Le domaine THAP de THAP1 : structure par RMN en solution et interaction avec l'ADN

Damien Bessiere

UNIVERSITE TOULOUSE III – Paul SABATIER

UFR SCIENCES DE LA VIE ET DE LA TERRE

THESE

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITE DE TOULOUSE
délivré par l'Université Toulouse III – Paul Sabatier

Discipline : **BIOLOGIE STRUCTURALE**

présentée et soutenue
par

Damien BESSIERE

le 01 février 2008

LE DOMAINE THAP DE THAP1 : STRUCTURE PAR RMN EN SOLUTION ET INTERACTION AVEC L'ADN

Directeurs de thèse :

Alain MILON et Virginie GERVAIS

JURY

Jean COGNET, professeur, Université Paris VI	Président
Eric GUITTET, directeur de recherche CNRS**,** ICSN, Gif/Yvette	Rapporteur
Michel KOCHOYAN, directeur de recherche CNRS, CBS, Montpellier	Rapporteur
Alain MILON, professeur, Université Toulouse III	Directeur de thèse
Virginie GERVAIS, chargée de recherche CNRS, IPBS, Toulouse	Co-directrice de thèse

Institut de Pharmacologie et de Biologie Structurale
CNRS UMR 5089 205 route de Narbonne, 31077 Toulouse cedex 4

UNIVERSITE TOULOUSE III – Paul SABATIER

UFR SCIENCES DE LA VIE ET DE LA TERRE

THESE

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITE DE TOULOUSE
délivré par l'Université Toulouse III – Paul Sabatier

Discipline : **BIOLOGIE STRUCTURALE**

présentée et soutenue
par

Damien BESSIERE

le 01 février 2008

LE DOMAINE THAP DE THAP1 : STRUCTURE PAR RMN EN SOLUTION ET INTERACTION AVEC L'ADN

Directeurs de thèse :

Alain MILON et Virginie GERVAIS

JURY

Jean COGNET, professeur, Université Paris VI	Président
Eric GUITTET, directeur de recherche CNRS**,** ICSN, Gif/Yvette	Rapporteur
Michel KOCHOYAN, directeur de recherche CNRS, CBS, Montpellier	Rapporteur
Alain MILON, professeur, Université Toulouse III	Directeur de thèse
Virginie GERVAIS, chargée de recherche CNRS, IPBS, Toulouse	Co-directrice de thèse

Institut de Pharmacologie et de Biologie Structurale
CNRS UMR 5089 205 route de Narbonne, 31077 Toulouse cedex 4

Remerciements

Je tiens à remercier tout d'abord les membres de mon jury, Messieurs Jean COGNET, Eric GUITTET et Michel KOCHOYAN d'avoir accepté d'évaluer mon travail de thèse et pour l'intérêt qu'ils y ont porté.

Je remercie Alain MILON de m'avoir accueilli dans son équipe et proposé ce sujet de thèse. Merci de m'avoir dirigé pendant ces quatre années.

Ce travail a été co-dirigé par Virginie GERVAIS que je tiens à remercier tout particulièrement pour son encadrement de tous les jours. Ca a été un véritable plaisir de travailler avec toi, merci pour tout…

Je tiens également à remercier tout les membres de l'équipe RMN et interactions protéines-membranes et en particulier Pascal DEMANGE pour son aide en production de protéine et Sebastien CAMPAGNE qui assure la suite de la réalisation de mon sujet de thèse.

Ce travail a fait l'objet d'une collaboration avec deux équipes de l'IPBS. Je remercie pour leur collaboration Jean-Philippe GIRARD, Vincent ECOCHARD et Chrystelle LACROIX de l'équipe de Biologie Vasculaire ainsi que Lionel MOUREY et Valérie GUILLET de l'équipe de Biophysique Structurale. Merci à toi Chrystelle pour ton travail considérable de mutagénese ; c'est avec plaisir que je partage la tête d'affiche de notre article.

Les expériences de RPS ont été réalisées en collaboration avec Jean-Pierre ESTEVE et Frédéric LOPEZ.

Pour la réalisation de cette thèse, j'ai pu bénéficier d'une bourse du Ministère de la Recherche sur trois ans et d'une bourse de quatrième année de la Fédération pour la Recherche Médicale.

Enfin, j'aimerais remercier mes amis de l'Institut ou « la bande à Lucie » pour leurs échanges scientifiques et autres… Merci également à Christel avec qui je partage le même parcours scientifique.

Je finirais par remercier mes amis et ma famille pour leur soutien, en particulier en ce jour du 1 février 2008.

RESUME

Le domaine THAP de THAP1 : structure par RMN en solution et interaction avec l'ADN

La famille des protéines THAP (THanatos Associated Protein) a été identifiée dans le laboratoire de Biologie Vasculaire de l'Institut de Pharmacologie et Biologie Structurale. Elle est caractérisée par la présence d'un nouveau motif protéique, le domaine THAP, conservé au cours de l'évolution et retrouvé dans une centaine de protéines chez l'homme et les organismes animaux modèles.
La protéine THAP1 définissant la famille des protéines THAP est impliquée dans la régulation du cycle cellulaire dans la voie pRb/E2F et dans la prolifération cellulaire. Il a été démontré que le domaine THAP de THAP1 est un nouveau motif de liaison à l'ADN de type C-X_{2-4}-C-X_{35-50}-C-X_2-H, séquence-spécifique et dépendant du zinc.

Nous avons déterminé la structure tridimensionnelle du domaine THAP de THAP1 par Résonance Magnétique Nucléaire en solution. Nous avons ainsi montré que ce domaine est un motif à doigt de zinc atypique d'environ 80 résidus, qui se distingue par la présence entre les deux paires de ligands de coordination au zinc de motif C_2CH, d'un court feuillet β anti-parallèle dans lequel s'intercale une longue insertion de type boucle-hélice-boucle.
Bien que les structures des domaines THAP de THAP2 et CtBP aient également été résolues, la structure du domaine THAP de THAP1 constitue la première structure d'un domaine doigt de zinc THAP pour lequel une activité biochimique et des fonctions biologiques associées ont été déterminées.

Nous avons également étudié la liaison du domaine THAP de THAP1 à sa séquence ADN spécifiquement reconnue par différentes approches biophysiques : la Résonance Magnétique Nucléaire (RMN), la fluorescence et la Résonance Plasmonique de Surface (RPS).
Ainsi nous avons déterminé une constante de dissociation spécifique par RPS et réalisé des expériences RMN de variation de déplacement chimique de façon à identifier les résidus impliqués dans la liaison à l'ADN.
En collaboration avec l'équipe de Biologie Vasculaire, des expériences de mutagénèse dirigée ont été réalisées afin de déterminer des résidus critiques pour la reconnaissance de l'ADN.
La combinaison des données de mutagénèse dirigée et de variation de déplacement chimique nous a permis de localiser l'interface de liaison à l'ADN du domaine THAP de THAP1 au niveau d'une zone fortement chargée positivement à la surface de la protéine.
Ces données présentent ainsi la première étude des relations structure/fonction d'un domaine THAP pour lequel une activité séquence spécifique a été clairement démontrée. Elles fournissent de plus une meilleure compréhension du mode de reconnaissance de l'ADN par ce doigt de zinc original et ont permis de proposer un modèle d'interaction.

PREAMBULE

La famille des protéines THAP a été découverte au début des années 2000 au sein de l'Institut de Pharmacologie et Biologie Structurale, dans l'équipe de biologie vasculaire (JP. Girard). L'identification de la protéine THAP1 et du domaine THAP a permis de définir une nouvelle famille de facteurs cellulaires, les protéines THAP.

De façon à mieux caractériser ces protéines, des études structurales et fonctionnelles ont été entreprises au sein de notre institut. Il a été démontré que le domaine THAP était un nouveau domaine de liaison à l'ADN séquence spécifique et dépendant du zinc.

Dans ce contexte, mes travaux de thèse, réalisés dans l'équipe de RMN biologique (A. Milon), ont consisté en la caractérisation structurale du domaine de liaison à l'ADN (domaine THAP) de la protéine THAP1, prototype de cette famille. Nous avons résolu la structure tridimensionnelle de ce domaine à doigt de zinc par Résonance Magnétique Nucléaire (RMN) et étudié la liaison à l'ADN de ce domaine en utilisant différentes méthodes biophysiques (RMN, fluorescence, Résonance Plasmonique de Surface). Ces travaux ont été réalisés en collaboration avec l'équipe de Biologie Vasculaire qui a mené des études de mutagénèse afin d'identifier les résidus importants pour la liaison.

L'ensemble de ces travaux a donné lieu à une publication (Bessiere et al; JBC, 2008) et constitue la première étude structure-fonction relative à cette nouvelle famille de protéines, l'objectif étant de comprendre le mécanisme de reconnaissance séquence-spécifique.

Je présenterai, dans un premier temps le domaine THAP et la famille qu'il définit. Nous développerons ensuite trois parties portant sur les protéines à doigt de zinc, la reconnaissance protéine-ADN et la détermination de structure par RMN.

Enfin, j'illustrerai les différentes études réalisées sur le domaine THAP de THAP1 et les résultats issus de mes travaux de thèse que nous discuterons pour finir par des perspectives sur ce travail.

TABLE DES MATIERES

INTRODUCTION

Le domaine THAP, un nouveau domaine de liaison à l'ADN

Identification de la protéine THAP1

La protéine humaine THAP1 a été identifiée par Myriam Roussigne et ses collaborateurs (Roussigne et al., 2003a). Elle est associée à des complexes multiprotéiques nucléaires, les PML-NBS (Promyelocytic leukemia Nuclear Bodies) impliqués dans la régulation de nombreux processus vitaux tels que la transcription, la croissance cellulaire, l'apoptose et la défense antivirale. Les premiers travaux réalisés montrent que la surexpression de GFP-THAP1 renforce le processus apoptotique des cellules en apoptose sous l'effet du facteur de nécrose tumorale TNFα ou du sevrage en sérum, soulignant le caractère pro-apoptotique de THAP1 (Roussigne et al., 2003a). C'est de là que provient le nom de cette protéine : THanatos Associated Protein 1, Thanatos étant le dieu grec de la mort.
Dans le cadre de cette étude, le facteur pro-apoptotique Par4 (Prostate apoptosis response 4) a été identifié comme partenaire protéique de THAP1 par des expériences de double hybride. De plus il a été montré que THAP1 interagit et colocalise avec Par4 au niveau des PML-NBs dans les cellules primaires endothéliales ou les fibroblastes.
Ces résultats mettent en évidence un lien entre la protéine THAP1, la protéine proapototique Par4 et les corps nucléaires PML-NBs (Roussigne et al., 2003a).

Le domaine THAP définit une nouvelle famille de protéines

La protéine THAP1 humaine est composée de 213 acides aminés. Des analyses de séquence ont permis de mettre en évidence que la région N-terminale de THAP1 présente de l'homologie avec de nombreuses protéines non caractérisées et a ainsi permis de définir un nouveau domaine protéique d'environ 90 acides aminés conservé dans ces protéines, le domaine THAP (Roussigne et al., 2003b). Le domaine THAP est restreint au règne animal, il a été recensé dans plus de 100 protéines chez différents organismes animaux vertébrés et invertébrés, mais il n'est

pas retrouvé dans les plantes, les levures, les champignons et les bactéries (Clouaire et al., 2005; Roussigne et al., 2003b).

Le domaine THAP n'est pas requis pour l'interaction avec Par4 ni pour la colocalisation aux PML-NBs, mais il est essentiel à l'activité pro-apoptotique identifiée pour la protéine THAP1 humaine (Roussigne et al., 2003a).

Chez l'homme, le domaine THAP définit une nouvelle famille de protéines, les protéines THAP. Cette famille comprend 12 membres, THAP0 à THAP11 (figure 1).

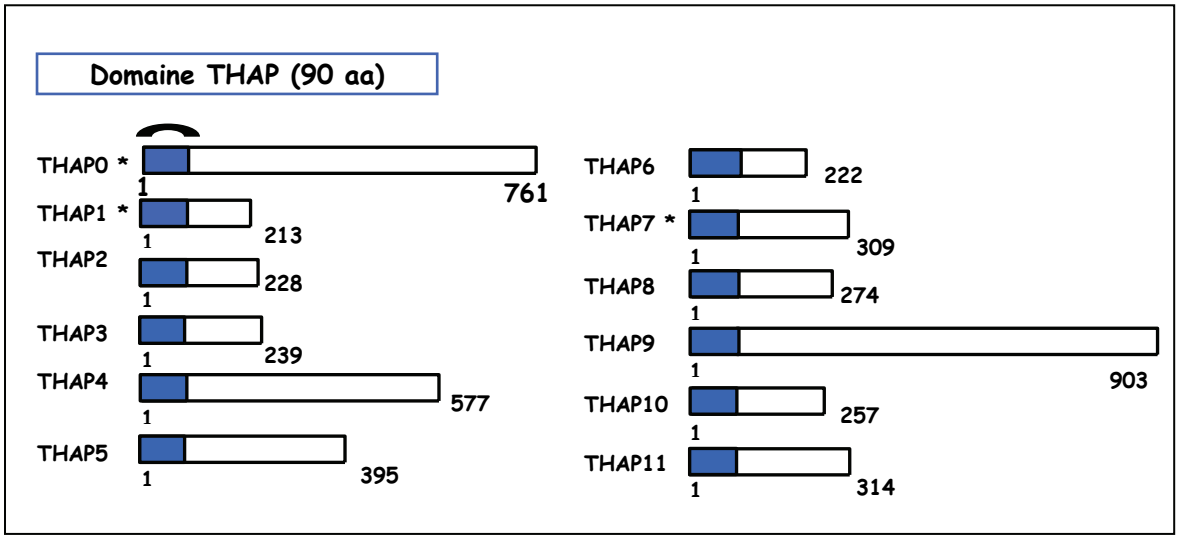


Figure1 : **La famille THAP chez l'homme**
Les rectangles bleus délimitent le domaine THAP de 90 acides aminés. Le nombre d'acides aminés est indiqué pour chaque protéine. Un astérisque indique les protéines pour lesquelles des données fonctionnelles sont disponibles.

La protéine THAP0 était déjà référencée sous le nom de DAP4 (Death Associated Protein 4). Elle est impliquée dans la régulation de l'apoptose (Deiss et al., 1995), notamment par son interaction avec la sérine/thréonine kinase MST1 (Lin et al., 2002). La co-expression de THAP0 et MST1 permet d'augmenter l'apoptose induite par MST1. MST1 est en effet impliquée dans la condensation de la chromatine caractéristique de l'apoptose par phosphorylation de l'histone H2B (Cheung et al., 2003). Aucune étude n'a pour l'instant mis en évidence le rôle du domaine THAP dans l'effet pro-apoptotique de THAP0.

Des données fonctionnelles supplémentaires sont aussi disponibles pour THAP1 qui, outre ses propriétés pro-apoptotiques en surexpression, est impliqué dans le contrôle

de la prolifération cellulaire et la régulation du cycle cellulaire (Cayrol et al., 2007) que nous développerons ci-après.

Enfin, des données fonctionnelles sont également disponibles pour THAP7. Ce facteur est impliqué dans la régulation transcriptionnelle et interagit avec des enzymes de modification de la chromatine (Macfarlan et al., 2005; Macfarlan et al., 2006). THAP7 réprime la transcription par recrutement de l'histone désacétylase HDAC3 et le corépresseur du récepteur nucléaire aux hormones NCoR (Macfarlan et al., 2005). De plus THAP7 interagit avec TAF-Iβ et inhibe l'acétylation des histones H3 et H4 pour réprimer la transcription (Macfarlan et al., 2006).

On retrouve également le domaine THAP dans plusieurs organismes animaux modèles (Table 1).

Model organism	THAP proteins	Putative orthologues
H. sapiens	12	THAP 0–THAP 11
M. musculus	7	THAP 0, 1, 2, 3, 4, 7, and 11
G. gallus (chicken)	6	THAP 0, 1, 4, 5, and 7
X. laevis	23	THAP 1, 4, 7, and 11
D. rerio (zebrafish)	32	THAP 0, 1, 7, 9, and 11
D. melanogaster	9	
C. elegans	8	

Table 1 : **Les protéines THAP dans les organismes modèles**
Adapté de (Clouaire et al., 2005)

L'étude de la famille des protéines THAP chez *C. elegans* est particulièrement riche en informations sur le rôle biologique de ces protéines. On retrouve ainsi des régulateurs du cycle cellulaire Lin-15B et Lin-36 qui sont des inhibiteurs de la transition de phase G1/S du cycle cellulaire (Boxem and van den Heuvel, 2002) ainsi qu'une protéine impliquée dans la modification de la chromatine, HIM-17 (Reddy and Villeneuve, 2004).

L'ensemble de ces données semble indiquer que les protéines THAP, à la fois chez l'homme et dans les organismes modèles animaux, sont des régulateurs de la prolifération cellulaire et du cycle cellulaire et qu'elles agissent au niveau de la régulation de la chromatine.

L'analyse des alignements de ces différentes séquences permet de caractériser le domaine THAP. Il présente une signature C2CH de consensus C-X_{2-4}-C-X_{35-50}-C-X_2-H, quatre autres résidus strictement conservés P26, W36, F58, P78 (numérotation

relative à THAP1) et une boîte AVPTIF en position C-terminale (Roussigne et al., 2003b) (figure 2).

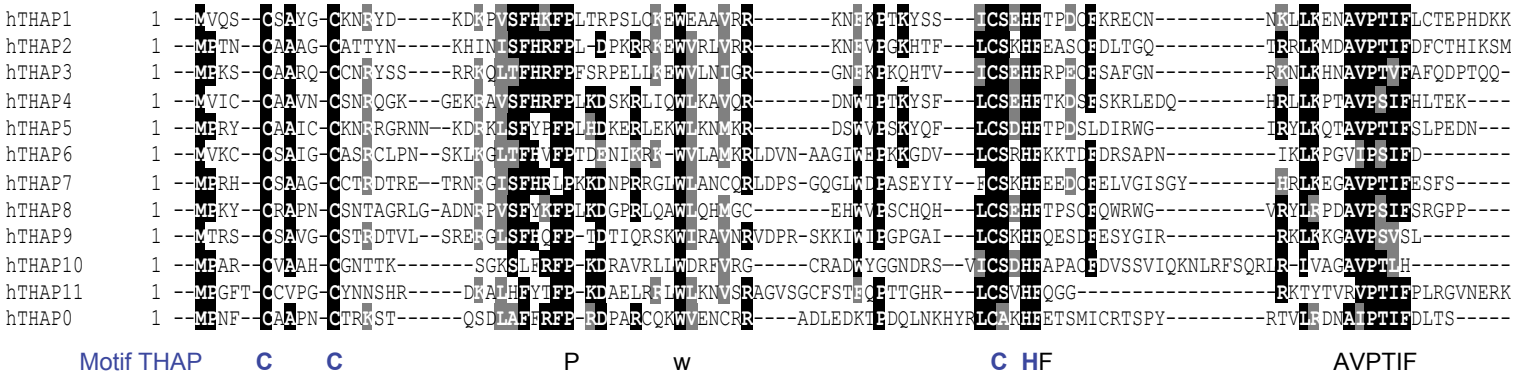


Figure 2 : **Alignement du domaine THAP pour les 12 protéines THAP humaines**
Les résidus conservés sont indiqués en dessous de l'alignement avec la signature C2CH en bleu.

De façon intéressante, le domaine THAP présente des similarités de séquence avec le domaine de liaison à l'ADN de l'élément P de la transposase (Lee et al., 1998; Roussigne et al., 2003b). L'homologie comprend la taille, la localisation N-terminale et la conservation des résidus constituant la signature du domaine THAP. Ces données ont permis alors de suggérer que les protéines THAP définissent une nouvelle famille de protéines de liaison à l'ADN jusqu'alors non caractérisée comprenant une signature C2CH, motif de coordination au zinc.

Il n'est pas si surprenant d'observer le partage d'un domaine de liaison à l'ADN entre des protéines cellulaires et la transposase d'un élément génétique mobile. Par exemple, le domaine de liaison BED, un domaine à doigt de zinc, est présent à la fois dans les protéines cellulaires BEAF et DREF et dans le domaine de liaison à l'ADN des transposases de la famille hAT (Aravind, 2000).

Le domaine THAP de THAP1, nouveau domaine de liaison à l'ADN

La liaison du domaine THAP au zinc ainsi que le rôle du domaine THAP de THAP1 en tant que motif d'interaction avec l'ADN ont été démontrés dans l'équipe de biologie vasculaire (Clouaire et al., 2005).

Des expériences de type SELEX (*Selective Evolution of Ligands by EXponential enrichment*) (Bouvet, 2001) ont permis d'isoler une séquence oligonucléotidique spécifiquement reconnue par le domaine THAP de THAP1 (figure 3).

Figure 3 : **Identification de la séquence consensus THABS**
Séquence oligonucléotidique reconnue par le domaine THAP de THAP1, obtenue par SELEX après 12 cycles. La taille des lettres représentant les bases identifiées est proportionnelle à leur probabilité de présence à chaque position. Adapté de (Clouaire et al., 2005).

Les résultats de SELEX ont été validés par des expériences de mutagénèse dirigée sur la séquence oligonucléotidique combinées à des expériences de gels retard (*Electrophoretic Mobility Shift Assay*) pour observer la liaison au domaine THAP. Cette séquence présente un cœur GGCA essentiel pour la reconnaissance par le THAP domaine de THAP1. Des mutations d'oligonucléotides sur ce motif abolissent la liaison au domaine THAP (figure 4). D'une façon générale, les expériences de mutagénèse dirigée sont en accord avec les données de SELEX sauf pour la thymine à la position 3 qui n'apparaît pas absolument nécessaire pour la reconnaissance protéine-ADN.
Cette étude a ainsi conduit à l'identification d'une séquence consensus d'ADN cible de 11 nucléotides du domaine THAP de la protéine humaine THAP1, nommée THABS (THAP1 Binding Sequence) : AGTAAGGGCAA (Clouaire et al., 2005)

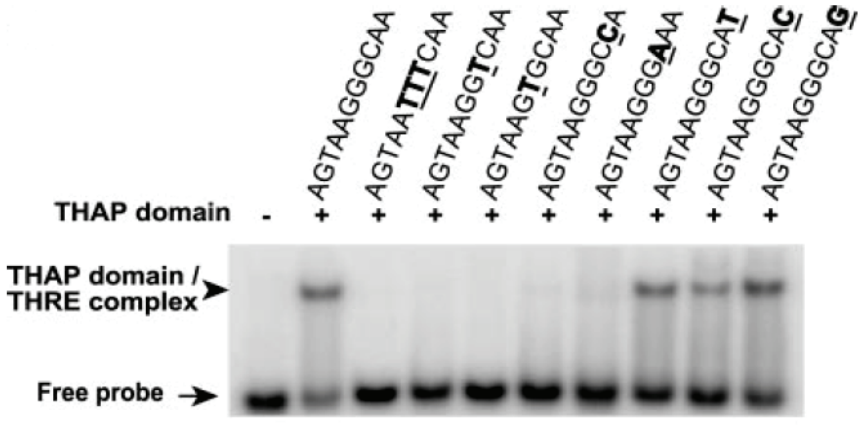


Figure 4 : **Le cœur GGCA est essentiel pour la reconnaissance THAP/THABS**
Gel retard réalisé avec des séquences THABS mutées sur certaines positions oligonucléotidiques. Adapté de (Clouaire et al., 2005)

La liaison au zinc a été démontrée par l'emploi de chélateurs comme l'EDTA ou la phénanthroline qui abolissent la liaison à l'ADN (observé par gel retard). La liaison peut être rétablie par ajout d'ions zinc spécifiquement. De plus, la mutation ponctuelle des quatre résidus de la signature C2CH, susceptibles de coordonner le zinc, abolit également la liaison à l'ADN (figure 5). Enfin, il est montré de la même façon que les résidus strictement conservés, P26, W36, F58 et P78 sont essentiels pour la liaison THAP-THABS (Clouaire et al., 2005).

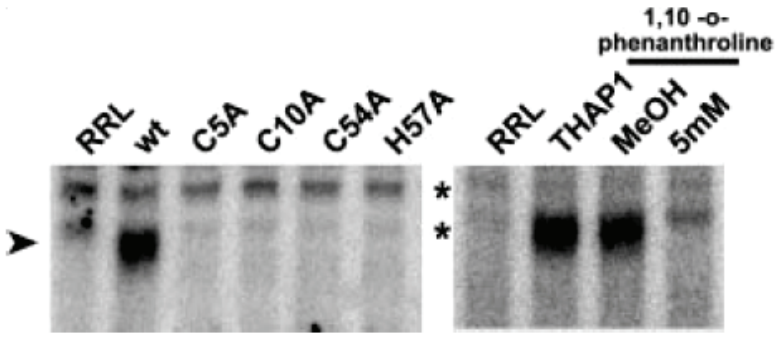


Figure 5 : **La signature C2CH est essentielle pour la liaison à l'ADN dépendante du zinc**
Gels retard du domaine THAP avec les mutations des résidus C2CH en alanine qui abolissent la liaison du domaine THAP à l'ADN. Le même résultat est observé par chélation du zinc par la phénanthroline à 5mM. Une astérisque indique des complexes non-spécifique, le complexe spécifique est repéré par une flèche. Adapté de (Clouaire et al., 2005).

Le domaine THAP constitue un nouveau motif de liaison à l'ADN séquence-spécifique, dépendant du zinc qui définit une nouvelle famille de protéines.

Le domaine THAP appartient à la famille des domaines à doigts de zinc. Le nombre de membres, à peu près 100 protéines identifiées, en fait l'un des domaines de liaison à l'ADN impliquant une coordination au zinc, les plus abondants, après le domaine à doigts de zinc C2H2 et les récepteurs nucléaires (Clouaire et al., 2005). Le domaine THAP a une taille particulière (90 acides aminés contre 30 pour le domaine C2H2) et surtout une signature C2CH originale. En effet, l'espacement des résidus entre les cystéines en position 2 et 3 du motif C2CH est particulièrement grand (35 à 53 acides aminés contre 10 à 12 pour le domaine C2H2). De plus, la taille de la séquence spécifiquement reconnue par ce domaine est particulière puisque THAP1 reconnaît une séquence de 11 nucléotides contre 3 pour le domaine C2H2 classique. Ces originalités suggèrent un repliement nouveau et rendent ce domaine et son complexe avec l'ADN intéressants d'un point de vue structural.

THAP1 est impliquée dans la régulation de la prolifération cellulaire et du cycle cellulaire

Pour compléter cette caractérisation biophysique du domaine THAP de THAP1, des études ont été menées sur le rôle biologique de THAP1 par l'équipe de Biologie Vasculaire, Cayrol et ses collaborateurs ont montré que THAP1 est un régulateur de la prolifération des cellules endothéliales et du cycle cellulaire (Cayrol et al., 2007). Ainsi la surexpression de THAP1 ou sa déplétion par RNAi (RNA interférence) dans des cellules endothéliales entraînent un arrêt de la prolifération cellulaire suggérant qu'un niveau optimal de l'expression de THAP1 est crucial pour la prolifération cellulaire. L'analyse du transcriptome de cellules endothéliales, par puce à ADN, a permis d'identifier des gènes sous-exprimés lors de la surexpression de THAP1. De façon intéressante, ces gènes sont impliqués dans la voie de régulation pRb/E2F du cycle cellulaire (Dimova and Dyson, 2005; Dyson and Wright, 2005)l. L'inhibition de THAP1 dans des cellules endothéliales entraîne la sous-expression de plusieurs de ces gènes, évaluée par PCR quantitative. En particulier, le gène *RRM1* qui est un gène activé à la transition G1/S, codant pour la sous-unité M1 de la réductase ribonucléotidique, et essentiel pour la synthèse de l'ADN en phase S (Bjorklund et al., 1993; Johansson et al., 1998), fait partie des gènes dont l'expression est modulée par THAP1. Des expériences de *ChiPs* (chromatin immunoprecipitation) et de *footprinting* ont montré que THAP1 se liait directement à l'ADN au niveau du promoteur de *RRM1* sur un site comprenant la séquence consensus THABS. Ces données complètent les données obtenues sur les organismes animaux modèles et permettent d'associer les protéines THAP à la voie pRB/E2F du cycle cellulaire et à la prolifération cellulaire.

Le domaine THAP de THAP1 fait l'objet de notre étude. Nous venons de voir qu'il s'agit d'un domaine de liaison à l'ADN faisant intervenir une coordination au zinc. Nous allons maintenant décrire les protéines à doigt de zinc et passer en revue les trois domaines les plus importants de liaison à l'ADN faisant intervenir une coordination au zinc : le doigt de zinc classique C2H2, le domaine GATA, et le domaine de liaison à l'ADN des récepteurs nucléaires.

Les protéines à doigt de zinc

Introduction

Le zinc est le métal le plus abondant dans le corps humain après le fer, il fut d'abord mis en évidence dans des métalloenzymes dans les années 1950 (Coleman, 1992) et en 1983 dans un facteur de transcription, TFIIIA (Hanas et al., 1983). Initialement, le terme *doigt de zinc* décrit le motif structural de coordination de l'ion zinc avec les quatre résidus Cys2His2 (C2H2) du facteur TFIIIA (Miller et al., 1985). Ce motif répété consécutivement au sein de la protéine se compose d'une trentaine de résidus. L'origine de cette désignation en *doigt de zinc* provient de la représentation topologique du domaine et en particulier de la région entre la deuxième cystéine et la première histidine du motif C2H2 déjà pressenti comme région liant l'ADN (figure 6). Le terme de *doigt* étant également associé au fait que cette région « s'accroche » à l'ADN (Klug and Schwabe, 1995).

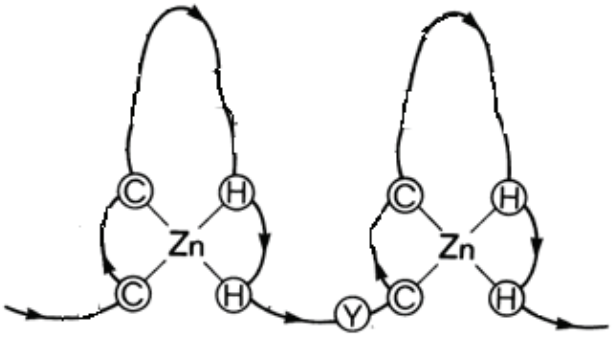


Figure 6 : **Topologie schématique de deux doigts de zinc consécutifs**
La coordination des résidus C2H2 au zinc est représentée. Le terme doigt provient de la boucle liant l'ADN, formée entre les résidus en position 2 et 3 du motif C2H2. Adapté de (Miller et al., 1985)

Le terme *doigt de zinc* désigne désormais l'ensemble des domaines protéiques comportant une coordination au zinc. Il définit un motif de petite taille, replié de manière autonome autour d'une coordination au zinc par plusieurs résidus (Klug and Schwabe, 1995), le plus souvent quatre, de type histidines ou cystéines (occasionnellement aspartates ou glutamates). Toutefois, certains préfèreront réserver le terme de *doigt de zinc* pour désigner le domaine de type C2H2 et

parleront plus généralement de motif structural liant le zinc. Nous faisons le choix de parler de *doigt de zinc* pour l'ensemble des motifs structuraux liant le zinc, considérant que le terme fait référence au schéma topologique de la coordination au zinc et puisqu'il semble être admis par la communauté scientifique (Krishna et al., 2003; Laity et al., 2001).

Ce motif structural est surtout connu pour son implication dans la reconnaissance d'ADN mais il est aussi retrouvé dans le cas d'interactions protéine-ARN (Kelly et al., 2007) et protéine-protéine (Gamsjaeger et al., 2007). Les membres de cette super famille protéique peuvent contenir un ou plusieurs de ces motifs.

Les protéines à doigts de zinc sont très nombreuses. Andreini et ses collaborateurs estiment la proportion de protéines liant le zinc à 10% du génome humain, dont 40% de protéines à doigt de zinc, associées quasi exclusivement à des facteurs de transcription (Andreini et al., 2006).

La coordination au zinc

La coordination à l'ion zinc augmente la stabilité conformationnelle de ces petits domaines et elle est indispensable au maintien de la structure. Le zinc a donc un rôle exclusivement structural dans les protéines à doigts de zinc, ce qui n'est pas le cas pour les metalloenzymes comme pour la carboxypeptidase ou l'anhydrase carbonique où il est associé à une fonction catalytique (Coleman, 1992). En particulier, le zinc n'interagit pas avec l'ADN reconnu par ces modules (Klug and Schwabe, 1995).

Le cation Zn^{2+} est lié aux chaînes latérales des cystéines ou histidines pour former un complexe organométallique. Ce complexe présente une coordination tétraédrique, avec le zinc au centre et les atomes de soufre (SG) ou d'azote (ND1 ou NE2) pour respectivement les cystéines et les histidines autour du zinc qui sont alors déprotonées (figure 7).

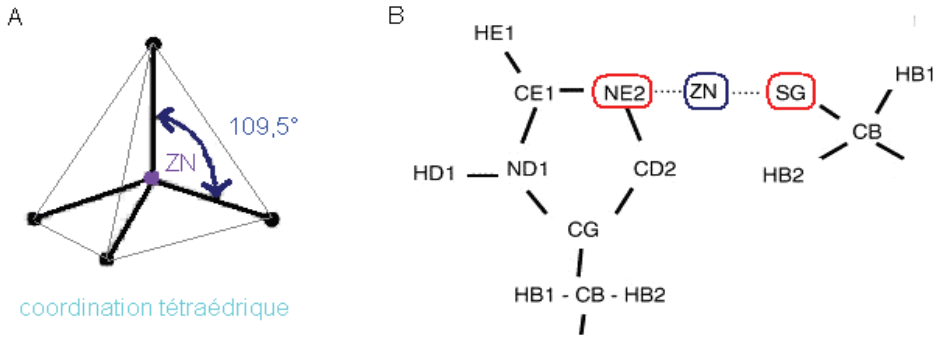


Figure 7 : **Coordination au zinc**

A : Géométrie de coordination au zinc en tétraèdre. B : Topologie de coordination au zinc par l'atome de soufre SG d'une cystéine et l'atome d'azote NE2 d'une histidine.

On trouve ainsi un angle proche de 109,5° entre le zinc et deux des ligands et une longueur caractéristique entre le zinc et les ligands d'environ 2.0 et 2.3 Å pour les distances Zn-N et Zn-S (Alberts et al., 1998; Lee et al., 1989). Ces paramètres sont classiquement utilisés pour placer le zinc dans la protéine lors de la génération des structures sous contraintes RMN comme nous le verrons pour la détermination de la structure du domaine THAP de THAP1 (cf p 97). De même, nous verrons que dans le cas des histidines il convient de déterminer lequel des atomes d'azote ND1 ou NE2 est coordonné au zinc.

Classification des protéines à doigts de zinc

L'identification et la classification des motifs à doigt de zinc ne sont pas évidentes compte tenu du nombre important de membres de cette super famille et de la variété de motifs structuraux qu'elle englobe ; elle en est donc d'autant plus intéressante. Toutefois, l'enjeu est de taille pour établir des relations structure-fonction et pouvoir prédire structures et fonctions de protéines.

Classification en fonction des ligands au zinc

Beaucoup de classes de protéines à doigts de zinc sont caractérisées en fonction du nombre et de la position des résidus cytéines et histidines impliqués dans la coordination au zinc. Une telle classification des motifs à doigts de zinc en fonction de la nature des ligands liant le zinc est ainsi proposée par Leon et collaborateurs (Leon and Roth, 2000). Elle comprend huit classes : Cys_2His_2 (ou CCHH), $Cys_2HisCys$ (ou CCHC), Cys_4 ribbon, Cys_4 GATA, Cys_6, Cys_8, Cys_3His Cys_4 RING fingers et His_2Cys_2 finger. (figure 8).

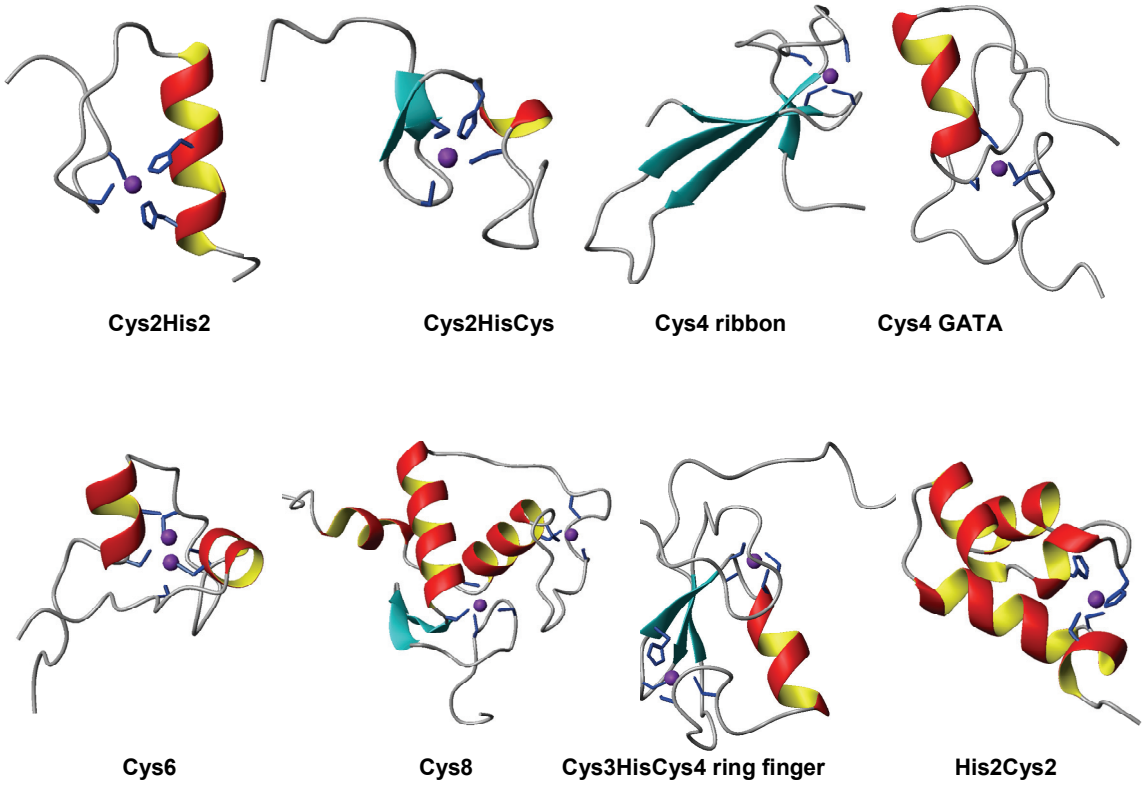


Figure 8 : **Structure des domaines à doigts de zinc**
Représentation des structures des doigts de zinc C2H2 : facteur de transcription Sp1 (code PDB : 1SP1) (Narayan et al., 1997), C2HC : nucleocapside de HIV-2 (code PDB : 1NC8) (Kodera et al., 1998), C4 ribbon : facteur de transcription TFIIS (code PDB : 1TFI) (Qian et al., 1993), C4 GATA : facteur de transcription GATA-1 (code PDB : 1GNF) (Kowalski et al., 1999), C6 facteur de transcription Gal4 (code PDB : 1AW6) (Baleja et al., 1997), C8 : recepteur rétinoïque X (code PDB : 1RXR) (Holmbeck et al., 1998), C3HC4 ring finger : protéine du virus de l'herpes (code PDB : 1CHC) (Barlow et al., 1994), H2C2 : domaine HHCC de l'intégrase de HIV-1 (code PDB : 1WJA) (Cai et al., 1997).
Les atomes de zinc sont représentés par des sphères violettes, les chaînes latérales des résidus cystéines et histidines coordinant le zinc sont figurées en bleu.

Les motifs de coordination les plus abondants dans le génome humain sont les motifs C4 et C2H2(Andreini et al., 2006) (figure 9)

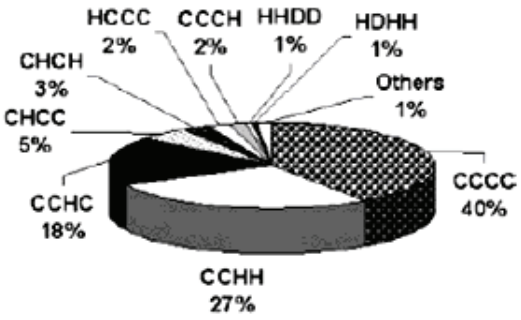


Figure 9 : **Distribution des motifs de coordination au zinc**
Représentation en camembert des proportions des différentes coordinations au zinc à 4 ligands seulement, chez l'homme. Adapté de (Andreini et al., 2006)

On remarquera également que le motif CCHH est beaucoup plus commun que le motif CCHC, qui l'est plus que le motif CCCH (respectivement 27%, 18% et 2%). En effet la nature du ligand n'est pas équivalente puisque l'histidine possède une chaîne latérale plus grande que celle de la cystéine. La coordination tétraédrique au zinc restant identique, le remplacement d'une histidine par une cystéine conduit à une déformation de la structure. C'est ainsi que l'on peut expliquer la déformation de l'hélice dans le motif CCCH du facteur de transcription amt1 (Turner et al., 1998) et le sous-domaine N-terminal du domaine de liaison à l'ADN de SBP (Yamasaki et al., 2004a) en comparaison avec le motif CCHH du doigt de zinc classique de la protéine χfin (Lee et al., 1989) (figure 10).

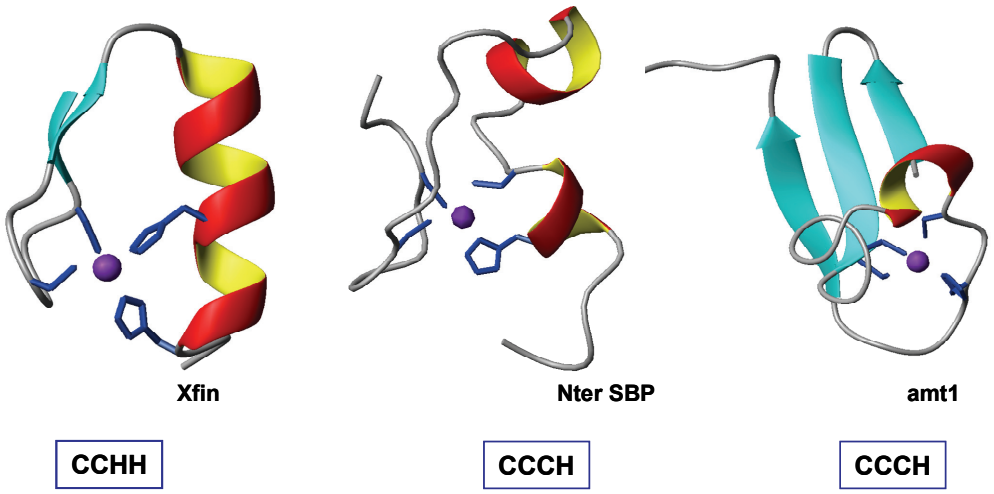


Figure 10 : **Substitution Cys/His, changement de structure**
Representation d'une structure CCHH, χfin (code PDB : 1ZNF), et de deux structures CCCH, sous-domaine N-terminal du domaine de liaison à l'ADN de SBP (code PDB : 1UL4) et amt1 (code PDB : 1CO4). La substitution Cys/His dans les ligands au zinc implique une destructuration de l'hélice α

On comprend aussi qu'une mutation de la deuxième histidine du motif CCHH sera moins pénalisante puisqu'elle se trouve à la fin de l'hélice α du motif doigts de zinc classique et non au milieu, d'où une proportion de CCHC supérieure à celle de CCCH. On trouve ainsi des variants naturels de CCHH avec une mutation His/Cys en quatrième position mais pas en troisième position (Krizek et al., 1991).

On retrouve le même genre de considérations dans le cas de la protéine NCp7 du virus HIV-1 qui comprend deux motifs doigts de zinc CCHC; la mutation de l'histidine en cystéine du premier doigt de zinc induit une perte de structure associée à une perte de fonctionnalité traduite par la production de virus non infectieux (Demene et al., 1994; Roques et al., 1997). Dans ce cas, la conclusion est d'autant plus significative puisqu'il s'agit d'une simple mutation sur la même séquence protéique (figure 11).

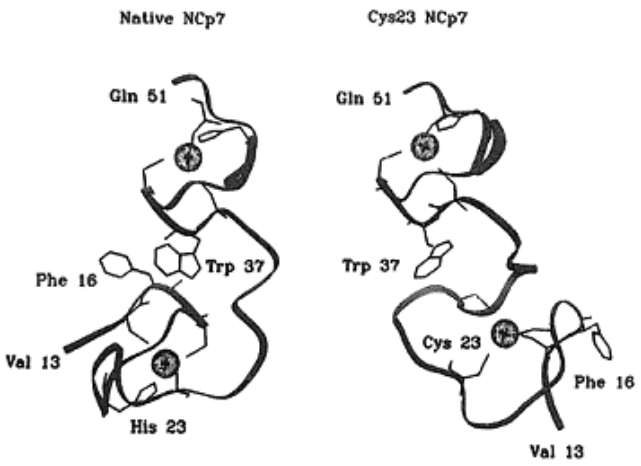


Figure 11 : **Structure des doigts de zinc de NCp7 sauvage et His23Cys mutant**
Le remplacement de l'histidine par une cystéine induit un changement de conformation de la liaison au zinc et une structuration différente. Adapté de (Roques et al., 1997).

Cette classification a ses limites puisque deux motifs ayant les mêmes ligands au zinc peuvent avoir des structures tridimensionnelles différentes, c'est le cas du domaine THAP de type CCCH qui n'a pas la même structure que amt1 comme nous l'avons remarqué lors de la détermination de sa structure. L'espacement entre ces différents ligands est à prendre en compte.

Classification structurale

Krishna et ses collaborateurs proposent une classification structurale des protéines à doigt de zinc (Krishna et al., 2003). Elle rassemble les protéines à doigt de zinc en groupe structuraux, où chaque membre d'un même groupe a le même placement de ses ligands au niveau des structures secondaires du motif. Elle comporte huit groupes : C2H2 like, Gag knuckle, treble clef, zinc ribbon, Zn2/Cys6, TAZ2 domain like, zinc binding loop, metallothionein (table 2).

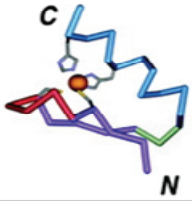
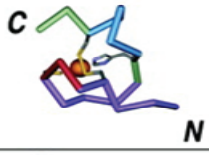
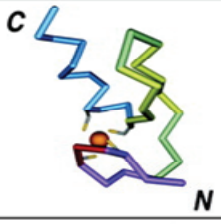
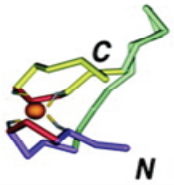
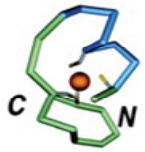
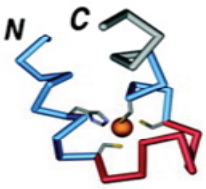
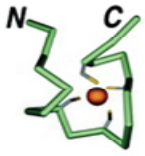
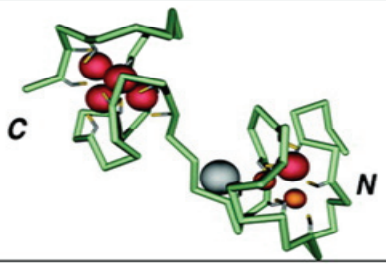
Fold group	Representative structure	Ligand placement	Members in the alignment
1. C2H2 like		Two ligands from a knuckle and two more from the C terminus of a helix	1ncsA, 1tf6A, 1tf6D, 1zfdA, 1ubdC, 2gliA, 1sp2A, 1rmdA, 1znfA, 2adrA, 1aayA, 1sp1A, 1bhiA, 1bboA, 2drpA, 1yuiA, 1ej6C, 1klrA, 1k2fA, 1fv5A, 1fu9A, 1g73C, 1jd5A, 1c9qA, 1e31A
2. Gag knuckle		Two ligands from a knuckle and two more from a short helix or loop	1a1tA, 1a6bB, 1dsvA, 1dsqA, 1fn9A, 1i3qA
3. Treble clef		Two ligands from a knuckle and two more from the N terminus of a helix	1chcA, 1borA,1jm7A, 1jm7B, 1rmdA, 1fbvA, 1g25A, 1ldjB, 1e4uA, 1dcqA, 1ptqA, 1faqA, 1kbeA, 1e53A, 1dvpA, 1vfyA, 1jocA, 1zbdB, 1fp0A, 1f62A, 1fjfN, 1jj2T, 1ee8A, 1k3wA, 1l2bA, 1ffyA, 1zfoA, 1xpaA, 4gatA, 2gatA, 1gnfA, 1b8tA, 1imlA, 1g47A, 1hcqA, 1kb6A, 1g2rA, 1en7A, 1bxiB, 1ql0A, 1a73A, 1mhdA, 1i3qJ, 1ef4A, 1lnrY, 1i3jA, 1hc7A, 1dgsA, 1lv3A
4. Zinc ribbon		Two ligands each from two knuckles	1jj2Z, 1jj2Y, 1d0qA, 1qypA, 1i50I, 1jj22, 1i5oD, 1qf8A, 1tfiA, 1i50B, 1pftA, 1i50A, 1aduA, 1yuaA, 1dfeA, 1gh9A, 1ileA, 1meaA, 1i50L, 1zinA, 1iciA, 1ma3A, 1a8hA, 1dx8A, 1irnA, 1dxgA, 2occF, 1freA, 1exkA,1b55A, 1f4lA, 1gaxA, 1lnrZ, 1lnr1, 1l1oF, 1e4vA, 1zakA, 1dl6A, 1b71A, 1j8fA, 1kjzA, 1ezvE, 1rfsA, 1g8kB, 1eg9A, 1fqtA
5. Zn2/Cys6		Two ligands from the N terminus of a helix and two more from a loop	1d66A, 1zmeC, 2hapC, 2alcA, 1co4A
6. TAZ2 domain like		Two ligands each from the termini of two helices.	1f81A, 1l8cA, 1wjbA, 1jr3A, 1jr3E
7. Zinc binding loops		Four ligands from a loop	A) 1hsoA, 1e3jA, 1i3qC, 1a5tA, 1cyqA1, 1gpcA B) 1enuA, 1iq8A, 1ia9A, 1cw0A, 1cyqA2, 1ldjB
8. Metallothionein		Cysteine rich metal binding loop	4mt2A

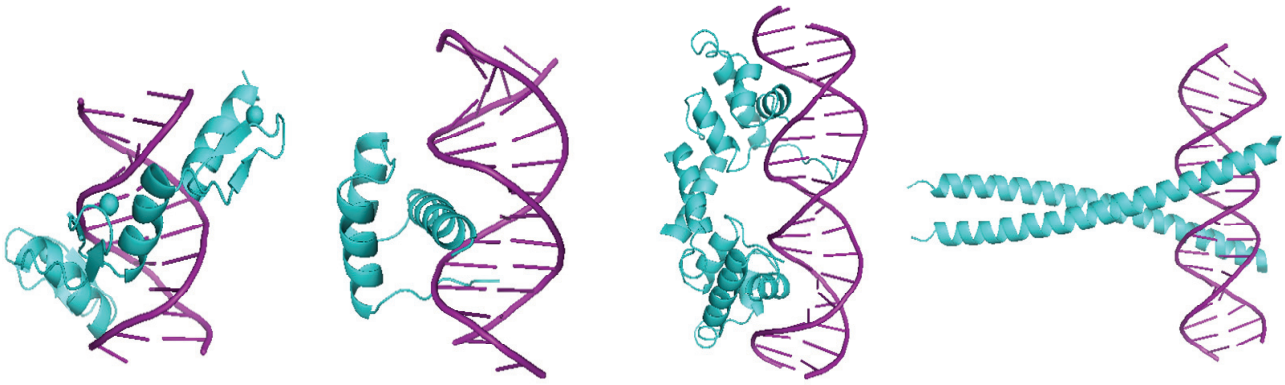
Table 2 : **Classification des motifs à doigts de zinc selon krishna et collaborateurs**
Les structures de motifs à doigts de zinc sont rassemblées en 8 groupes selon les propriétés structurales autour du site de liaison au zinc. Le zinc est représenté par une sphère orange, les chaînes latérales des ligands au zinc sont représentées en gris. Adapté de (Krishna et al., 2003)

Le but recherché est d'identifier des relations structure-activité et pouvoir prédire la fonction d'une protéine à partir de son groupe d'appartenance. Elle a été réalisée manuellement à partir d'alignements structuraux. En effet, la classification automatique de cette famille n'est pas concluante, la difficulté provenant de la petite taille du domaine à doigts de zinc.
Nous allons maintenant considérer quelques exemples de sous familles des protéines à doigts de zinc pour lesquelles des informations structurales sont connues et en nous focalisant plus particulièrement sur les domaines de liaison à l'ADN qui définissent les plus grandes familles de facteurs de transcription.

Les doigts de zinc liant l'ADN

Introduction

Les protéines de liaison à l'ADN jouent un rôle central en biologie et sont impliquées dans des processus majeurs comme la réplication du génome, la transcription des gènes ou la réparation de l'ADN endommagé. Elles représentent 2-3% du génome eucaryote et 6-7% du génome procaryote (Luscombe et al., 2000).
Dans cette grande famille de protéines reconnaissant l'ADN, les facteurs de transcription constituent la classe la plus importante et la plus variée. On y trouve ainsi la famille la plus étudiée et la plus vaste regroupant les protéines à doigt de zinc ainsi que les domaines hélice-tour-hélice, homéodomaine, bZIP (basic leucine zipper) pour ne citer que les principaux (Garvie and Wolberger, 2001; Luscombe et al., 2000; Pabo and Sauer, 1992).



Doigt de zinc **Homéodomaine** **Hélice tour hélice** **Leucine zipper**

Figure 12 : **Exemples des principales familles de domaine de liaison à l'ADN**
Representation des complexes protéine (bleu) ADN (violet) pour les domaines doigt de zinc Zif 268 (code PDB : 1ZAA) (Pavletich and Pabo, 1991), homéodomaine engrailed (code PDB : 1HDD) (Kissinger et al., 1990), hélice-tour-hélice du répresseur λ (code PDB : 1LMB) (Beamer and Pabo, 1992), bzip de Jun (code PDB : 2H7H)

Le doigt de zinc classique Cys2His2 (CCHH)

Introduction

Il s'agit du motif structural initialement identifié dans TFIIIA (Miller et al., 1985). Ce motif a depuis été retrouvé dans beaucoup de protéines de liaison à l'ADN. C'est l'un des domaines les plus abondants dans les protéines eucaryotes et plus particulièrement dans le génome humain où il a une proportion de 3%. Cette proportion augmente avec la complexité de l'organisme sur l'échelle de l'évolution (Klug, 2005b) et en fait le motif de liaison à l'ADN le plus commun chez l'homme. Même si l'on sait que des domaines à doigts de zinc CCHH peuvent lier l'ARN (Searles et al., 2000) et que d'autres participent à des interactions protéine-protéine (Sun et al., 1996), le domaine CCHH reste majoritairement impliqué dans la reconnaissance protéine-ADN.

Structure du domaine

Le motif CCHH est composé d'environ 30 résidus, lié au zinc par deux cystéines et deux histidines conservées. Il est caractérisé par la séquence consensus X_2-C-$X_{2,4}$-C-X_{12}-H-$X_{3,4,5}$-H (Pabo et al., 2001), où X représente n'importe quel acide aminé, C et H les deux cystéines et histidines liant le zinc et où l'espace entre les deux cystéines et les deux histidines est variable. La région de 12 acides aminés entre la

deuxième cystéine et la première histidine est généralement de la forme $-X_3-\Phi\ X_5-\Phi-X_2-$ où Φ représente un résidu hydrophobe. Cette région principalement polaire et basique est impliquée dans la liaison à l'ADN.

Des études de RMN ont permis de définir la structure du motif doigt de zinc C2H2 de type $\beta\beta\alpha$ (Lee et al., 1989). Les deux ligands cystéines sont situés à l'extrémité d'un feuillet β antiparallèle, et les deux ligands histidines sont situés dans la partie C-terminale de l'hélice α (figure 13). Ces quatre ligands maintiennent le feuillet et l'hélice ensemble par coordination à un unique ion zinc. La structure est aussi stabilisée par un petit cœur hydrophobe.

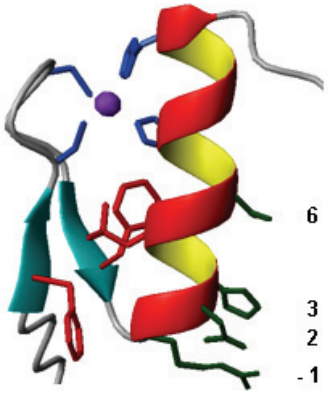


Figure 13 : **Le doigt de zinc classique**
Motif $\beta\beta\alpha$ du deuxième doigt de zinc de Zif268 (code PDB : 1ZAA). Le zinc est représenté par une sphère en violet, les chaînes latérales des résidus CCHH coordonnés au zinc en bleu, des résidus hydrophobes conservés en rouge, des résidus en position -1, 2, 3,6 en contact avec l'ADN en vert.

Liaison à l'ADN

Les premières informations structurales concernant l'interaction domaine C2H2 avec l'ADN furent apportées par la structure cristallographique du complexe Zif268, résolue à 2.1 Á en 1991 (Pavletich and Pabo, 1991). La structure fut ensuite affinée à une résolution de 1.6 Á (Elrod-Erickson et al., 1996). Elle sert de modèle général pour le mécanisme de reconnaissance de l'ADN pour de multiples doigts de zinc.

Le domaine de liaison à l'ADN de Zif268 comprend trois motifs C2H2 qui interagissent avec le grand sillon de l'ADN par l'intermédiaire de leur hélice (figure 14). La séquence consensus a été déterminée par Barbara Christy et ses collaborateurs (Christy and Nathans, 1989). Chaque doigt de zinc reconnaît 3-4

paires de bases (Pavletich and Pabo, 1991). L'hélice α ou hélice de reconnaissance contacte le grand sillon de l'ADN par sa partie N-terminale.

Plus particulièrement, les résidus en position -1, 2, 3 et 6 (figures 13 et 14) de chaque hélice sont impliqués dans des contacts spécifiques. Les résidus en positions -1,3 et 6 vont reconnaître 3 bases consécutives du même premier brin d'ADN (G/C riche) tandis que le résidu en position 2 reconnaît une base sur le deuxième brin d'ADN, menant ainsi à une reconnaissance croisée au niveau de l'ADN par deux doigts de zinc consécutifs (figure 14).

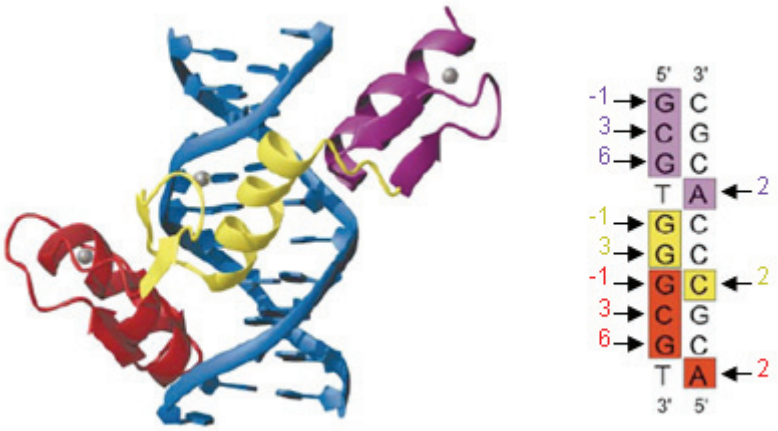


Figure 14 : **Liaison à l'ADN du domaine C2H2**
Représentation du complexe Zif268 (code PDB : 1ZAA) comprenant trois motifs doigt de zinc C2H2 représentés respectivement en rouge, jaune et violet et une molécule d'ADN représentée en bleu. Sur la séquence ADN reconnue, les bases reconnues par chacun des motifs doigt de zinc sont repérées par un code couleur respectif (rouge, jaune,violet), sont aussi figurées les positions d'acides aminés impliqués dans la reconnaissance de chaque base (-1,2,3,6). Adapté de (Pabo et al., 2001).

La région linker TGEKP

Deux doigts de zinc consécutifs sont communément reliés par une séquence conservée TGEKP (Jacobs, 1992). Des mutations sur cette région conduisent à une perte d'affinité pour l'ADN (Choo and Klug, 1993) suggérant l'importance de cette région dans la reconnaissance de l'ADN. Cette région est désordonnée dans la protéine libre en solution, mais apparaît structurée lorsque la protéine est en complexe avec l'ADN, comme le montrent des études de dynamique par RMN pour TFIIIA seul et en complexe avec l'ADN (Foster et al., 1997; Wuttke et al., 1997). Des études comparatives des valeurs de déplacement chimique $^{13}C\alpha$ pour les protéines WT1 et TFIIIA dans leurs formes libre et liée avec l'ADN ont permis de montrer que

la liaison à l'ADN conduit à une extension de l'hélicité en amont des régions TGEKP. Ce phénomène est attribué à la formation d'une coiffe de la région TGEKP sur la partie C-terminale de l'hélice α (Laity et al., 2000) par la formation d'interactions spécifiques permettant la stabilisation de l'hélice en interaction avec l'ADN. Ainsi cette région charnière flexible à l'état libre permettrait aux doigts de zinc successifs de pouvoir diffuser librement le long de l'ADN pour ensuite stabiliser le complexe spécifique ADN-protéine. Ainsi la stabilisation par la région TGEKP serait essentielle pour la reconnaissance spécifique de l'ADN.

Des doigts de zinc synthétiques

Le mode de reconnaissance de l'ADN du domaine C2H2 est le plus étudié et le mieux connu et en fait un candidat de choix pour la conception de protéines liant des cibles désirées d'ADN. Des doigts de zinc C2H2 sont ainsi synthétisés dans le but de contrôler l'expression des gènes. L'enjeu est de taille pour le biologiste moléculaire comme en témoigne le nombre important de revues relatives à ce sujet (Klug, 2005a; Klug, 2005b; Pabo et al., 2001; Papworth et al., 2006; Wolfe et al., 2000). En plus de son mode de reconnaissance simple, le domaine doigt de zinc classique ne requiert pas de séquence de reconnaissance ADN palindromique et il est fonctionnel sous forme monomérique contrairement à d'autres domaines de liaison à l'ADN. De plus, il peut être associé en tandem et répété plusieurs fois pour reconnaître spécifiquement des séquences ADN plus grande. Il offre de ce fait une possibilité de permutation et d'association combinatoire. Ainsi la méthode *phage display* (présentation de peptides à la surface de phages, (Smith, 1985)) a permis de concevoir des modules à multiples doigts de zinc (entre 3 et 6) pouvant reconnaître une séquence ADN spécifique. Ces domaines peuvent être utilisés seuls comme inhibiteurs compétitifs en masquant des sites de cibles de facteurs de transcription comme pour l'interruption du cycle infectieux du virus de l'herpes (Papworth et al., 2003). Fusionnés avec un domaine fonctionnel, les domaines à doigts de zinc C2H2 peuvent conduire à de nombreuses applications thérapeutiques (Jamieson et al., 2003), telles que l'inhibition de l'expression du virus HIV-1 (Reynolds et al., 2003), l'activation de l'expression de l'érythropoïétine (Zhang et al., 2000) ou encore l'activation de l'expression du facteur de croissance VEGF-A (Liu et al., 2001; Rebar et al., 2002). Ainsi la connaissance du mode de reconnaissance du motif C2H2 a

permis l'émergence d'une nouvelle technologie avec des applications croissantes dans le domaine de la médecine.

Le doigt de zinc GATA, Cys4

Introduction

La famille des domaines GATA regroupe des facteurs de transcription dont on retrouve des membres dans un grand nombre d'espèces allant du champignon à l'homme, chez les vertébrés et les invertébrés (Lowry and Atchley, 2000). Leur nom provient de leur capacité à lier la séquence ADN consensus (A/T) GATA (A/G). Les facteurs GATA contiennent un ou deux domaines à doigts de zinc C4 de consensus C-X_2-CX_{17}-C-X_2-C nommés N- et C- terminal le cas échéant. Pour les protéines contenant deux doigts de zinc, le domaine GATA de liaison à l'ADN comprend le doigt de zinc C-terminal suivi d'une extension comprenant des résidus basiques et conservée dans cette famille de protéines. Les facteurs GATA jouent des rôles clefs dans le processus biologique du développement, comme dans la spécification des tailles des cellules, la régulation de la différentiation et le contrôle de la prolifération et le mouvement des cellules (Patient and McGhee, 2002). Une modulation de leur expression peut d'ailleurs entraîner des maladies chez l'homme. La famille GATA n'est pas très étendue en comparaison avec d'autres familles de facteurs nucléaires. On compte six facteurs GATA chez les vertébrés, trois chez la *drosophile*, et onze chez *C. Elegans*.

Structure du domaine

Les modules à zinc de cette famille adoptent une structure similaire composée de deux épingles à cheveux β souvent irrégulières suivies d'une longue hélice α (figure 15)

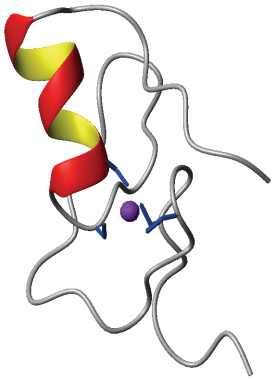


Figure 15 : **Structure du domaine C4 GATA**
(code PDB : 1GNF) (Kowalski et al., 1999)

L'atome de zinc est coordonné aux quatre cystéines de la séquence consensus qui se situent sur le premier feuillet de la première épingle, au niveau de la boucle de cette même épingle et au début de l'hélice pour les deux cystéines restantes. Cette coordination assure l'orientation correcte de la première épingle sur l'hélice. Des interactions hydrophobes stabilisent le reste de la structure notamment entre les deux épingles à cheveux, l'hélice et la deuxième épingle à cheveux et enfin la partie C-terminale avec l'hélice et la première épingle à cheveux (Omichinski et al., 1993).

Liaison à l'ADN

La structure du domaine de liaison à l'ADN GATA-1 avec sa séquence ADN cible a été résolue par RMN (Omichinski et al., 1993). C'est l'une des premières structures de complexe ADN-protéine résolue par RMN. L'organisation du complexe présentée sur la figure 16. L'hélice et la boucle en dessous de l'hélice, entre les deux épingles à cheveux β, interagissent avec le grand sillon de l'ADN tandis que la queue C-terminale contacte le petit sillon de l'ADN, directement opposé à l'hélice. Contrairement aux autres domaines de liaison à l'ADN, la majorité des contacts avec les bases de l'ADN se font par des interactions hydrophobes avec seulement trois interactions par liaison hydrogène. Cette interface hydrophobe avec des interactions de type van der Waals est favorable pour la résolution de ce complexe par RMN. En effet, les contacts hydrophobes entre les deux partenaires qui en découlent permettent d'identifier un plus grand nombre de NOEs intermoléculaires que dans le cas d'une interface polaire faisant intervenir des liaisons hydrogène.

Figure 16 : **Le domaine C4 des facteurs GATA en complexe avec l'ADN**
Représentation du domaine C-terminal cGATA-1 en complexe avec l'ADN (code PDB : 2GAT), la protéine est représentée en bleu, l'ADN en violet.

Le doigt de zinc N-terminal que l'on retrouve dans les six protéines GATA des vertébrés n'est pas nécessaire pour l'interaction avec l'ADN mais des travaux sur GATA-1 montrent qu'il augmente la stabilité et la spécificité du complexe (Martin and Orkin, 1990) et qu'il peut lier un double site (A/T)GATA(A/T) en conjugaison avec le doigt de zinc N-terminal (Trainor et al., 1996). Il a été également décrit que le doigt de zinc N-terminal des protéines GATA-1, GATA-2 et GATA-3 pouvait reconnaître de manière autonome la séquence GATC alors que le doigt de zinc C-terminal reconnaît plus spécifiquement la séquence GATA (Newton et al., 2001). Il a été aussi montré que le doigt de zinc N-terminal des protéines GATA était impliqué dans des interactions protéine-protéine avec d'autres protéines à doigt de zinc, notamment sp1 (Newton et al., 2001), EKLF (Erythroid Krüppel Like Factor) (Newton et al., 2001), FOG (Friend Of GATA) (Tsang et al., 1997), et GATA-1 elle-même (Crossley et al., 1995). Ces études montrent que les protéines à doigt de zinc sont impliquées dans des interactions protéine-protéine et ne sont pas restreintes aux interactions protéine-ADN (Liew et al., 2005; Mackay and Crossley, 1998). La structure du complexe entre le doigt de zinc C2HC (dérivant du doigt de zinc classique C2H2) de FOG et le doigt de zinc N-terminal C4 de GATA-1 a été déterminée avec le logiciel HADDOCK (Dominguez et al., 2003) en utilisant des contraintes de NOEs intermoléculaires et des données de cartographie par variations de déplacements chimiques (*chemical shift mapping*) (Liew et al., 2005) (figure 17). FOG se lie à GATA-1 par son hélice α. Ainsi, la surface de reconnaissance protéine-protéine de

FOG est la même que celle normalement utilisée par les doigts de zinc classiques pour la reconnaissance à l'ADN. Au contraire, dans le cas de GATA-1 la reconnaissance protéine-protéine met en jeux les épingles à cheveux β, particulièrement la première. Les surfaces de reconnaissance de l'ADN et de FOG ne sont donc pas superposées. Le doigt de zinc N-terminal de GATA-1 est donc l'exemple d'un motif de reconnaissance à la fois d'ADN et de protéine et présentant deux surfaces différentes pour ces deux fonctions, malgré sa petite taille (30 acides aminés).

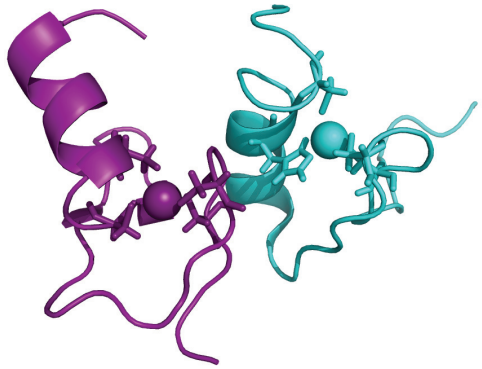


Figure 17 : **Structure du complexe entre FOG et GATA-1**
Le doigt de zinc de FOG est représenté en bleu et celui de GATA-1 en violet. Les atomes de zinc sont figurés par des sphères. Les chaînes latérales des résidus liant le zinc sont représentées. (code pdb : 1YoJ) (Liew et al., 2005)

Le domaine de liaison à l'ADN des récepteurs nucléaires, Cys8

Introduction

Les récepteurs nucléaires constituent une des plus grandes familles de facteurs de transcription. On les retrouve chez les vertébrés, les nématodes et les arthropodes où ils contrôlent un grand nombre de processus biologiques impliqués dans le développement, la croissance, la différenciation cellulaire, l'apoptose (Renaud and Moras, 2000). Dans la plupart des cas, ils régulent des gènes par transcription en réponse à une interaction avec un ligand, en général lipophile. On retrouve ainsi parmi les récepteurs nucléaires connus, les récepteurs des hormones stéroïdiennes, des hormones thyroïdiennes, des acides rétinoïques, et de la vitamine D (Mangelsdorf et al., 1995). Leurs activités dépendantes de la liaison d'un ligand en font des cibles pharmacologiques pour lesquelles des agonistes et antagonistes ont

été développés. Ils sont utilisés par exemple en contraception, dans le contrôle de l'inflammation et dans le traitement de nombreuses maladies dont le diabète, certaines maladies hormonales et certains cancers. Néanmoins, il existe un ensemble de récepteurs, définis comme orphelins, pour lesquels aucun ligand n'a été identifié (Mangelsdorf and Evans, 1995).

Structure du domaine de liaison à l'ADN

Les récepteurs nucléaires sont constitués de différents domaines dont un domaine de liaison au ligand et un domaine de liaison à l'ADN qui sont conservés dans la famille et qui peuvent fonctionner de manière autonome. Le domaine de liaison à l'ADN comprend 66 acides aminés fortement conservés dans la famille et se situe au centre des protéines de cette famille. Les premières structures de ce domaine furent résolues en 1990 par RMN (Hard et al., 1990; Schwabe et al., 1990). Une coordination autour de deux atomes de zinc met en jeu quatre cystéines, pour chaque atome de zinc, comme ligands. Cela permet de définir ce domaine Cys4-Cys4. La structure est composée de deux boucles liant un atome de zinc et de deux hélices α situées après chacune de ces boucles, formant ainsi deux motifs séparés boucle-hélice. Les deux hélices sont orientées perpendiculairement l'une par rapport à l'autre, et se croisent en leur centre avec la formation d'un cœur hydrophobe à leur interface (figure 18).

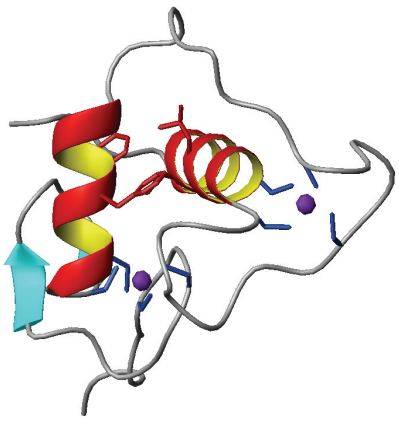


Figure 18 : **Le domaine de liaison à l'ADN Cys8 des récepteurs nucléaires**
Représentation de la structure du domaine de liaison à l'ADN du récepteur aux œstrogènes (code PDB : 1HCP) (Schwabe et al., 1993b). Les ions zinc sont représentés en violet, les ligands au zinc en bleu, les résidus du cœur hydrophobe en rouge.

Liaison à l'ADN

Les récepteurs nucléaires régulent l'expression de gènes en se liant au niveau de séquences spécifiques, les éléments de réponse aux hormones. Deux motifs consensus de six paires de bases ont été identifiés, AGAACA reconnu principalement par les récepteurs aux hormones stéroïdiennes et AGGTCA reconnu par l'ensemble des autres récepteurs aux hormones non stéroïdiennes et les récepteurs orphelins (Khorasanizadeh and Rastinejad, 2001). Les récepteurs nucléaires se lient aux éléments de réponse sous forme de monomères, d'homo-dimères ou d'hétéro-dimères. La disposition des séquences pour un dimère peut se configurer en palindrome, en palindrome inversé ou en répétition directe (figure 19) (Mangelsdorf et al., 1995). L'arrangement qui en découle confère une spécificité et une modulation de la liaison des récepteurs à l'ADN.

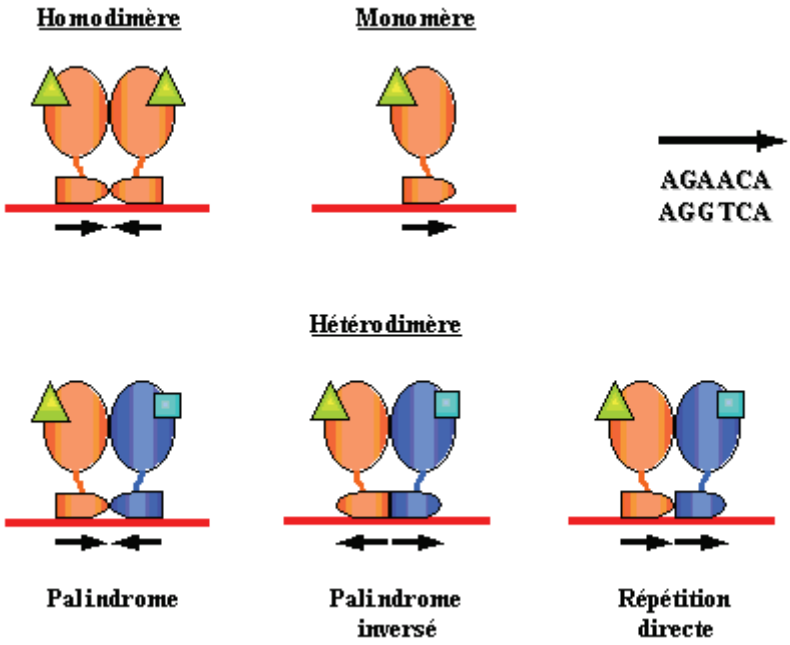


Figure 19 **: Différents modes de liaison des récepteurs nucléaires à l'ADN**
Les récepteurs peuvent exercer leur fonction régulatrice de la transcription sous forme d'homodimère, de monomère ou d'hétérodimère. Les séquences consensus ADN reconnues, illustrées par des flèches, peuvent être disposées en palindrome, en palindrome inversés ou en répétition directe.

La première structure d'un complexe du domaine de liaison à l'ADN d'un facteur nucléaire avec son ADN a été résolue en 1991 par radiocristallographie, correspondant au domaine de liaison à l'ADN du récepteur aux glucocorticoïdes (Luisi et al., 1991).

L'hélice 1, nommée hélice de reconnaissance, s'insère dans le grand sillon de l'ADN, permettant à ses chaînes latérales en surface de faire des contacts spécifiques avec les bases de l'ADN consensus. Dans le cas de dimères, les deux molécules de protéines contactent de cette même façon l'ADN sur des grands sillons adjacents. La protéine fait également un grand nombre de contacts avec le squelette phosphate de l'ADN de chaque côté, permettant d'orienter correctement l'hélice de reconnaissance dans le grand sillon. De plus dans le cas du dimère, une dimérisation coopérative s'effectue lors de la complexation à l'ADN, médiée par la région C-terminale du domaine de liaison à l'ADN, contenant la deuxième coordination au zinc (figure 20).

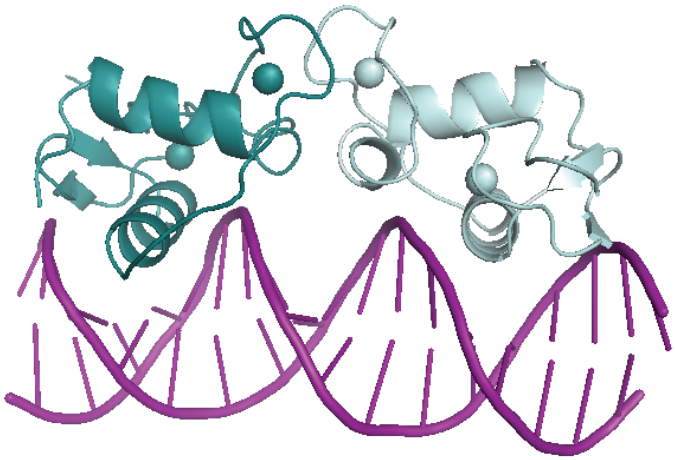


Figure 20 : **Complexe du récepteur nucléaire avec l'ADN**
Représentation du domaine de liaison à l'ADN du récepteur nucléaire aux œstrogènes avec sa séquence cible (code PDB : 1HCQ) (Schwabe et al., 1993a). La reconnaissance se fait avec un dimère, représenté en bleu, sur une séquence répétée palindromique, représentée en violet

Nous venons de décrire plusieurs domaines de liaison à l'ADN. Nous pouvons désormais étudier plus particulièrement la reconnaissance protéine-ADN et les différents aspects de ce processus.

La reconnaissance protéine-ADN

Introduction

La reconnaissance protéine-ADN joue un rôle important dans différents processus biologiques comme la réplication, la transcription ou encore la réparation de l'ADN. Parmi les protéines interagissant avec l'ADN, les facteurs de transcription représentent une des classes les plus variées et les plus abondantes (Pabo and Sauer, 1992), qui permettent la régulation de l'expression des gènes. Dans ce contexte l'étude de la reconnaissance protéine-ADN est fondamentale pour comprendre comment l'information génétique est utilisée au niveau de la transcription. Ainsi la définition d'un code de reconnaissance protéine-ADN a toujours été une question centrale ; les premières études ont été menées par Seeman et ses collaborateurs en 1976 (Seeman et al., 1976). Même si nous verrons qu'un tel code universel, analogue au code génétique, n'existe pas, beaucoup d'efforts sont menés pour comprendre la spécificité d'interaction et notamment dans le but de construire des domaines synthétiques de liaison à une séquence donnée d'ADN (Pabo et al., 2001).

Les données structurales sur les complexes protéine-ADN fournissent des clefs pour comprendre les principes de reconnaissance protéine-ADN. Malgré un nombre important de structures de complexes protéine-ADN résolues, les mécanismes expliquant leur reconnaissance spécifique restent toujours peu connus (Sarai and Kono, 2005). Néanmoins les principes généraux de leur reconnaissance peuvent être dégagés pour expliquer la spécificité de reconnaissance.

Thermodynamique de l'interaction

Comme dans pour toute interaction entre deux molécules, les molécules de protéine et d'ADN interagissent s'il y a une baisse de l'énergie libre lors de la formation du complexe. La variation d'énergie libre (ΔG) lors de la formation du complexe dépend de la variation d'enthalpie (ΔH) et d'entropie (ΔS) selon :

$$\Delta G = \Delta H - T.\Delta S$$

Dans le cas des complexes protéine-ADN une baisse d'enthalpie, favorable à la formation du complexe, provient principalement de la formation de nombreuses liaisons non-covalentes entre les deux partenaires. Une hausse d'entropie, favorable aussi à la formation du complexe, provient elle principalement d'une libération de molécules d'eau ordonnée à la surface des partenaires qui sont libérées lors de la formation du complexe (Rhodes et al., 1996). De ce fait, une contribution favorable au ΔG est assurée par une complémentarité de forme des deux partenaires qui va permettre la formation des liaisons non-covalentes à courte distance et la déshydratation des surfaces en contact. Certaines molécules d'eau vont être emprisonnées entre les deux molécules et vont permettre la formation de liaisons hydrogène. De plus, la thermodynamique de formation de complexes spécifiques protéine-ADN est également régie par des phénomènes de complémentarité de charges et de potentiels électrostatiques ainsi que des réarrangements conformationnels et des changements de dynamique (Hard and Lundback, 1996). L'ensemble de ces phénomènes participant à la thermodynamique de formation du complexe protéine-ADN est schématisé sur la figure ci-dessous (figure 21)

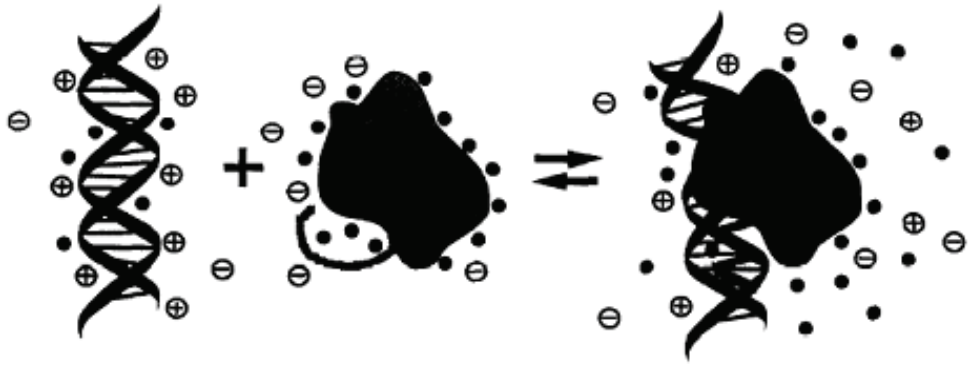


Figure 21 : **Thermodynamique de formation du complexe protéine-ADN**
Représentation schématique des phénomènes impliqués dans la thermodynamique de formation du complexe. L'image représente une molécule d'ADN chargée négativement, entourée de contre-ions (cations) et solvatée par des molécules d'eau (disques noirs) et une molécule de protéine chargée positivement, également solvatée et entourée d'anions. Le complexe formé présente des complémentarités de forme et de charge entre les deux partenaires. La formation du complexe implique des changements de conformation des deux molécules, une déshydratation des surfaces d'interaction mais également des changements de dynamique (non représenté). Adapté de (Hard and Lundback, 1996)

L'étude thermodynamique de tels complexes participe à la compréhension de la reconnaissance protéine-ADN, en particulier pour étudier les spécificités d'interaction

et les affinités des complexes protéine-ADN (Jen-Jacobson, 1997; Privalov et al., 2007; Spolar and Record, 1994).

La reconnaissance de forme

L'ensemble des structures de complexes protéine-ADN présentent une complémentarité de forme remarquable. Nous pouvons nous en rendre compte sur une structure de complexe avec une représentation des surfaces moléculaires (figure 22)

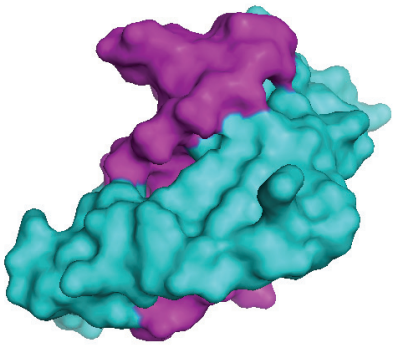


Figure 22 : **La complémentarité de forme**
Représentation du complexe Zif268 (code PDB : 1ZAA) avec la surface de la protéine en bleu et celle de l'ADN en violet.

Ainsi dans beaucoup de structures de complexes connues, on retrouve une complémentarité de forme entre une hélice α de la protéine, appelée hélice de reconnaissance, et le grand sillon de l'ADN. C'est le cas pour les domaines que nous avons présentés jusqu'à présent comme le doigt de zinc classique et l'homéodomaine (figure 23). Ces deux éléments de structure, hélice α et grand sillon, ont une compatibilité remarquable. Les chaînes latérales des résidus exposés sur l'une des faces de l'hélice vont ainsi pouvoir interagir directement avec les bases de l'ADN. Cependant, de nombreux contacts sont généralement formés par des régions de la protéine en dehors de cette hélice et participent également à la spécificité de l'interaction. De plus, l'orientation précise de cette hélice par rapport au grand sillon varie grandement entre les différents domaines de liaison à l'ADN (Garvie and Wolberger, 2001) et ce mode de reconnaissance n'est en aucun cas universel même si il est très récurrent. Il existe des domaines de liaison à l'ADN utilisant des feuillets

β, par exemple la protéine TBP (Kim et al., 1993a; Kim et al., 1993b), ou des boucles, comme NF-κB (Ghosh et al., 1995; Muller et al., 1995), pour former des contacts spécifiques nécessaires à la reconnaissance de leur séquence ADN cible (figure 23).

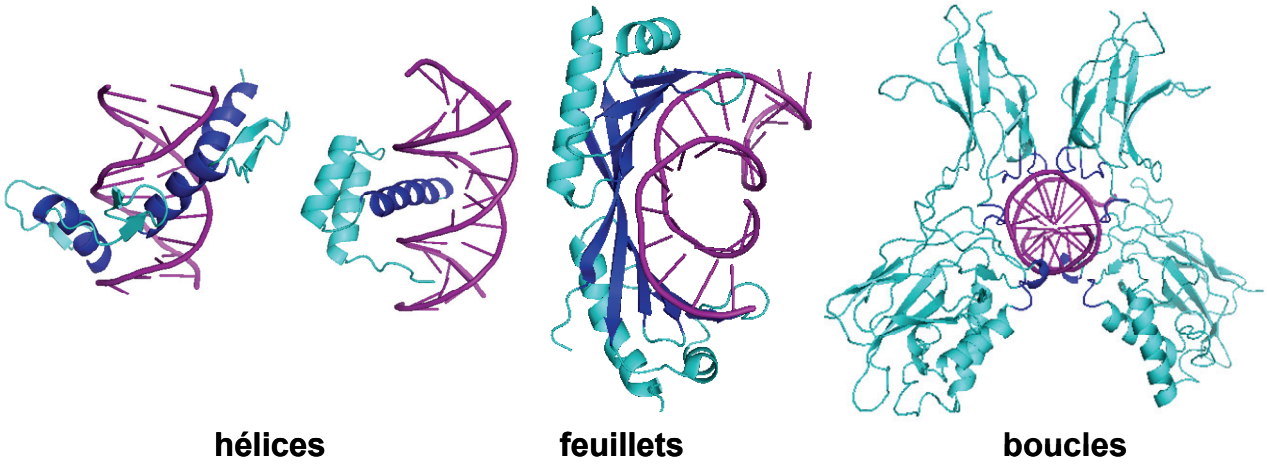


Figure 23 : **Différents éléments de reconnaissance**

Représentation de complexe protéine-ADN mettant en jeux des éléments de structure différents pour la reconnaissance d'ADN, des hélices : doigt de zinc Zif268 (code PDB : 1ZAA) (Pavletich et al., 1991) et homéodomaine engrailed (code PDB : 1HDD) (Kissinger et al., 1990), des feuillets : TBP (code PDB : 1YTB) (Kim et al., 1993b) et des boucles : NFκB (code PDB : 1NFK) (Ghosh et al., 1995). Les éléments de reconnaissance sont représentés en bleu foncé, les molécules de protéine en bleu clair et d'ADN en violet.

Les surfaces d'interaction des complexes protéine-ADN ont donc des formes très complémentaires. L'ADN ayant une forme assez uniforme, il n'est pas étonnant que différents domaines de liaisons à l'ADN aient employé les mêmes stratégies architecturales pour parvenir à une forme complémentaire de celle de l'ADN. Voyons désormais la nature des interactions entre ces surfaces complémentaires qui définissent la reconnaissance chimique.

La reconnaissance chimique

L'ADN double brin présente un squelette sucre/phosphate chargé négativement et des paires de bases empilées qui sont exposées dans le grand et le petit sillon, rendant ainsi accessibles les groupes fonctionnels de chacune de ces bases. Une protéine peut ainsi reconnaître une séquence spécifique si elle possède une surface chimiquement compatible avec celle de l'ADN (Garvie and Wolberger, 2001).

Les interactions mises en jeu comprennent les liaisons hydrogène, les forces de van der Waals, les interactions hydrophobes, les interactions électrostatiques et les ponts salins. Les structures des complexes protéine-ADN révèlent le réseau tridimensionnel d'interactions qui tient les deux molécules ensemble. Des contacts avec le squelette sucre/phosphate de l'ADN permettent d'orienter la protéine de façon à positionner les éléments de structure secondaire impliqués dans des interactions spécifiques. Ces contacts non-spécifiques avec le squelette de l'ADN représentent la majorité des interactions protéine-ADN et assurent la stabilité du complexe (Luscombe et al., 2001).

La spécificité de reconnaissance directe provient essentiellement d'un réseau complexe de liaisons hydrogène entre les chaînes latérales des acides aminés de la protéine et les groupements fonctionnels des bases. La majeure partie des contacts formés implique des liaisons hydrogène de type bidenté (un acide aminé forme plusieurs contacts avec une base ou une paire de bases) ou complexe (un acide aminé interagit avec plusieurs bases). Ce type de liaisons hydrogène permet d'augmenter le nombre de contacts entre la protéine et les bases de l'ADN et améliore la spécificité d'interaction (Luscombe et al., 2001).

L'observation de ces interactions spécifiques pose la question de l'existence d'un code de reconnaissance reliant spécifiquement un acide aminé et une base.

Existence d'un code de reconnaissance ?

L'existence d'un code de reconnaissance protéine-ADN analogue au code génétique est une question récurrente. En 1976, Seeman et ses collaborateurs (Seeman et al., 1976) mettent en évidence que les protéines liant l'ADN de façon séquence spécifique semblent interagir avec les bases dans le grand sillon de l'ADN, où la répartition des donneurs et accepteurs de liaisons hydrogène est unique pour chaque paire de bases. L'étude prévoit ainsi que les résidus asparagine et acide glutamique contactent des adénines et que le résidu arginine contacte les guanines. Bien que ces contacts soient observés dans plusieurs complexes, ces relations présentent des variations. En effet, les interfaces protéine-ADN sont beaucoup trop complexes et il n'est pas possible de définir un code de reconnaissance universel reliant un acide aminé à une base.

Il est toutefois possible de dégager des principes de reconnaissance qui s'appliquent à une famille donnée. C'est le cas pour la famille des doigts de zinc C2H2 qui présentent un mode de reconnaissance bien connu et qui fait l'objet de construction de domaines synthétiques liant une séquence spécifique, comme nous l'avons vu précédemment. Ainsi Desjarlais et Berg (Desjarlais and Berg, 1992) puis Choo et Klug ont défini des règles pour la reconnaissance des séquences ADN par les domaines C2H2 (Choo and Klug, 1994; Choo and Klug, 1997). Ces règles mettent en relation le type d'acide aminé impliqué dans la reconnaissance spécifique selon leur position -1,2,3 ou 6 (figure 12) avec une base préférablement reconnue (table 3).

Position on Helix	T	C	A	G
	Base Preference			
6	Lys?		Gln?	**Arg*** Lys*
3	Ser Ala	**Asp*** Thr Glu Ser*	**Asn*** His*	**His*** Lys*
-1	**Thr** Leu* His	**Asp*** His	**Gln***	**Arg***
2		Asp*	Asp*	

Table 3 : **Code de reconnaissance pour les doigts de zinc C2H2**
La table indique le type d'acide aminé préférentiel en fonction de sa position spécifique sur l'hélice de reconnaissance et selon le type de base reconnu. Les acide aminés en gras sont ceux qui apparaissent le plus fréquemment dans les expériences de phage display, ceux marqués d'un astérisque on été observés dans des études structurales. Adapté de (Pabo et al., 2001)

Cependant, cette correspondance semble plus efficace pour concevoir des domaines C2H2 liant une certaine séquence d'ADN plutôt que pour prévoir réellement une séquence reconnue par un domaine C2H2 donné.

Plus récemment, Luscombe et ses collaborateurs ont répertorié les interactions de type van der Waals, les liaisons hydrogène et les liaisons à travers les molécules d'eau pour 129 complexes protéine-ADN (Luscombe et al., 2001). Bien que les contacts avec le squelette de l'ADN soient les plus nombreux, des préférences d'interaction entre les chaînes latérales des protéines, selon le type d'acide aminé, et les ADN, selon le type de bases, ont pu être dégagées. Benos et collaborateurs

proposent aussi un code basé en se basant sur des probabilités de reconnaissance (Benos et al., 2002).
Ainsi, bien qu'un code de reconnaissance universel reliant une base à un acide aminé n'existe pas, des préférences d'interaction sont souvent observées, pour une famille de doigt de zinc donnée, même si elles ne constituent pas une règle absolue.

Rôle de la structure de l'ADN

Etant donnée la grande diversité des structures des protéines liant l'ADN comparée à celles de l'ADN, l'interaction spécifique entre les protéines et l'ADN est souvent analysée du point de vue de la protéine. Toutefois plusieurs considérations sont à prendre au niveau de la structure de l'ADN. La double hélice d'ADN adopte des conformations différentes selon la séquence et le degré d'hydratation (Arnott and Selsing, 1974a; Arnott and Selsing, 1974b). Deux modèles pour les conformations extrêmes ont été définis, les formes A et B. Toutefois ces modèles ne reflètent pas complètement l'ADN en solution qui a une flexibilité et une structure en double hélice constamment variables qui sont à prendre en considération dans le mécanisme de reconnaissance protéine-ADN. La protéine qui reconnaît spécifiquement une séquence ADN doit reconnaître le squelette sucre/phosphate de l'ADN mais surtout les bases qui sont spécifiques. Ces bases ne sont accessibles directement, par des liaisons hydrogène, que dans le petit et le grand sillon. L'accessibilité aux bases dépend alors de la déformation et de la variation de structure de l'ADN. Ainsi un ADN sous la forme B présente un grand sillon plus grand et plus accessible que le petit sillon à l'inverse d'un ADN sous la forme A (Rhodes et al., 1996). De plus la répartition des accepteurs et donneurs de liaisons hydrogène est unique pour chaque type de base dans le grand sillon mais pas dans le petit sillon. Il n'est donc pas étonnant de constater qu'un grand nombre d'interactions protéine-ADN se fait via le grand sillon d'un ADN sous forme B par une hélice α de forme complémentaire.
La reconnaissance directe de la séquence par les bases est importante pour expliquer la spécificité, mais il ne faut pas négliger la reconnaissance indirecte par le squelette sucre/phosphate qui peut avoir des rôles différents selon les complexes.

Enfin la déformation de l'ADN peut s'avérer nécessaire pour la reconnaissance spécifique, pour adopter une surface complémentaire à celle de la protéine où lorsqu'il y a des contacts de la protéine avec les bases dans le petit sillon. Dans le cas extrême de la protéine TBP qui reconnait la séquence TATAAA, l'ADN présente deux coudes de 90°, permettant à la protéine de se lier à l'ADN au niveau du petit sillon (Kim et al., 1993a). L'interaction se fait alors avec un feuillet β de forme compatible avec la taille du petit sillon (figure 24). Dans le cas d'ADN courbé, la perte d'énergie que représente cette déformation est compensée par la formation de liaisons non covalentes entre la protéine et l'ADN.

Figure 24 : **Déformation de l'ADN**
Illustration de la déformation de la double hélice d'ADN, représentée en violet, dans le cas de la liaison à la protéine TBP, représentée en bleu. (code PDB : 1YTB).

La reconnaissance non spécifique

Si la résolution de nombreux complexes protéine-ADN permet de comprendre comment une protéine peut interagir spécifiquement avec une molécule d'ADN, il reste à savoir comment une protéine est capable d'identifier sa cible parmi l'énorme quantité d'ADN non-spécifique.

De façon analogue au paradoxe de Levinthal dans le cas de la structuration des protéines, les protéines semblent trouver leur cible ADN beaucoup plus rapidement que ne le permet la simple diffusion tridimensionnelle des molécules. Ainsi dès 1970, Riggs et ses collaborateurs ont mesuré un taux d'association du répresseur au lactose d'*E .Coli* à son opérateur de ~10^{10} $M^{-1}S^{-1}$ soit 100 à 1000 fois supérieur à ce qui est prévu par la diffusion simple en trois dimensions (Riggs et al., 1970a; Riggs et

al., 1970b). Ce paradoxe peut être résolu si on prend en compte deux modes de liaisons (Halford and Marko, 2004; Slutsky and Mirny, 2004). Un premier mode de liaison non-spécifique faisant intervenir des liaisons électrostatiques permettrait à la protéine de lier la protéine à l'ADN pour ensuite inspecter la molécule d'ADN jusqu'à la formation d'un complexe spécifique. Cette inspection pourrait se faire par un processus de diffusion à une dimension (von Hippel and Berg, 1989) en glissant le long de la molécule ou par une diffusion à trois dimensions (Gowers and Halford, 2003; Halford and Marko, 2004; Slutsky and Mirny, 2004) par une succession d'associations et dissociations sur la même molécule d'ADN. De cette façon, la recherche de la séquence spécifique serait plus rapide que si elle était réalisée aléatoirement.

Ainsi la formation d'un complexe intermédiaire non-spécifique est importante dans le processus de reconnaissance de l'ADN. Dans le cas du répresseur au lactose, les structures du complexe non spécifique et spécifique ont été résolues par RMN (Kalodimos et al., 2004a; Kalodimos et al., 2004b) en utilisant des ADN de séquences non spécifiquement et spécifiquement reconnues.

Globalement la protéine, qui interagit sous forme dimérique, garde la même structure. Elle s'incline de ~25° entre les complexes non-spécifique et spécifique engendrant une perte de contacts protéine-ADN dans le cas du complexe non-spécifique (figure 25).

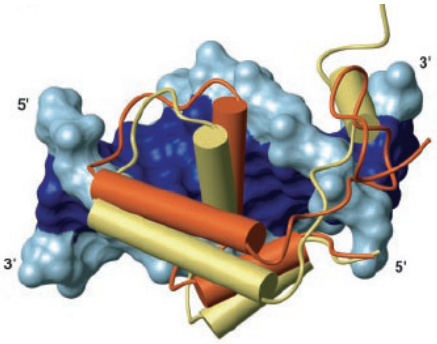


Figure 25 : **Modes d'interaction spécifique et non-spécifique**
Représentation du complexe non-spécifique et spécifique du domaine de liaison à l'ADN du répresseur au lactose, respectivement en jaune et rouge, avec l'ADN, en bleu. La structure globale de la surface d'interaction avec le grand sillon de l'ADN est identique, mais la protéine est basculée de 25° entre les deux types de complexe. Adapté de (Kalodimos et al., 2004a)

De plus un changement conformationnel local se produit dans le cas du complexe spécifique, avec la structuration en hélice α de la partie C-terminale qui va s'insérer dans le petit sillon et induire une coubure de l'ADN (figure 26).

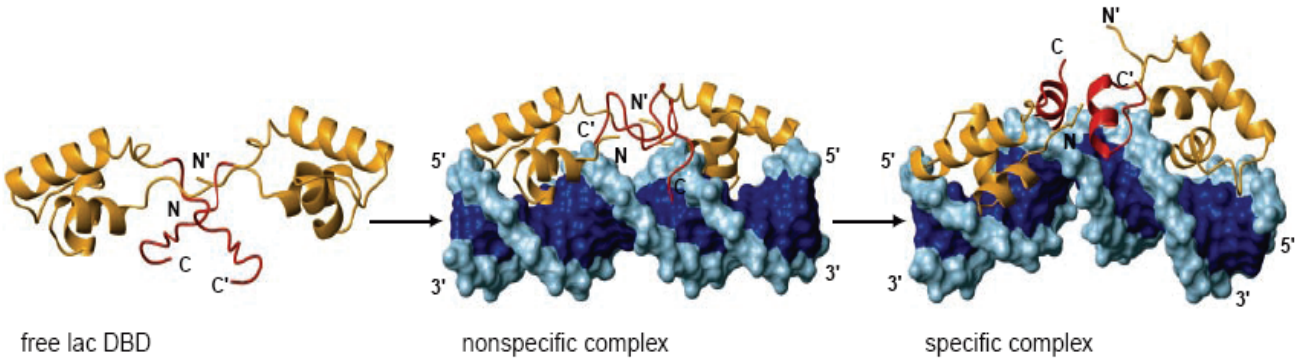


Figure 26 : **Mécanisme structural de la liaison à l'ADN du répresseur au lactose**
Lors de la formation du complexe non-spécifique, l'ADN et la protéine ont des structures similaires à leur état libre. Lors de la formation du complexe spécifique, la région C-terminale, en rouge, se structure en hélice α et s'insère dans le petit sillon. Adapté de (Kalodimos et al., 2004a)

Ainsi la formation du complexe non-spécifique permet de positionner correctement les éléments de structure de la protéine par rapport à l'ADN via des contacts principalement électrostatiques, indépendants de la séquence. Les résidus impliqués dans la reconnaissance spécifique sont alors en contact du grand sillon. Des mutations de ces résidus sur les hélices de reconnaissance affectent l'affinité du complexe spécifique mais aussi du complexe non-spécifique, révélant le double rôle des hélices qui assurent la spécificité d'interaction et la stabilité du complexe non spécifique (Kalodimos et al., 2004a). Ainsi la mutation de la tyrosine 17, située sur l'hélice de reconnaissance, en phénylalanine, soit la déplétion d'un groupement OH, abaisse de ~100 fois l'affinité à l'ADN de séquence spécifique et de ~10 fois l'affinité à la séquence non-spécifique. L'affinité pour l'ADN non-spécifique est par ailleurs 10^7 fois plus faible que celle pour l'ADN spécifique.

Des études de relaxation ^{15}N, avec la détermination des valeurs de vitesses d'échange, permettent de mettre en évidence une grande flexibilité des résidus impliqués dans la reconnaissance spécifique, qui leur permet d'échantillonner différents environnements au niveau des bases de l'ADN pour passer en interaction spécifique lorsqu'ils se trouvent au contact de la séquence cible. Des études d'échange deutérium-proton ont montré que la protéine est plus protégée de l'échange chimique de ses protons labiles lors de la formation du complexe spécifique que lors de la formation du complexe non-spécifique (figure 27). De même la protéine est plus rigide lors du passage du complexe non-spécifique au complexe spécifique.

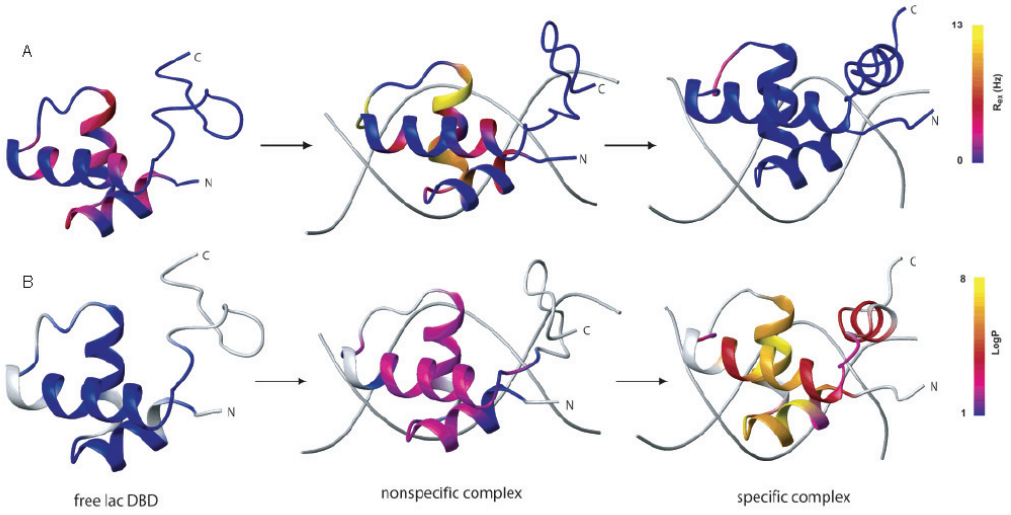


Figure 27 : **Etude dynamique et d'échange hydrogène lors du processus de reconnaissance**
A Représentation des valeurs des vitesses d'échange selon un code de couleur. L'hélice de reconnaissance présente des valeurs d'échange élevées (jaune) dans le complexe non-spécifique. Les valeurs d'échange sont nulles (bleu) dans le complexe spécifique. B Représentation du facteur de protection selon un code de couleur. La protéine est de plus en plus protégée au cours du processus de reconnaissance. Adapté de (Kalodimos et al., 2004b)

Cette étude permet de progresser dans la compréhension du mécanisme de reconnaissance protéine-ADN et montre de façon intéressante que des changements conformationnels, comme la formation d'hélice α en interaction dans le petit sillon, permettent de stabiliser le complexe spécifique.

Cinétique de l'interaction protéine-ADN

Nous venons de voir quelques grands principes du mécanisme de reconnaissance protéine-ADN avec la formation d'un complexe non-spécifique intermédiaire. Les principes plus détaillés du mécanisme de reconnaissance et de la cinétique qui en découle restent encore peu connus (Kalodimos et al., 2004b). En effet il s'agit d'un processus biologique complexe, avec des phases d'association et de dissociation (Gowers and Halford, 2003) faisant intervenir des intermédiaires, comme des complexes non spécifiques mais également des changements conformationnels.
Le mécanisme d'association est ainsi composé par au moins deux étapes, classiquement une étape de liaison suivie par un réarrangement structural, soit selon le schéma (Halford and Marko, 2004):

$$A+B \leftrightarrow AB^* \leftrightarrow AB$$

Ainsi par exemple, des mesures de fluorimétrie ont mis en évidence un mécanisme à deux étapes avec la succession des phénomènes de liaison de la protéine sur l'ADN puis de courbure de l'ADN dans le cas de la protéine IHF (Khrapunov et al., 2006; Kuznetsov et al., 2006; Sugimura and Crothers, 2006) (figure 28)

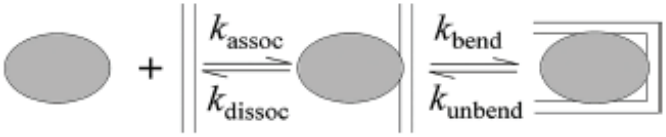


Figure 28 : **Mécanisme de liaison protéine-ADN en deux étapes : association protéine-ADN puis courbure de l'ADN**
Adapté de (Kuznetsov et al., 2006).

Ce genre d'étude permet de progresser dans la connaissance du mécanisme de reconnaissance, et même si une courbure de l'ADN n'a pas forcément lieu lors de la reconnaissance protéine-ADN, elle est très fréquente.
Il est donc possible d'envisager des mécanismes à plusieurs étapes faisant intervenir des intermédiaires non-spécifiques et des changements de conformation de la protéine et / ou de l'ADN.
Mais l'étude cinétique d'un système protéine-ADN n'est pas aisée car les intermédiaires intervenant dans le mécanisme ne sont pas ou peu connus et il n'est pas toujours possible de mesurer leur présence en vue de l'étude cinétique complète du système (Gutfreund, 1987). De même la mise en œuvre d'un système d'étude comprenant un complexe spécifique et non-spécifique reste difficile.
Toutefois des approches sont développées pour la détermination et l'étude cinétique du mécanisme de reconnaissance protéine-ADN. Par exemple, Iwhara et ses collaborateurs ont étudié la liaison à l'ADN de l'homéodomaine HOXD9. Ils ont étudié la liaison de ce domaine sur différentes séquences d'ADN de façon à former des complexes spécifiques et non-spécifiques en utilisant différentes techniques. Des expériences de mesure de taux de relaxation R2 avec des centres paramagnétiques (Paramagnetic Relaxation Enhancement) ont permis de mettre en évidence des intermédiaires transitoires dans le processus de reconnaissance ainsi que des phénomènes de translocation le long de la même molécule d'ADN (diffusion à une dimension) et d'une molécule d'ADN à une autre (diffusion en trois dimensions) (Iwahara and Clore, 2006a). Des expériences d'échange d'aimantation longitudinale sur le noyau ^{15}N (Farrow et al., 1994) sur deux cibles ADN spécifiques différentes ont

permis de caractériser cette cinétique. Les constantes cinétiques mesurées sont en accord avec un processus impliquant un transfert direct d'une molécule d'ADN à une autre, par collision plutôt qu'un simple mécanisme de dissociation/réassociation (Iwahara and Clore, 2006b) (figure 29). Des mesures du taux de relaxation R2 sur des cibles d'ADN non-spécifiques ont permis d'étudier la cinétique de translocation d'une cible à une autre en accord avec le mécanisme proposé dans le cas d'ADN spécifique, c'est-à-dire sans intermédiaire avec la protéine libre (Iwahara et al., 2006).

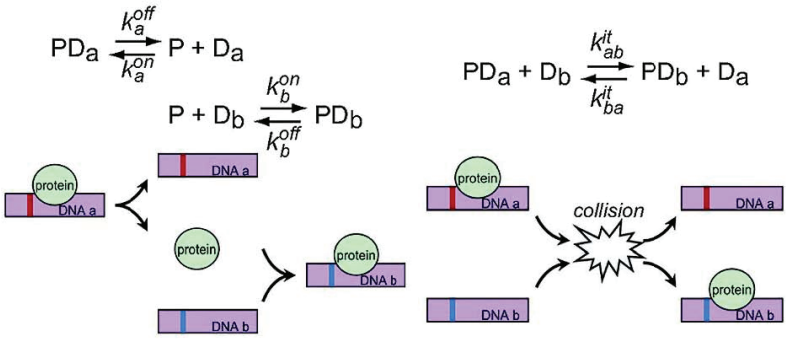


Figure 29 : **Mécanisme de translocation d'un site ADN à un autre**
Deux mécanismes proposés pour la translocation d'un domaine de liaison à l'ADN d'un site à un autre. Le mécanisme par collision, à droite, est en accord avec les données de cinétique déterminées pour l'homéodomaine HOXD9 dans le cas de cibles spécifiques et non-spécifiques. Adapté de (Iwahara and Clore, 2006b)

Ce mécanisme par collision est intéressant car il permet d'expliquer la rapidité qu'ont les protéines de liaison à l'ADN à trouver leur cible. Il est démontré dans ces études que les constantes de vitesse sont proportionnelles à la concentration en ADN, en accord avec le mécanisme par collision, mettant en évidence l'importance de la liaison à l'ADN non-spécifique, en concentration élevée dans le noyau, dans le processus de reconnaissance rapide protéine-ADN.

Détermination de structure par RMN

Introduction

Deux méthodes principales existent pour la détermination de la structure tridimensionnelle à résolution atomique des protéines : la diffraction de rayons X sur des cristaux de protéines et la Résonance Magnétique Nucléaire. Les méthodes fondées sur la diffraction électronique fournissant généralement des structures à plus basse résolution, de l'ordre de 2 à 5 Å. Si la radiocristallographie n'est pas limitée par la taille de la protéine et est la méthode qui a fourni et continue de fournir le plus grand nombre de structures 3D, elle nécessite néanmoins l'obtention de monocristaux qui diffractent, ce qui est souvent une étape limitante. La RMN permet l'étude des protéines en solution (Wüthrich, 1986), mais elle reste limitée par la taille de la protéine étudiée. Ces deux techniques sont à plusieurs titres complémentaires. La RMN offre l'avantage de donner accès à la dynamique interne des protéines (Boehr et al., 2006; Fischer et al., 1998; Korzhnev et al., 2001). Elle permet également l'étude structurale, thermodynamique et cinétique d'interactions entre protéine-protéine (Bonvin et al., 2005; Clore and Gronenborn, 1998; Zuiderweg, 2002), ou protéine-ADN ou protéine-ligand en solution. Ainsi il est possible de concevoir par RMN de nouveaux ligands d'affinité accrue à l'issue d'études de relations de structure activité, pour la recherche d'inhibiteurs dans le cadre de la conception de médicaments (Carlomagno, 2005; Pellecchia et al., 2002; Shuker et al., 1996).

Dans ce chapitre, nous nous efforcerons de décrire la stratégie que nous avons utilisée pour résoudre la structure du domaine THAP de THAP1. D'autres approches que nous n'avons pas mises en œuvre telles que l'usage des couplages dipolaires résiduels (Bax, 2003; de Alba and Tjandra, 2002; Prestegard et al., 2004; Tolman, 2001) ou des stratégies impliquant des marquages sélectifs (Kainosho et al., 2006; Staunton et al., 2006) ne seront pas développées. De même, nous n'avons pas développé les principes de base de la Résonance Magnétique Nucléaire, les séquences d'impulsion, l'acquisition des spectres, le traitement du signal que l'on peut retrouver par exemple dans les ouvrages suivants : (Keeler, 2002; Sattler et al., 1999; Wüthrich, 1986). Par ailleurs, nous illustrerons nos propos avec les données

que nous avons recueillies pour la détermination de la structure du domaine THAP de THAP1. Cette stratégie a été utilisée pour la première fois en 1984 pour la détermination structurale de BUSI (Williamson et al., 1985) (figure 30), toutefois sans marquage isotopique.

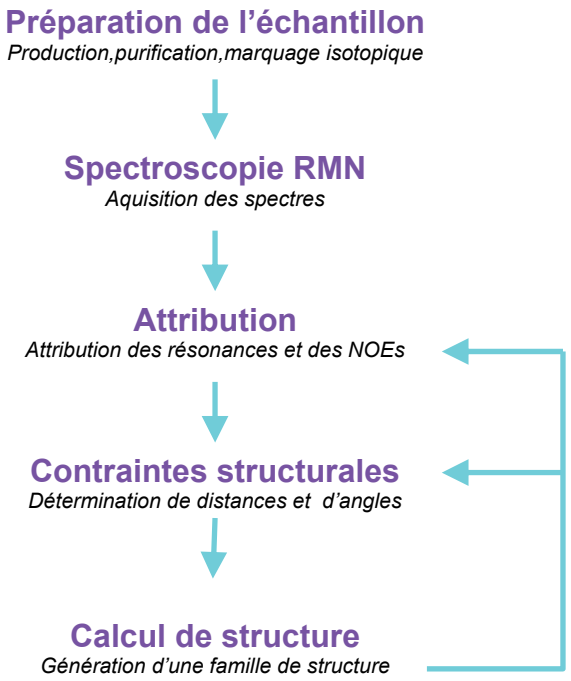


Figure 30 : **Protocole standard pour la détermination de structure de protéines par RMN**
Les étapes d'attribution et de détermination des contraintes structurales sont complétées après le calcul de structure de façon itérative jusqu'à une bonne qualité des structure générée

La structure de BUSI avait été résolue en utilisant la RMN du proton à deux dimensions ; depuis beaucoup d'améliorations ont été apportées avec notamment l'utilisation de la RMN héteronucléaire à plusieurs dimensions (Sattler et al., 1999) permettant d'étudier des protéines et des complexes de tailles supérieures (Foster et al., 2007) et avec également de nombreuses innovations dans les séquences d'impulsion. Ainsi, par exemple, la structure de la malate synthase de 81.4 Kda (figure 31) (Tugarinov et al., 2005) a été résolue à partir de données RMN (NOEs, couplages dipolaires résiduels, valeur de déplacement chimique) en adoptant une stratégie de marquage spécifique et en utilisant des expériences à quatre dimensions utilisant le principe TROSY (Transverse Relaxation-Optimized Spectroscopy) (Pervushin et al., 1997).

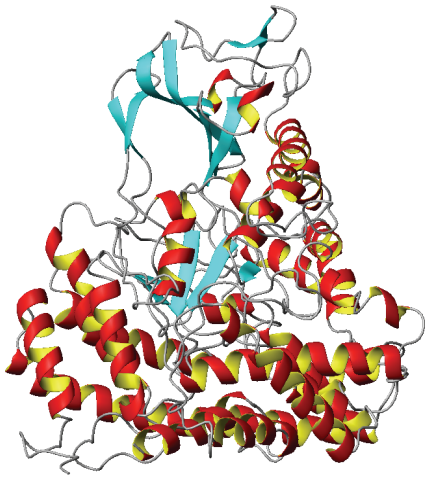


Figure 31 : **Structure en solution par RMN de la malate synthase de 724 résidus** (code PDB : 1Y8B) (Tugarinov et al., 2005)

De plus en plus de structures sont résolues par RMN, comme le montre le nombre croissant de structures RMN déposées sur la PDB (Protein Data Bank), traduisant l'essor de cette technique (figure 32)

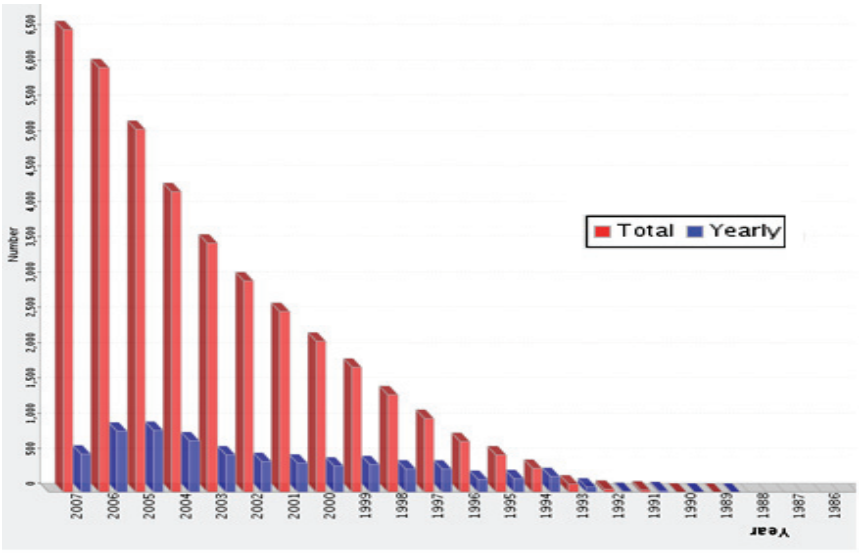


Figure 32 : **Structures déterminées par RMN déposées sur la PDB**
Adapté de http://www.pdb.org

La détermination de la structure d'une protéine est un processus complexe et passionnant car il fait intervenir des champs de compétences très différents : biologie moléculaire, biochimie, spectroscopie RMN, modélisation moléculaire.

Préparation de l'échantillon

La première étape de la détermination de structure est la préparation de l'échantillon. C'est une étape longue car les analyses par RMN nécessitent des grandes quantités de protéines à l'état pur. En effet l'échantillon doit avoir un volume de 300 µL et une concentration d'environ 1 mM pour des spectromètres dotés d'équipements classiques. Ce qui représente 3 mg de matériel pur pour une protéine de 10 kDa. Une étude structurale exige donc des capacités de production bien supérieures à celle requise pour une étude classique de biochimie.

Il y a deux modes généraux d'obtention de la protéine : par extraction ou purification à partir d'un organisme ou d'un organe ou elle est naturellement présente en grande quantité ou par purification à partir de la protéine clonée et surexprimée.

La mise au point du protocole de production et de purification peut durer de quelques semaines à plusieurs années. La technique de surexpression de protéines recombinantes permet le marquage isotopique. Le marquage aux isotopes stables (^{13}C, ^{15}N et ^{2}H) est à l'origine d'avancées considérables en RMN des protéines. Il permet de réaliser des expériences hétéronucléaires nécessaires pour l'attribution de protéines de taille supérieure à 10 kDa. Le marquage s'effectue classiquement en utilisant les techniques de clonage, de surexpression et de purification. La surexpression peut avoir lieu dans un milieu minimum qui ne contient qu'une seule source du noyau dans lequel on veut enrichir la protéine (par exemple : NH_4Cl ^{15}N, glucose ^{13}C) ou un milieu riche enrichi en isotope produit à partir d'une première culture. Ce marquage est nécessaire pour observer les noyaux azote et carbone dont les isotopes observables par RMN ; ^{15}N et ^{13}C respectivement n'ont pas une abondance naturelle suffisante. Pour l'étude des grosses protéines, supérieures à 20 KDa, il peut être utile de remplacer les protons par des deutériums de façon à simplifier les spectres de grosses protéines et diminuer les vitesses de relaxation transverses. En effet, quand on augmente la taille de la protéine, l'augmentation du couplage dipolaire entre protons conduit à une diminution de l'intensité du signal observé sur les spectres. Le deutérium possède un moment magnétique environ 6 fois plus faible que le proton et sa présence va donc diminuer la fuite de l'aimantation due aux interactions dipolaires entre protons ou entre protons et hétéroatomes. Pour ce marquage, on utilise un milieu de culture contenant du D_2O permettant d'effectuer la deutériation complète de la protéine. Le passage dans une solution d'H_2O permet

ensuite la reprotonation sélective des hydrogènes amides et autres protons échangeables.

Les conditions d'étude (tampon, force ionique, température, pH) doivent être déterminées de façon à avoir une protéine stable et correctement structurée. L'étude des protéines par RMN se fait généralement à pH acide ou neutre afin de ralentir l'échange des protons amides avec l'eau car leur observation est particulièrement importante pour l'analyse de la structure.

RMN des protéines en solution : stratégie d'attribution

La RMN des protéines utilise quatre types d'isotopes : ^{1}H, ^{2}H, ^{15}N, ^{13}C. Le proton est l'isotope le plus observé, pour les autres isotopes un marquage est nécessaire étant donné leur faible abondance naturelle.

Les spectres RMN de protéines sont enregistrés dans l'eau protonée contenant 5 à 10 % de D_2O pour le verrouillage champ-fréquence, l'utilisation d'eau deutérée entraînant la perte des signaux des protons amides échangeables. Une atténuation du signal de l'eau est donc indispensable dans les séquences d'impulsions pour pouvoir observer les signaux de la protéine, celui de l'eau étant bien plus intense. Une autre caractéristique des spectres RMN de protéines est leur extrême complexité due au grand nombre de noyaux observables qui induit un encombrement des spectres. L'observation de plusieurs noyaux qui résonnent à des fréquences différentes permet de lever ces problèmes de superpositions des déplacements chimiques.

L'analyse et l'attribution des spectres consistent à relier tous les pics croisés observés sur un spectre à une interaction entre deux noyaux définis de la protéine. Deux phénomènes d'interaction sont utilisés en RMN des protéines : le couplage scalaire qui est observé entre deux noyaux séparés par une ou plusieurs liaisons covalentes et le couplage dipolaire qui est observé entre deux noyaux proches dans l'espace et qui définit une valeur de l'effet Overhauser nucléaire (NOE) qui sera converti en une distance entre les deux noyaux utilisée pour le calcul de structure.

Les premières déterminations de structures de protéines ont été réalisées par RMN homonucléaire ^{1}H 2D qui ne concerne que la détermination de structures de protéines de moins de 10 kDa. Les deux expériences homonucléaires les plus utilisées sont la TOCSY (TOtal Correlated SpectroscopY) et la NOESY (Nuclear

Overhauser Effect SpectroscopY). Les différents types d'hydrogènes (amides, aliphatiques, aromatiques) résonnent à des fréquences spécifiques puisqu'ils ont des environnements électroniques différents; on les retrouve dans des régions distinctes comme on peut le voir sur un spectre 1D du domaine THAP de THAP1 (figure 33).

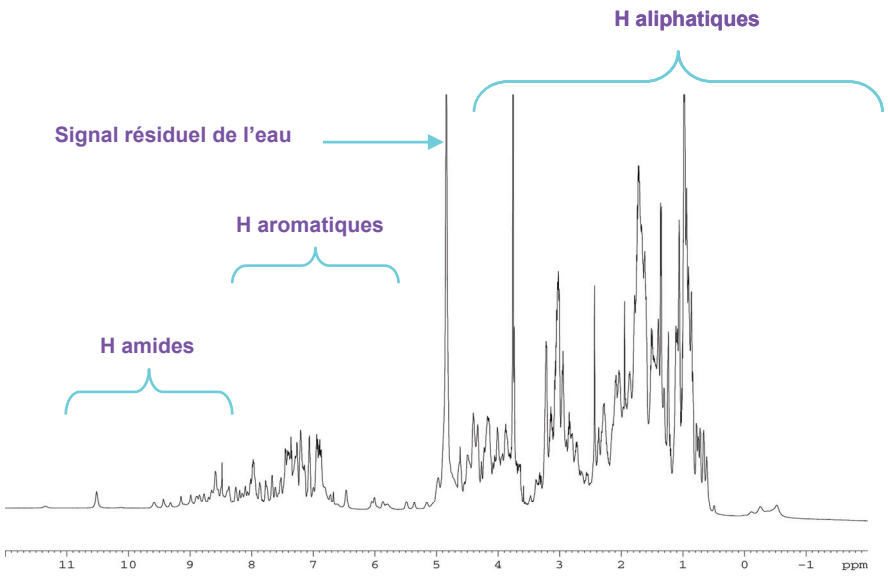


Figure 33 : **Spectre1D proton du domaine THAP de THAP1**

Sur un spectre 2D on distingue différentes régions qui vont nous aider lors de l'attribution (figure 34).

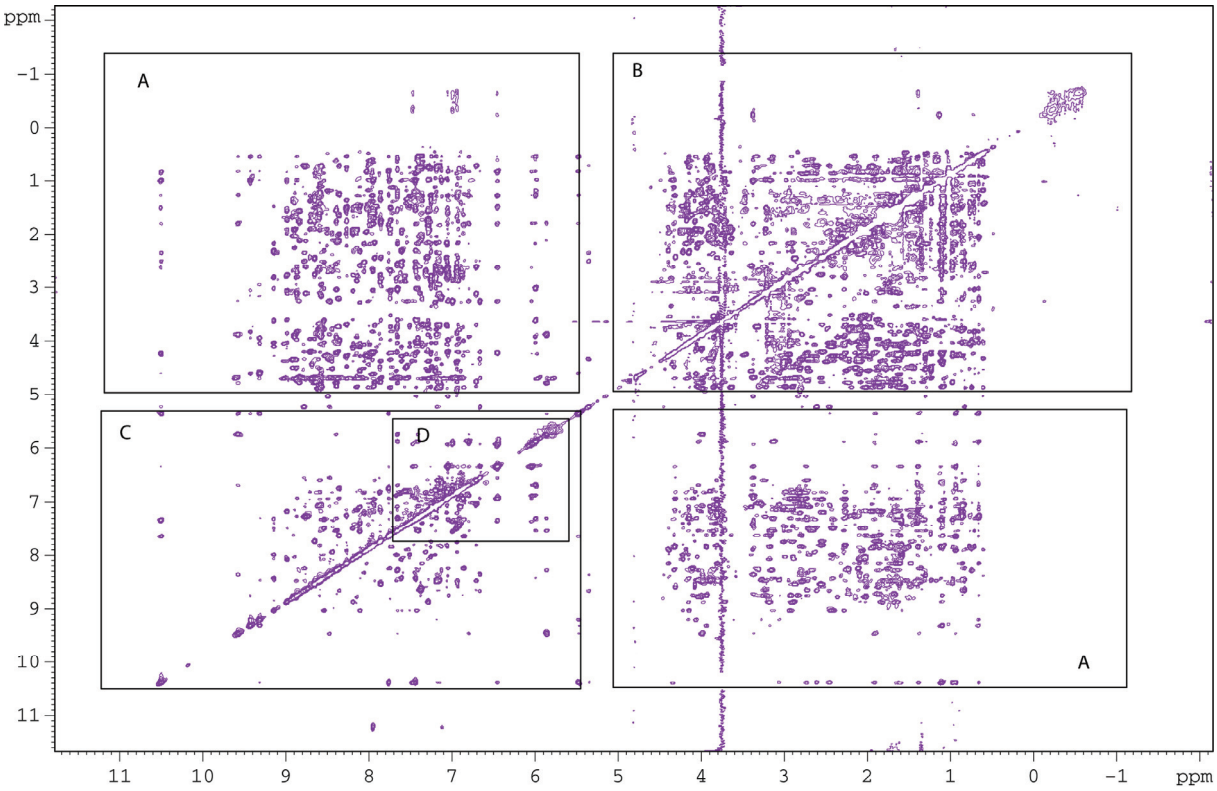
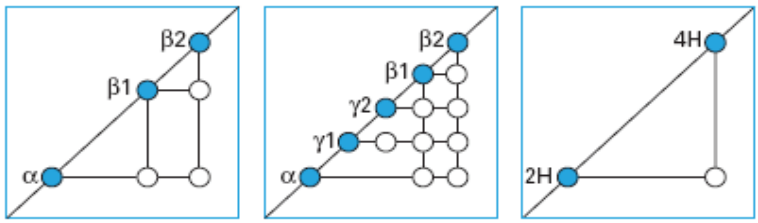


Figure 34 : **Spectre NOESY du domaine THAP de THAP1**
Pics de corrélation entre :

A. H amides – H aliphatiques
B. H aliphatiques
C. H amides
D. H aromatiques

L'expérience TOCSY permet d'observer des couplages scalaires entre hydrogènes et donc l'identification des acides aminés qui ont des systèmes de spins différenciables en fonctions de leurs chaînes latérales (figure 35).



Chaîne courte (AMX) (Ser, Cys, Asp, Asn, Phe,Tyr, His, trp)

Chaîne moyenne (Gln, Glu, met)

Chaîne aromatique (His par exemple)

Figure 35 : **Systèmes de spin des chaînes latérales**
Adapté de (Malliavin and Dardel, 2002)

L'expérience NOESY permet d'observer des NOEs (issus des couplages dipolaires à travers l'espace) entre deux protons à moins de 5 Á l'un de l'autre environ, selon le

rapport signal sur bruit. Sur un spectre NOESY, on observe des pics supplémentaires correspondant à des hydrogènes proches dans l'espace. En particulier les contacts entre hydrogènes amides de deux résidus consécutifs ou de l'hydrogène amide d'un résidu avec l'hydrogène α du résidu précédent vont permettre d'avoir des informations sur l'ordre des systèmes de spin dans la séquence. On réalise ainsi l'attribution séquentielle de la protéine comme l'a été proposé par Wüthrich et ses collaborateurs (Wüthrich, 1986).

La RMN homonucléaire présente une limitation; l'étude de protéines de taille supérieure à 100 résidus est difficile avec seulement les expériences protons pour deux raisons : plus le nombre d'hydrogènes est grand, plus le cas de superpositions de déplacements chimiques augmente et l'élargissement des raies avec l'augmentation de la taille des protéines détériore la qualité des expériences homonucléaires de type TOCSY, les vitesses de relaxation longitudinale accrues des protons entrant en compétition avec les mécanismes de transfert de cohérence. Pour des protéines plus grosses que 10 kDa, il devient indispensable d'avoir recours en supplément aux expériences hétéronucléaires ^{15}N et / ou ^{13}C.

Les expériences hétéronucléaires font intervenir des constantes de couplage plus intenses (figure 36) et de ce fait vont permettre un bon transfert d'aimantation qui leur confère une appréciable sensibilité (Sattler et al., 1999).

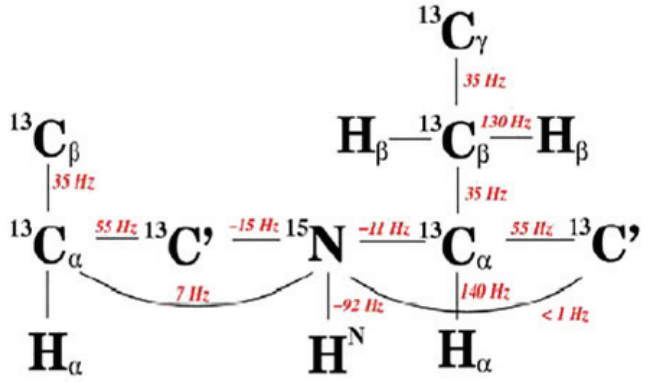


Figure 36 : **Couplage ^{2}J et ^{1}J dans les protéines**

On peut ainsi réaliser des expériences tridimensionnelles hétéronucléaires (et même de dimension supérieure).

Trois expériences principales utilisent la résonance ^{15}N. L'expérience 2D ^{15}N-^{1}H HSQC (Heteronuclear Single Quantum Correlation) permet d'observer les pics de corrélation dus au couplage scalaire hétéronucléaire entre l'hydrogène amide et l'azote de chaque résidu (figure 37).

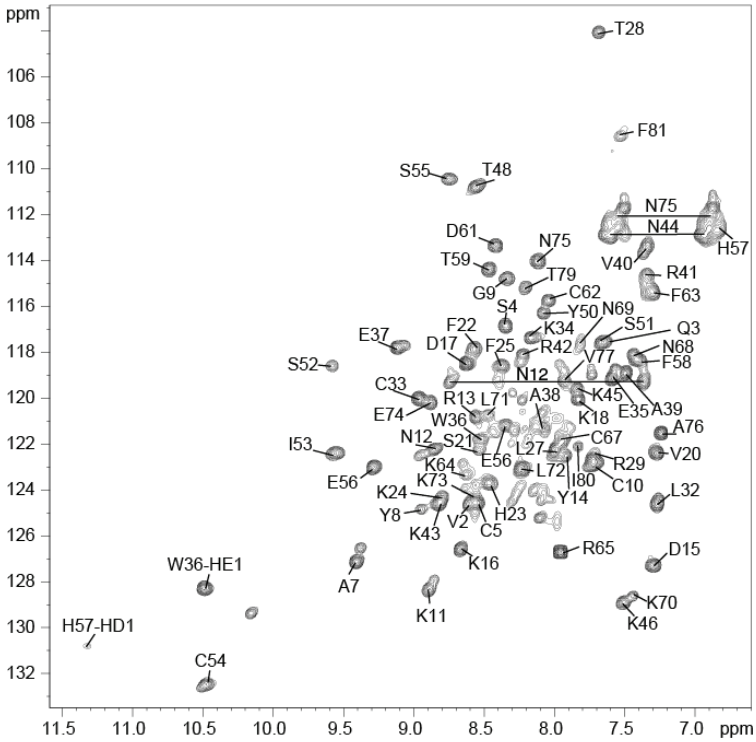


Figure 37 : **Spectre 2D ^{15}N HSQC du domaine THAP de THAP1**
Chaque pic correspond à une corrélation entre un proton amide et l'azote auquel il est directement lié. L'expérience utilise leur couplage hétéronucléaire de 90 Hz (figure 32). On retrouve un pic par acide aminé, en dehors des chaînes latérales ; cette attribution est indiquée sur le spectre

L'expérience 3D TOCSY-HSQC permet d'observer les systèmes de spin de chaque hydrogène amide tandis que l'expérience 3D NOESY-HSQC permet d'observer, pour chaque hydrogène amide, les effets NOEs avec les autres atomes d'hydrogènes de la molécule. Le principe d'attribution utilisant ces deux expériences est présenté sur la figure 38, il illustre l'emploi des expériences 3D en général.

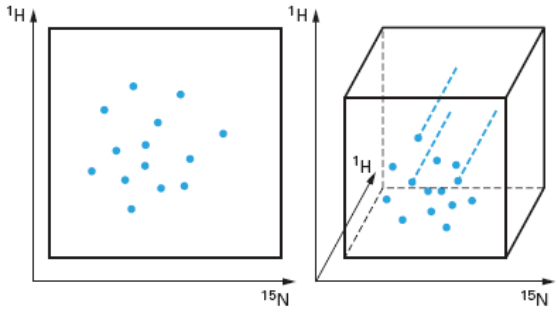


Figure 38 : **Principe de l'attribution utilisant les expériences 3D TOCSY – HSQC (systèmes de spin) et 3D NOESY – HSQC (attribution séquentielle)**
L'expérience 2D HSQC (à gauche) est utilisée pour étiqueter les résidus selon des déplacements chimiques ^{15}N et ^{1}H de leur hydrogène amide ; les expériences 3D HSQC (à droite) sont utilisées pour différencier les hydrogènes qui sont couplés scalairement à l'hydrogène amide de ceux qui lui sont couplés dipolairement. Adapté de (Malliavin and Dardel, 2002)

Les expériences hétéronucléaires utilisant les noyaux ^{15}N et ^{13}C sont nommées d'après les noms des atomes qu'elles permettent de relier. En guise d'exemple, nous avons répertorié les types d'expériences tridimensionnelles utilisées pour l'attribution du domaine THAP de THAP1 (Table 4).

Expérience	Noyaux observés
HNCO	HN(i) N(i) C'(i-1)
HNCA	HN(i) N(i) Ca(i) Ca(i-1)
HNCACB	HN(i) N(i) Ca(i) Ca(i-1) Cb(i) Cb(i-1)
HN(CO)CA	HN(i) N(i) Ca(i-1)
CBCA(CO)NH	HN(i) N(i) Ca(i-1) Cb(i-1)
HBHA(CO)NH	HN(i) N(i) Ha(i-1) Hb(i-1)
HNHA	HN(i) N(i) HA(i)
NOESY-1H 15N HSQC	H(i) -> H(j) N(j)
TOSCY-1H 15N HSQC	H(i) -> H(j) N(j)

Table 4 : **Expériences tridimensionnelles réalisées pour l'attribution du domaine THAP de THAP1**

Certaines de ces expériences permettent d'observer les couplages à l'intérieur d'un même résidu ; d'autres à travers le carbone du carbonyle, les couplages entre noyaux appartenant à deux résidus successifs. Un exemple est donné (figure 36) sur l'utilisation des expériences HNCACB (Wittekind and Mueller, 1993) et CBCACONH (Grzesiek and Bax, 1993) qui permettent de réaliser l'attribution séquentielle. Comme un plus grand nombre de déplacements chimiques est utilisé pour définir un système de spin donné, de nombreuses ambiguïtés dues aux superpositions de déplacements chimiques ^{1}H sont levées.

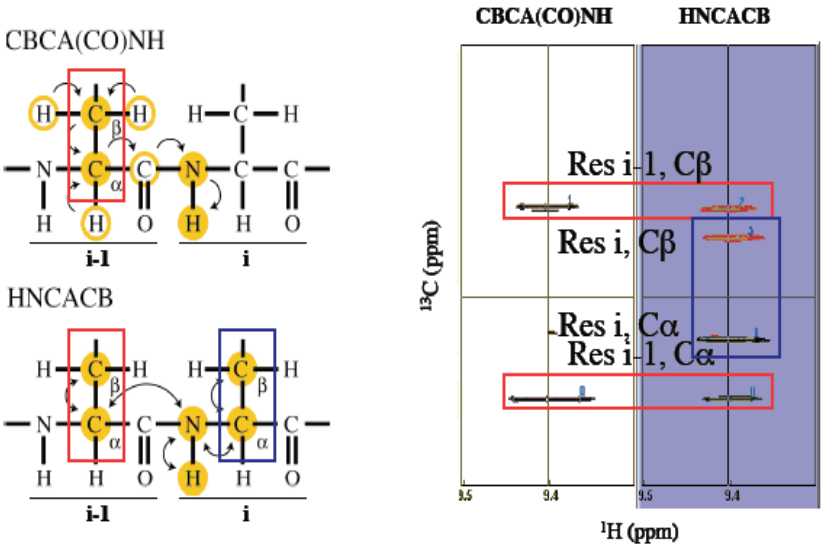


Figure 39 : **Utilisation des expériences CBCA(CO)NH et HNCACB**
Ces expériences 3D (N H C) permettent d'observer pour chaque couple proton amide – azote, les corrélations avec les carbones du résidu précédent pour l expérience CBCA(CO)NH , et les carbones du résidu précédent et propre pour l'expérience HNCACB.
Ainsi il est possible d'attribuer séquentiellement ces atomes en cherchant des correspondances entre les fréquences carbones

Les données fournies par la RMN : contraintes structurales

Les données provenant des expériences RMN que nous avons utilisées sont de deux types : les contraintes de distances provenant des mesures d'effets Overhauser et les contraintes d'angles dièdres provenant des mesures de constantes de couplage scalaire. Ces données sont utilisées comme contraintes pour le calcul de structure.

Contraintes de distances

Lorsque deux protons sont proches dans l'espace, typiquement distants de moins de 5 Á, la mesure de l'intensité de l'effet NOE (Nuclear Overhauser Effect) observé donne une estimation de leur distance (Noggle and Shirmer, 1971). L'intensité de l'effet NOE entre deux protons dépend de la dynamique interne et globale de la protéine et de la distance r entre ces deux protons. Sous couvert d'un certain nombre d'approximations, absence de diffusion de spin notamment, cette relation est inversement proportionnelle en $1/r^6$ x $f(\tau_c)$ où τ_c est le temps de corrélation

caractérisant les mouvements du vecteur internucléaire. Ce qui donne la relation de proportionnalité suivante (Wüthrich, 1986) :

$$NOE \propto \frac{1}{r^6}$$

Cette relation entre distances et intensités NOEs reste assez qualitative du fait de la sensibilité et la difficulté d'interprétation des effets NOE. En effet la superposition de certains pics rend la mesure des intensités NOEs souvent difficile, à cela s'ajoute le phénomène de diffusion de spin qui correspond à un transfert d'aimantation indirect entre deux hydrogènes et conduit à une intensité différente de celle donnée dans la relation ci-dessus (Linge et al., 2004). On préfère donc en pratique classer les intensités en trois groupes : faible, moyen et fort. Une distance maximale est associée à chaque groupe d'intensité, par exemple 2.2, 3.3 et 5 Å pour respectivement les groupes à effet fort, moyen et faible. La distance minimale est définie à 1.8 Å, qui correspond à la somme des rayons de van der Waals de deux atomes d'hydrogène. Cette distance minimale peut être prise dans chaque catégorie car la dynamique interne peut conduire à des intensités NOEs plus faibles pour une distance moyenne donnée.

Un obstacle à l'attribution des NOEs est la présence d'ambigüités dans le cas où deux hydrogènes résonnent à la même fréquence. Il faut alors décider lequel est impliqué dans l'interaction observée. Pour résoudre ces ambigüités, on réalise l'attribution des NOEs en même temps que le calcul de structure (figure 30), et lorsque celui-ci est avancé on réalise les modifications d'attribution en accord avec la structure calculée. Ce processus itératif est conduit jusqu'à une attribution du spectre quasi complète. Le logiciel ARIA (Ambiguous Restraints for Iterative Assignment) propose une automatisation de ce processus et la gestion des NOEs ambigus (Linge et al., 2001).

Structures secondaires

Une analyse qualitative des NOEs permet de déterminer la structure secondaire. Les hélice α sont en particulier caractérisées par la présence d'effets NH(i)-NH(i+1), HA(i)-NH(i+3) et les feuillets β par des effet NOEs inter-brins impliquant leurs protons HA et HN. L'ensemble de ces contacts particuliers des éléments de structure

secondaire sont répertoriés dans l'ouvrage de Wüthrich (Wüthrich, 1986). Nous les avons détaillés dans notre cas dans la partie suivante (cf p 96)

L'identification des structures secondaires peut également être prédite à partir des déplacements chimiques (Wishart et al., 1991). En effet les structures en hélice α et feuillet β induisent des déplacements chimiques caractéristiques par rapport aux déplacements chimiques moyens d'un résidu donné. Dans ce cadre, on étudie les différences de déplacement chimique d'un acide aminé avec les déplacements chimiques du même aminé en pelote statique, sans structure particulière ou encore *random coil*, pour déterminer les régions en feuillet ou en hélice (Wishart et al., 1992).

Afin d'automatiser et de rendre simple cette identification des structures secondaires, Wishart a développé un algorithme, appelé *Chemical Shift Index* (CSI) ou index de déplacement chimique. Le principe est d'associer à chaque résidu un nombre (-1, 0, +1) en fonction de l'écart de déplacement chimique par rapport au déplacement chimique du même résidu en random coil :

- 0 si le déplacement chimique observé est compris dans un intervalle autour du random coil,
- +1 si le déplacement chimique observé est plus grand que la borne supérieure de l'intervalle,
- -1 si le déplacement chimique observé est plus petit que la borne inférieure de l'intervalle.

Par exemple pour les protons Hα, le déplacement chimique se trouve en moyenne décalé vers les hauts champs dans une hélice α, (-0,39 ppm par rapport au *random coil*), dans un brin β, il est décalé en moyenne vers les bas champs (+0,37 ppm par rapport au *random coil*), l'intervalle autour du *random coil* est fixée à ±0, 1ppm et ainsi un index + 1 indique une structuration en hélice α et -1 en feuillet β.

Ce type de prévisions est utile pour réaliser l'attribution des NOEs et contraindre ces régions lors du calcul.

Cette prévision des structures secondaires est corroborée par les données cinétiques d'échange des hydrogènes amides mesurées dans une protéine en solution dans le D_2O (Wuthrich and Wagner, 1979). Les protons des régions structurées vont s'échanger plus lentement en deutérium et rester visibles plus longtemps sur une expérience RMN ^1H (figure 40).

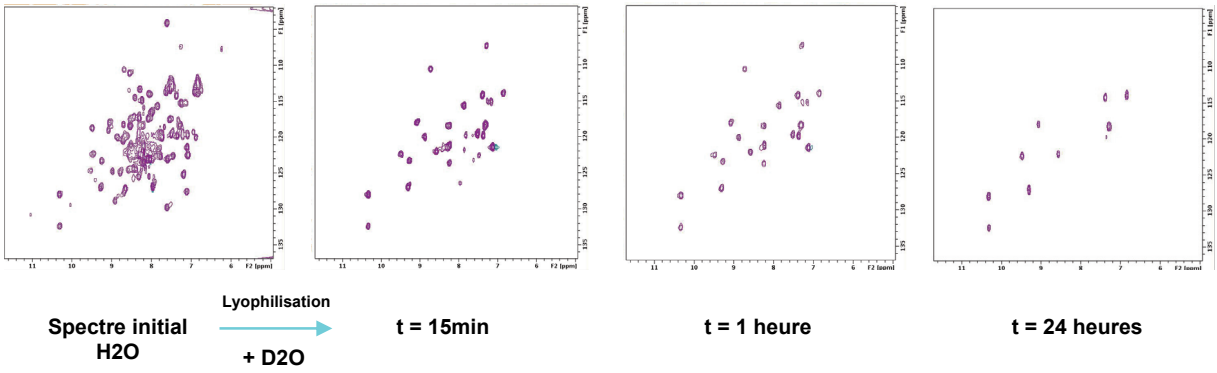


Figure 40 : **Effet sur le spectre HSQC du domaine THAP de THAP1 de l'échange des hydrogènes amides en deutérium**
Après lyophilisation, la protéine est reprise dans le D_2O et des spectres HSQC sont enregistrés en fonction du temps, de façon à déterminer la cinétique d'échange des protons amides en deutérium.

On identifie ainsi les protons protégés, généralement impliqués dans une liaison hydrogène et potentiellement caractéristique d'une structure secondaire.

Les angles dièdres

Les angles dièdres sont déterminés à partir des constantes de couplage scalaire en utilisant les relations de Karplus (Karplus, 1959). Par exemple le couplage $^3J_{H\alpha, HN}$ permet d'accéder à la valeur de l'angle phi dans le cas d'une protéine (Pardi et al., 1984) :

$$^3J_{H\alpha, HN}(\phi) = 6{,}4\cos^2(\phi - 60°) - 1{,}4\cos(\phi - 60°) + 1{,}9$$

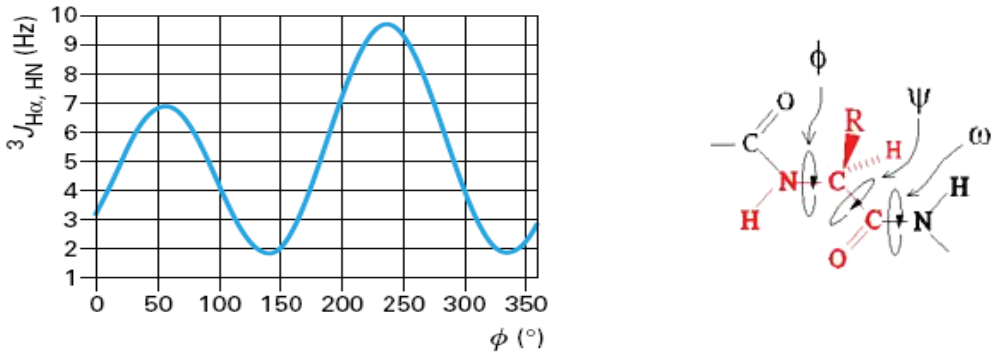


Figure 41 : **Relation de Karplus et définition des angles dièdres du squelette peptidique**

La constante de couplage $^3J_{H\alpha,HN}$ est la plus fréquemment utilisée, elle est déterminée à partir de l'expérience 3D HNHA (Vuister et al., 1993) à partir du rapport d'intensité du pic de corrélation HN/HA et du pic diagonal HN/HN pour chaque résidu d'après la relation : $\frac{I_{HNHA}}{I_{HNHN}} = -\tan^2(2\pi J_{H\alpha HN}\zeta)$

où ζ est un délai de valeur 13.05 ms, *J* est mesuré en Hertz.

Une autre approche pour la détermination des angles dièdres consiste en l'utilisation de la relation entre le déplacement chimique et la structure secondaire et les constantes de couplage. Le programme TALOS (Torsion Angle Likelihood Obtained from Shift and sequence similarity) permet ainsi d'évaluer les angles ψ et ϕ à partir de la séquence de la protéine et des déplacements chimiques : Hα, Cα, Cβ, CO et N comparée à une base de déplacements chimiques construite à partir des données sur 20 protéines (Cornilescu et al., 1999).

Le principe est que si une succession d'acides aminés montre une similarité de déplacements chimiques avec une séquence d'une protéine de structure connue, alors les résidus centraux des deux chaînes ont probablement les mêmes angles de torsion ψ et ϕ. Lors de la prédiction des angles ψ et ϕ d'un résidu, les informations du résidu précédent et du résidu suivant sont prises en compte. En pratique TALOS utilise les données pour trois résidus consécutifs pour faire la prédiction sur le résidu central du triplet. Pour chaque triplet, TALOS recherche dans la base de données les 10 meilleures valeurs d'angles ψ et ϕ pour lesquels la similarité de déplacement chimiques est la meilleure. L'ensemble des solutions et leur moyenne sont visualisées sur un diagramme de Ramachandran (figure 44). Si ces 10 solutions sont acceptables, c'est à dire qu'elles sont dans une même zone et dans une région permise alors leur moyenne et leur déviation standard sont utilisées pour la prédiction de la valeur des angles ψ et ϕ du résidu. Dans le cas contraire, il n'y a pas de prédiction effectuée. Ce critère de sélection est ajusté manuellement à l'aide d'une interface présentant le diagramme de Ramachandran.

Attribution stéréospécifique

L'attribution stéréospécifique des résonances, par exemple des protons β ou des groupes méthyles des valines et des leucines, peut être importante pour définir des

conformations locales des résidus. Cette attribution peut s'effectuer en combinant les NOEs avec les constantes de couplage mesurées (figure 42) (Guntert et al., 1989; Nilges et al., 1990). Lorsqu'il n'est pas possible de résoudre les attributions stéréospécifiques, on utilise alors des pseudo-atomes pour remplacer chaque groupe, avec des intervalles de contraintes augmentés.

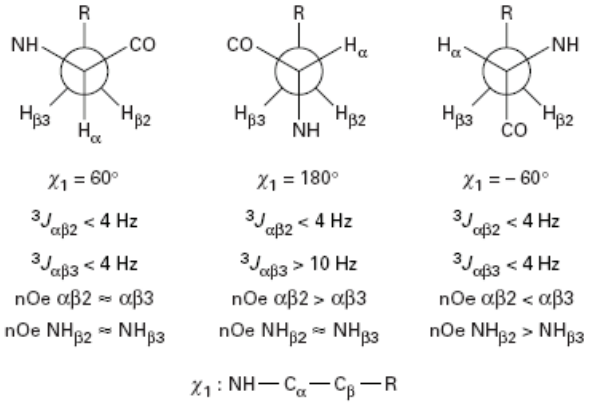


Figure 42 **: Variation des constantes de couplage et des NOEs entre Hα et Hβ et entre H et Hα en fonction de la valeur de l'angle dièdre χ_1**

Génération des structures

Calcul de structure

L'étape finale de la détermination de structure est le calcul des coordonnées des atomes de la protéine à partir des contraintes de distance et d'angle déterminées par RMN. Il existe de nombreux programmes informatiques permettant le calcul des structures RMN, et grâce à l'amélioration constante de la puissance des ordinateurs, le calcul proprement dit des structures n'est plus une étape limitante.

Le calcul de structure d'une protéine par RMN est la détermination d'un ensemble de conformères, compatibles avec les données expérimentales, qui peuvent aussi le cas échéant rendre compte d'une dynamique interne de la protéine en solution. La famille de structures générées tient compte des contraintes expérimentales en respectant la structure covalente des acides aminés. Cela permet de réduire le nombre de degrés de liberté du système et de rendre le problème soluble.

Le calcul des structures RMN se fait dans le cadre du formalisme de la mécanique moléculaire. Les potentiels d'interactions entre atomes sont décrits par des fonctions empiriques simples :

$$V = \sum_{\text{liaisons}} \frac{1}{2} K_\ell (\ell - \ell_0)^2 + \sum_{\text{angles}} \frac{1}{2} K_\theta (\theta - \theta_0)^2$$
$$+ \sum_{\text{dièdres}} K_\phi [1 + \cos(n\phi - \delta)] + \sum_{i,j} \left(\frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^6} \right) + \sum_{i,j} \frac{q_i q_j}{4 \pi \varepsilon_0 \varepsilon_r r_{ij}}$$

Les deux premiers termes (potentiels de liaisons et d'angles covalents) sont harmoniques. Un terme sinusoïdal représente les potentiels d'angles dièdres. Des potentiels de Lennard-Jones (quatrième terme) sont utilisés pour les interactions de van der Waals entre atomes non liés, des potentiels de type coulombien (cinquième terme) pour les interactions électrostatiques entre atomes non liés. La forme du potentiel et les valeurs des paramètres le caractérisant forment le champ de force de la méthode utilisée. Nous avons utilisé le champ de force de CNS (Crystallography & NMR system) (Brunger et al., 1998).

Une caractéristique des calculs empiriques effectués pour la détermination de structure par RMN est donc l'utilisation d'un terme harmonique supplémentaire dans l'énergie potentielle, qui permet d'appliquer à la structure les contraintes provenant des mesures RMN. Les allures des potentiels entre atomes non liés et du potentiel des contraintes dues aux mesures de NOE sont représentés figure 43.

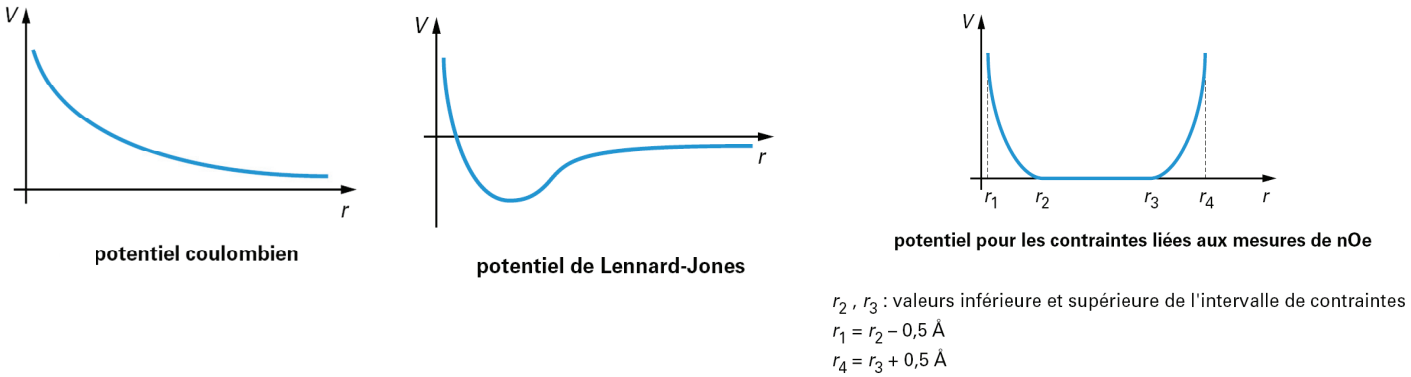


Figure 43 : **Allures des potentiels**

Le calcul se fait classiquement en deux étapes. Une première méthode, utilisant la théorie mathématique de géométrie des distances (Crippen and Havel, 1988),

permet de générer un ensemble de structures qu'il convient d'optimiser par une étape supplémentaire utilisant la dynamique moléculaire sous champs de force. Le but de l'optimisation est d'obtenir un minimum global de la surface d'énergie potentielle du système, comprenant l'énergie interne plus l'énergie des contraintes expérimentales. Pour sortir des minima locaux et franchir les barrières de potentiel qui les entourent, on introduit de l'énergie cinétique au système en faisant varier la température, c'est le principe du recuit simulé (figure 44) (Nilges et al., 1988).

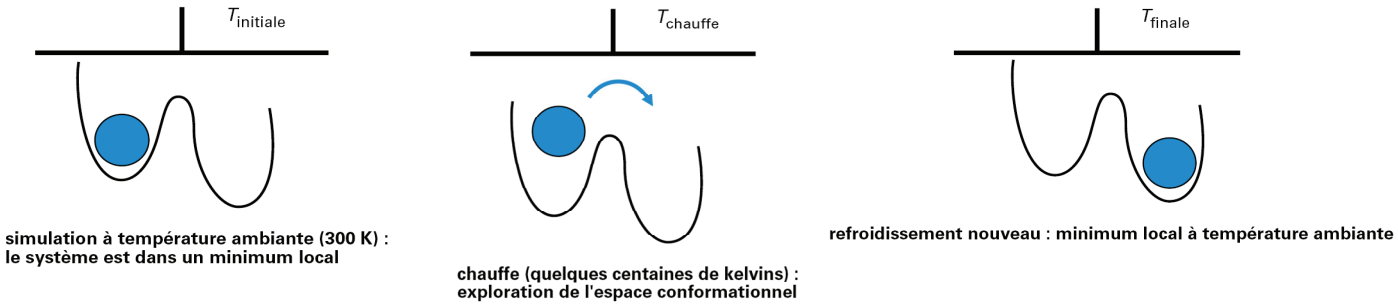


Figure 44 : **Méthode du recuit simulé**
L'introduction d'énergie cinétique permet de franchir des barrières de potentiel et d'accéder au minimum global

Qualité de la structure obtenue

La qualité de la structure obtenue est obtenue en observant, pour l'ensemble des conformères de plus basse énergie : le nombre de contraintes de distances, la valeur de l'écart quadratique moyen des coordonnées (rmsd), le nombre de contraintes non respectées, les valeurs des angles phi et psi sur un diagramme de Ramachandran (Ramachandran et al., 1963) (figure 45).

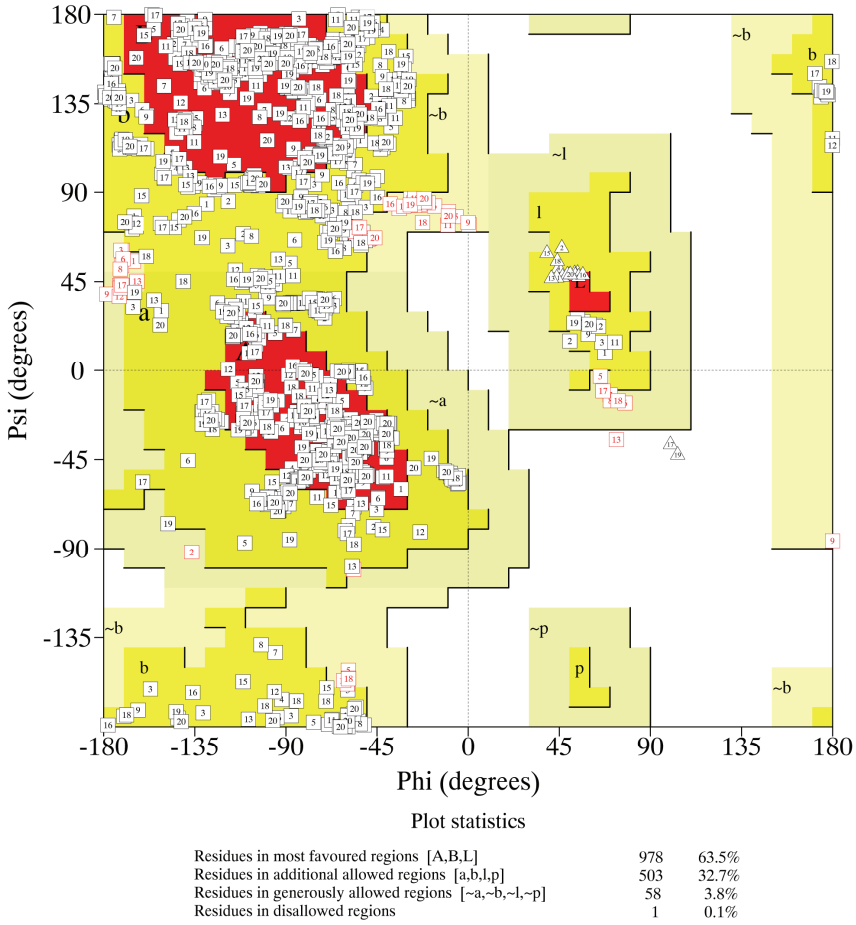


Figure 45 : **Diagramme de Ramachandran du domaine THAP de THAP1**
L'ensemble des valeurs (phi,psi) est reporté sur un diagramme de Ramachandran pour les 20 structures de plus basse énergie générées lors du calcul du domaine THAP. On comptabilise ainsi les valeurs (phi,psi) dans les zones interdites, permises et favorables. Réalisé avec le programme Procheck (Laskowski et al., 1996) .

Nous avons utilisé cette stratégie pour déterminer la structure du domaine THAP de THAP1. Pour finir d'illustrer cette technique, nous pouvons retrouver dans l'article l'ensemble des 20 structures de plus basses énergie générées par le calcul sous contraintes RMN, la structure représentative de plus basse énergie ainsi qu'une table résumant l'ensemble des statistiques structurales.

La procédure expérimentale que nous avons développée est décrite dans le chapitre suivant, *Production des échantillons et détermination de la structure du domaine THAP de THAP1*. La structure ainsi que l'étude de la liaison à l'ADN du domaine THAP de THAP1 est présentée dans le chapitre, *Etude de la structure et de la liaison à l'ADN du domaine THAP de THAP1*, dans lequel nous avons inséré l'article publié sur ce sujet, *Structure-Function analysis of the THAP-zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways*, ainsi que des résultats complémentaires et discussions.

PRODUCTION DES ECHANTILLONS ET DETERMINATION DE LA STRUCTURE DU DOMAINE THAP DE THAP1

Production et purification du domaine THAP de THAP1

Nous détaillons ici la production et la purification du domaine THAP de THAP1 contenant une étiquette de 6 histidines clivable. La production de la protéine recombinante est réalisée avec le système T7 RNA polymérase (Studier and Moffatt, 1986) dans des bactéries d'*E.Coli.* La protéine est purifiée du surnageant protéique par deux colonnes, une colonne d'affinité au nickel et une colonne d'exclusion sur gel, l'étiquette de 6 histidines est clivée par la thrombine.

Construction

Pour l'étude structurale du domaine THAP de THAP1, nous avons été amenés à produire des domaines de différentes tailles, d'abord le domaine 1-90 puis le domaine 1-81 dont le choix est détaillé plus en détail ci-après. Le protocole de production et purification a été conçu et optimisé sur le domaine 1-90 et ensuite appliqué sans changements pour le domaine 1-81 et qui permet d'obtenir les mêmes rendements de production et degrés de pureté.

Nous avons conçu ce protocole de production et de purification avec les impératifs qu'exigent la biologie structurale et l'étude par RMN : production en grande quantité, en milieu synthétique pour le marquage aux isotopes, d'une protéine de grande pureté.

Pour cela nous avons utilisé successivement trois constructions, une construction du domaine sauvage, une construction avec une étiquette de 6 histidines en position C-terminale et une construction avec une étiquette de 6 histidines clivable à la thrombine.

Nous avons finalement opté pour cette dernière construction qui donne des taux d'expression en milieu minimum supérieurs à la construction sauvage, ce qui n'est pas le cas en milieu complet Luria Broth (Bertani, 1951) comme nous pouvons le voir sur gel *SDS PAGE* (sodium dodecyl sulfate polyacrylamide gel electrophoresis) (Laemmli, 1970) (figure 46). Cet exemple illustre la difficulté du passage d'une production en milieu complet à une production en milieu minimum où les bactéries sont dans des conditions particulièrement stringentes.

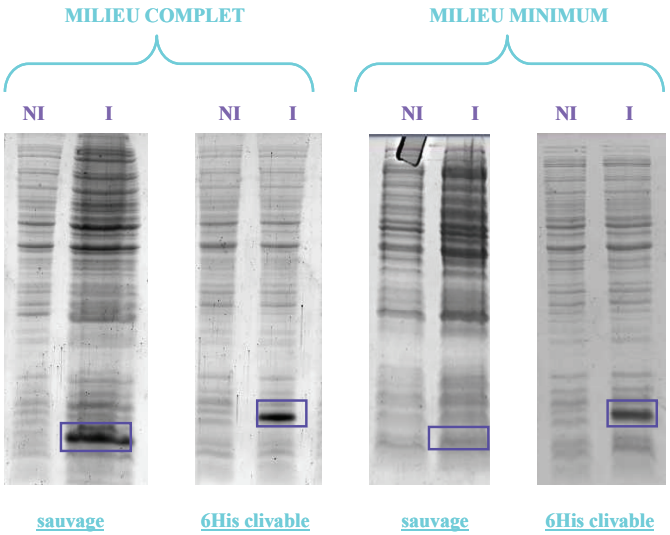


Figure 46 : **Production des constructions 6His et sauvage en milieu complet et minimum**
Gel sds PAGE 16%, des bactéries non-induites (NI) et induite (I) transfectées avec les plasmides des constructions 6His clivable et sauvage. La bande correspondant au domaine THAP surexprimé est encadrée. La construction 6His donne un meilleur taux de surexpression en milieu minimum.

De plus, l'étiquette histidine facilite et améliore l'étape de purification avec l'emploi d'une chromatographie d'affinité au nickel. D'autre part, le clivage de l'étiquette histidine facilite l'étape d'attribution et nous affranchit de toute interaction néfaste entre les histidines et le zinc présent dans la structure de la protéine. Enfin la position C-terminale de l'étiquette nous permet de ne purifier que des protéines entièrement traduites, ce qui ne serait pas le cas avec une étiquette en position N-terminale.

La séquence utilisée est la suivante pour le domaine 1-81 qui est le fragment dont nous avons déterminé la structure :

MVQSCSAYGC KNRYDKDKPV SFHKFPLTRP SLCKEWEAAV RRKNFKPTKY SSICSEHFTP DCFKRECNNK LLKENAVPTI FLELVPR|EGSV HHHHHH

Elle correspond aux 81 acides aminés du domaine THAP de THAP1 (GenBank #NP_060575), jusqu'au domaine AVPTIF, suivis du site de coupure à la thrombine LELVPR et de l'étiquette de 6 histidines qui est clivable. Après coupure, on obtient une protéine de 87 acides aminés de poids moléculaire 10202.8 Da.

Le plasmide utilisé est le suivant : pET-26BII, il a été construit au sein du laboratoire par modification du plasmide pET-26 (Novagen) de façon à obtenir une étiquette 6 histidines clivable en position C-terminale.

Production

La protéine est surexprimée dans des bactéries d'*E.Coli* BL21(DE3) par induction à l'IPTG (IsoPropyl β-D-1-ThioGalactopyranoside) avec le système d'expression utilisant la T7 RNA polymérase (Studier and Moffatt, 1986).
Il est utile de faire des cultures à partir de bactéries fraîchement transformées pour un taux de production maximal. Nous utilisons des bactéries BL21(DE3) chimio-compétentes. Les bactéries transformées avec le plasmide Pet26 présentent une résistance à la kanamycine, on utilise donc cet antibiotique pour les sélectionner. Le protocole de transformation est le suivant :

Protocole 1 : transformation

Etapes à réaliser stérilement sous flamme :

- décongeler la solution de plasmide et de bactéries dans la glace
- ajouter 5µL de plasmide Pet26bII THAP6H clivable à 100µL de bactéries BL21 (DE3) chimiocompétentes
- Laisser dans la glace 30 min à 4°C
- Effectuer un choc thermique : 1 min 30 sec à 42°C
- Laisser dans la glace 5 min à 4 °C
- Ajouter 900 µL milieu Luria Broth (sans antibiotique) et laisser pousser les bactéries 60 min à 37°C
- étaler sur boite Luria Broth agar avec de la kanamycine à 50 mγ. Laisser pousser une nuit à 37°C

Le milieu de culture utilisé est du milieu M9 à 2g/L en glucose, chlorure d'ammonium et poudre Celtone (milieu riche marqué, Spectra) qui sont les seules sources en carbone et azote et qui seront donc utilisés pour le marquage isotopique ^{15}N et ^{13}C. L'ajout de Celtone s'est avéré primordial pour améliorer le taux de surexpression. De même nous ajoutons des vitamines et des oligoéléments pour atteindre des meilleurs rendements de production.

Protocole 2 : Milieu de culture

Solutions mères :

- M9 5 X: Na_2HPO_4 64 g , KH_2PO_4 15g, NaCl 2.5g, Qsp 1L H_2O MQ à stériliser.
- Oligoéléments : pour 200 mL dissoudre 1 a 1:
  1 g EDTA, 1.2 g $CaCl_2$, $2H_2O$, 88 mg CuSO4, $5H_2O$, 240 mg MnCl2, $4H_2O$, 4 mg H_3BO_3, 140 mg $ZnSO_4$,$7H_2O$, en dernier: 1.2g $FeSO_4$, $7H_2O$ + 40 mg acide ascorbique (antioxydant). Passer sur filtre 0.22μM sous flamme et conserver à l'abri de la lumière au froid.
- Vitamines :

vitamine	**stock(mg/mL)**	**volume (µL) pour 10 mL**
Thiamine	500	500
Riboflavine	100	100
Niacinamide	100	100
Pyridoxal	10	100
Cobalamine	10	10
D-Biotine	100	100
Ac panthothénique	100	100
Ac folique	10	100

- Pour 500ml de culture (une fiole de 2L) mélanger :

  400 mL H_2O sterile
  100 mL M9 5X
  1 mL $MgCl_2$ 1M
  50 µL $CaCl_2$ 1M
  5 mL EDTA 50g/L (0.5g/L final)
  5 mL glucose 200g/L (2g/L final)
  2 mL de (NH_4) Cl 500g/L (2g/L final)
  5mL d'oligoéléments
  500 µL vitamines
  1 g celtone (2g/L final)

On réalise une préculture sur la nuit à 37°C à partir d'un clone avec le milieu de culture en présence de kanamycine (50 mg/mL). La culture est réalisée à 37°C à partir de 20 mL de préculture pour 500 mL de milieu, soit une densité optique (DO) à une longueur d'onde de 600 nm de 0,1. Les bactéries sont cultivées en Erlenmeyer à ailettes pour favoriser leur oxygénation; dans le même but, on choisira typiquement un Erlenmeyer de 2 L pour 500mL de culture et remué sur table agitante à 180 rpm. Les bactéries sont induites à une DO de 0,8 avec de l'IPTG, de plus on rajoute à ce moment du zinc pour assurer la présence de cet ion dans le milieu de culture au moment de la synthèse de la protéine à doigt de zinc.

Protocole 3 : induction

Solutions mères :

- IPTG 1M
- $ZnCl_2$ 0,1M

- Lorsque la DO à 600nm atteint une valeur de 0,8 ajouter au 500 mL de culture 500 µL de solution d'IPTG et de $ZnCl_2$. Laisser à 37°C, sous agitation à 180 rpm, pendant 4heures.

Des tests d'expression réalisés à différentes températures ont montré un meilleur rendement à 37°C qu'à 20°C. Quatre heures d'induction sont suffisantes pour obtenir une quantité maximale de protéine d'intérêt comparée aux autres protéines produite chez *E.Coli*. La protéine est produite sous forme soluble, présente dans le surnageant protéique après lyse des bactéries et centrifugation. La détection de la protéine d'intérêt se fait par électrophorèse sur gel polyacrylamide en présence de dodécylsulfate de sodium (SDS PAGE) (Laemmli, 1970) (figure 49). Nous avons utilisé des gels à 16% d'acrylamide avec un ratio acrylamide/bisacrylamide de 29/1. Ce gel est relativement concentré et réticulé (classiquement on utilise un ratio de 37.5/1) de façon à pouvoir observer notre protéine d'intérêt de ~10 KDa.

Protocole 4 : lyse des bactéries

- Centrifugation des cultures à 6 000 rpm pendant 15 min
- Les culots sont repris dans (100 mL/ 2L de culture):
  Tris 50 mM pH = 7,9
  NaCl 300 mM
  β-Mercaptoethanol 2,8 mM
- Choc thermique : les culots sont plongés dans l'azote liquide (les culots peuvent alors être conservés à -80°C)
- Ajouter pour 100 mL :
  0,1% triton X 100 (1 mL 10%) détergent
  0,5 mM Benzamidine (100 µL à 0.5 M) antiprotéase
  0,5 mM PMSF (100 µL à 0.5 M) antiprotéase
  0,1 mg/mL lysozyme (10 mg)
- passages au vortex puis laisser 30 minutes sous agitation a 4°C
- sonication : 6 cycles de 30 secondes à 50% actif
- centrifugation 15 000 rpm pendant 60 min

Purification

Le surnageant est récupéré, la protéine est purifiée par une première colonne d'affinité au nickel, éluée par un gradient d'imidazole. Le chromatogramme obtenu, absorption à 280 nm en fonction du volume d'élution, est représenté figure 47.

Protocole 5 : colonne d'affinité au nickel

- Nous utilisons une phase sépharose (*Chelatine Sépharose fast flow* d'Amersham Bioscience) sur laquelle nous avons chélaté des ions nickel sur un système ÄKTA FPLC (Amersham Bioscience) à 4°C.
- Nous effectuons une première étape de lavage à 20 mM imidazole (solution A) sur 100 mL puis un gradient d'élution de 20 à 300 mM Imidazole sur 200 mL (solution B de 0 à 100%) à un débit de 1 ml/min.
- Solutions utilisées

A :
Tris 50 mM pH = 7,9
NaCl 300 mM
β-Mercaptoethanol 2,8 mM
Imidazol 20 mM

B :
Tris 50 mM pH = 7,9
NaCl 300 mM
β-Mercaptoethanol 2,8mM
Imidazol 300 mM

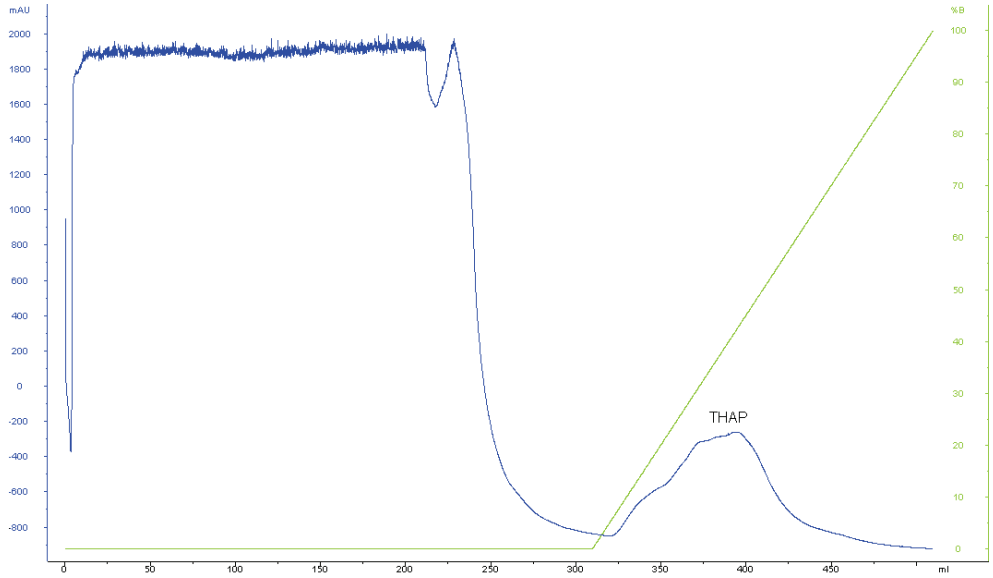


Figure 47 : **Chromatogramme : purification d'affinité au nickel**
L'absorption à 280 nm est représentée en bleue en fonction du volume élué. La proportion de solution B par rapport à la solution A est représentée en vert. La protéine est éluée au cours du gradient d'imidazole.

Les fractions correspondant à la protéine d'intérêt et vérifiées sur gel (figure 49) sont concentrées à 2 ml pour être ensuite purifiées sur une colonne d'exclusion sur gel. Toutes les étapes de concentration sont réalisées par ultracentrifugation sur membrane avec un seuil de coupure de 5000 Da (Vivaspin HY). Le chromatogramme obtenu est représenté figure 48.

Protocole 6 : colonne d'exclusion sur gel

- Nous utilisons une phase composée de dextran et d'agarose (superdex 75, Amersham Bioscience) sur un système ÄKTA FPLC (Amersham Bioscience) à 4°C.
- La protéine est chargée avec une boucle de 2 mL. L'élution se fait avec du tampon à 0.6mL/min :

NaCl 150 mM
Tris 50 mM pH = 7,4
DTT 1 mM

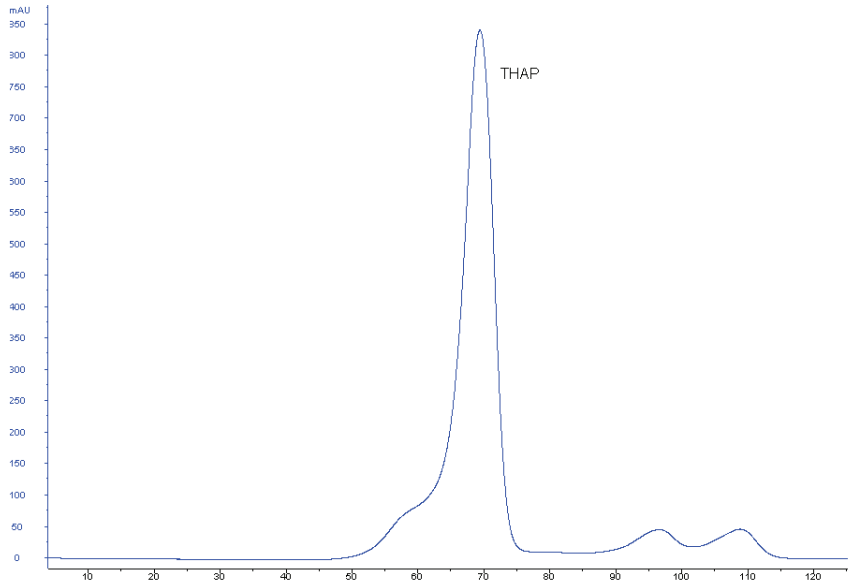


Figure 48 : **Chromatogramme : purification d'exclusion sur gel**
L'absorption à 280 nm est représentée en bleue en fonction du volume élué. La protéine est éluée vers 70 mL.

La protéine est concentrée à 2 mL dosée par absorption à 280 nm (avec un coefficient d'absorption théorique de 9970 $M^{-1}.cm^{-1}$ calculé sur EXPASY) puis subit une coupure à la thrombine afin d'éliminer l'étiquette de 6 histidines (figure 46). On réalise ensuite une dernière colonne d'exclusion sur gel pour purifier à nouveau la protéine. Cette dernière purification s'est avérée nécessaire pour l'élimination d'une proportion de protéine dégradée lors de la coupure à la thrombine, comme nous avons pu le voir par enregistrement de spectres ^{15}N HSQC avant et après cette étape. L'étiquette histidine est quant à elle déjà éliminée lors des concentrations sur membranes.

Protocole 7 : coupure à la thrombine

- 4 unités de thrombine (Novagen) par mg de protéine
- 2 H à 20°C, une nuit à 4°C sous agitation

L'échantillon est finalement préparé pour l'enregistrement des expériences RMN.

Protocole 8 : préparation de l'échantillon RMN

- Lavages dans le tampon RMN : (dilutions et concentrations successives pour que la concentration en tris non deutéré passe de 50mM à 0,01 mM)
  Tris deutéré 50 mM pH = 6,7
  NaCl 10 mM
  DTT deutéré 1 mM
  0,01% NaN3
- La protéine est concentrée dans le tampon RMN jusqu'à un volume final de 300 µL et transférée dans un tube RMN Shigemi avant l'adition de 10% de D_2O (30 µL) et passage sous argon

Le rendement obtenu est de 1 à 6 mg de protéine pure par litre de culture. Ce rendement est assez faible malgré nos nombreux efforts d'optimisation et rend compte de la réelle difficulté à produire cette protéine. Nous avons remarqué de plus que le rendement de production n'était pas toujours reproductible. Cela peut être expliqué par un certain degré de toxicité de cette protéine endogène pour la bactérie. La fonctionnalité et l'état de la protéine peuvent être vérifiés par électrophorèse sur gel (figure 49), enregistrement de spectre ^{15}N HSQC ou 1D ^{1}H et encore gel retard. La présence d'ion zinc coordonné à la protéine est d'ailleurs directement liée à la bonne fonctionnalité de la protéine produite, vérifiée par gel retard avec l'ADN cible. En effet, des expériences de gel retard montrent que le domaine THAP ne lie plus l'ADN lorsque l'on ajoute des chélateurs du zinc tel que l'EDTA et la phénantroline (Clouaire et al., 2005) (figure 5). De plus des analyses par spectrométrie de masse couplée à un plasma inductif (ICP-MS) ont été réalisées par l'équipe de Bio-cristallographie de notre institut et qui ont permis de montrer que le domaine THAP est lié à un ion zinc.

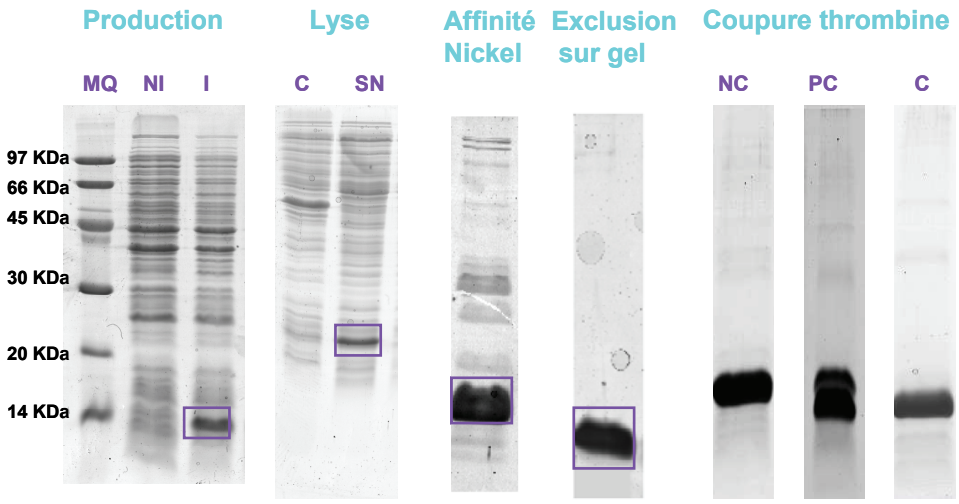


Figure 49 : **Production et purification du domaine THAP de THAP1**
Gels SDS PAGE 16% des différentes étapes du protocole de production et purification. Légende : MQ (marqueur de taille), NI (non induit), I (induit), C (culot), SN (surnageant), NC (non coupée), PC (partiellement coupée), TC (totalement coupée).

Formation du complexe THAP-ADN

La formation du complexe THAP avec sa cible ADN en vue des études de titration par RMN s'est avérée une étape délicate.

En effet, lors du premier test de formation du complexe, nous avons observé l'apparition d'un précipité blanc irréversible. Le spectre HSQC de cette solution après centrifugation rend compte de la complète disparition de la protéine.

Ce phénomène de précipitation protéine-ADN, aux hautes concentrations qu'exige la RMN (au moins 100 µM avec l'utilisation d'une cryosonde), a été décrit dans a littérature (Hard, 1999). Ce problème de solubilité a été résolu de façon empirique en faisant varier les conditions de formation du complexe, comme cela est suggéré (Hard, 1999). Nous avons notamment fait varier les concentrations en protéine et en ADN, la force ionique, le pH, la température, l'utilisation d'EDTA et de Tween pour finalement trouver les conditions de formation du complexe sans précipitation.

Les premiers essais de formation du complexe ont été réalisés sur lame de verre, sur des gouttes de 5 µL (de façon à ne pas consommer trop de protéine ni d'ADN), afin d'observer l'éventuel précipité sur loupe binoculaire.

La protéine est initialement dans le tampon de RMN décrit au protocole 8. Nous avons utilisé un ADN double brin de 14 paires (MWG Biotech) de bases comprenant la séquence consensus :

5' CAAGTATGGGCAAG 3 '
3' GTTCATACCCGTTC 5'

L'hybridation se fait dans l'eau par chauffage à 60°C puis refroidissement à température ambiante. L'hybridation de l'ADN a été vérifiée sur gel d'agarose.

De façon intéressante, lorsque nous utilisons une ADN de séquence aléatoire (cf p 121 pour la séquence) nous n'observons pas de formation de précipité. Cette réaction est donc spécifique et traduit la liaison du domaine THAP à la séquence ADN consensus.

Pour former le complexe spécifique, nous avons déterminé une concentration nécessaire en NaCl de 250 mM pour éviter la précipitation. Cet effet de la force ionique n'est pas surprenant étant donné la charge de la protéine à un pH 6.8 (charge théorique de +10 calculée avec EXPASY, E**x**pert Protein Analysis SYstem) et la large surface positive qu'elle présente (cf fig 4 F de l'article), l'ADN étant chargé

négativement. Nous avons d'ailleurs étudié l'effet de la force ionique sur la formation du complexe par RPS (Résonance Plasmonique de Surface). Les résultats corroborent nos observations puisqu'à haute force ionique, on mesure une perte de l'interaction protéine-ADN (cf p 134).

Nous avons vérifié que THAP liait toujours l'ADN en effectuant une expérience de gel retard dans ces mêmes conditions (tampon d'étude avec une concentration en NaCl de 250mM), puis nous avons réalisé l'étude de la liaison à l'ADN par titration par RMN et RPS.

Initialement nous avons travaillé avec de l'ADN en solution. Nous avons cherché à concentrer au maximum cette solution pour limiter les variations de volume lors de la titration (en ajoutant de la solution d'ADN sur la protéine). Nous avons déterminé une concentration en ADN au maximum 5 fois supérieure à celle de la protéine, un rapport de concentration plus élevé entraînant une précipitation du complexe. Nous travaillons ainsi classiquement avec une concentration en protéine de 300 µM et en ADN de 1,5 mM.

Plus récemment, nous avons finalement travaillé avec de l'ADN lyophilisé directement ajouté dans la solution de protéine ce qui donne un résultat satisfaisant et offre l'avantage d'avoir un volume quasiment constant au cours de la titration.

Nous avons réalisé des spectres ^{15}N HSQC de la protéine à 10 mM NaCl (condition d'étude structurale du domaine seul) à 150 mM NaCl (condition physiologique) et à 250 mM NaCl (condition d'étude du complexe). Nous avons observé des variations de déplacement chimique uniformes selon les résidus mais pas de variation majeure ce qui indique que la protéine garde la même conformation globale lors de l'ajout de sel.

Détermination de la structure du domaine THAP de THAP1

Conditions d'étude

Après avoir produit et conditionné la protéine d'intérêt dans le tampon d'étude RMN, les expériences pour la détermination de structure sont réalisées dans les conditions suivantes :

- pH = 6.8
- température = 296K
- Tampon : Tris 50 mM
  NaCl 10 mM
  DTT deutéré 1 mM
  0.01% NaN_3

Ces conditions ont été optimisées de façon à obtenir une protéine stable et bien structurée. Nous avons ainsi enregistré des spectres ^{15}N HSQC, véritable *empreinte* de la protéine, de façon à apprécier le bon repliement de la protéine caractérisé par une bonne dispersion des signaux. De plus, nous avons également apprécié la quantité de NOEs observables sur des spectres 2D homonucléaires dans différentes conditions d'étude de façon à optimiser ces conditions. Cette étude a été menée sur le domaine THAP 1-90, qui est le domaine sur lequel nous avons initialement travaillé. Nous avons gardé les mêmes conditions d'étude lorsque nous avons ensuite travaillé avec le domaine plus court THAP 1-81.

Effet du pH

Pour l'étude RMN, nous nous plaçons à un pH légèrement acide de façon à pouvoir observer les protons amides en ralentissant leur échange avec l'eau. Nous réalisons ainsi l'étude à un pH de 6.8. La protéine possède un point isoélectrique théorique de 9 et donc est globalement chargée positivement dans les conditions de l'étude. Un pH plus bas, de 6, entraîne une déstructuration de la protéine comme nous pouvons le voir sur les spectres ^{15}N HSQC (figure 50). A pH 6 le spectre présente dans la zone centrale des superpositions des pics de corrélation mettant en évidence une déstructuration partielle de la protéine.

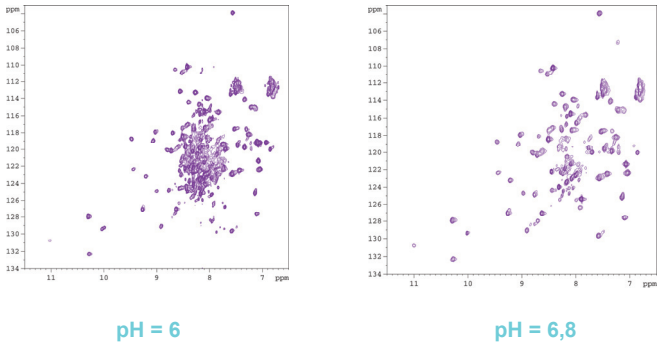


Figure 50 : **Optimisation du pH**
Spectres HSQC du domaine THAP de THAP1 à pH 6 et 6,8. La présence d'une zone de recouvrement à pH 6 est caractéristique d'une déstructuration de la protéine.

Ajout de zinc

Nous avons testé l'ajout d'un excès de zinc dans le tampon final (en plus de l'équivalent de zinc présent dans le site de coordination de la protéine) comme cela peut se faire pour ce genre d'étude sur des protéines à zinc (Kulczyk et al., 2004). L'ajout d'un excès de $ZnCl_2$ à 0.3 mM, pour une solution à 1 mM en protéine, entraîne un élargissement des raies comme nous l'avons observé sur un spectre 1D ^1H (figure 51). Cet élargissement peut par exemple s'expliquer par une oligomérisation de la protéine par les cystéines libres qui pourraient coordonner le zinc ajouté. Nous n'ajouterons donc pas de zinc dans le tampon d'étude.

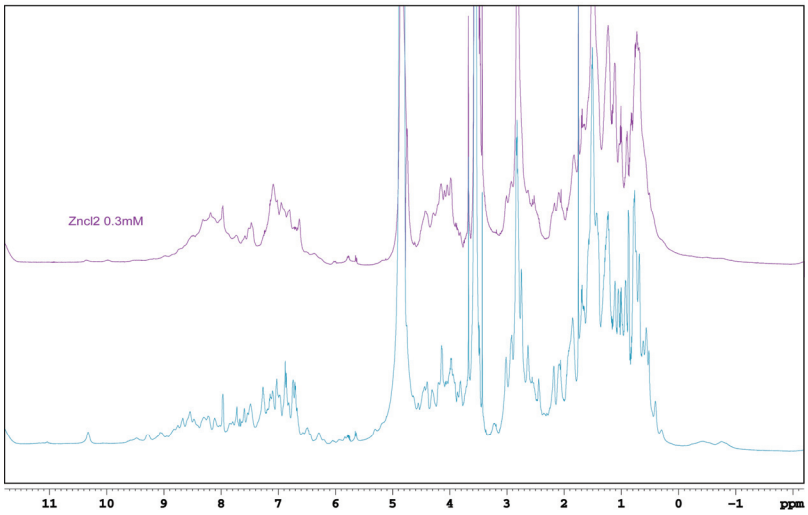


Figure 51 : **Effet du Zinc**
Spectres 1D 1H en absence d'ajout de zinc en bleu et avec 0.3 mM de $ZnCl_2$ en excès en violet pour 1mM de protéine. L'ajout de zinc entraîne un élargissement des raies.

Effet de la température

Nous avons mené une première étude à température égale à 285 K sur le domaine 1-90 et nous avons attribué les fréquences du domaine THAP et identifié les NOEs. Nous avons dans un premier temps choisi une température basse de façon à conserver la protéine stable suffisamment longtemps pour réaliser la campagne d'enregistrement des expériences 3D (cf table 4 p 20) sur le même échantillon doublement marqué. Toutefois le nombre de NOEs recueillis s'est avéré trop faible pour un calcul de structure de qualité. Finalement nous avons réalisé la même étude à 296 K qui donne un nombre plus important de NOEs et a permis de résoudre la structure du domaine THAP de THAP1.

Concentration en NaCl

Nous avons réalisé l'étude à basse force ionique, 10 mM en NaCl. En effet une haute concentration en sel augmente la conductivité de la solution et entraîne une baisse de la sensibilité de la sonde, particulièrement pour une cryosonde (Flynn et al., 2000). Par ailleurs des études par diffusion des RX aux petits angles (Small Angle X-rays Scattering) réalisés dans l'équipe de L. Mourey à l'IPBS indiquent des possibilités de dimérisation à haute force ionique. Toutefois, le spectre ^{15}N HSQC réalisé à 250 mM NaCl est identique à de légères variations de déplacements chimiques près. De même à une concentration en NaCl égale à 150 mM proche des conditions physiologiques nous n'observons pas de variation majeure du spectre ^{15}N HSC.

Enfin nous rajoutons du DTT (dithiothréitol) pour éviter l'oxydation de la protéine et du NaN_3 pour inhiber la croissance bactérienne.

Acquisition et transformation des spectres

Les expériences RMN ont été enregistrées sur un spectromètre Bruker Avance 600 MHz muni d'une cryosonde et sur un spectromètre 700 MHZ Bruker Avance disponibles au laboratoire. Nous avons également eu accès au spectromètre 800

MHz Bruker Avance à l'ICSN (Gif/Yvette) et y avons enregistré deux expériences NOESY et TOCSY.

Nous avons optimisé les temps de mélange pour l'expérience NOESY (120 ms pour une expérience 2D à 600MHz) et l'expérience TOCSY (60 ms pour une 2D à 600 MHz)

Nous avons utilisé le logiciel Topspin (Bruker) pour l'acquisition des spectres. Le traitement des spectres a été réalisé avec le logiciel NMRPipe (Delaglio et al., 1995). Les différentes fonctions pour la transformation du spectre sont appliquées successivement selon un schéma de *pipelines*, les dimensions sont transformées les unes après les autres. Un exemple de macro utilisée pour le traitement d'un spectre en trois dimensions est figuré ci-dessous (figure 52)

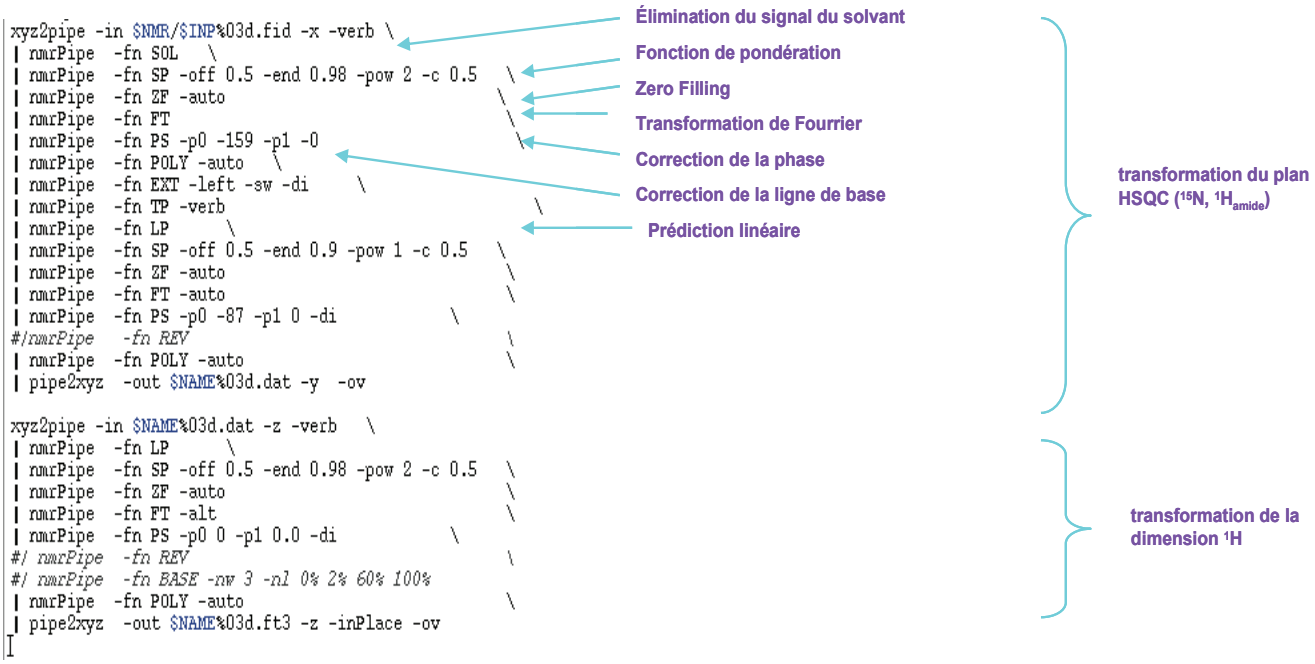


Figure 52 : **Transformation des données**
Exemple de macro utilisé sous NMRPipe pour le traitement d'un spectre en trois dimensions (NOESY ^{15}N HSQC), les différentes fonctions sont appliquées successivement.

Attribution

Nous avons dans un premier temps réalisé l'attribution du domaine THAP de THAP1 1-90 tel qu'il a été défini dans l'équipe de J.P Girard (Roussigne et al., 2003b). Nous avons produit un échantillon doublement marqué ^{15}N ^{13}C et réalisé l'attribution séquentielle à l'aide des fréquences carbones puis l'attribution de l'ensemble des

systèmes de spins et des NOEs selon la stratégie standard telle que nous l'avons décrite (figure 38). Une figure des bandes extraites des expériences 3D des expériences CBCACONH et HNCACB pour l'attribution séquentielle des résidus 6 à 11 est représenté sur la figure 53.

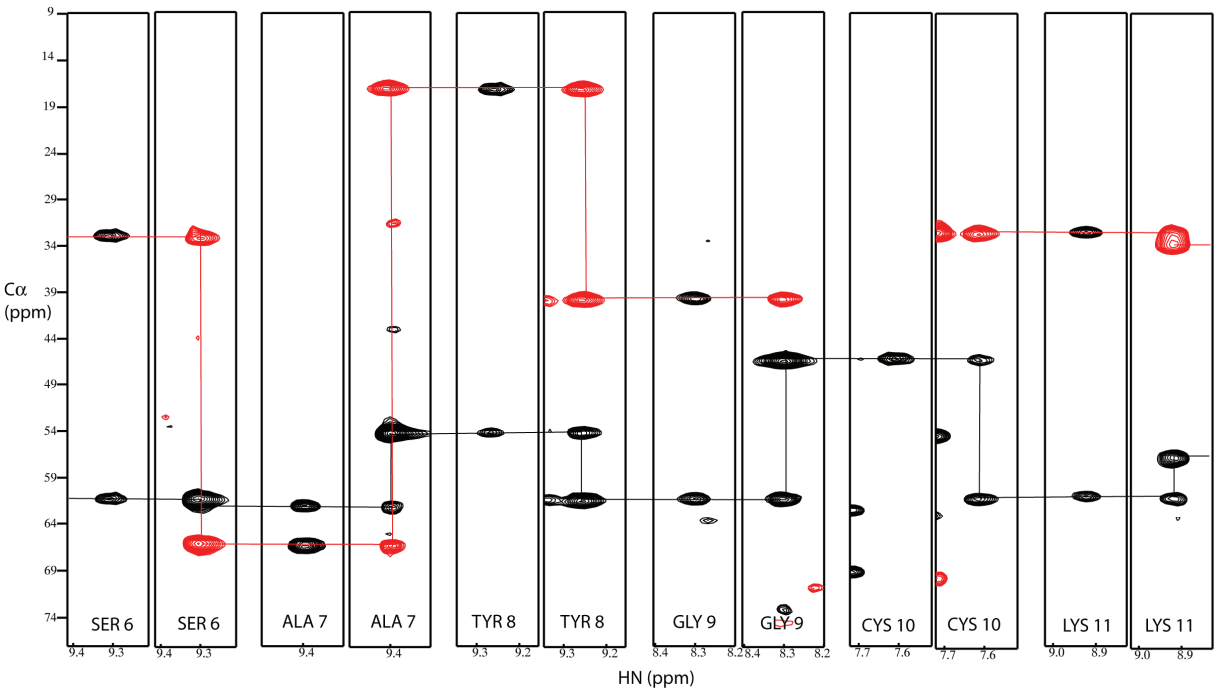


Figure 53 : **Attribution séquentielle des résidus 6 à 11**
Pour chaque résidu, les bandes extraites des plans HSQC des expériences CBCACONH (à gauche, corrélations NHi - CBCA i-1) et HNCACB (à droite, corrélations NHi – CBCA i et i-1) sont figurées côte à côte (une bande par NH amide de la protéine). Pour l'expérience HNCACB les taches des Cα sont positives (noires) et celles des Cβ négatives (rouge). Les correspondances de fréquences sont représentées par des traits.

Nous avons utilisé les logiciels XEASY (Bartels et al., 1995) pour l'attribution des expériences 2D et NMRView (Johnson, 2004) pour l'attribution des expériences 3D. La conversion des fichiers pour passer d'un logiciel à un autre a été effectuée avec le logiciel Format converter de la suite CCPNMR (Collaborative Computing Project for NMR). Il faut néanmoins être vigilant sur les nomenclatures utilisées. Par exemple, dans le cas des protons Hβ, NMRview utilise les dénominations HB1 et HB2 tandis que XEASY utilise les dénominations HB2 et HB3. Le logiciel de conversion change HB1 en HB3 pour passer des fichiers NMRview à des fichiers XEASY et ne change donc pas selon le schéma : HB1 en HB2 et HB2 en HB3.

Les attributions des fréquences ^{15}N, ^{1}H et ^{13}C du domaine {Met1-Lys90} ont été déposées dans la BMRB (Biological Magnetic Resonance Data Bank) sous le code d'accès 15300.

Par contre, la partie C-terminale (78 à 90) présente peu de NOEs du fait de sa dynamique interne et un fort recouvrement spectral et nous n'avons pu obtenir qu'une attribution partielle de cette région.

Nous avons donc été amenés à redéfinir le domaine THAP {Met1-Phe81} comme nous l'expliquons dans la partie suivante. Les spectres des domaines 90 et 81 sont superposables et nous nous sommes largement aidés de l'attribution du domaine 1-90 pour reprendre l'attribution du domaine 1-81. Ainsi nous n'avons pas eu besoin des expériences tridimensionnelles avec le noyau carbone pour l'attribution du domaine 1-81 mais seulement d'un marquage ^{15}N. Nous avons également déposé les fréquences ^{1}H et ^{15}N du domaine {Met1-Phe81} dans la banque de données BMRB sous le code d'accès 15289. Un spectre HSQC annoté avec l'attribution des résidus est présenté (figure 37).

L'attribution des NOES et la détermination des contraintes structurales du domaine THAP 1-81 ont été réalisées à partir de spectres homonucléaires NOESY et 3D ^{15}N HSQC enregistrés à 296K. Deux autres jeux de données homonucléaires ont été enregistrés à des températures de 288k et 280K, afin de lever des ambiguïtés en cas de recouvrement. Nous avons également enregistré des spectres dans le D_2O de façon à observer spécifiquement les contacts des protons aromatiques et aliphatiques en s'affranchissant du signal résiduel de H_2O.

L'attribution stéréospécifique de 15 protons Hβ a été réalisée en utilisant un spectre COSY (Correlated SpectroscopY) (Rance et al., 1983) et avec la stratégie décrite (figure 42).

Contraintes structurales

Contraintes de distance

Les contraintes de distance issues des NOEs ont été extraites à partir d'une expérience 2D NOESY et d'une expérience 3D ^{15}N HSQC-NOESY. Les volumes définissant la valeur des NOEs ont été intégrés avec les deux logiciels XEASY et

NMRview respectivement pour la 2D et la 3D. Une attention toute particulière a été apportée à l'intégration des volumes de la 2D NOESY, c'est l'expérience qui donne l'ensemble des distances et qui présente des superpositions de pics de corrélations. Il s'agit donc de bien définir les rectangles de définition de chaque pic pour une intégration correcte. Comme nous l'avons vu dans le chapitre d'introduction, les intensités de NOEs ont été classées en trois groupes : faible, moyen et fort auxquels on a associé une distance et un intervalle de valeurs possibles, respectivement 5 $\pm$ 1 ; 3,3 $\pm$ 0,8 et 2,2 $\pm$ 0,3 Å. La marge supérieure est majorée dans le cas des pseudo-atomes, de 1 Å dans le cas de deux pseudo-atomes (par exemple lorsque les protons Hβ sont confondus) et de 2,5 dans le cas de trois pseudo-atomes (un groupe méthyle). Les distances sont limitées à 6 Å. Pour la calibration, il convient de déterminer les intensités définissant les deux bornes du groupe moyen.

Toutes les distances extraites sont non-ambigües, c'est-à-dire qu'elles correspondent à une attribution unique effectuée par nos soins.

Nous avons ainsi identifié 801 contraintes de distance intrarésiduelle, 329 contraintes séquentielles, 103 distances à moyenne portée et 306 distances à longue portée (cf table 1 de l'article)

Nous avons également paramétré 10 liaisons hydrogène à partir de l'identification des structures secondaires définis à partir d'observation de NOEs caractéristiques, de la valeur des déplacements chimiques (Chemical Shift Index) et de la cinétique d'échange des hydrogènes amides (figure 40) en deutérium comme nous l'avons vu dans le chapitre d'introduction.

Voyons un exemple pour le paramétrage du feuillet β antiparallèle entre les résidus 22-23-24 et 52-53-54, nous utilisons la syntaxe compatible avec le logiciel CNS (http://cns.csb.yale.edu) pour le calcul de structure.

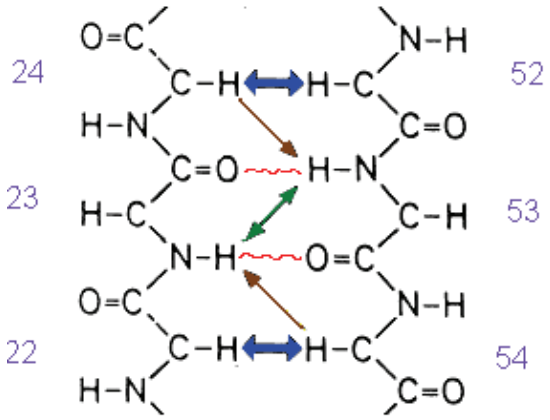


Figure 54 : **Distances caractéristiques pour le feuillet β antiparallèle 22-23-24 52-53-54**

Les deux liaisons hydrogène en rouge sont paramétrées de la sorte :

```
assign (resid  53 and name     O ) ( resid  23 and name     N )  2.90  0.30  0.50 !
assign (resid  53 and name     O ) ( resid  23 and name    HN )  1.90  0.50  0.10 !
assign (resid  23 and name     O ) ( resid  53 and name     N )  2.90  0.30  0.50 !
assign (resid  23 and name     O ) ( resid  53 and name    HN )  1.90  0.50  0.10 !
```

Le premier nombre correspond à la valeur de la distance en Á le deuxième et le troisième les incertitudes inférieure et supérieure. Ces valeurs théoriques, ainsi que les suivantes, peuvent être retrouvées dans l'ouvrage de Wüthrich (Wüthrich, 1986). Ces distances avec des oxygènes ne sont pas mesurables directement par les NOEs. Mais on peut également définir des liaisons spécifiques des feuillets β antiparallèles entre deux hydrogènes, qui doivent être retrouvées sur nos cartes NOEs, d'après les couleurs de la figure 54 :

```
assign (resid  54  and name   HA ) ( resid  22 and name    HA )  2.3   0.50  0.30 !
assign (resid  52  and name   HA ) ( resid  24 and name    HA )  2.3   0.50  0.30 !
assign (resid  53  and name   HN ) ( resid  23 and name    HN )  3.3   1.50  0.75 !
assign (resid  54  and name   HA ) ( resid  23 and name    HN )  3.2   1.50  0.75 !
assign (resid  24  and name   HA ) ( resid  53 and name    HN )  3.2   1.50  0.75 !
```

De même pour les régions en hélice α on définit une liaison hydrogène entre les résidus i et i+3 ainsi qu'une liaison HN - Ha, comme par exemple entre les résidus 36 et 33 :

```
assign (resid  36 and name    HN ) ( resid  33 and name     O )  1.90  0.50  0.10 !
assign (resid  36 and name     N ) ( resid  33 and name     O )  2.90  0.30  0.50 !
assign (resid  36 and name    HN ) ( resid  33 and name    HA )  3.40  0.50  0.50 !
```

Nous avons ainsi paramétré l'ensemble des contraintes de distances caractéristiques des structures secondaires; soit pour les résidus prédits dans une hélice α (résidus 31 à 41) soit les résidus prédits dans un feuillet β (résidus 22 à 24 et 52 à 54).

Enfin nous avons utilisé 14 contraintes de coordination au zinc, entre l'atome de zinc et les résidus 5, 10, 54 et 57 formant la signature C2CH. Pour les cystéines, le ligand au zinc est l'atome de soufre SG pour l'histidine nous avons déterminé l'azote NE2 plutôt que ND1 pour former le tétraèdre de coordination au zinc (figure 7). Les paramètres de distances sont ainsi les suivants :

```
assign (resid   1 and name    ZN ) ( resid   5 and name    SG )  2.35  0.02  0.02 !
assign (resid   1 and name    ZN ) ( resid   5 and name    CB )  3.38  0.13  0.02 !
assign (resid   1 and name    ZN ) ( resid  10 and name    SG )  2.35  0.02  0.02 !
assign (resid   1 and name    ZN ) ( resid  10 and name    CB )  3.38  0.13  0.02 !
assign (resid   1 and name    ZN ) ( resid  54 and name    SG )  2.35  0.02  0.02 !
assign (resid   1 and name    ZN ) ( resid  54 and name    CB )  3.38  0.13  0.02 !
assign (resid   1 and name    ZN ) ( resid  57 and name   NE2 )  2.00  0.05  0.05 !
assign (resid   1 and name    ZN ) ( resid  57 and name   ND1 )  4.15  0.03  0.03 !
assign (resid   5 and name    SG ) ( resid  10 and name    SG )  3.76  0.20  0.20 !
assign (resid   5 and name    SG ) ( resid  54 and name    SG )  3.76  0.20  0.20 !
assign (resid  54 and name    SG ) ( resid  10 and name    SG )  3.76  0.20  0.20 !
assign (resid   5 and name    SG ) ( resid  57 and name   NE2 )  3.52  0.20  0.20 !
assign (resid  10 and name    SG ) ( resid  57 and name   NE2 )  3.52  0.20  0.20 !
assign (resid  54 and name    SG ) ( resid  57 and name   NE2 )  3.52  0.20  0.20 !
```

Il convient dans un premier temps de générer des structures sans ces contraintes, de vérifier l'existence de ces coordinations pour ensuite intégrer ces paramètres au calcul de façon à améliorer les structures. Pour déterminer quelle solution est en accord avec l'ensemble des contraintes structurales, on compare l'énergie des structures générées avec la coordination à ND1 ou NE2.

L'état de protonation de ND1 ou NE2 et le choix du ligand au zinc qui en découle (l'azote déprotoné), peut aussi être observé sur des expériences long-range ^{15}N-HSQC (observation des corrélations proton amide-azote sur une large fenêtre spectrale) (Pelton et al., 1993). Nous avons ainsi observé sur une telle expérience un pic de corrélation à (11,29 ppm ; 166,36 ppm) aussi observé sur les expériences 2D NOESY et 3D ^{15}N HSQC NOESY et qui a pu être attribué à (HD1 ; ND1) d'où une coordination avec l'atome NE2 déprotoné (comme sur la figure 7). L'observation de ce proton HD1 est inhabituel car il est en général en échange rapide avec l'eau (Liew et al., 2007) (figure 55).

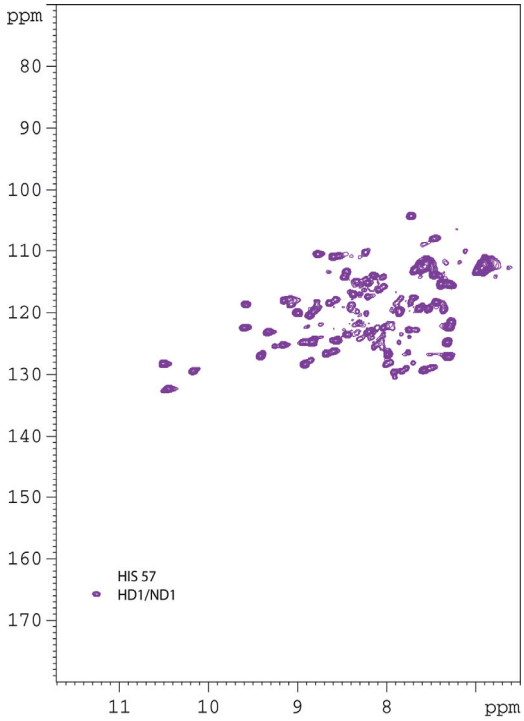


Figure 55 : **HSQC à large fenêtre spectrale**

De plus, on relève des contacts NOEs entre les résidus de la signature C2CH mettant en évidence leur proximité spatiale :

```
assign (resid   5 and name   HB2 ) ( resid  54 and name   HB1 )  1.60  0.30  0.30 !
assign (resid  10 and name    HA ) ( resid  57 and name   HE1 )  5.00  1.00  1.00 !
assign (resid   5 and name   HB1 ) ( resid  54 and name   HB1 )  3.30  0.80  0.70 !
assign (resid  54 and name    HA ) ( resid  57 and name   HD2 )  5.00  1.00  1.00 !
assign (resid  54 and name    HN ) ( resid   6 and name    HN )  5.00  1.00  1.00 !
assign (resid  54 and name    HA ) ( resid  56 and name    HN )  5.00  1.00  1.00 !
```

Contraintes d'angles

Des contraintes d'angles ψ et ϕ ont été paramétrées pour les résidus prédits dans des structures secondaires : résidus 31 à 40 prédits dans une hélice α, résidus 22 à 24 et 52 à 54 prédits dans un feuillet β. Les valeurs canoniques ont été utilisées (Wüthrich, 1986) ce qui donne par exemple pour le résidu 36 dans l'hélice α et le résidu 23 en feuillet β :

```
assign (resid  35  and name c ) (resid  36  and name n )
       (resid  36  and name ca) (resid  36  and name c )   1.0   -57.0 20.0 2
assign (resid  36  and name n ) (resid  36  and name ca)
       (resid  36  and name c ) (resid  37  and name n )   1.0   -47.0 20.0 2

assign (resid  22  and name c ) (resid  23  and name n )
       (resid  23  and name ca) (resid  23  and name c )   1.0   -140.0 20.0 2
 assign (resid  23  and name n ) (resid  23  and name ca)
       (resid  23  and name c ) (resid  24  and name n )   1.0   135.0 20.0 2
```

Des constantes d'angle dièdre ψ et ϕ ont été recueillies à partir de la valeur des déplacements chimiques avec le programme TALOS. Pour les résidus situés dans des structures secondaires, nous n'avons pas utilisé ces valeurs prédites par TALOS mais les valeurs canoniques.

Nous avons également déterminé des valeurs de constantes de couplage $^3J_{H\alpha,HN}$ déterminées à partir de l'expérience HNHA (cf p 71). Nous avons utilisé pour ce faire l'implémentation du logiciel NMRview. Ces valeurs nous ont permis de confirmer les structures secondaires (un résidu situé dans feuillet β est caractérisé par un couplage supérieur à 8 Hz et un résidu situé dans hélice α est caractérisé par un couplage entre 2 et 6 Hz). Nous n'avons pas utilisé ces valeurs comme contraintes dans le calcul final de structure puisqu'elles sont redondantes avec les autres contraintes et qu'elles n'apportaient pas d'amélioration au calcul.

Nous n'avons pas tenu compte des valeurs prédites par Talos et des couplages $^3J_{H\alpha,HN}$ mesurés pour les résidus dans les régions flexibles (résidus 1 à 3, 17 à 20, 41 à 45, et 64 à 68), déterminées par des mesures de dynamique et l'absence de NOE (fig 2 C et fig suppl 1 de l'article)

L'ensemble des valeurs ψ et ϕ déterminées par Talos et les valeurs de couplage $^3J_{H\alpha,HN}$ sont reportées dans la table suivante.

résidu	J (Hz)	ΔJ (Hz)	ϕ (°) Talos	Δϕ (°)	ψ (°) Talos	Δψ (°)
4						
5						
6	8,1	0,1	-77	24	-41	54
7	5,2	0,1				
8			-68	24	138	10
9						
10			-76	30	137	28
11	9,4	0				
12						
13	9,3	0,1				
14	3,8	0,1	-68	10	131	50
15	9,3	0	-90	35	140	54
16	3,8	0,2	-73	20	-43	22
21	5,7	0,1	-106	21	154	38
22	9,1	0,1	-113	36	130	15
23	9	0	-85	25	127	18
24	8,5	0,1	-93	26	141	25
25	6,3	0,4				
26						
27			-73	19	-37	28
28	9	0,2				
29	8,8	0,2				
30			-63	30	-36	30
31	5,3	0,1	-66	15	-40	30
32			-64	17	-44	20
33			-66	16	-42	22
34	3,8	0,3	-64	15	-37	18
35	6,6	0	-66	15	-43	17
36			-64	16	-43	23
37			-65	15	-41	25
38	5,4	0	-70	13	-38	26
39	5,3	0	-65	15	-38	19
40	5,8	0,4	-68	10	-36	20
46	9,3	0,1	-104	24	125	32

résidu	J (Hz)	ΔJ (Hz)	ϕ (°) Talos	Δϕ (°)	ψ (°) Talos	Δψ (°)
47			-76	42	148	18
48	10,7	0				
49			-79	21	-14	16
50	10,2	0,1	-89	19	-15	15
51	5	0,1				
52	7,9	0	-139	18	137	14
53	10,1	0	-127	18	128	18
54	6,1	0,6	-85	17	160	17
55			-71	19	-37	13
56	7,2	0,1	-67	11	-31	20
57			-95	13	-6	20
58	10,2	0,1	-111	27	144	19
59	6,6	0,1	-86	24	141	27
60			-64	18	-34	13
61			-70	19	-34	16
62	7,5	0				
63	9,2	0				
65	8,6	0,2				
66						
67						
68						
69	6,8	0,1				
70	10	0,1				
71						
72	8	0,2				
73						
74	2,9	0,2	-71	16	-29	12
75	10	0,2	-96	18	-6	19
76	4	0,1	-82	16	119	24
77	9,7	0,1	-114	29	137	32
78			-70	14	-27	15
79			-84	15	-30	17
80	6,5	0,1	-61	10	-44	28
81						

Table 3 : **Constantes de couplage $^3J_{H\alpha,HN}$ et valeurs des angles ϕ et ψ déterminées par TALOS**
Les résidus prédits situés dans des feuillets β sont en vert et ceux dans l'hélice α en rouge. Les valeurs des résidus dans les parties flexibles ne sont pas répertoriées.

Nous avons ainsi recueilli 1539 contraintes de distances et 104 contraintes d'angles pour le calcul de la structure du domaine THAP de THAP1 dont le détail peut être retrouvé dans la table 1 de l'article, *structural statistics of the THAP-zinc finger of THAP.*

Calcul de structure

Pour le calcul de structure, nous avons utilisé le programme CNS (Crystallography and NMR System) (Brunger et al., 1998)

Nous avons utilisé dans un premier temps un protocole de géométrie des distances de façon à générer un premier ensemble de 20 structures. La structure d'énergie la plus basse issue de ce premier calcul a été utilisée comme structure de référence pour effectuer le calcul avec un protocole de recuit simulé (figure 44). De cette façon, nous partons de structures de départ repliées plutôt que de structures étendues. Le calcul s'est fait sur 500 structures et nous sélectionnons les 20 meilleures structures d'énergies les plus basses.

Les calculs ont été effectués en faisant varier les angles de torsion et non les cordonnées cartésiennes pour augmenter l'efficacité du calcul (Stein et al., 1997) en diminuant le nombre de degrés de liberté du système.

Nous avons également utilisé le logiciel ARIA (Linge et al., 2001) pour effectuer le calcul de structure en même temps qu'une attribution automatique de contraintes ambigües. Toutefois, cela ne nous a pas permis de définir d'autres contraintes. En effet, les attributions proposées par ce logiciel ne sont pas en accord avec la structure globale de la protéine et la structure qui en découle n'est donc pas de bonne qualité. Nous pensons que cela est dû à un recouvrement critique des fréquences dans le cas de notre protéine qui engendre des erreurs lors de l'attribution automatique par ce logiciel, même lorsque nous baissons les tolérances d'attribution, jusqu'à 0.01 ppm pour les protons. La superposition des pics nous a donc amenés à effectuer l'attribution exclusivement manuellement avec un processus itératif en retour de calculs de structures en cours d'attribution comme nous l'avons expliqué (figure 30).

Une des particularités de ce calcul est la présence de l'ion zinc qu'il faut ajouter dans le fichier de topologie et les fichiers pdb. Il convient également de fixer les coordonnées de cet atome et en particulier de définir sa position par rapport aux atomes de la protéine, sans quoi le calcul utilisant les angles de torsion n'est pas possible. On utilise pour se faire les paramètres suivants pour la coordination au zinc, définissant les distances et les angles de la géométrie tétraédrique :

```
{- ZN coordination  -}

bond      ZN  SH1E       1000 2.3
bond      ZN  NH1        1000 2.0
angles SH1E ZN SH1E 500 109.5
angles ZN SH1E CH2E 500 110
angles SH1E ZN NH1  500 109.5
angles NH1  ZN SH1E  500 109.5
angles ZN  NH1 CR1H 500 126
```

Nous avons ensuite utilisé le programme Procheck pour apprécier la qualité des structures générées (Laskowski et al., 1996). Le diagramme de Ramachandran est représenté figure 45 et un tableau regroupant les statistiques structurales est présenté dans l'article (table1). Nous obtenons ainsi un écart quadratique moyen des coordonnées sur les résidus (4-16,21-40,46-63,69-81) de 0.530 Å sur les atomes du squelette et 1.297 Å sur l'ensemble des atomes lourds. Nous avons exclu dans ce calcul les régions flexibles d'après les données de dynamique et de recueil de contraintes NOEs comme nous le verrons dans le chapitre suivant.

Les structures générées présentent les structures secondaires prédites, c'est-à-dire l'hélice α du résidu 31 à 40 et les deux brins β formés par les résidus 22 à 24 et 52 à 54. Toutefois les résidus 31 et 32 ne seront pas définis comme faisant partie de l'hélice d'après l'analyse du logiciel Procheck. On note également la présence de 3 hélices 3_{10}, définis par les résidus 55 à 57, 60 à 63 et 79 à 81. Nous avons pu observer des NOE caractéristiques pour ces éléments de structure secondaire.

L'ensemble des 20 structures de plus basse énergie ont été déposées dans la PDB (Protein Data Bank) sous le code d'accès 2jtg.

ETUDE DE LA STRUCTURE ET DE LA LIAISON A L'ADN DU DOMAINE THAP DE THAP1

Présentation des résultats publiés

L'ensemble des résultats de notre étude sur la structure du domaine THAP de THAP1 et de son interaction avec l'ADN a donné lieu à une publication. Nous proposons de reprendre brièvement les résultats publiés dans cet article qui est inséré ensuite au paragraphe suivant.

L'article présente l'identification du domaine THAP restreint à 81 acides aminés au lieu de 90, sa structure en solution, l'identification de l'interface de liaison à l'ADN par mutagénèse dirigée et variations de déplacement chimique et également l'étude par RPS (résonance plasmonique de surface) de l'interaction du domaine THAP de THAP1 avec sa séquence ADN cible.

Ce travail est l'objet d'une collaboration avec l'équipe de biologie vasculaire de l'IPBS. Les expériences de mutagénèse dirigée et de gel retard ont été réalisées par Chrystelle Lacroix, doctorante dans cette équipe.

Dans ce chapitre nous faisons référence aux figures de l'article.

Caractérisation biophysique du domaine THAP de THAP1

Le domaine THAP, comme nous l'avons vu, est à l'origine défini par les 90 résidus N-terminaux de THAP1 (Roussigne et al., 2003b).

Nous avons réalisé les premières expériences de RMN sur ce domaine ainsi que des expériences de RPS.

Pour les expériences de RPS, nous avons enregistré les sensogrammes (mesure du signal RPS en fonction du temps) correspondant à l'injection de protéine sur les molécules d'ADN immobilisées. Nous avons travaillé simultanément avec une séquence d'ADN comprenant la séquence consensus reconnue par le domaine THAP de THAP1 et une séquence d'ADN aléatoire. De cette façon, nous pouvons étudier la réponse spécifique en réalisant la soustraction des deux sensogrammes. Nous avons réalisé l'étude de la liaison en injectant la protéine à différentes concentrations (fig 1 A). Nous avons ainsi pu déterminer une constante de dissociation (K_D) globale de 8 µM, spécifique, et une stœchiométrie de 1 : 1, à partir de la courbe des réponses à l'équilibre en fonction de la concentration en protéine (fig 1 B).

Nous avons réalisé l'attribution du domaine 1-90 avec la stratégie déjà décrite. Mais nous n'avons qu'une attribution partielle de la partie C-terminale (78 à 90) pour laquelle nous avons observé très peu de NOEs non séquentiels. Cette partie présente de plus des valeurs de NOE heteronucléaires négatives indiquant que l'extrémité C-terminale du domaine n'est pas structurée. Nous avons donc cherché à déterminer un domaine de taille plus courte, capable de lier spécifiquement la cible ADN consensus. Nous avons pu identifier un domaine de 81 résidus {Met1-Phe81} dont nous avons résolu la structure.

Structure par RMN en solution du doigt de zinc THAP de THAP1

Nous avons déterminé la structure tridimensionnelle du domaine THAP {Met1-Phe81} de THAP1 comme nous l'avons précédemment décrit. Les contraintes de distances et d'angles sont répertoriées ainsi que les statistiques structurales (Table 1).

Le cœur de la structure adopte une architecture $\beta\alpha\beta$ (fig 2 A-B). L'hélice α est composée des résidus 33 à 40 et les deux brins β des résidus 22 à 24 et 52 à 54. Les résidus Cys5, Cys10, Cys54, et His57 forment le site de coordination au zinc. Trois courtes hélices 3_{10} sont formées par les résidus 55 à 57, 60 à 63 et 79 à 81. L'analyse des NOEs hétéronucléaires nous a permis d'apprécier les parties mobiles (sur l'échelle des temps ps-ns) en accord avec les régions moins bien définies sur l'ensemble des structures générées (fig 2 C et fig supplémentaire 1).

Les protons amides protégés de l'échange H_2O/D_2O correspondent à des résidus dans l'hélice α (résidus 35, 36, 37, 39 et 40) le feuillet β (résidus 53 et 54) et des résidus proches du site de liaison au zinc (résidus 6, 7 10 et 12). Ces protons protégés sont situés dans des régions rigides (fig 2 C) présentant des éléments de structure secondaire ou étant structurés par la liaison au zinc. Tous les autres protons amides s'échangent rapidement avec le solvant.

Parmi les résidus conservés du domaine THAP, le tryptophane 36 situé sur l'hélice α est un élément clef de la structure, au centre d'un cœur hydrophobe. Le motif conservé AVPTIF joue un rôle essentiel dans le repliement de la protéine en reliant l'hélice α et la partie C-terminale.

Une séquence ADN cible non partagée parmi les domaines THAP

Les structures en solution de deux autres domaines THAP ont été récemment déterminées, la structure du domaine THAP de THAP2 humaine (code PDB : 2d8r) et le domaine THAP de CtBP (*Caenorhabditis elegans* C-terminal binding protein) (code pdb 2jm3) (Liew et al., 2007).

Nous avons cherché à savoir si ces domaines qui semblent partager la même structure globale tridimensionnelle (fig3 A) étaient capables de reconnaître la même séquence consensus reconnue par THAP1. Nous avons montré que les domaines THAP de THAP2 et THAP3 ne liaient pas cette séquence ADN de façon spécifique (fig 3 B). Le domaine THAP de CtBP se lie à cette séquence mais pas de façon spécifique tout comme le domaine THAP de GON-14 (fig 3 C).

Analyse structure-fonction du domaine THAP de THAP1 par mutagénèse dirigée

Dans le but de définir les résidus nécessaires pour la liaison à l'ADN, des analyses par mutagénèse dirigée ont été menées dans l'équipe de biologie Vasculaire par C. Lacroix. Les liaisons à l'ADN de 30 simples mutants et 6 triples mutants de résidus en alanines ont été testées par gel retard (Fig 4 A-D et table 2).

Six résidus sont identifiés dont la mutation ponctuelle abolit complètement la liaison à l'ADN : les résidus 36, 24, 29, 42, 45 et 48. Il apparaît évident que la mutation du tryptophane 36 au centre du cœur hydrophobe entraîne la déstructuration de la protéine qui n'est plus capable de lier l'ADN. De même les résidus 29 et 45 sont orientés à l'intérieur de la protéine et ont un rôle structural, leur mutation pourrait provoquer une déstructuration de la protéine (fig 4 E). Par contre les résidus 24, 42 et 48 sont exposés à la surface de la protéine et sont situés sur une aire chargée positivement, présentant de nombreuses lysines et arginines, (fig 4 F) qui suggère que ces trois résidus sont directement impliqués dans la liaison à l'ADN.

Identification de la surface de liaison du domaine THAP avec l'ADN par variations de déplacements chimiques

Dans le but de caractériser la surface de liaison à l'ADN, nous avons réalisé la titration du domaine THAP marqué ^{15}N par l'ajout d'ADN en enregistrant des spectres ^{15}N HSQC à différents rapports ADN/Protéine.

Lorsque nous ajoutons de l'ADN de séquence aléatoire nous n'observons pas de variation de déplacement chimique. Par contre lors de l'ajout d'ADN de séquence consensus, plusieurs pics de corrélations sont affectés (fig 5A).

Après l'ajout d'ADN à un ratio protéine:ADN de 1:1 plus aucune variation n'est observée, confirmant la stœchiométrie 1:1 du complexe.

Les acides aminés affectés par la liaison à l'ADN sont identifiés par comparaison des déplacements chimiques ^{15}N et de ^{1}H amides de la protéine libre et en complexe avec l'ADN (fig 5 B). Ils sont regroupés en trois zones, formées des résidus (7, 8, 11, 13, 20, 52, 53, 54, 56) proche du site de liaison au zinc, (29, 36, 38, 39) proche de l'hélice α et (68, 69, 70, 72, 76) dans une boucle rigide (L4, fig 2 B de l'article) située entre les deux hélices 3_{10} C-terminales. De façon intéressante, les résidus présentant les plus grandes variations de déplacement chimiques sont situés sur la surface chargée positivement identifiée également par mutagénèse (fig 4 F), soulignant le rôle pressenti de cette région comme surface d'interaction avec l'ADN.

Structure-function analysis of the THAP-zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 283, NO. 7, pp. 4352–4363, February 15, 2008

Structure-Function Analysis of the THAP Zinc Finger of THAP1, a Large C2CH DNA-binding Module Linked to Rb/E2F Pathways*[S]

Received for publication, September 10, 2007, and in revised form, October 24, 2007 Published, JBC Papers in Press, December 11, 2007, DOI 10.1074/jbc.M707537200

Damien Bessière[‡§1], Chrystelle Lacroix[‡¶1], Sébastien Campagne[‡§], Vincent Ecochard[‡¶], Valérie Guillet[‡‖], Lionel Mourey[‡‖], Frédéric Lopez[**], Jerzy Czaplicki[‡§], Pascal Demange[‡§], Alain Milon[‡§], Jean-Philippe Girard[‡¶2], and Virginie Gervais[‡§3]

From the [‡]University of Toulouse, Institute of Pharmacology and Structural Biology, the [§]Laboratory of NMR and Protein-Membrane Interactions, the [¶]Laboratory of Vascular Biology, and the [‖]Laboratory of Structural Biophysics, IPBS-CNRS-UPS, 31077 Toulouse, France and [**]INSERM-IFR31, CHU Rangueil, 31059 Toulouse, France

THAP1, the founding member of a previously uncharacterized large family of cellular proteins (THAP proteins), is a sequence-specific DNA-binding factor that has recently been shown to regulate cell proliferation through modulation of pRb/E2F cell cycle target genes. THAP1 shares its DNA-binding THAP zinc finger domain with *Drosophila* P element transposase, zebrafish E2F6, and several nematode proteins interacting genetically with the retinoblastoma protein pRb. In this study, we report the three-dimensional structure and structure-function relationships of the THAP zinc finger of human THAP1. Deletion mutagenesis and multidimensional NMR spectroscopy revealed that the THAP domain of THAP1 is an atypical zinc finger of ~80 residues, distinguished by the presence between the C2CH zinc coordinating residues of a short antiparallel β-sheet interspersed by a long loop-helix-loop insertion. Alanine scanning mutagenesis of this loop-helix-loop motif resulted in the identification of a number of critical residues for DNA recognition. NMR chemical shift perturbation analysis was used to further characterize the residues involved in DNA binding. The combination of the mutagenesis and NMR data allowed the mapping of the DNA binding interface of the THAP zinc finger to a highly positively charged area harboring multiple lysine and arginine residues. Together, these data represent the first structure-function analysis of a functional THAP domain, with demonstrated sequence-specific DNA binding activity. They also provide a structural framework for understanding DNA recognition by this atypical zinc finger, which defines a novel family of cellular factors linked to cell proliferation and pRb/E2F cell cycle pathways in humans, fish, and nematodes.

Zinc finger proteins represent the most abundant class of DNA-binding proteins in the human genome. Zinc fingers have been defined as small, functional, independently folded domains that require coordination of a zinc atom to stabilize their structure (1). The zinc finger superfamily includes the C2H2-type zinc finger, a compact ~30-amino acid DNA-binding module repeated in multiple copies in the protein structure (2, 3), the C4-type zinc finger found in the GATA family of transcription factors (4), and the zinc-coordinating DNA-binding domain of nuclear hormone receptors (5). We recently described an atypical zinc finger motif, characterized by a large C2CH module (Cys-X_{2-4}-Cys-X_{35-53}-Cys-X_2-His) with a spacing of up to 53 amino acids between the zinc-coordinating C2 and CH residues (6). This motif, designated THAP domain or THAP zinc finger, defines a previously uncharacterized large family of cellular factors with more than 100 distinct members in the animal kingdom (6, 7). We showed that the THAP domain of THAP1, the prototype of the THAP family (8), possesses zinc-dependent sequence-specific DNA binding activity and recognizes a consensus DNA target sequence of 11 nucleotides (THABS, for the THAP1 binding sequence) (7), considerably larger than the 3–4 nucleotides motif typically recognized by classical C2H2 zinc fingers (2, 7). Interestingly, the consensus C2CH signature of the THAP domain was identified in the sequence-specific DNA-binding domain of *Drosophila* P element transposase, suggesting the THAP zinc finger constitutes a novel example of a DNA-binding domain shared between cellular proteins and transposons from mobile genomic parasites (6, 9).

Although the biological roles of cellular THAP proteins remain largely unknown, data supporting an important function in cell proliferation and cell cycle control have recently been provided. We found that human THAP1 is an endogenous physiological regulator of endothelial cell proliferation and G_1/S cell cycle progression, which modulates expression of sev-

* This work was supported in part by the French Research Ministry (ACI BCMS), Ligue Nationale Contre le Cancer (Equipe labelisée, to J. P. G.) and by FRM (to D. B.) and ARC (to C. L.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 2jtg) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

Chemical Shift information for the THAP domains (Met[1]–Lys[90]) and (Met[1]–Phe[81]) are available from the BioMagRes Data Bank (http://www.bmrb.wisc.edu), under the accession numbers 15300 and 15289, respectively.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Fig. S1.

[1] Both authors contributed equally to the work.

[2] To whom correspondence may be addressed: University of Toulouse, Inst. of Pharmacology and Structural Biology, IPBS-CNRS-UPS, Toulouse, France. E-mail: jean-philippe.girard@ipbs.fr.

[3] To whom correspondence may be addressed: University of Toulouse, Inst. of Pharmacology and Structural Biology, IPBS-CNRS-UPS, Toulouse, France. E-mail: virginie.gervais@ipbs.fr.

eral pRb[4]/E2F cell cycle target genes. In addition, we identified *RRM1*, a G_1/S-regulated gene required for S-phase DNA synthesis, as a direct transcriptional target of endogenous THAP1 (10). These data provided the first links in mammals between THAP proteins, cell proliferation, and pRB/E2F cell cycle pathways and complemented genetic data previously obtained in model animal organisms. Indeed, in zebra fish and other fish species, the ortholog of cell cycle transcription factor E2F6, a repressor of E2F-dependent transcription during S phase (11) was found to contain a THAP zinc finger at its N terminus (7). In the nematode *Caenorhabditis elegans*, five distinct THAP zinc finger proteins (LIN-36, LIN-15B, LIN-15A, HIM-17, and GON-14) (7) were shown to interact genetically with LIN-35/Rb, the sole *C. elegans* retinoblastoma homolog (12–16). Among these, GON-14 appeared to function as a positive regulator of cell proliferation, because cell division defects were observed in the intestine, gonad, and vulva of *gon-14* null mutant (16). In contrast, LIN-36 and LIN-15B, initially characterized for their role in the specification of vulval cell fates (synthetic Multivulva class B genes, synMuvB) (12, 13), were found to function as inhibitors of the G_1/S cell cycle transition (14). LIN-36 behaved most similar to LIN-35/Rb and Efl-1/E2F, the ortholog of mammalian cell cycle transcription factors E2F4/5, and was therefore proposed to act in a transcriptional repressor complex with these factors to repress G_1/S control genes (14, 17, 18). However, LIN-36, LIN-15B, and THAP1 were not found in the evolutionary conserved pRb/E2F protein complexes (DREAM or DRM complexes) that have recently been described in *Drosophila*, *C. elegans*, and human cells and that contain pRb/p130, E2F4/5, DP, and five other synMuvB gene products LIN-9, LIN-37, LIN-52, LIN-53, LIN-54 (19–22). This suggests that THAP zinc finger proteins may function in distinct transcriptional regulatory complexes to regulate E2F target genes. Although not associated with Rb complexes, THAP zinc finger proteins may still act at the level of chromatin regulation because several *C. elegans* THAP family members have been found to interact genetically with components of diverse chromatin-modifying and/or chromatin-remodeling complexes, including members of the Nucleosome Remodeling Deacetylase (NuRD) complexes and components of the Tip60/NuA4 histone acetyltransferase complex (12–16, 23, 24). In addition, the human THAP7 protein has also been shown to interact with chromatin-modifying enzymes (25). Together, these observations indicate that both in humans and model animal organisms, THAP zinc finger proteins appear to be critical regulators of cell proliferation and cell cycle progression, likely to act at the level of chromatin regulation.

Solution structures of the THAP domains from two previously uncharacterized proteins, human THAP2 and *C. elegans* CtBP, have recently been reported (THAP2, PDB code 2D8R; CtBP, PDB code 2JM3 (26)). However, sequence-specific DNA binding properties have not yet been demonstrated for these two domains. Here, we report the first structure-function analysis of a functional THAP domain, the THAP zinc finger of human THAP1. The three-dimensional structure of the domain was determined by multidimensional NMR spectroscopy and its DNA binding interface was characterized by a combination of alanine scanning mutagenesis and NMR chemical shift perturbation analysis. Together, these data provide a better understanding of the structure-function relationships of this atypical zinc finger.

EXPERIMENTAL PROCEDURES

Plasmid Constructions—The THAP domain of human THAP1 (Met^1–Phe^{81} or Met^1–Lys^{90}; GenBank™ NP_060575) was amplified by PCR and cloned in-frame with a C-terminal His tag into a modified pET-26 plasmid (Novagen). The pET-21c-THAP domain expression vectors for human THAP2 (residues 1–90; GenBank™ NP_113623), human THAP3 (residues 1–92; GenBank™ AAH92427), *C. elegans* CtBP (residues 1–88; GenBank™ NP_508983), and *C. elegans* GON-14 (residues 1–84; GenBank™ NP_741558) were generated by PCR as previously described for the pET-21c-THAP1 (Met^1-Lys^{90}) plasmid (7). Construction of the pcDNA3-THAP1 eukaryotic expression vector has previously been described (7). The THAP1 alanine-scanning single point and triple mutants were obtained by PCR using specific primers containing the corresponding mutations, and cloned as EcoRI-XbaI fragments in the pcDNA3 expression vector.

Protein Expression and Purification—For NMR experiments, recombinant THAP domains of human THAP1 (Met^1-Lys^{90}) and (Met^1-Phe^{81}) were produced as His tag fusion proteins in *Escherichia coli* BL21(DE3). Cells were grown in LB medium at 37 °C to an A_{600} of 0.8 before induction with 1 mM isopropyl-1-thio-β-D-galactopyranoside, to obtain an unlabeled sample. $ZnCl_2$ was added at this step (final concentration of 0.1 mM). Isotopically $^{15}N/^{13}C$-labeled THAP domain (Met^1-Lys^{90}) and ^{15}N-labeled THAP domain (Met^1–Phe^{81})) were expressed in minimal (M9) medium containing $^{15}NH_4Cl$ and ^{15}N Celtone and either [^{13}C]glucose or [^{12}C]glucose. Proteins were purified using a Ni-NTA column (HiTrap, Amersham Biosciences) followed by gel-filtration chromatography on Sephadex G75 (Amersham Biosciences). After digestion with thrombin (Novagen), proteins were further purified on a gel filtration column. NMR samples were concentrated to 0.4–1.7 mM in 50 mM deuterated Tris-HCl, pH 6.8, 1 mM DTT with either 10 mM NaCl (for protein structure determination) or 250 mM NaCl (for DNA binding studies).

For gel-shift assays, the recombinant THAP domain of human THAP1 was produced as previously described (7). The recombinant THAP domains of THAP2, THAP3, Ce-CtBP, or GON-14 were produced in *E. coli* strain BL21*trxB*(DE3), transformed with the different pET-21c-THAP domain expression vectors. Protein expression and purification was performed according to the manufacturer's instructions (Novagen, Madison, WI), as previously described (7). The purity of the different THAP domains was assessed by SDS-PAGE, and protein concentrations were determined using the Bradford protein assay (Bio-Rad). Full-length THAP1 wild-type and mutants were synthesized *in vitro* in rabbit reticulocyte lysate (RRL). The corresponding pcDNA3 expression vectors were used with the

[4] The abbreviations used are: pRb, retinoblastoma protein; DTT, dithiothreitol; RRL, rabbit reticulocyte lysate; SPR, surface plasmon resonance; EMSA, electrophoretic mobility shift assay; wt, wild type; r.m.s.d., root mean square deviation.

TNT-T7 kit (Promega). Protein production was performed in the presence of ^{35}S-labeled methionine and verified by SDS-PAGE and autoradiography.

Surface Plasmon Resonance Experiments—DNA interaction kinetics was investigated by Surface Plasmon Resonance (SPR) assays using a four channel BIAcore 3000 optical biosensor instrument (BIAcore AB, Uppsala, Sweden). Immobilization of biotinylated single-stranded DNA probes was performed on a streptavidin-coated sensorchip (BIAcore SA sensorchip) in HBS-EP buffer (10 mM Hepes pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.0005% Surfactant P20) (BIAcore AB, Uppsala, Sweden). All immobilization steps of biotinylated single-stranded DNA probes were performed at a final DNA concentration of 100 ng/ml and at a flow rate of 2 μl/min. Hybridization of complementary DNA strands was performed in HBS-EP buffer supplemented with 200 mM NaCl. Biotinylated oligonucleotide sequences and complementary DNA strands were purchased from MWG Biotech.

Binding analyses were performed with multiple injections of THAP domain (Met^1–Lys^{90}) at different protein concentrations over the immobilized surfaces at 25 °C. A second DNA probe with unrelated sequence was used as a control. All samples were diluted in the running buffer containing 50 mM Tris, pH 6.8, 250 mM NaCl, 1 mM DTT, and were injected over the sensor surface for 4 min at a flow rate of 20 μl/min. No binding was observed to the DNA probe with unrelated sequence. The SPR signal was therefore analyzed as difference sensorgrams between the two DNA sequences immobilized to separate channels of the sensor chip (the signal from the unrelated DNA used as control was subtracted).

NMR Spectroscopy—NMR experiments were recorded at 296 K on Bruker Avance 800, 700, and 600 MHz spectrometers. Backbone and side-chain resonances (^{1}H, ^{15}N, and ^{13}C) of the THAP domain (Met^1-Lys^{90}) were assigned using a set of heteronuclear experiments (27). This information was partially used to assign ^{1}H and ^{15}N resonances of the THAP domain (Met^1–Phe^{81}) using a combination of homonuclear and ^{15}N heteronuclear spectra. ϕ/ψ torsion angles were derived using TALOS (28) and the $^3J_{\mathrm{HN\text{-}H}\alpha}$ coupling constants obtained from three-dimensional HNHA (29). Four X-Pro bonds were identified as being in the trans-configuration on the basis of strong NOEs between the Hδ proton of each Pro residue and the Hα protons of the preceding residues (30) and confirmed by ^{13}C chemical shifts (31). From characteristic NOEs, the residue Pro^{26} was identified as being in the cis form (30). 15 stereospecific assignments of Hβ methylene were obtained using the DQF-COSY spectrum (32). NMR data were processed using TOPSPIN software (Bruker) and NMRPipe (33) and analyzed using XEASY (34) and NMRView (35).

^{15}N relaxation data were recorded at 296 K on a 0.9-mM protein sample at 10 mM and 250 mM NaCl using standard pulse sequences. The heteronuclear NOEs were determined from two ^{15}N HSQC spectra recorded in the presence and absence of ^{1}H presaturation period of 3 s and with a recycling delay of 5 s (36).

Structure Calculations—To solve the structure of the THAP domain (Met^1–Phe^{81}), a set of distances was extracted from integration of two-dimensional ^{1}H and three-dimensional ^{15}N heteronuclear NOESY spectra. The secondary structure elements were derived from analysis of coupling constants, from identification of slowly exchanging amide protons and from characteristic NOEs. To maintain well-defined secondary structure elements, hydrogen bonds were added with restraints of 1.8 to 2.4 imposed on the distance between hydrogen and acceptor oxygen and restraints of 2.3 to 3.2 imposed on the distance between the donor nitrogen and acceptor oxygen.

Preliminary structure calculations run either with Nδ1 or Nϵ2 of the His^{57} ring allowed us to identify Nϵ2 as the zinc-bound atom. Subsequent structural refinement including a zinc ion together with constraints defining tetrahedral coordination (3, 37) was performed. The structures were calculated using a torsion angle dynamics simulated annealing protocol using the CNS software suite (38). From 500 initial structures, a set of 20 structures were selected as accepted structures, based on the following criterions: low total energy, no distance violation larger than 0.2 Å and no torsion angle violation greater than 2°. Their structural quality was analyzed using PROCHECK (39).

NMR Chemical Shift Perturbation Analysis—The 14-bp duplex DNA containing THABS was reconstituted by hybridizing the following oligonucleotides, 5′-CAAGTATGGGCAAG-3′ and 5′-CTTGCCCATACTTG-3′ in a 1:1 ratio. For NMR titration, two-dimensional ^{15}N HSQC spectra of the THAP domain at the concentration of 0.4 mM were collected at 296 K and 250 mM NaCl after each incremental addition of lyophilized DNA. One-dimensional ^{1}H spectra were recorded after each DNA addition and the DNA/protein ratio was followed from integration of ^{1}H protein and DNA signals on the ^{1}H spectra. A DNA fragment with an unrelated sequence was reconstituted by hybridizing the oligonucleotides, 5′-GATTTGCATTTTAA-3′ and 5′-TTAAAATGCAAATC-3′, and added to the THAP domain following the same procedure as described above. Normalized chemical shift changes were calculated as: $\Delta\delta = [(\Delta\delta_{\mathrm{HN}})^2 + (\Delta\delta_{\mathrm{N}} \times 0.154)^2]^{1/2}$.

Electrophoretic Mobility Shift Assays (EMSA)—EMSA were performed with purified recombinant THAP domains produced in *E. coli* or with full-length THAP1 wild type or THAP1 mutants synthesized *in vitro* in RRL, using the following THABS probes, 25-bp (5′-AGCAAGTAAGGGCAAACTACTTCAT-3′) and 36-bp (5′-TATCAACTGTGGGCAAACTACGGGGCAACAGGTAATG-3′), as previously described (7). Increasing amounts of purified recombinant THAP domains were incubated for 20 min at room temperature in 20 μl of binding buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 0.1% Nonidet P-40, 100 μg/ml bovine serum albumin, 2.5 mM DTT, 5% glycerol, 10 μg/ml poly(dI/dC)). For *in vitro* translated proteins, 3 μl of RRLs expressing full-length THAP proteins or THAP1 mutants were incubated in 20 μl of binding buffer supplemented with 50 μg/ml of poly(dI/dC) and 50 μg/ml salmon sperm DNA. Electrophoresis was performed and gels were exposed as previously described (7). Supershift experiments were performed using 1 μg of anti-THAP1 affinity-purified rabbit polyclonal antibodies (10).

Model Building of the Complex between the THAP Zinc Finger and the THABS Sequence—Computational docking of the THABS DNA target onto the THAP zinc finger (Met^1–Phe^{81}) of THAP1 was performed using HADDOCK1.3 (High Ambi-

guity Driven DOCKing) (40), in conjunction with CNS (38). The docking was performed using an ensemble of THAP zinc finger structures of THAP1 and the 14-bp double-stranded DNA containing THABS built as a B-DNA template using Insight II (Accelrys). The active residues used to define the Ambiguous Interaction restraints (AIR) included residues showing relative solvent accessibility higher than 40% (as calculated by NACESS, Hubbard and Thornton, University College London) and either displaying a chemical shift perturbation higher than 0.2 ppm upon DNA binding or giving rise to loss of DNA binding from site-directed mutagenesis. Briefly, active residues were Lys^{24}, Glu^{37}, Arg^{42}, Lys^{46}, and Thr^{48}. Solvent accessible residues that were surface neighbors of the active residues were defined as passive residues including Lys^{34}, Glu^{35}, Ala^{38}, Arg^{41}, Lys^{43}, Asn^{44}, Lys^{49}, and Tyr^{50}.

For DNA, bases corresponding to the core GGCA motif of the THABS sequence together with the thymine upstream were defined as active residues, as based on already reported data obtained from scanning mutagenesis (7).

The docking was performed on a SGI cluster equipped with 40 processors. Starting from 15 THAP domain structures of THAP1 and a model of the THABS target in B-DNA conformation, 1000 rigid-body solutions were generated. The best 200 solutions according to the HADDOCK rigid-body score were selected for semi-flexible refinement in torsion angle space; the top 200 structures were finally refined in explicit water. The final ensemble of 200 solutions was analyzed and clustered based on a pair-wise r.m.s.d. matrix calculated over the backbone atoms.

RESULTS

Biophysical Characterization of the THAP Domain of THAP1 and Identification of a Shorter Functional Fragment—The THAP domain was originally assigned to the first 90 N-terminal residues of THAP1 (6). A corresponding fragment was initially expressed and purified for Surface Plasmon Resonance (SPR) and NMR studies. Inductively Coupled Plasma-Mass Spectrometry (ICP-MS) experiments indicated that the domain includes a zinc ion (data not shown). The SPR experiments (Fig. 1, *A* and *B*) showed that the THAP zinc finger of THAP1 binds to a 14-bp THAP domain binding site (THABS DNA probe) that includes the GGCA core motif at sequence positions 7–10, previously found to be critical for recognition by the THAP domain of THAP1 (7).

A series of triple-resonance NMR experiments allowed us to assign unambiguously residues 3–63 and 74–77 of the domain. However, only partial assignment for residues 64–73 and 78–90 in the C-terminal tail could be performed due to severe lack of connectivity for these residues. These data together with characteristic negative heteronuclear NOE values (data not shown) indicated that the C-terminal tail of the domain is unstructured. These results led us to search for the minimal size of the functional THAP zinc finger. The cysteine residues at the N terminus of the THAP domain (THAP1 Cys^{5} and Cys^{10}) have previously been shown to be required for the functional activity of the domain, for both human THAP1 (7) and *Drosophila* P element transposase (41), thus defining the N-terminal boundary of the domain. In contrast, nothing was known about the

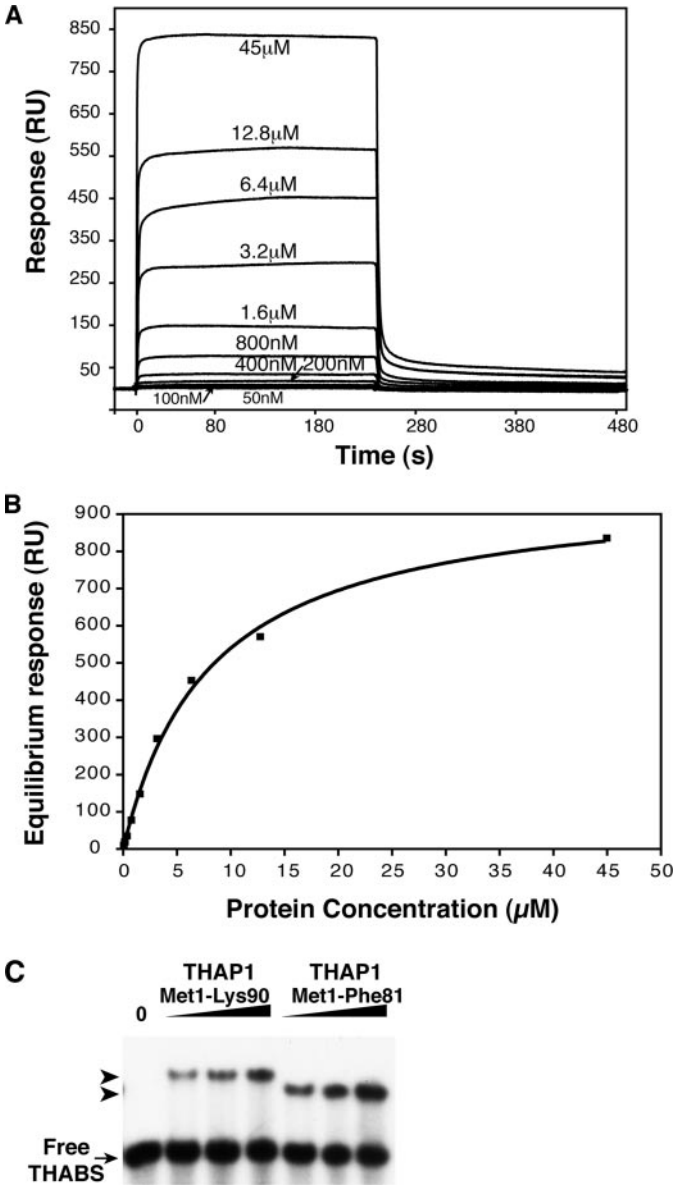


FIGURE 1. **Specific THABS-DNA binding of the THAP domain of THAP1 as observed by SPR and EMSA.** *A*, SPR difference sensorgrams with increasing concentrations of protein from 50 nM to 45 μM as indicated on the sensorgram lines (responses with the unrelated DNA sequence were subtracted). *B*, K_D determination. Data points represent the equilibrium responses values as function of the protein concentration for each of the experiments shown in *A*. The global binding constant obtained by fitting the SPR data to a 1:1 stoichiometry-binding model was found to be 8 μM. The *solid line* represents the fitting theoretical curve calculated for the 1:1 binding model. *C*, first 81 residues of the THAP zinc finger of THAP1 are sufficient for sequence-specific DNA binding activity. EMSA experiments were performed using a consensus 25-bp THABS probe and increasing amounts (5, 10, and 100 nM) of recombinant THAP zinc finger Met^{1}–Lys^{90} or Met^{1}–Phe^{81} of human THAP1. *Black arrows*, THAP1-THABS DNA complex.

requirements at the C terminus downstream of the conserved AVPTIF motif (6) containing the essential Pro^{78} residue (7). Alanine-scanning mutagenesis was therefore performed and revealed that residues 82–90 are not required for DNA binding activity of the THAP domain of THAP1 (data not shown). This was confirmed using a THAP1 deletion mutant, THAP1Δ82–90, that exhibited a similar activity in EMSA experiments than wild-type THAP1 (data not shown). In addition, recombinant

TABLE 1
Structural statistics of the THAP-zinc finger (Met[1]-Phe[81]) of THAP1

Restraints for calculation	
Intraresidue	801
Sequential	329
Medium-range (2≤|i-j|≤4)	103
Long range (|i-j|>4)	306
Dihedral angle restraints (TALOS)	104
angle ϕ	52
angle ψ	52
Hydrogen bond restraints	10
Zinc coordination	14
Characteristics	
R.m.s deviation from constraints	
NOE restraints (Å)	0.0348 ± 0.0004
Dihedral angle restraints (°)	0.27 ± 0.02
R.m.s deviation from idealized geometry (±SD)	
Bond length (Å)	0.0048 ± 0.0005
Bond angle (°)	0.60 ± 0.05
Improper angle (°)	0.39 ± 0.07
Final energies (kcal mol^{-1} ± SD)	
Overall	477.6 ± 6.7
van der Waals	96.44 ± 3.3
Bonds	33.3 ± 0.7
Angles	144.2 ± 2.5
NOE	185.33 ± 4.2
Dihedrals	0.61 ± 0.1
Coordinate precision[a]	
R.m.s deviation of backbone atoms (Å)	0.530
R.m.s deviation of all heavy atoms (Å)	1.297
Ramachandran plot[b]	
Residues in most favored region (%)	64.0
Residues in additional allowed regions (%)	32.2
Residues in generously allowed regions (%)	3.8
Residues in disallowed regions (%)	0.0

[a] Average r.m.s deviation from the mean structure (residues 4–16, 21–40, 46–63, 69–81).
[b] The ϕ and ψ dihedral angles were analyzed using the PROCHECK program (39).

THAP domains (Met[1]–Phe[81]) and (Met[1]–Lys[90]) exhibited similar activities in EMSA experiments indicating that residues 1–81 are sufficient for sequence-specific DNA binding to the THABS motif (Fig. 1*C*). A deletion mutant referred to fragment (Met[1]–Phe[63]) was also tested both by NMR and EMSA experiments, but the HSQC spectrum corresponded to that of an unfolded protein and the fragment did not possess any DNA binding activity (data not shown). Therefore, the recombinant THAP domain (Met[1]–Phe[81]) was selected for all structural studies.

NMR Solution Structure of the THAP Zinc Finger of Human THAP1—We solved the three-dimensional solution structure of the zinc-containing form of the THAP domain (Met[1]-Phe[81]) using a set of distances extracted from two-dimensional and three-dimensional NOESY spectra. For the THAP domain in its DNA-free state, spectra were recorded in a buffer containing 10 mM NaCl. In these conditions, 1539 distance restraints obtained from two-dimensional ^{1}H-NOESY and three-dimensional ^{15}N HSQC-NOESY spectra and 104 angle restraints were used for calculations (Table 1).

The core of the THAP zinc finger of THAP1 adopts a $\beta\alpha\beta$ fold (Fig. 2, *A–D*). Residues Cys[5], Cys[10], Cys[54], His[57] form a single zinc-binding site that begins with a long loop L1 (residues Gln[3]–Ser[21]), which precedes the first β-strand (β1; residues Phe[22]–Lys[24]). This portion is followed by a second loop L2 (Phe[25]–Lys[32]), which continues into a α-helix H1 (residues Cys[33]–Val[40]). An additional loop L3 (residues Arg[41]–Ser[51]) is followed by the second β-strand (β2; residues Ser[52]–Cys[54]) running anti-parallel to β1. The zinc-binding site is completed by a short 3_{10} helix (H2; residues Ser[55]–His[57]). A second 3_{10} helix (H3; residues Pro[60]–Phe[63]) is followed by a flexible loop L4 (Lys[64]–Asn[68]) that continues into an extended region (Asn[69]–Pro[78]) followed by an additional 3_{10} helix (H4) including residues Thr[79]-Phe[81] that form part of the AVPTIF motif (6). Some regions of the loops are relatively well defined (L1: 6–16; L2: 26–32; L3: 46–50; L4: 69–78) with restricted mobility as judged by the heteronuclear NOEs, whereas the other parts of the loops are less ordered and display a mobility on ns-ps time scale (Fig. 2*C* and supplemental Fig. S1). In particular, beginning of loop L4 displays a high degree of structural disorder although its heteronuclear NOEs are not much lower than those of loop L1. This is because of the scarcity of NOEs involving this region despite extensive search in the NOESY maps.

Only a few amide protons are protected as observed from H_2O/D_2O experiments. They mainly correspond to some residues located within the helix H1 (Glu[35], Trp[36], Glu[37], Ala[39], Val[40]) and the two-stranded β-sheet (Ile[53], Cys[54]) as well as few additional residues (Ser[6], Ala[7], Cys[10], Asn[12]) in the vicinity of the zinc-binding site. Apart from these residues, the THAP domain amide protons exchange rapidly with the solvent (data not shown).

A structure-based sequence alignment of the THAP zinc finger of THAP1 with representative THAP domains is shown in Fig. 2*E*. Besides the strictly conserved C2CH motif that provides ligands for the zinc ion, the THAP domain is defined by a C-terminal AVPTIF motif and four residues (Pro[26], Trp[36], Phe[58], and Pro[78], numbering refers to THAP1) that are invariant in more than hundred THAP domain sequences and that are absolutely required for DNA binding activity (6, 7). The unique tryptophan (Trp[36]) located in the α-helix H1 is a key element of the THAP zinc finger structure and constitutes the anchoring residue that makes hydrophobic contacts with the conserved Phe[58] residue (Fig. 2*D*) and the surrounding aromatic residues, namely Phe[25] and Phe[63]. In addition, NOEs are detected between Trp[36] at the center of the hydrophobic core and the two invariant prolines (Pro[26] in the loop L2 and Pro[78] in the AVPTIF motif). Both of the prolines display strongly upfield-shifted resonances because of the proximity of Trp[36]. NOEs are also observed between Phe[81] in the AVPTIF motif and Ala[39]-Val[40] in the helix H1 (data not shown). Therefore, the AVPTIF motif appears to play an essential role in the folding of the THAP zinc finger by bringing together the C terminus and the α-helix H1 (Fig. 2*D*).

The THAP Zinc Fingers Share the Same Three-dimensional Fold but Not the Same DNA Target Sequence—Comparison of the structure of the THAP zinc fingers from human THAP1, THAP2, and *C. elegans* CtBP revealed that the overall fold and the packing around the tetrahedral zinc-coordinating site are similar for the three THAP domains (Fig. 3*A*). The structural homology is higher between the THAP domains of THAP1 and THAP2, as expected for closely related sequences (48%). Indeed, the solution structure of the THAP domain of THAP1 can be superimposed onto that of THAP2 for 80 Cα equivalent residues with an r.m.s.d. value of 2.8 Å. A weaker score is found for the superimposition of the THAP domain of THAP1 onto that of CtBP with 66 Cα equivalent atoms that could be super-

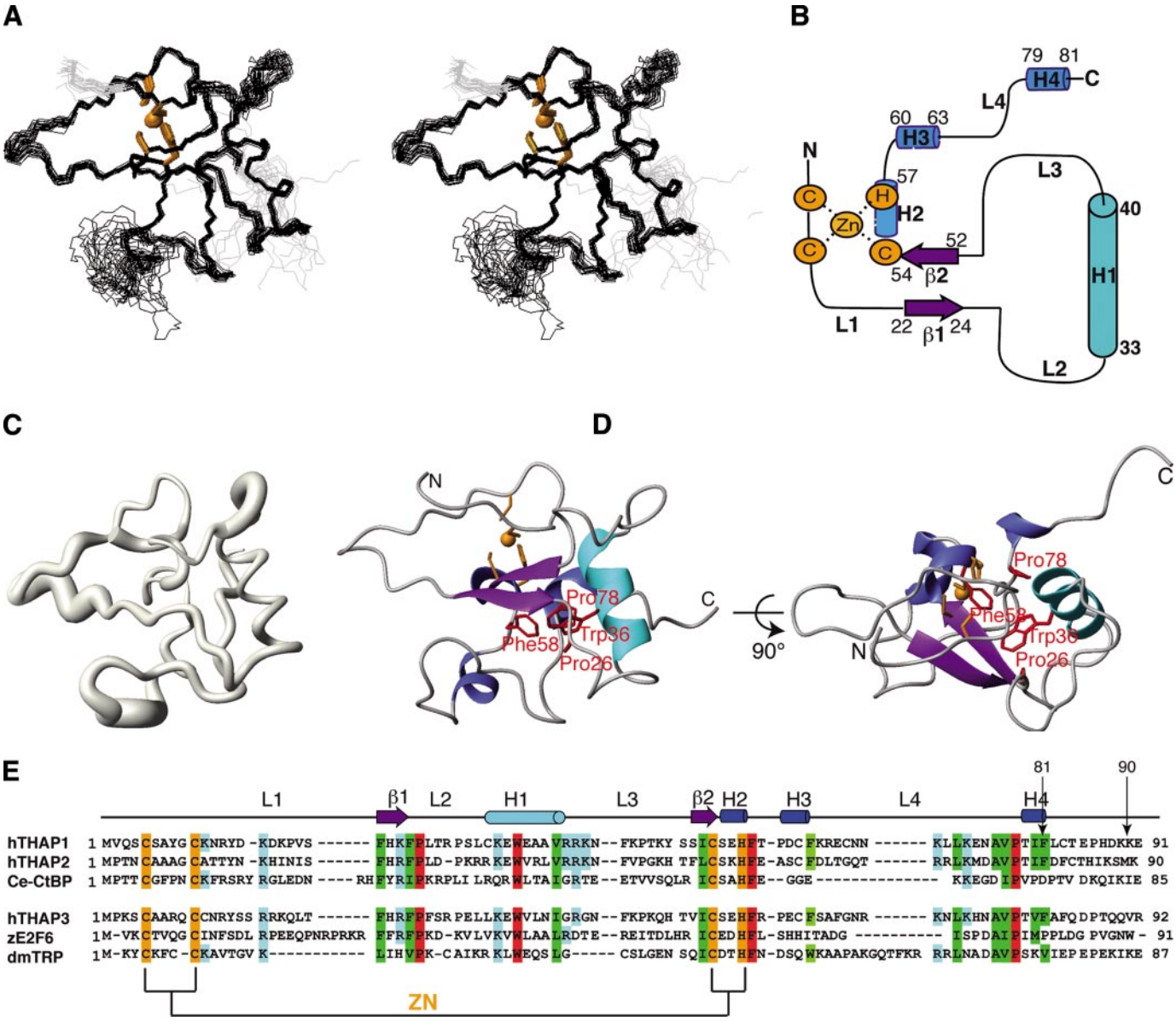


FIGURE 2. **Solution structure of the THAP zinc finger of human THAP1.** *A*, *backbone traces* of the NMR ensemble for the 20 lowest energy structures in stereo view. The zinc and the four ligands are shown in *orange*. *B*, topology diagram showing the secondary structure elements. The α-helix, the 3_{10} helices and the β-sheet are depicted in *cyan*, *blue*, and *purple*, respectively. The zinc and the four ligands are shown in *orange*. *C*, spine representation of the lowest energy structure (residues 4–81) with variable radius, the radius representing the mobility as judged by heteronuclear NOEs (supplemental Fig. S1). The orientation is the same as in *A*. *D*, *ribbon diagram* of the THAP zinc finger of THAP1 showing side chains of the four invariant residues in *red*. The left side is in the same orientation as in *A*. The right side is rotated 90° around horizontal axis. *E*, structure-based sequence alignment of THAP zinc fingers from human (hTHAP1–3), zebrafish (zE2F6), *Drosophila melanogaster* (*dmTRP*) and *C. elegans* (*Ce-CtBP*) members of the THAP family. The secondary structure elements are represented with the same color as in *1B*. The zinc ion together with the four zinc ligands is depicted in *orange*. Highly conserved hydrophobic residues are shown in *green*. The four invariant residues (Pro^{26}, Trp^{36}, Phe^{58}, and Pro^{78}) are shown in *red*. The conserved basic residues are shown in *blue*.

imposed with an r.m.s.d. value of 3.1 Å. It is noteworthy that the sequence identity between these two domains is only of 27%. The core fold consisting of the anti-parallel β-sheet with the two strands separated by a loop-helix-loop insertion is conserved among the three THAP domains. Nevertheless, the C terminus displays structural variability. Indeed, the THAP domain of THAP1 shows two additional short 3_{10} helices H3 and H4, encompassing residues 60–63 and 79–81, respectively, whereas the THAP domains of THAP2 and CtBP display additional two-stranded anti-parallel β-sheets. In the structure of the THAP domain of THAP2, the second anti-parallel β-sheet is formed by residues that would correspond to residues Phe^{63}-Lys^{64} and Leu^{71}-Leu^{72} in THAP1. In the CtBP structure instead, the second anti-parallel β-sheet involves residues that would correspond to THAP1 residues Ala^{76}-Val^{77} in the AVPTIF motif and the two residues Leu^{82}-Cys^{83} that follow the AVPTIF motif. Because the Met^{1}–Phe^{81} fragment of THAP1 retains its capacity to bind DNA (Fig. 1*C*), the second β-sheet observed for the THAP domain of CtBP is unlikely to be important in the molecular scaffold of the THAP zinc finger.

It is noteworthy that the flexible loop L4 (residues 64–68) between H3 and H4 in THAP1 is also observed in THAP2 but is absent in CtBP because it corresponds to an 8-residue sequence insertion compared with the THAP domain of CtBP. Despite

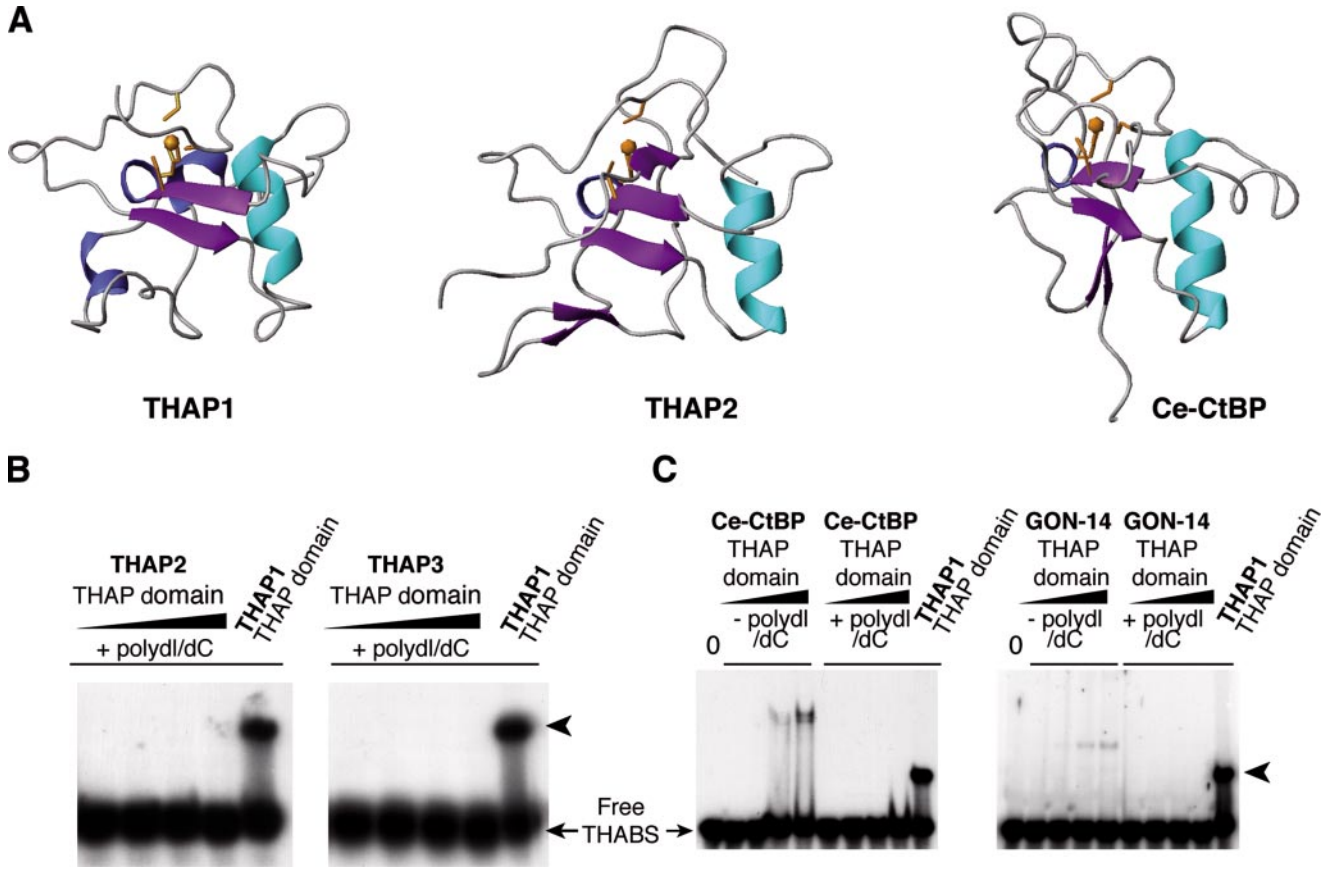


FIGURE 3. **The THAP zinc fingers share a similar fold but do not recognize the same DNA target sequence.** *A*, *ribbon diagrams* of the THAP domains of THAP1, THAP2, and CtBP. *B*, recombinant THAP zinc fingers from human THAP2 and THAP3 do not bind to the THABS. EMSA experiments were performed by incubating the 25-bp THABS probe in the presence of poly(dI/dC) with increased amounts (0.01, 0.1, 1, and 5 μM) of recombinant THAP domains from THAP2 and THAP3 produced in *E. coli*. As control, EMSA was performed with THABS probe and the recombinant THAP zinc finger of THAP1 (0.1 μM) in the presence of poly(dI/dC). *C*, THAP zinc fingers of Ce-CtBP and GON-14 bind DNA but do not interact specifically with the THABS motif recognized by human THAP1. Increasing amounts (0.1, 1, 5, and 10 μM) of recombinant THAP zinc fingers of Ce-CtBP or GON-14 were incubated with the THABS probe in the presence or absence of the nonspecific competitor poly(dI/dC) and analyzed by EMSA. As a control, EMSA was performed with the recombinant THAP zinc finger of THAP1 (0.1 μM) in the presence of poly(dI/dC).

these discrepancies in the secondary structure elements, residues in the C-terminal region keep mostly equivalent positions in the three structures and positions of residues Ala^{76}-Phe^{81} that form part of the AVPTIF motif in THAP1 are identical when compared with that of THAP2.

Because different THAP zinc fingers appear to exhibit a similar three-dimensional fold, we then studied the possibility they may recognize the same DNA target sequence. Among the 12 human THAP proteins, THAP2 and THAP3 are the most closely related to THAP1 and for instance, the THAP domains of THAP1 and THAP3 exhibit up to 50% identity (6). Therefore, we tested the ability of these two proteins to bind to the THABS motif specifically recognized by THAP1 (7). As shown in Fig. 3*B*, we found that the recombinant THAP domains of THAP2 and THAP3 did not bind to the THABS probe in gel shift assays. In contrast, the recombinant THAP domain of THAP1, used in the same conditions, exhibited strong binding to the THABS sequence (Fig. 3*B*). These results were confirmed using *in vitro* translated full-length THAP proteins, which provided an independent source of THAP domains. In contrast to THAP1, the full-length THAP2 and THAP3 proteins did not bind to the THABS probe in gel-shift assays (data not shown).

Liew *et al.* (26) recently reported binding of the THAP domain of *C. elegans* CtBP (Ce-CtBP) to the THABS sequence recognized by human THAP1. However, their gel-shift assays were performed in the absence of competitor DNA, and we considered the possibility that their observations may correspond to a non-sequence specific DNA binding activity of the THAP domain of Ce-CtBP. We therefore performed gel shift assays with the recombinant Ce-CtBP THAP domain in the presence or absence of the synthetic poly(dI/dC) nonspecific competitor DNA. In agreement with the results reported by Liew *et al.* (26), we observed binding of the THAP domain of Ce-CtBP to the THABS motif in the absence of competitor (Fig. 3*C*). However, no specific protein-DNA complex was observed in the presence of the poly(dI/dC) competitor. Similar results were obtained with the recombinant THAP domain from another *C. elegans* THAP protein, the cell proliferation and developmental regular GON-14 (Fig. 3*C*). In contrast, strong binding of the THAP domain of THAP1 to the THABS sequence was observed in the presence of competitor (Fig. 3*C*). We concluded that Ce-CtBP and GON-14 THAP zinc fingers do not bind specifically to the THABS motif recognized by THAP1. Together with the findings on human THAP2 and

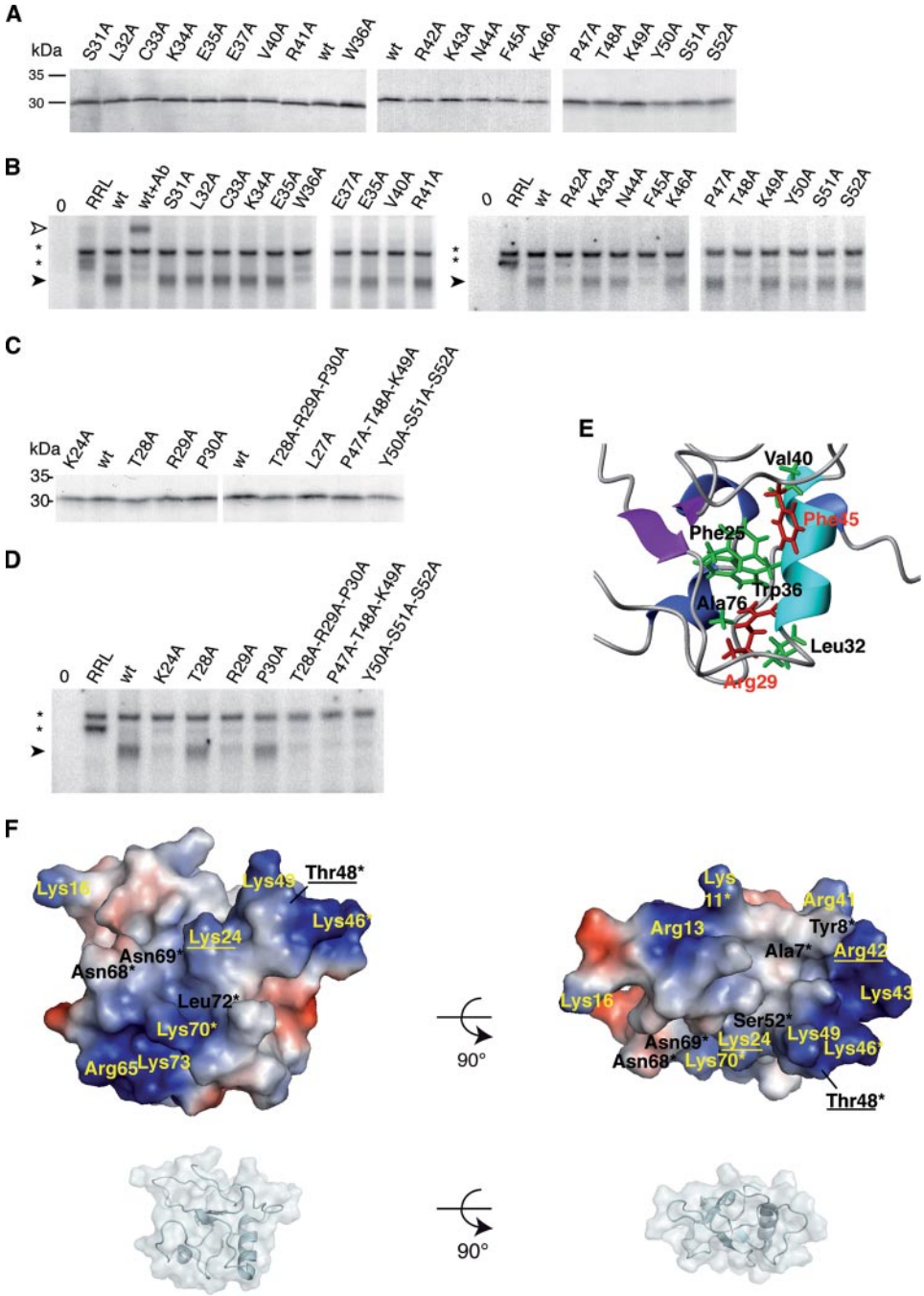


FIGURE 4. **Site-directed mutagenesis of the THAP zinc finger of THAP1 reveals critical residues for DNA binding activity.** *A–D*, alanine-scanning mutagenesis. *A* and *C*, wild-type THAP1 (*wt*) and single point or triple mutants in the THAP zinc finger were *in vitro* translated in rabbit reticulocyte lysate in the presence of ^{35}S-labeled methionine and analyzed by SDS-PAGE and autoradiography. Molecular weight markers are shown on the *left* (kDa). *B* and *D*, single point mutation of residues Lys^{24}, Arg^{29}, Arg^{42}, Phe^{45}, or Thr^{48} abrogates DNA binding activity of the THAP zinc finger of THAP1. EMSA were performed with the 36-bp THABS probe and THAP1 wt or the indicated mutant proteins, in the presence of poly(dI/dC) and salmon sperm DNA competitors. The previously described Trp36A mutant (7) was included as a control for loss of DNA binding activity. *Wt + Ab*, supershift experiment with anti-THAP1 antibody to demonstrate the specificity of the protein-DNA complexes. RRL, unprogrammed rabbit reticulocyte lysate; *black arrowhead*, THAP1-THABS DNA complex; *white arrowhead*, antibody-THAP1-THABS DNA complex; *asterisks*, nonspecific complexes. *E*, view illustrating the interactions of Arg^{29} and Phe^{45} (in *red*) with Leu^{32}, Trp^{36}, Ala^{76}, and Phe^{25}, Val^{40}, respectively (in *green*). *F*, representations of the electrostatic surface potential of the THAP domain (Met^{1}–Phe^{81}) showing the exposed residues that are found to be critical for DNA binding from site-directed mutagenesis experiments (*underlined*) or that undergo more than 0.2 ppm chemical shift change (marked with an *asterisk*). Exposed residues that are positively charged are indicated on the surface and colored *yellow*, otherwise *black*. Representations of the two corresponding ribbons are shown for clarity (in *gray*).

THAP3, these results indicate that the different THAP zinc fingers share the same three-dimensional fold but not the same DNA target sequence.

Structure-Function Analysis of the THAP Zinc Finger of THAP1 by Site-directed Mutagenesis—We have previously shown that the eight invariant residues that define the THAP domain are absolutely required for DNA binding activity (7). To get further insights into the role of other residues in DNA recognition, thirty additional residues were individually mutated to alanine and the resulting mutants were tested in gel-shift assays (Fig. 4, *A–D* and Table 2). These included twenty-four consecutive residues (Leu^{27} to Ser^{52}) from the long loop-helix-loop motif (L2-H1-L3) inserted into the anti-parallel β-sheet, one of the most distinctive features of the THAP zinc finger (Fig. 2). Single-point mutation of the invariant Trp^{36} in the center of the α-helix was used as a control and, as expected, this mutation completely abolished the interaction. Similarly to mutation of Trp^{36}, mutation of residues Lys^{24}, Arg^{29}, Arg^{42}, Phe^{45}, and Thr^{48} led to a complete loss of DNA-binding activity whereas mutation of residues Lys^{11}, Leu^{27}, Glu^{37}, Val^{40}, and Tyr^{50} decreased but did not abrogate the interaction of the THAP zinc finger with its THABS DNA target sequence (Fig. 4, *A–D* and Table 2). Triple mutations of residues Thr^{28}-Arg^{29}-Pro^{30}, Arg^{41}-Arg^{42}-Lys^{43} and Pro^{47}-Thr^{48}-Lys^{49} to alanines were also performed and confirmed the importance of these regions for DNA binding activity of the THAP domain of THAP1 (Fig. 4, *C* and *D* and Table 2). Interestingly, triple alanine mutation of residues Tyr^{50}, Ser^{51}, and Ser^{52} revealed a critical role for these residues that was less apparent in the single point mutants. Mapping of the essential residues on the THAP domain structure of THAP1 revealed that residues Arg^{29} and Phe^{45} make contacts with the hydrophobic core of the domain. Indeed, the Arg^{29} resi-

due in the loop L2 is shown to make several NOE contacts with protons of residues Leu[32] in the loop L2, Trp[36] in the helix H1 and Ala[76] that is part of the AVPTIF motif (Fig. 4*E*). The residue Phe[45] that is part of the loop L3 gives NOEs to Val[40] in the helix H1 and to Phe[25] in the loop L2 (Fig. 4*E*). Therefore the loss in DNA binding after mutation of the two residues Arg[29] and Phe[45] could be due to a disruption of local structure.

In contrast, residues Lys[24], Arg[42], and Thr[48] are exposed at the surface of the THAP domain. Interestingly, they map onto the area of the domain that is highly positively charged due to the presence of several exposed basic side chains of lysines and arginines consistent with DNA interaction (Fig. 4*F*). These data strongly suggest that the three residues Lys[24], Arg[42], and Thr[48] may be directly involved in DNA binding.

TABLE 2

Alanine-scanning mutagenesis of the THAP zinc finger of human THAP1

DNA binding activity of the different mutants was tested using gel-shift assays. The eight evolutionary conserved residues, which define the THAP zinc finger and have previously been shown to be essential for DNA binding (7) are indicated in bold.

DNA binding to THABS probe	THAP1 mutants
Severely affected	**C5A, C10A, C54A, H57A, P26A, W36A, F58A, P78A**, K24A, R29A, R42A, F45A, T48A Triple mutants: T28A/R29A/P30A R41A/R42A/K43A P47A/T48A/K49A Y50A/S51A/S52A
Partially affected	K11A, L27A, E37A, V40A, Y50A
Not affected[a]	S4A, S6A, Y8A, K16A, T28A, P30A, S31A, L32A, C33A, K34A, E35A, R41A, K43A, N44A, K46A, P47A, K49A, S51A, S52A, S55A

[a] Within the limits of detection of the present assay; this qualitative classification does not imply that the corresponding residues play no role at all in the binding affinity and selectivity.

Identification of the DNA Binding Interface of the THAP Zinc Finger by NMR Chemical Shift Perturbation Analysis—To further characterize the DNA binding interface, binding of the THAP domain of THAP1 to the THABS DNA target sequence was probed by NMR chemical shift perturbation analysis. The ^{15}N-labeled THAP domain dissolved in NMR buffer containing 250 mM NaCl was titrated with 14-bp duplex DNA containing THABS. In the absence of DNA, the spectrum recorded at 250 mM NaCl was similar to the one recorded at 10 mM NaCl, except a few peaks that slightly shifted (data not shown). In addition, based on ^{15}N longitudinal and transverse relaxation times, rotational correlation times (τ_c) were determined to be 6.03 ± 0.1 ns and 6.89 ± 0.4 ns at 10 mM and 250 mM NaCl, respectively, indicating that the protein is monomeric in both salt conditions (data not shown). In the presence of increasing concentrations of the THABS oligonucleotide, several cross-peaks were significantly affected during titration (Fig. 5*A*). A similar two-dimensional ^{15}N HSQC spectrum recorded in the presence of an unrelated 14-bp duplex DNA did not reveal any significant shift (data not shown). The chemical shift perturbations observed in the presence of the specific THABS sequence were not further affected when DNA:protein ratio was above 1:1 suggesting a 1:1 binding stoichiometry, in agreement with SPR experiments (Fig. 1, *A* and *B*).

During titration, the majority of the affected signals could be followed as the fast-exchange manner, *i.e.* a single cross peak with intermediate chemical shift between that of the free and bound forms (Fig. 5*A*). Signals with the largest chemical shift changes could be followed as the slow-exchange manner with two peaks corresponding to the free and bound forms with intensities proportional to the free/bound ratio. Finally, the

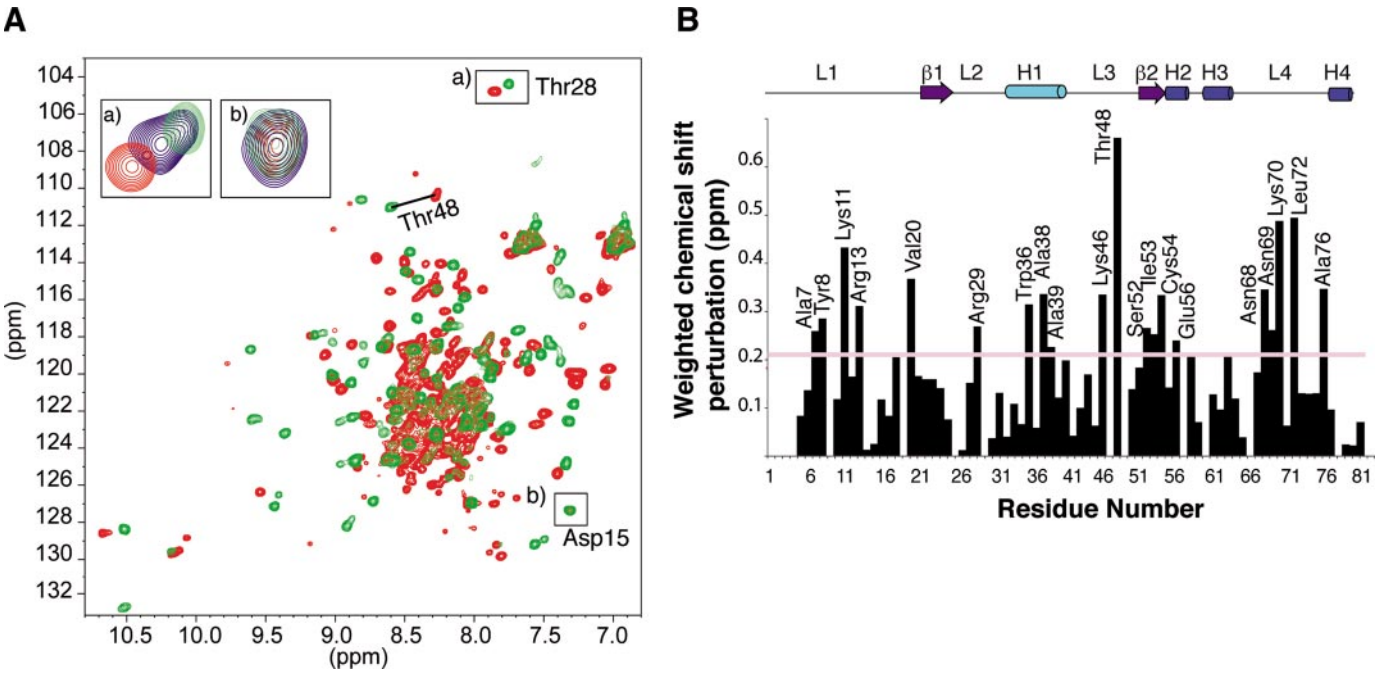


FIGURE 5. **Chemical shift perturbation analysis of the THAP zinc finger of THAP1 upon DNA binding.** *A*, overlay of selected regions of ^{1}H-^{15}N HSQC spectra of the THAP domain in the absence (*green*) and in the presence (*red contour levels*) of an equimolar concentration of the 14-mer oligonucleotide (THABS). Two examples of residues affected (Thr[28]) and not (Asp[15]) are represented in *A* and *B*, respectively with additional *blue* contour levels for half an equimolar concentration of DNA (*blue contour levels*). Residue Thr[48] that displays the largest chemical shift change is also indicated. *B*, histogram of chemical shift changes upon DNA binding as a function of the THAP zinc finger residue number. Reported chemical shifts $\Delta\delta$ represent combined ^{15}N and ^{1}H chemical shifts ($\Delta\delta = [(\Delta\delta_{HN})^2 + (\Delta\delta_N \times 0.154)^2]^{1/2}$). Secondary structure elements are depicted with the same color as in Fig. 2.

coalescence could be observed for a couple of residues (Arg^{29}, Leu^{32}) with a single signal (larger line-width at intermediate DNA/protein ratio) at a chemical shift variation of about 50 Hz. The average chemical shift changes ($\Delta\delta = [(\Delta\delta_{HN})^2 + (\Delta\delta_N \times 0.154)^2]^{1/2}$) between the free and the DNA-bound state of the THAP domain were plotted *versus* the THAP1 residue numbers (Fig. 5*B*). Several protein residues experienced significant chemical shift perturbation ($\Delta\delta$ higher than 0.2 ppm) and these were mainly organized into three different patches. The first one includes residues Ala^7, Tyr^8, Lys^{11}, Arg^{13}, Val^{20}, Ser^{52}, Ile^{53}, Cys^{54}, and Glu^{56} located in the region encompassing the zinc atom. The second patch reveals residues Arg^{29}, Trp^{36}, Ala^{38}, and Ala^{39} within or nearby the α-helix H1. Residues in the third patch (Asn^{68}, Asn^{69}, Lys^{70}, Leu^{72}, Ala^{76}) are located in the loop L4. Two additional residues undergoing large chemical shift changes are located in the loop L3 following the α-helix H1 and correspond to residues Lys^{46} and Thr^{48}. Notably, most of these residues are located either in structured regions of the domain (H1, H2, and β2), or in ordered regions of the loops L1 (residues 7–11), L2 (Arg^{29}), L3 (Thr^{48}), and L4 (residues 69–76) that exhibit restricted motions (Fig. 2*C* and supplemental Fig. S1).

Interestingly, the clusters of large chemical shift changes upon DNA binding map to the highly positively charged area of the THAP domain (Fig. 4*F*), further supporting the potential role of this region as the DNA binding interface of the THAP zinc finger.

Data Deposition—the atomic coordinates for the structures of the THAP domain (Met^1–Phe^{81}) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (code 2jtg). Chemical Shift informations for the THAP domains (Met^1–Lys^{90} and Met^1–Phe^{81}) are available from the BioMagRes Data Bank, under the accession numbers 15300 and 15289, respectively.

DISCUSSION

We report here the three-dimensional structure and structure-function analysis of the sequence-specific DNA-binding THAP zinc finger of human THAP1, the prototype of a novel family of cellular factors involved in pRb/E2F cell cycle pathways. We recently demonstrated that THAP1 is a physiological regulator of cell proliferation. Silencing of THAP1 by RNA interference in human primary endothelial cells resulted in inhibition of G_1/S cell cycle progression and down-modulation of several pRb/E2F cell cycle target genes, including *RRM1,* a gene activated at the G_1/S transition and essential for S-phase DNA synthesis (10). We showed that the THAP zinc finger of THAP1 recognizes a consensus THAP1-binding site in the *RRM1* promoter and that endogenous THAP1 associates *in vivo* with this site, indicating that *RRM1* is a direct target gene of THAP1. The solution structure of the THAP domain of THAP1 is therefore the first structure of a THAP zinc finger with demonstrated biochemical activity as a sequence specific DNA-binding domain (7) and associated with a known biological function, *i.e.* recruitment of THAP1 on the pRb/E2F target gene *RRM1* (10). In contrast, although the structure of the THAP domains from human THAP2 (PDB code 2D8R) and *C. elegans* CtBP have been determined (26), these two proteins have not been functionally characterized, and it is not yet known whether their THAP domains possess sequence-specific DNA binding properties.

The structure of the THAP zinc finger differs from that of other DNA-binding modules belonging to the zinc finger superfamily. For instance, the $\beta\alpha\beta$ topology, the long spacing between the two pairs of zinc ligands (up to 53 residues in some THAP domains) distinguish the THAP zinc finger from the classical DNA-binding C2H2 zinc finger, which exhibits a $\beta\beta\alpha$ topology with a shorter spacing (10–12 residues) between the two pairs of zinc-coordinating residues (3). The position of the zinc is also an interesting feature; in the classical zinc finger, the zinc atom plays a central role in the structure by coordinating four ligands that anchor one end of the helix to one end of the β-sheet, whereas in the C2CH THAP motif, the zinc is not buried in the interior of the protein, and it links the N terminus of the domain to the second β-strand, without involving the α-helix that is distal to the zinc ion. The presence of the long loop-helix-loop insertion in the two-stranded anti-parallel β-sheet is one of the most distinctive features of the THAP zinc finger. It explains the atypical spacing of the C2 and CH residues in the C2CH zinc-coordinating module and the relatively large size of the THAP domain (~80 residues) compared with the C2H2 zinc finger (~30 residues). The above features are very unique and are not found in other classes of zinc-coordinating DNA-binding modules. Surprisingly, however, the THAP zinc finger exhibits structural similarities with a protein-protein interaction module, the Zinc Finger-Associated Domain (ZAD, PDB code 1PZW) of the *Drosophila* transcription factor Grauzone (26, 42). These structural homologies include the presence between the zinc-coordinating residues of a similar loop-helix-loop insertion into the two-stranded anti-parallel β-sheet. However, despite these similarities in their molecular scaffolds, the THAP zinc finger and the ZAD domain are linked to different functions. The ZAD domain mediates protein-protein interactions and exhibits a highly negative electrostatic potential inconsistent with DNA-binding properties (42). In contrast, the THAP zinc finger of THAP1 functions as a sequence specific DNA-binding module with a highly positively charged surface (Fig. 4*F*).

Our previous mutagenesis studies have revealed that mutation of any of the eight residues that define the THAP zinc finger motif (including the C2CH residues) abrogate DNA binding activity of the domain (7). Mapping of these residues on the THAP zinc finger structure indicates that these amino acids play an essential role in the folding of the domain (Fig. 2*D*). In the present study, we identified five additional residues that are essential for DNA binding activity. Two of these residues (Arg^{29} and Phe^{45}) could play a structural role by anchoring the loops of the loop-helix-loop motif to the hydrophobic core of the domain, potentially limiting the motions of these loops (Fig. 4*E*). The three other essential residues (Lys^{24}, Arg^{42}, Thr^{48}) are exposed on the positively charged surface of the THAP domain of THAP1 (Fig. 4*F*) and are therefore less likely to contribute to the folding or structure of the domain. Rather, they may play a direct role in DNA binding. Although Lys^{24} and Arg^{42} do not display significant chemical shift changes upon DNA binding, the residue Thr^{48} is clearly affected and undergoes the largest

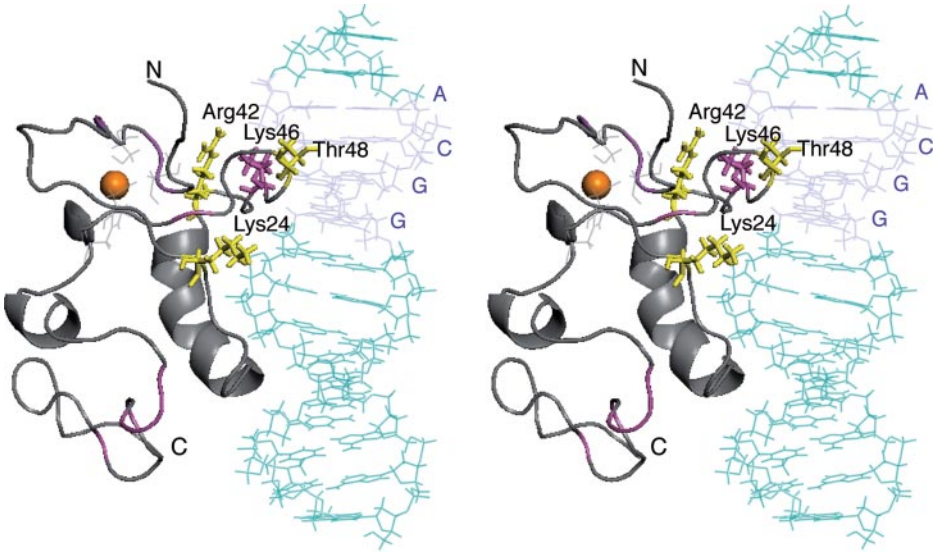


FIGURE 6. **A proposed model of the complex between the THAP zinc finger of THAP1 (*gray*, *pink*, *yellow*) and its DNA target (*green*, *blue*) in stereo view.** Amino acids with backbone chemical shifts that undergo more than 0.2 ppm chemical shift change upon DNA binding are colored *pink*. Side chains of residues that are found critical from site-directed mutagenesis experiments are depicted in *yellow*. The side chain of Lys[46] is colored in *pink*. The THABS DNA molecule is colored *green* except bases at the GGCA core motif shown in *blue*. The THABS molecule was docked onto the THAP zinc finger of THAP1 using HADDOCK1.3 (High Ambiguity Driven Protein Docking) (40).

change in chemical shift, consistent with it being directly involved in DNA interactions (Fig. 5*B*). Furthermore, the residue Thr[48] is poorly conserved among the THAP domains (Fig. 2*E*) suggesting a key role in binding specificity.

Surprisingly, with the exception of Trp[36], mutation of residues located in the helix of the loop-helix-loop motif did not abrogate DNA binding. Therefore, despite the fact the α-helix is the most common protein structural element used for DNA recognition in zinc fingers and other types of DNA-binding domains (2, 43, 44), our results strongly suggest that the α-helix of the THAP zinc finger may not be the main DNA recognition element. In contrast, the positively charged region is likely to play a key role (electrostatic contacts) together with more specific contacts involving residues in the loops, namely the loop L3 of the loop-helix-loop motif (Arg[42], Thr[48]) and the loop L4 at the C terminus. Based on the NMR and mutagenesis data, a model of the complex between the THAP zinc finger of THAP1 and its DNA target was built using HADDOCK1.3 (40) (Fig. 6). The proposed model shows a good shape complementarity between the loop-helix-loop motif (L2-H1-L3) and DNA. The helix does not appear as the major recognition element but is located along the DNA chain so that the two loops on its sides fit into the DNA grooves. Remarkably, the loop L3 enters into the major groove to contact DNA and in particular, side chain of Thr[48] gives a polar contact with the GGCA core. Although the EMSA assays allowed us to define critical residues in this loop, these qualitative assays may be too insensitive to detect the role of other residues in DNA binding. Additional experiments may reveal, for instance, a role for Lys[46], a residue that undergoes significant chemical shift changes upon DNA binding (Fig. 5), and is predicted to be in close proximity to DNA in the model of the protein-DNA complex (Fig. 6).

Interestingly, genetic data obtained in *C. elegans* for THAP family members LIN-36 and HIM-17 have revealed several single-point mutations which affect the functional activity of the THAP zinc finger. Most of these mutations concern residues that are critical for the folding of the domain. For instance, mutation of the second Cys of the C2CH motif was found in one of the THAP domains of HIM-17 (15), while two independent mutations were found in the last Pro residue of the THAP motif in the LIN-36 protein (13). However, other mutations have been found to affect residues that do not appear to be part of the hydrophobic core of the domain, and these may correspond to residues exposed on the surface and directly involved in DNA binding. Finally, a double alanine mutation introduced into the THAP zinc finger of *Drosophila* P element transposase at the level of residues His[18] and Cys[22] (corresponding to THAP1 residues Lys[24] and Arg[29]) has previously been shown to abrogate sequence-specific DNA binding activity of the protein (41). This suggests that the essential residues we have identified in the present study are likely to be also critical for the functional activity of other THAP zinc fingers.

In this study, we show that the different THAP zinc fingers, despite sharing some structural homologies, do not recognize the same DNA target sequence. We found that recombinant THAP domains from human THAP2 and THAP3, and *C. elegans* CtBP and GON-14 do not exhibit sequence-specific DNA binding activity toward the DNA sequence motif recognized by human THAP1 (Fig. 3). Although Ce-CtBP and GON-14 were able to bind the THAP1 target sequence, this DNA binding activity was completely eliminated in the presence of nonspecific competitor DNA. Together with the observation that distinct THAP domains sequences within a single species exhibit less than 50% identity between each other (7), this suggests that each THAP zinc finger may possess its own specific DNA-binding site. This possibility is further supported by the observation that the DNA target sequence of the THAP zinc finger of THAP1 does not share homology with the AT-rich motif recognized by the THAP zinc finger of P element transposase (7, 45). However, we cannot exclude at this stage that some THAP zinc fingers may lack sequence specificity or even DNA binding activity, and may rather function as protein-protein interaction modules.

Finally, protein-protein interactions mediated by other domains of the THAP proteins may be critical to increase the DNA binding activity of the THAP zinc finger, which appears to be relatively weak. In this respect, the C-terminal coiled-coil domain found in THAP1, as well as several other human THAP proteins, may enhance the affinity of the full-length protein for DNA by allowing dimerization or multimerization. Future studies will help to resolve these issues and will provide impor-

tant new insights about the structure and functions of THAP zinc finger proteins both in humans and model animal organisms.

Acknowledgments—We thank E. Guittet at the ICSN Institute, Gif/Yvette for providing access to the 800-MHz National Facility. We wish to acknowledge O. Saurel and P. Ramos at the IPBS, Toulouse for NMR technical assistance and J. P. Estève at the CHU Rangueil, Toulouse for help with SPR experiments. The 600 MHz cryoprobe and 700 MHz IPBS spectrometers were acquired within the CPER 2000–2006 involving CNRS, University of Toulouse, Région Midi-Pyrénées and European Structural Funds (FEDER).

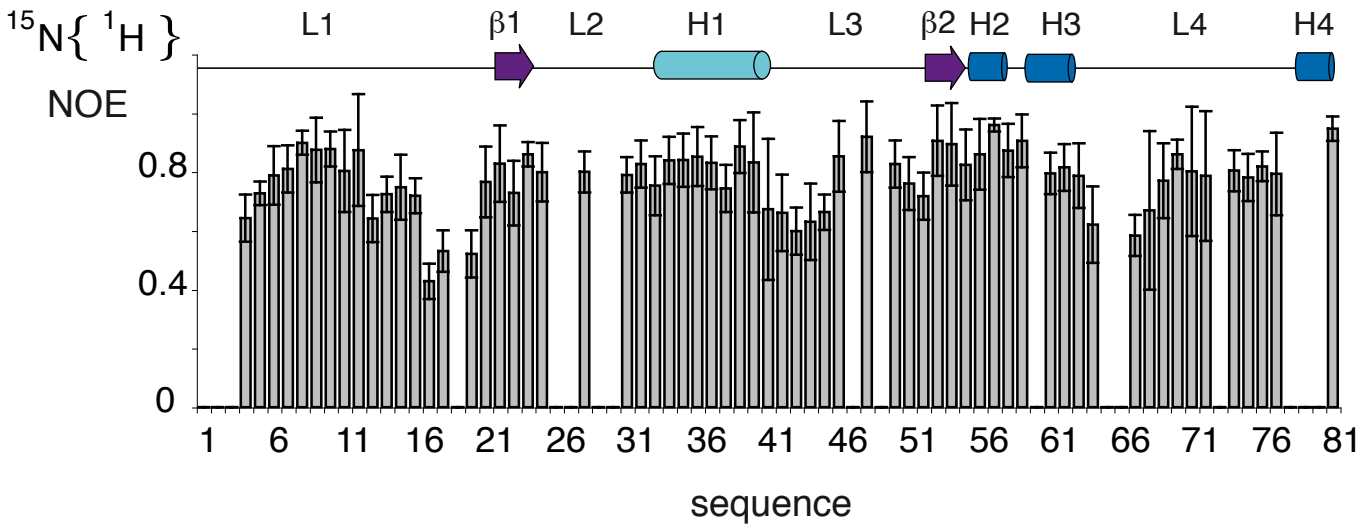
REFERENCES

1. Klug, A. (1999) *J. Mol. Biol.* **293,** 215–218
2. Pavletich, N. P., and Pabo, C. O. (1991) *Science* **252,** 809–817
3. Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A., and Wright, P. E. (1989) *Science* **245,** 635–637
4. Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. J., and Gronenborn, A. M. (1993) *Science* **261,** 438–446
5. Schwabe, J. W., Neuhaus, D., and Rhodes, D. (1990) *Nature* **348,** 458–461
6. Roussigne, M., Kossida, S., Lavigne, A. C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F., and Girard, J. P. (2003) *Trends Biochem. Sci.* **28,** 66–69
7. Clouaire, T., Roussigne, M., Ecochard, V., Mathe, C., Amalric, F., and Girard, J. P. (2005) *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6907–6912
8. Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F., and Girard, J. P. (2003) *Oncogene* **22,** 2432–2442
9. Hammer, S. E., Strehl, S., and Hagemann, S. (2005) *Mol. Biol. Evol.* **22,** 833–844
10. Cayrol, C., Lacroix, C., Mathe, C., Ecochard, V., Ceribelli, M., Loreau, E., Lazar, V., Dessen, P., Mantovani, R., Aguilar, L., and Girard, J. P. (2007) *Blood* **109,** 584–594
11. Giangrande, P. H., Zhu, W., Schlisio, S., Sun, X., Mori, S., Gaubatz, S., and Nevins, J. R. (2004) *Genes Dev.* **18,** 2941–2951
12. Ferguson, E. L., and Horvitz, H. R. (1989) *Genetics* **123,** 109–121
13. Thomas, J. H., and Horvitz, H. R. (1999) *Development* **126,** 3449–3459
14. Boxem, M., and van den Heuvel, S. (2002) *Curr. Biol.* **12,** 906–911
15. Reddy, K. C., and Villeneuve, A. M. (2004) *Cell* **118,** 439–452
16. Chesney, M. A., Kidd, A. R., 3rd, and Kimble, J. (2006) *Genetics* **172,** 915–928
17. Fay, D. S., Keenan, S., and Han, M. (2002) *Genes Dev.* **16,** 503–517
18. Koreth, J., and van den Heuvel, S. (2005) *Oncogene* **24,** 2756–2764
19. Lewis, P. W., Beall, E. L., Fleischer, T. C., Georlette, D., Link, A. J., and Botchan, M. R. (2004) *Genes Dev.* **18,** 2929–2940
20. Korenjak, M., Taylor-Harding, B., Binne, U. K., Satterlee, J. S., Stevaux, O., Aasland, R., White-Cooper, H., Dyson, N., and Brehm, A. (2004) *Cell* **119,** 181–193
21. Litovchick, L., Sadasivam, S., Florens, L., Zhu, X., Swanson, S. K., Velmurugan, S., Chen, R., Washburn, M. P., Liu, X. S., and DeCaprio, J. A. (2007) *Mol. Cell* **26,** 539–551
22. Harrison, M. M., Ceol, C. J., Lu, X., and Horvitz, H. R. (2006) *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16782–16787
23. Ceol, C. J., and Horvitz, H. R. (2004) *Dev. Cell* **6,** 563–576
24. Poulin, G., Dong, Y., Fraser, A. G., Hopper, N. A., and Ahringer, J. (2005) *EMBO J.* **24,** 2613–2623
25. Macfarlan, T., Kutney, S., Altman, B., Montross, R., Yu, J., and Chakravarti, D. (2005) *J. Biol. Chem.* **280,** 7346–7358
26. Liew, C. K., Crossley, M., Mackay, J. P., and Nicholas, H. R. (2007) *J. Mol. Biol.* **366,** 382–390
27. Sattler, M., Schleucher, J., and Griedinger, C. (1999) *Progr. Nucl. Magn. Reson. Spectr.* **34,** 93–158
28. Cornilescu, G., Delaglio, F., and Bax, A. (1999) *J. Biomol. NMR* **13,** 289–302
29. Ponstingl, H., and Otting, G. (1998) *J. Biomol. NMR* **12,** 319–324
30. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, pp. 122–125, Wiley Press
31. Schubert, M., Labudde, D., Oschkinat, H., and Schmieder, P. (2002) *J. Biomol. NMR* **24,** 149–154
32. Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R., and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.* **117,** 479–485
33. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) *J. Biomol. NMR* **6,** 277–293
34. Bartels, C., Xia, T.-H., Billeter, M., Güntert, M., and Wüthrich, K. (1995) *J. Biomol. NMR* **5,** 1–10
35. Johnson, B. A. (2004) *Methods Mol. Biol.* **278,** 313–352
36. Auguin, D., Barthe, P., Auge-Senegas, M. T., Stern, M. H., Noguchi, M., and Roumestand, C. (2004) *J. Biomol. NMR* **28,** 137–155
37. Daikun, G. P., Fairall, L., and Klug, A. (1986) *Nature* **324,** 688–699
38. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 905–921
39. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) *J. Biomol. NMR.* **8,** 477–486
40. Dominguez, C., Boelens, R., and Bonvin, A. M. (2003) *J. Am. Chem. Soc.* **125,** 1731–1737
41. Lee, C. C., Beall, E. L., and Rio, D. C. (1998) *EMBO J.* **17,** 4166–4174
42. Jauch, R., Bourenkov, G. P., Chung, H. R., Urlaub, H., Reidt, U., Jackle, H., and Wahl, M. C. (2003) *Structure* **11,** 1393–1402
43. Laity, J. H., Lee, B. M., and Wright, P. E. (2001) *Curr. Opin. Struct. Biol.* **11,** 39–46
44. Garvie, C. W., and Wolberger, C. (2001) *Mol. Cell* **8,** 937–946
45. Kaufman, P. D., Doll, R. F., and Rio, D. C. (1989) *Cell* **59,** 359–371

SUPPLEMENTARY FIGURES

Supplementary Fig. 1. Histogram of ^{15}N $\{^{1}H\}$NOE enhancement ratios as a function of sequence.

Suppl. Fig. 1.


15N{ 1H }
NOE
L1
β1
L2
H1
L3
β2
H2
H3
L4
H4
0.8
0.4
0
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
sequence

Résultats complémentaires et discussions

Le domaine THAP de THAP1

Redéfinition du domaine d'un point de vue structural

Comme nous l'avons vu, nous avons travaillé dans un premier temps avec le domaine THAP {Met1-Lys90} tel qu'il avait été défini par Roussigne et ses collaborateurs (Roussigne et al., 2003b) par des analyses de séquence.
Nous avons réalisé l'attribution de ce domaine ainsi que les premiers calculs de structure. Toutefois la partie C-terminale de ce domaine est apparue peu structurée, nous n'avons qu'une attribution partielle des résidus 80 à 90 qui présentent très peu de NOEs. Nous avons donc cherché à déterminer un domaine de taille plus courte.
Si la proline 78 du motif conservé AVPTIF a été démontrée comme essentielle pour la liaison à l'ADN (Clouaire et al., 2005), le rôle des résidus au delà de ce motif n'avait pas été clairement établi (résidus 82 à 90). Nous avons donc cloné deux constructions plus courtes 1-81 et 1-63 pour lesquelles nous avons testé à la fois la fonctionnalité par liaison à la cible ADN consensus observée sur gel retard et la structuration par enregistrement d'une expérience ^{15}N HSQC (figure 56).
Comme nous pouvions l'attendre, la construction 1-63 ne comprenant pas le motif AVPTIF ne présente aucune activité sur gel retard et son spectre ^{15}N HSQC montre une faible dispersion caractéristique d'une protéine non structurée. Ce domaine est néanmoins produit de façon soluble.
Par contre, de façon intéressante, la construction 1-81 se lie spécifiquement à la séquence ADN cible et présente un spectre ^{15}N HSQC comparable à celui du domaine 1-90 mais présentant moins de recouvrement dans la partie centrale.
L'emploi de l'expérience ^{15}N HSQC est un critère de sélection pour une protéine bien structurée dans des projets de protéomique structurale (Yee et al., 2002).

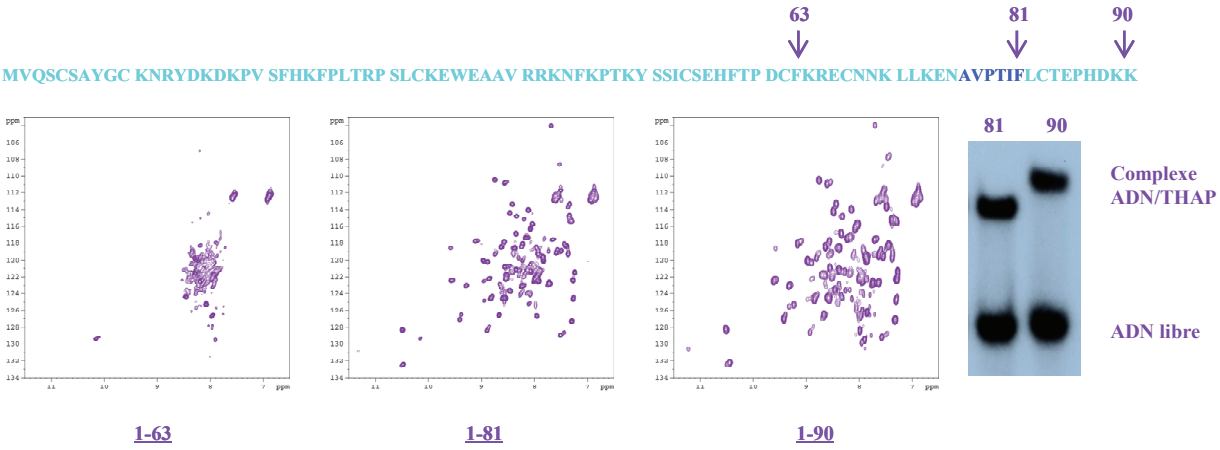


Figure 56 : **Identification du domaine THAP 1-81**
Spectres ^{15}N HSQC des constructions 1-63, 1-81 et 1-90 et gel retard avec la séquence ADN consensus pour les constructions 1-81 et 1-90.

Les 81 premiers résidus sont donc suffisants à la fonction de reconnaissance spécifique et au repliement du domaine. Ce travail nous a permis de redéfinir le domaine THAP aux résidus 1 à 81 de la protéine THAP1 et nous avons donc continué l'étude structurale en nous concentrant sur ce domaine.

La définition des domaines protéiques est un enjeu majeur dans les études de biologie structurale et l'identification de domaine en absence d'information structurale reste un problème important. Cette identification à partir des données de séquences reste subjective et peut parfois mener à une définition différente de celle basée sur des données structurales (Elofsson and Sonnhammer, 1999). La prédiction de domaines est néanmoins d'une grande importance étant donné l'écart entre le nombre de séquence et de structures de protéines connues.

Comparaison des domaines de THAP1, THAP2 et CtBP

Nous avons résolu la structure du domaine THAP de THAP1. Deux autres structures du domaine THAP ont également été résolues au cours de nos travaux, la structure du domaine THAP de THAP2 (code PDB : 2d8r) et de CtBP (Liew et al., 2007) (code PDB : 2jm3) . Ces structures présentent un motif similaire; on retrouve en commun, un repliement de topologie $\beta\alpha\beta$, la coordination tétraédrique au zinc avec les ligands C_2CH et la présence d'une hélice 3_{10} entre les résidus en position 3 et 4 de la signature C_2CH.

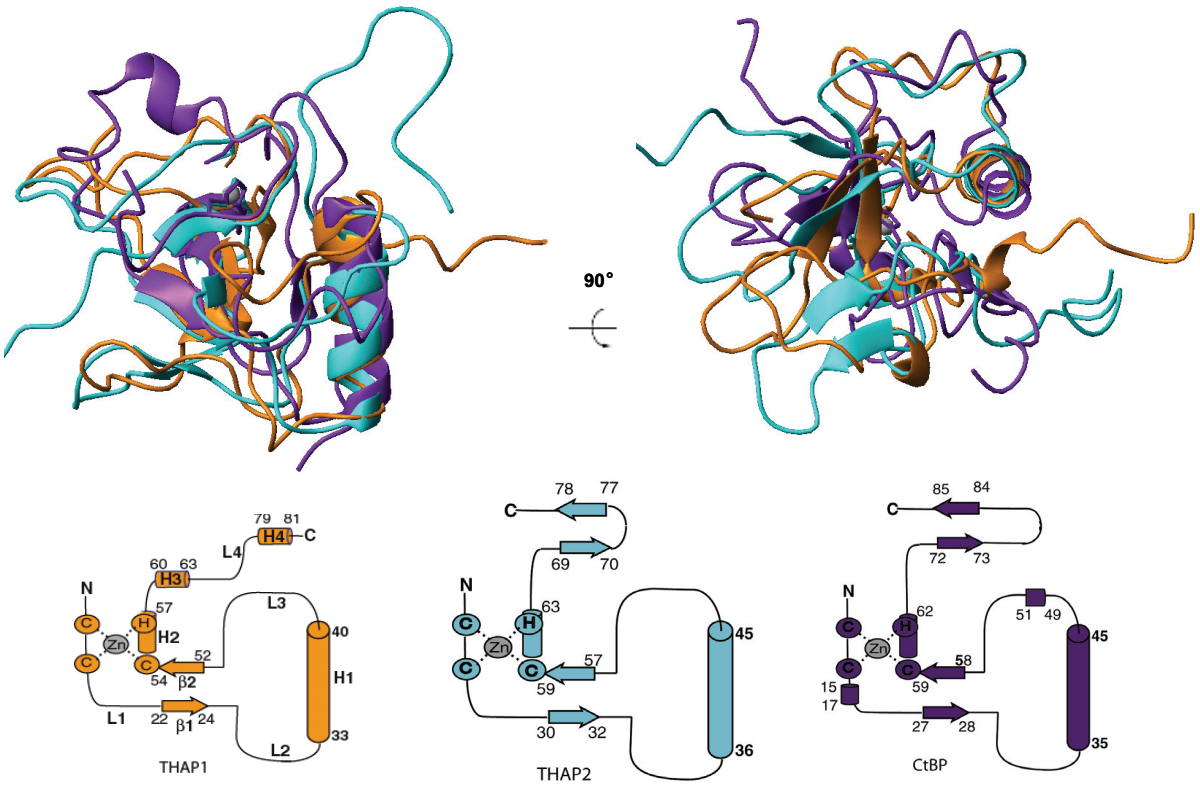


Figure 57 : **Superposition des domaines THAP de THAP1, THAP2 et CtBP**
Représentation des structures des domaines THAP de THAP1 (code PDB : 2jtg) en violet, de THAP2 en bleu (code PDB : 2d8r) et CtBP (code pdb : 2jm3), superposées selon un ajustement sur les structures des Cα des structures secondaires du motif βαβ.

L'homologie structurale la plus grande est entre THAP1 et THAP2 qui présentent des séquences proches (48% d'identité contre 27% entre THAP1 et CtBP) et pour lesquelles on compte 80 C_α équivalents superposables avec un rmsd de 2.8 Å contre 66 C_α et un rmsd de 3.1 Å entre les domaines THAP de THAP1 et de CtBP. La détermination des atomes superposables a été réalisée avec le programme DALI (Distance matrix ALIgnment) (Holm and Sander, 1993).

Néanmoins, des variabilités structurales sont observées dans la partie C-terminale. En effet, le domaine THAP de THAP1 présente deux courtes hélices 3_{10} supplémentaires qui comprennent les résidus 60 à 63 et 79 à 81 tandis que les domaines THAP de THAP2 et CtBP présentent un feuillet β antiparallèle supplémentaire, formé par les résidus équivalent à THAP1 63 à 64 et 71 à 72 pour THAP2 et 76 à 77 et 82 à 83 pour CtBP. Ces feuillets supplémentaires dans les structures de THAP2 et CtBP ne sont donc pas similaires. On notera que le deuxième brin de CtBP est formé par des résidus 82 et 83 équivalents chez THAP1

qui ne font pas parti du domaine redéfini aux résidus 1 à 81. De plus les 7 résidus au début de la boucle L4 ne sont pas présents chez CtBP.

Un nouveau motif en doigt de zinc

Nous pouvons définir le domaine THAP d'un point de vue structural à partir des éléments identifiés en commun dans les trois structures connues; soit un repliement de topologie $\beta\alpha\beta$, une coordination au zinc et la présence d'une hélice 3_{10} entre les résidus en position 3 et 4 de la signature C_2CH. Il est judicieux de considérer également le type de coordination au zinc (C_2CH) et l'espacement entre les ligands au zinc ($C\text{-}X_{2\text{-}4}\text{-}C\text{-}X_{35\text{-}53}\text{-}C\text{-}X_2\text{-}H$) pour définir complètement cette structure en doigt de zinc.

A partir de ces critères de définition, nous avons cherché à classer ce domaine parmi les domaines à doigts de zinc en fonction des ligands au zinc ou selon une classification structurale. Il ne nous est pas apparu possible d'intégrer ce domaine à une des classes définies : la structure du domaine THAP est différente de celles des domaines de la super famille des doigts de zinc.

En particulier le domaine THAP diffère du doigt de zinc classique C_2H_2 qui présente un repliement de topologie $\beta\beta\alpha$ et un espacement plus réduit (10-12 acides aminés) entre les deux paires de résidus coordonnés au zinc. La position du zinc est différente pour les deux domaines. Dans le doigt de zinc classique, le zinc joue un rôle central dans la structure en coordonnant deux résidus situé à une extrémité de l'hélice α et deux autres sur un des brins du feuillet β. Dans le domaine THAP, le zinc n'est pas enfoui au cœur de la structure, il coordonne deux résidus de l'extrémité C-terminale à deux résidus sur un brin du feuillet β et sur l'hélice 3_{10}, il ne relie pas, en particulier, l'hélice α qui n'est pas à proximité de l'ion zinc (figure 58). La présence de cette hélice entre les deux brins du feuillet β est une des caractéristiques principales du domaine THAP. La présence de cet élément de structure secondaire est à mettre en relation avec le grand espacement atypique entre les deux paires de ligand au zinc C2 et CH (entre 35 et 53 résidus pour le domaine THAP contre 12 pour le doigt de zinc classique) et la relative grande taille du domaine THAP (~80 acides aminés contre ~30 pour le domaine classique).

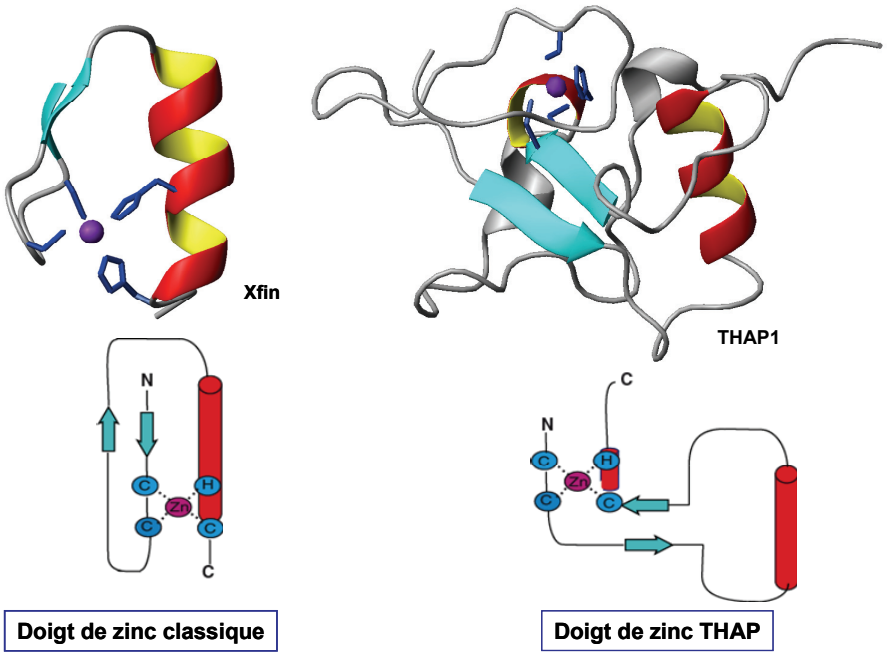


Figure 58 **: Le domaine doigt de zinc THAP C_2CH est différent du domaine classique C_2H_2**
Représentation du domaine doigt de zinc classique (Xfin code PDB : 1znf) et du domaine THAP (THAP1 code PDB : 2jtg) et de leur topologie.

Le domaine THAP diffère également des motifs C_2CH identifiés (le domaine du facteur de transcription amt1 (Turner et al., 1998) et le sous-domaine N-terminal du domaine de liaison à l'ADN de SBP (Yamasaki et al., 2004a)) qui dérivent du motif classique C_2H_2 (figure 10)

Liew et ses collaborateurs proposent tout de même une classification possible du domaine THAP dans la classe des doigts de zinc *treble clef* (clef de sol : 𝄞) (table 2 p26). Mais, comme le remarquent Liew et ses collaborateurs, pour le domaine THAP les deux premiers ligands au zinc ne sont pas situés sur une épingle à cheveux β, comme c'est le cas pour les membres de la classe des doigts de zinc *treble clef,* et les deux autres ligands au zinc ne sont situés qu'à l'extrémité d'une hélice à un seul tour et non entièrement sur une hélice. De plus, cette classification du domaine THAP ne rend pas compte du repliement βαβ et donc de la présence de ces deux structures secondaires (feuillet β et hélice α) qui forment le cœur de la structure du domaine THAP. Cette classification du domaine THAP en doigt de zinc *treble clef* n'est donc pas parfaite.

Nous pensons donc pour ces raisons que le domaine THAP adopte un repliement en doigt de zinc original qui ne peut être rattaché à aucune classe de doigt de zinc

décrite. Nous employons d'ailleurs la désignation *domaine doigt de zinc THAP* pour identifier ce motif structural.

Homologues structuraux

La recherche d'homologues structuraux parmi la Protein Data Bank avec l'aide du programme DALI (Holm and Sander, 1993) ne donne aucun résultat, confirmant que le domaine THAP a une structure originale.

Toutefois, des similarités structurales ont pu être trouvées avec le domaine ZAD (zinc finger associated domain) de grauzone (code PDB : 1pzw) (Liew et al., 2007) tout particulièrement en ce qui concerne la topologie (l'insertion d'une hélice α entre les deux brins du feuillet β) et la position des ligands au zinc (figure 59). Par comparaison des deux structures sur DALI, 52 atomes C_α équivalents ont pu être identifiés avec un rmsd de 3.3 Å. Les régions superposables comprennent le site de coordination au zinc, le premier brin du feuillet β ainsi que le motif boucle hélice boucle (L2 H1 L3). La région correspondant aux résidus 58 à 77 de THAP1 est quant à elle non superposable. La structure de ZAD présente une grande hélice (H2) qui est impliquée dans la dimérisation de la protéine et des interactions protéine-protéine (Jauch et al., 2003). Nous pouvons remarquer que la petite hélice 3_{10} (H2) de THAP correspond au début de cette hélice H2 chez ZAD et pourrait donc être un vestige de cette longue hélice.

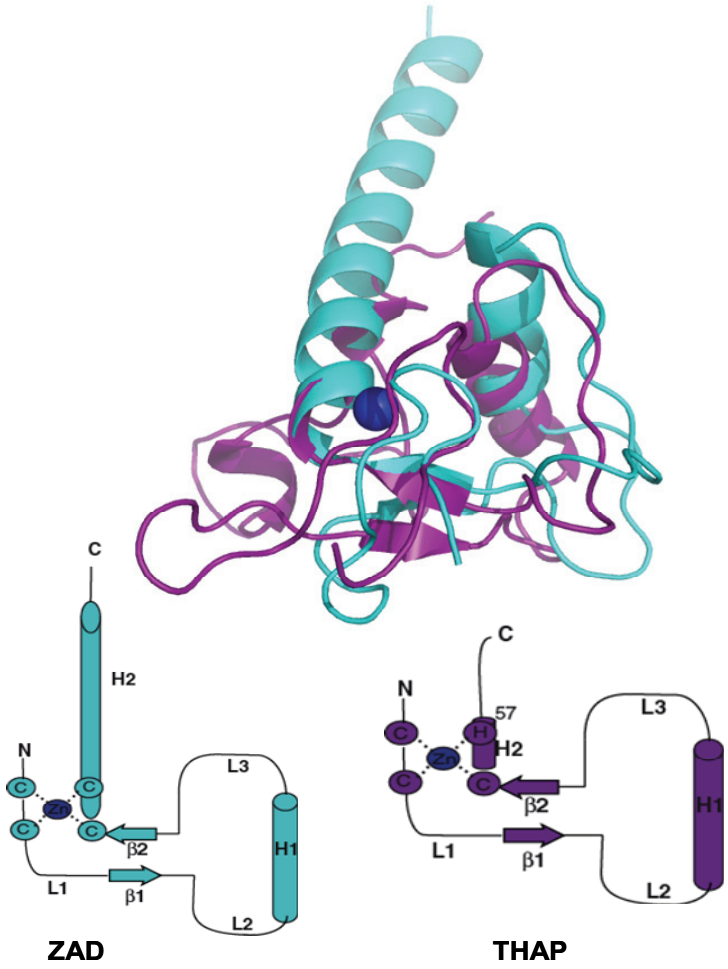


Figure 59 : **Superposition des domaines Zad et THAP**

Représentation du domaine ZAD de grauzone en bleu (code PDB : 1pzw) et du domaine THAP de THAP1 (code PDB : 1znf) et de leur topologie. L'atome de zinc est représenté par une sphère bleue, il est superposable dans les deux structures.

Cependant même si ces deux domaines présentent des similarités structurales, ils possèdent des fonctions différentes. Le domaine ZAD est impliqué dans des interactions protéine-protéine et possède une surface chargée négativement incompatible avec des interactions protéine-ADN alors que le domaine THAP présente une surface chargée positivement et que le domaine THAP de THAP1 est impliqué dans des interactions protéine-ADN.

Dynamique du domaine THAP

La RMN offre l'avantage de pouvoir réaliser l'étude dynamique de la protéine en solution et ainsi déterminer les parties figées et mobiles et participer à la

compréhension du mécanisme d'interaction. Nous avons, à ce propos, mesuré les paramètres de relaxation ^{15}N que sont les temps de relaxation longitudinal (T1) et transversal (T2) et les valeurs des NOEs hétéronucléaires ^{1}H-^{15}N. L'exploitation de ces données nous permet d'avoir accès à la dynamique globale et locale de la protéine. (pour revues : (Boehr et al., 2006; Fischer et al., 1998; Korzhnev et al., 2001).

Dans le cas de notre étude, il est intéressant de mesurer ces paramètres sur le domaine seul et sur le complexe de façon à réaliser une comparaison utile pour la compréhension du mécanisme de reconnaissance de l'ADN par la protéine.

Ces études ont été initialement entreprises sur le domaine THAP 1-90 et ont été ensuite reprises sur le domaine THAP 1-81. Ce travail est désormais poursuivi par S. Campagne, doctorant dans l'équipe. De mesures ont été ainsi réalisées à 600 MHz sur le domaine 1-81, à 296K, dans deux conditions de force ionique, à 10 mM et 250 mM NaCl. Elles sont actuellement complétées par des mesures à différentes températures et différents champs. L'analyse de la dynamique du complexe est également en prévision.

Nous exposons ici les valeurs de vitesse de relaxation longitudinale R1 (1/T1) et transversale R2 (1/T2) et des NOEs hétéronucléaires ^{15}N-^{1}H pour le domaine 1-81 à 296K, 600 MHz et 10 mM NaCl (figure 60).

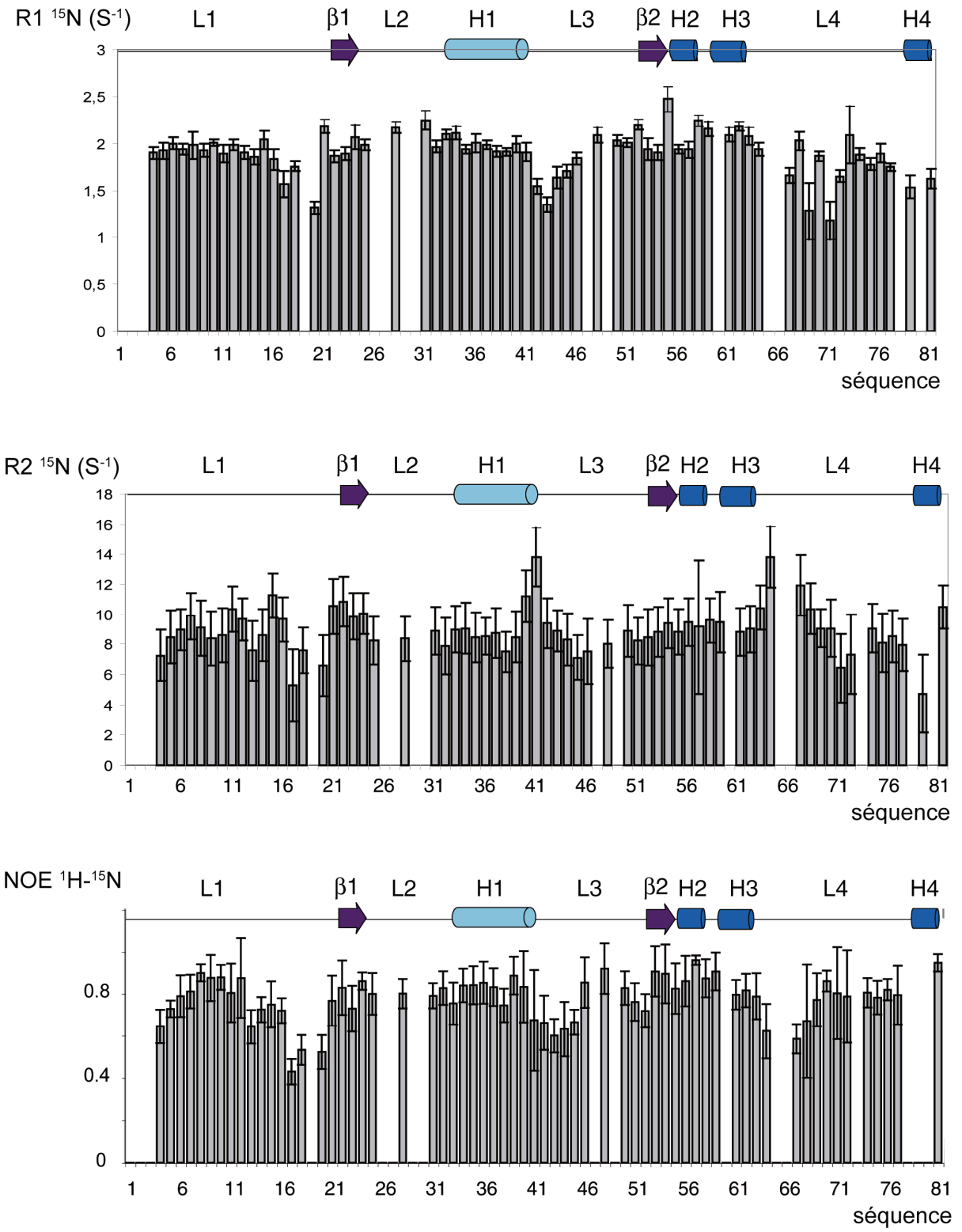


Figure 60 : **Valeurs de R1, R2 et de NOEs hétéronucléaires**
Valeurs mesurées à 296 K, 10 mM NaCl et 600 MHz sur le domaine THAP 1-81

Les valeurs de T1, T2 ont été ajustées sur un modèle monoexponentiel avec le module *rate analysis* du logiciel NMRView (Johnson, 2004). Nous utilisons également ce module pour l'analyse des NOEs hétéronucléaires.

Nous avons déterminé à partir de l'analyse des données de T1 et T2 le temps de corrélation de diffusion rotationnelle globale (τ_c) du domaine 1-81 à 10 mM et 250 mM NaCl dont les valeurs sont respectivement de 6.0 ± 0.1 ns et 6.9 ± 0.4 ns. Ces valeurs ont été calculées à partir du rapport R1/R2 sur les résidus présentant le moins de flexibilité (paramètre d'ordre S^2 proche de 1) et présentant des mouvements internes de faible amplitude (temps de corrélation des mouvements rapides négligeables, τ_e<10 ps) de telle sorte que l'approximation ci-dessous, de la fonction de densité spectrale ($J(\omega)$) puisse être faite. Il s'agit de la fonction de densité globale pour une molécule sphérique rigide en rotation globale isotrope et aléatoire.

$$J(\omega) = \frac{2}{5} \cdot \frac{\tau_c}{1 + \omega^2 {\tau_c}^2}$$

Une estimation théorique du temps de corrélation globale peut être calculée en fonction de la température (T en K) et du nombre de résidus (N) à partir de la relation suivante (Daragan and Mayo, 1998) :

$$\tau_c = \frac{0.0098}{T} . e^{(\frac{2416}{T})} . N^{0.93}$$

Ainsi, la valeur estimée du temps de corrélation pour le domaine 1-81 (comprenant 87 résidus) est de 7.4 ns à 296K pour la forme monomérique et de 14,7 ns pour une espèce dimérique. Nous concluons donc que le domaine est monomérique aux deux concentrations en NaCl, 10 mM et 250 mM.

De plus, l'analyse qualitative des données de relaxation pour le domaine 1-81 à 10 mM et 296 K, nous a permis de définir les zones flexibles de la protéine. En particulier, nous avons figuré dans l'article la valeur des NOEs hétéronucléaires sur la structure, à mettre en relation avec l'ensemble des 20 structures de plus basse énergie générées. Les zones flexibles déterminées par les données de relaxation correspondent aux régions moins bien définies lors du calcul de structure. Ces régions présentent des valeurs de NOE hétéronucléaires plus faibles, comparées aux régions structurées en hélice ou en feuillet. Les zones flexibles sont constituées des acides aminés 1 à 3, 17 à 20, 41 à 45, 64 à 69 et 82 à 87.

La région correspondant aux acides aminés 64 à 68 (début de la boucle L4) apparaît la plus désordonné, elle ne présente pas en effet pas de NOEs inter-résiduels. Elle présente des valeurs de NOE hétéronucléaires faibles, comparables à celles de la

région formée des acides aminées 16 à 20 (fin de la boucle L1) mais qui n'est pas aussi désordonnée. Cela peut s'expliquer par une mobilité de la région 64 à 68 supérieure au ns et donc non détectable par les expériences de relaxation ^{15}N. Cette région présente des structures secondaires différentes qui ne sont pas retrouvées avec CtBP et THAP2 (figure 57) et elle pourrait avoir un rôle dans la liaison à l'ADN comme nous l'expliquons ci-après. Il est possible que cette région se structure au contact de l'ADN, soulignant un rôle possible des régions non-structurées dans les interactions protéiques (Dyson and Wright, 2005; Zhang et al., 2007).
D'autre part, la présence de ces régions mobiles permet d'expliquer pourquoi un grand nombre de protons amides s'échangent rapidement avec le solvant lors des expériences d'échange isotopique. Des essais de cristallisation du domaine THAP ont également été menés au sein de notre institut dans l'équipe de L. Mourey qui n'ont à ce jour pas donné de cristaux de qualité suffisante; ceci peut également être dû à la présence de ces régions flexibles.

Liaison à l'ADN

Nous avons étudié la liaison du domaine THAP de THAP1 avec un oligonucléotide de 14 paires de bases, comprenant la séquence consensus spécifiquement reconnue par le domaine THAP de THAP1 (Clouaire et al., 2005). Pour étudier la spécificité d'interaction nous avons utilisé un deuxième oligonucléotide de séquence aléatoire comprenant également 14 paires de bases. Nous rappelons ici les deux séquences de ces ADN doubles brins.

ADN de séquence consensus : 5' CAAGTATGGGCAAG 3 '
3' GTTCATACCCGTTC 5'

ADN de séquence aléatoire : 5' GATTTGCATTTTAA 3'
3' CTAAACGTAAAATT 5'

Nous avons étudié la liaison du domaine à l'ADN en utilisant trois techniques différentes, la Résonance Magnétique Nucléaire (RMN), la fluorescence, la Résonance Plasmonique de Surface (RPS). Les études par RMN et RPS font partie de l'article publié, nous exposons ici des résultats complémentaires. Enfin nous

présentons les résultats préliminaires concernant la modélisation du complexe ADN-protéine avec le logiciel HADDOCK (van Dijk et al., 2006).
Nous avons déterminé les conditions de formation du complexe qui permettent d'éviter la précipitation de la protéine en contact avec l'ADN, et nous réalisons ainsi ces études avec une concentration en NaCl de 250 mM dans le tampon tris à pH 6,8 à la température de 296 K.

Etude de la liaison par RMN

Nous avons réalisé la titration du domaine THAP (1-90 et 1-81) marqué ^{15}N par l'ajout d'ADN consensus en enregistrant des spectres ^{15}N HSQC à différents rapports ADN/Protéine. Par comparaison des déplacements chimiques de la protéine libre et en complexe avec l'ADN, nous avons déterminé les acides aminés affectés par la liaison à l'ADN et ainsi cartographié une zone d'interaction. Ces résultats sont publiés dans l'article.
Nous avons démontré, à nouveau, avec cette technique, la spécificité de l'interaction au niveau de la séquence d'ADN en réalisant cette titration avec un ADN aléatoire qui n'a conduit à aucune variation significative des déplacements chimiques de la protéine.

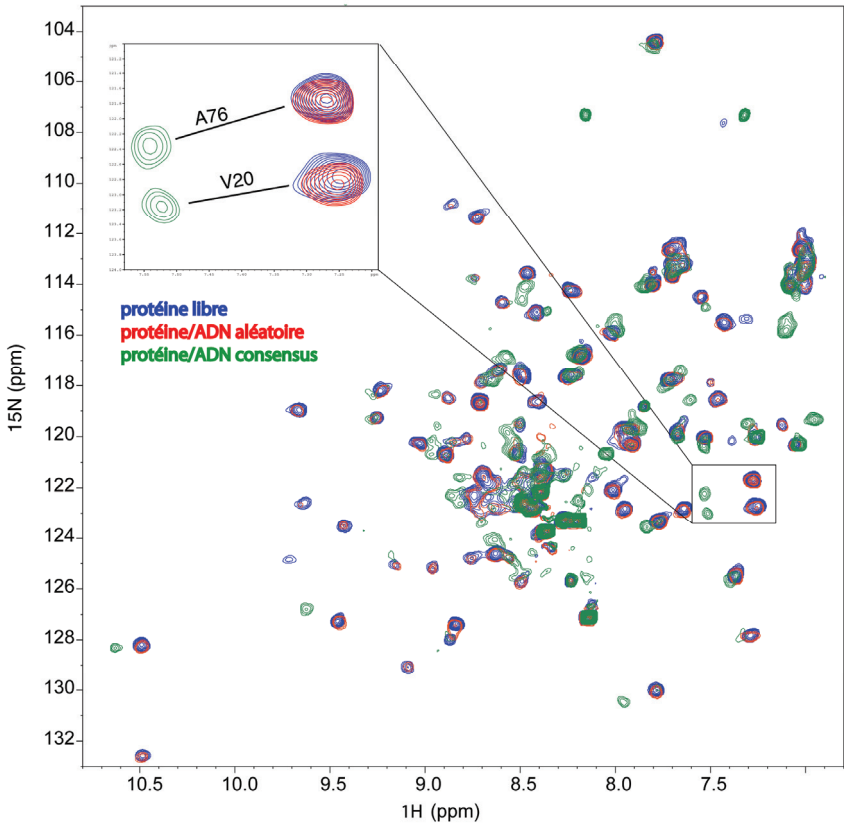


Figure 61 : **Spécificité d'interaction**
Superposition des spectres ^{15}N HSQC de la protéine libre (bleu) ou avec un équivalent d'ADN de séquence aléatoire (rouge) ou consensus (vert).

Au cours de la titration de la protéine par l'ADN spécifique nous observons plusieurs régimes d'échange. En effet, les composants du complexe sont en régime d'échange rapide ou lent selon l'échelle de variation de déplacement chimique qui dépend du résidu affecté et de sa variation de déplacement chimique (figure 62).

Pour une variation entre la forme libre et le complexe supérieur à 50 Hz, on observe au cours de la titration deux tâches de corrélation aux déplacements chimiques de la forme libre et liée avec des poids en intensité en accord avec la proportion des deux formes. Il s'agit d'échange lent.

Pour une variation entre la forme libre et le complexe supérieur à 50 Hz, on observe au cours de la titration une seule tâche de corrélation aux déplacements chimiques intermédiaires de la forme libre et liée. Il s'agit d'échange rapide.

Il existe un cas frontière entre ces deux régimes, la coalescence pour une variation égale à environ 50 Hz. On observe alors une seule tâche de corrélation au cours de la titration recouvrant les tâches des formes libre et liée.

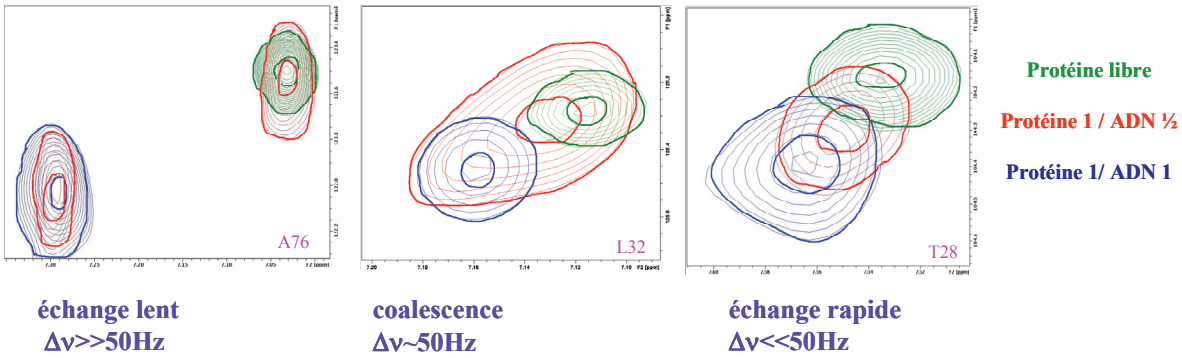


Figure 62 : **Différents régimes d'échange**
Représentation de taches de corrélation superposées en absence ou en présence d'un demi équivalent d'ADN et d'un équivalent d'ADN pour des résidus en échange lent, rapide ou à la coalescence.

Il est possible d'estimer la valeur de la durée de vie du complexe à partir de la relation suivante (Günther, 1995):

$$\tau = \sqrt{2}/(\pi.\Delta\nu)$$

Ainsi pour la coalescence observée à 50 Hz nous calculons un temps de vie du complexe de 9 ms.

Nous pouvons de même estimer la constante cinétique de dissociation (k_{off}) comme l'inverse de ce temps de vie à 111 s^{-1}. Nous utilisons les dénominations suivantes pour les constantes, où K_D est la constante de dissociation:

$$\text{THAP} + \text{ADN} \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} \text{THAP-ADN}$$

$$K_d = \frac{[\text{THAP}]\,[\text{ADN}]}{[\text{THAP-ADN}]} = \frac{k_{off}}{k_{on}} \qquad (1)$$

Si on considère que l'association est limitée par les phénomènes de diffusion, la valeur maximum de k_{on} se situe entre 10^7 et 10^9 M^{-1} S^{-1} (Emerson et al., 1995). Il s'agit là d'une approximation qui considère en particulier qu'aucun réarrangement structural n'a lieu lors de la formation du complexe. De même cette estimation standard ne prend pas en compte un éventuel mécanisme de collision/diffusion le

long de l'ADN susceptible d'augmenter k_{on}. A ces réserves près, ceci donne une constante de dissociation K_D comprise entre 111 nM et 11 µM.

Nous avons suivi la variation de déplacement chimique au cours de la titration, en fonction de la concentration en ADN ajouté. Nous avons donc réalisé ces mesures pour les résidus affectés en échange rapide, pour les résidus en échange lent il aurait fallu mesurer le volume des taches de corrélation correspondant aux formes libres et liées.

Nous avons représenté la courbe de variation relative de déplacements chimiques en fonction du ratio [ADN]/[protéine].

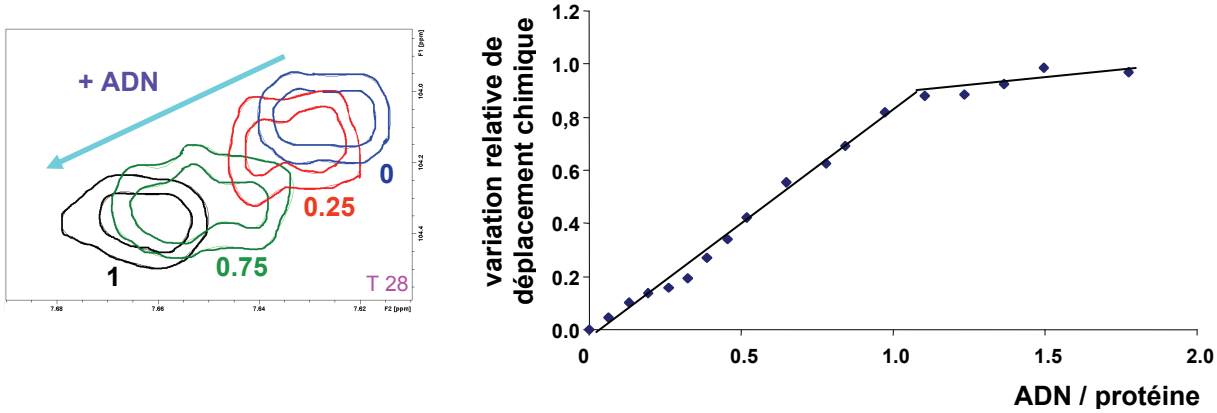


Figure 63 : **Titration par RMN : suivi des variations de déplacement chimiques**
Courbe de variation de déplacement chimique pour les résidus affectés en régime rapide un exemple est donné pour le résidu T28 dont les taches de corrélation sont représentées à différents équivalents d'ADN 0, 0.25, 0.75 et 1.

Toutefois il n'a pas été possible de déterminer la constante de dissociation (K_D) à partir de l'ajustement de ces valeurs avec un modèle de liaison *un pour un*, issu de la relation (1) (L'équation utilisée pour l'ajustement des données des variations de déplacement chimique peut être retrouvée dans la revue suivante (Fielding, 2007). En effet, les valeurs croissent linéairement pour atteindre un plateau au ratio 1:1. Cela s'explique par le fait que la concentration en protéine (~500 µM) est beaucoup plus forte que la valeur du K_D, ce qui revient à ce que tout l'ADN ajouté lors de la titration forme du complexe jusqu'à saturation et la courbe de titration n'est pas sensible au K_D dans ces conditions.

La seule information que l'on peut extraire de cette courbe est la stœchiométrie du complexe (1:1) qui est le ratio [ADN]/[protéine] à partir duquel on n'observe plus de variation de déplacement chimique. Ce ratio est défini par le rapport des concentrations en protéine et ADN déterminées par absorbance. Pour confirmer ce

rapport, nous avons mesuré le rapport d'intégration sur des spectres 1D ^{1}H du complexe d'un pic correspondant à un proton amide de la protéine et d'un pic correspondant à un proton imino de l'ADN.
Comme nous sommes limités par la concentration avec la technique de RMN (au minimum 100 µM avec une cryosonde) nous avons eu recours à la fluorescence puis à la RPS pour déterminer la valeur de la constante de dissociation.

Etude de la liaison par Fluorescence

Dans le but de déterminer une constante d'affinité, nous avons étudié la fluorescence intrinsèque du domaine THAP de THAP1 et réalisé une titration avec des concentrations croissantes d'ADN. Nous avons ainsi pu travailler avec des concentrations en protéine plus faibles de l'ordre du µM.
Nous avons étudié la fluorescence du tryptophane 36 du domaine THAP. Le fait que la séquence ne comporte qu'un seul tryptophane présente un avantage considérable pour l'analyse des données. De plus, nous savons que ce tryptophane est affecté par la liaison à l'ADN d'après les expériences de variations de déplacement chimique. Par contre, nous ne savions pas, lorsque nous avons entrepris ces expériences, si le tryptophane était orienté vers l'extérieur de la protéine et accessible ou s'il était dans le cœur de la protéine. Mis à part le tryptophane, seuls les acides aminées tyrosine et phénylalanine fluorescent. Les caractéristiques de ces fluorophores, en particulier les longueurs d'onde d'émission et d'excitation peuvent être retrouvés dans l'ouvrage suivant (Janin, 1985).
Nous avons réalisé les spectres d'émission et d'excitation du domaine THAP seul de façon à optimiser la longueur d'onde d'émission et d'excitation du tryptophane (figure 64)

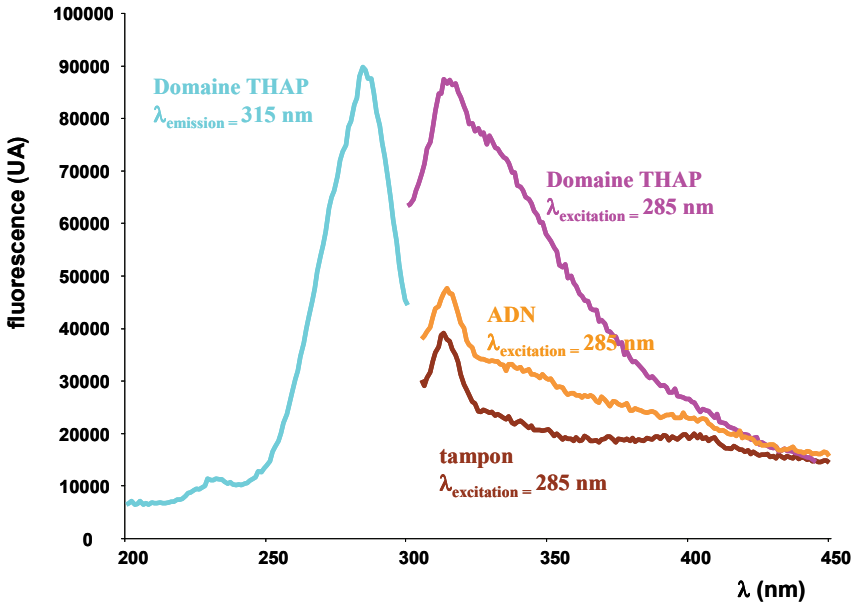


Figure 64 : **Spectres de fluorescence d'émission et d'excitation**

Spectres d'excitation (en bleu) et d'émission (en rose) du domaine THAP de THAP1 à 1µM, et spectres d'émission de l'ADN à 1µM (en orange) et du tampon (en marron). Les spectres d'émission se font à une longueur d'onde d'excitation de 285 nm et les spectres d'excitation à une longueur d'onde d'émission de 315 nm.

Nous avons ainsi déterminé une longueur d'onde de maximum d'excitation de 285 nm (cf spectre d'excitation) spécifique du tryptophane (la longueur d'onde maximum d'excitation classique d'une tyrosine étant de 275 nm et de 257 pour la phénylalanine contre 280 nm pour le tryptophane). A cette longueur d'onde les tyrosines sont encore excitées mais leur coefficient d'excitation est beaucoup plus faible que celui du tryptophane (Ross et al., 1997). Nous avons donc choisi d'utiliser cette longueur d'onde d'excitation de 285 nm pour nos expériences.

Nous avons déterminé une longueur d'onde de maximum d'émission de 315 nm (cf spectre d'émission) qui est spécifique d'un tryptophane dans un environnement apolaire (Janin, 1985). Ce résultat est en accord avec la structure que nous avons résolue où le tryptophane 36 est situé dans le cœur hydrophobe de la protéine.

Nous avons également réalisé les spectres d'émission du tampon utilisé et de l'ADN seul pour vérifier que le signal observé correspondait bien à la protéine. Il est connu que l'ADN ne possède pas d'activité de fluorescence (Roy, 2004).

Nous avons également mesuré l'intensité de fluorescence à 315 nm en fonction de la concentration en protéine et vérifié que ces deux grandeurs étaient bien proportionnelles (figure 66)

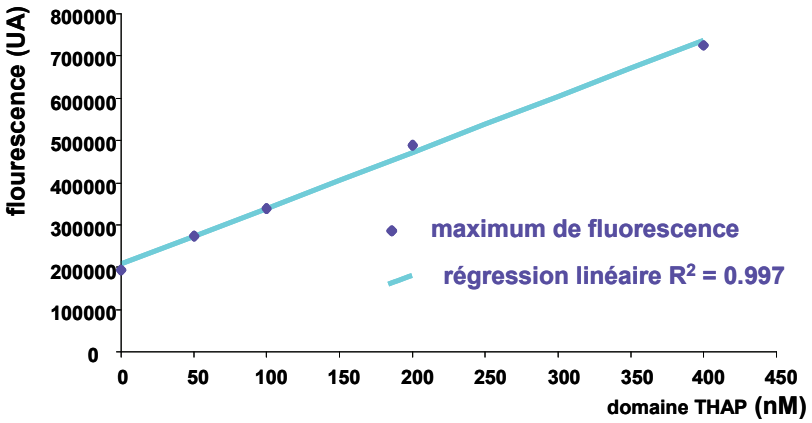


Figure 66 : **Fluorescence en fonction de la concentration en protéine**
Intensité de fluorescence à 315 nm (excitation à 285 nm) pour le domaine THAP à différentes concentrations.

Nous avons réalisé la titration du domaine THAP à 1 µM en ajoutant des concentrations croissantes d'ADN de séquence consensus et en observant l'intensité du maximum de fluorescence (figure 67).

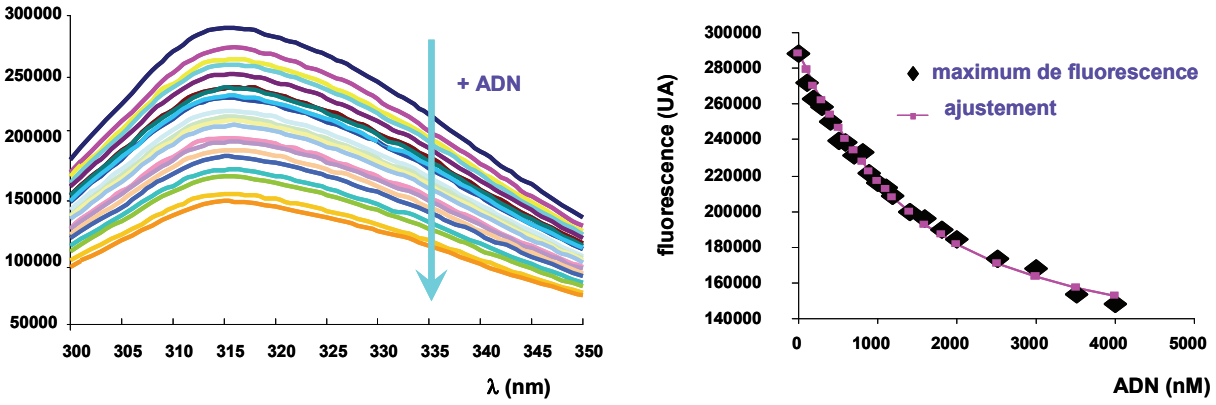


Figure 67 : **Titration par fluorescence intrinsèque du tryptophane 36**
A gauche : Spectres d'émission du domaine THAP à 1µM (longueur d'onde d'excitation 285 nm) avec des concentrations croissantes d'ADN consensus (de 0 à 4 µM). A droite : Intensité de fluorescence à 315 nm en fonction de la concentration en ADN et courbe d'ajustement.

Nous n'avons pas observé de variation de la longueur d'onde d'émission maximum mais une décroissance de l'intensité de fluorescence lorsque l'on rajoute de l'ADN. Nous avons ajusté la courbe de maximum d'intensité en fonction de la concentration en ADN avec un modèle de liaison *un pour un* dont l'équation peut être retrouvée dans la revue suivante (Roy, 2004). Nous avons ainsi déterminé une constante de dissociation de 600 ± 100 nM toujours dans les mêmes conditions, à 250 mM NaCl, 296K et à un pH de 6,8.

Nous avons également réalisé cette expérience avec l'ADN de séquence aléatoire et un ADN pour lequel le cœur GGCA est muté en TTTT (figure 4) (Clouaire et al., 2005). Nous avons trouvé des constantes de dissociation du même ordre de grandeur alors que ces séquences ne sont pas spécifiques de la liaison du domaine THAP de THAP1 à l'ADN. Nous remarquons également que l'intensité de fluorescence continue de baisser même lorsque la stœchiométrie 1:1 est atteinte. Nous savons de plus que le tryptophane n'est pas exposé à la surface de la protéine et n'est pas en interaction directe avec l'ADN.

Ces observations nous ont amenés à mettre en doute les conditions que nous avions mises au point pour cette étude. En effet, la baisse d'intensité de fluorescence pourrait être due à un effet de filtrage de l'ADN qui, même s'il n'émet pas de fluorescence, absorberait de la lumière à 285 nm ; ce qui engendrerait une baisse de la quantité de lumière pour l'excitation de la protéine et ainsi une baisse artéfactuelle de l'intensité de fluorescence émise. Ce phénomène est décrit pour des longueurs d'ondes de 280 nm même si l'ADN a une longueur d'onde d'absorbance maximum de 260 nm (Roy, 2004).

Toutefois, cet effet devrait être proportionnel à la concentration en ADN. Nous proposons donc une expérience supplémentaire qui consiste à réaliser la titration avec de l'AMP dans les mêmes proportions que l'ADN et regarder si l'on observe une baisse de l'intensité de fluorescence. Dans le cas d'une réponse linéaire, il serait alors judicieux de soustraire cette composante à nos courbes pour calculer alors une composante non spécifique. Une réponse comparable à celle que l'on a obtenue précédemment montrerait au contraire que l'on ne mesure pas un effet sur le tryptophane mais bien une réponse fonction de la quantité d'ADN.

Il pourrait également être intéressant de réaliser ces expériences de titration à une longueur d'onde de 295 nm à laquelle l'ADN n'absorbe plus et qui en plus serait plus spécifique du tryptophane par rapport aux tyrosines (Roy, 2004).

Suite aux expériences de fluorescence, nous avons réalisé des gels retards dans des conditions identiques (concentration en protéine de 1 µM, tampon Tris pH 6.8, NaCl 250 mM et sans compétiteur) et nous avons observé une liaison avec l'ADN de séquence consensus mais pas avec l'ADN de séquence aléatoire ou muté (GGCA → TTTT). Ce qui nous a amenés une fois de plus à remettre en cause nos mesures de K_D par fluorescence.

Etude de la liaison par RPS

Nous avons utilisé la technique de Résonance Plasmonique de Surface au sein de la plateforme d'interactions moléculaires (Institut Louis Bugnard, Toulouse) dont les résultats sont publiés dans l'article et présentés dans le chapitre précédent. Nous présentons ici des résultats complémentaires et des précisions d'informations.

Détermination de la constante de dissociation

La Résonance Plasmonique de Surface (RPS) est une méthode de mesure de la liaison entre deux partenaires, l'un immobilisé à la surface d'une couche métallique, l'autre injecté en flux à différentes concentrations. Le système de détection est basé sur une variation de l'indice de réfraction de l'interface quand les deux partenaires se fixent (pour revues : (Malmqvist, 1993; Schuck, 1997a; Schuck, 1997b)).
Nous avons ainsi choisi d'immobiliser nos deux sondes ADN consensus et aléatoire en utilisant une extrémité biotinylée et des surfaces greffées en streptavidines. Nous avons ainsi réalisé l'hybridation du double brin en maîtrisant la quantité de matériel immobilisé. Nous avons mesuré la réponse RPS en fonction de la concentration de protéine injectée sur l'ADN immobilisé. Nous avons pu travailler avec des gammes de concentration plus faibles qu'en RMN.
En outre, nous avons pu analyser la liaison spécifique en soustrayant le signal obtenu avec l'ADN consensus de celui de l'ADN aléatoire (réalisé en parallèle sur deux canaux différents avec les mêmes solutions de protéine injectées). Nous avons représenté les sensogrammes de réponse avec les deux séquences et la réponse différentielle calculée sur la figure 56. Nous avons travaillé avec 10 concentrations en protéine comprises entre 50 nM à 45 µM (cf article) toujours dans les mêmes conditions (296K, tampon Tris pH 6.8, NaCl 250mM).

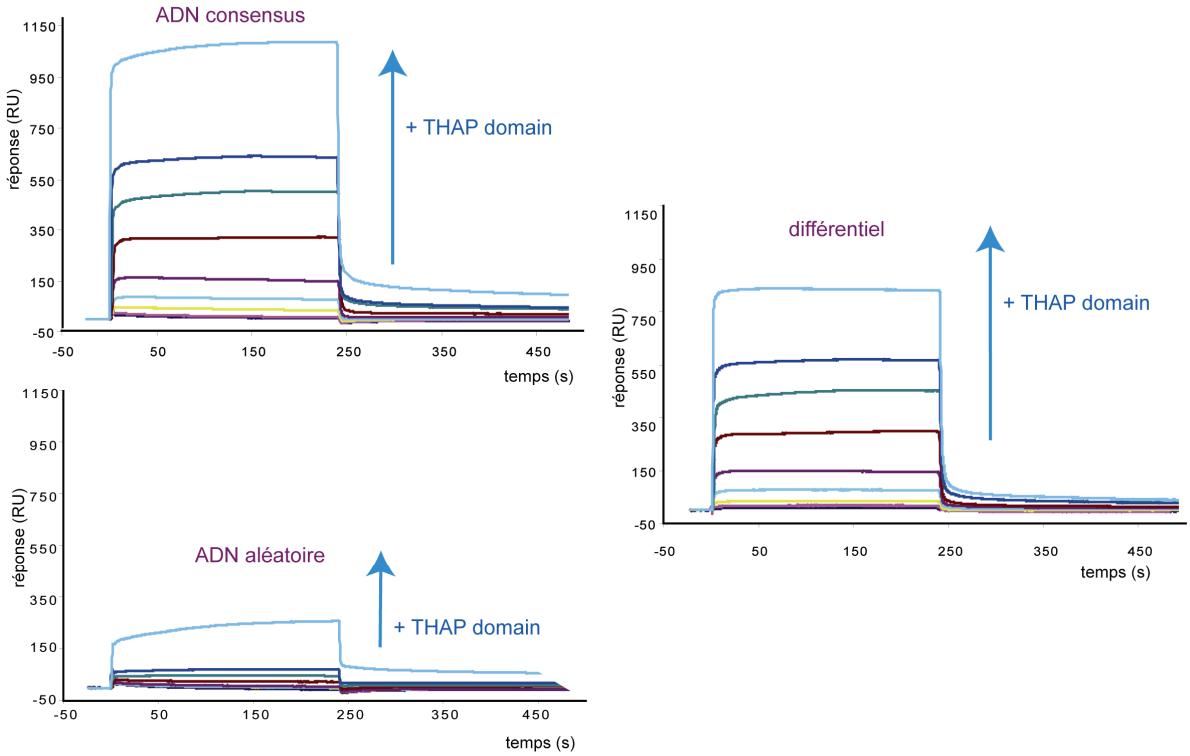


Figure 68 : **Sensogrammes de RPS sur les sondes aléatoires et consensus et réponse différentielle**

La réponse sur l'ADN aléatoire est beaucoup plus faible. De plus si on représente la réponse à l'équilibre (Réq, au plateau avant la dissociation) en fonction de la concentration en protéine (figure 69), on s'aperçoit que la réponse non-spécifique (ADN aléatoire) est proportionnelle à la concentration en protéine et qu'il n'est pas possible d'ajuster une valeur de K_D non-spécifique à ces valeurs.

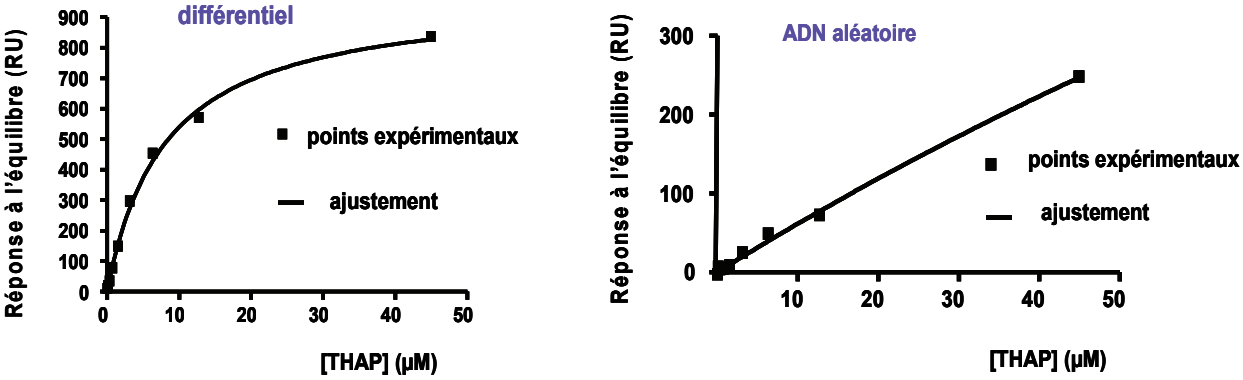


Figure 69 : **Ajustement de la réponse différentielle et de la réponse non-spécifique**
Les points expérimentaux de la réponse à l'équilibre ainsi que la solution d'ajustement sont représentés dans le cas des valeurs en différentiel et des valeurs pour l'ADN aléatoire.

Il est possible par contre de calculer une constante de dissociation globale en ajustant avec un modèle à un site les valeurs à l'équilibre trouvées pour la réponse différentielle en fonction de la concentration en ADN. Nous avons ainsi utilisé une équation de la forme suivante (Schuck, 1997b) :

$$Réq = \frac{R\max.C}{(KD + C)}$$

Avec Réq la réponse à l'équilibre en RU (unité de réponse), Rmax la valeur de Réq maximum et C la concentration en protéine. Les valeurs de Kd et Rmax ont été ajustées par la méthode des moindres carrés.

Nous avons pu confirmer la stoechiométrie 1:1 et mesurer une constante d'affinité de 8 ± 1 µM spécifique de l'ADN consensus par rapport à l'ADN aléatoire.

L'affinité mesurée est assez faible, on trouve classiquement des valeurs de l'ordre de 100 nM pour des interactions spécifiques mesurées par RPS de doigt de zinc avec un ADN spécifique (Yamaguchi et al., 2006; Yamasaki et al., 2004a).

Cette affinité reste néanmoins spécifique et comme il s'agit d'une affinité globale nous n'excluons pas que le K_D puisse être décomposé en plusieurs composantes de valeurs différentes comme nous le suggèrent les études cinétiques.

Il est probable de plus que cette affinité soit augmentée *in vivo* par des phénomènes d'interaction protéine-protéine ou de dimérisation de la protéine. En particulier les protéines THAP humaine présente dans leur partie C-terminale un domaine coiled-coil (résidus 142 à 190 pour THAP1), connu pour son implication dans les interactions protéine-protéine (Burkhard et al., 2001 ; Lupas, 1996).

La RPS permet en effet l'étude de la cinétique de liaison puisque nous mesurons un signal en fonction du temps. Il est ainsi possible d'ajuster les phases d'association, d'équilibre et de dissociation pour mesurer les constantes cinétiques k_{off} et k_{on} et thermodynamique K_D. Les différentes équations sont retrouvées dans la revue (Schuck, 1997b) et de façon plus détaillée sur le site *www.sprpages.nl* .

Nous avons essayé plusieurs modèles, du modèle le plus simple (un pour un), jusqu'à des modèles avec deux voire trois constantes de dissociation, incluant plusieurs étapes d'association et de dissociation soit de façon parallèle soit de façon continue. Nous avons également pris en compte le transfert de masse qui consiste à un transfert préalable de la protéine à la surface avant la liaison à l'ADN. Tous ces

essais ne nous ont pas permis d'obtenir un ajustement convenable pour ces mécanismes.

L'allure des sensogrammes est d'ailleurs atypique, les phases d'association et de dissociation sont très rapides quelle que soit la concentration en protéine et sont multi-exponentielles. Cette allure de courbe peut traduire une cinétique particulière avec des intermédiaires. Toutefois, le domaine de liaison à l'ADN de SATB1 (Yamaguchi et al., 2006) présente des sensogrammes de même allure par opposition par exemple à des allures plus classiques pour le domaine de liaison à l'ADN de RAV1 (Yamasaki et al., 2004b) (figure 70). Dans ces deux cas, des constantes de dissociation globale de l'ordre de 100 nM ont été déterminées mais la cinétique n'a pas été étudiée.

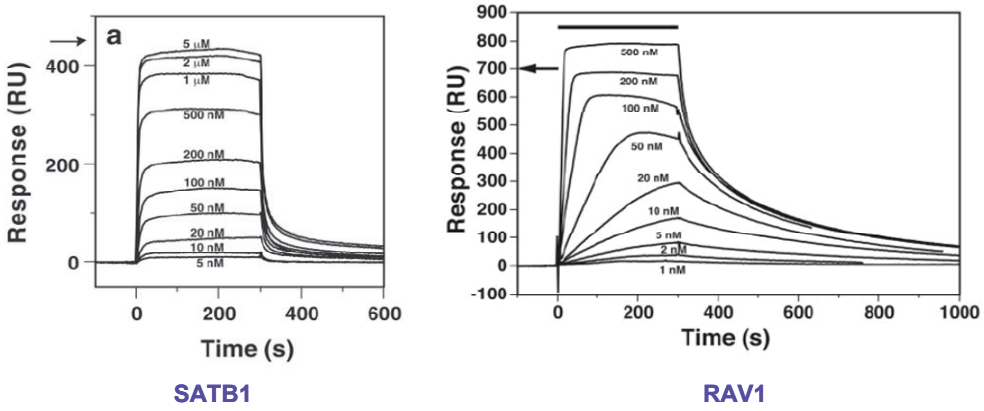


Figure 70 : **Sensogrammes pour les domaines de liaison à l'ADN de SATB1 et RAV1**

Influence de la force ionique

Nous avons également étudié l'influence de la force ionique par RPS en enregistrant des sensogrammes en différentiels avec le domaine THAP concentré à 3,2 µM pour des concentrations en NaCl de 250 mM à 500 mM.

Nous avons observé une diminution de la réponse avec l'augmentation en sel jusqu'à une perte d'interaction à haute force ionique (figure 71)

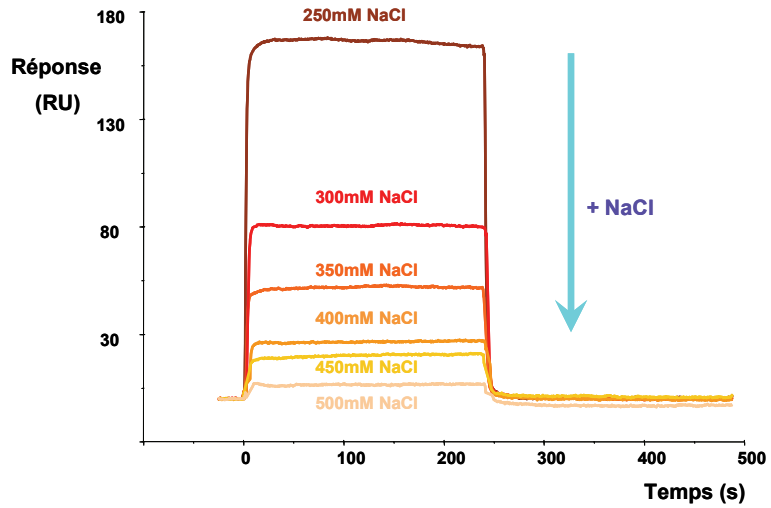


Figure 71 **: Influence de la force ionique**

Sensogrammes de RPS avec des concentrations croissantes en NaCl pour une concentration en protéine de 3,2 µM.

Cet effet de la force ionique est à mettre en parallèle avec la précipitation à basse force ionique pouvant traduire une interaction plus forte. C'est effet n'est pas surprenant étant donné la charge positive de la protéine interagissant avec l'ADN chargé négativement. Toutefois cette reconnaissance de charge est souvent associée à la composante non-spécifique (Kalodimos et al., 2004b) mais semble dans notre cas participer à la reconnaissance spécifique.

Résidus impliqués dans l'interaction avec l'ADN

Pour la compréhension du mode de reconnaissance, nous avons donc identifié des résidus impliqués dans l'interaction avec l'ADN suivant deux approches. Nous avons détecté les résidus affectés par la liaison à l'ADN par variation de déplacement chimique et nous avons étudié la liaison à l'ADN de mutant du domaine sauvage. Ces données peuvent être retrouvées sur les histogrammes suivants (figure 73) représentant la variation de déplacement chimique en fonction de la séquence (fig 5 B de l'article) et l'effet de la mutation ponctuelle d'un résidu en alanine sur la liaison à l'ADN par gel retard (table 2 de l'article).

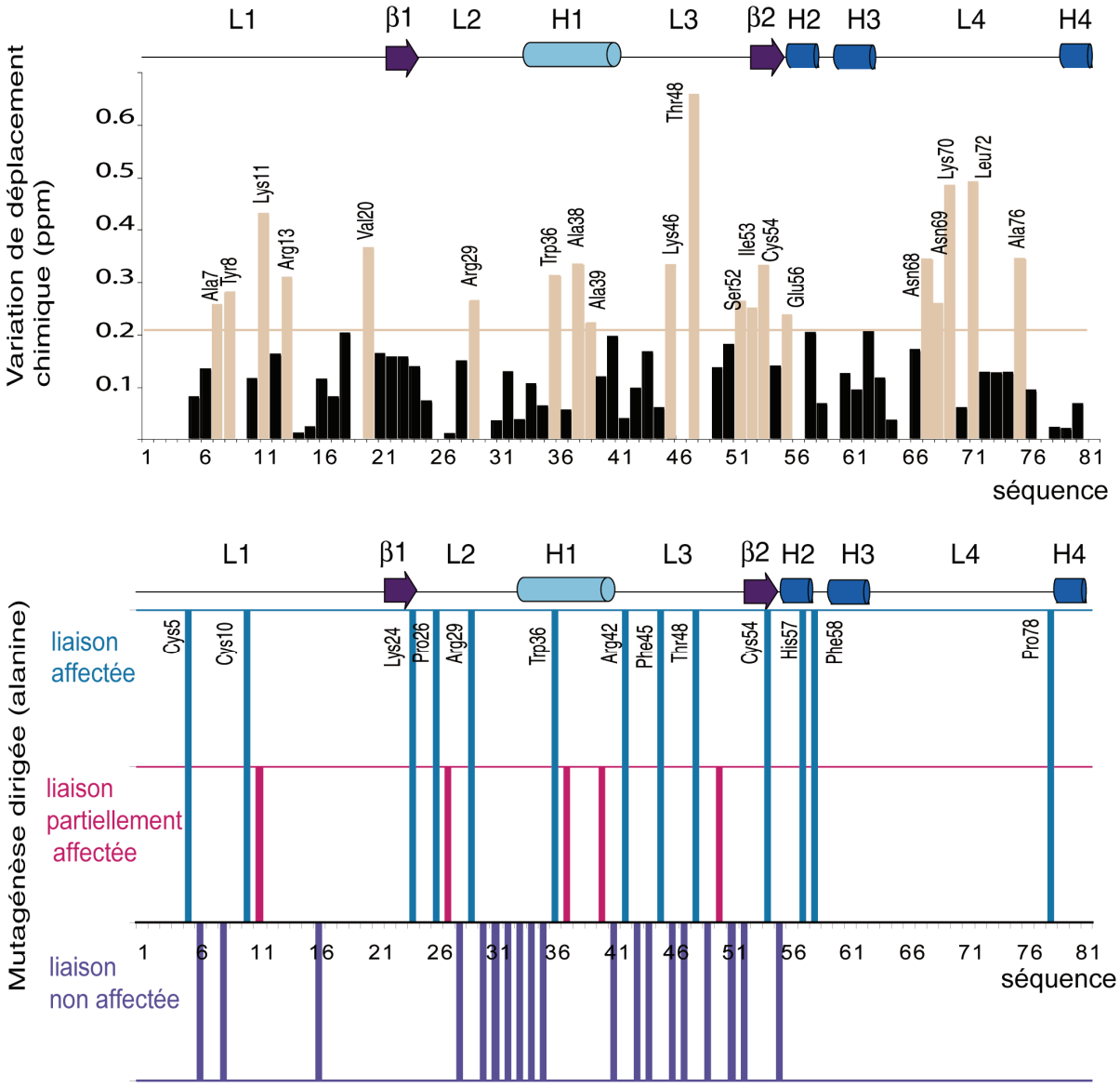


Figure 73 : **Identification des résidus affectés dans la liaison à l'ADN**
Représentation en histogramme, selon la séquence, de la variation de déplacement chimique lors de l'ajout de l'ADN et de l'effet de la mutation ponctuelle d'un résidu en alanine sur la liaison observé sur gel retard.

Les expériences de variation de déplacement chimique nous ont permis d'identifier trois régions qui subissent les plus grandes variations de déplacement chimique (fig 5 de l'article). Nous détectons par cette méthode les groupes NH qui voient leur environnement changé au cours de la titration. (Nous n'avons pas d'information sur les prolines). Nous détectons donc en théorie des résidus en interaction directe avec l'ADN mais également des résidus sur des régions subissant des changements conformationnels ou à proximité. En pratique il est important d'identifier les résidus accessibles à la surface de la protéine qui pourraient rentrer en contact avec l'ADN.

Par exemple le tryptophane 36 est affecté par la liaison à l'ADN mais nous savons qu'il est enfoui au cœur de la protéine, il ne peut donc vraisemblablement pas être en interaction avec l'ADN. Nous comprenons assez bien d'autre part qu'il soit affecté étant donné son rôle clef dans le maintien de la structure. D'autre part il est possible que des résidus en interaction avec l'ADN voient globalement le même environnement dans la protéine seule ou complexée. Il n'est donc pas impossible que certains résidus en interaction ne soient pas identifiés par cette technique, d'autant plus que nous observons des déplacements chimiques ^{15}N et ^{1}H des groupes NH et donc pas des chaînes latérales qui sont en interaction directe avec l'ADN. Il est possible de réaliser le même genre d'expérience sur des spectres ^{1}H (ou ^{13}C) de façon à observer les variations de déplacement chimiques des chaînes latérales (Schmiedeskamp et al., 1997) mais l'observation est plus délicate que sur une expérience ^{15}N HSQC classiquement utilisée.

Nous remarquons de plus que beaucoup de résidus sont affectés, nous avons fixé une valeur seuil à 0.2 ppm pour distinguer les variations significatives. Une variation globale des déplacements chimiques peut indiquer une restructuration de la protéine. Dans notre cas, même si nous n'excluons pas que certaines régions puissent changer de structure au contact de l'ADN, nous pensons que la protéine garde globalement la même structuration. Les corrélations sur l'expérience HSQC peuvent être suivies lors de la titration et l'analyse préliminaire des NOEs sur le complexe indiquent la conservation des structures secondaires lors de l'ajout d'ADN. Le nombre important de variations de déplacement chimique peut être dû à l'effet de l'ADN et des courants de cycle nombreux, d'autant plus que la molécule d'ADN est de taille comparable à celle de la protéine (en comparaison à des expériences de variation de déplacement chimique lors de l'ajout d'une molécule organique de petite taille).

Il a donc été, pour ces raisons, particulièrement intéressant de confronter ces données aux données de mutagénèse dirigée. Les expériences de mutagénèse dirigée nous ont permis d'identifier les résidus qui affectent la liaison à l'ADN lorsqu'ils sont mutés an alanine. Nous explorons alors la fonctionnalité des chaînes latérales des résidus. Les résidus identifiés peuvent être impliqués directement dans la reconnaissance à l'ADN mais peuvent aussi être impliqués dans la déstructuration de la protéine qui mène au même résultat observé sur gel retard. Il est donc important, au vu de la structure, d'observer si les résidus mutés ont un rôle structural.

C'est le cas par exemple pour le tryptophane 36 ou les résidus liant le zinc (CCCH). Nous pouvons ainsi confronter de façon intéressante ces données de mutagénèse à nos données de structure; nous avons pour autre exemple identifié les résidus 45 et 29 par mutagénèse dirigée qui ont un rôle structural dans la stabilisation de boucles (fig 4 E de l'article).

Nous identifions pour la compréhension du mode de reconnaissance protéine-ADN les résidus dont la mutation affecte la liaison et qui sont accessibles à la surface de la protéine. Ainsi les résidus 24, 42 et 48 semblent être directement impliqués dans la liaison à l'ADN.

Nous complétons actuellement ces données par des expériences de transfert de saturation (Lane et al., 2001; Ramos et al., 2000). Elles consistent en la saturation des protons iminos de l'ADN et à l'observation de transfert de saturation sur les protons amides de la protéine sur un spectre ^{15}N HSQC.

Un mode de reconnaissance original

Nous avons pu définir, d'après les expériences de mutagénèse dirigée et de RMN, une surface d'interaction du domaine THAP de THAP1 avec l'ADN consensus. Cette surface correspond à la surface chargée positivement de la protéine (fig 4 F de l'article). L'importance des interactions électrostatiques, et de cette face chargée positivement, est d'ailleurs confirmée par l'étude de la liaison en fonction du sel par RPS.

Nous avons déterminé trois résidus essentiels à la liaison exposés sur cette surface : la lysine 24, l'arginine 42 et la thréonine 48. Le résidu 48 nous apparaît intéressant puisqu'il montre une grande variation de déplacement chimique lors de l'ajout d'ADN et qu'il n'est pas conservé parmi les domaines THAP, suggérant un rôle clef dans la reconnaissance spécifique.

Nous avons vu que la stratégie la plus connue et sûrement la plus utilisée pour la reconnaissance spécifique de l'ADN par une protéine impliquait l'interaction d'une hélice dans le grand sillon. Même si le domaine THAP présente une hélice α, Il apparaît clairement qu'elle ne constitue pas l'élément de reconnaissance principal de l'ADN. En effet, la mutation des résidus de l'hélice, mis à part le tryptophane 36 impliqué dans les interactions hydrophobes du cœur de la protéine, n'affecte pas la liaison à l'ADN. C'est en ce sens que nous pensons que le domaine THAP de

THAP1 se lie à l'ADN avec un mode de reconnaissance original pouvant faire intervenir des boucles, comme cela été observé pour d'autres doigt de zinc. En particulier, les boucles L3 et L4 présentent des variations de déplacement chimique lors de l'ajout d'ADN (cf article) et correspondent à la surface positive de la protéine et pourraient participer à la reconnaissance de l'ADN. Des mutations sur la boucle L3 affectent la liaison à l'ADN ; l'effet des mutations sur la boucle L4 est actuellement à l'étude.

Cette surface d'interaction que nous avons déterminée est relativement grande. Cela peut s'expliquer par la grandeur de l'ADN consensus reconnu spécifiquement. En effet, le domaine THAP de THAP1 reconnaît une séquence de 11 nucléotides (Clouaire et al., 2005) contre seulement 3 pour le doigt de zinc classique. Selon ce point de vue, une courbure ou une déformation de l'ADN dans le complexe serait à prendre en compte pour satisfaire la complémentarité des deux partenaires dans un complexe ADN-protéine tel que nous l'avons vu (figure 22).

Modélisation du complexe THAP-ADN

Afin de proposer un modèle d'interaction THAP-THABS, nous nous sommes tournés vers le logiciel Haddock (High Ambiguity Driven protein-protein Docking) dont l'approche utilise des données biochimiques et biophysiques qui sont introduites dans le calcul en tant que contraintes d'interaction ambiguës (AIR) pour guider le *docking* de molécules (Dominguez et al., 2003; van Dijk et al., 2006). Dans cette approche, nous avons utilisé les données de variation de déplacement chimique et de mutagénèse dirigée. Ce calcul par *docking* nécessite la connaissance des structures des deux molécules de départ, en supposant qu'elles ne subissent pas de réarrangements structuraux lors de la formation du complexe, ainsi que l'identification des résidus impliqués dans l'interaction.

Le protocole comprend trois étapes. Une première étape consiste en la génération d'un ensemble d'orientations aléatoires et d'une minimisation d'énergie en corps rigide. La deuxième étape consiste à réaliser un calcul par recuit simulé semi-rigide; le squelette est rigide, seules les chaînes latérales peuvent bouger mais également des petites régions peuvent être définies comme totalement flexibles. La dernière étape est un affinement dans l'eau.

Nous présentons ici quelques résultats issus des travaux réalisés par S. Campagne, doctorant dans l'équipe. La structure 3D du domaine THAP de THAP1 telle que

nous l'avons résolue par RMN a été utilisée comme structure de départ ainsi qu'une structure modélisée de l'ADN sous forme B.

Les résidus les plus affectés par l'interaction mis en évidence par les expériences de variation de déplacement chimique (nous avons fixé comme valeur seuil 0.2 ppm) et par les expériences de mutagénèse dirigée ont été utilisés. Un deuxième filtre est ensuite appliqué en retenant les résidus accessibles à la surface de la protéine déterminés à l'aide du programme NACCESS (Atomic solvent accessible areas) (Hubbard and Thornton, 1993) avec une valeur seuil de 40% (figure 74). Les résidus ainsi déterminés forment les résidus actifs de la protéine.

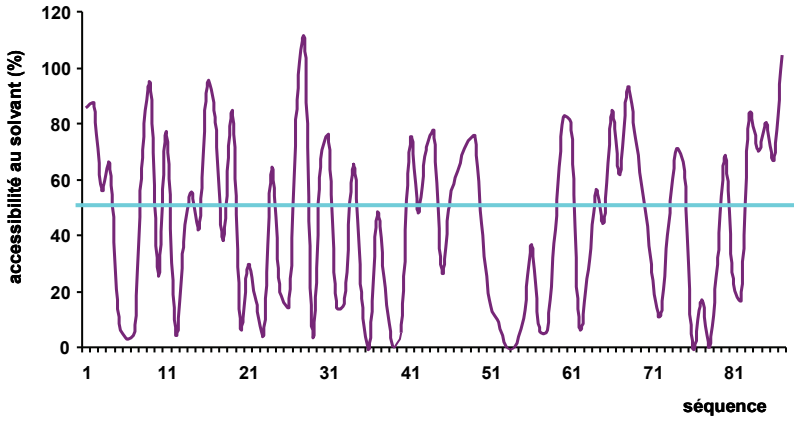


Figure 74 : **Accessibilité au solvant du domaine THAP calculé avec NACCESS**

Pour la molécule d'ADN les résidus actifs sont formés par le cœur GGCA et la base T à quatre positions en 5', dont l'importance a été évaluée par SELEX (figure 3) et mutagénèse dirigée. Les résidus proches des résidus actifs et accessibles forment les résidus passifs. Le principe du calcul avec HADDOCK se fonde sur l'emploi de contraintes d'interactions ambiguës définies entre ces résidus actifs et passifs (Dominguez et al., 2003).

Une structure du modèle ainsi construit est présentée dans l'article (fig 6). Elle a été réalisée avec les résidus 24, 37, 42, 46 et 48 définis comme actifs et les résidus 34, 35, 38, 41, 43, 44, 49 et 50 définis comme passifs.

Le modèle présente une bonne complémentarité de forme des deux partenaires en particulier du motif boucle-hélice-boucle (L2-H1-L3). De plus, l'hélice α ne semble pas être l'élément majeur de la reconnaissance protéine-ADN.

Nous complétons actuellement nos données par la recherche de NOEs intermoléculaires et par l'identification de résidus à la surface d'interaction par des

expériences de transfert de saturation. Cela devrait permettre d'observer plus finement les résidus impliqués dans la liaison à l'ADN. En effet, un des problèmes apparu est la complexité de la répartition des résidus affectés par variation de déplacement chimique. Cela est certainement dû à la grandeur de l'ADN par rapport à la protéine et des courants de cycles des bases de l'ADN affectant l'ensemble des résidus de la protéine.

Nous envisageons également de réaliser les expériences d'échange isotopique proton/deutérium sur le complexe ainsi que l'étude de la dynamique du complexe pour compléter nos données.

De plus, nous allons réaliser des calculs avec de l'ADN courbé, comme on le retrouve dans des complexes protéine-ADN (figure 24). Nous cherchons d'ailleurs à évaluer un changement de conformation de l'ADN lors de la formation du complexe, par des mesures de variations de déplacement chimiques ^{31}P des groupes phosphates de l'ADN entre les formes libre et complexée; en particulier une dispersion des résonances ^{31}P lors de l'ajout d'ADN (Castagne et al., 2000).

Enfin, nous n'excluons pas la possibilité qu'il soit difficile de modéliser ce complexe du fait de changements structuraux importants ou encore d'un mécanisme de reconnaissance faisant intervenir des intermédiaires, comme cela peut être suggéré par la complexité de la cinétique de liaison observé par SPR. En effet, la présence d'intermédiaire nous amène peut-être à identifier des résidus qui ne sont pas impliqués dans la structure du complexe final mais dans la structure d'états intermédiaires. Enfin, nous devons toujours prendre en considération le fait que l'identification des résidus par mutagénèse dirigé peut être due à l'obtention d'une protéine déstructurée plutôt qu'à la perte réelle de liaison observée par gel retard. Nous envisageons à ce propos d'étudier par RMN les mutants clefs déterminés, comme le mutant sur la thréonine 48. De plus, il est possible que la méthode de détection par gel retard ne soit pas assez sensible pour identifier tous les résidus clefs, par exemple la lysine 46 qui n'est pas détectée par les expériences de mutagénèse semble avoir un rôle important dans la liaison à l'ADN d'après le modèle du complexe et ses variations de déplacement chimique.

Parallèlement aux efforts entrepris dans l'équipe de RMN, des essais de diffraction des RX sont conduits dans l'équipe de L. Mourey à l'IPBS. Ces études ont permis d'obtenir des cristaux du complexe protéine–ADN qui diffractent à environ 3 Å de résolution. Malheureusement à ce jour le problème des phases n'a pas encore pu

être résolu. Il est possible que l'utilisation d'un modèle plus affiné à partir des données de RMN permette à terme de résoudre le problème des phases par remplacement isomorphe, fournissant ainsi une structure à haute résolution du complexe.

Une fonction différente associé à chaque domaine THAP

Si nous avons confirmé par RMN et RPS la spécificité au niveau de la séquence consensus reconnu par le domaine THAP de THAP1 (Clouaire et al., 2005), nous avons également montré dans le cadre de cette étude que cette séquence n'était pas reconnue spécifiquement par d'autres domaines THAP (CtBP, THAP2, THAP3 et GON-14). De plus le domaine THAP de l'élément P de la drosophile reconnaît une séquence A-T riche différente de la séquence reconnue par THAP (Kaufman et al., 1989). Ces résultats suggèrent que chaque domaine THAP ne reconnaît pas la même séquence d'ADN et serait donc impliqué dans des fonctions biologiques propres. De plus, il n'est pas exclu que certains domaines THAP puissent être associés à des fonctions de reconnaissance d'ARN ou même de protéines. C'est pourquoi il est intéressant d'étudier les différents membres de cette famille, même entre domaines de séquence proche, comme entre THAP1 et THAP2.

CONCLUSIONS ET PERSPECTIVES

Nous avons, dans le cadre de nos travaux, déterminé la structure du domaine THAP de THAP1 définissant la famille des protéines THAP. Il s'agit de la première structure d'un domaine doigt de zinc THAP pour lequel une activité biochimique et des fonctions biologiques associées ont été déterminées. Le domaine THAP de THAP1 est en effet un domaine de liaison à l'ADN séquence spécifique et impliqué dans la régulation du cycle cellulaire dans la voie pRb/E2F et dans la prolifération cellulaire.
Nous avons déterminé la structure de ce domaine par Résonance Magnétique Nucléaire. La structure de ce domaine, en accord avec les deux autres structures de domaines THAP connues, se caractérise par un repliement $\beta\alpha\beta$ et une coordination au zinc par le motif C2CH. Cette structure est originale et définit un nouveau motif de doigt de zinc.
Seules trois structures de domaine THAP sont connues, la résolution de structures de domaines THAP d'autres protéines de la famille peut être envisagée. En effet chaque domaine THAP est lié à une fonction particulière. L'étude des variations structurales entre les différents domaines THAP et des relations structure-activité semble d'un grand intérêt pour la compréhension et la caractérisation de ce domaine et de la famille THAP.
Nous avons également étudié la liaison à l'ADN du domaine THAP de THAP1 à sa séquence consensus cible. Nous avons confirmé la spécificité de reconnaissance vis-à-vis de la séquence ADN par RMN et déterminé une constante de dissociation pour ce complexe protéine-ADN. Des études RPS complémentaires pourraient être réalisées avec des oligonucléotides de séquences différentes, comme celles utilisées par Clouaire et ses collaborateurs (Clouaire et al., 2005) de façon à quantifier de manière plus fine l'importance des bases de la séquence consensus identifiées par SELEX. Cette étude pourrait être intéressante dans la compréhension de la reconnaissance spécifique et non-spécifique de certaines séquences, d'autant plus que la séquence ADN consensus est relativement grande (11 nucléotides). La détermination d'une composante non-spécifique par RPS n'a d'ailleurs pas été possible. Il pourrait être intéressant de confronter ces mesures d'affinité déterminées par RPS avec d'autres mesures utilisant d'autres techniques, comme l'anisotropie de fluorescence (Heyduk and Lee, 1990) ou la calorimétrie de titration isotherme (*Isothermal Tiration Calorimetry*) (Fisher and Singh, 1995). Il pourrait également être

intéressant de travailler avec des ADN de séquence plus grande et présentant des répétitions de la séquence reconnue.

Dans le cadre de nos travaux, nous avons cherché à comprendre le mécanisme de reconnaissance THAP-ADN. Nous avons ainsi déterminé des résidus participant à la liaison à l'ADN et déterminé une surface d'interaction. Nous modélisons actuellement le complexe protéine-ADN à l'aide de ces données.

La résolution de la structure du complexe est évidemment une étape importante pour la compréhension du mode de reconnaissance de l'ADN par le domaine THAP de THAP 1. Des essais de cristallisation du complexe par l'équipe de cristallographie de notre institut sont entrepris mais sans résultats encourageants. Toutefois, l'utilisation de la nouvelle construction 1-81, correspondant à un domaine défini structuralement, est une piste actuellement à l'étude. Nous envisageons également de résoudre la structure de ce complexe par RMN ; des spectres NOESY ont été enregistrés, mais leur exploitation est très compliquée devant les nombreux recouvrements de fréquence que présente la protéine complexée. Il est donc important d'essayer d'améliorer les conditions de formation du complexe pour obtenir des données de meilleure qualité. Nous avons déjà travaillé sur ce sujet pour éviter la précipitation de la protéine. Nous poursuivons ce travail en étudiant la complexation de mutants de la protéine. Un travail est actuellement réalisé dans l'équipe sur la mutation des cystéines libres de la protéine (c'est-à-dire qui ne sont pas liées au zinc) en sérines. Il est en effet possible d'imaginer que la protéine, au moins en partie, polymérise par des ponts disulfures lorsqu'elle est complexée, engendrant une dégradation des spectres RMN.

D'autre part, la modélisation du complexe doit être améliorée en utilisant des données supplémentaires. Des expériences de transfert de saturation sont actuellement en cours, elle devrait permettre d'identifier les résidus directement impliqués dans la liaison à l'ADN. Nous projetons également d'étudier la dynamique du complexe et de la comparer à celle du domaine seul et également des expériences d'échange isotopique proton deutérium sur le complexe. Enfin il est également possible d'utiliser des données de couplages dipolaires résiduels pour améliorer la structure du complexe ou de son modèle.

Une étude supplémentaire intéressante à mener est l'étude des mutants qui ont été déterminés, par RMN et RPS. L'étude par RMN de ^{15}N HSQC permettrait de savoir si le mutant correspond à une structure correcte de la protéine. La mesure d'affinité des

mutants permettrait de quantifier l'importance des résidus. L'étude du mutant de la thréonine 48 est particulièrement intéressante puisque ce résidu est peu conservé et impliqué dans la liaison à l'ADN. Les mutants des résidus 29 et 45 conduisent à une forme non structurée de la protéine sans affinité pour l'ADN puisque nous avons pressenti un rôle structural pour ces résidus (fig 4 E de l'article).

Enfin, nos travaux font partie d'un projet de recherche fondamental sur le domaine THAP qui devrait permettre de mieux définir les fonctions des protéines THAP et participer à une meilleure connaissance des mécanismes moléculaires contrôlant la prolifération et le cycle cellulaire. Dans des perspectives à plus long terme, la conception de ligands à partir de la structure de ce domaine pourrait avoir des applications thérapeutiques importantes.

LISTE DES FIGURES

ABREVIATIONS

ADN	Acide DésoxiriboNucléique
AMP	Adénosine MonoPhosphate
ARIA	Ambiguous Restraints for Iterative Assignment
ARN	Acide RiboNucléique
BMRB	Biological Magnetic Resonance data Bank
BUSI	BUll Seminal Inhibitor
bZIP	basic leucine ZIPper
CCPNMR	Collaborative Computing Project for NMR
ChiP	Chromatin immunoPrecipitation
CNS	Crystallography & NMR System
COSY	COrrelated SpectrscopY
CSI	Chemical Shift Index
CtBP	C-terminal Binding Protein
DALI	Distance matrix ALIgnment
DAP4	Death Associated Protein 4
DO	Densité Optique
DTT	DiThioThreitol
EDTA	Acide Ethylène-Diamine-Tetraacétique
EKLF	Erythroid Krüppel Like Factor
EXPASY	Expert Protein Analysis System
FOG	Friend Of GATA
G1/S	Gap 1/ Synthese
GFP	Green Fluorescent Proteine
HADDOCK	High Ambiguity Driven biomolecular DOCKing
HDAC3	Histone DeACetylase 3
HIV	Human Immunodeficiency Virus
HSQC	Heteronuclear Single Quantum Correlation
IPTG	IsoPropyl-β-D-ThioGalacopyrasonide
Lin	abnormal cell LINeage
MST1	Mitochondrial aminoacyl-tRNA Synthetase Threonine 1
NCoR	Nuclear receptor CoRepressor
NCp7	Nucleocapsid protein 7
NF	Nuclear Factor
NOE	Nuclear Overhausser Effect
NOESY	Nuclear Overhauser Effect SpectroscopY
PML-NBS	Promyelocytic leukemia Nuclear Bodies

pRb	Protéine du RétinoBlastome
Rb	RétinoBlastome
RMN	Résonance Magnétique Nucléaire
Rmsd	Root Mean Square Deviation
RNAi	RNA interférence
RPS	Résonance Plasmonique de Surface
RRM1	ribonucleotide reductase M1
Par4	Prostate Apoptosis Response gene-4
PDB	Protein Data Bank
RU	Response Unit
SDS PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SELEX	Selective Evolution of Ligands by EXponential enrichment
Sp1	Specificity Protein 1
TBP	TATA Binding Protein
TAF-Iβ	Template Activating Factor-I β
TALOS	Torsion Angle Likelihood Obtained from Shift and sequence similarity
TF	Transcription Factor
THABS	THAP1 Binding Sequence
THAP	THanatos Associated Protein
TNFα	Tumosis Necrophile Factor alpha
TOCSY	TOtal Correlated SpectroscopY
VEGF	Vascular Endothelial Growth Factor
ZAD	Zinc finger Associated Domain

CODES D'ACCES

Déplacements chimiques

Les valeurs des déplacements chimiques ont été déposées dans la BMRB (Biological Magnetic Resonance Data Bank). Les attributions des fréquences ^{15}N ^{1}H et ^{13}C du domaine {Met1-Lys90} ont été déposées dans la BMRB sous le code d'accès 15300 et les fréquences ^{1}H et ^{15}N du domaine {Met1-Phe81} sous le code d'accès 15289. Ces attributions ont été réalisés à 296K, à pH 6.8 et dans un tampon Tris 50 mM, NaCl 10mM, DTT 1 mM, 0.01% NaN_3.

Attribution des fréquences ^{15}N ^{1}H et ^{13}C du domaine THAP 1-90

D. Bessiere, S. Campagne, A. Milon and V. Gervais *1H, 15N, 13C chemical shift assignment of the THAP domain 1-90 from human THAP1 protein*.

 # BMRB 15300

Attribution des fréquences ^{15}N et ^{1}H du domaine THAP 1-81

D. Bessiere, S. Campagne, A. Milon and V. Gervais 1H*, 15N chemical shift assignment of the THAP domain 1-81 from human THAP1 protein*.

#BMRB 15289

Coordonnées de structure

Le fichier PDB (Protein Data Bank) du domaine THAP 1-81 a été déposé sous le code 2jtg.

D. Bessiere, S. Campagne, A. Milon and V. Gervais *solution strucuture of the THAP domain 1-81 from human THAP1 protein*.

 # PDB 2jtg

REFERENCES BIBLIOGRAPHIQUES

Alberts, I.L., Nadassy, K. and Wodak, S.J. (1998) Analysis of zinc binding sites in protein crystal structures. *Protein Sci*, **7**, 1700-1716.

Andreini, C., Banci, L., Bertini, I. and Rosato, A. (2006) Counting the zinc-proteins encoded in the human genome. *J Proteome Res*, **5**, 196-201.

Aravind, L. (2000) The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem Sci*, **25**, 421-423.

Arnott, S. and Selsing, E. (1974a) Letter: The structure of polydeoxyguanylic acid with polydeoxycytidylic acid. *J Mol Biol*, **88**, 551-552.

Arnott, S. and Selsing, E. (1974b) Structures for the polynucleotide complexes poly(dA) with poly (dT) and poly(dT) with poly(dA) with poly (dT). *J Mol Biol*, **88**, 509-521.

Baleja, J.D., Thanabal, V. and Wagner, G. (1997) Refined solution structure of the DNA-binding domain of GAL4 and use of 3J(113Cd,1H) in structure determination. *J Biomol NMR*, **10**, 397-401.

Barlow, P.N., Luisi, B., Milner, A., Elliott, M. and Everett, R. (1994) Structure of the C3HC4 domain by 1H-nuclear magnetic resonance spectroscopy. A new structural class of zinc-finger. *J Mol Biol*, **237**, 201-211.

Bartels, C., Xia, T.-H., Billeter, M., Güntert, P. and Wüthrich, K. (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *j. Biomol. NMR*, **6**, 1-10.

Bax, A. (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci*, **12**, 1-16.

Beamer, L.J. and Pabo, C.O. (1992) Refined 1.8 A crystal structure of the lambda repressor-operator complex. *J Mol Biol*, **227**, 177-196.

Benos, P.V., Lapedes, A.S. and Stormo, G.D. (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays*, **24**, 466-475.

Bertani, G. (1951) Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. *J Bacteriol*, **62**, 293-300.

Bjorklund, S., Hjortsberg, K., Johansson, E. and Thelander, L. (1993) Structure and promoter characterization of the gene encoding the large subunit (R1 protein) of mouse ribonucleotide reductase. *Proc Natl Acad Sci U S A*, **90**, 11322-11326.

Boehr, D.D., Dyson, H.J. and Wright, P.E. (2006) An NMR perspective on enzyme dynamics. *Chem Rev*, **106**, 3055-3079.

Bonvin, A.M., Boelens, R. and Kaptein, R. (2005) NMR analysis of protein interactions. *Curr Opin Chem Biol*, **9**, 501-508.

Bouvet, P. (2001) Determination of nucleic acid recognition sequences by SELEX. *Methods Mol Biol*, **148**, 603-610.

Boxem, M. and van den Heuvel, S. (2002) C. elegans class B synthetic multivulva genes act in G(1) regulation. *Curr Biol*, **12**, 906-911.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, **54**, 905-921.

Burkhard, P., Stetefeld, J. and Strelkov, S.V. (2001) Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol*, **11**, 82-88.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M. and Gronenborn, A.M. (1997) Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat Struct Biol*, **4**, 567-577.

Carlomagno, T. (2005) Ligand-target interactions: what can we learn from NMR? *Annu Rev Biophys Biomol Struct*, **34**, 245-266.

Castagne, 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*, **267**, 1223-1229.

Cayrol, C., Lacroix, C., Mathe, C., Ecochard, V., Ceribelli, M., Loreau, E., Lazar, V., Dessen, P., Mantovani, R., Aguilar, L. and Girard, J.P. (2007) The THAP-zinc finger protein THAP1 regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes. *Blood*, **109**, 584-594.

Cheung, W.L., Ajiro, K., Samejima, K., Kloc, M., Cheung, P., Mizzen, C.A., Beeser, A., Etkin, L.D., Chernoff, J., Earnshaw, W.C. and Allis, C.D. (2003) Apoptotic phosphorylation of histone H2B is mediated by mammalian sterile twenty kinase. *Cell*, **113**, 507-517.

Choo, Y. and Klug, A. (1993) A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nucleic Acids Res*, **21**, 3341-3346.

Choo, Y. and Klug, A. (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A*, **91**, 11168-11172.

Choo, Y. and Klug, A. (1997) Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol*, **7**, 117-125.

Christy, B. and Nathans, D. (1989) DNA binding site of the growth factor-inducible protein Zif268. *Proc Natl Acad Sci U S A*, **86**, 8737-8741.

Clore, G.M. and Gronenborn, A.M. (1998) Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol*, **16**, 22-34.

Clouaire, T., Roussigne, M., Ecochard, V., Mathe, C., Amalric, F. and Girard, J.P. (2005) The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. *Proc Natl Acad Sci U S A*, **102**, 6907-6912.

Coleman, J.E. (1992) Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins. *Annu Rev Biochem*, **61**, 897-946.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*, **13**, 289-302.

Crippen, G.M. and Havel, T.F. (1988) *Distance Geometry and Molecular Conformation*, New York.

Crossley, M., Merika, M. and Orkin, S.H. (1995) Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domains. *Mol Cell Biol*, **15**, 2448-2456.

Daragan, V.A. and Mayo, K.H. (1998) Motional model analyses of protein and peptide dynamics using Image and Image NMR relaxation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **31**, 63-105.

de Alba, E. and Tjandra, N. (2002) NMR dipolar couplings for the structure determination of biopolymers in solution. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **40**, 175-197.

Deiss, L.P., Feinstein, E., Berissi, H., Cohen, O. and Kimchi, A. (1995) Identification of a novel serine/threonine kinase and a novel 15-kD protein as potential mediators of the gamma interferon-induced cell death. *Genes Dev*, **9**, 15-30.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*, **6**, 277-293.

Demene, H., Dong, C.Z., Ottmann, M., Rouyez, M.C., Jullian, N., Morellet, N., Mely, Y., Darlix, J.L., Fournie-Zaluski, M.C., Saragosti, S. and et al. (1994) 1H NMR structure and biological studies of the His23-->Cys mutant nucleocapsid protein of HIV-1 indicate that the conformation of the first zinc finger is critical for virus infectivity. *Biochemistry*, **33**, 11707-11716.

Desjarlais, J.R. and Berg, J.M. (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc Natl Acad Sci U S A*, **89**, 7345-7349.

Dimova, D.K. and Dyson, N.J. (2005) The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, **24**, 2810-2826.

Dominguez, C., Boelens, R. and Bonvin, A.M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, **125**, 1731-1737.

Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**, 197-208.

Elofsson, A. and Sonnhammer, E.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, **15**, 480-500.

Elrod-Erickson, M., Rould, M.A., Nekludova, L. and Pabo, C.O. (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171-1180.

Emerson, S.D., Madison, V.S., Palermo, R.E., Waugh, D.S., Scheffler, J.E., Tsao, K.L., Kiefer, S.E., Liu, S.P. and Fry, D.C. (1995) Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry*, **34**, 6911-6918.

Farrow, N.A., Zhang, O., Forman-Kay, J.D. and Kay, L.E. (1994) A heteronuclear correlation experiment for simultaneous determination of 15N longitudinal decay and chemical exchange rates of systems in slow equilibrium. *J Biomol NMR*, **4**, 727-734.

Fielding, L. (2007) NMR methods for the determination of protein–ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy*, in press.

Fischer, M.W., Majumdar, A. and Zuiderweg, E.R. (1998) Protein NMR relaxation : theory, applications and outlook. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **33**, 207-272.

Fisher, H.F. and Singh, N. (1995) Calorimetric methods for interpreting protein-ligand interactions. *Methods Enzymol*, **259**, 194-221.

Flynn, P.F., Matiello, D.L., Hill, H.D.W. and Wand, A.J. (2000) Optimal Use of Cryogenic Probe Technology in NMR
Studies of Proteins. *J. Am. Chem. Soc.*, **122**, 4823-4824.

Foster, M.P., McElroy, C.A. and Amero, C.D. (2007) Solution NMR of large molecules and assemblies. *Biochemistry*, **46**, 331-340.

Foster, M.P., Wuttke, D.S., Radhakrishnan, I., Case, D.A., Gottesfeld, J.M. and Wright, P.E. (1997) Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. *Nat Struct Biol*, **4**, 605-608.

Gamsjaeger, R., Liew, C.K., Loughlin, F.E., Crossley, M. and Mackay, J.P. (2007) Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem Sci*, **32**, 63-70.

Garvie, C.W. and Wolberger, C. (2001) Recognition of specific DNA sequences. *Mol Cell*, **8**, 937-946.

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.

Gowers, D.M. and Halford, S.E. (2003) Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *Embo J*, **22**, 1410-1418.

Grzesiek, S. and Bax, A. (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J Biomol NMR*, **3**, 185-204.

Guntert, P., Braun, W., Billeter, M. and Wüthrich, K. (1989) Automated stereospecific proton NMR assignments and their impact on the precision of protein structure determinations in solution. *J Am Chem Soc*, **111**, 3997-4004.

Günther, H. (1995) *NMR spectroscopy*.

Gutfreund, H. (1987) Reflections on the kinetics of substrate binding. *Biophys Chem*, **26**, 117-121.

Halford, S.E. and Marko, J.F. (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res*, **32**, 3040-3052.

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.

Hard, T. (1999) NMR studies of protein-nucleic acid complexes: structures, solvation, dynamics and coupled protein folding. *Q Rev Biophys*, **32**, 57-98.

Hard, T., Kellenbach, E., Boelens, R., Maler, B.A., Dahlman, K., Freedman, L.P., Carlstedt-Duke, J., Yamamoto, K.R., Gustafsson, J.A. and Kaptein, R. (1990) Solution structure of the glucocorticoid receptor DNA-binding domain. *Science*, **249**, 157-160.

Hard, T. and Lundback, T. (1996) Thermodynamics of sequence-specific protein-DNA interactions. *Biophys Chem*, **62**, 121-139.

Heyduk, T. and Lee, J.C. (1990) Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. *Proc Natl Acad Sci U S A*, **87**, 1744-1748.

Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123-138.

Holmbeck, S.M., Foster, M.P., Casimiro, D.R., Sem, D.S., Dyson, H.J. and Wright, P.E. (1998) High-resolution solution structure of the retinoid X receptor DNA-binding domain. *J Mol Biol*, **281**, 271-284.

Hubbard, S.J. and Thornton, J.M. (1993) *'NACCESS', computer program*. University College, London.

Iwahara, J. and Clore, G.M. (2006a) Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature*, **440**, 1227-1230.

Iwahara, J. and Clore, G.M. (2006b) Direct observation of enhanced translocation of a homeodomain between DNA cognate sites by NMR exchange spectroscopy. *J Am Chem Soc*, **128**, 404-405.

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**, 15062-15067.

Jacobs, G.H. (1992) Determination of the base recognition positions of zinc fingers from sequence analysis. *Embo J*, **11**, 4507-4517.

Jamieson, A.C., Miller, J.C. and Pabo, C.O. (2003) Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov*, **2**, 361-368.

Janin, J. (1985) *Méthodes biophysiques pour l'étude des macromolécules*. Herman, Paris.

Jauch, R., Bourenkov, G.P., Chung, H.R., Urlaub, H., Reidt, U., Jackle, H. and Wahl, M.C. (2003) The zinc finger-associated domain of the Drosophila transcription factor grauzone is a novel zinc-coordinating protein-protein interaction module. *Structure*, **11**, 1393-1402.

Jen-Jacobson, L. (1997) Protein-DNA recognition complexes: conservation of structure and binding energy in the transition state. *Biopolymers*, **44**, 153-180.

Johansson, E., Hjortsberg, K. and Thelander, L. (1998) Two YY-1-binding proximal elements regulate the promoter strength of the TATA-less mouse ribonucleotide reductase R1 gene. *J Biol Chem*, **273**, 29816-29821.

Johnson, B.A. (2004) *Using NMRView to Visualize and Analyze the NMR Spectra of Macromolecules*.

Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Mei Ono, A. and Guntert, P. (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature*, **440**, 52-57.

Kalodimos, C.G., Biris, N., Bonvin, A.M., Levandoski, M.M., Guennuegues, M., Boelens, R. and Kaptein, R. (2004a) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386-389.

Kalodimos, C.G., Boelens, R. and Kaptein, R. (2004b) Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. *Chem Rev*, **104**, 3567-3586.

Karplus, M. (1959) Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys*, **30**, 11-15.

Kaufman, P.D., Doll, R.F. and Rio, D.C. (1989) Drosophila P element transposase recognizes internal P element DNA sequences. *Cell*, **59**, 359-371.

Keeler, J. (2002) *Understanding NMR Spectrocsopy*, Cambridge.

Kelly, S.M., Pabit, S.A., Kitchen, C.M., Guo, P., Marfatia, K.A., Murphy, T.J., Corbett, A.H. and Berland, K.M. (2007) Recognition of polyadenosine RNA by zinc finger proteins. *Proc Natl Acad Sci U S A*, **104**, 12306-12311.

Khorasanizadeh, S. and Rastinejad, F. (2001) Nuclear-receptor interactions on DNA-response elements. *Trends Biochem Sci*, **26**, 384-390.

Khrapunov, S., Brenowitz, M., Rice, P.A. and Catalano, C.E. (2006) Binding then bending: a mechanism for wrapping DNA. *Proc Natl Acad Sci U S A*, **103**, 19217-19218.

Kim, J.L., Nikolov, D.B. and Burley, S.K. (1993a) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**, 520-527.

Kim, Y., Geiger, J.H., Hahn, S. and Sigler, P.B. (1993b) Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**, 512-520.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, **63**, 579-590.

Klug, A. (2005a) The discovery of zinc fingers and their development for practical applications in gene regulation. *proc Japan Acad*, **81**, 87-102.

Klug, A. (2005b) Towards therapeutic applications of engineered zinc finger proteins. *FEBS Lett*, **579**, 892-894.

Klug, A. and Schwabe, J.W. (1995) Protein motifs 5. Zinc fingers. *Faseb J*, **9**, 597-604.

Kodera, Y., Sato, K., Tsukahara, T., Komatsu, H., Maeda, T. and Kohno, T. (1998) High-resolution solution NMR structure of the minimal active domain of the human immunodeficiency virus type-2 nucleocapsid protein. *Biochemistry*, **37**, 17704-17713.

Korzhnev, D.M., Billeter, M., Arseniev, A.S. and Orekhov, V.Y. (2001) NMR studies of Brownian tumbling and internal motion in proteins. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **38**, 197-266.

Kowalski, K., Czolij, R., King, G.F., Crossley, M. and Mackay, J.P. (1999) The solution structure of the N-terminal zinc finger of GATA-1 reveals a specific binding face for the transcriptional co-factor FOG. *J Biomol NMR*, **13**, 249-262.

Krishna, S.S., Majumdar, I. and Grishin, N.V. (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res*, **31**, 532-550.

Krizek, B.A., Amann, B.T., Kilfoil, V.J., Merkle, D.L. and Berg, J.M. (1991) A consencus zinc finger peptide : hight-affinity metal binding, a pH-dependent structure, and a his to cys sequence variant. *J. Am. Chem. Soc.*, **113**, 4518-4523.

Kulczyk, A.W., Yang, J.C. and Neuhaus, D. (2004) Solution structure and DNA binding of the zinc-finger domain from DNA ligase IIIalpha. *J Mol Biol*, **341**, 723-738.

Kuznetsov, S.V., Sugimura, S., Vivas, P., Crothers, D.M. and Ansari, A. (2006) Direct observation of DNA bending/unbending kinetics in complex with DNA-bending protein IHF. *Proc Natl Acad Sci U S A*, **103**, 18515-18520.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680-685.

Laity, J.H., Dyson, H.J. and Wright, P.E. (2000) DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers. *J Mol Biol*, **295**, 719-727.

Laity, J.H., Lee, B.M. and Wright, P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol*, **11**, 39-46.

Lane, A.N., Kelly, G., Ramos, A. and Frenkiel, T.A. (2001) Determining binding sites in protein-nucleic acid complexes by cross-saturation. *J Biomol NMR*, **21**, 127-139.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**, 477-486.

Lee, C.C., Beall, E.L. and Rio, D.C. (1998) DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. *Embo J*, **17**, 4166-4174.

Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A. and Wright, P.E. (1989) Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science*, **245**, 635-637.

Leon, O. and Roth, M. (2000) Zinc fingers: DNA binding and protein-protein interactions. *Biol Res*, **33**, 21-30.

Licata, V.J. and Wowor, A.J. (2008) Applications of Fluorescence Anisotropy to the Study of Protein-DNA Interactions. *Methods Cell Biol*, **84**, 243-262.

Liew, C.K., Crossley, M., Mackay, J.P. and Nicholas, H.R. (2007) Solution structure of the THAP domain from Caenorhabditis elegans C-terminal binding protein (CtBP). *J Mol Biol*, **366**, 382-390.

Liew, C.K., Simpson, R.J., Kwan, A.H., Crofts, L.A., Loughlin, F.E., Matthews, J.M., Crossley, M. and Mackay, J.P. (2005) Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. *Proc Natl Acad Sci U S A*, **102**, 583-588.

Lin, Y., Khokhlatchev, A., Figeys, D. and Avruch, J. (2002) Death-associated protein 4 binds MST1 and augments MST1-induced apoptosis. *J Biol Chem*, **277**, 47991-48001.

Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2004) Correction of spin diffusion during iterative automated NOE assignment. *J Magn Reson*, **167**, 334-342.

Linge, J.P., O'Donoghue, S.I. and Nilges, M. (2001) Automated assignment of ambiguous nuclear overhauser effects with ARIA. *Methods Enzymol*, **339**, 71-90.

Liu, P.Q., Rebar, E.J., Zhang, L., Liu, Q., Jamieson, A.C., Liang, Y., Qi, H., Li, P.X., Chen, B., Mendel, M.C., Zhong, X., Lee, Y.L., Eisenberg, S.P., Spratt, S.K., Case, C.C. and Wolffe, A.P. (2001) Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J Biol Chem*, **276**, 11323-11334.

Lowry, J.A. and Atchley, W.R. (2000) Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. *J Mol Evol*, **50**, 103-115.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R. and Sigler, P.B. (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, **352**, 497-505.

Lupas, A. (1996) Coiled coils: new structures and new functions. *Trends Biochem Sci*, **21**, 375-382.

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.

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.

Macfarlan, T., Kutney, S., Altman, B., Montross, R., Yu, J. and Chakravarti, D. (2005) Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. *J Biol Chem*, **280**, 7346-7358.

Macfarlan, T., Parker, J.B., Nagata, K. and Chakravarti, D. (2006) Thanatos-associated protein 7 associates with template activating factor-Ibeta and inhibits histone acetylation to repress transcription. *Mol Endocrinol*, **20**, 335-347.

Mackay, J.P. and Crossley, M. (1998) Zinc fingers are sticking together. *Trends Biochem Sci*, **23**, 1-4.

Malliavin, T. and Dardel, F. (2002) Structure des protéines par RMN. *Techniques de l'ingénieur*.

Malmqvist, M. (1993) Biospecific interaction analysis using biosensor technology. *Nature*, **361**, 186-187.

Mangelsdorf, D.J. and Evans, R.M. (1995) The RXR heterodimers and orphan receptors. *Cell*, **83**, 841-850.

Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. and Evans, R.M. (1995) The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835-839.

Martin, D.I. and Orkin, S.H. (1990) Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev*, **4**, 1886-1898.

Miller, J., McLachlan, A.D. and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *Embo J*, **4**, 1609-1614.

Muller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L. and Harrison, S.C. (1995) Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature*, **373**, 311-317.

Narayan, V.A., Kriwacki, R.W. and Caradonna, J.P. (1997) Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition. *J Biol Chem*, **272**, 7801-7809.

Newton, A., Mackay, J. and Crossley, M. (2001) The N-terminal zinc finger of the erythroid transcription factor GATA-1 binds GATC motifs in DNA. *J Biol Chem*, **276**, 35794-35801.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Lett*, **239**, 129-136.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1990) 1H-NMR stereospecific assignments by conformational data-base searches. *Biopolymers*, **29**, 813-822.

Noggle, J.H. and Shirmer, R.E. (1971) *The Nuclear Overhausser Effect*, New York.

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.

Pabo, C.O., Peisach, E. and Grant, R.A. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem*, **70**, 313-340.

Pabo, C.O. and Sauer, R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem*, **61**, 1053-1095.

Papworth, M., Kolasinska, P. and Minczuk, M. (2006) Designer zinc-finger proteins and their applications. *Gene*, **366**, 27-38.

Papworth, M., Moore, M., Isalan, M., Minczuk, M., Choo, Y. and Klug, A. (2003) Inhibition of herpes simplex virus 1 gene expression by designer zinc-finger transcription factors. *Proc Natl Acad Sci U S A*, **100**, 1621-1626.

Pardi, A., Billeter, M. and Wuthrich, K. (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein.

Use of 3JHN alpha for identification of helical secondary structure. *J Mol Biol*, **180**, 741-751.

Patient, R.K. and McGhee, J.D. (2002) The GATA family (vertebrates and invertebrates). *Curr Opin Genet Dev*, **12**, 416-422.

Pavletich, N.P. and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809-817.

Pellecchia, M., Sem, D.S. and Wuthrich, K. (2002) NMR in drug discovery. *Nat Rev Drug Discov*, **1**, 211-219.

Pelton, J.G., Torchia, D.A., Meadow, N.D. and Roseman, S. (1993) Tautomeric states of the active-site histidines of phosphorylated and unphosphorylated IIIGlc, a signal-transducing protein from Escherichia coli, using two-dimensional heteronuclear NMR techniques. *Protein Sci*, **2**, 543-558.

Pervushin, K., Riek, R., Wider, G. and Wuthrich, K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A*, **94**, 12366-12371.

Prestegard, J.H., Bougault, C.M. and Kishore, A.I. (2004) Residual dipolar couplings in structure determination of biomolecules. *Chem Rev*, **104**, 3519-3540.

Privalov, P.L., Dragan, A.I., Crane-Robinson, C., Breslauer, K.J., Remeta, D.P. and Minetti, C.A. (2007) What drives proteins into the major or minor grooves of DNA? *J Mol Biol*, **365**, 1-9.

Qian, X., Jeon, C., Yoon, H., Agarwal, K. and Weiss, M.A. (1993) Structure of a new nucleic-acid-binding motif in eukaryotic transcriptional elongation factor TFIIS. *Nature*, **365**, 277-279.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol*, **7**, 95-99.

Ramos, A., Kelly, G., Hollingworth, D., Pastore, A. and Frenkiel, T.A. (2000) Mapping the interfaces of proteinnucleic acid complexes using cross-saturation. *J Am Chem Soc*, **122**, 11311-11314.

Rance, M., Sorensen, O.W., Bodenhausen, G., Wagner, G., Ernst, R.R. and Wuthrich, K. (1983) Improved spectral resolution in cosy 1H NMR spectra of proteins via double quantum filtering. *Biochem Biophys Res Commun*, **117**, 479-485.

Rebar, E.J., Huang, Y., Hickey, R., Nath, A.K., Meoli, D., Nath, S., Chen, B., Xu, L., Liang, Y., Jamieson, A.C., Zhang, L., Spratt, S.K., Case, C.C., Wolffe, A. and Giordano, F.J. (2002) Induction of angiogenesis in a mouse model using engineered transcription factors. *Nat Med*, **8**, 1427-1432.

Reddy, K.C. and Villeneuve, A.M. (2004) C. elegans HIM-17 links chromatin modification and competence for initiation of meiotic recombination. *Cell*, **118**, 439-452.

Renaud, J.P. and Moras, D. (2000) Structural studies on nuclear receptors. *Cell Mol Life Sci*, **57**, 1748-1769.

Reynolds, L., Ullman, C., Moore, M., Isalan, M., West, M.J., Clapham, P., Klug, A. and Choo, Y. (2003) Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. *Proc Natl Acad Sci U S A*, **100**, 1615-1620.

Rhodes, D., Schwabe, J.W., Chapman, L. and Fairall, L. (1996) Towards an understanding of protein-DNA recognition. *Philos Trans R Soc Lond B Biol Sci*, **351**, 501-509.

Riggs, A.D., Bourgeois, S. and Cohn, M. (1970a) The lac repressor-operator interaction. 3. Kinetic studies. *J Mol Biol*, **53**, 401-417.

Riggs, A.D., Suzuki, H. and Bourgeois, S. (1970b) Lac repressor-operator interaction. I. Equilibrium studies. *J Mol Biol*, **48**, 67-83.

Roques, B.P., Morellet, N., de Rocquigny, H., Demene, H., Schueler, W. and Jullian, N. (1997) Structure, biological functions and inhibition of the HIV-1 proteins Vpr and NCp7. *Biochimie*, **79**, 673-680.

Ross, J.B., Szabo, A.G. and Hogue, C.W. (1997) Enhancement of protein spectra with tryptophan analogs: fluorescence spectroscopy of protein-protein and protein-nucleic acid interactions. *Methods Enzymol*, **278**, 151-190.

Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F. and Girard, J.P. (2003a) THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. *Oncogene*, **22**, 2432-2442.

Roussigne, M., Kossida, S., Lavigne, A.C., Clouaire, T., Ecochard, V., Glories, A., Amalric, F. and Girard, J.P. (2003b) The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem Sci*, **28**, 66-69.

Roy, S. (2004) Fluorescence quenching methods to study protein-nucleic acid interactions. *Methods Enzymol*, **379**, 175-187.

Sarai, A. and Kono, H. (2005) Protein-DNA recognition patterns and predictions. *Annu Rev Biophys Biomol Struct*, **34**, 379-398.

Sattler, M., Schleucher, J. and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **34**, 93-158.

Schmiedeskamp, M., Rajagopal, P. and Klevit, R.E. (1997) NMR chemical shift perturbation mapping of DNA binding by a zinc-finger domain from the yeast transcription factor ADR1. *Protein Sci*, **6**, 1835-1848.

Schuck, P. (1997a) Reliable determination of binding affinity and kinetics using surface plasmon resonance biosensors. *Curr Opin Biotechnol*, **8**, 498-502.

Schuck, P. (1997b) Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules. *Annu Rev Biophys Biomol Struct*, **26**, 541-566.

Schwabe, J.W., Chapman, L., Finch, J.T. and Rhodes, D. (1993a) The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*, **75**, 567-578.

Schwabe, J.W., Chapman, L., Finch, J.T., Rhodes, D. and Neuhaus, D. (1993b) DNA recognition by the oestrogen receptor: from solution to the crystal. *Structure*, **1**, 187-204.

Schwabe, J.W., Neuhaus, D. and Rhodes, D. (1990) Solution structure of the DNA-binding domain of the oestrogen receptor. *Nature*, **348**, 458-461.

Searles, M.A., Lu, D. and Klug, A. (2000) The role of the central zinc fingers of transcription factor IIIA in binding to 5 S RNA. *J Mol Biol*, **301**, 47-60.

Seeman, N.C., Rosenberg, J.M. and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A*, **73**, 804-808.

Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, **274**, 1531-1534.

Slutsky, M. and Mirny, L.A. (2004) Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys J*, **87**, 4021-4035.

Smith, G.P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315-1317.

Spolar, R.S. and Record, M.T., Jr. (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science*, **263**, 777-784.

Staunton, D., Schlinkert, R., Zanetti, G., Colebrook, S.A. and Campbell, L.D. (2006) Cell-free expression and selective isotope labelling in
protein NMR. *Magn. Reson. Chem.*, **44**, 2-9.

Stein, E.G., Rice, L.M. and Brunger, A.T. (1997) Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J Magn Reson*, **124**, 154-164.

Studier, F.W. and Moffatt, B.A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol*, **189**, 113-130.

Sugimura, S. and Crothers, D.M. (2006) Stepwise binding and bending of DNA by Escherichia coli integration host factor. *Proc Natl Acad Sci U S A*, **103**, 18510-18514.

Sun, L., Liu, A. and Georgopoulos, K. (1996) Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. *Embo J*, **15**, 5358-5369.

Tolman, J.R. (2001) Dipolar couplings as a probe of molecular dynamics and structure in solution. *Curr Opin Struct Biol*, **11**, 532-539.

Trainor, C.D., Omichinski, J.G., Vandergon, T.L., Gronenborn, A.M., Clore, G.M. and Felsenfeld, G. (1996) A palindromic regulatory site within vertebrate GATA-1 promoters requires both zinc fingers of the GATA-1 DNA-binding domain for high-affinity interaction. *Mol Cell Biol*, **16**, 2238-2247.

Tsang, A.P., Visvader, J.E., Turner, C.A., Fujiwara, Y., Yu, C., Weiss, M.J., Crossley, M. and Orkin, S.H. (1997) FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell*, **90**, 109-119.

Tugarinov, V., Choy, W.Y., Orekhov, V.Y. and Kay, L.E. (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci U S A*, **102**, 622-627.

Turner, R.B., Smith, D.L., Zawrotny, M.E., Summers, M.F., Posewitz, M.C. and Winge, D.R. (1998) Solution structure of a zinc domain conserved in yeast copper-regulated transcription factors. *Nat Struct Biol*, **5**, 551-555.

van Dijk, M., van Dijk, A.D., Hsu, V., Boelens, R. and Bonvin, A.M. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res*, **34**, 3317-3325.

von Hippel, P.H. and Berg, O.G. (1989) Facilitated target location in biological systems. *J Biol Chem*, **264**, 675-678.

Vuister, G.W., Wang, A.C. and Bax, A. (1993) Measurement of three-bond nitrogen-carbon J couplings in proteins uniformly enriched in nitrogen-15 and carbon-13. *J. Am. Chem. Soc.*, **115**, 5334-5335.

Williamson, M.P., Havel, T.F. and Wuthrich, K. (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J Mol Biol*, **182**, 295-315.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol*, **222**, 311-333.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, **31**, 1647-1651.

Wittekind, M. and Mueller, L. (1993) HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. *J. Magn. Reson.*, **B101**, 1993.

Wolfe, S.A., Nekludova, L. and Pabo, C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, **29**, 183-212.

Wüthrich, K. (1986) *NMR of proteins and nucleic acids*. Wiley-Interscience, New York.

Wuthrich, K. and Wagner, G. (1979) Nuclear magnetic resonance of labile protons in the basic pancreatic trypsin inhibitor. *J Mol Biol*, **130**, 1-18.

Wuttke, D.S., Foster, M.P., Case, D.A., Gottesfeld, J.M. and Wright, P.E. (1997) Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *J Mol Biol*, **273**, 183-206.

Yamaguchi, H., Tateno, M. and Yamasaki, K. (2006) Solution structure and DNA-binding mode of the matrix attachment region-binding domain of the transcription factor SATB1 that regulates the T-cell maturation. *J Biol Chem*, **281**, 5319-5327.

Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Nunokawa, E., Ishizuka, Y., Terada, T., Shirouzu, M., Osanai, T., Tanaka, A., Seki, M., Shinozaki, K. and Yokoyama, S. (2004a) A novel zinc-binding motif revealed by solution structures of DNA-binding domains of Arabidopsis SBP-family transcription factors. *J Mol Biol*, **337**, 49-63.

Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Tomo, Y., Hayami, N., Terada, T., Shirouzu, M., Osanai, T., Tanaka, A., Seki, M., Shinozaki, K. and Yokoyama, S. (2004b) Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1. *Plant Cell*, **16**, 3448-3459.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G.M., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C.H., Cort, J.R., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L.P., Gehring, K., Kennedy, M.A., Edwards, A.M. and Arrowsmith, C.H. (2002) An NMR approach to structural proteomics. *Proc Natl Acad Sci U S A*, **99**, 1825-1830.

Zhang, L., Spratt, S.K., Liu, Q., Johnstone, B., Qi, H., Raschke, E.E., Jamieson, A.C., Rebar, E.J., Wolffe, A.P. and Case, C.C. (2000) Synthetic zinc finger transcription factor action at an endogenous chromosomal site. Activation of the human erythropoietin gene. *J Biol Chem*, **275**, 33850-33860.

Zhang, Y., Stec, B. and Godzik, A. (2007) Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure*, **15**, 1141-1147.

Zuiderweg, E.R. (2002) Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, **41**, 1-7.

AUTEUR : Damien BESSIERE

TITRE : Le domaine THAP de THAP1 : structure par RMN en solution et interaction avec l'ADN

DIRECTEURS DE THESE : Alain MILON et Virginie GERVAIS

LIEU ET DATE DE SOUTENANCE : Salle Fernand Gallais (LCC, 205 route de Narbonne, 31077 Toulouse, France), le 01 février 2008

Résumé

La famille des protéines THAP est caractérisée par la présence d'un motif protéique, le domaine THAP, conservé au cours de l'évolution et retrouvé dans une centaine de protéines chez l'homme et les organismes animaux modèles. La protéine THAP1 humaine est impliquée dans la régulation du cycle cellulaire dans la voie pRb/E2F et dans la prolifération cellulaire. Le domaine THAP de THAP1 est un motif de liaison à l'ADN de type C-X_{2-4}-C-X_{35-50}-C-X_2-H, séquence-spécifique et dépendant du zinc. Nous avons déterminé la structure tridimensionnelle du domaine THAP de THAP1 par Résonance Magnétique Nucléaire en solution. Ce doigt de zinc atypique de ~ 80 résidus se distingue par la présence d'un long motif $\beta\alpha\beta$ entre les deux paires de ligands de coordination au zinc C2CH. Nous avons étudié la liaison du domaine THAP de THAP1 à sa séquence ADN spécifiquement reconnue en déterminant une constante de dissociation spécifique par Résonance Plasmonique de Surface et en réalisant des expériences d'empreinte RMN de façon à identifier les résidus impliqués dans la liaison à l'ADN. La combinaison des données de variation de déplacement chimique avec des données de mutagénèse dirigée nous a permis de localiser l'interface de liaison à l'ADN du domaine, correspondant à une zone chargée positivement, et de construire un modèle d'interaction protéine-ADN rendant compte d'une reconnaissance originale.

Abstract

The THAP proteins family is characterised by the presence of a protein motif, designed the THAP domain, conserved during evolution and found in a thousand of human and model animal organism proteins. Human THAP1 protein is a regulator of the cell cycle through pRB/E2F pathway. The THAP domain of THAP1 is a C-X2-4-C-X35-50-C-X2-H type of zinc binding module with sequence specific DNA-binding properties. We solved the three-dimensional structure of the THAP domain of THAP1 by solution Nuclear Magnetic Resonance. This atypical zinc finger of ~ 80 residues is distinguished by the presence of a long $\beta\alpha\beta$ motif between the two pairs of the C2CH zinc coordinating residues. We studied the DNA-binding of the THAP domain of THAP1 by the determination of a specific dissociation constant with Surface Plasmon Resonance studies and by performing chemical shift mapping in order to identify DNA-binding affected residues. Combining the chemical shift perturbation data with mutagenesis data allowed us to map the DNA-binding interface of the domain to a highly positively charged area and to build a DNA-protein interaction model showing an original way of recognition.

MOTS-CLES : structure par Résonance Magnétique Nucléaire, doigt de zinc, domaine de liaison à l'ADN, interaction protéine-ADN, Résonance Plasmonique de Surface

DISCIPLINE : Biologie structurale

INTITULE ET ADRESSE DU LABORATOIRE : Institut de Pharmacologie et de Biologie Structurale (IPBS) – CNRS UMR 5089
205 route de Narbonne, 31077 Toulouse

AUTEUR : Damien BESSIERE

TITRE : Le domaine THAP de THAP1 : structure par RMN en solution et interaction avec l'ADN

DIRECTEURS DE THESE : Alain MILON et Virginie GERVAIS

LIEU ET DATE DE SOUTENANCE : Salle Fernand Gallais (LCC, 205 route de Narbonne, 31077 Toulouse, France), le 01 février 2008

Résumé

La famille des protéines THAP est caractérisée par la présence d'un motif protéique, le domaine THAP, conservé au cours de l'évolution et retrouvé dans une centaine de protéines chez l'homme et les organismes animaux modèles. La protéine THAP1 humaine est impliquée dans la régulation du cycle cellulaire dans la voie pRb/E2F et dans la prolifération cellulaire. Le domaine THAP de THAP1 est un motif de liaison à l'ADN de type C-X_{2-4}-C-X_{35-50}-C-X_2-H, séquence-spécifique et dépendant du zinc. Nous avons déterminé la structure tridimensionnelle du domaine THAP de THAP1 par Résonance Magnétique Nucléaire en solution. Ce doigt de zinc atypique de ~ 80 résidus se distingue par la présence d'un long motif $\beta\alpha\beta$ entre les deux paires de ligands de coordination au zinc C2CH. Nous avons étudié la liaison du domaine THAP de THAP1 à sa séquence ADN spécifiquement reconnue en déterminant une constante de dissociation spécifique par Résonance Plasmonique de Surface et en réalisant des expériences d'empreinte RMN de façon à identifier les résidus impliqués dans la liaison à l'ADN. La combinaison des données de variation de déplacement chimique avec des données de mutagénèse dirigée nous a permis de localiser l'interface de liaison à l'ADN du domaine, correspondant à une zone chargée positivement, et de construire un modèle d'interaction protéine-ADN rendant compte d'une reconnaissance originale.

Abstract

The THAP proteins family is characterised by the presence of a protein motif, designed the THAP domain, conserved during evolution and found in a thousand of human and model animal organism proteins. Human THAP1 protein is a regulator of the cell cycle through pRB/E2F pathway. The THAP domain of THAP1 is a C-X2-4-C-X35-50-C-X2-H type of zinc binding module with sequence specific DNA-binding properties. We solved the three-dimensional structure of the THAP domain of THAP1 by solution Nuclear Magnetic Resonance. This atypical zinc finger of ~ 80 residues is distinguished by the presence of a long $\beta\alpha\beta$ motif between the two pairs of the C2CH zinc coordinating residues. We studied the DNA-binding of the THAP domain of THAP1 by the determination of a specific dissociation constant with Surface Plasmon Resonance studies and by performing chemical shift mapping in order to identify DNA-binding affected residues. Combining the chemical shift perturbation data with mutagenesis data allowed us to map the DNA-binding interface of the domain to a highly positively charged area and to build a DNA-protein interaction model showing an original way of recognition.

MOTS-CLES : structure par Résonance Magnétique Nucléaire, doigt de zinc, domaine de liaison à l'ADN, interaction protéine-ADN, Résonance Plasmonique de Surface

DISCIPLINE : Biologie structurale

INTITULE ET ADRESSE DU LABORATOIRE : Institut de Pharmacologie et de Biologie Structurale (IPBS) – CNRS UMR 5089
205 route de Narbonne, 31077 Toulouse