13 and no significant sequence preference for the central 4 bp (although there is a very weak A/T selectivity visible at positions 9 and 12). Analyzing this result in the light of existing experimental data is instructive. The central AATT selectivity has been interpreted as indirect recognition resulting from the formation of a B' structure characterized by a narrow minor groove and increased helical twist. While the MD simulation indeed sees such changes, no sequence selectivity occurs. A second recognition factor mentioned in the experimental studies was the probability of cations in the central 'tunnel' region electrostatically favoring A/T base pairs. Although Tl^+/Rb^+ cations were tested as 'visible' substitutes for K^+, no ions were found in the crystal structure (possibly due to substitution by $NH4^+$ cations) (39). The role of elec-

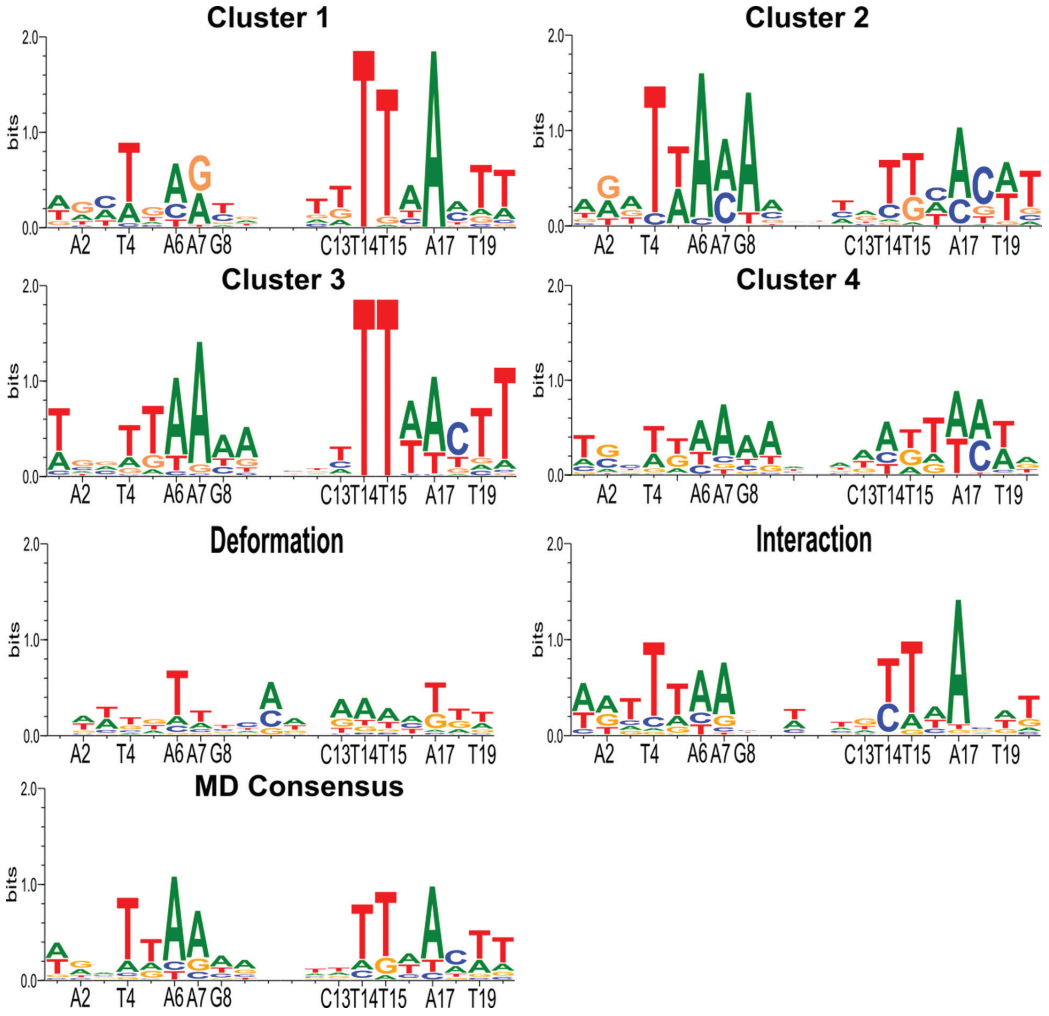


Figure 9. PWM logos for the P22/DNA complex obtained from the analysis of MD trajectory. Glutamine 37 (Q37) dynamics in each monomer (termed M1 and M2) generates four substates: (i) Q37/M1 bound to backbone, (ii) Q37/M2 bound to backbone, (iii) no backbone interactions, (iv) both Q37 bound to backbone. Components of recognition: indirect from DNA deformation energy (third row left), direct from DNA–protein interaction energy (third row right), overall (fourth row left).

trostatics was however supported by the loss of selectivity when either E44 or E48 were substituted by neutral residues (40). The present MD studies further support this analysis, confirming a strong K^+ presence in the central tunnel region with two strong density regions close to the bases in the minor groove that would certainly favor AT base pairs. Similar densities are observed in the major groove close to the positions 8 and 13 which would favor the GC base pairs seen in the native operator sequences. Unfortunately, given the computational effort necessary in ADAPT, the environmental of water and ions can only be represented in a simplified manner and thus the effect of explicit ion densities is not taken into account. This is also true for specific water molecules that have also been proposed as favoring the G/C preference at positions 8 and 13 via bridged hydrogen bonds to E42 (8). Such an effect is also beyond the range of our threading procedure and it is consequently not surprising that we see no selectivity at these positions.

CONCLUSIONS

We have extended our earlier studies of the role of dynamics in protein–DNA recognition to three new transcription factors: TBP, SRY and P22. The results show that the protein–DNA interfaces are dynamic in all three cases. Interactions with the DNA backbones and the DNA bases, involving salt bridges or hydrogen bonds, have lifetimes that are typically of the order of tens to hundreds of picoseconds. This is in line with recent NMR and simulations studies of the dynamics of lysine salt bridges in protein/DNA complexes (5,41,42). A very recent extension of this work shows that, in contrast, arginines bound to guanine in a bidentate manner within a Zn-finger complex are much less dynamic (43). The proteins we have studied here have no such cases, but we do see the almost permanent presence of interactions from a single arginine (R7) to two adjacent bases, and a similar double interaction involving an asparagine (N10) within the SRY complex.

Table 3. P22 interactions with the DNA backbone (A) and the bases (B) showing percentage presence during the 1 μs MD trajectory and the average lifetime (ps) for each monomer (M1 and M2)

A

M1	Backbone	%pres.	Lifetime	M2	Backbone	%pres.	Lifetime
R11(NH1)	T17'(O1P)	24	70	R11(NH1)	T4(O1P)	12	80
R11(NH2)	T17'(O1P)	32	185	R11(NH2)	T4(O1P)	16	160
R11(NH2)	T17'(O2P)	14	40	R11(NH2)	T4(O2P)	8	35
R14(NH1)	T18'(O1P)	89	85	R14(NH1)	T3(O1P)	92	610
R14(NH2)	T18'(O1P)	61	80	R14(NH2)	T3(O1P)	67	95
R20(NE)	A19'(O1P)	14	55	R20(NE)	A2(O1P)	29	135
R20(NE)	A19'(O5')	10	40	R20(NE)	A2(O5')	5	20
R20(NH1)	A19'(O2P)	29	60	R20(NH1)	A2(O2P)	27	65
R20(NH1)	A19'(O1P)	18	50	R20(NH1)	A2(O1P)	17	45
R20(NH2)	A19'(O1P)	19	95	R20(NH2)	A2(O1P)	26	25
R40(NE)	T17'(O2P)	24	25	R40(NE)	T4(O2P)	30	25
R40(NH1)	T16'(O2P)	13	80	R40(NH1)	T5(O2P)	6	55
Q21(N)	T18'(O2P)	95	35 535	Q21(N)	T3(O2P)	100	67 910
Q21(NE2)	T17'(O2P)	98	249 260	Q21(NE2)	T4(O2P)	100	5020
S31(N)	T14(O2P)	100	166 170	S31(N)	T7'(O2P)	100	83 080
S31(OG)	T14(O2P)	100	498 530	S31(OG)	T7'(O2P)	100	498 530
S36(OG)	T17'(O2P)	99	9575	S36(OG)	T4(O2P)	99	8120
Q37(NE2)	C13(O2P)	11	75	Q37(NE2)	C8'(O2P)	3	60
W38(NE1)	C13(O2P)	95	705	W38(NE1)	C8'(O2P)	94	860
W38(NE1)	C13(O1P)	15	25	W38(NE1)	C8'(O1P)	17	25
T43(OG1)	C13(O2P)	85	305	T43(OG1)	C8'(O2P)	81	290
N46(ND2)	T9'(O1P)	42	165	N46(ND2)	T12(O1P)	15	195
N46(N)	C13(O1P)	2	140	N46(N)	C8'(O1P)	1	2725
N49(ND2)	C13(O1P)	3	137	N49(ND2)	C8'(O1P)	<1	-
N49(ND2)	C13(O5')	<1	-	N49(ND2)	C8'(O5')	<1	-

B

M1	Base	%pres.	Lifetime	M2	Base	%pres.	Lifetime
Q37(NE2)	C13(O4)	3	8	Q37(NE2)	C8'(O4)	<1	-
Q37(NE2)	T14(O4)	3	125	Q37(NE2)	T7'(O4)	1	160

Bold horizontal lines indicate the separation between salt bridges (above) and hydrogen bonds (below). Interactions in black are common to the experimental structure and the MD trajectory, those in red only occur in the MD trajectory and those in green only occur in the experimental structure.

Many of these interactions we have analyzed, both salt bridges and hydrogen bonds, not only break and reform regularly, but also involve changes in the DNA sites contacted by given amino acids. How much these fluctuations subsequently modify recognition of the DNA sequence varies: TBP is completely unaffected, SRY is moderately affected due to a single interface residue and P22 is significantly affected due to changes indirectly coupled to a single interface residue. At least for the complexes studied here, changes in amino acid interactions seem to have little impact on DNA conformation and where they modify sequence selectivity, this occurs because of the changes in direct amino-acid base interactions.

The case of P22 also underlines one limitation of our ADAPT sequence threading approach. While the ion distributions seen during the MD simulation clearly support the observed sequence preference in the center of the binding site (that is not directly in contact with the protein), these environmental effects cannot be taken into account by ADAPT that, for computational reasons, cannot deal with explicit ions or water molecules. However, the fact that ADAPT fails to predict any recognition in the central region of P22 also suggests that the changes in DNA geometry (involving a B → B' transition) that we indeed observe are not themselves sufficient to explain the associated sequence recognition.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors wish to acknowledge GENCI for a generous allocation of supercomputer resources at the CINES center in Montpellier.

FUNDING

ANR project CHROME [ANR-12-BSV5-0017-01]; Rhône-Alpes ARC 1 Santé Doctoral Grant (to L.E.). Funding

for open access charge: ANR project CHROME [ANR-12-BSV5-0017-01].
Conflict of interest statement. None declared.

REFERENCES

1. Etheve,L., Martin,J. and Lavery,R. (2016) Dynamics and recognition within a protein-DNA complex. *Nucleic Acids Res.*, **44**, 1440–1448.
2. Esadze,A., Li,D.W., Wang,T., Brüschweiler,R. and Iwahara,J. (2011) Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear 1H−15N NMR spectroscopy. *J. Chem. Soc.*, **133**, 909–919.
3. Anderson,K.M., Esadze,A., Manoharan,M., Bru schweiler,R., Gorenstein,D.G. and Iwahara,J. (2013) Direct observation of the ion-pair dynamics at a protein–DNA interface by NMR spectroscopy. *J. Am. Chem. Soc.*, **135**, 3613–3619.
4. Zandarashvili,L., Esadze,A. and Iwahara,J. (2013) NMR studies on the dynamics of hydrogen bonds and ion pairs involving lysine side chains of proteins. *Adv. Protein Chem. Struct. Biol.*, **93**, 37–80.
5. Chen,C., Esadze,A., Zandarashvili,L., Nguyen,D., Pettitt,B.M. and Iwahara,J. (2015) Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein–DNA Complexes. *J. Phys. Chem. Lett.*, **6**, 2733–2737.
6. Nikolov,D.B., Chen,H., Halay,E.D., Hoffman,A., Roeder,R.G. and Burley,S.K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 4862–4867.
7. Murphy,E.C., Zhurkin,V.B., Louis,J.M., Cornilescu,G. and Clore,G.M. (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.*, **312**, 481–499.
8. Watkins,D., Hsiao,C., Woods,K.K., Koudelka,G.B. and Williams,L.D. (2008) P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. *Biochemistry*, **47**, 2325–2338.
9. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
10. Deremble,C., Lavery,R. and Zakrzewska,K. (2008) Protein-DNA recognition: Breaking the combinatorial barrier. *Comput. Phys. Commun.*, **179**, 112–119.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
12. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (Junction Minimization of Nucleic-Acids). *Comput. Phys. Commun.*, **91**, 135–158.
13. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
14. Pearlman,D.A., Case,D.A., Caldwell,J.W., Ross,W.S., Cheatham,T.E., DeBolt,S., Ferguson,D., Seibel,G. and Kollman,P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.*, **91**, 1–41.
15. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
16. Cheatham,T.E. 3rd, Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
17. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
18. Dang,L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous-solutions of 18-crown-6 ether—a molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
19. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
20. Darden,T., Perera,L., Li,L. and Pedersen,L. (1999) New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, **7**, R55–R60.
21. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
22. Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1977) Numerical-integration of cartesian equations of motion of a system with constraints—molecular-dynamics of N-alkanes. *J. Comput. Phys.*, **23**, 327–341.
23. 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.
24. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
25. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2413–2423.
26. Goddard,T.D., Huang,C.C. and Ferrin,T.E. (2007) Visualizing density maps with UCSF Chimera. *J. Struct. Biol.*, **157**, 281–287.
27. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
28. Case,D.A., Berryman,J., Betz,R.M., Cerutti,D., Cheatham,T. III, Darden,T., Duke,R., Glese,T., Gohlke,H. *et al.* (2015) AMBER 2015.
29. Krause,E.F. (2012) *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation.
30. Ward,J.H. Jr (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
31. Kaufman,L. and Rousseeuw,P.J. (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
32. Murtagh,F. and Legendre,P. (2014) Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *J. Classif.*, **31**, 274–295.
33. R Development Core Team (2009) R: a language and environment for statistical computing.
34. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
35. Faiger,H., Ivanchenko,M., Cohen,I. and Haran,T.E. (2006) TBP flanking sequences: asymmetry of binding, long-range effects and consensus sequences. *Nucleic Acids Res.*, **34**, 104–119.
36. 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.
37. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A. and Ienasescu,H. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
38. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A. and Chen,X. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
39. Watkins,D., Mohan,S., Koudelka,G.B. and Williams,L.D. (2010) Sequence recognition of DNA by protein-induced conformational transitions. *J. Mol. Biol.*, **396**, 1145–1164.
40. Harris,L.A., Watkins,D., Williams,L.D. and Koudelka,G.B. (2013) Indirect readout of DNA sequence by p22 repressor: roles of DNA and protein functional groups in modulating DNA conformation. *J. Mol. Biol.*, **425**, 133–143.
41. Iwahara,J., Esadze,A. and Zandarashvili,L. (2015) Physicochemical properties of Ion Pairs of biological macromolecules. *Biomolecules*, **5**, 2435–2463.
42. Anderson,K.M., Nguyen,D., Esadze,A., Zandrashvili,L., Gorenstein,D.G. and Iwahara,J. (2015) A chemical approach for site-specific identification of NMR signals from protein side-chain NH3+ groups forming intermolecular ion pairs in protein–nucleic acid complexes. *J. Biomol. NMR*, **62**, 1–5.
43. Esadze,A., Chen,C., Zandarashvili,L., Roy,S., Pettitt,B.M. and Iwahara,J. (2016) Changes in conformational dynamics of basic side chains upon protein–DNA association. *Nucleic Acids Res.*, **44**, 6961–6970.

Supplementary Data

Protein-DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors

Loïc Etheve, Juliette Martin and Richard Lavery

Table S1. Percentage presence of salt bridges during the MD simulations of the TBP, SRY and P22 complexes with DNA, ignoring changes in the closest interacting atom pairs between arginine or lysine and the DNA phosphate groups.

TBP	% pres.
ARG192-T10'	77
ARG192-T11'	43
ARG199-T9'	95
ARG199-T10'	25
ARG204-T8'	47
ARG290-T7	38
ARG290-A8	97
ARG295-A9	15

SRY	% pres.
ARG4-A8	51
LYS6-A9	78
ARG7-A8	46
ARG17-A10	79
ARG17-A9	46
ARG21-C11	89
ARG31-C14	7
LYS37-T9'	45
LYS44-T8'	22
ARG66-G7'	9
LYS73-C4'	28
ARG75-A3'	50
ARG77-A6	76
ARG77-C7	73
ARG78-G2'	8
LYS79-A6	28
LYS81-C5	12

P22 M1	% pres.	P22 M2	% pres.
ARG11-T17'	40	ARG11-T4	20
ARG14-T18'	95	ARG14-T3	97
ARG20-A19'	73	ARG20-A2	79
ARG40-T17'	24	ARG40-T4	30
ARG40-T16'	13	ARG40-T5	6

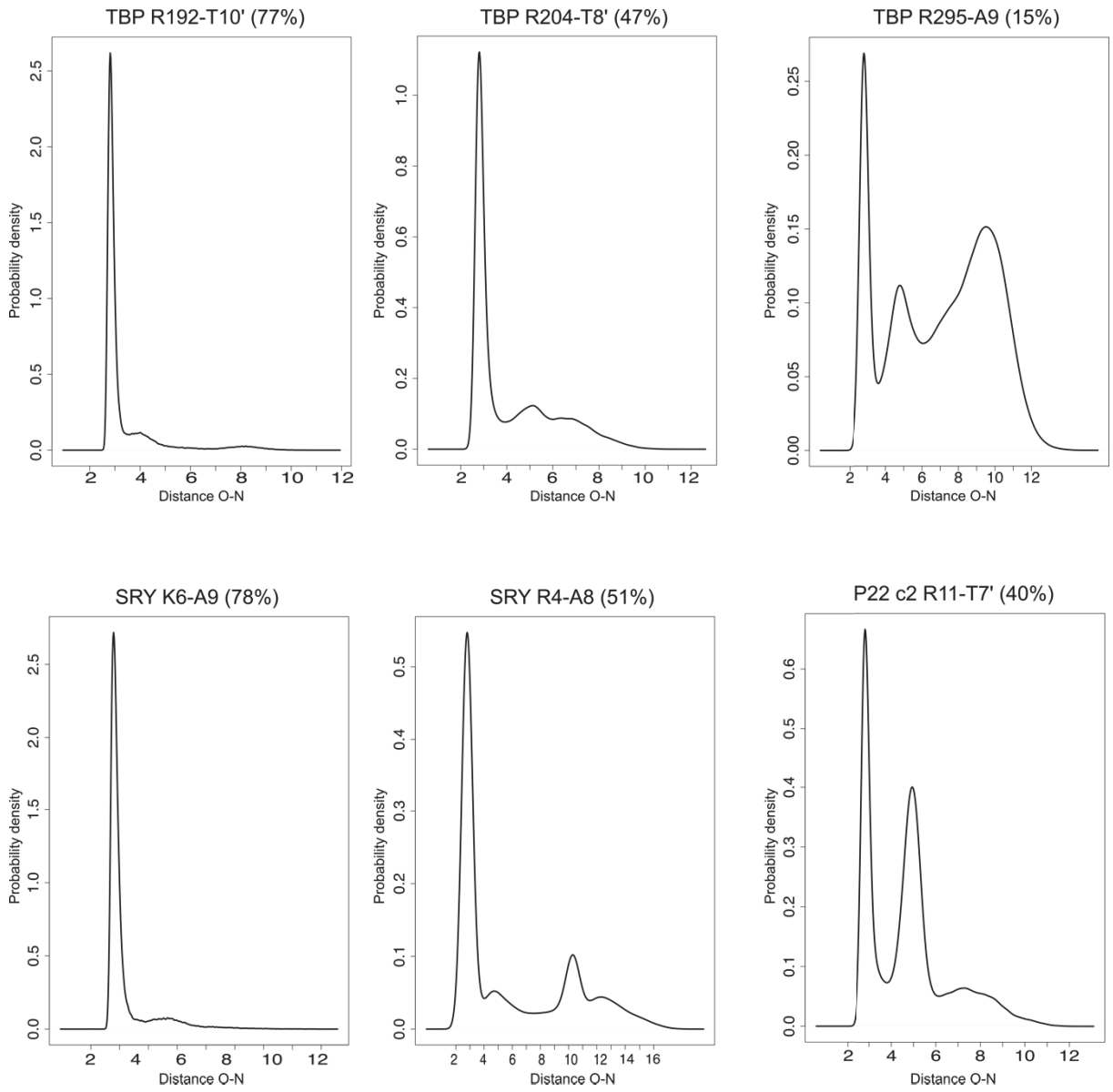


Figure S3. Probability densities for the Lys/Arg-phosphate interactions (shortest distance between lysine NZ or arginine NH1/NH2 and phosphate O1P/O2P) for a number of salt bridges drawn from the MD trajectories of the TBP, SRY and P22 complexes. Direct interactions are characterized by distances below 3.5 Å. Their percentage presence during the trajectory is indicated in brackets. Most cases show indirect interactions in the range 3.5-6 Å and even longer-range interactions may occur.

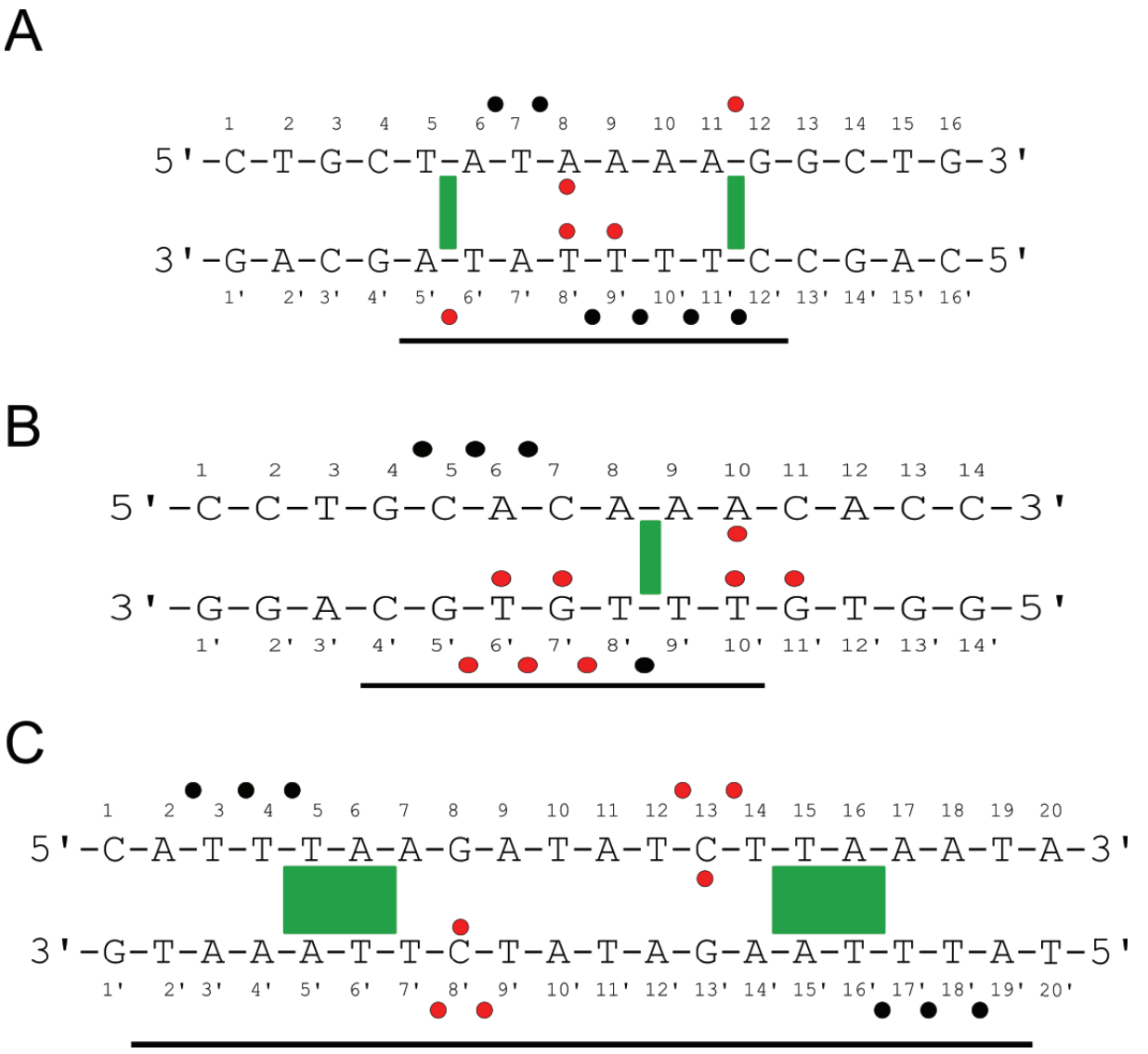


Figure S2. Salt bridges (black dots) and hydrogen bonds (red dots) and important apolar contacts (green rectangles) in the experimental structures of the complexes: A) TBP, B) SRY and C) P22. The DNA binding sites are delimited by horizontal black lines.

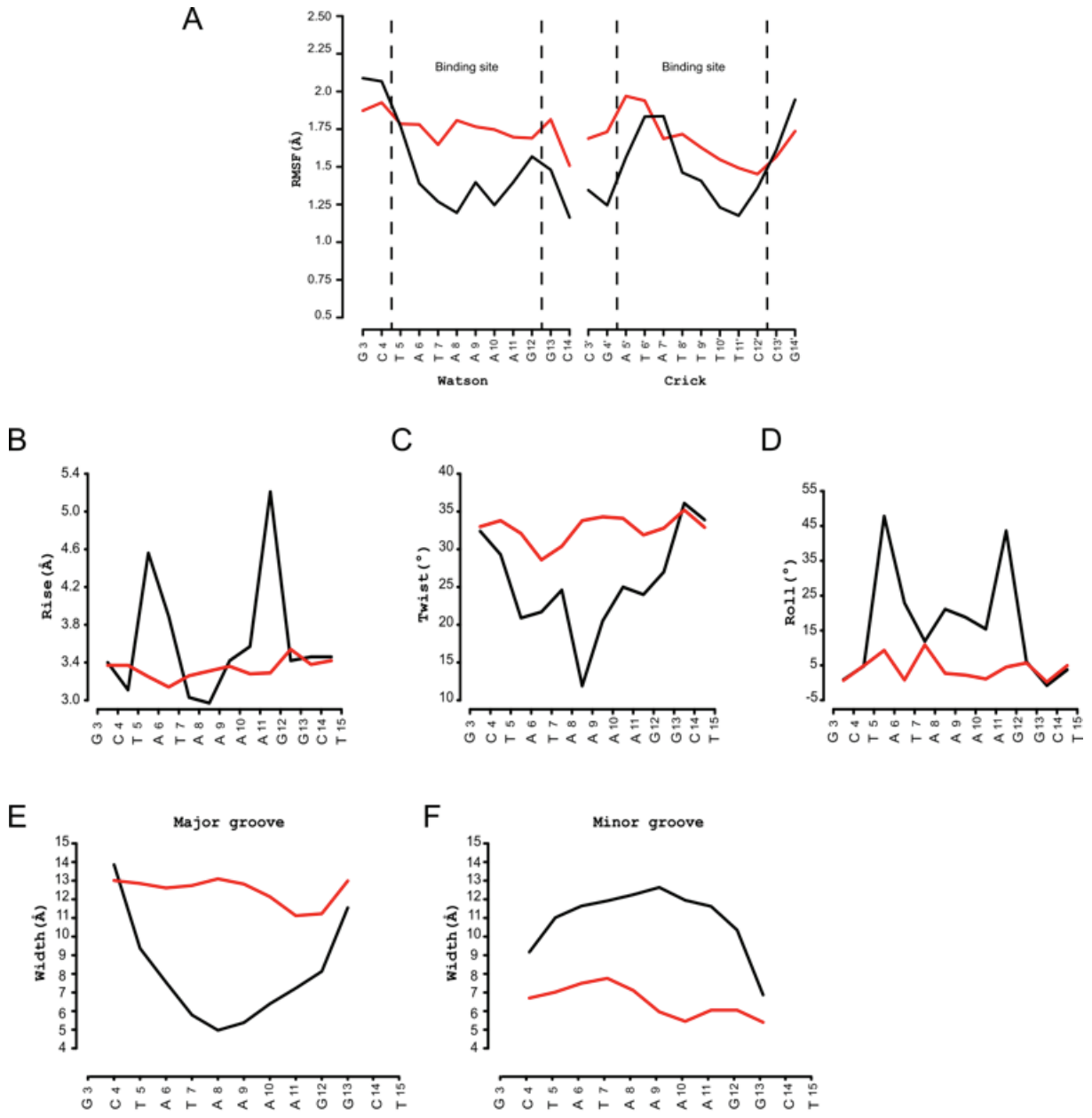


Figure S3. Structural information for the DNA/TBP complex (black lines) and for the isolated DNA oligomer (red lines): A) RMSF (Å) of phosphorus atoms. Vertical dashed lines indicate the protein-binding site; B, C and D) average values of rise (Å), twist (°) and roll (°) along DNA oligomer during MD simulation; E and F) major and minor groove widths (Å). Sequences are shown for the Watson strand in the 5'-3' direction.

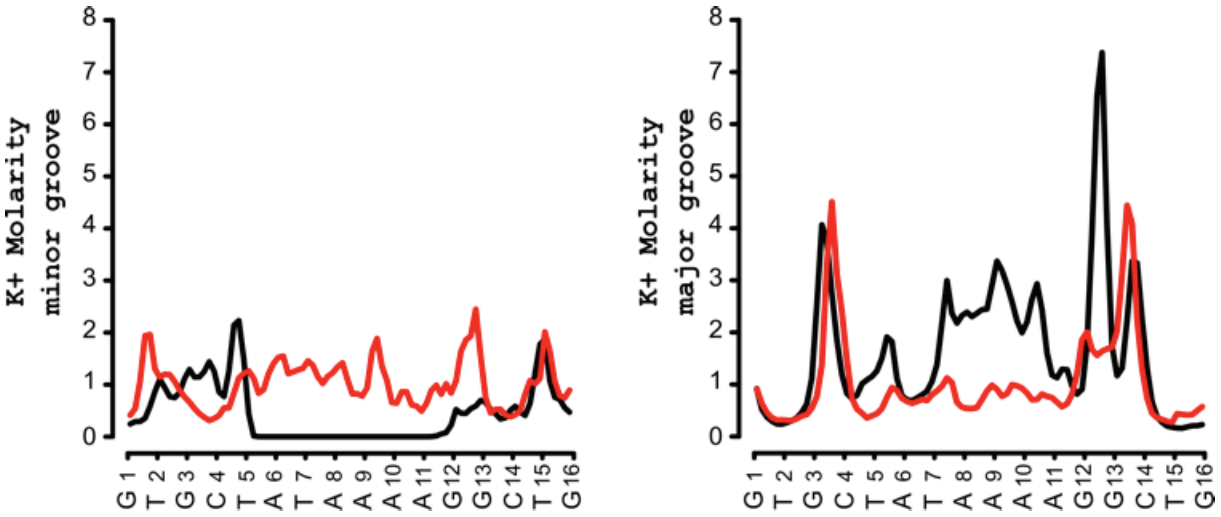


Figure S4. Potassium ion molarity for the DNA/TBP complex (black lines) and for the isolated DNA oligomer (red lines) in the minor groove (left panel) and in the major groove (right panel). Sequences are shown for the Watson strand in the 5'-3' direction.

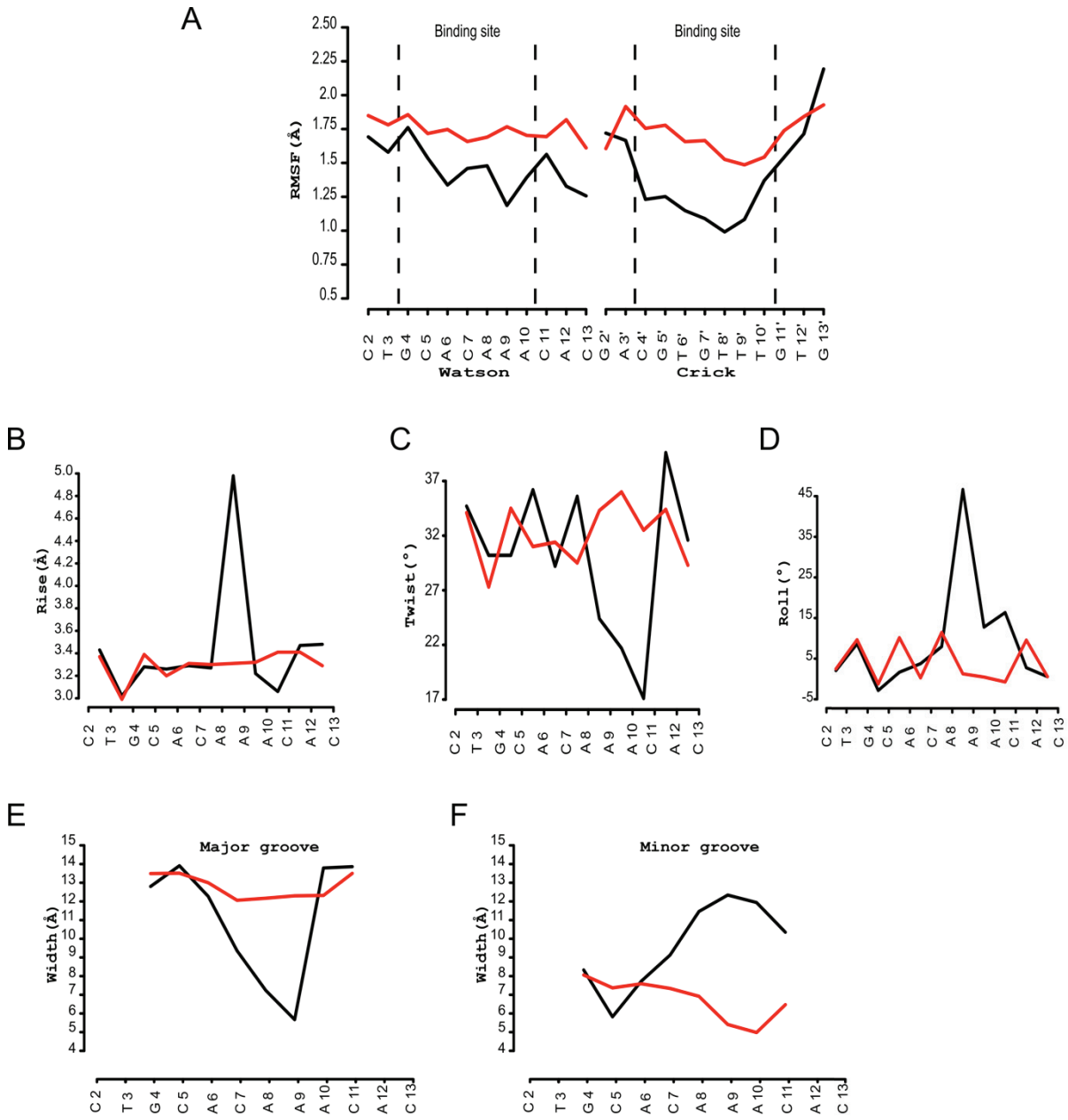


Figure S5. Structural information for the DNA/SRY complex (black lines) and for the isolated DNA oligomer (red lines): A) RMSF (Å) of phosphorus atoms. Vertical dashed lines indicate the protein-binding site; B, C and D) average values of rise (Å), twist (°) and roll (°) along DNA oligomer during MD simulation; E and F) major and minor groove widths (Å). Sequences are shown for the Watson strand in the 5'-3' direction.

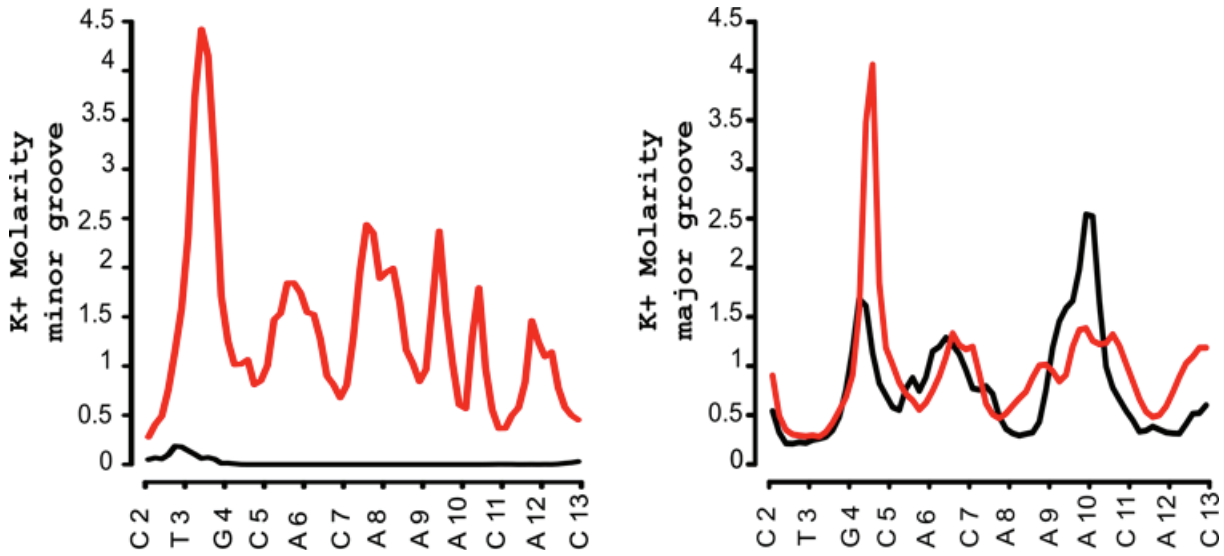


Figure S6. Potassium ion molarity for the DNA/SRY complex (black lines) and for the isolated DNA oligomer (red lines) in the minor groove (left panel) and in the major groove (right panel). Sequences are shown for the Watson strand in the 5'-3' direction.

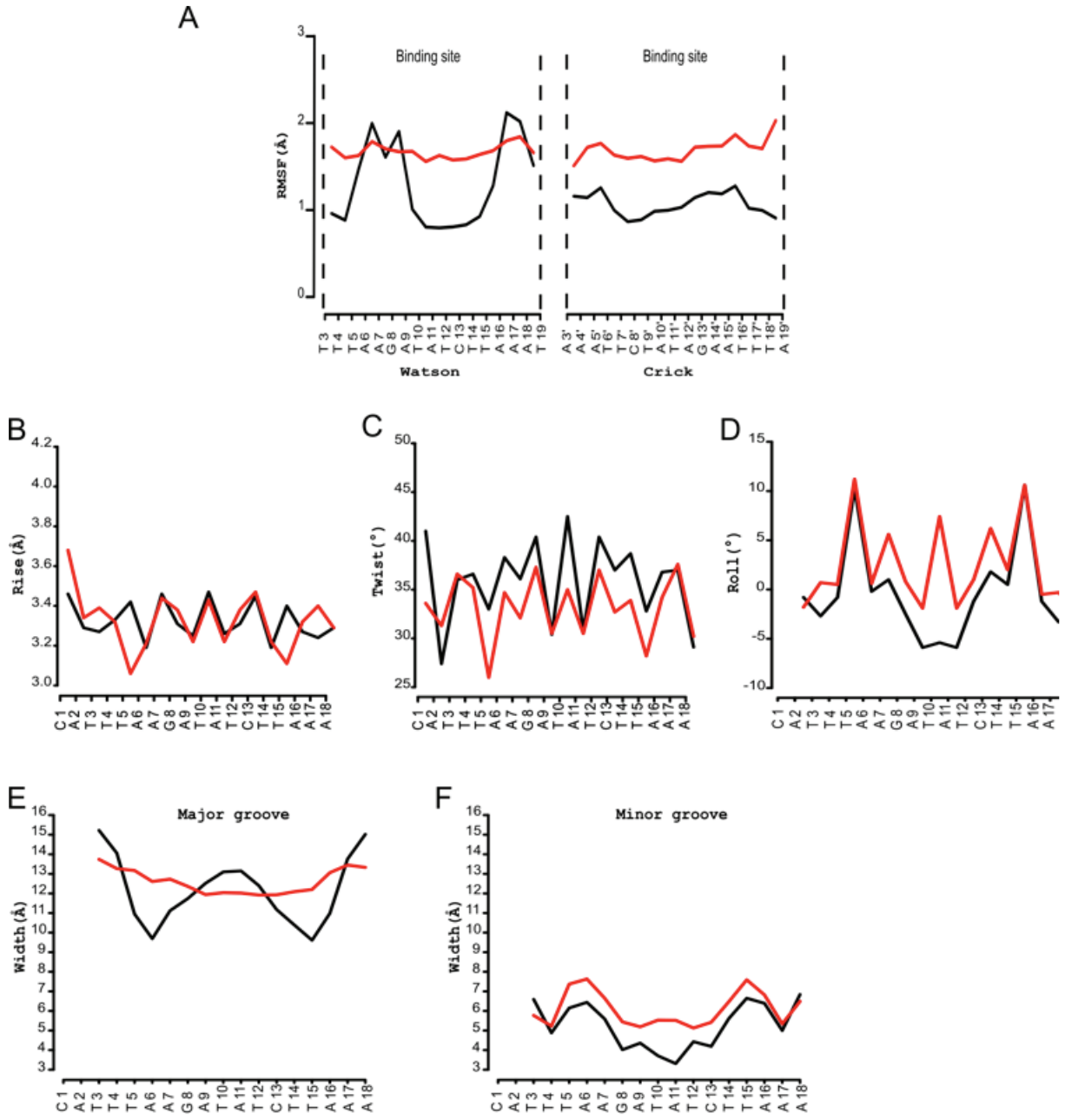


Figure S7. Structural information for the DNA/P22 complex (black lines) and for the isolated DNA oligomer (red lines): A) RMSF (Å) of phosphorus atoms. Vertical dashed lines indicate the protein-binding site; B, C and D) average values of rise (Å), twist (°) and roll along DNA oligomer during MD simulation E and F) major and minor groove widths. Sequences are shown for the Watson strand in the 5'-3' direction.

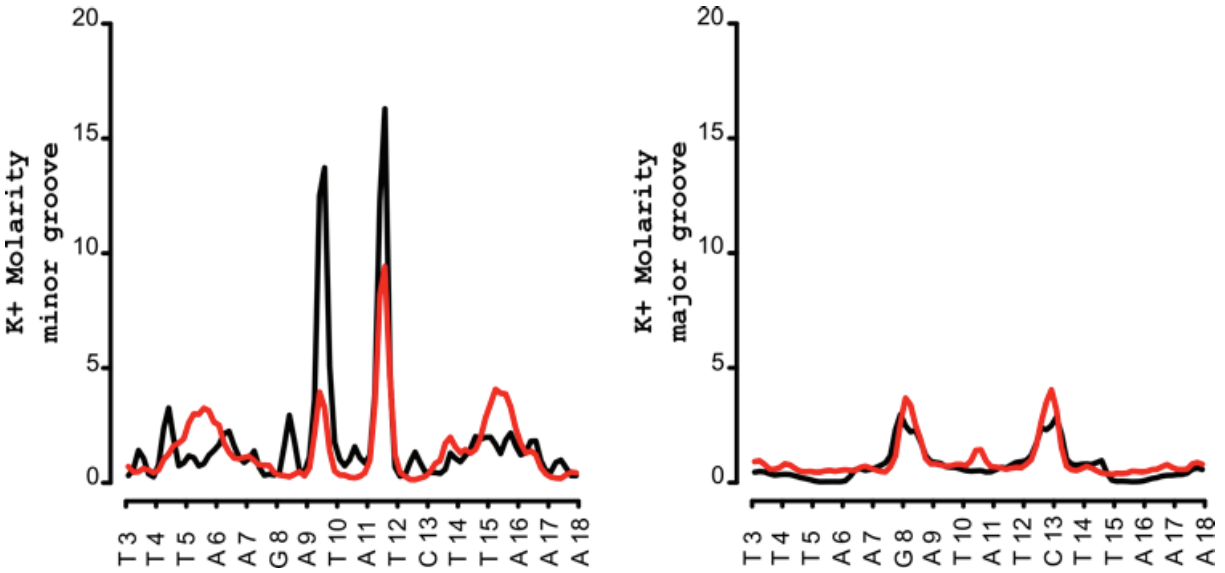


Figure S8. Potassium ion molarity for the P22 complex (black lines) and for the isolated DNA oligomer (red lines) in the minor groove (left panel) and in the major groove (right panel). Sequences are shown for the Watson strand in the 5'-3' direction