2013

Structural basis of DNA binding complexes

Ninad Walavalkar
Virginia Commonwealth University

STRUCTURAL BASIS OF DNA BINDING COMPLEXES

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy at Virginia Commonwealth University.

By

Ninad Maheshwar Walavalkar, M.Sc.

Advisor: David C. Williams Jr, MD, PhD

Assistant Professor

Department of Pathology

Virginia Commonwealth University

Richmond, Virginia

May, 2013

Acknowledgements

This work would not have been possible without the help and support of numerous people that have guided me through various phases of graduate school. I am very grateful to them and would like to thank them from bottom of my heart.

First and foremost, I would like to express my deep gratitude to my advisor Dr. David Williams for his unparalleled support and guidance. This piece of work would not have been accomplished without his knowledge, insight and wisdom. He has been a great role model and I have learned valuable lessons from him that have given me the right attitude to pursue a career in science. I am thankful for his patience and support through all the ups and downs faced during this journey.

I thank my supervisory committee members Dr. Jessica Bell, Dr. Qinglian Liu, Dr. Darrell Peterson and Dr. Neel Scarsdale for their constant support and useful critiques, which have been crucial in achieving my research goals. Also, I would like to thank Dr. Gordon Ginder, Dr. Matthew Hartman, Dr. Ellis Bell and Dr. Carlos Escalante for their help and guidance. I would also like to thank directors of the Integrative Life Sciences program, Dr. Robert Tombes and Dr. William Eggleston for having faith in me as a doctoral candidate. I am thankful to all current and past members of Williams laboratory who have helped in my scientific training.

Also, I would like to thank all my current and past teachers, who have played a key role in shaping me as a scientist, without their words of encouragement this day would not have been possible. Finally, I would like to acknowledge my ‘family and friends’ for their love and moral support. Thank you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two dimensional
3D	Three dimensional
Å	Angstrom units
APBS	Adaptive Poisson-Boltzmann solver
ATP	Adenosine triphosphate
AUC	Analytical ultracentrifugation
BAH	Bromo-adjacent homology
BER	Base excision repair
BMRB	Biological magnetic resonance bank
bp	Base pair
BRCA1	Breast cancer 1
C	Celsius
CAF-1	Chromatin assembly factor 1
CC	Coiled-coil
CD	Circular Dichroism
CHD4	Chromodomain helicase DNA binding protein 4
cMBD2	Chicken methyl DNA binding protein 2
CpG	Cytosine-guanine dinucleotide
CR1/2	Conserved region 1/2
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase

DS DNA	Double stranded DNA
DTT	Dithiothreitol
DUF	Domain of unknown function
EDTA	Ethylenediaminetetraacetic acid
GATAD2A	GATA zinc finger domain containing 2A
GC (rich)	Guanine-Cytosine pair
GCN4	General Control Nonderepressible
GSTP1	Glutathione S-transferase pi gene
HDAC	Histone deacetylases
HMT	Histone methyltransferases
HSQC	Heteronuclear Single Quantum Coherence
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
K	Kelvin units
Ka	Association constant
K_D	Dissociation constant
kDa	Kilo dalton
M9	Minimal media
MBD	Methyl binding domain
MBD3L1/2	MBD3-like-1/2
mCpG	Methylated cytosine guanine dinucleotide
MeCP2	Methyl CpG binding protein 2
MRE	Mean residue ellipticity

mRNA	Messenger Ribonucleic acid
MST1	Macrophage stimulating protein 1
MTA	Metastasis-associated protein
NMR	Nuclear magnetic resonance
NOE	Nuclear overhauser effect
NOESY	Nuclear overhauser effect spectroscopy
NuRD	Nucleosome remodeling and histone deacetylase
PAGE	Polyacrylamide gel electrophoresis
PDB	Protein data bank
pH	Potency of hydrogen
PHD	Plant homeo domain
ppm	Parts per million
psf	Protein structure file
RbAp	Retinoblastoma protein associated protein
RDC	Residual dipolar coupling
rMD	Restrained molecular dynamics
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
SARAH	Sav/Rassf/Hpo
SDS	Sodium dodecyl sulfate
SETDB1	SET Domain, Bifurcated 1
siRNA	Small interfering RNA
SPR	Surface plasmon resonance

TDG	Thymine DNA glycosylase
TET	Ten eleven translocation
TpG	Thymine-guanine dinucleotide
TRD	Transcription repression domain
TROSY	Transverse relaxation optimized spectroscopy
UV	Ultraviolet
ΔG	Change in free energy
ΔH	Change in enthalpy
ΔS	Change in entropy

Preface

With the completion of 'Human Genome Project', we have in our hands a list of all human genes that define the blueprint of human life. These genes encode a variety of molecules which are responsible for biological processes like growth, division, movement, interaction *etcetera*. Acting alone or in groups, these biological macromolecules constitute a complex and intricate network that is turning out to be quite challenging for researchers to study. Scientists probe biological systems using a variety of techniques including genetics, bioinformatics, molecular biology, biochemistry work together to understand these complex biological systems. Here, I present a study involving techniques of structural biology and biophysics to understand molecular details. The reductionist method of dissecting biological systems into their constituent parts is the basis for the structural biology and biophysics and it has proven to be very effective in explaining the chemical basis of numerous living processes. One may argue about the 'logical depth' of reductionism and its use in finding structure based solutions to a biological problem (for example, drugs for a disease). This approach has given clear answers to several of the biological questions, especially those that address structure-function relationships. Keeping this in mind, here I present a study of *'Structural basis of DNA binding complexes.'*

Although, passage of information from DNA to Protein via RNA is considered as 'central dogma' of molecular biology, the field of epigenetics has emerged and has added a new dimension to this concept. In addition to the information encoded by the DNA sequence, distinct epigenetic mechanisms have evolved to store and propagate the information. DNA methylation, covalent and non-covalent histone modifications are a few examples of epigenetic regulation which are maintained by the action of epigenetic modifiers including DNA methyltransferase

(DNMTs), Histone methyltransferases (HMTs) and Histone deacetylases (HDACs). In mammalian cells, methylation occurs at the 5' position of the cytosine from CpG dinucleotides. These methylated CpG islands are involved in transcriptional silencing of imprinted genes, genes located on the inactive X-chromosome, and a number of tumor suppressor genes. DNA methylation, as an epigenetic mark, is the main focus of this dissertation project. How this mark is 'read', the structural basis for 'reader' selectivity and how these 'readers' bring along other cellular machinery to carry out a task at hand are discussed in subsequent chapters with an appropriate example under study.

The nucleosome remodeling and deacetylase (NuRD) complex is an abundant deacetylase complex, which couples histone deacetylation and chromatin remodeling ATPase activities, and has a broad cellular and tissue distribution. Although the working model of how this complex forms and functions is not well known, we have demonstrated that the coiled-coil interaction between two proteins (MBD2 and p66α) is critical for DNA methylation dependent gene silencing *in vivo.* Chapter one: '*Unique features of the anti-parallel, heterodimeric coiled-coil interaction between methyl-cytosine binding domain 2 (MBD2) homologues and p66α dictate high affinity binding'* describes this unique coiled coil interaction. Coiled-coils were studied using a variety of biophysical techniques including analytical ultracentrifugation (AUC), isothermal titration calorimetry (ITC) and circular dichroism (CD). Results were compared across homologues and mutation studies were carried out to test our hypotheses. The studies reported in this chapter add to our understanding of coiled-coil interaction and thereby facilitate development of small peptide based drugs which target such interactions in nature.

A number of proteins have been identified in humans that specifically bind to methylated CpG via a methyl binding domain (MBD). The human genome encodes at least five MBD

proteins: MeCP2 and MBD1 through MBD4, which are homologous in their methyl binding domains but not many similarities are seen outside the MBD. Out of the five MBDs, MBD4 has a c-terminal glycosylase domain through which it recognizes mCpG.TpG mismatch and is important for base excision repair system. Chapter two: 'Dynamic behavior *of MBD4 in methylated DNA recognition'* focuses on MBD4 and its preference for DNA methylation mark. Techniques of surface plasmon resonance (SPR), nuclear magnetic resonance (NMR) spectroscopy are used to study binding affinity for variations of methylated DNA mark. Chemical exchange studies are used to demonstrate how MBD4 scans for methylation mark and these studies have added a new dimension to our understanding of how MBD proteins 'read' DNA methylation marks.

Chapter three*: 'Solving the solution structure of MBD domain of MBD4 on methylated DNA by NMR'* describes a process of structure determination using NMR spectroscopy. The focus of this chapter is not on developing a new technique but rather on using current resources to solve a protein structure, which can be used to further understand our biological system. Here, I have discussed the workflow used to determine a final three-dimensional structure starting from sample preparation, data collection, data analysis to structure calculation.

I hope this dissertation project summarizing my work with DNA binding complexes provides some useful insights into understanding the complex field of epigenetic regulation.

Chapter 1

Unique features of the anti-parallel, heterodimeric coiled-coil interaction between methyl-cytosine binding domain 2 (MBD2) homologues and p66α dictate high affinity binding.

1.1 Abstract

The methyl-cytosine binding domain 2 (MBD2)-nucleosome remodeling and deacetylase (NuRD) complex recognizes methylated DNA and silences expression of the associated genes through histone deacetylase and nucleosome remodeling functions. Our previous structural work demonstrated that a coiled-coil interaction between the MBD2 and GATA zinc finger domain containing 2A (GATAD2A/p66α) proteins recruits the chromodomain helicase DNA-binding protein (CHD4/Mi2β) to the NuRD complex and is necessary for MBD2-mediated DNA methylation dependent gene silencing *in vivo (1)*. The p66α-MBD2 interaction differs from most coiled-coils studied to date by forming a high-affinity anti-parallel heterodimeric complex between two peptides that are largely monomeric in isolation. To further characterize the unique features of this complex that drive heterodimeric specificity and high affinity binding, we carried out biophysical analyses of MBD2 and related homologues MBD3, MBD3-like protein 1(MBD3L1), and MBD3-like protein 2 (MBD3L2) as well as specific mutations that modify charge-charge interactions and helical propensity of the coiled-coil domains. Analytical ultracentrifugation analyses show that the individual peptides remain monomeric in isolation, even at 300 μM in concentration for MBD2. Circular dichroism analyses of the different mutations and homologues demonstrate a direct correlation between helical content of the coiled-coil domains in isolation and binding affinity for p66α. Furthermore, complementary electrostatic surface potentials and inherent helical content of each peptide are necessary to maintain high-affinity association. These factors lead to a binding affinity hierarchy of p66α for the different MBD2 homologues (MBD2 ≈ MBD3 > MBD3L1 ≈ MBD3L2) and suggest a hierarchical regulatory model in tissue and life cycle stage specific silencing by NuRD complexes.

1.2 Introduction

1.2.1 DNA methylation as an epigenetic tool

DNA methylation is the major modification of eukaryotic genome and is shown to be essential for development of mammals *(2)*. DNA methylation involves methylation at the 5' position of symmetrically opposed cytosine bases in a double stranded cytosine-guanosine sequence (CpG). Clusters of CpGs, called CpG islands are GC rich regions of DNA with average length of 1kb and are often found associated with promoters and first exons of genes. The DNA methylation pattern of the CpG islands dictates the expression levels of the associated genes, methylation usually leads to repression *(3, 4)*. In cancer, DNA methylation patterns are dramatically different, as there is a global decrease in CpG methylation but promoters of tumor suppresser genes become hypermethylated *(5, 6)*.

1.2.2 MBD2-NuRD complex

At least five methyl-cytosine binding domain (MBD) proteins have been identified in humans, which share a homologous MBD that selectively binds methylated DNA: MeCP2 and MBD1 through MBD4. Each of these proteins contains a unique sequence outside of the MBD and recruit distinct co-regulatory complexes to silence expression of the associated gene, with the notable exception of MBD2 and MBD3 which share > 65% identity *(5, 7-10)*. The nucleosome remodeling and deacetylase (NuRD) complex is an abundant deacetylase complex with a broad cellular and tissue distribution. MBD2 and MBD3 recruit NuRD which couples histone deacetylation and chromatin remodeling ATPase activities in the same complex. The working model of how this complex forms and functions is unknown. In MBD2 containing NuRD complexes, MBD2 binds methylated DNA selectively and is responsible for recruiting the

NuRD complex to methylated CpG islands. On the other hand, mammalian MBD3 does not specifically recognize methylated DNA and MBD3-NuRD complexes are not preferentially associated with methylated CpG islands. The core components of the NuRD complex include Mi2α/β, HDAC1 and 2, RbAp46/48, MTA1/2, p66α/β, and MBD2 or MBD3 (Figure 1.1A). Information about all components is summarized in Table 1.1.

1.2.3 Domain organization of MBD2 and p66α

MBD2 mRNA can be read to give two isoforms, MBD2a and MBD2b. MBD2b shares high degree of homology with MBD3, whereas MBD2a has an additional N-terminal 140 amino acids which has several glycine-arginine rich repeats; the function of which is unknown *(11)*. Human MBD2b is 262 amino acid long and has two distinct domains: N-terminal MBD, which binds to methylated DNA *(12)* and C-terminal coiled coil which has been shown to interact with p66α *(1)*. The p66α/β subunits from NuRD are 66kDa transcriptional repressors that interact directly with MBD2, MBD3, and histones *(13-15)*. p66α has a N-terminal conserved region (CR1) that covers a coiled coil domain and a C-terminus conserved region (CR2) that has a GATA like zinc finger domain. Domain organization of MBD2 and p66α are shown in Figure 1.1B and 1.1C, respectively.

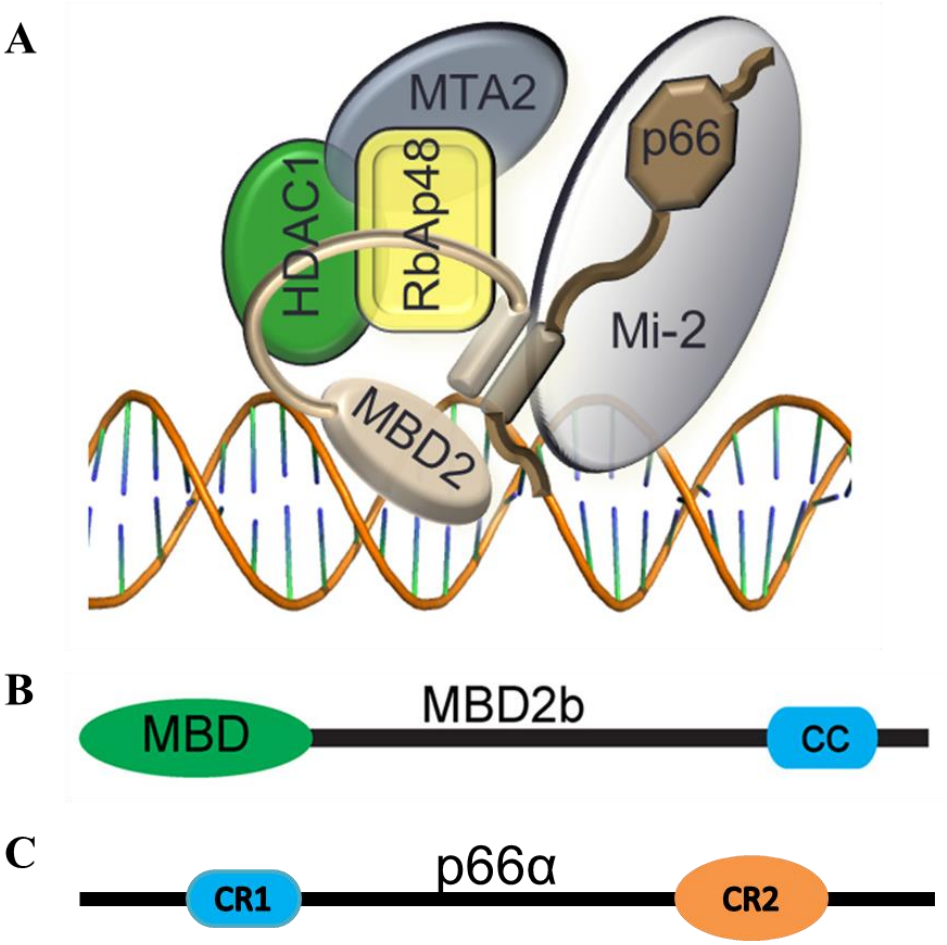


Figure 1.1: Components of MBD2-NuRD. A) Schematic representation of the MBD2-NuRD complex. B) Domain organization of MBD2b: MBD is methyl binding domain; CC is coiled-coil domain C) Domain organization of p66α: two conserved regions are seen. CR1 is a coiled coil domain; CR2 includes a GATA zinc finger domain.

Table 1.1: Summary of MBD2-NuRD complex components.

Protein	Function	Conserved Domains
Mi2α/β (CHD3/CHD4)	large multi-domain protein with a central helicase-like ATPase that functions in chromatin remodeling	SNF2_N (DEXHc), HELICc, CHDN, PHD, CHROMO, CHDCT2, DUF
HDAC1/2	histone deacetylase enzyme	HDAC
RbAp46/48 (RBBP7/RBBP4)	bind HDACs and histone tails and facilitate deacetylase activity	WD40
MTA1/ 2	associated with histone as well as non-histone deacetylase complexes	BAH, ELM2, SANT, GATA zinc finger
p66α/β (GATAD2A/B)	bind MBD2 and histones	Coiled-coil, GATA zinc finger
MBD2	binds methylated CpG	MBD and coiled-coil

1.2.4 Pivotal role of MBD2-p66α interaction in MBD2 mediated silencing

MBD2 is associated with repression of several genes. For example, repression of chicken ρ-globin expression is under control of MBD2 mediated silencing. siRNA knock-down of chicken MBD2 leads to a 25-fold increase in ρ-globin expression. Knockout of the MBD2 gene in βYAC transgenic mice leads to persistent γ-globin expression in the adult mouse *(16, 17)*. Coiled coil interaction between MBD2 and p66α is central to formation of functional NuRD complex. We recently demonstrated that the highly conserved and homologous C-terminal coiled-coil regions of MBD2 and MBD3 form a high-affinity heterodimeric complex with p66α critical for recruitment of the Mi2 protein and methylation dependent gene silencing *in vivo.* Also, enforced expression of the isolated p66α coiled-coil domain relieves MBD2-mediated globin gene silencing and the expressed peptide interacts only with a subset of components of the MBD2-NuRD complex that does not include native p66α or Mi-2. These results demonstrate the central importance of the coiled-coil interaction and suggest that MBD2-dependent DNA methylation-driven gene silencing can be disrupted by selectively targeting this coiled-coil complex *(1)*.

1.2.5 Coiled-coil: a common structural motif in nature

The coiled-coil domain represents a relatively simple yet common protein:protein interaction motif found in as many as 10% of all eukaryotic proteins *(18)*. Recent work has shown that selective disruption of coiled-coil complexes can target specific protein complexes for potential therapeutic benefit *(1, 19)*. Coiled-coils form specific homo or hetero-oligomeric complexes involving 2-7 α-helices in parallel or anti-parallel arrangements important for a wide variety of cellular functions either on their own or as a part of larger protein complexes *(18, 20-*

22). Most studies to date have described the formation of parallel homo-oligomeric coiled-coils while the anti-parallel heterodimeric coiled-coil complexes are relatively understudied *(19, 22)*. Note that coiled-coil interaction between MBD2 and p66α is heterodimeric and anti-parallel thus relatively understudied. A coiled-coil domain can be identified by a regular seven amino acid (**a-g**) repeat of hydrophobic and charged residues. In this heptad repeat, a branched hydrophobic residue is present at **a** and **d** positions while charged/polar residues are present at **e** and **g**. These seven residues form approximately two turns of a typical α-helix, generating a hydrophobic face (**a** and **d**) bordered by charged/polar residues (**e** and **g)**. Two or more of these α-helices bind along this hydrophobic interface and, due to the natural rotation of this surface, the helices tend to wrap around one another forming a 'coil of coils.' This arrangement is capable of forming either parallel or anti-parallel hetero- or homo-oligomeric complexes ranging from 2 to 7 helices. Despite this seemingly simple paradigm, subtle variations in sequence can have dramatic consequences on binding specificity, stoichiometry, and parallel vs. anti-parallel alignment *(23)*.

1.2.6 Coiled-coils of MBD2, p66α and homologues of MBD2

Out of family of five MBD binding proteins, MBD2 and MBD3 are the only proteins with a coiled-coil domain which is approximately 34 amino acids long. There are two other proteins, MBD3-like protein 1 (MBD3L1) and MBD3-like protein 2 (MBD3L2) which contain homologous coiled-coil domain but lack the methyl-CpG-binding domain. MBD3L1 and MBD3L2 have been shown to be capable of recruiting an intact NuRD complex. MBD3L2 is expressed in germ cell tumors and some somatic tissues while MBD3L1 is testis specific and expressed in post-meiotic spermatids *(24-26)*. The coiled-coil of p66α is approximately 41 amino acids long and binds in anti-parellel fashion to MBD2 homologues. Domains organization of coiled coils of MBD2 homologues is depicted in Figure 1.2A. In Figure 1.2B shows sequence

alignment and the key residues that are involved in making hydrophobic and polar/ionic interactions with the coiled-coil of p66α are highlighted. Given the similarities and differences between these homologous domains, we have pursued detailed analysis of the different homologues to gain a better understanding of the structural determinants for high affinity binding.

In the studies presented here, we show that high affinity binding requires pre-formed helical content as well as specific charged residues on the individual coiled-coil domains. The reduction in helical content of the isolated MBD3L1 and MBD3L2 homologues reduces binding affinity for p66α. We previously demonstrated that changing the charge of three residues in p66α eliminates binding to wild type MBD2 *(1)*. Introducing complementary charge changes in MBD2 restores binding, but not with the same high affinity as wild type. Based on electrostatic potential calculations, we suggest that the uniquely high affinity association of the wild type complex depends on complementary alternating positive and negative electrostatic potential surfaces. Hence variations in both the helical content and electrostatic interactions between MBD2 homologues lead to a relative binding affinity hierarchy for p66α (MBD2 ≈ MBD3 > MBD3L1 ≈ MBD3L2).

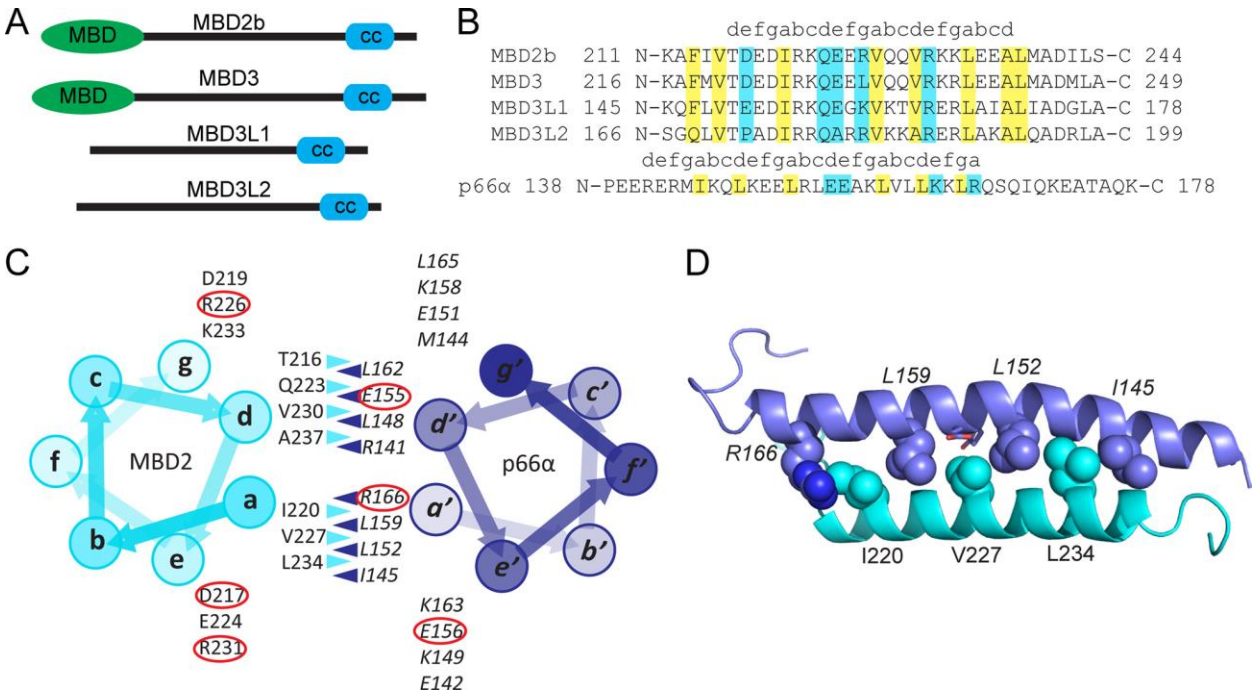


Figure 1.2: The coiled-coil interaction between MBD2 homologues and p66α: The domain organization (A) is diagramed for MBD2 and homologues, which shows that the MBD3L1 and MBD3L2 proteins lack a methyl-cytosine binding domain. A sequence alignment (B) of the coiled-coil (*cc*) domains from p66α and MBD2 homologues is shown with key hydrophobic (*yellow*) and ionic/polar (*cyan*) contact residues highlighted and the heptad repeat (**a-g**) indicated above the amino acid sequences. A helical wheel diagram of the complex (C) highlights the interacting residues at positions **a**, **d**, **g**, and **e** of the heptad repeat with key charged residues circled in red. A ribbon diagram of the p66α-MBD2 coiled-coil complex (D) is shown with the branched hydrophobic residues at the **a** position on each chain depicted as spheres and the two central glutamates (*E155* and *E156*) of p66α depicted as sticks.

1.3 Materials and methods

1.3.1 Protein expression and purification

The coiled-coil regions of human MBD2b (amino acids 211-244), MBD3 (amino acids 216-249), MBD3L1 (amino acids 145-178), MBD3L2 (amino acids 166-199) and p66α (amino acids 138-178) were cloned and expressed with a hexahistidine tag and as thioredoxin fusion proteins in a modified pET32a vector *(27)*. The expression vectors were transformed into the BL21(DE3) *E. coli* strain, grown in Luria Bertani medium at 37 ^{0}C and induced with 1 mM isopropyl-β-d-thiogalactopyranoside at an A_{600} ~ 0.8. The bacteria were harvested after 2 hours of induction and lysed with the B-PER reagent (Thermo Scientific). The soluble fraction was passed over a nickel-sepharose column, protein eluted with a step gradient of imidazole and further purified by gel filtration over a Superdex-75 column (GE Healthcare). The thioredoxin fusion proteins were used directly for analytical ultracentrifugation (AUC) and isothermal titration calorimetry (ITC) studies. For circular dichroism (CD) studies, clones were modified to incorporate a tyrosine residue just after the thrombin cleavage site (for quantification of the isolated peptide by UV) and were expressed in a similar manner. After purification over a nickel-sepharose column, the peptides were cleaved by thrombin digest and isolated by gel filtration chromatography over a Superdex-75 column (GE Healthcare). Specific mutations were introduced using the QuickChange® site-directed mutagenesis kit (Stratagene) following the manufacturer's protocol. The final concentrations of all protein samples were determined by UV absorbance at 280 nm.

1.3.2 Analytical ultracentrifugation

Sedimentation velocity experiments were carried out using a Beckman Optima XL-I analytical ultracentrifuge (Beckman Coulter Inc.) equipped with a four and eight-position AN-60Ti rotor. Sedimentation was performed at 40,000 rpm, 20 ºC, under physiological buffer conditions (20 mM Tris pH 8.0, 150 mM NaCl). Sedimentation profiles were recorded using UV absorption (280 nm) and interference scanning optics. The partial specific volume ($\bar{V}$) of the sample, density (ρ) and viscosity (η) of the buffer were calculated using the SEDNTERP program *(28)*. Data were fit using a continuous size distribution (c(s)) and the effective molecular weight determined from the resulting sedimentation coefficients with the SEDFIT software *(29)*.

1.3.3 Isothermal titration calorimetry

Protein samples were prepared in standard buffer (20 mM Tris pH 8.0, 150mM NaCl) and binding analyzed with an iTC_{200} Microcalorimeter (GE Healthcare). A total of 24 injections (1.5 µL each) of the p66α coiled-coil (100 µM) were injected into MBD2 and homologues (10 µM, 298 K, stir speed of 400 rpm, 120 seconds time delay between injections). The resulting isotherms were auto adjusted for baseline and fit to a one-site binding model using Origin 7.0 software to determine binding constant (K_D , Ka^{-1}) and enthalpy (ΔH) while the Gibbs free energy (ΔG) and entropy (ΔS) of binding were calculated according to Equation 1.1,

$$-RTln(K) = \Delta G = \Delta H - T\Delta S \qquad \text{(Equation 1.1)}$$

where T is the temperature in Kelvin and R is the gas constant.

1.3.4 Circular dichroism

CD spectra were collected on purified peptide samples (~33 µg/mL total protein in 10mM sodium phosphate, pH 6.5) with a JASCO J-715 CD spectrometer (JASCO Corp) at 293 K, with

a 1 cm path length, scanning from 190-260 nm with 0.5 nm interval at a scanning speed of 50 nm/min. CD spectra were normalized to give molar ellipticity values (θ) in degrees·cm^{-2}·$dmol^{-1}$$residue^{-1}$. Helical content for each peptide was calculated from the ratio of the observed θ_{222nm} to the expected θ_{222nm} for 100% helix as given by 40,000 × [(n − 4)/n], where n is the number of residues *(30)*. Thermal denaturation was followed at θ_{222nm} from 277-368K at 1 K intervals with a heating rate of 1 K/min. The data were fit to a simple two state thermodynamic model of unfolding as described by Koepf *et al. (31)* .

1.3.5 Helical content prediction

The expected helical content for each peptide was calculated using the AGADIR *(32, 33)* algorithm with the N- and C-termini 'free', at 293 K, ionic strength of 0.02, and pH 6.5 to closely match the experimental conditions for CD. The predicted helical content was used to help design amino acid changes that stabilize helix formation.

1.3.6 Electrostatic surface potential

The coordinates of the isolated wild type coiled-coil domains were extracted from the previously determine solution structure (PDB# 2L2L) and the surface potential calculated with the Adaptive Poisson-Boltzmann Solver (APBS) *(34)*. The coiled-coil domains of the different mutants and homologous domains were derived from the wild type MBD2 structure by introducing appropriate sequence differences with the mutagenesis function of PyMOL *(35)* and choosing a sidechain rotamer that did not sterically collide with neighboring residues. The calculated electrostatic potential was mapped to the solvent accessible surface with the APBS plugin tool in PyMOL and colored from red to blue (-1 eV to +1 eV, respectively).

1.4 Results

1.4.1 Key contact residues are conserved among MBD2 homologues

The key contact residues in the MBD2 coiled-coil domain are conserved across all homologues, with MBD3L2 composed of the most divergent sequence (44% identity between MBD2 and MBD3L2 coiled-coil domains), whereas MBD3 and MBD2 are nearly identical (93% identity). Anti-parallel coiled-coils form sequential intermolecular interactions between branched hydrophobic residues at **a** and **a'** as well as **d** and **d'** positions of the heptad repeat in the two chains. Recent work indicates that select triplet repeats at the **a'-a-a'** positions favor heterodimeric coiled-coil formation, with LIL or ILI triplets the most favored combination *(36)*. As can be seen in Figure 1.2, the **a** positions are identical across all MBD2 homologues with the **a-a'** interactions composed by *RILVLLI* (p66α residues are in italics throughout). This arrangement includes two favorable I-L pairings and one of the more favorable triplets, LVL. The highly conserved valine residue of this triplet inserts into a pocket between two conserved glutamate residues at a central bend in the p66α helix (Figure 1.2D). The shorter valine side chain (as compared with isoleucine and leucine residues) allows close approximation of the two helices, which likely contributes to close intermolecular ionic interactions involving the glutamate residues. The conserved **d-d'** interactions *(37)*, which are composed of *L*Q*E*V(A)*L*A, also place a valine (or alanine in MBD3L2) at the bend on the p66α helix near these same glutamate residues.

1.4.2 Coiled-coils of MBD2, MBD3, MBD3L1 and MBD3L2 are largely monomeric in isolation

Although coiled-coil domains often form homo-oligomeric interactions, we previously demonstrated that both MBD2 and p66α remain monomeric in isolation *(1)*. To test for homo-oligomerizaton of the different homologues as well as the concentration dependence of homodimerization of MBD2 and p66α, we carried out sedimentation velocity AUC studies. The MBD2 coiled-coil domain remains monomeric even at concentrations up to 300 μM (Figure 1.3A). On the other hand, the p66α coiled-coil shows a tendency to form a homodimer at concentration beyond 50 μM (Figure 1.3B); however, the monomer remains the dominant species up to 300 μM. Given the low nanomolar binding constant between MBD2 and p66α coiled-coil domains, p66α preferentially forms a stable heterodimer with MBD2 rather than a homodimer. Similarly, AUC analyses showed that the coiledcoil domains of MBD3, MBD3L1, and MBD3L2 homologues remain stable monomers at 50 μM concentrations (Figure 1.3C).

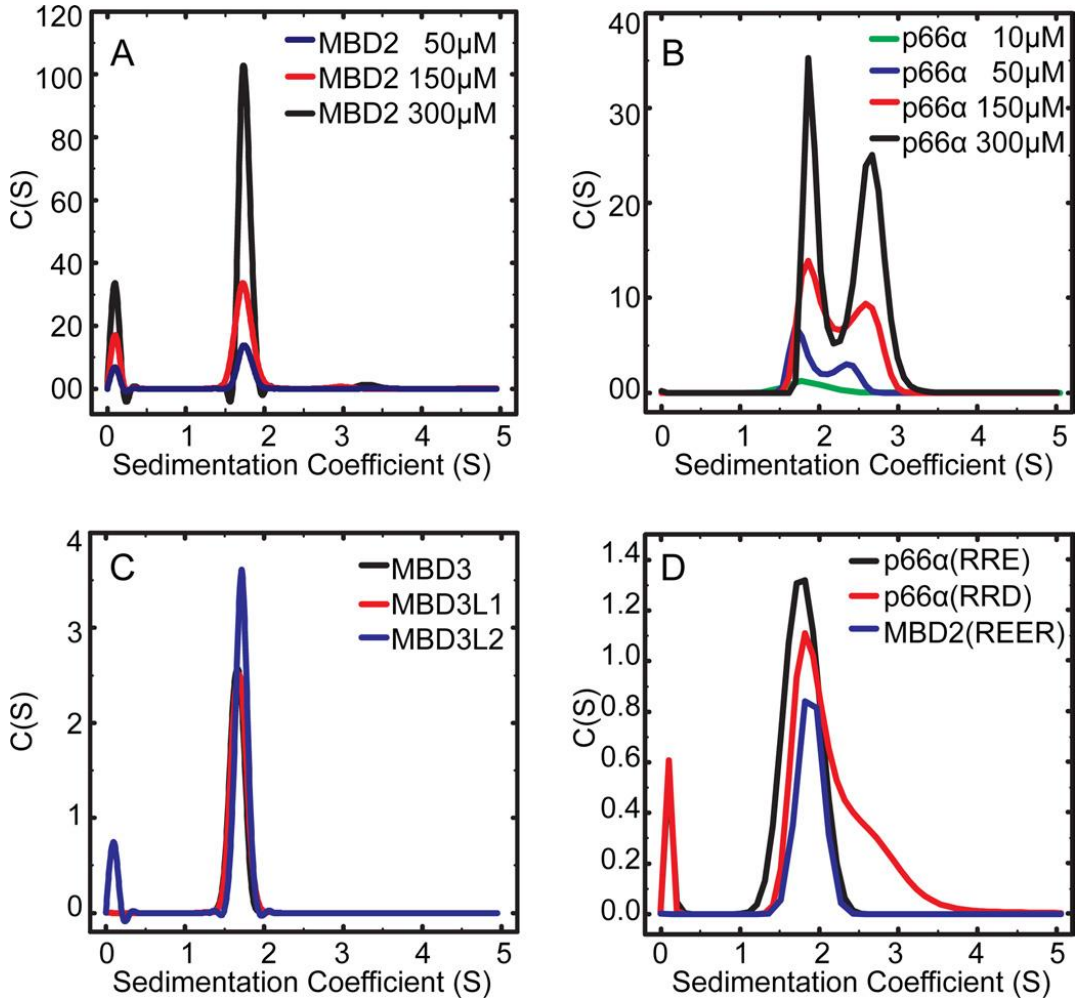


Figure 1.3: The coiled-coil domains remain largely monomeric in isolation. Analytical ultracentrifugation analysis was performed on the individual coil domains and the sedimentation velocity fit using a continuous size distribution (*c*(*s*)). The results are shown for increasing concentrations of MBD2 (*A*) and p66α (*B*) coiled-coil domains as well as for 50 μM concentrations of the coiled-coil domains from MBD3, MBD3L1, and MBD3L2 (*C*) and MBD2 and p66α mutants (*D*).

1.4.3 Binding analysis of MBD2 homologues reveals a hierarchial affinity preference for the p66α coiled-coil

ITC was performed using thioredoxin fusion constructs of the coiled-coil domains. Exothermic heat was generated with each injection in all experiments. The binding isotherms are shown in Figure 1.4, and the measured binding affinity (K_D), free energy (ΔG), enthalpy (ΔH), and entropy ($-T\Delta S$) for each homologue is provided in Table 1.2. These results show that p66α binds with higher affinity to MBD3 ($K_D = 23 \pm 3$ nM) and MBD2 ($K_D = 42 \pm 9$ nM) as compared with MBD3L1 ($K_D = 377 \pm 34$ nM) and MBD3L2 ($K_D = 268 \pm 32$ nM). Each complex binds with a stoichiometry of ~ 1:1 (n ranges from 0.7 to 1.4, Table 1.2) consistent with heterodimer formation. The reduced binding affinity of p66α for the MBD3L1 and MBD3L2 homologues reflects a more unfavorable change in entropy upon binding ($-T\Delta S$ = 0.17 and 6.3 kcal/mol for MBD2 and MBD3L1, respectively) that is not fully compensated by a more favorable change in enthalpy (ΔH = -10.2 and -15.1 kcal/mol for MBD2 and MBD3L1, respectively).

Although the coiled-coil domains do not contain a histidine residue or other titratable protons at a pH of 8.0, the high ionization enthalpy of Tris buffer (11.4 kcal/mol) *(38)* could contribute to the apparent enthalpy change upon binding. To address this possibility, we repeated ITC for MBD2-p66α in PIPES buffer (20 mM PIPES, pH 7.5, 150mM NaCl), which has a much lower ionization enthalpy (2.7 kcal/mol) *(38)*. The binding constant and change in enthalpy are very similar in PIPES ($K_D = 30 \pm 11$ nM, $\Delta H = -10.8 \pm 0.2$ kcal/mol, $-T\Delta S$ = 0.56 kcal/mol), which indicates that complex formation does not involve net transfer of a proton.

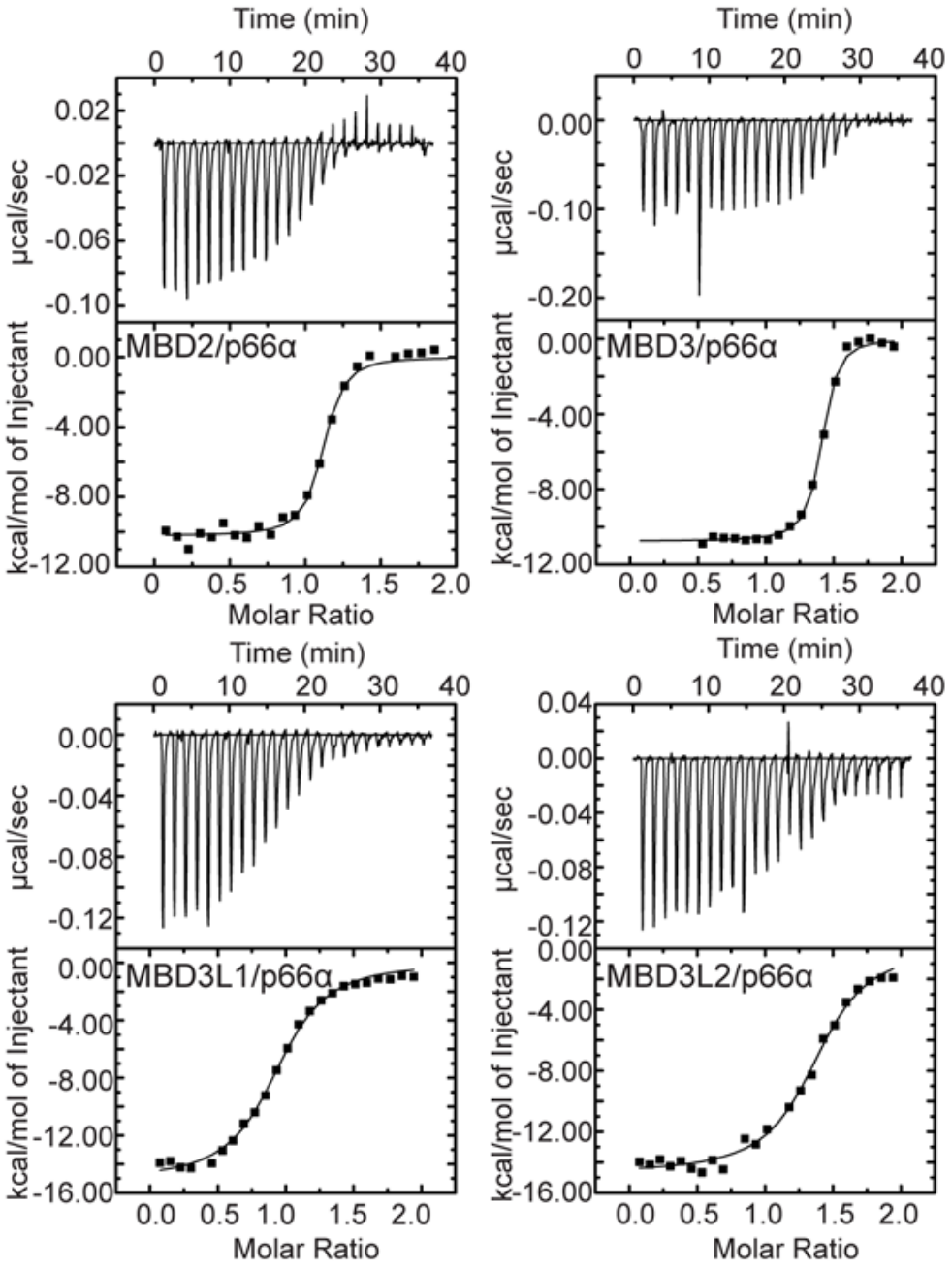


Figure 1.4: Binding analysis of MBD2 homologues. Isothermal titration calorimetry studies were performed and the experimental data (*top panel*) and resulting fit (*bottom panel*) are shown for MBD2, MBD3, MBD3L1, and MBD3L2 coiled-coil domains binding to the p66α coiled-coil domain.

Table 1.2: Binding affinity analyses. The dissociation constant (K_D), change in enthalpy (ΔH), entropy ($-T\Delta S$), Gibbs free energy (ΔG), and apparent stoichiometry (n) derived from isothermal titration calorimetry studies are given for wild type and mutant coiled-coil complexes between p66α and MBD2 homologues.

Coiled-coil complex	n	K_D (nM)	ΔH (kcal/mol)	-TΔS (kcal/mol)	ΔG (kcal/mol)
MBD2/p66α	1.1	42 ± 9	-10.2 ± 0.1	0.17	-10.0 ± 0.1
MBD3/p66α	1.4	23 ± 3	-10.8 ± 0.1	0.35	-10.4 ± 0.1
MBD3L1/p66α	0.9	377 ± 34	-15.1 ± 0.2	6.3	-8.8 ± 0.1
MBD3L2/p66α	1.4	268 ± 32	-14.7 ± 0.2	5.7	-9.0 ± 0.2
MBD2(REE)/p66α(RRE)	0.7	10,800 ± 400	-27.1 ± 0.4	20.3	-6.8 ± 0.4
MBD2(REE)/p66α(RRD)	1.2	38,000 ± 2000	-13.9 ± 1.1	7.9	-6.0 ± 1.1
MBD2(REER)/p66α(RRE)	0.8	5,400 ± 500	-20.0 ± 0.7	12.8	-7.2 ± 0.7
MBD2(REER)/p66α(RRD)	0.7	33,000 ± 9800	-16.3 ± 1.1	10.2	-6.1 ± 1.1
MBD2(E225G)/p66α		≥ 50,000			
MBD3L1(G159E)/p66α	1.3	44 ± 20	-22.2 ± 0.5	12.2	-10.0 ± 0.5

1.4.4 High affinity binding depends on the helical content of the isolated coiled-coil domains

We previously demonstrated that the MBD2 and p66α coiled-coil domains show a strong tendency to form monomeric helices in isolation *(1)*. An algorithm based on helix-coil transition theory (AGADIR *(32, 33)*) predicts that the MBD3L1 and MBD3L2 homologues do not have the same tendency to form α-helices in isolation (Table 1.3). CD analyses were performed on the isolated domains, which confirmed the relative helical content of the homologous coiled-coil domains in isolation. MBD2 (25%) and MBD3 (28%) are more helical than MBD3L1 (7%) and MBD3L2 (11%) (Figure 1.5*A*). The thermal stability of the different coiled-coil complexes was determined by following molar ellipticity at 222 nm (θ_{222} nm) as a function of temperature. Complexes involving MBD2 and MBD3 melt at a higher temperature than those involving MBD3L1 and MBD3L2, consistent with the higher binding affinities of MBD2 and MBD3.

To test whether helical content dictates high affinity association, we introduced mutations at residues opposite the binding interface of the coiled-coil domains of MBD2 and MBD3L1 that reduce or increase helical content, respectively. A glycine for glutamate substitution in the middle of the helix opposite the binding interface could contribute to the reduced helical content of MBD3L1 (Figure 1.5*B*). Consistent with *in silico* calculations, the G159E mutation of MBD3L1 increases helical content (7% to 16%, Table 1.3), whereas the E225G mutation of MBD2 reduces helical content (25% to 9%, Table 1.3). As expected, the binding affinity for the p66α coiled-coil domain (Table 1.2) and the melting temperature of the complex (Table 1.4) increased for MBD3L1 G159E (K_D = 44 nM, T_m = 331 K), which is close to the affinity of wild-type MBD2 (K_D = 42 nM, T_m = 338 K) and much greater than wild-type MBD3L1 (K_D = 377

nM, T_m = 319 K). In contrast, the binding affinity and melting temperature of MBD2 E225G (K_D > 50 μM, T_m = 313 K) was greatly decreased as compared with wild type. Because this residue is on the side of the helix opposite the binding surface and does not directly interact with p66α, these findings support the hypothesis that high affinity association requires pre-existing helical content.

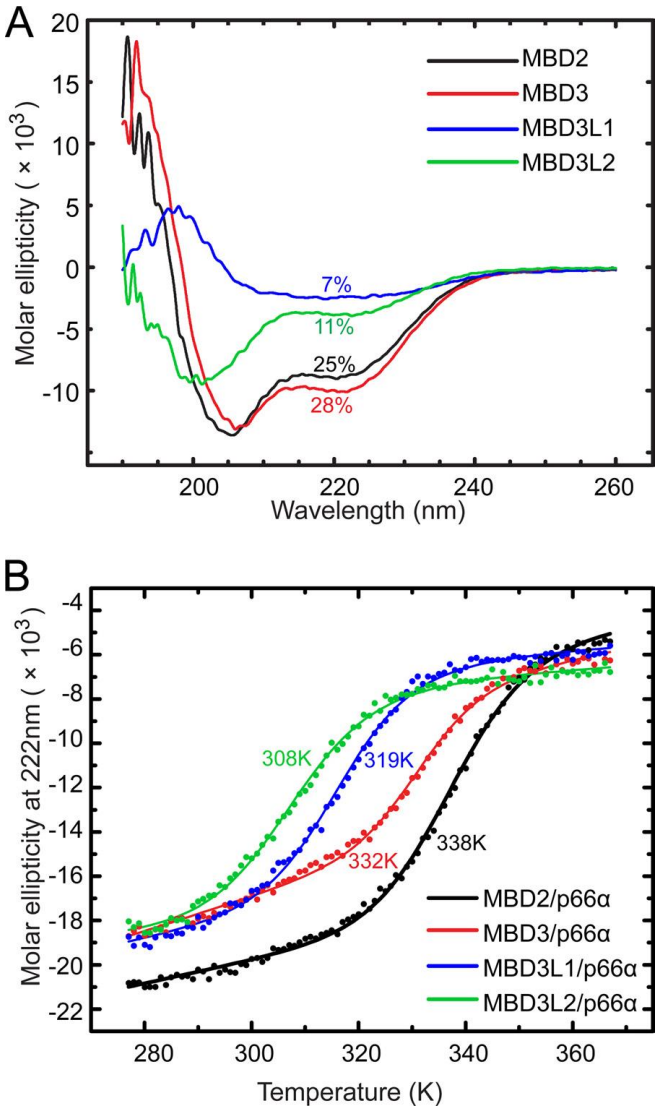


Figure 1.5: The helical content and thermal denaturation of MBD2 homologues. Circular dichroism spectra (*A*) of the coiled-coil domains from MBD2 homologues in isolation are shown and labeled with the helical content as calculated from θ_{222nm}. The temperature dependence of θ_{222nm} (*B*) is shown from 277 to 368 K for coiled-coil complexes between the MBD2 homologues and p66α. The data were fit to a simple two-state unfolding model *(31)*, and the resulting thermal denaturation curves labeled with the melting temperature (T_m).

Table 1.3: Helical contents of the isolated coiled-coil domains. The percent helix as predicted by AGADIR and calculated from the circular dichroism molar ellipticity at 222 nm ($\theta_{222\ nm}$) is given for the wild type and mutant coiled-coil domains.

Coiled-coil domain	Helical propensity in isolation	
	Predicted* (%)	Calculated** (%)
MBD2	39.5	24.6
MBD3	48.0	28.3
MBD3L1	11.2	6.7
MBD3L2	7.5	10.7
MBD2(E225G)	11.1	8.9
MBD3L1(G159E)	24.8	15.5
MBD2(REE)	7.0	9.1
MBD2(REEE)	20.6	35.8
P66α	55.3	66.0
P66α(RRE)	56.1	59.6

*Based on AGADIR algorithm
**Based on circular dichroism measurements

Table 1.4: Thermal stability of the complexes. The melting temperatures (T_m) derived from circular dichroism studies are given for wild type and mutant coiled-coil complexes between p66α and MBD2 homologues.

Coiled-coil complex	T_m (K)
MBD2/p66α	338
MBD3/p66α	332
MBD3L1/p66α	319
MBD3L2/p66α	308
MBD3L1(G159E)/p66α	331

1.4.5 Specific ionic interactions are required for high affinity coiled-coil complex formation

We recently demonstrated that mutating specific charged residues of p66α reduces binding affinity by 3 orders of magnitude when introduced separately (either *E155R/E156R* or *R166E*) and abolishes complex formation when introduced together (*E155R/E156R/R166E*)*(1)*.These residues form close ionic interactions with three charged residues in MBD2 (Asp-217, Arg-226, and Arg-231) that are conserved across all MBD2 homologues (Figure 1.2B) and likely provide specificity for the coiled-coil interaction. Based on this observation, we hypothesized that introducing complementary changes in MBD2 would restore high affinity binding.

Introducing the D217R/R226E/R231E mutations in MBD2 (MBD2(REE)) does restore binding to the and p66α *E155R/E156R/R166E* mutant (p66α(RRE)). However, ITC analysis (Figure 1.6) indicates a much lower affinity between the mutant proteins (*KD* = 10.8 μM) than wild type. *In silico* calculations with AGADIR as well as CD measurements show that mutating these residues decreased the α-helical content of MBD2 (9% *versus* 25%, Table 1.3). These changes introduced an unfavorable charge interaction between Glu-231 and Glu-235 residues of MBD2(REE). To increase helicity, a fourth mutation (E235R) was introduced (as predicted by AGADIR) without disrupting intermolecular contacts. The D217R/R226E/R231E/E235R mutant MBD2 (MBD2(REER)) did show an increase in helical content (36% *versus* 9%, Table 1.3), and ITC revealed a slight increase in binding affinity (5.4 μM) for p66α(RRE) as compared with MBD2(REE) but still approximately 2 orders of magnitude lower than for the wild type complex.

One potential difference between the wild type and mutant complexes is that the favorable Arg-166-Asp-217 p66α-MBD2 ionic interaction was replaced by Glu-166-Arg-217 in the mutant complex. The additional carbon atom in glutamate (as compared with aspartate) could alter the geometry and prohibit favorable interaction. However, both MBD2(REE) and MBD2(REER) bound to p66α(RR**D**) with lower affinity than p66α(RRE) (Table 1.2), demonstrating that this difference was not responsible for the lower binding affinity.

Alternatively, swapping glutamate and arginine residues between chains could alter the geometric relationship between the charged residues by changing the relative positions of the two residues on the helical backbone. Changes in the relative positions of these residues could preclude ideal interaction and prevent high affinity association. Geometrical restraints are particularly important for hydrogen bond formation. Hence, to probe this possibility we tested whether geometrically restrained bidentate hydrogen bonds could be formed between the p66α-MBD2 intermolecular ionic pairs of Arg-166-Asp-217 or Glu-156-Arg-231 and the respective charge swap mutations. We introduced hydrogen bond distance and angle restraints and performed simulating annealing calculations (XPLOR_NIH *(39)*) while keeping the coordinates fixed for all backbone atoms and the side-chain atoms for all amino acids exclusive of the four under consideration. As can be seen in Figure 1.7, reasonable hydrogen-bond distances and angles can be established between these side chains in both the wild type and mutant complexes, suggesting that geometric constraints do not prevent optimal interaction in either case.

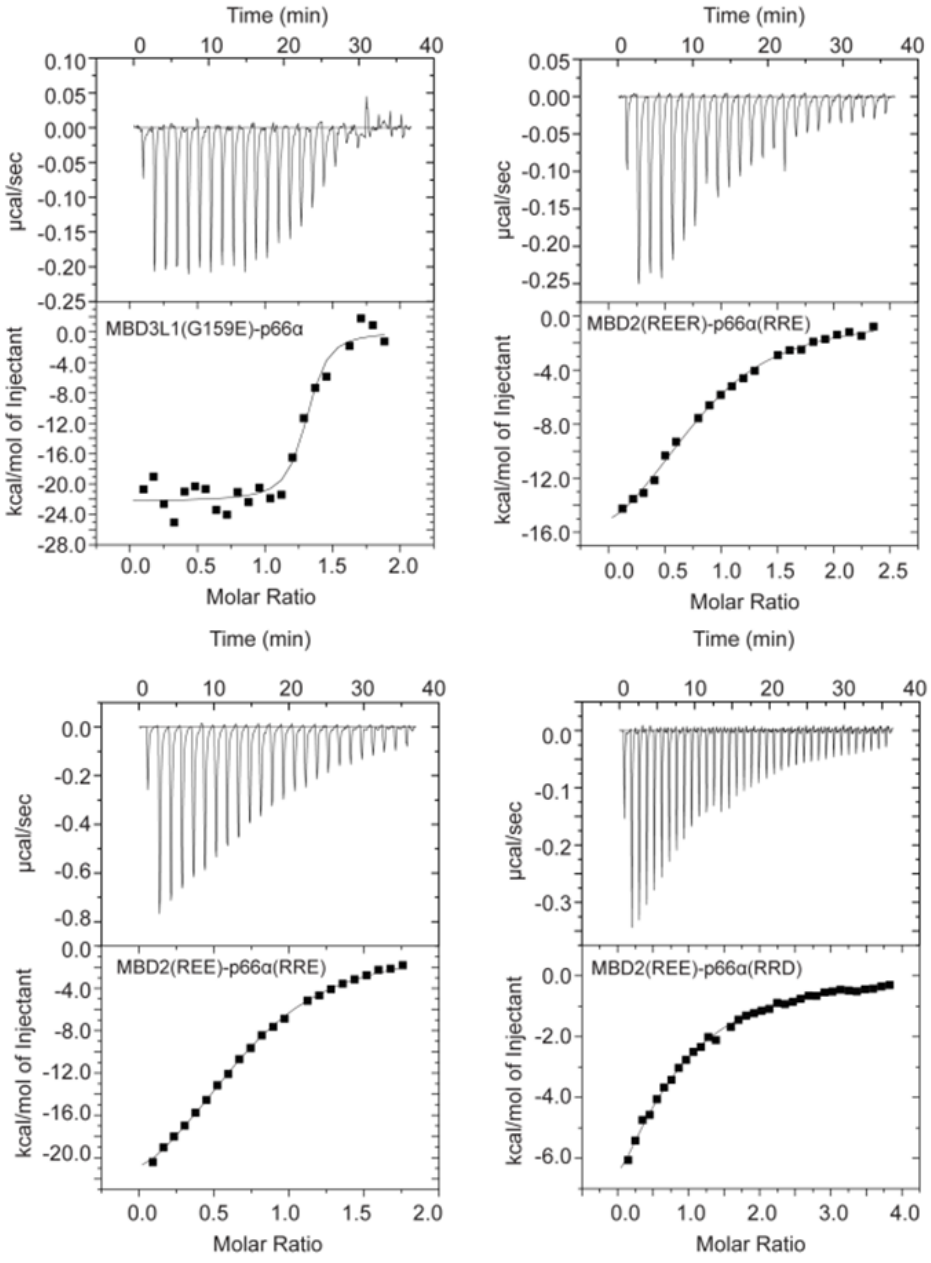


Figure 1.6: Binding analyses of mutants of MBD2 and MBD3L1. Isothermal titration calorimetry studies were performed and the experimental data (*top panel*) and resulting fit (*bottom panel*) are shown for the coiled-coil domains of MBD3L1(G159E) binding to p66α, MBD2(REE), and MBD2(REER) binding to p66α(RRE) and MBD2(REE) binding to p66α(RRD).

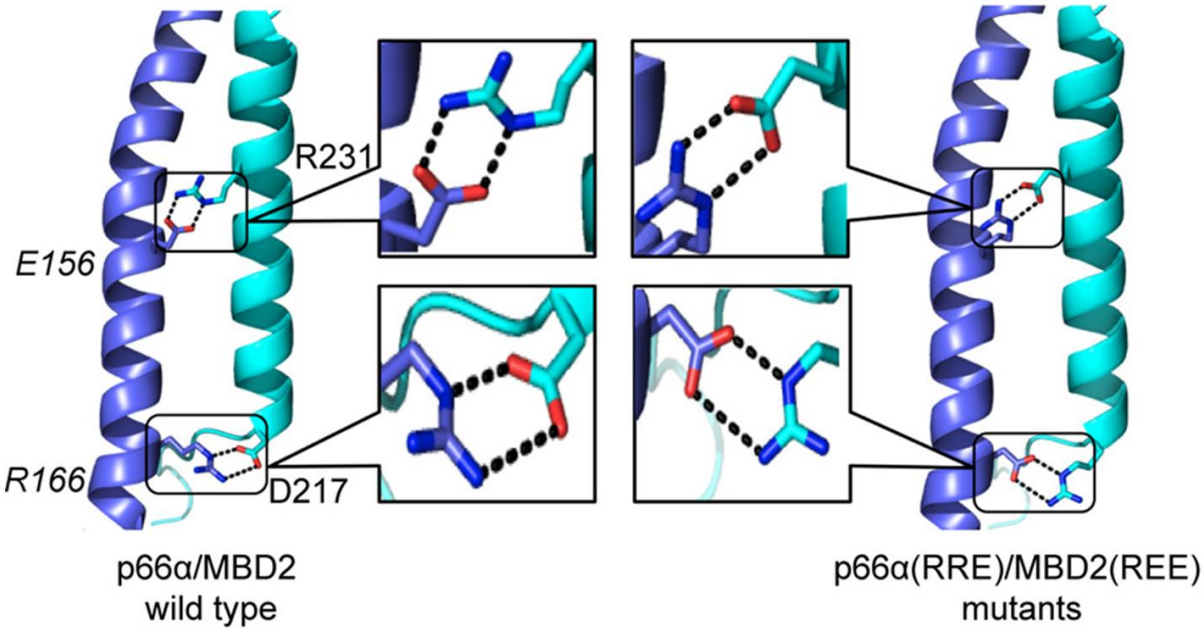


Figure 1.7: Ionic interactions of the coiled-coil complex**.** To test whether geometric constraints prevent close interaction between charged residues in the charge swap mutant proteins, bidentate hydrogen bond restraints between p66α-MBD2 residues Arg-166-Asp-217 and Glu-156-Arg-231 (and the respective charge swap mutations) were incorporated, and the side-chain conformations were minimized by simulated annealing molecular dynamic calculations. Representative minimized structures are shown for both the wild type and mutant complexes; the minimized charged side chains are depicted as *sticks*, and the bidentate hydrogen bonds are indicated by *dashed lines*.

1.4.6 The MBD2 and p66α coiled-coil domains have highly complementary electrostatic surface potentials

To investigate differences between the wild type and charge swap complexes, the electrostatic potential of the individual peptides were calculated by the APBS *(34)* and mapped to the surface with the APBS plugin tool in PyMOL *(35)*. This analysis reveals that the wild type peptides have complementary alternating positive (*blue*) and negative (*red*) surface potentials (Figure 1.8A). The MBD3, MBD3L1, and MBD3L2 homologues show a very similar pattern (Figure 1.8C). The **a** and **a'** hydrophobic "knobs" are largely positioned where the surface potentials change from positive to negative (indicated by *arrows* in Figure 1.8A).

The electrostatic surface potential of the charge swap mutations are highly complementary as well (Figure 1.8B), indicating that the chosen mutations did restore the specific charge interactions. However, the interaction surfaces on each of the mutant peptides are more uniformly positive (p66α) or negative (MBD2). Hence the hydrophobic knobs are now positioned within a more uniform electrostatic charge potential.

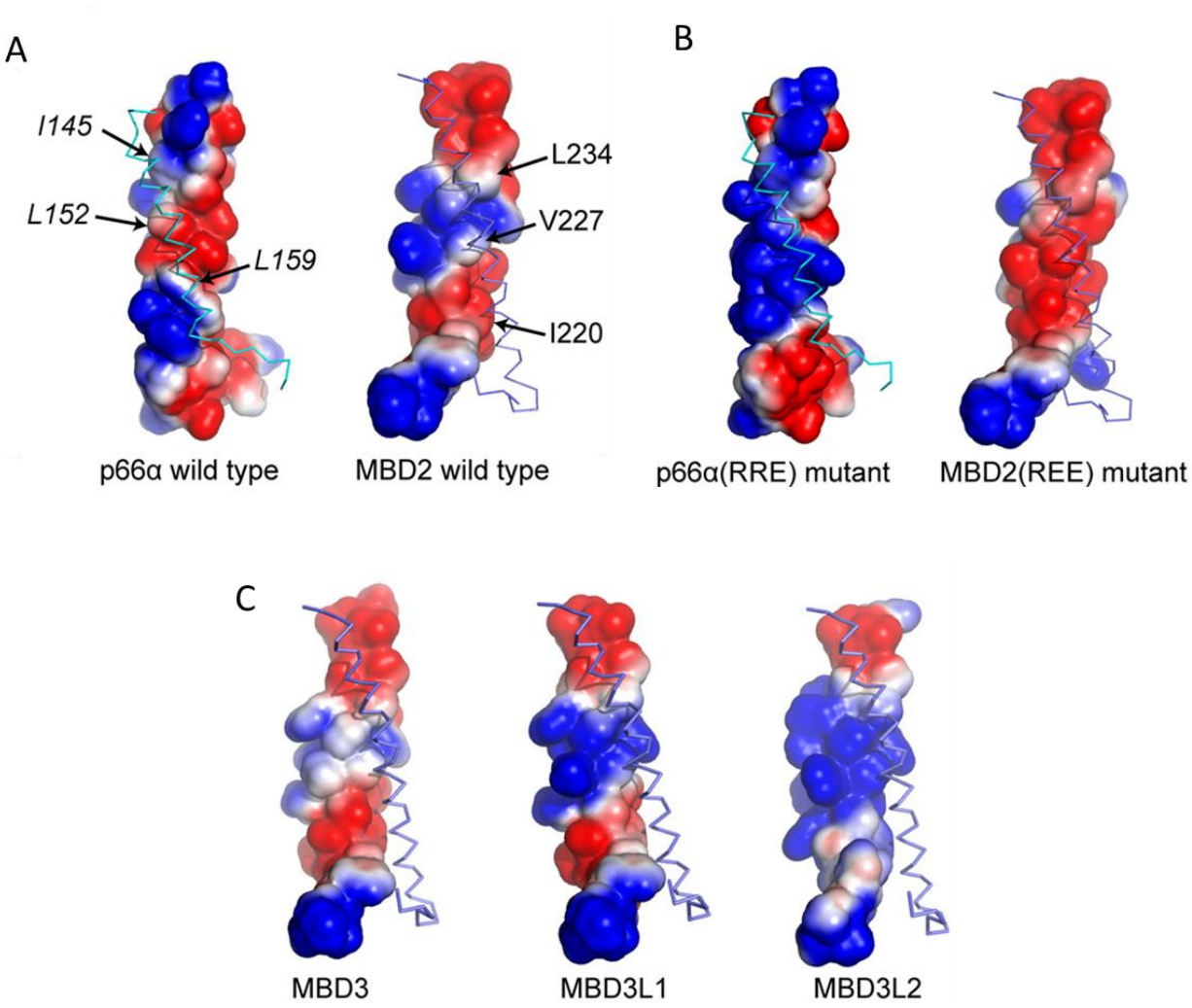


Figure 1.8: Electrostatic surface potentials of the coiled-coil complex. The electrostatic surface potentials for wild type (A) and mutant (*B*) p66α and MBD2 coiled-coil domains as well as for MBD3, MBD3L1, and MBD3L2 coiled-coil domains (C) were calculated with the APBS *(34)* tool in PyMOL *(35)* and the surface potential colored from r*ed* to *blue* (-1 to +1 eV). The intermolecular contact surface is shown for each with the binding partners depicted as a ribbon diagram for orientation. The location of hydrophobic residues at **a** and **a'** positions are indicated with *arrows* (*A*) for the wild type peptides.

1.5 Discussion

Here we have presented biophysical analyses of the heterodimeric anti-parallel coiled-coil complex between p66α and MBD2 homologues. This represents one of the few studies of a native anti-parallel coiled-coil complex and, to our knowledge, the only such study comparing a family of homologous coiled-coil domains. The p66α-MBD2 complex demonstrates several unusual features including the propensity of the individual domains to form monomeric helices in isolation, a clear lack of oligomerization for MBD2 even at fairly high concentrations (300 μM), and a requirement for minimum helical content in isolation to bind with high affinity.

Comparing the different homologous coiled-coil domains, we find that the MBD2 and MBD3 domains bind with an ~10-fold greater affinity than either the MBD3L1 or MBD3L2 domains. This difference reflects a larger unfavorable change in entropy that is not fully compensated by a more favorable change in enthalpy when comparing MBD3L1 and MBD3L2 with MBD2. These changes correlate with the observation that MBD3L1 contains less preformed helical content than MBD2 (7% *versus* 25%, respectively). Hence MBD3L1 binding to p66α involves a coil to helix transition that reduces internal degrees of freedom yielding a large unfavorable change in entropy while at the same time forming backbone hydrogen bonds of an α- helix providing a favorable enthalpy change. Furthermore, mutating a non-contact glutamate of MBD2 to a glycine (E225G) reduces helical content (from 25% to 9%) and greatly reduces the binding affinity, whereas the reverse mutation in MBD3L1 (G159E) increases helical content (from 7% to 16%) and increases binding affinity. Taken together, these experiments show that high affinity binding requires minimal preformed helical content of the individual coiled-coil domains.

This requirement for preformed helical content differs from other coiled-coil complexes that often show evidence of a coil to helix transition upon binding *(40)*. For example, binding and folding of the GCN4 coiled-coil dimerization domain has been well characterized with conflicting results. In general, thermodynamic unfolding of GCN4 reflects a two-state transition such that binding appears coupled to folding *(41)*. However, NMR *(42)* and mutation analyses *(43)* indicate that pre-existing helical content promotes complex formation. In contrast to earlier studies *(44)*, the latter results suggest that the transition state involves interaction between preformed helical segments *(45)*.

In more recent studies of the oligomerization domain (SARAH) from serine/threonine mammalian sterile 20-like kinase (MST1), thermodynamic analyses show that this coiled-coil domain remains unstructured in isolation and folds upon binding *(40)*. The unstructured state of the monomeric SARAH domain allows the protein to adopt different structures and bind different partners. In contrast to these prior studies, the coiled-coil domains of MBD2 and homologues do not form homodimeric complexes. The lack of homodimerization simplified analysis and allowed us to probe the relationship between helical content of the MBD2 monomer with heterodimer formation.

Given the propensity to form helices in isolation and the large hydrophobic surface of the coiled-coil domains, one would anticipate that the individual peptides should homo-oligomerize in isolation, especially at high concentration. However AUC analyses show that the coiled-coil domains of MBD2 and each of the homologues do not homo-oligomerize. In fact, the MBD2 remains entirely monomeric even at 300 μM concentration (Figure 1.3A). Therefore, the coiled-coil domains do not form homo- or heterodimeric complexes between the different homologues

as has been suggested previously. Instead these domains remain isolated monomeric helices until binding p66α as a 1:1 complex.

As we demonstrated previously *(1)*, charged residues contributed to binding specificity such that reversing the charge of three residues in p66α eliminates binding to MBD2. Here we have shown that introducing complementary charge changes in MBD2 restores binding to the charge mutant of p66α but not with the same high affinity as the wild type complex. The inability to bind with high affinity does not reflect a lack of helical content nor does it reflect geometric restraints on the relative positioning of the charged residues. Therefore, we hypothesize that the electrostatic surface potential generated by these charged residues contributes to high affinity association. The wild type electrostatic surface potential alternates between positive and negative regions such that the branched hydrophobic residues (at position **a**) fall at the interface between these oppositely charged regions. In contrast, the surface potentials for the mutant proteins are more uniformly positive (p66α(ERR)) or negative (MBD2(REE)). Even though the charge swap mutations generate complementary electrostatic surface potentials, these surface potentials are qualitatively different from the wild type proteins. Based on these observations we suggest that optimal high affinity binding between MBD2 and p66α depends on the alternating surface potential. One possibility for this requirement is that the position of the hydrophobic residues between alternating surface potentials may stabilize induced dipole moments and increase van der Waals interactions between the two chains.

These studies underscore how small changes in helical content and electrostatic interactions can modulate the binding affinity of the coiled-coil domains. In this case, the changes led to a 10-fold binding affinity preference of p66α for MBD2 and MBD3 over the MBD3L1 and MBD3L2 homologues. MBD3L1 and MBD3L2 homologues are expressed in specific tissue

types *(24)*, whereas MBD2 and MBD3 are more ubiquitous. Each of these proteins recruits the same NuRD chromatin remodeling complex; however, MBD3L1 and MBD3L2 lack a methyl-cytosine binding domain (Figure 1.2*A*) and as such target the complex to distinct regions *(24-26, 46)*.In previous studies we showed that the coiled-coil interaction between p66α and MBD2 was critical for the formation of a functional NuRD complex. Enforced expression of the isolated p66α peptide blocked recruitment of the native p66α protein and the Mi2 chromatin remodeling protein *(1)*. Consequently the p66α peptide blocked DNA methylation dependent gene silencing by the MBD2 protein. The relative binding hierarchy of the MBD3 homologues indicates that the ubiquitously expressed MBD2 and MBD3 should effectively compete with MBD3L1 and MBD3L2 for a functional NuRD in those cell types that co-express the homologous proteins. In this manner, fine-tuning of coiled-coil domain binding affinity can be used to establish hierarchical binding networks for tissue specific gene regulation and chromatin remodeling.

In summary, we have shown that differences in helical content and charge distribution dictate high affinity anti-parallel heterodimeric coiled-coil complex formation between MBD2 homologues and p66α. The MBD2 homologues remain monomeric helices in isolation, even at high concentrations, poised to bind p66α with high affinity and specificity. Although the coiled-coil domain represents a relatively simple binding motif, subtle variations in sequence can modify binding affinity and specificity. Understanding the determinants of high affinity binding will inform the development of inhibitors of coiled-coil complexes for potential therapeutic benefit *(19)*.

1.6 References

1. M. N. Gnanapragasam *et al.*, p66Alpha-MBD2 coiled-coil interaction and recruitment of Mi-2 are critical for globin gene silencing by the MBD2-NuRD complex. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7487-7492 (2011).

2. E. Li, T. H. Bestor, R. Jaenisch, Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell.* **69**, 915-926 (1992).

3. J. K. Kim, M. Samaranayake, S. Pradhan, Epigenetic mechanisms in mammals. *Cell Mol. Life Sci.* **66**, 596-612 (2009).

4. J. Berger, A. Bird, Role of MBD2 in gene regulation and tumorigenesis. *Biochem. Soc. Trans.* **33**, 1537-1540 (2005).

5. B. Hendrich, A. Bird, Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538-6547 (1998).

6. L. Lopez-Serra, M. Esteller, Proteins that bind methylated DNA and human cancer: reading the wrong words. *Br. J. Cancer.* **98**, 1881-1885 (2008).

7. Y. Zhang *et al.*, Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13**, 1924-1935 (1999).

8. M. Fatemi, P. A. Wade, MBD family proteins: reading the epigenetic code. *J. Cell. Sci.* **119**, 3033-3037 (2006).

9. B. Hendrich, J. Guy, B. Ramsahoye, V. A. Wilson, A. Bird, Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev.* **15**, 710-723 (2001).

10. B. Hendrich, S. Tweedie, The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269-277 (2003).

11. N. J. Bowen, N. Fujita, M. Kajita, P. A. Wade, Mi-2/NuRD: multiple complexes for many purposes. *Biochim. Biophys. Acta.* **1677**, 52-57 (2004).

12. J. N. Scarsdale, H. D. Webb, G. D. Ginder, D. C. Williams Jr, Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res.* **39**, 6741-6752 (2011).

13. M. Brackertz, J. Boeke, R. Zhang, R. Renkawitz, Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *J. Biol. Chem.* **277**, 40958-40966 (2002).

14. M. Brackertz, Z. Gong, J. Leers, R. Renkawitz, p66alpha and p66beta of the Mi-2/NuRD complex mediate MBD2 and histone interaction. *Nucleic Acids Res.* **34**, 397-406 (2006).

15. Z. Gong, M. Brackertz, R. Renkawitz, SUMO modification enhances p66-mediated transcriptional repression of the Mi-2/NuRD complex. *Mol. Cell. Biol.* **26**, 4519-4528 (2006).

16. J. W. Rupon, S. Z. Wang, K. Gaensler, J. Lloyd, G. D. Ginder, Methyl binding domain protein 2 mediates gamma-globin gene silencing in adult human betaYAC transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6617-6622 (2006).

17. E. P. Kransdorf *et al.*, MBD2 is a critical component of a methyl cytosine-binding protein complex isolated from primary erythroid cells. *Blood.* **108**, 2836-2845 (2006).

18. G. Grigoryan, A. E. Keating, Structural specificity in coiled-coil interactions. *Curr. Opin. Struct. Biol.* **18**, 477-483 (2008).

19. H. M. Strauss, S. Keller, Pharmacological interference with protein-protein interactions mediated by coiled-coil motifs. *Handb. Exp. Pharmacol.* **(186)**, 461-482 (2008).

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

21. M. O. Steinmetz *et al.*, Molecular basis of coiled-coil formation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7062-7067 (2007).

22. O. D. Testa, E. Moutevelis, D. N. Woolfson, CC+: a relational database of coiled-coil structures. *Nucleic Acids Res.* **37**, D315-22 (2009).

23. M. K. Yadav *et al.*, Coiled coils at the edge of configurational heterogeneity. Structural analyses of parallel and antiparallel homotetrameric coiled coils reveal configurational sensitivity to a single solvent-exposed amino acid substitution. *Biochemistry.* **45**, 4463-4473 (2006).

24. C. L. Jiang *et al.*, MBD3L1 and MBD3L2, two new proteins homologous to the methyl-CpG-binding proteins MBD2 and MBD3: characterization of MBD3L1 as a testis-specific transcriptional repressor. *Genomics.* **80**, 621-629 (2002).

25. S. G. Jin, C. L. Jiang, T. Rauch, H. Li, G. P. Pfeifer, MBD3L2 interacts with MBD3 and components of the NuRD complex and can oppose MBD2-MeCP1-mediated methylation silencing. *J. Biol. Chem.* **280**, 12700-12709 (2005).

26. C. L. Jiang, S. G. Jin, G. P. Pfeifer, MBD3L1 is a transcriptional repressor that interacts with methyl-CpG-binding protein 2 (MBD2) and components of the NuRD complex. *J. Biol. Chem.* **279**, 52456-52464 (2004).

27. M. Cai *et al.*, Solution structure of the phosphoryl transfer complex between the signal-transducing protein IIAGlucose and the cytoplasmic domain of the glucose transporter IICBGlucose of the Escherichia coli glucose phosphotransferase system. *J. Biol. Chem.* **278**, 25191-25206 (2003).

28. J. Lebowitz, M. S. Lewis, P. Schuck, Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.* **11**, 2067-2079 (2002).

29. P. Schuck, Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606-1619 (2000).

30. J. R. Kumita, O. S. Smart, G. A. Woolley, Photo-control of helix content in a short peptide. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3803-3808 (2000).

31. E. K. Koepf, H. M. Petrassi, M. Sudol, J. W. Kelly, WW: An isolated three-stranded antiparallel beta-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **8**, 841-853 (1999).

32. V. Munoz, L. Serrano, Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**, 399-409 (1994).

33. E. Lacroix, A. R. Viguera, L. Serrano, Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **284**, 173-191 (1998).

34. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037-10041 (2001).

35. L. Schrödinger, *The PyMOL Molecular Graphics System, Version 1.3 r1.Schrödinger, LLC, Portland, OR.*(2010).

36. E. B. Hadley, O. D. Testa, D. N. Woolfson, S. H. Gellman, Preferred side-chain constellations at antiparallel coiled-coil interfaces. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 530-535 (2008).

37. J. D. Steinkruger *et al.*, The d'--d--d' vertical triad is less discriminating than the a'--a--a' vertical triad in the antiparallel coiled-coil dimer motif. *J. Am. Chem. Soc.* **134**, 2626-2633 (2012).

38. M. L. Doyle, Titration microcalorimetry. *Current Protocols in Protein Science.* **20.4.1-20.4.24**(2001).

39. C. D. Schwieters, J. J. Kuszewski, N. Tjandra, G. M. Clore, The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65-73 (2003).

40. D. Constantinescu Aruxandei, C. Makbul, A. Koturenkiene, M. B. Ludemann, C. Herrmann, Dimerization-induced folding of MST1 SARAH and the influence of the intrinsically unstructured inhibitory domain: low thermodynamic stability of monomer. *Biochemistry.* **50**, 10990-11000 (2011).

41. A. I. Dragan, P. L. Privalov, Unfolding of a leucine zipper is not a simple two-state transition. *J. Mol. Biol.* **321**, 891-908 (2002).

42. J. K. Myers, T. G. Oas, Reinterpretation of GCN4-p1 folding kinetics: partial helix formation precedes dimerization in coiled coil folding. *J. Mol. Biol.* **289**, 205-209 (1999).

43. J. A. Zitzewitz, B. Ibarra-Molero, D. R. Fishel, K. L. Terry, C. R. Matthews, Preformed secondary structure drives the association reaction of GCN4-p1, a model coiled-coil system. *J. Mol. Biol.* **296**, 1105-1116 (2000).

44. T. R. Sosnick, S. Jackson, R. R. Wilk, S. W. Englander, W. F. DeGrado, The role of helix formation in the folding of a fully alpha-helical coiled coil. *Proteins.* **24**, 427-432 (1996).

45. L. B. Moran, J. P. Schneider, A. Kentsis, G. A. Reddy, T. R. Sosnick, Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10699-10704 (1999).

46. S. G. Jin, W. Tsark, P. E. Szabo, G. P. Pfeifer, Haploid male germ cell- and oocyte-specific Mbd3l1 and Mbd3l2 genes are dispensable for early development, fertility, and zygotic DNA demethylation in the mouse. *Dev. Dyn.* **237**, 3435-3443 (2008).

Chapter 2

Dynamic behavior of MBD4 in methylated DNA recognition

2.1 Abstract

Having a C-terminal glycosylase domain in addition to N-terminal methyl-cytosine binding domain (MBD) makes methyl-cytosine binding domain 4 (MBD4) a unique member of family of MBD proteins which recognize methylated Cytosine-Guanine (mCpG) dinucleotide. MBD4 serves as a potent DNA glycosylase in DNA mismatch repair and is also shown to be involved in transcriptional repression. The MBD domain of MBD4 can bind TpG dinucleotide in addition to mCpG and thus drives specificity for the glycosylase to act on, therefore understanding how MBD binds methylated DNA sequence is crucial. Here, we present a solution structure of the MBD of human MBD4 bound to DNA. Based on chemical shift changes and binding analyses, we show that the MBD of MBD4 can bind methylated as well as unmethylated, hydroxymethylated and mismatched (TpG) DNA with preference for mCpG. Further, with the help of chemical exchange studies we demonstrate the dynamics of methylated DNA recognition by MBD4. MBD4 exchanges slowly on two DNA binding sites on two separate pieces of DNA (inter-molecular) but if linked together on a single DNA (intra-molecular), MBD4 exhibits fast exchange between two sites. This suggests that MBD4 prefers to move along DNA using either sliding or hopping until it finds its target rather than searching through random three-dimensional diffusion. Introducing more bases or a defect between two sites on the same DNA molecule does not affect the fast exchange rate, indicating that MBD4 may use hopping mechanism for moving along the DNA. Furthermore, we demonstrate the effect of NaCl concentrations on inter and intra molecular exchange of MBD4. This is the first time we demonstrate a hopping mechanism of MBD4 in targeting a relatively diverse DNA methylation mark.

2.2 Introduction

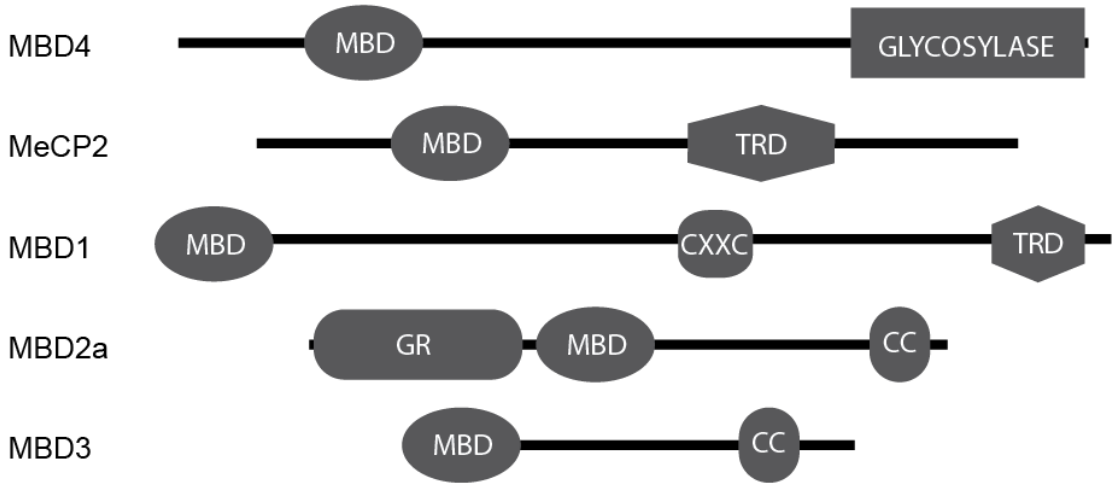
2.2.1 Methylated DNA binding proteins

DNA methylation represents a key epigenetic signal involved in developmental and tissue type specific gene silencing, chromatin modifications, and aberrant silencing of tumor suppressor genes in cancer. Central to these regulatory functions is a family of proteins that selectively bind symmetrically methylated CG dinucleotide sequences (mCpG) through a common methylcytosine binding domain (MBD). The MBD was first described as a ~70 amino acid region in the MeCP2 protein *(1)* and subsequently identified by homology in six additional proteins, MBD1-6 *(2)*. Each MBD protein has a unique domain architecture and amino acid sequences outside of the MBD itself (with the exception of MBD2 and MBD3, Figure 2.1).

MeCP2 (methyl CpG binding protein 2), is a 50 kDa protein encoded by four exons that lead to two different splice variants called MeCP2α and MeCP2β depending on inclusion of exon 2. It has MBD in its N-terminal region and a transcriptional repressor domain (TRD) in its C-terminal region. The TRD domain of MeCP2 causes repression in response to a methylated DNA sequence signal seen by the MBD domain. Sin3A, a complex containing histone deacetylase (HDAC1 and HDAC2), is recruited by the TRD domain. Mutations in the MBD or TRD lead to Rett syndrome, severe autism spectrum disorder *(3)*.

MBD1 is a 55 kDa protein which contains MBD at the N-terminal region, a TRD at the C-terminal region and at least 2 (most likely 3) CXXC domains between MBD and TRD. Repression by MBD1 is mediated by K9 of Histone H3 methylation which recruits complexes like CAF-1, SETDB1. MBD1 knockout mice show affected abnormal differentiation and chromosome instability *(4)*.

A)



B)

```
                                  β1        L1     β2          β3     α1
MBD4   81  ------ECRK SVPCGWERVV KQRLFGKTAG RFDVYFISPQ GLKFRSKSSL  124
MeCP2  95  -----YDD-P TLPEGWTRKL KQRKSGRSAG KYDVYLINPQ GKAFRSKVEL  138
MBD1   1   MAEDWLDC-P ALGPGWKRRE VFRKSGATCG RSDTYYQSPT GDRIRSKVEL   49
MBD2a 151  ------DC-P ALPPGWKKEE VIRKSGLSAG KSDVYYFSPS GKKFRSKPQL  193
MBD3   1   MERKRWEC-P ALPQGWEREE VPRRSGLSAG HRDVFYYSPS GKKFRSKPQL   49

              α1
MBD4   125 ANYLHKNGET SLKPEDFDFT VLSK----  148
MeCP2  138 IAYFEKVGDT SLDPNDFDFT VTGR----  162
MBD1   50  TRYLGPACDL TL----FDFK -QGIL---   69
MBD2a  194 ARYLGNTVDL SS----FDFR -TGKMM--  214
MBD3   50  ARYLGGSMDL ST----FDFR -TGKMLMS   72
```

Figure 2.1: Comparison of MBD proteins: A) Domain organization of MBD proteins is proportionally depicted. Note that all proteins have at least one domain in addition to the MBD. B) CLUSTAL format alignment of MBD domains is created using MAFFT (v7.029b). Key contact residues are shown in bold and predicted secondary structure motifs are shown above.

MBD2 and MBD3 are two MBD proteins which share homology outside their MBD domains. A shorter variant of MBD2, MBD2b (262 amino acids) and MBD3 share 71% sequence identity throughout *(2)*. Both MBD2 and MBD3 have a MBD domain at its N-terminus and a coiled-coil domain at the C-terminus (Figure 2.1A). Both recruit a large nucleosome remodeling and deacetylation complex (NuRD) that contains a histone deacetylase (HDAC1 or HDAC2) and a nucleosome remodeling protein Mi2 *(5)*. MBD2 selectively associates with mCpG while MBD3 does not, reflecting amino acid differences within DNA contacting regions of the MBD *(6, 7)*. A single tyrosine (MBD2) to phenylalanine (MBD3) substitution alters interaction with the methyl group of methylcytosine and abolishes selectivity for methylated DNA. Importantly, based on mass spectrometry analyses of purified complexes, MBD2 and MBD3 are mutually exclusive components of the NuRD complex *(8)*. Experimental knockout mouse of MBD3 is shown to embryonically lethal *(9)*, but MBD2 knockout mice are viable and only show mild changes in behavior and reduced susceptibility to tumorigenesis *(10)*.

MBD4 (also known as methyl CpG binding endonuclease 1, MED1) is the most unique of the lot, as it is a part of DNA repair machinery and has a glycosylase domain at its C-terminus in addition to its N-terminus MBD domain. The glycosylase domain of MBD4 has DNA repair function and removes the thymine in a mCpG·TpG double-stranded mismatch *(2, 11, 12)*. The mCpG·TpG mismatch arises from hydrolytic deamination of a methylcytosine to thymine, which is one of the more common sources of germline and somatic DNA point mutations. The MBD from MBD4 preferentially binds mCpG·TpG, but it also recognizes mCpG sequences *(11)*. MBD4 knock-out mice show increased tumor susceptibility *(13)*.

A few proteins that lack a MBD domain are also shown to recognize methylated DNA. Kaiso is a member of the BTB/POZ zinc finger domain family of proteins that bind methylated

CpG dinucleotide. Repressive activity of Kaiso involves N-COR corepressor complex which includes HDAC3 *(4)*. MBD proteins constitute a major group of proteins that recognize DNA methylation mark. Genome wide promoter occupancy studies show that MBD proteins can be found at distinct but overlapping subsets of genes silenced by DNA methylation in cancer *(14)*. Furthermore, silencing of specific methylated genes has been attributed to individual MBD proteins (i.e. MBD2 silences BRCA1 and GSTP1) *(15-22)*. In this study, the focus is on MBD4 and investigating its methylated DNA recognition abilities.

2.2.2 Comparison of MBD domains from MBD proteins

Figure 2.1B shows a sequence alignment for MBD domains from human MeCP2, MBD1, MBD2a, MBD3 and MBD4. Note that all MBD domains are homologous and are about 70 amino acids long. Due to high degree of homology and similar target recognition (mCpG dinucleotide), MBD domains are thought to have similar structure and mechanism of action. The crystal structure of the MBD domain from MeCP2 *(23)*, the NMR structure of MBDs from MBD1 *(24)* and MBD2 *(25)* proteins bound to a methylated DNA are known. All three structures show similar folds and also demonstrate that the key DNA contacting residues are conserved and form a similar DNA recognition interface. Secondary structure elements are depicted at the top of the sequence alignment and all MBD domains may exhibit these folds when bound to the DNA. The MBD domain when bound to a DNA has two prominent β strands (β1 and β2) connected by a loop (L1), a small β strand (β3) and followed by a helix (α1). Two key interacting arginine residues and one tyrosine residue are conserved among all MBDs and are shown in bold (Figure 2.1B).

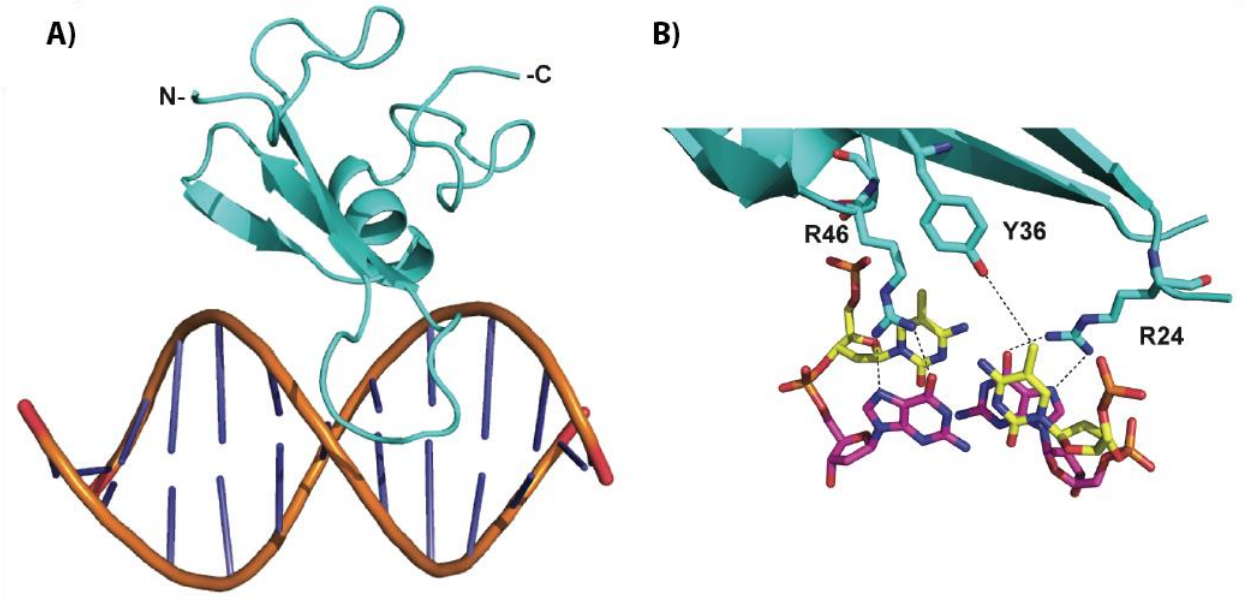


Figure 2.2: Solution structure of the methyl binding domain of chicken MBD2 bound to a methylated promoter target sequence. A mixed rendering diagram (A) of the solution structure is presented. The methyl binding domain is shown as a ribbon diagram in cyan, DNA is shown as a tube model. A more detailed view of three critical residues involved in the protein:DNA interface (B) is shown as stick diagram with protein residues colored cyan and DNA bases colored yellow (methylcytosine) and magenta (guanosine) with select residues labeled. Adapted from Scarsdale *et al.* (2011) *(25)*.

Figure 2.2, adapted from Scarsdale *et al. (25)*, demonstrates the structure of the MBD from chicken MBD2 bound to a ten base pair methylated sequence from the ρ-globin promoter. Figure 2.2A demonstrates that MBD2 is a major groove binder whereas Figure 2.2B illustrates importance of key contacting arginine and tyrosine residues. Binding of MBD to a specific DNA involves complex network of electrostatic and hydrophobic interactions. Arginine residues form a hydrogen bond with guanine and the aliphatic portion of the side-chains pack against the adjacent methyl group of methyl cytosine. A cation-π interaction between the guanidinium group of Arginine and pyrimidine ring is also shown to stabilize the interaction *(26, 27)*. A key tyrosine residue specifically interacts with the methyl groups of one methylcytosine. Despite these similarities, the MBD proteins target different non-overlapping subsets of genes. Here, we study how MBD4 differs in its mode of recognition of methylated DNA variants.

2.2.3 MBD4: multifaceted MBD protein with unique glycosylase activity

MBD4 was first identified in a bioinformatics study by its N-terminal MBD, which shares sequence homology with other MBD proteins. MBD4 also has another highly conserved C-terminal DNA binding glycosylase domain and a long spacer domain linking it with the N-terminal MBD domain *(2)*. The glycosylase domain of MBD4 removes thymine and uracil paired with guanine base by base excision repair (BER) pathway. The hydrolytic deamination of 5-methylcytosine and cytosine to thymine and uracil respectively (Figure 2.3), is a spontaneous event that causes 2-300 lesions in a cell per day; If not corrected, it leads to C:G to T:A transitions in the following round of DNA replication *(28)*. Glycosylase of MBD4 is also shown to remove halogenated pyrimidines like 5-chlorouracil and 5-bromouracil paired with G that result from peroxidase-mediated inflammatory processes *(29)*. Rai *et al* have also shown that the

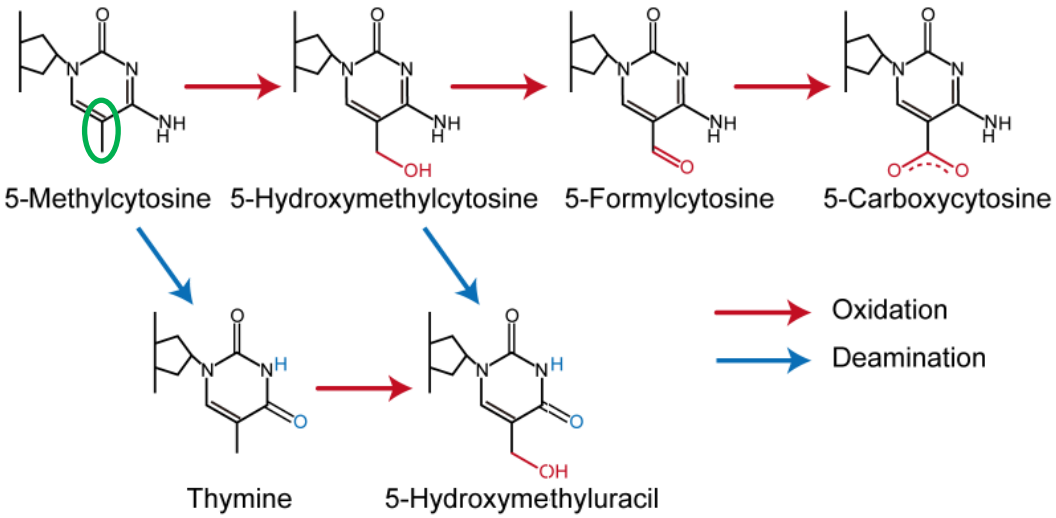


Figure 2.3: A schematic representation of deamination and oxidation products of 5-methylcytosine. The methyl group attached to the 5' position of cytosine is circled in green. Oxidation and deamination are shown in red and blue, respectively. TET mediated oxidation products, 5-Formylcytosine and 5-Carboxycytosine are not excised by MBD4 but Thymine DNA glycosylase (TDG) has shown to have activity towards these substrates *(31)*. Figure adapted from Otani *et al* (2013) *(32)*.

MBD4 can cause active DNA demethylation and thereby compliment DNMTs in carefully regulating DNA methylation mark *(30)*

Unlike other members of the family of MBD proteins, MBD4 is the only one that contains a DNA glycosylase catalytic domain and therefore was classically thought to be involved in DNA repair rather than transcriptional suppression, but in 2005, Kondo and colleagues demonstrated that MBD4 represses transcription of certain genes in human cells and this involved Sin3A and HDAC activity. Though MBD4 has been demonstrated to be localized to hypermethylated promoters of affected genes, knockdown of MBD4 did not change the repression status of either of these promoters. This suggests that other MBD proteins may substitute for repression activity caused by MBD4 *(33)*. It is also noteworthy that the MBD domain is a major grove binder whereas the glycosylase domain is a minor groove binder, and they both are separated by long spacer region (~280 amino acids) allowing them to work independently or complement each other.

In the studies presented here, the focus is on the MBD domain of MBD4 and how it compares to other MBD domains studied to date in terms of structure and function. We have solved a solution structure of MBD domain of MDB4 and shown that it has similar structural folds (two β strands connected by a loop, a small β strand and a helix) exhibited by other MBD domains. Though similar, it exhibits certain features (a longer and more stable alpha helix, for example) which make it more closely related to MeCP2 than MBD1 and MBD2. Being considered as a mismatch repair enzyme, MBD4 is expected to bind not only methylated CpG but also TpG mismatched sites. Based on chemical shift changes and binding analyses, we show that MBD of MBD4 can bind methylated as well as unmethylated, hydroxymethylated and

mismatched (TpG) DNA with little preference for mCpG, thereby defining target specificity for MBD4.

It has always been intriguing how DNA binding proteins accomplish the remarkable feat of finding their correct target sequences within huge genomic DNA. It has been proposed that, in order to reach their target site, proteins first translocate along nonspecific DNA, i.e. move along DNA that does not contain a cognate binding site *(34)*. The mechanism underlying this so-called 'facilitated diffusion', however, is still under debate. Here, using 'MBD4 recognizing methylated DNA mark' as a model we provide new insights on DNA-protein recognition. With the help of chemical exchange studies we demonstrate the dynamics of methylated DNA recognition by MBD4. MBD4 exchanges slowly on two DNA binding sites on two separate molecules of DNA (inter-molecular) but if linked together on a single DNA (intra-molecular), MBD4 exhibits fast exchange between the two sites. This suggests that MBD4 prefers to slide or hop across DNA until it finds its target rather than search through three-dimensional diffusion. Introducing more bases or a defect between the two sites on the same DNA does not affect the fast exchange rate indicating that MBD4 may prefer hopping over sliding as a mode of facilitated diffusion. Furthermore, we demonstrate that NaCl qualitatively affects inter and intra molecular exchange of MBD4. This is the first time we demonstrate the hopping mechanism of MBD4 in targeting a relatively diverse DNA methylation mark.

2.3 Materials and methods

2.3.1 Protein expression and purification

The methyl DNA binding domain of human MBD4 (amino acids 80-148) was cloned and expressed with a hexahistidine tag and as a thioredoxin fusion proteins in a modified pET32a vector *(35)*. The expression vector was transformed into the BL21(DE3) *E. coli* strain, grown at 37 ^{0}C and induced with 1 mM isopropyl-β-d-thiogalactopyranoside at an A_{600} ~ 0.8. Induced bacteria were harvested and lysed with the B-PER reagent (Thermo Scientific). The soluble fraction was passed over a nickel-sepharose column and protein was eluted with a step gradient of imidazole. After removing the thioredoxin and hexahistidine using thrombin cleavage, the protein was further purified by gel filtration over a Superdex-75 column (GE Healthcare). For additional purity, protein was cleaned using reverse phase column SOURCE-15RPC (GE Healthcare) and eluent was dialyzed against physiological buffer. The thioredoxin fusion protein was used for surface plasmon resonance (SPR) studies. The fusion protein was expressed in a similar manner and after purification over a nickel-sepharose column, it was passed over MonoS 10/100 GL (GE Healthcare). Size exclusion chromatography was used for additional purification. Resulting thioredoxin fused MBD was > 95% pure as estimated by SDS-PAGE analysis. Uniform double (^{13}C, ^{15}N) and triple (^{13}C, ^{15}N, ^{2}H) labeled protein samples were generated by standard techniques. The final concentrations of all protein samples were determined by monitoring UV absorbance at 280 nm.

2.3.2 DNA purification

Complimentary DNA oligonucleotides were purchased from Integrated DNA technologies. Forward and reverse oligonucleotides were dissolved in standard buffer (20 mM Tris pH 8.0)

and mixed in equimolar concentrations. After heating at > 363K for 10 minutes, mixture was cooled slowly to room temperature (annealed) allowing formation of double stranded (DS) DNA. Subsequently, DS DNA was purified by ion exchange chromatography on MonoQ 10/100 column (GE Healthcare). 3' biotinylated forward oligonucleotides (purchased from Integrated DNA technologies) were mixed with regular unlabeled complimentary reverse oligonucleotides, annealed and further purified using ion exchange for binding studies using surface plasmon resonance. The final concentration of DS DNA was determined by monitoring UV absorbance at 260 mm.

2.3.3 Surface plasmon resonance

Protein and DNA samples were prepared in standard buffer (10 mM HEPES pH 6.5, 50 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT). Binding affinities of MBD domain with 3' biotinylated DNA variants were studied using a Sensor SA chip on Biacore T200 (GE Healthcare). Biotinylated DS DNAs were immobilized to the experimental channels of SA chip using biotin-streptavidin chemistry until the final response units were in the range of ~ 50-100, control channels were blocked without linking DNA. Various concentrations of thioredoxin fused MBD4 was passed over control and experimental channels (at a flow rate of 30 µl/min) in running buffer (10 mM HEPES pH 6.5, 50 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, 0.05% polysorbate 20). At least one concentration of MBD4 was repeated in triplicate to estimate error and thereby determine quality of data. The sensogram which resulted from subtracting the control channel response from the experimental channel response was fit by a steady state analysis using manufacturer's software.

2.3.4 NMR experiments

Purified ^{15}N-MBD4 protein was mixed with purified DS DNA oligonucleotides, buffer-exchanged into 10 mM $NaPO_4$, pH 6.5, 1 mM dithiothreitol, 10% D_2O and 0.02% sodium azide. Standard NMR experiments for resonance assignments, distance and torsional angle restraints were collected on a Bruker Avance III^{TM} 700MHz NMR spectrometers at 25^0C. Analysis and structure calculation (described in Chapter 3) resulted in a three dimensional solution structure of MBD4 on a methylated DNA.

To compare binding preference with HSQC experiments, ^{15}N-MBD4 and respective DNA were mixed at 1:1.1 ratio (10% excess DNA), buffer exchanged into standard buffer (mentioned above) and concentrated to 200 µM. ^{15}N-HSQC spectra (amide and Arg Nε region) for MBD4 when bound to 17 bp methylated (1xmCpG) DNA, unmethylated, hydroxymethylated and mismatch DNA (Table 2.1) were collected. For chemical exchange studies MBD4 was mixed with excess DNA in a ratio corresponding to four methylated DNA binding sites for each MBD4. HSQC and TROSY HSQC experiments were collected for studying dynamics. The effect of increasing NaCl were measured by titrating a 300 mM stock of NaCl in standard buffer into previously used sample to give final salt concentration ranging from 20-100 mM.

2.4 Results

2.4.1 Solution structure of methyl binding domain of MBD4 on methylated DNA

We present a solution structure of the methyl binding domain from human MBD4 (amino acids 80–148) bound to a 10bp DNA fragment containing a methylated CpG using multidimensional NMR spectroscopy. The 10 bp DNA used for this studies is the same sequence used by Scarsdale *et al (25)* and is a known target for MBD proteins and has one centrally located mCpG dinucleotide (Table 2.1). The reported solution structure is in the final stages of refinement based on ~800 NOE restraints and were calculated by simulated annealing using the Xplor-NIH software package *(36)*. In Figure 2.4 a best-fit superimposition cartoon diagram of MBD4 methyl binding domain when bound to a methylated DNA is shown for an ensemble of 10 calculated lowest energy structures. Chapter three describes the process of solving the three dimensional protein structure using NMR spectroscopic techniques.

The crystal structure of MBD domain from MeCP2 *(23)*, the NMR structure of MBDs from MBD1 *(24)* and MBD2 *(25)* proteins bound to a methylated DNA are known. All three structures show similar folds and also demonstrate that the key DNA contacting residues are conserved and form a similar DNA recognition interface (Figure 2.1). Like these domains, MBD4 has two prominent β strands (β1 {Residue E91-Q96} and β2 {R105-I111}) connected by a loop (L1 {R97-G104}), a small β strand (β3 {L116-K117}) followed by an α helix (α1 {K121-L128}). The MBD domain of MBD4 is most similar to that of MeCP2 and proposed to be evolved closely with MeCP2 *(2, 37)*. Similar to MeCP2, the helix α1 is longer and well defined in MBD4 than in MBD2 and MBD1. Recently, Otani *et al*, have released a crystal structure of mouse MBD4 methyl binding domain bound to methylated DNA and our findings are congruent with

their structure *(32)*. The MBD domain of MBD4 is also relatively well folded in isolation (not bound to DNA). It undergoes further structural changes upon binding to DNA but these changes are not as dramatic as seen in case of MBD2. MBD2 is proposed to fold upon binding to DNA whereas MBD4 is inherently structured even in the absence of DNA.

Table 2.1: Table of all DNA constructs used. Complimentary strands were annealed as described in 2.3.2. Central CpG dinucleotide and its modifications are highlighted in bold (m = methylated, hm = hydroxymethylated). 5'- 3' (forward) oligonucleotides were biotinylated at 3' end for surface plasmon resonance experiment (not shown in the table).

Oligonucleotide name	**Sequence**
17bp 1xmCpG	5'-GAGGCGC TmCGG CGGCAG-3' 3'-CTCCGCG AGmCC GCCGTC-5'
17bp unmethylated	5'-GAGGCGC TCGG CGGCAG-3' 3'-CTCCGCG AGCC GCCGTC-5'
17bp hydroxymethylated	5'-GAGGCGC ThmCGG CGGCAG-3' 3'-CTCCGCG AGhmCC GCCGTC-5'
17bp mismatch	5'-GAGGCGC TmCGG CGGCAG-3' 3'-CTCCGCG AGTC GCCGTC-5'
10bp 1xmCpG	5'-GGA TmCGG CTC-3' 3'-CCT AGmCC GAG-5'
10bp Inverted	5'-GGA CmCGA CTC-3' 3'-CCT GGmCT GAG-5'
Tandem short (20bp)	5'-GGA TmCGG CTC GGA CmCGA CTC-3' 3'-CCT AGmCC GAG CCT GGmCT GAG-5'
Tandem long (30bp)	5'-CACGGA TmCGG CT CCCC CGAG TmCGG TCCCGC-3' 3'-GTGCCT AGmCC GA GGGG GCTC AGmCC AGGGCG-5'
Tandem long nicked (30bp)	5'-CACGGA TmCGG CT CCCC CGAG TmCGG TCCCGC-3' 3'-GTGCCT AGmCC GA GG - - GCTC AGmCC AGGGCG-5'

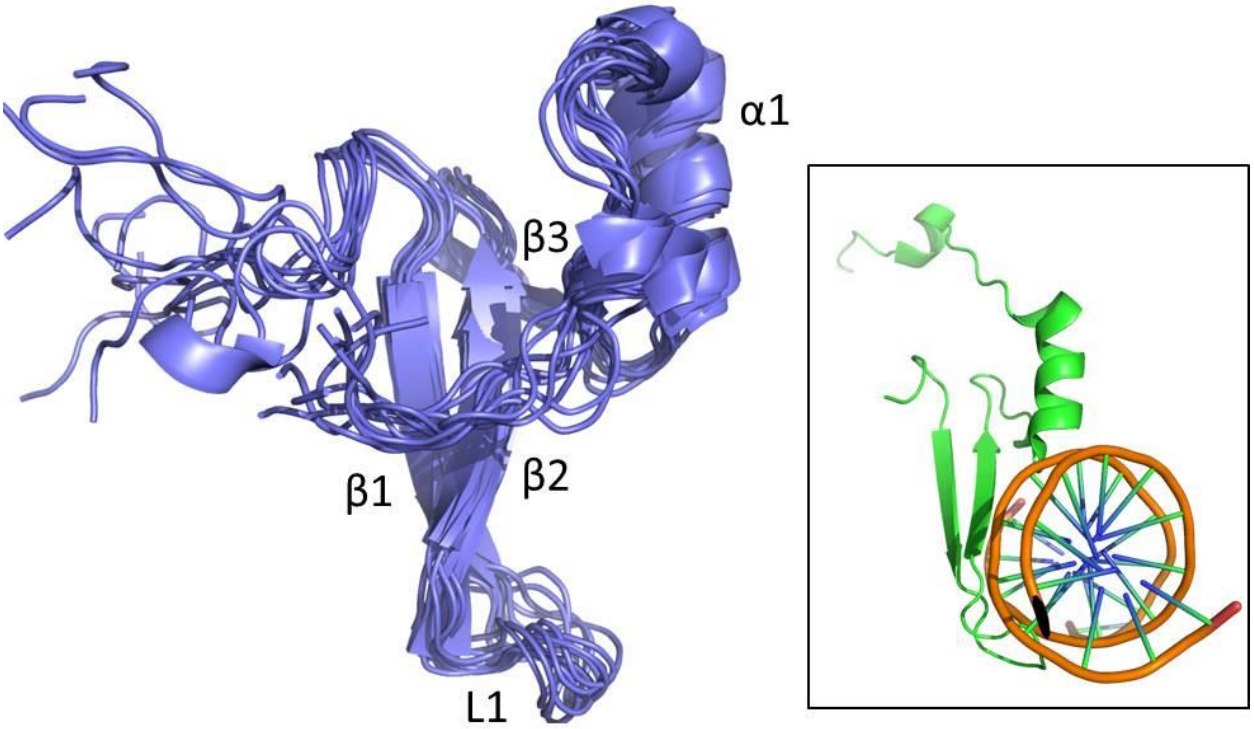


Figure 2.4: Solution structure of MBD4 methyl binding domain bound to a methylated DNA. A best-fit superimposition cartoon diagram of MBD4 methyl binding domain (blue) when bound to a methylated DNA (not shown) is shown for the ensemble of 10 calculated lowest energy structures. Like other MBD domains, it has two prominent β strands (β1 and β2) connected by a loop (L1), a small β strand (β3) and an α helix (α1) that follows. Figure was generated using PyMOL *(38)* In the inset, a recent crystal structure of mouse MBD4 methyl binding domain with a methylated DNA *(32)* is depicted which shares similar binding domain as the human variant.

2.4.2 MBD4 prefers mismatch and methylated DNA over unmethylated and hydroxymethylated DNA

Proposed to being involved in rather diverse biological roles, including repression, base excision repair *et cetera*, the most preferred target sequence for the MBD domain of MBD4 was not clear. In this study, we analyzed MBD4 preference for methylated, unmethylated, mismatch and hydroxymethylated double stranded DNA oligonucleotides. A 17 base pair oligonucleotide with centrally methylated CpG was used for methylated DNA and the central bases were changed for other DNA variants (Table 2.1). Binding constant analysis was carried out with surface plasmon resonance and results are summarized in Table 2.2. MBD4 binds methylated DNA sequence with a fast on-rate and off- rate with a K_D ~725 nM (Figure 2.5). Affinity for mismatched (TpG) DNA was ~15 fold less than methylated DNA but it bound ~3 fold and ~7 fold tighter than unmethylated and hydroxymethylated DNA. This shows that MBD4 has relatively higher preference for methylated and mismatch DNA over unmethylated and hydroxymethylated DNA. MBD4 showed a similar binding affinity towards tandem DNA (20 bp long oligonucleotide with two methylated CpG sites) to that of methylated DNA with one binding site. Stoichiometry for this interaction was calculated to be 1.0 indicating that MBD4 occupies just one of the two binding sites. This observation indicates that having two closely spaced binding sites does not work synergistically to improve MBD4 binding. Although, MBD4 has higher affinity for methylated DNA, binding is weaker compared to other MBD domains, e.g., MBD2 exhibits stronger binding (K_D ~ 50 nM) towards the same 17 bp DNA with one methylated CpG site *(25)*. Therefore, although MBD4 may not be the best MBD protein for targeting CpG it can easily target mismatch DNA, towards which other MBD proteins have weaker affinity.

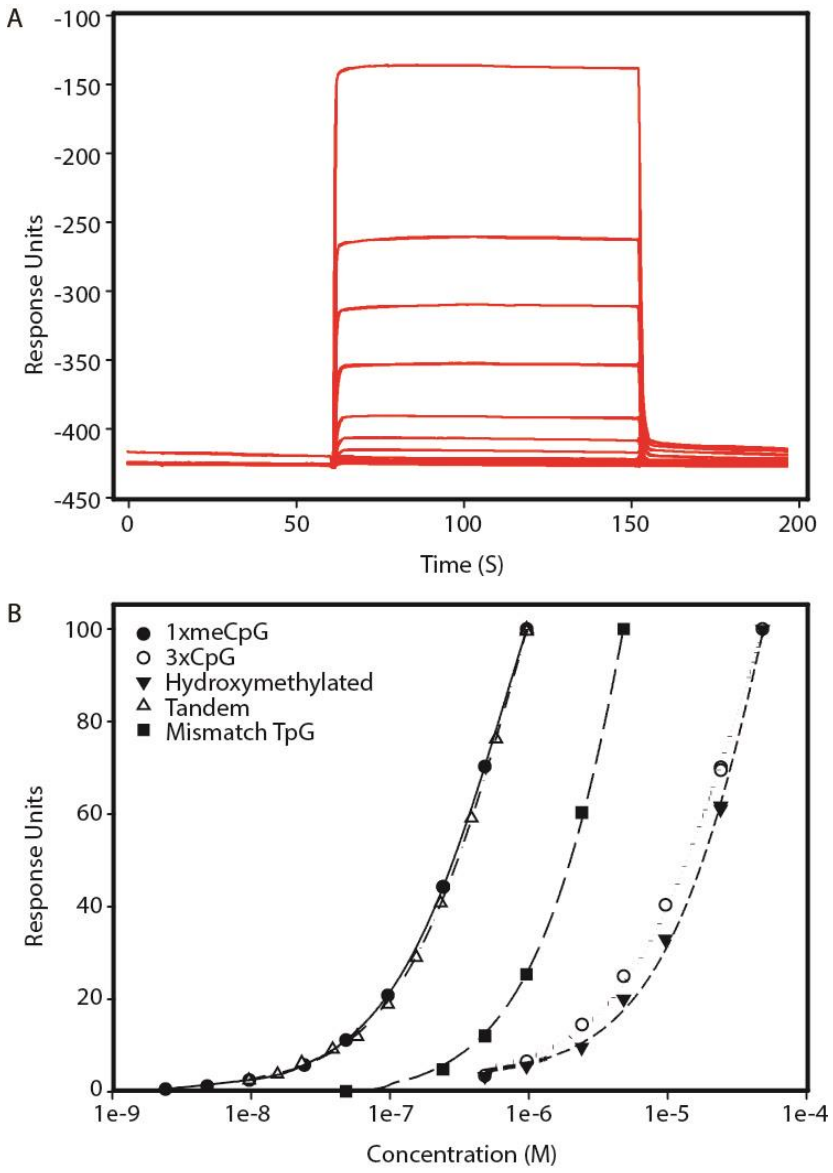


Figure 2.5: Binding affinity of MBD4 to methylated and other DNAs. (A) Sensogram of surface plasmon resonance analysis for varying concentrations of MBD4 binding to a 3'-biotinylated DS DNA coupled to a Sensor Chip SA measured on a Biacore T200 (GE Healthcare). (B) Steady state binding response for MBD4 binding to methylated, unmethylated, hydroxymethylated, tandem, and TpG DNA. The data were fit using the Biacore T200 evaluation software. For comparison, the response units for each were normalized to an $R_{max} = 100$ (Table 2.2).

Table 2.2: Binding analyses of MBD4 to methylated DNA. Surface plasmon resonance studies were performed and the dissociation constants (K_D) between MBD4 methylated DNAs were calculated.

Methylated DNA in complex with MBD4	**K_D**	**Rmax**	**Chi square**
17bp 1xmCpG	725±3.4 nM	278.8	0.0189
17bp mismatch	11.2±1.3 µM	764.5	5.75
17bp unmethylated	34.47±2.6 µM	941.7	72.4
17bp hydroxymethylated	80.6±14 µM	856	31.6
Tandem short (20bp)	881±23 nM	85.54	0.0396

Further, with the help of 2D HSQC analysis of MBD4 bound to various DNA oligonucleotides, we demonstrated that MBD4 prefered methylated and mismatch DNA over unmethylated and hydroxymethylated DNA. 2D HSQC spectra of ^{15}N-MBD4 showed characteristic chemical shifts for certain reporter residues when bound to methylated DNA. Figure 2.6A shows an HSQC for MBD4 when bound to methylated DNA and reporter residues (R97Hε and G100HN) are circled. Subsequent HSQC spectra of MBD4 with unmethylated, hydroxymethylated and mismatch DNA were also collected and changes in chemical shifts of R97Hε and G100HN were observed. Hε of R97 shifts downfield upon binding to methylated DNA (shown in red Figure 2.6B), but its chemical shifts are different when bound to hydroxymethylated and unmethylated DNA. Also, peak for R97Hε is not seen when on mismatch DNA indicating that it is binding to mismatch DNA in relatively different manner and thereby giving a broad peak which is not visible. G100HN, on the other hand, clearly shows that chemical shifts are similar when bound to methylated and mismatch DNA. Unmethylated and hydroxymethylated DNA do not change chemical shifts for G100HN (Figure 2.6C). This clearly demonstrates that MBD4 has higher binding preference for methylated and mismatch (TpG) DNA. Considering the role of glycosylase domain of MBD4 in excising TpG mismatch bases, affinity towards TpG dinucleotide may help with targeting the glycosylase domain to mismatch sites.

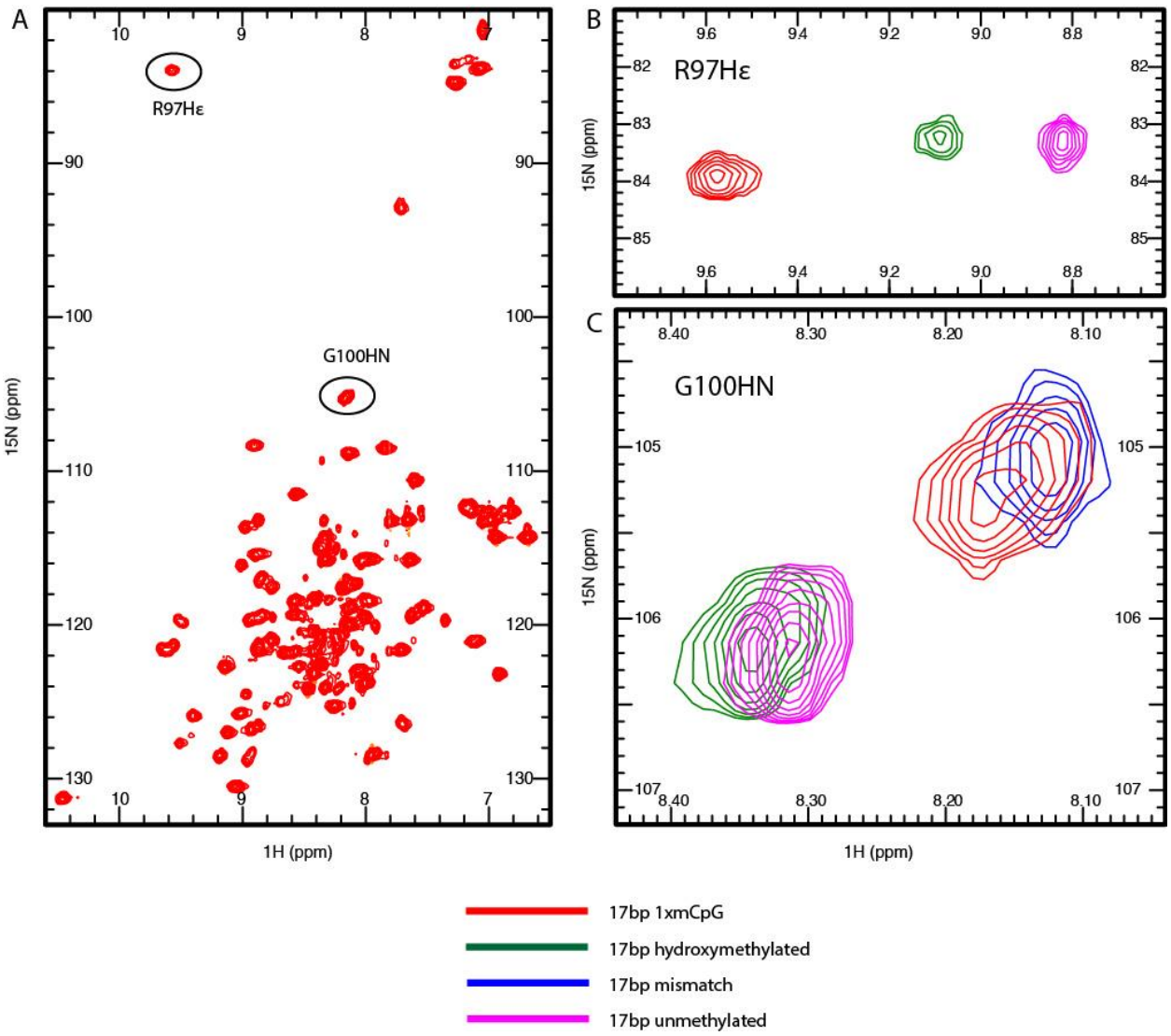


Figure 2.6: NMR spectra of ^{15}N-MBD4. 2D ^{15}N-HSQC spectra (amide and Arg Nε region) for MBD4 when bound to methylated (1xmCpG) DNA is shown in panel A. Reporter residues that have characteristic chemical shift changes when bound to various DNAs are marked with a circle. R97Hε and G100HN are two reporter residues whose distinct chemical shifts are shown in panel B and C.

2.4.3 MBD4 exchanges slowly between two separate DNA binding (inter-molecular) sites on two DNA molecules

Methylated CpG dinucleotide is symmetrical and protein recognizing it can bind in two orientations. However, it was shown in the past that MBD2 predominantly prefers one orientation over the other depending on residues just outside the mCpG pair and inverting those residues can change the orientation preference *(25)*. This subtle change in binding is reflected in a characteristic change in chemical shifts of certain reporter residues and thus serves as a readout for the orientation in which the MBD domain is bound, which in turn indicates the specific sequence to which MBD4 is bound. In a mixed population of two such DNA molecules, the final chemical shift observed for the reporter residue depends on the relative distribution of the population of MBD protein, and how fast it is exchanging between the two states. Here, we use this phenomenon, to explore how the MBD of MBD4 exchanges between two mCpG on different DNA oligonucleotides.

10 base pair double stranded DNA oligonucleotide with central four bases in regular and inverted position (Table 2.1) were used for the study. MBD4 shows a single orientation preference on a given DNA and thus have specific chemical shifts for reporter residues when bound to a regular versus inverted DNA sequence. When we bound ^{15}N-MBD4 methyl binding domain to a mixture of the regular and inverted sequences (1:2:2 molar ratio of MBD4, regular and inverted DNA), two separate peaks were observed for reporter residues (Figure 2.8). These two peaks corresponding to two binding states, indicated that MBD4 exchanges slowly (millisecond or longer time regime) on intermolecular DNA binding sites. Further with the help of the HSQC based zz-exchange spectroscopy, we determined the exchange rate for this slow

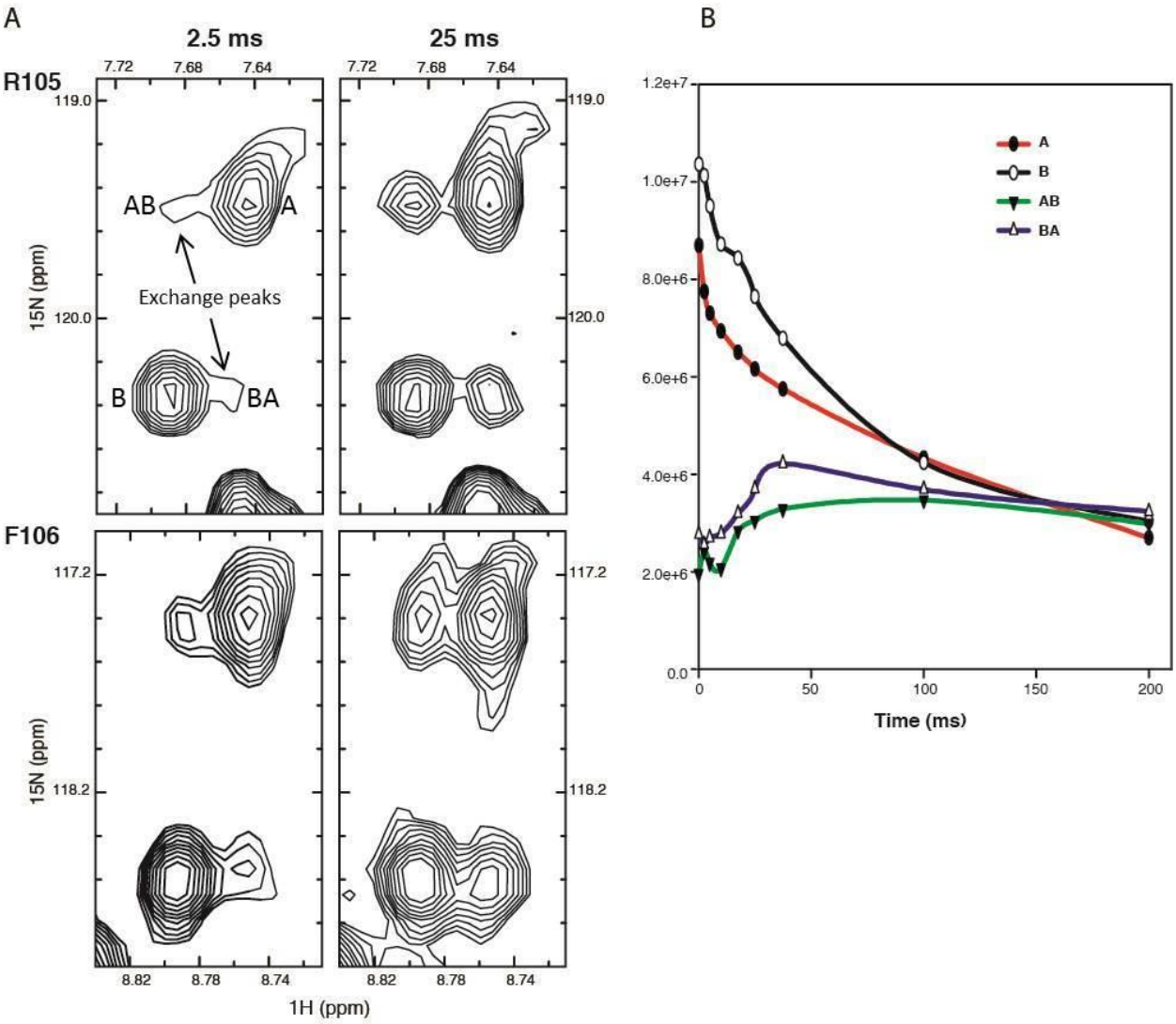


Figure 2.7: The zz-exchange experiment for the inter-molecular exchange of MBD4 between mCpG sites. (A) ^{15}N-MBD4 was bound to a 1:1 mixture of methylated 10bp 1xmCpG and 10bp inverted DS DNA, which show distinct chemical shifts for select amide resonances (R105, F106) are shown for varying z-exchange delays (2.5ms and 25ms). Autopeaks A (regular) and B (inverted) and crosspeaks AB and BA are identified and represent slow chemical exchange between the two binding sites during the z-exchange delay. (B) The intensities for auto and exchange peaks of F106 as a function of mixing time are shown.

inter-molecular exchange. 2D ^{15}N-HSQC z-exchange spectra with varying time delays reveal a buildup of exchange peaks consistent with the exchange rates in the ~20 ms time regime (Figure 2.7).

In Figure 2.7A chemical shifts for two reporter residues R106HN and F106HN are shown. In z-exchange experiments *(39)*, delays are added to allow transfer of magnetization between two states, A (when bound to regular) and B (when bound to inverted DNA). In this mixing period 'T', due to dynamic exchange between two states, nuclei experience different magnetic environment and give rise to exchange cross-peaks. The magnetization is transferred back to the proton for detection. As a result of this experiment, four peaks (two autopeaks A and B corresponding to regular and inverted binding mode and two crosspeaks AB and BA) emerge and are shown in Figure 2.7A. Delays of 0.1, 2.5, 5, 10, 17.5, 25, 37.5, 50, 100, 200 ms were used and buildup of crosspeaks was observed. Intensities of the four peaks after each mixing time were recorded and plotted as a function of time (Figure 2.7B) and used for calculating exchange rate which was found to be ~20 ms. We measured the exchange rate at 200 µM and 333 µM concentrations of MBD4 and found no difference in exchange rates.

2.4.4 MBD4 exhibits fast exchange between two binding sites (intra-molecular) linked on the same DNA oligonucleotide

To test whether MBD4 exchanges more rapidly between two binding sites in the same DNA molecule, we bound MBD4 to a 20 base pair oligonucleotide resulted from tandem covalent linking of 10 bp regular and 10 bp inverted DNA used in the earlier experiment (Table2.1). 2D ^{15}N-HSQC no longer showed distinct peaks for two bound states, instead a single peak was observed for each amide backbone consistent with fast exchange between sites (~µs

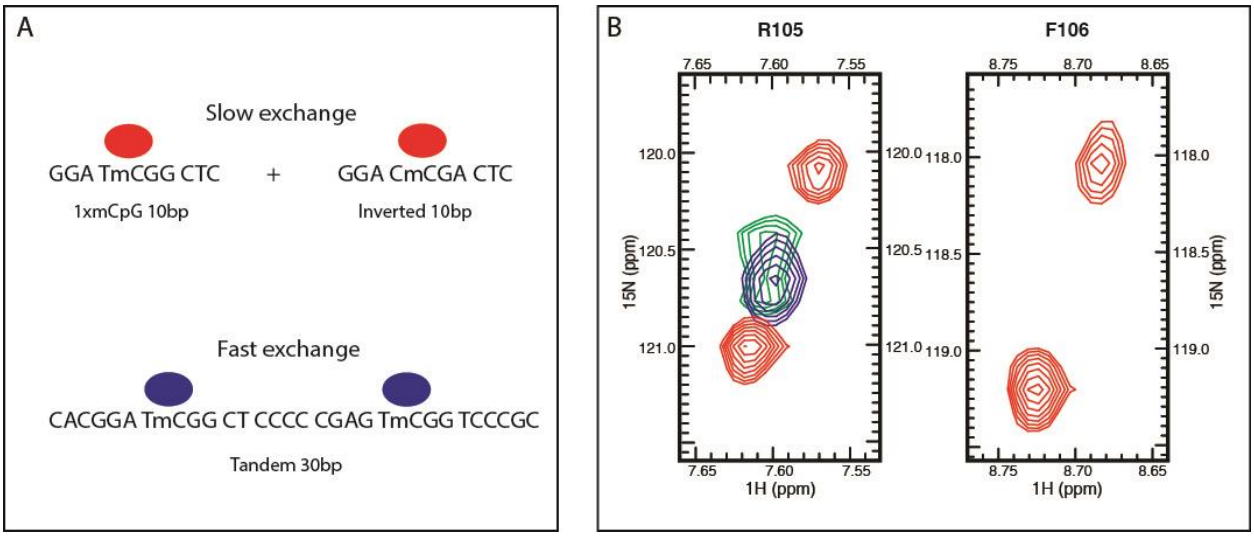


Figure 2.8: Inter-molecular and intra-molecular exchange of MBD4 between mCpG sites. (A) Schematic representation demonstrating that the MBD4 when mixed with a 1:1 mixture of methylated 10bp 1xmCpG and 10bp inverted DS DNA showed slow exchange (shown in red), whereas exhibit fast exchange when bound to a tandem 30bp DNA (shown in blue). (B) TROSY HSQC spectra showing distinct chemical shifts for select amide resonances (R105, F106) when bound to 1:1 mixture of methylated 10bp 1xmCpG and 10bp inverted DS DNA (red), tandem 30bp (blue) and tandem nicked 30bp (green).

time regime). This data suggested that MBD4 showed facilitated diffusion and changed orientation after finding inverted methylated CpG. In this construct MBD4 was changing orientation with respect to the 10bp DNA and still showing a fast exchange rate. Another 30 base pair tandem DNA was designed where in entire 10bp inverted DNA was flipped so that the direction of MBD4 need not change while moving along the DNA, even though orientation with respect to bases outside mCpG can change. A few bases were added to avoid self-annealing; this construct (Tandem long 30bp) is listed in Table 2.1. As the longer DNA was mixed with MBD4, the overall complex was getting larger and therefore, modified TROSY based HSQC experiments were used for these studies. Peaks observed in TROSY experiments are shifted diagonally compared to corresponding peaks from HSQC.

We observed two reporter residues, R105 and F106, after mixing ^{15}N-MBD4 (200 μM) with 30 bp Tandem DNA (400 μM) and resultant peaks in blue can be seen in Figure 2.8B. For both residues, two peaks corresponding to the slow exchange are not observed, indicating that it does not exhibit slow exchange but instead only a single peak for R105 is observed intermediate between the two indicating that MBD4 exhibits fast exchange between the linked sites. Due to lack of an exchange peak for F106, it can be deduced that MBD4 exchanges in an intermediate regime. In conclusion, MBD4 exchanges faster between two DNA binding sites if sites are covalently linked. MBD4 can easily find its next target if it is intra-molecular versus inter-molecular. This data suggest that MBD4 uses facilitated diffusion mechanism along the DNA to help find its binding site.

2.4.5 Addition of small defects in double stranded DNA does not affect facilitated diffusion exhibited by MBD4

Based on earlier experiment, MBD4 can find the next binding site quicker if it is linked covalently, indicating that MBD4 either slides or hops along the DNA sequence rather than using random 3D-diffusion. Once in proximity of DS DNA, MBD4 does not come off and keeps moving along the stretch of non-specific DNA binding sites until it finds a high affinity binding site (mCpG in this case). To test whether a defect in the DNA can impact MBD4 movement along the DNA sequence, we introduced a small defect in a double stranded DNA and observed the two reporter residues, R105 and F106. We used a nicked version of 30bp tandem DNA used earlier. It has two base pairs missing on one of the strands and the construct is listed in Table 2.1. Figure 2.8B shows result of this experiment (green) and it is very similar to 30 bp Tandem DNA. Adding a small defect in double stranded DNA (missing nucleotides in this case) does not cause MBD4 to fall off the DNA, and the overall exchange rate is similar to that of MBD4 on Tandem 30bp DNA. This data suggest that MBD4 does not always slide along the double stranded DNA and it may translocate with the help of small jumps along the DNA (hopping mechanism).

2.4.6 Effect of Salt concentration on MBD4 exchanging on inter-molecular and intra-molecular DNA binding sites

In an earlier work, it has been shown that salt concentration used in experimental procedures has dramatic effect on exchange rates of protein HoxD9 when bound to a DNA *(40)*. Here, we study an effect of increasing concentrations of NaCl on exchange rates of MBD4 when on a i) 1:1 mixture of 10bp regular mCpG and 10bp inverted DNA, ii) tandem 30 bp, and iii) tandem 30 bp nicked DNA. Figure 2.9 summarizes results for reporter residues R105, F106. As shown earlier, R105 (top panel Figure 2.9) shows slow exchange for inter-molecular sites and exchanges rapidly on intra-molecular sites in standard buffer (0 mM NaCl added). As increasing concentrations of NaCl are added, intra-molecular peaks become sharper (blue and green)

indicating a qualitative change in exchange rate. With increasing concentrations of NaCl, inter-molecular peaks (red) undergo profound changes, two peaks (indicating slow exchange) collapse into one peak (indicating fast exchange) at an average chemical shift. The other reporter residue F106 (bottom panel Figure 2.9) shows slow exchange for inter-molecular sites and intermediate exchange on intra-molecular sites in standard buffer (0 mM NaCl added). As increasing concentrations of NaCl are added, the two inter-molecular peaks (red) indicating slow exchange collapse into one peak at an average chemical shift indicating fast exchange. Peaks for inter-molecular exchange were not visible and thought to have broadened because of intermediate exchange. With addition of NaCl, F106 goes from intermediate exchange to fast exchange; peaks appear at an average position (green and blue). Also, note that there is no noticeable difference between tandem DNA with and without a defect (blue versus green). Higher salt concentration pushes MBD4 to undergo rapid exchange between the two sites.

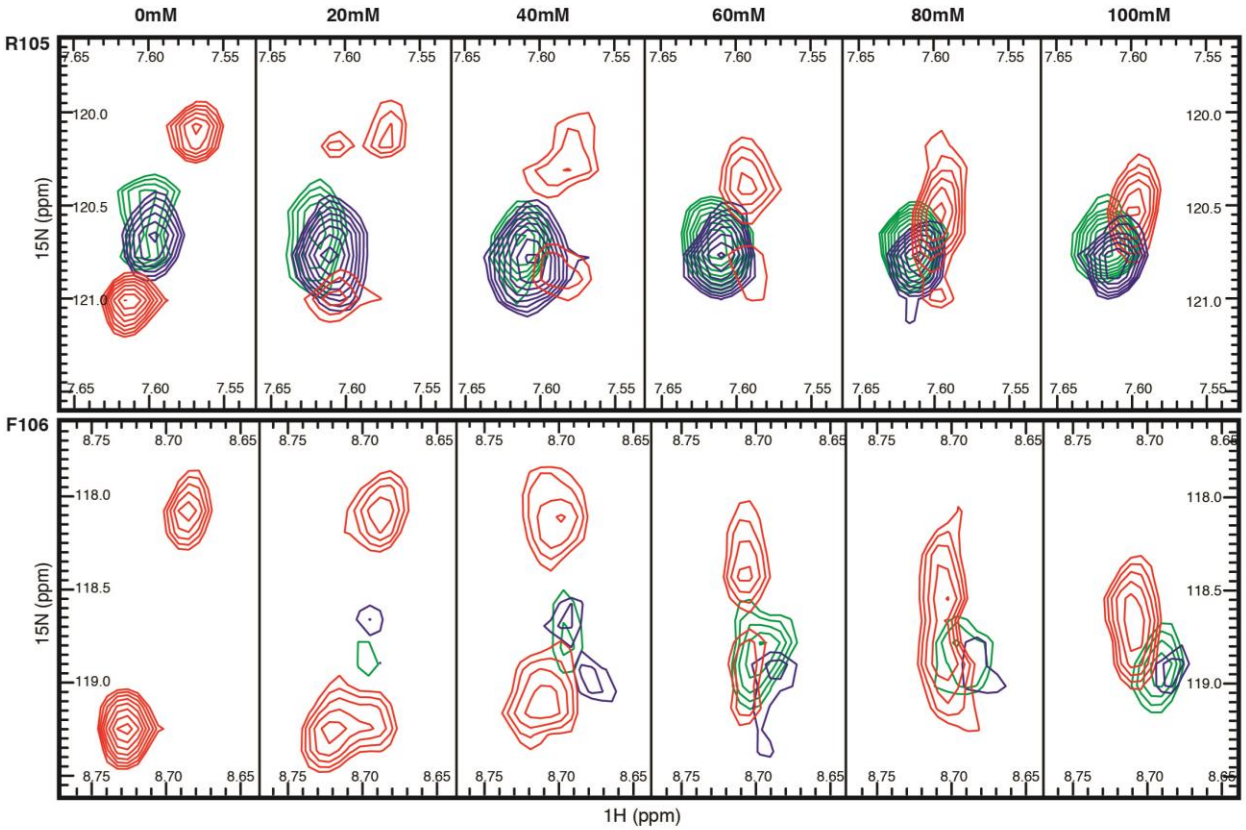


Figure 2.9: Effect of NaCl on inter-molecular and intra-molecular exchange of MBD4 between mCpG sites. Distinct chemical shifts for select amide resonances, R105 (top panel), F106 (bottom panel) are shown when bound to 1:1 mixture of methylated 10bp 1xmCpG and 10bp inverted DS DNA (red), tandem 30bp (blue) and tandem nicked 30bp (green). TROSY-HSQC experiments were carried out in standard buffer described in section 2.3.4. Increasing concentrations of NaCl (0 – 100 mM shown at the top panel) were added to the sample and spectra were recorded.

2.5 Discussion

Multifaceted MBD4 is proposed to have diverse biological roles and a broad spectrum of binding targets. In the studies presented here, we demonstrate that MBD domain of MBD4 has ability to bind multiple substrates. Unlike other MBD proteins, e.g. MBD2, MBD4 not only can bind mCpG but it also can target mismatch, hydroxymethylated and unmethylated DNA marks. Although, overall structural motifs and binding recognition mode of MBD4 are very similar to known MBDs to date, there can be local structural differences that account for diverse DNA binding abilities of MBD4. The recently solved crystal structure of the mouse MBD domain of MBD4 shows that the DNA interface of MBD4 has flexible structural features and can accommodate various bases due to the extensive water network involved *(32)*. Our structure of human MBD4 is in the final stages of refinement and has a highly similar DNA interface to that of highly homologues mouse MBD4. We also show that MBD4 can bind methylated, mismatch, hydroxymethylated and unmethylated DNA with little preference for methylated and mismatch DNA. The 17 bp oligonucleotide with one methylated CpG used in this experiment showed a stronger binding (~50 nM) with MBD domain of MBD2. This suggests that MBD4 may not be the most preferred biological reader for methylated CpG mark. In contrast, the relatively strong affinity towards mismatch DNA suggests that MBD4 is likely to be the most preferred reader for TpG sites. Considering the role of the glycosylase domain of MBD4 in excising TpG mismatch bases, a higher affinity towards TpG dinucleotide seems logical and may help with targeting the glycosylase domain to mismatch sites.

The MBD and glycosylase domains of MBD4 are linked with ~280 amino acid long largely disordered region of unknown function *(12, 41)*. The MBD and glycosylase domains are self-

sufficient for their action but are always linked with this linker domain. Previous work has shown that the glycosylase domain has no effect on its glycosylase activity towards a mismatch TpG whether it is acting alone or in context of full length MBD4 *(12)*. Thymine DNA glycosylase (TDG) is another glycosylase which was thought to have very similar biological role as that of glycosylase of MBD4 but it does not have methyl DNA binding domain. TDG can bind methylated and mismatch DNA with greater affinity than the glycosylase of MBD4. 5-formylcytosine and 5-carboxylcytosine, intermediates involved in the TET (ten-eleven-translocation) dependent deamination pathway are also primary targets for TDG. On the other hand, Manvilla *et al,* have shown that the structure of glycosylase itself has little selectivity for acting on TpG dinucleotide over methylated or hydroxymethylated DNAs *(31)*. Therefore, we hypothesize, having a MBD domain attached to glycosylase of MBD4 makes it better equipped to read TpG mismatches over TDG. Thus MBD4's better binding preference for TpG sites provides specificity for the glycosylase to act on. This finding is important and provides a functional difference between action of TDG and MBD4. Also, MBD4 glycosylase is targeted to TpG mismatch sites within CpG islands but glycosylase itself cannot discriminate whether adjacent CpG sites are methylated or not *(31)*. Therefore, MBD domain having higher affinity for methylated CpG over unmethylated CpG is crucial. The MBD domain can rapidly locate itself on mCpG islands and can very easily recognize TpG mismatch sites within the island. This suggests that the MBD domain can effectively recruit the glycosylase domain of MBD4 to its target (TpG) and thus underling its importance with respect to glycosylase activity of MBD4.

The role of MBD4 in repression is questionable. Kondo and colleagues proposed a role of MBD4 in transcription repression. Though MBD4 has been demonstrated to be localized to hyper-methylated promoters of affected genes, knockdown of MBD4 did not change the

repression status of these promoters. Since the MBD4 has a weaker binding affinity for mCpG when compared other MBD proteins, it is likely that repression was not a direct effect of MBD4 or other MBD proteins can easily substitute for the repression activity caused by MBD4. MBD4 has also shown to accumulate on hydroxymethylated CpG sites and was thereby proposed to be active in DNA demethylation *(41)*. From this study, the MBD domain of MBD4 can recognize hydroxymethylated DNA and, therefore, under certain circumstances (depending on stimulus and property of linker region of MBD4) it can localize on hydroxymethylated site, recruit glycosylase and thereby can cause erasure of DNA methylation mark.

It has always been intriguing how DNA binding proteins translocate into nuclei and accomplish the remarkable feat of finding their correct target sequences within huge genomic DNA. There are many examples in the literature about non-specific contacts made by DNA binding proteins with non-cognate DNA through phosphate backbone and base stacking *(42, 43)*. It has long been proposed that, in order to reach their target site, proteins first translocate along nonspecific DNA, i.e. move along DNA that does not contain a cognate binding site *(34)* and then find its cognate DNA targets (facilitated diffusion). Time spent by a DNA binding protein on a non-cognate DNA helps protein with facilitated diffusion but it can also slow down the target search process if a protein spends too much time on non-cognate sites. A correct balance of facilitated diffusion and random 3D jumps seems to be the answer for 'speed versus specificity' exhibited by DNA binding proteins. Here, using the recognition of mCpG sites by MBD as a model we provide new insights into DNA-protein recognition.

With the help of chemical exchange studies we demonstrate that MBD4 exchanges slowly on two DNA binding sites on two separate pieces of DNA (inter-molecular), but if linked together on a single DNA (intra-molecular), MBD4 exhibits fast exchange between the two sites.

This suggests that MBD4 prefers to hop across the DNA sequence until it finds its target rather than searching freely using three-dimensional diffusion. When sliding, DNA binding proteins once associated with non-specific DNA, do not dissociate and keep moving along the DNA in one dimension (reduced search space) to find its target site. When a defect (missing bases on one strand) was introduced on a nonspecific DNA sequence linking two cognate sites, the defect did not affect the fast exchange rate of MBD4. Not having two bases in a nicked version of Tandem 30 bp DNA ensures that MBD4 is not always in direct contact with non-specific DNA. Though not in contact with DNA it manages to scan along the DNA and therefore still exchanges rapidly between two sites. Over a short distance, protein stays in close proximity (within the 'electrostatic field' of DNA) without making direct contacts with it. These small 'hops' can be distinguished from 3D jumps where protein dissociates from the DNA completely and translocates to another DNA molecule in 3D space. Restriction enzyme EcoRV has shown to alternate between sliding and jumping on a non-cognate DNA sequence *(34)*. Here, we demonstrate that while scanning for its cognate binding site, MBD4 moves along the non-specific sites with the help of small 'hops'. This is a crucial phenomenon, because in the nucleus, DNA is always associated with small proteins, e.g. cofactors, and these small obstacles can be overcome by the proposed hopping mechanism.

DNA-protein interactions are highly dependent on ionic conditions, especially concentrations of monovalent ions. An increased NaCl concentration, for example, is shown to speed up association and dissociation between DNA and proteins *(34, 40)*. Here, we demonstrate that increasing the salt concentration forces MBD4 to exhibit faster exchange rates; both inter-molecular and intra-molecular. Salts have great effect on dielectric constants of the buffer and thereby greatly reduce association and dissociation times. This also weakens the electrostatic

field of the DNA and makes hopping less prevalent. The protein undergoes direct intermolecular transfer with the help of 3D jumps. In the experimental setup here, we have excess DNA binding sites and therefore it is very easy for MBD4 to find its cognate binding site. Inside the cell, however, if we allow protein to translocate only with the aid of 3D diffusion and calculate the chances of it colliding with its cognate binding site, proteins are thought to need hours before they can find their cognate sites. This is contrary to a millisecond timescale required for most DNA binding proteins to locate and activate their specific response genes *in vivo (44)*. It was shown that the *lac* repressor finds its target site 1000 times faster than time predicted based on simple 3D diffusion and collision processes *(45)*, suggesting facilitated diffusion mechanism.

In summary, we show that MBD4 has a wide range of binding specificity, but prefers methylated and mismatched DNA sequence over hydroxymethylated and unmethylated CpGs. We propose that the MBD domain of MBD4 helps carry the glycosylase cargo to the appropriate mismatch DNA sites within mCpG islands. Also, the scanning mechanism of MBD domain of MBD4 while recognizing mCpG is discussed and this is the first time the hopping mechanism is demonstrated for MBD4 in targeting a relatively diverse DNA methylation mark. This study provides useful insights into protein-DNA recognition mechanism. Further studies involving various lengths of the linker DNA between two sites and diverse defects within the DNA are necessary to fully understand how MBD4 translocates on DNA while scanning for its target sites.

2.6 References

1. X. Nan, R. R. Meehan, A. Bird, Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res.* **21**, 4886-4892 (1993).

2. B. Hendrich, A. Bird, Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538-6547 (1998).

3. R. E. Amir *et al.*, Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185-188 (1999).

4. R. J. Klose, A. P. Bird, Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89-97 (2006).

5. Y. Zhang *et al.*, Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13**, 1924-1935 (1999).

6. M. Saito, F. Ishikawa, The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J. Biol. Chem.* **277**, 35434-35439 (2002).

7. M. F. Fraga *et al.*, The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res.* **31**, 1765-1774 (2003).

8. X. Le Guezennec *et al.*, MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Mol. Cell. Biol.* **26**, 843-851 (2006).

9. B. Hendrich, J. Guy, B. Ramsahoye, V. A. Wilson, A. Bird, Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev.* **15**, 710-723 (2001).

10. J. Berger, O. Sansom, A. Clarke, A. Bird, MBD2 is required for correct spatial gene expression in the gut. *Mol. Cell. Biol.* **27**, 4049-4057 (2007).

11. B. Hendrich, U. Hardeland, H. H. Ng, J. Jiricny, A. Bird, The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature.* **401**, 301-304 (1999).

12. P. Wu *et al.*, Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4. *J. Biol. Chem.* **278**, 5285-5291 (2003).

13. E. Wong *et al.*, Mbd4 inactivation increases Cright-arrowT transition mutations and promotes gastrointestinal tumor formation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14937-14942 (2002).

14. L. Lopez-Serra *et al.*, A profile of methyl-CpG binding domain protein occupancy of hypermethylated promoter CpG islands of tumor suppressor genes in human cancer. *Cancer Res.* **66**, 8342-8346 (2006).

15. S. Laget *et al.*, The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. *PLoS One.* **5**, e11982 (2010).

16. X. Lin, W. G. Nelson, Methyl-CpG-binding domain protein-2 mediates transcriptional repression associated with hypermethylated GSTP1 CpG islands in MCF-7 breast cancer cells. *Cancer Res.* **63**, 498-504 (2003).

17. J. Qian *et al.*, Aberrant methylation of the death-associated protein kinase 1 (DAPK1) CpG island in chronic myeloid leukemia. *Eur. J. Haematol.* **82**, 119-123 (2009).

18. M. Sidiropoulos, G. Pampalakis, G. Sotiropoulou, D. Katsaros, E. P. Diamandis, Downregulation of human kallikrein 10 (KLK10/NES1) by CpG island hypermethylation in breast, ovarian and prostate cancers. *Tumour Biol.* **26**, 324-336 (2005).

19. O. Y. Mian *et al.*, Methyl-binding domain protein 2-dependent proliferation and survival of breast cancer cells. *Mol. Cancer. Res.* **9**, 1152-1162 (2011).

20. M. Chahrour *et al.*, MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science.* **320**, 1224-1229 (2008).

21. H. H. Ng, P. Jeppesen, A. Bird, Active repression of methylated genes by the chromosomal protein MBD1. *Mol. Cell. Biol.* **20**, 1394-1406 (2000).

22. E. Auriol, L. M. Billard, F. Magdinier, R. Dante, Specific binding of the methyl binding domain protein 2 at the BRCA1-NBR2 locus. *Nucleic Acids Res.* **33**, 4243-4254 (2005).

23. K. L. Ho *et al.*, MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol. Cell.* **29**, 525-531 (2008).

24. I. Ohki *et al.*, Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell.* **105**, 487-497 (2001).

25. J. N. Scarsdale, H. D. Webb, G. D. Ginder, D. C. Williams Jr, Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res.* **39**, 6741-6752 (2011).

26. M. Rooman, J. Lievin, E. Buisine, R. Wintjens, Cation-pi/H-bond stair motifs at protein-DNA interfaces. *J. Mol. Biol.* **319**, 67-76 (2002).

27. X. Zou, W. Ma, I. A. Solov'yov, C. Chipot, K. Schulten, Recognition of methylated DNA through methyl-CpG binding domain proteins. *Nucleic Acids Res.* **40**, 2747-2758 (2012).

28. B. K. Duncan, J. H. Miller, Mutagenic deamination of cytosine residues in DNA. *Nature.* **287**, 560-561 (1980).

29. J. P. Henderson, J. Byun, J. Takeshita, J. W. Heinecke, Phagocytes produce 5-chlorouracil and 5-bromouracil, two mutagenic products of myeloperoxidase, in human inflammatory tissue. *J. Biol. Chem.* **278**, 23522-23528 (2003).

30. K. Rai *et al.*, DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell.* **135**, 1201-1212 (2008).

31. B. A. Manvilla, A. Maiti, M. C. Begley, E. A. Toth, A. C. Drohat, Crystal structure of human methyl-binding domain IV glycosylase bound to abasic DNA. *J. Mol. Biol.* **420**, 164-175 (2012).

32. J. Otani *et al.*, Structural basis of the versatile DNA recognition ability of the methyl-CpG binding domain of methyl-CpG binding domain protein 4. *J. Biol. Chem.* **288**, 6351-6362 (2013).

33. E. Kondo, Z. Gu, A. Horii, S. Fukushige, The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. *Mol. Cell. Biol.* **25**, 4388-4396 (2005).

34. I. Bonnet *et al.*, Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Res.* **36**, 4118-4127 (2008).

35. M. Cai *et al.*, Solution structure of the phosphoryl transfer complex between the signal-transducing protein IIAGlucose and the cytoplasmic domain of the glucose transporter IICBGlucose of the Escherichia coli glucose phosphotransferase system. *J. Biol. Chem.* **278**, 25191-25206 (2003).

36. C. D. Schwieters, J. J. Kuszewski, N. Tjandra, G. M. Clore, The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65-73 (2003).

37. E. Ballestar, A. P. Wolffe, Methyl-CpG-binding proteins. Targeting specific gene repression. *Eur. J. Biochem.* **268**, 1-6 (2001).

38. L. Schrödinger, *The PyMOL Molecular Graphics System, Version 1.3 r1.Schrödinger, LLC, Portland, OR.*(2010).

39. N. A. Farrow, O. Zhang, J. D. Forman-Kay, L. E. Kay, 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 (1994).

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

41. G. Thillainadesan *et al.*, TGF-beta-dependent active demethylation and expression of the p15ink4b tumor suppressor are impaired by the ZNF217/CoREST complex. *Mol. Cell.* **46**, 636-649 (2012).

42. C. Jin, I. Marsden, X. Chen, X. Liao, Dynamic DNA contacts observed in the NMR structure of winged helix protein-DNA complex. *J. Mol. Biol.* **289**, 683-690 (1999).

43. C. O. Pabo, R. T. Sauer, Protein-DNA recognition. *Annu. Rev. Biochem.* **53**, 293-321 (1984).

44. T. Misteli, Protein dynamics: implications for nuclear architecture and gene expression. *Science*. **291**, 843-847 (2001).

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

Chapter 3

Solving the solution structure of MBD domain of MBD4 on methylated DNA by NMR.

3.1 Abstract

Nuclear magnetic resonance (NMR) spectroscopy has emerged as a powerful tool for structural studies of biological macromolecules in solution. Over the years it has undergone dramatic developments in both instrument hardware and methodology making it one of the most complex and constantly evolving biophysical techniques. Different groups of researchers use widely different approaches and sets of experiments to derive get the final three dimensional structures. Here, I describe the workflow used in solving the solution structure of MBD domain of human MBD4 bound to a methylated DNA. After a brief introductory section that describes overall process and pipeline strategy, NMR sample preparation, data analysis and structure calculation will be discussed. The section on NMR sample preparation describes an overall sample preparation strategy used in making labeled proteins which often is crucial and can be a bottleneck for the success of NMR project. Later, experiments collected for resonance assignment of backbone atoms and side-chain atoms are briefly described. How to extract structural restraints to use in NMR structure calculation are explained. The final structure is calculated by simulated annealing using the XPLOR-NIH software package.

3.2 Introduction

3.2.1 NMR spectroscopy in structural biology

Knowing the three-dimensional structure of a protein has always been insightful in understanding its biological function and mechanism of action. As of May 8, 2013, there are 90,424 structures deposited in protein data bank (PDB), out of which 79,970 (~88.5%) were solved with X-ray crystallography methods, 9,944 (~11.0%) with NMR spectroscopy methods and the one remaining with the help of electron microscopy. Although X-ray crystallography remains dominant, NMR spectroscopy is the only biophysical technique that can study dynamics and determine three dimensional (3D) structures in solution. Structural information provided by NMR spectroscopy can add to already existing information based on a crystal structure; NMR methods are sometimes used for refining existing x-ray structures. Along with these advantages, NMR also has few drawbacks; the main limitation of solution NMR spectroscopy is a limitation on the size of the molecule being investrigated. Larger protein spectra not only lead to significant resolution problems due to overlapping signals, but also face severe broadening of signals since larger molecule tumbles slowly in solution. This in turn leads to an increase of relaxation processes and thereby affecting overall resolution. Development of cryo-probes and high field strength magnets help with another drawback of NMR: insensitivity. Despite these shortcomings, NMR spectroscopy remains a great tool for studying 3D structures and dynamics. Here, we use NMR spectroscopy to solve the solution structure of the MBD domain on MBD4 bound to a DNA.

3.2.2 NMR spectroscopy pipeline for protein structure determination

Even today, NMR protein structure determination remains a tedious, costly and time consuming process. Figure 3.1 enlists standard processes involved in NMR structure determination pipeline. Target selection and cloning the most feasible region for study is a relative straightforward but crucial step and can have a huge impact on the success of the NMR experiments. Overexpression of labeled protein can be easily achieved with the help of standard molecular biology techniques. Making a labeled protein and screening the quality of the sample is described in section 3.3. Once you have a necessary quantity of a pure and stable protein, the next step of collecting NMR spectra is purely physical and depends on concentration of sample and NMR spectrometer. The set of experiments collected depends on the information being sought. All experiments were collected on Bruker Avance III^{TM} 700MHz NMR spectrometers at 25^{0}C. Initial data processing was done on Topspin^{TM}, software provided by Bruker. The more powerful and widely used software NMRPipe *(1)*, a script based NMR spectral processing and analysis package, was used for data processing. The CCPNMR analysis software *(2)* was used for displaying spectra, analyzing, peak picking, resonance assignments and NOE picking. The assignment process is briefly described in section 3.4. Later chapters discuss the process of structure determination. NOE distance restraints exported from CCPNMR and torsion angle restraints calculated using TALOS+ *(3)* were used for structure calculation by simulated annealing using the XPLOR-NIH software package *(4)*. With MBD4, pertinent steps are briefly described and resulting 3D structure is shown.

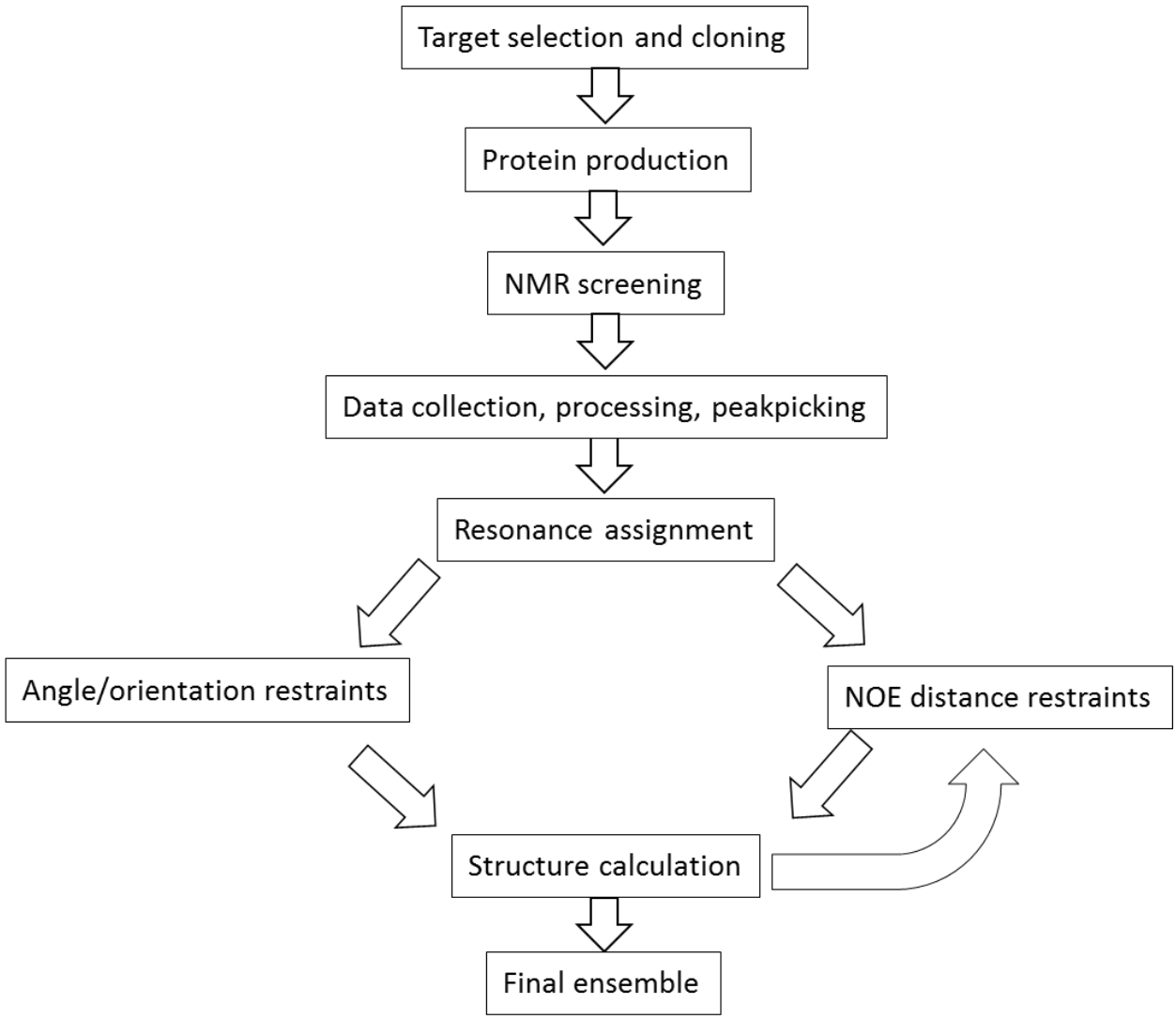


Figure 3.1: Standard process of NMR structure determination.

3.3 NMR sample preparation

3.3.1 General considerations

NMR is an inherently insensitive technique and also has a size limitation (< 40kD), thereby making the choice of protein construct under study very critical. The smaller and more stable the domain chosen for structure determination, the easier it is to get higher concentrations and better signal. On several occasions, researchers have found that sample preparation is the bottleneck to the structure determination and can be costly and time consuming. In x-ray crystallography, getting a good quality diffracting crystal is a challenging stage. But, in NMR, use of expensive isotopes (^{13}C, ^{15}N, ^{2}H) add cost and warrant a need for growing protein expressing bacteria in minimal media which can affect the final yield. Also, NMR experiments are conducted over longer periods (weeks) at room temperature (unless cryo-magnets are used), making the sample susceptible to degradation. Therefore, an extremely pure and stable sample is needed for structure determination.

3.3.2 Overexpression of Isotope labeled MBD4.

The MBD domain of human MBD4 (Figure 2.1) is a well-defined domain and known to bind a methylated CpG containing DNA *(5)*. The MBD domain was cloned as explained in section 2.3.1 keeping in mind that no unnecessary amino acids will be added to the final protein, thereby keeping the overall complex as small as possible. A minimal medium with defined nitrogen and carbon sources was used for making uniformly labeled protein. Bacteria containing the plasmid were grown on $^{15}NH_4Cl$ and ^{13}C-glucose to have ^{15}N and ^{13}C labeled protein. The recipe for minimal media (M9) is listed in Table 3.1.

With MBD4, bacteria containing MBD4 expression vector were streaked on Luria Bertani (LB) agar plate. Streak of these bacteria was used to inoculate 1 ml of LB media for several hours. When turbid, only 100 µL of this was added to 2 ml M9 media and incubated at 37^0C until the cells entered log phase. This 2ml was used as inoculum for 10ml and then 50 ml M9 media for overnight growth at 37^0C. This was used to inoculate 1L M9 and was grown until OD_{600} reached 0.8. It was then induced with 1mM IPTG for 4 hours and harvested as described in 2.3.1. To make ^2H labeled sample, all sources of H_2O were avoided and sample was grown in heavy water (^2H_2O). The bacterial doubling time in M9 is slower than in enriched media like LB, and it slows down further when heavy water is used for making triple labeled proteins. The protein purification process was optimized and is described in 2.3.1.

A 10 base pair oligonucleotide with one centrally methylated CpG (Table 2.1) was used as a binding partner for MBD4 and was purified as described in 2.3.2. The MBD domain was added to DNA at 1:1.1 ratio (10% excess DNA) and then buffer exchanged into 10 mM $NaPO_4$, pH 6.5, 1 mM dithiothreitol, 10% D_2O and 0.02% sodium azide. At least 0.5 ml of this mixture at concentration of 200 µM was added to a regular NMR tube and the quality of the sample was tested using ^{15}N-HSQC.

3.3.3 NMR sample screening with ^1H-^{15}N-HSQC experiment.

^1H-^{15}N-HSQC is the most standard 2D heteronuclear experiment and requires at least ^{15}N labeling. Magnetization is transferred from a proton to attached ^{15}N nuclei via J-coupling. Chemical shift is evolved on nitrogen and then transferred back to the more sensitive proton for detection. This experiment shows all N-H pairs, mainly amide in a peptide bond. The HSQC spectrum contains a cross peak for each backbone amide group (except Pro), the NH_2 side chain

groups of Asn & Gln and the aromatic NH groups of Trp. The $N_{\varepsilon}H_{\varepsilon}$ of Arg is also visible, but chemical shift generally lies outside the area usually covered (103 ppm -133ppm for MBD4) by the HSQC spectrum. A wide spectral width (103 ± 30 ppm in the case of MBD4) was used to detect these ARG $N_{\varepsilon}H_{\varepsilon}$ resonances. A typical HSQC spectrum of MBD4 on a 10 bp DNA is shown in Figure 3.2.

HSQC spectrum provides a footprint of each protein and provides an overall idea about the nature of the protein under study. A widely spread HSQC spectrum indicates that protein is folded, more likely to be stable and worth pursuing further structure solving process. Although it is not possible to assign cross peaks to specific residues using an HSQC alone, the spectrum provides valuable information about the quality of data to expect from a given sample.

Table 3.1: Recipe for 1L minimal medium (M9).

H_2O (D_2O if needed)	970.0 ml
KH_2PO_4 (Anhydrous)	13.0 g
K_2HPO_4 (Anhydrous)	10.0 g
Na_2HPO_4 (Anhydrous)	9.0 g
K_2SO_4 (Anhydrous)	2.4 g
NH_4Cl (^{15}N-labeled if needed)	1.0 g
D-Glucose (^{13}C-labeled if needed)	3.0-5.0 g
Dissolve and sterile filter above and add below	
Trace Elements	10.0 ml
Trace Elements Recipe per 100ml H_2O $FeSO_4$ ($7H_2O$) 0.60 g $CaCl_2$ ($2H_2O$) 0.60 g $MnCl_2$ ($4H_2O$) 0.12 g $CoCl_2$ ($6H_2O$) 0.08 g $ZnSO_4$ ($7H_2O$) 0.07 g $CuCl_2$ ($2H_2O$) 0.03 g H_3BO_3 2 mg $(NH_4)_6$ Mo_7 O_{24} ($4H_2O$) 0.025 g EDTA 0.50 g Sterile filter. (lyophilize to remove water if needed)	
1M $MgCl_2$ stock (powdered form if needed)	10.0 ml
5 mg/ml Thiamine stock (powered form if needed)	6.0 ml
100mg/ml ampicillin (appropriate antibiotic)	1.0 ml

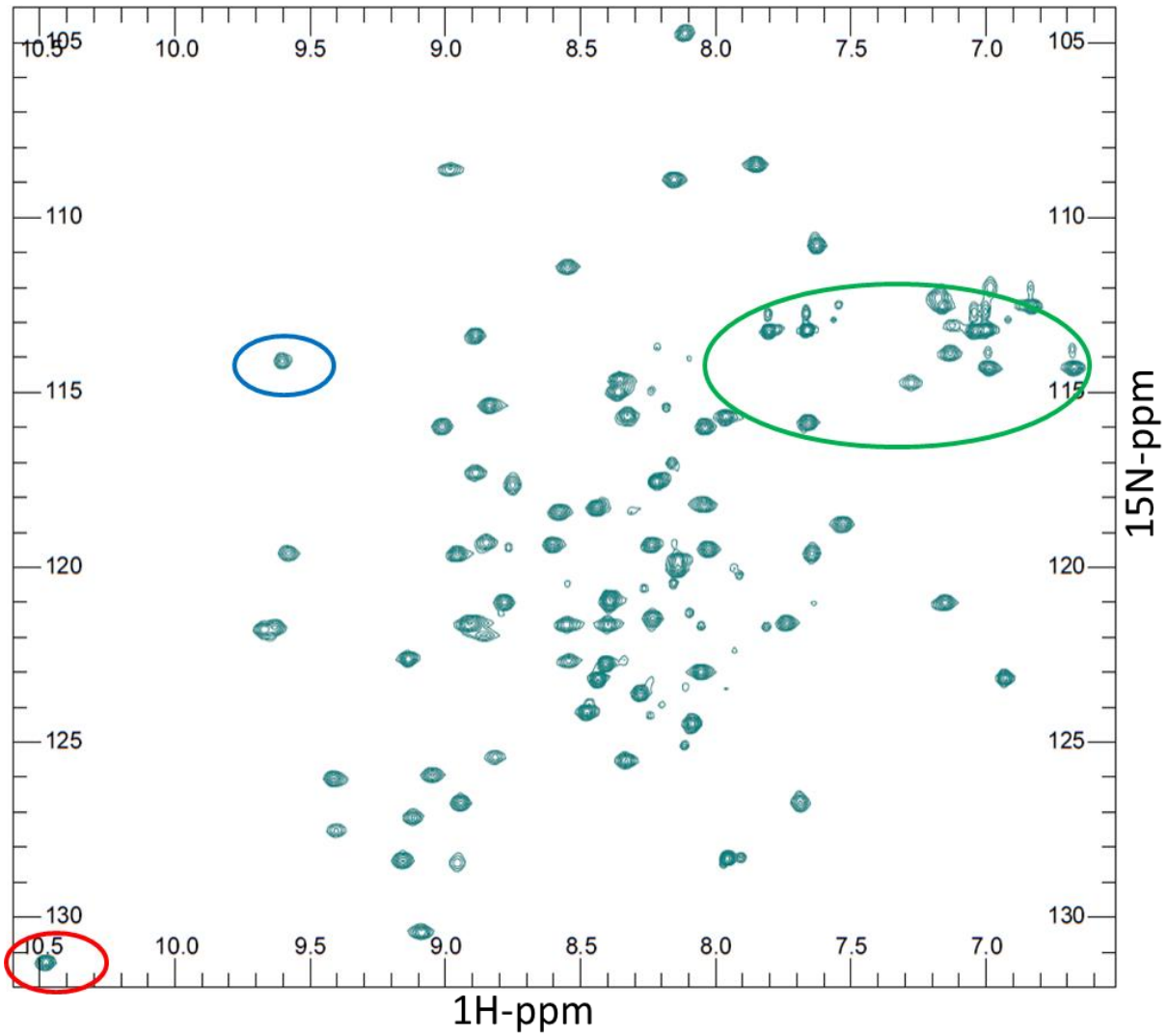


Figure 3.2: ^{1}H-^{15}N-HSQC of MBD4 on methylated DNA. A peak for each backbone amide group (except Pro) is visible. The NH_2 side chain groups of Asn & Gln are also visible and generally present in region circled in green. The aromatic NH groups of Trp are detected in region circled in red. The $N_{\varepsilon}H_{\varepsilon}$ of Arg is also visible but generally folded (circled in blue).

3.4 Resonance Assignment

3.4.1 Backbone resonance assignment

The protein backbone is comprised of H, N, CA, CB and C atoms and the assignment of the backbone resonances is a prerequisite for further investigation by NMR spectroscopy. A series of heteronuclear experiments correlating inter and intra residual H, N, Cα, Cβ and C through scalar couplings along bonds are collected in uniformly ^{13}C and ^{15}N labeled proteins. This results in a set of sequential resonances corresponding to backbone atoms. Assigning these resonances to a particular amino acid in native sequence is a vital step in NMR protein structure determination. Typically, backbone resonance assignment is divided into three steps; forming spin systems, linking spin systems into fragments, and mapping the fragments to the target sequence. A spin system denotes a group of coupled nuclei that can be observed as cross-peaks in one or more spectra. Usually spin systems contain both inter-residue and intra-residue information. For MBD4, the set of experiments mentioned in Table 3.2 were collected and processed data were imported in CCPNMR. After initializing HSQC, peaks were picked in all spectra depending on correlation observed. The type and number of peaks expected from each spectra are mentioned in Table 3.2. This process results in list of resonances that are linked (i-1, i, i+1 *etcetera*). The Biological Magnetic Resonance data Bank (BMRB) has statistically calculated chemical shift possibilities for all atoms in 20 amino acids. Within CCPNMR; with the help of sequence specific assignment, resonance tables and distributions of chemical shifts by amino acid type as found in BMRB; resonances for backbone atoms can be linked and assigned to the native sequence. Backbone assignment is important as it provides the basic framework for

structure determination. Chemical shifts of backbone residues are also used for secondary structure prediction and thereby used in deriving angle restraints.

3.4.2 Side-chain resonance assignment

The assignment of side-chain resonances is the next step and depends on prior knowledge of the backbone spin system. A series of 3D experiments were collected and are listed in Table 3.2. The CCHTOCSY experiments correlates all aliphatic ^{1}H and ^{13}C spins within residues and is used to assign all observable side-chain resonances to the backbone resonances. For larger proteins, CCHTOCSY spectra can be complicated due to overlap and more experiments are needed to sort through all peaks (Table 3.2). Prior knowledge of average chemical shifts for observable side chain protons is very helpful and is shown in Figure 3.3. Protons and carbons of aliphatic side-chain were assigned as mentioned above. For the assignment of aromatic side-chain proton experiments HBCBCGCDHD and HBCBCGCHHDHE experiments were collected which correlate carbon CB to proton HD or HE of an aromatic side-chain. Also ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC spectra can be used for resonance assignments with the assumption that protons close enough in space should be visible in these spectra. For detailed information about assignment procedures, please refer to a tutorial written by Victoria Higman (http://www.protein-nmr.org.uk/solution-nmr/assignment-theory/). For MBD4, almost all backbone residues were assigned and high percentage of side-chain resonances were assigned. All assignments are summarized in Table 3.3. Assignments of side-chain resonances are crucial as they are most likely to provide long distance NOE cross-peaks, thereby helping with three-dimensional structure determination.

3.4.3 Assignment of NOESY cross-peaks

After the complete assignments of backbone and side-chain resonances, ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC spectra were collected. These correlate protons with N-H or C-H pair that are proximate in space. Assigning these complex spectra yields vital structural restraints used in structure determination. In the initial stage of assignment, usually only a fraction of total NOESY cross-peaks are assigned, but additional NOESY peaks are assigned in later stages of structure calculations with a help of initial structural folds.

Table 3.2: Experiments utilized in backbone and side-chain resonance assignment.

Experiments utilized in assignment of backbone atoms

Experiment Name	Dimension	Peaks observed.
1H-15N-HSQC	2 (15N,1H)	Amide NH pairs, also side chains of Arg, Trp, Asn,Gln)
HNCA	3 (13C, 1H, 15N)	Strong CA(i) and weak CA(i-1)
HNCO	3 (13C, 1H, 15N)	C (i-1)
HNCACO	3 (13C, 1H, 15N)	Strong C(i) and weak C(i-1)
HNCACB	3 (13C, 1H, 15N)	Strong CA(i), CB(i), and weak CA(i-1), CB (I-1)
CBCACONH	3 (13C, 1H, 15N)	CA (i-1) and CB (i-1)

Experiments utilized in assignment of side-chain atoms

Experiment Name	Dimension	Peaks observed.
HBHACONH	3 (1H,1H,15N)	HA (i-1) and HB (i-1)
CCHTOCSY	3 (13C, 1H, 13C)	CA (i), CB (i), CG (i), CD(i), CE(i) *etcetera*
CDIPSI	3 (13C, 1H, 15N)	CA(i-1), CB (i-1), CG (i-1), CD(i-1), CE(i-1) *etcetera*
HDIPSI	3 (1H, 1H, 15N)	HA (i-1), HB(i-1), HG(i-1), HD(i-1) *etcetera*

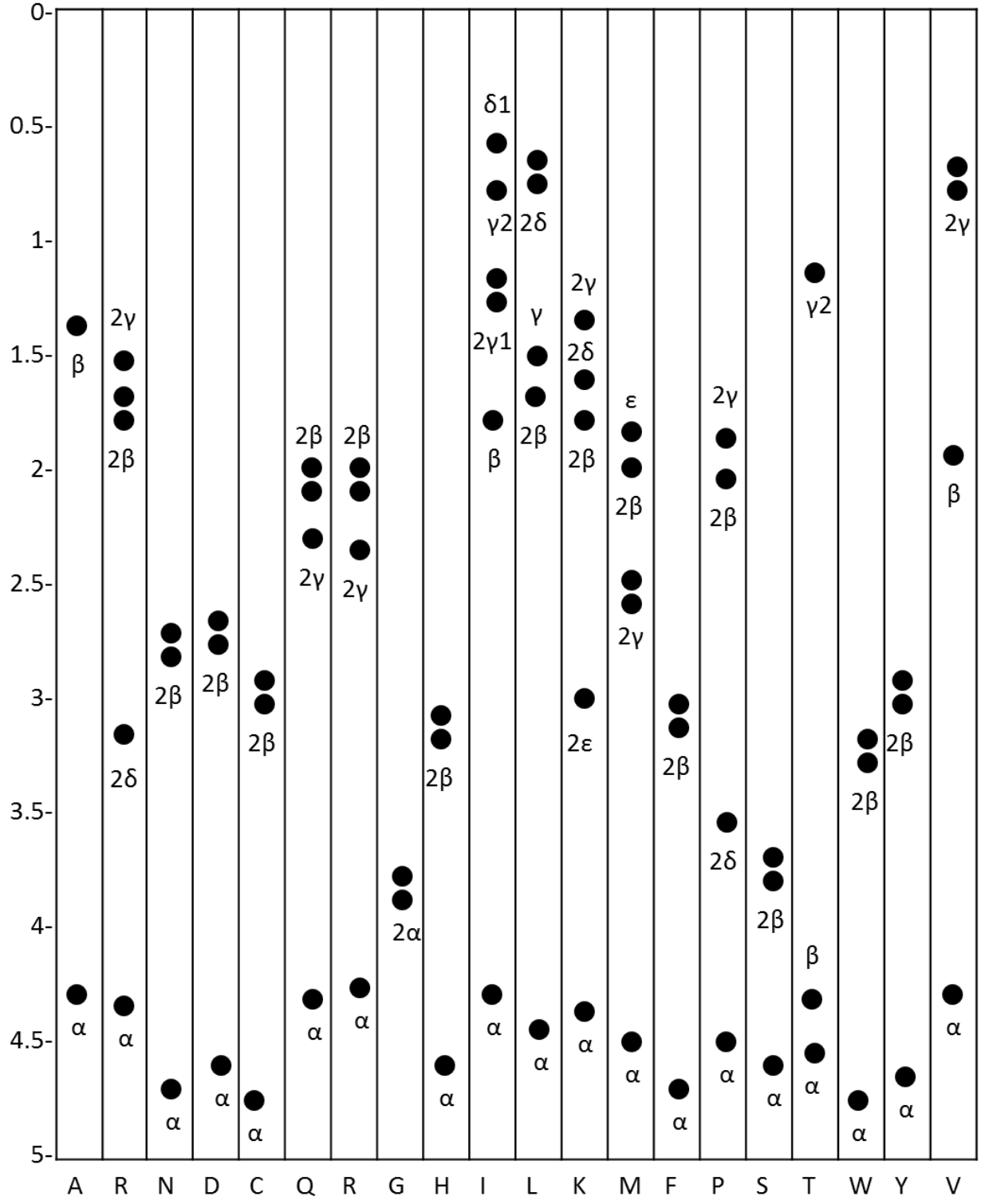


Figure 3.3: Average chemical shifts for observable side-chain protons of twenty amino acids.

Table 3.3: Chemical shift table in ppm for backbone and side chain resonances of amino acids in MBD4.

!		H	N	HA	HB	HG	HD	C	CA	CB	CG	CD
79	S	-	-	-	-	-	-	174.96	58.54	64.01	-	-
80	T	8.34	115.66	4.38	4.29	1.22	-	174.57	62.14	69.64	21.31	-
81	E	8.43	123.11	4.33	1.97,2.05	2.27,2.33	-	176.12	56.67	30.26	36.34	-
82	C	8.39	120.92	4.48	2.90	-	-	174.34	58.46	27.99	-	-
83	R	8.48	124.15	4.37	1.90,1.79	1.63	3.21	176.14	56.27	30.86	27.10	43.31
84	K	8.41	122.77	-	-	-	-	176.56	55.97	33.33	-	-
85	S	8.44	118.29	4.41	3.81	-	-	173.67	58.35	63.93	-	-
86	V	8.28	123.63	4.29	2.10	1.10	-	-	60.18	33.38	21.79	-
87	P	-	-	4.20	2.04,1.47	1.16,0.59	3.45,2.68	175.67	63.62	32.04	27.25	51.04
88	C	8.39	120.99	4.24	2.98	-	-	176.34	61.16	26.92	-	-
89	G	9.01	115.99	4.41,3.78	-	-	-	174.17	45.44	-	-	-
90	W	8.41	121.61	4.99	3.45,3.28	-	7.26	175.33	57.49	29.51	-	127.19 131.31:Ne1
10.48:He1	7.56:Hz2	6.81:Hz3	6.63:Hh2	114.39:Cz2	121.64:Cz3	124.29:Ch2						
91	E	8.96	119.67	4.85	1.97,1.88	2.10,2.26	-	174.31	55.07	33.41	35.98	-
92	R	8.94	126.74	-	-	-	-	174.64	54.94	33.25	-	-
93	V	9.10	130.44	4.13	1.18	0.82,0.73	-	173.39	61.55	34.33	20.83,21.15	-
94	V	8.34	125.53	4.66	1.95	1.42	-	175.83	61.30	33.09	21.43	-
95	K	9.16	128.34	4.88	1.83,2.00	1.51,1.42	1.84	175.17	54.74	36.17	25.31	28.94
3.09:Hea	3.16:Heb	42.21:Ce										
96	Q	8.94	128.36	4.47	1.83	2.00,1.20	-	175.67	54.82	30.08	33.44	-
97	R	9.12	127.15	4.44	1.84,1.85	1.63	2.83,3.00	178.21	57.72	30.37	26.95	44.37
98	L	9.05	125.93	4.34	0.96,1.46	0.90	0.68	176.65	55.70	44.06	25.08	22.50
99	F	8.25	119.36	4.95	3.00,3.22	-	7.34	173.85	56.70	42.76	-	132.58
100	G	8.11	104.71	4.54,3.76	-	-	-	176.17	44.10	-	-	-
101	K	9.13	122.60	4.11	2.00	1.67,1.59	1.83	179.61	59.66	32.77	25.35	29.84
3.08:Hea	41.76:Ce											
102	T	8.98	108.65	4.34	4.46	1.37	-	172.97	61.32	68.97	21.67	-
103	A	6.93	123.17	3.37	1.07	-	-	178.08	54.12	17.94	-	-
104	G	8.54	111.43	3.42,4.48	-	-	-	175.27	44.48	-	-	-
105	R	7.65	119.64	-	-	-	-	174.18	56.78	30.64	-	-
106	F	8.75	117.62	5.42	2.75,2.90	-	-	175.44	57.19	42.53	-	-
107	D	8.86	119.34	5.18	2.52,2.45	-	-	173.61	53.68	45.29	-	-
108	V	8.55	121.64	4.92	1.98	0.53,0.84	-	174.76	60.81	33.93	21.78,22.67	-
109	Y	8.89	121.62	5.42	3.02,2.90	-	6.89	173.43	55.46	42.04	-	133.36
6.66:He*	118.03:Ce*											
110	F	8.90	117.34	5.85	3.44,2.98	-	-	175.42	56.71	44.12	-	-
111	I	9.66	121.79	5.23	1.83	1.58,1.05,0.95	0.87	176.36	59.98	39.90	17.64,27.94	13.65
112	S	9.42	126.06	3.05	3.25,3.43	-	-	-	56.70	63.15	-	-
113	P	-	-	4.26	1.77,2.39	2.00,1.65	1.61,2.90	177.21	64.91	31.44	27.48	49.06
114	Q	7.17	112.29	4.38	2.25,1.81	2.46,2.37	-	175.99	55.82	28.65	34.59	-
115	G	8.15	108.91	4.15,3.43	-	-	-	174.16	45.32	-	-	-
116	L	7.17	121.03	4.18	1.08,1.42	1.29	0.88,0.83	175.19	54.77	42.54	27.56	23.33,25.22

Table 3.3 continued…

117	K	8.24	121.58	5.27	1.52,1.66	1.23,1.42	1.40,1.56	175.98	55.43	34.56	26.14	29.49	2.73:Hea 41.73:Ce	
118	F	9.64	121.82	4.91	3.27,3.09	-	-	176.45	58.01	42.98	-	-		
119	R	9.59	119.54	5.11	2.02,2.12	1.56	3.13	174.70	55.17	31.82	27.43	43.96		
120	S	7.62	110.78	4.34	3.44,3.94	-	-	172.79	56.03	66.49	-	-		
121	K	8.82	125.44	3.64	1.80,1.62	1.39,1.45	1.77	178.83	59.65	31.15	25.06	29.22		
122	S	8.83	115.39	-	-	-	-	177.43	61.35	62.34	-	-		
123	S	8.04	118.25	-	-	-	-	178.02	61.62	63.07	-	-		
124	L	7.69	126.77	3.65	1.40,1.62	0.81	-0.36,-0.11	177.99	58.30	42.31	27.19	24.33,23.97		
125	A	8.91	121.66	3.69	1.44	-	-	180.03	55.69	17.88	-	-		
126	N	7.96	115.71	4.50	2.96,2.98	-	-	177.17	56.37	38.52	-	-		
127	Y	7.75	121.60	4.29	3.41,3.07	-	7.08	177.89	61.58	39.05	-	133.71	6.97:He* 118.19:Ce*	
128	L	8.60	119.36	3.80	1.73,1.43	1.83	0.82,0.42	179.42	57.75	40.89	26.31	25.39,21.94		
129	H	8.04	115.99	4.51	3.31	-	-	178.01	58.09	29.35	-	-		
130	K	8.14	119.80	4.08	1.88,1.92	1.55,1.50	1.72	177.32	58.36	32.57	25.01	29.19	3.04:Hea	42.02:Ce
131	N	7.66	115.91	4.66	2.67,2.27	-	-	175.34	53.59	39.28	-	-		
132	G	7.85	108.48	3.91	-	-	-	174.65	46.33	-	-	-		
133	E	8.14	120.05	4.36	2.05,2.10	2.30	-	177.32	56.71	29.89	36.54	-		
134	T	8.36	114.99	4.36	4.38	1.22	-	174.93	62.08	69.98	21.71	-		
135	S	8.58	118.37	-	-	-	-	174.28	59.51	63.81	-	-		
136	L	8.05	123.00	4.46	1.67,1.34	0.93	0.78	175.10	54.52	44.47	25.75	23.50		
137	K	8.79	121.09	4.77	1.82,1.70	1.68	1.67	-	53.13	33.23	24.78	28.87	2.96:Hea 2.98:Heb	41.89:Ce
138	P	-	-	3.96	2.10,1.99	2.31,1.76	3.71,3.87	177.65	66.04	31.47	28.31	49.98		
139	E	8.89	113.38	4.14	2.04,1.97	2.32,2.26	-	176.69	58.77	28.53	36.89	-		
140	D	8.03	119.52	4.30	2.21,2.54	-	-	174.54	55.91	40.87	-	-		
141	F	-	-	-	-	-	6.05	-	-	-	-	-		
141	F	7.54	118.83	3.87	2.75,1.75	-	-	173.52	57.27	39.48	-	132.01		
142	D	8.54	122.73	-	-	-	-	176.42	53.72	41.68	-	-		
143	F	9.41	127.53	4.51	3.57,3.10	-	7.41	174.88	59.00	39.50	-	132.59		
144	T	8.36	114.74	4.59	4.27	1.30	-	-	62.52	70.26	21.46	-		
145	V	-	-	4.00	1.98	0.90,0.65	-	176.10	62.70	32.59	21.19,21.27	-		
146	L	8.08	124.46	4.37	1.56,1.62	1.58	0.92,0.85	176.92	54.96	42.44	27.02	24.70,23.65		
147	S	8.22	117.56	-	-	-	-	173.42	58.23	63.95	-	-		
148	K	7.96	128.34	4.16	1.81,1.69	1.36,1.38	1.68	-	57.72	33.75	24.69	29.06	2.96:Hea 2.99:Heb	41.93:Ce

3.5 Structural parameters for NMR

Several of NMR experiments generate structural restraints used for structure calculations. The most widely used restraints are inter-proton distances measured by NOE, Hydrogen bond restraints predicted from secondary structure and backbone torsion angles defined by chemical shift analysis. Orientation restraints generated by residual dipolar coupling (RDC) are not required but are being used much more commonly.

3.5.1 Hydrogen bond restrains

Protein secondary structures (α helix or β sheet) are well studied and are known to exhibit geometrically driven hydrogen bonding. In the case of an α helix, a linear hydrogen bond is set between the amide proton and the oxygen of carboxyl group of $(n-4)^{th}$ amino acid. In the case of a β sheet, hydrogen bonding exists between two parallel or anti-parallel strands. This information is very useful and can have a large impact on accuracy of final resulting structure. Hydrogen bonding information is taken only from region with well-defined secondary structure. With MBD4, we defined one α helix and two anti-parallel β strands using TALOS+.

TALOS+ breaks the sequence into overlapping amino acid triplicates and compares their chemical shifts with a database of homologous polypeptides with known torsion angles and chemical shifts. It predicts backbone torsion angles based on their chemical shift. Chemical shift data from CCPNMR was exported using format converter tool. The format used for TALOS+ calculation is shown in Table 3.4. TALOS+ predictions of backbone torsion angles and secondary structure for each residue were calculated. The Degree of prediction accuracy is depicted in green (good), yellow (ambiguous) and blue (dynamic) in Figure3.4.

Table 3.4: A chemical shifts table for TALOS+. (Note that residue ID starts from 1.)

```
REMARK File written by CcpNmrFormat converter.
DATA FIRST_RESID  1

DATA SEQUENCE GSTECRKSVP CGWERVVKQR LFGKTAGRFD VYFISPQGLK FRSKSSLANY
DATA SEQUENCE LHKNGETSLK PEDFDFTVLS K

VARS   RESID RESNAME ATOMNAME SHIFT
FORMAT %4d %1s %4s %8.3f

   2 S    C  174.960
   2 S   CA   58.539
   2 S   CB   64.005
   3 T    C  174.574
   3 T   CA   62.185
   3 T   CB   69.650
   3 T  CG2   21.358
   3 T   HN    8.340
   3 T   HA    4.374
   3 T   HB    4.279
   3 T    N  115.664
   3 T HG2#    1.219
   4 E    C  176.123
   4 E   CA   56.672
   4 E   CB   30.260
   4 E   CG   36.338
   4 E   HN    8.432
   4 E   HA    4.327
   4 E  HB2    2.055
   4 E  HB1    1.975
   4 E  HG2    2.328
   4 E  HG1    2.269
   4 E    N  123.107
   5 C    C  174.341
   5 C   CA   58.546
   5 C   CB   27.941
   5 C   HN    8.392
   5 C   HA    4.486
   5 C  HB1    2.899
   5 C    N  120.920
   6 R    C  176.140
   6 R   CA   56.215
   6 R   CB   30.862
   6 R   CD   43.339
   6 R   CG   27.099
   6 R   HN    8.478
   6 R   HA    4.386
   6 R  HB2    1.788
   6 R  HB1    1.901
   6 R  HD1    3.211
   6 R  HG1    1.635
   6 R    N  124.149
   7 K    C  176.555
INCOMPLETE
```

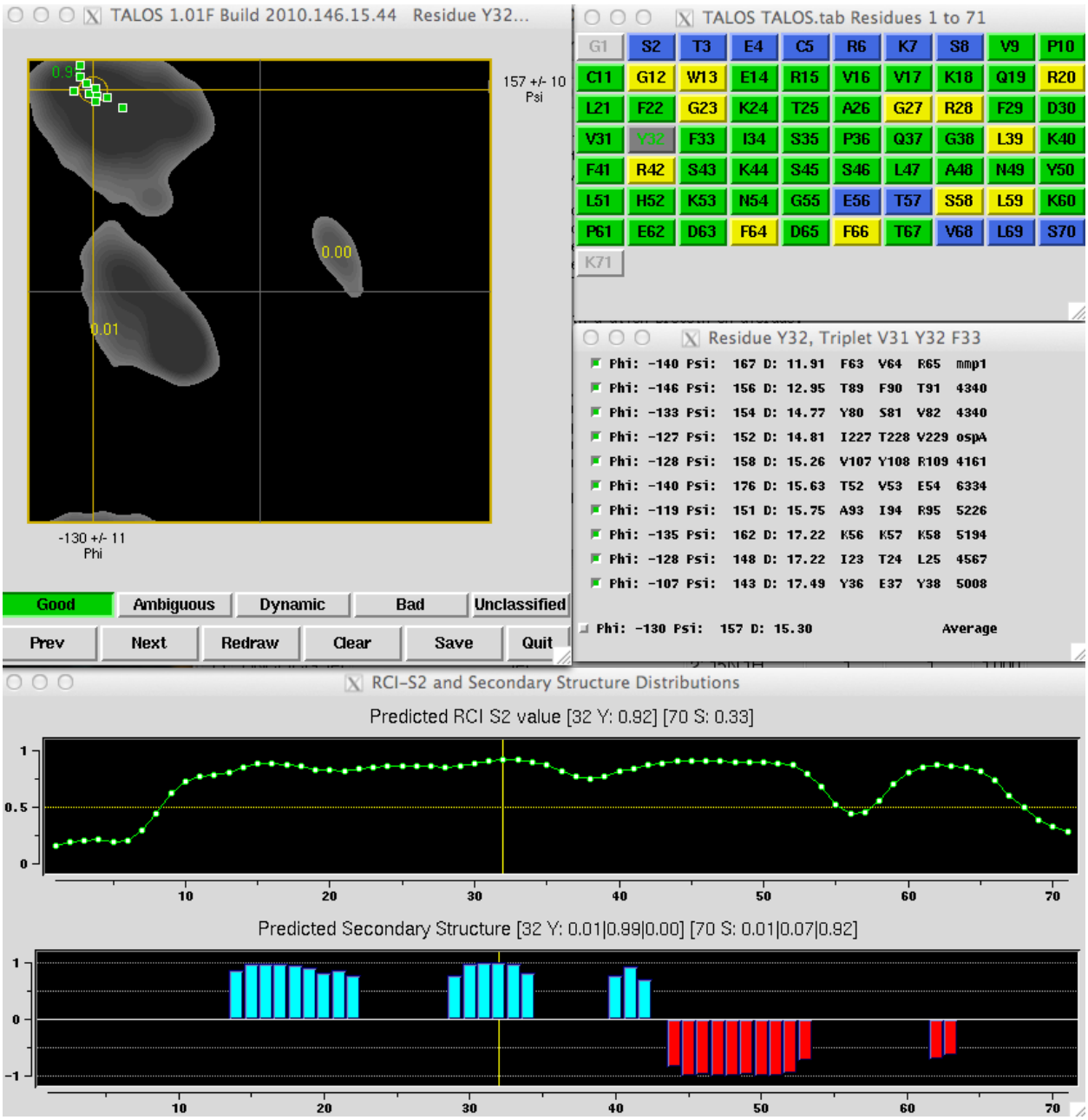


Figure 3.4: TALOS+ analysis for MBD4. Backbone torsion angles (ψ and φ) are predicted using chemical shifts. Predicted secondary structures are shown in blue (β strands) and red (α helices).

Secondary structures predicted by TALOS+ can be seen in Figure 3.4. Three β strands (blue and positive) and two α helices (red and negative) are predicted by TALOS+. NOE data were used to confirm the secondary structure predictions. The short β strand and α helix were not included in the structure calculation, thereby using only two strands (91E-99F, 106E-111I) and one helix (121K-130K) for generating structural restraints. Hydrogen bond distance and Hydrogen bond distance angle tables were generated for structured residues and are listed in Table 3.5 and Table 3.6.

3.5.2 Torsion angle restraints

Proteins have backbone dihedral angles (φ, ψ, ω) and side chain dihedral angles (χ1-2-3…) and are depicted in Figure 3.5. Torsion angles can be determined from several J scalar coupling interactions between atoms and they provide geometric information between atoms in a molecule. TALOS+ analysis predicts psi and phi torsion angles which range from -180^0 to $+180^0$. Torsion angle ω is approximately equal to 180^0 or 0^0 because of the planer nature of peptide bond so it is not often measured. The φ and ψ angles from well predicted residues (shown in green, Figure 3.4) are used for structure calculation. The format for the table is shown in Table 3.7. Additional scalar coupling experiments are conducted to determine the side-chain torsion angles. Experiments HNCG_ARO and HNCOCG_ARO are used to measure χ1 for F, W, Y and H amino acids. Side chain torsion angles are typically near 180^0 (trans), 60^0 (gauche+) and -60^0 (gauche-). A side-chain torsion angle table was generated and these parameters were used as structural restraints (Table 3.8).

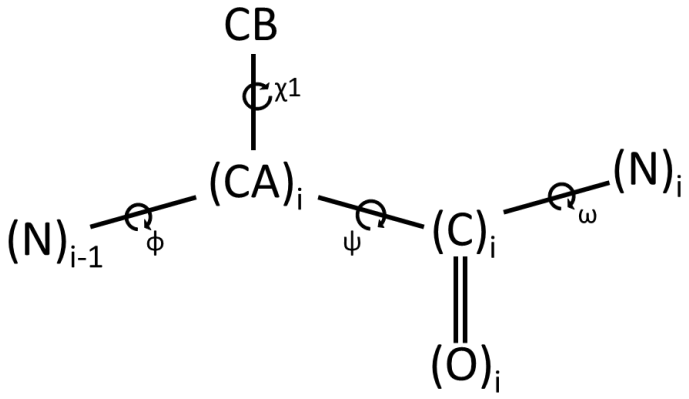


Figure 3.5: Torsion angles φ, ψ, ω and χ1 in proteins.

Table 3.5: Hydrogen bond table. Hydrogen bond restraints are defined based on TALOS+ prediction of secondary structures.

hb_alpha.tbl

```
assign (resid 121 and name O )(resid 125 and name HN ) 2.3 1.0 0.2
 assign (resid 121 and name O )(resid 125 and name N ) 3.3 1.0 0.2

assign (resid 122 and name O )(resid 126 and name HN ) 2.3 1.0 0.2
 assign (resid 122 and name O )(resid 126 and name N ) 3.3 1.0 0.2

assign (resid 123 and name O )(resid 127 and name HN ) 2.3 1.0 0.2
 assign (resid 123 and name O )(resid 127 and name N ) 3.3 1.0 0.2

assign (resid 124 and name O )(resid 128 and name HN ) 2.3 1.0 0.2
 assign (resid 124 and name O )(resid 128 and name N ) 3.3 1.0 0.2

assign (resid 125 and name O )(resid 129 and name HN ) 2.3 1.0 0.2
 assign (resid 125 and name O )(resid 129 and name N ) 3.3 1.0 0.2

assign (resid 126 and name O )(resid 130 and name HN ) 2.3 1.0 0.2
 assign (resid 126 and name O )(resid 130 and name N ) 3.3 1.0 0.2
```

hb_beta.tbl

```
assign (resid 91 and name O )(resid 111 and name HN ) 2.3 1.0 0.2
 assign (resid 91 and name O )(resid 111 and name N ) 3.3 1.0 0.2
assign (resid 111 and name O )(resid 91 and name HN ) 2.3 1.0 0.2
 assign (resid 111 and name O )(resid 91 and name N ) 3.3 1.0 0.2

assign (resid 93 and name O )(resid 109 and name HN ) 2.3 1.0 0.2
 assign (resid 93 and name O )(resid 109 and name N ) 3.3 1.0 0.2
assign (resid 109 and name O )(resid 93 and name HN ) 2.3 1.0 0.2
 assign (resid 109 and name O )(resid 93 and name N ) 3.3 1.0 0.2

assign (resid 95 and name O )(resid 107 and name HN ) 2.3 1.0 0.2
 assign (resid 95 and name O )(resid 107 and name N ) 3.3 1.0 0.2
assign (resid 107 and name O )(resid 95 and name HN ) 2.3 1.0 0.2
 assign (resid 107 and name O )(resid 95 and name N ) 3.3 1.0 0.2
```

Table 3.6: Hydrogen bond distance angle table. Hydrogen bond distance angles restraints are defined based on TALOS+ prediction of secondary structures.

hbda_alpha.tbl

```
assign (resid 125 and name n)(resid 125 and name hn)(resid 121 and name o)

assign (resid 126 and name n)(resid 126 and name hn)(resid 122 and name o)

assign (resid 127 and name n)(resid 127 and name hn)(resid 123 and name o)

assign (resid 128 and name n)(resid 128 and name hn)(resid 124 and name o)

assign (resid 129 and name n)(resid 129 and name hn)(resid 125 and name o)

assign (resid 130 and name n)(resid 130 and name hn)(resid 126 and name o)
```

hbda_beta.tbl

```
assign (resid 91 and name n)(resid 91 and name hn)(resid 111 and name o)
assign (resid 111 and name n)(resid 111 and name hn)(resid 91 and name o)

assign (resid 93 and name n)(resid 93 and name hn)(resid 109 and name o)
assign (resid 109 and name n)(resid 109 and name hn)(resid 93 and name o)

assign (resid 95 and name n)(resid 95 and name hn)(resid 107 and name o)
assign (resid 107 and name n)(resid 107 and name hn)(resid 95 and name o)
```

Table 3.7: Backbone dihedral angle restraint table. Generated using TALOS+ for Phi (φ) and Psi (ψ) angles. Information from only well predicted residues is used. Rests are commented out.

```
! talos_pred2xplor.nawk Table !   Date: Fri Jan 25 15:48:39 EST 2013
!   Path: /Users/ninadmw/StructureCalc/MBD4/TALOS+_10bp
!
! G78 Dihedrals

! S79 Dihedrals
!assign (resid  78 and name c)  (resid  79 and name n)
!       (resid  79 and name ca) (resid  79 and name c)  1.0  -72.4  84.4 2      ! S79 Talos Phi
!assign (resid  79 and name n)  (resid  79 and name ca)
!       (resid  79 and name c) (resid  80 and name n)  1.0  103.7  58.0 2      ! S79 Talos Psi

! T80 Dihedrals
!assign (resid  79 and name c)  (resid  80 and name n)
!       (resid  80 and name ca) (resid  80 and name c)  1.0  -95.3  42.6 2      ! T80 Talos Phi
!assign (resid  80 and name n)  (resid  80 and name ca)
!       (resid  80 and name c) (resid  81 and name n)  1.0   -9.2  28.1 2      ! T80 Talos Psi

! E81 Dihedrals
!assign (resid  80 and name c)  (resid  81 and name n)
!       (resid  81 and name ca) (resid  81 and name c)  1.0  -73.7  73.5 2      ! E81 Talos Phi
!assign (resid  81 and name n)  (resid  81 and name ca)
!       (resid  81 and name c) (resid  82 and name n)  1.0  135.8  23.3 2      ! E81 Talos Psi

! C82 Dihedrals
!assign (resid  81 and name c)  (resid  82 and name n)
!       (resid  82 and name ca) (resid  82 and name c)  1.0  -82.5  57.5 2      ! C82 Talos Phi
!assign (resid  82 and name n)  (resid  82 and name ca)
!       (resid  82 and name c) (resid  83 and name n)  1.0  117.7  22.6 2      ! C82 Talos Psi

! R83 Dihedrals
!assign (resid  82 and name c)  (resid  83 and name n)
!       (resid  83 and name ca) (resid  83 and name c)  1.0  -74.5  69.7 2      ! R83 Talos Phi
!assign (resid  83 and name n)  (resid  83 and name ca)
!       (resid  83 and name c) (resid  84 and name n)  1.0  126.5  40.1 2      ! R83 Talos Psi

! K84 Dihedrals
!assign (resid  83 and name c)  (resid  84 and name n)
!       (resid  84 and name ca) (resid  84 and name c)  1.0  -87.6  76.5 2      ! K84 Talos Phi
!assign (resid  84 and name n)  (resid  84 and name ca)
!       (resid  84 and name c) (resid  85 and name n)  1.0  130.0  44.1 2      ! K84 Talos Psi

! S85 Dihedrals
!assign (resid  84 and name c)  (resid  85 and name n)
!       (resid  85 and name ca) (resid  85 and name c)  1.0  -79.0  57.3 2      ! S85 Talos Phi
!assign (resid  85 and name n)  (resid  85 and name ca)
!       (resid  85 and name c) (resid  86 and name n)  1.0  131.5  22.1 2      ! S85 Talos Psi

! V86 Dihedrals
assign (resid  85 and name c)  (resid  86 and name n)
       (resid  86 and name ca) (resid  86 and name c)  1.0  -89.8  51.2 2      ! V86 Talos Phi
assign (resid  86 and name n)  (resid  86 and name ca)
       (resid  86 and name c) (resid  87 and name n)  1.0  137.9  36.5 2      ! V86 Talos Psi

! P87 Dihedrals
assign (resid  86 and name c)  (resid  87 and name n)
       (resid  87 and name ca) (resid  87 and name c)  1.0  -72.0  30.0 2      ! P87 Talos Phi
assign (resid  87 and name n)  (resid  87 and name ca)
       (resid  87 and name c) (resid  88 and name n)  1.0  156.2  20.5 2      ! P87 Talos Psi

! C88 Dihedrals
assign (resid  87 and name c)  (resid  88 and name n)
       (resid  88 and name ca) (resid  88 and name c)  1.0  -70.0  32.1 2      ! C88 Talos Phi
assign (resid  88 and name n)  (resid  88 and name ca)
       (resid  88 and name c) (resid  89 and name n)  1.0  131.8  20.0 2      ! C88 Talos Psi

! G89 Dihedrals
```

Table 3.7 continued…

```
!assign (resid  88 and name c)  (resid  89 and name n)
!       (resid  89 and name ca) (resid  89 and name c)  1.0   89.7  30.0 2      ! G89 Talos Phi
!assign (resid  89 and name n)  (resid  89 and name ca)
!       (resid  89 and name c) (resid  90 and name n)  1.0   -9.7  26.8 2      ! G89 Talos Psi

! W90 Dihedrals
!assign (resid  89 and name c)  (resid  90 and name n)
!       (resid  90 and name ca) (resid  90 and name c)  1.0  -80.8  49.4 2      ! W90 Talos Phi
!assign (resid  90 and name n)  (resid  90 and name ca)
!       (resid  90 and name c) (resid  91 and name n)  1.0  132.6  60.0 2      ! W90 Talos Psi

! E91 Dihedrals
assign (resid  90 and name c)  (resid  91 and name n)
       (resid  91 and name ca) (resid  91 and name c)  1.0 -129.2  30.0 2      ! E91 Talos Phi
assign (resid  91 and name n)  (resid  91 and name ca)
       (resid  91 and name c) (resid  92 and name n)  1.0  133.5  31.4 2      ! E91 Talos Psi

! R92 Dihedrals
assign (resid  91 and name c)  (resid  92 and name n)
       (resid  92 and name ca) (resid  92 and name c)  1.0 -108.5  32.2 2      ! R92 Talos Phi
assign (resid  92 and name n)  (resid  92 and name ca)
       (resid  92 and name c) (resid  93 and name n)  1.0  117.7  20.0 2      ! R92 Talos Psi

! V93 Dihedrals
assign (resid  92 and name c)  (resid  93 and name n)
       (resid  93 and name ca) (resid  93 and name c)  1.0 -121.0  30.0 2      ! V93 Talos Phi
assign (resid  93 and name n)  (resid  93 and name ca)
       (resid  93 and name c) (resid  94 and name n)  1.0  126.7  28.5 2      ! V93 Talos Psi

! V94 Dihedrals
assign (resid  93 and name c)  (resid  94 and name n)
       (resid  94 and name ca) (resid  94 and name c)  1.0 -118.1  30.0 2      ! V94 Talos Phi
assign (resid  94 and name n)  (resid  94 and name ca)
       (resid  94 and name c) (resid  95 and name n)  1.0  126.7  20.0 2      ! V94 Talos Psi

! K95 Dihedrals
assign (resid  94 and name c)  (resid  95 and name n)
       (resid  95 and name ca) (resid  95 and name c)  1.0 -122.5  38.0 2      ! K95 Talos Phi
assign (resid  95 and name n)  (resid  95 and name ca)
       (resid  95 and name c) (resid  96 and name n)  1.0  126.3  26.7 2      ! K95 Talos Psi

! Q96 Dihedrals
assign (resid  95 and name c)  (resid  96 and name n)
       (resid  96 and name ca) (resid  96 and name c)  1.0 -105.1  39.3 2      ! Q96 Talos Phi
assign (resid  96 and name n)  (resid  96 and name ca)
       (resid  96 and name c) (resid  97 and name n)  1.0  116.9  25.8 2      ! Q96 Talos Psi

! R97 Dihedrals
!assign (resid  96 and name c)  (resid  97 and name n)
!       (resid  97 and name ca) (resid  97 and name c)  1.0  -81.3  54.1 2      ! R97 Talos Phi
!assign (resid  97 and name n)  (resid  97 and name ca)
!       (resid  97 and name c) (resid  98 and name n)  1.0  126.6  40.9 2      ! R97 Talos Psi
```

INCOMPLETE

Table 3.8: Side-chain dihedral angle (CHI (χ)) restraint table. Generated for χ1 angle for F, W, Y and H (highlighted in red).

```
!Xplor Dihed Table from make_chi_xplr.nawk
! USER  ninadmw
! DATE  Sun Jan 27 13:48:26 EST 2013
! PATH  /Users/ninadmw/StructureCalc/MBD4/CONSTRAINTS/DIHED


! S79 Chi-1
!assign (resid  79 and name n  ) (resid  79 and name ca )
!       (resid  79 and name cb ) (resid  79 and name og )  1.0    0.0   20.0 2

! T80 Chi-1
!assign (resid  80 and name n  ) (resid  80 and name ca )
!       (resid  80 and name cb ) (resid  80 and name og1)  1.0    0.0   20.0 2

! E81 Chi-1
!assign (resid  81 and name n  ) (resid  81 and name ca )
!       (resid  81 and name cb ) (resid  81 and name cg )  1.0    0.0   20.0 2

! C82 Chi-1
!assign (resid  82 and name n  ) (resid  82 and name ca )
!       (resid  82 and name cb ) (resid  82 and name sg )  1.0    0.0   20.0 2

! R83 Chi-1
!assign (resid  83 and name n  ) (resid  83 and name ca )
!       (resid  83 and name cb ) (resid  83 and name cg )  1.0    0.0   20.0 2
! R83 Chi-2
!assign (resid  83 and name ca ) (resid  83 and name cb )
!       (resid  83 and name cg ) (resid  83 and name cd )  1.0    0.0   30.0 2

! K84 Chi-1
!assign (resid  84 and name n  ) (resid  84 and name ca )
!       (resid  84 and name cb ) (resid  84 and name cg )  1.0    0.0   20.0 2
! K84 Chi-2
!assign (resid  84 and name ca ) (resid  84 and name cb )
!       (resid  84 and name cg ) (resid  84 and name cd )  1.0    0.0   30.0 2

! S85 Chi-1
!assign (resid  85 and name n  ) (resid  85 and name ca )
!       (resid  85 and name cb ) (resid  85 and name og )  1.0    0.0   20.0 2

! V86 Chi-1
!assign (resid  86 and name n  ) (resid  86 and name ca )
!       (resid  86 and name cb ) (resid  86 and name cg1)  1.0    0.0   20.0 2

! P87 Chi-1
!assign (resid  87 and name n  ) (resid  87 and name ca )
!       (resid  87 and name cb ) (resid  87 and name cg )  1.0    0.0   20.0 2

! C88 Chi-1
!assign (resid  88 and name n  ) (resid  88 and name ca )
!       (resid  88 and name cb ) (resid  88 and name sg )  1.0    0.0   20.0 2


! W90 Chi-1
!assign (resid  90 and name n  ) (resid  90 and name ca )
!       (resid  90 and name cb ) (resid  90 and name cg )  1.0    -60.0   20.0 2
```

Table 3.8 continued…

```
! E91 Chi-1
!assign (resid  91 and name n  ) (resid  91 and name ca )
!       (resid  91 and name cb ) (resid  91 and name cg )  1.0    0.0   20.0 2

! R92 Chi-1
!assign (resid  92 and name n  ) (resid  92 and name ca )
!       (resid  92 and name cb ) (resid  92 and name cg )  1.0    0.0   20.0 2
! R92 Chi-2
!assign (resid  92 and name ca ) (resid  92 and name cb )
!       (resid  92 and name cg ) (resid  92 and name cd )  1.0    0.0   30.0 2

! V93 Chi-1
!assign (resid  93 and name n  ) (resid  93 and name ca )
!       (resid  93 and name cb ) (resid  93 and name cg1)  1.0    0.0   20.0 2

! V94 Chi-1
!assign (resid  94 and name n  ) (resid  94 and name ca )
!       (resid  94 and name cb ) (resid  94 and name cg1)  1.0    0.0   20.0 2

! K95 Chi-1
!assign (resid  95 and name n  ) (resid  95 and name ca )
!       (resid  95 and name cb ) (resid  95 and name cg )  1.0    0.0   20.0 2
! K95 Chi-2
!assign (resid  95 and name ca ) (resid  95 and name cb )
!       (resid  95 and name cg ) (resid  95 and name cd )  1.0    0.0   30.0 2

! Q96 Chi-1
!assign (resid  96 and name n  ) (resid  96 and name ca )
!       (resid  96 and name cb ) (resid  96 and name cg )  1.0    0.0   20.0 2

! R97 Chi-1
!assign (resid  97 and name n  ) (resid  97 and name ca )
!       (resid  97 and name cb ) (resid  97 and name cg )  1.0    0.0   20.0 2
! R97 Chi-2
!assign (resid  97 and name ca ) (resid  97 and name cb )
!       (resid  97 and name cg ) (resid  97 and name cd )  1.0    0.0   30.0 2

! L98 Chi-1
!assign (resid  98 and name n  ) (resid  98 and name ca )
!       (resid  98 and name cb ) (resid  98 and name cg )  1.0    0.0   20.0 2
! L98 Chi-2
!assign (resid  98 and name ca ) (resid  98 and name cb )
!       (resid  98 and name cg ) (resid  98 and name cd1)  1.0    0.0   30.0 2

! F99 Chi-1
!assign (resid  99 and name n  ) (resid  99 and name ca )
!       (resid  99 and name cb ) (resid  99 and name cg )  1.0   -60.0   20.0 2
! F99 Chi-2
!assign (resid  99 and name ca ) (resid  99 and name cb )
!       (resid  99 and name cg ) (resid  99 and name cd1)  1.0    0.0   30.0 2
```

INCOMPLETE

3.5.3 Inter-proton distances

Inter-proton distances are the most important NMR structural restraints for structure determination because they provide both long-range and short-range distance information. Inter-proton distances are usually measured with multi-dimensional NOESY (^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC) spectra where the Nuclear Overhausar Effect (NOE) provides correlation between pair of protons which are separated by less than 5Å. Short-range NOEs are useful for confirming secondary structure elements such as α-helix and β-sheet. Long-range NOEs provide crucial tertiary structure information. The intensity of the NOE cross peak is inversely proportional to the sixth power of the distance between the two protons. A weaker NOE cross peak can also result from chemical exchange or dynamics in proteins. Therefore, weaker peaks can also be seen between protons that are closer than 4Å, while, a stronger peak is indicative of two protons close in space. Within CCPNMR, NOE cross peaks were assigned and intensities of the peaks were recorded which were used to determine qualitative distance restraints to be used for the structure calculation. Cross peak intensities (heights) were grouped into four different categories: 1.8Å-2.5Å (strong), 1.8Å-3.8Å (medium), 1.8Å-5.0Å (weak) and 1.8Å – 6.0Å (very weak). NOE tables from ^{15}N and ^{13}C NOE experiments were merged, duplicates were removed and the format was adjusted for XPLOR calculations. The NOE table with more than 800 NOEs for MBD4 is shown in Table 3.9. Generally, ten NOEs per amino acid residue is enough for the structure determination.

Table 3.9: A combined NOE table. Both ^{13}C and ^{15}N NOE are demonstrated.

!^{15}N NOE

assi ((resid 81 and name HB#)) ((resid 82 and name HN)) 4.000 2.200 1.000

assi ((resid 82 and name HN)) ((resid 82 and name HA)) 4.000 2.200 1.000

assi ((resid 83 and name HN)) ((resid 82 and name HB#)) 4.000 2.200 1.000

assi ((resid 83 and name HN)) ((resid 82 and name HA)) 4.000 2.200 1.000

assi ((resid 83 and name HB#)) ((resid 83 and name HN)) 4.000 2.200 1.000

assi ((resid 83 and name HN)) ((resid 83 and name HG#)) 4.000 2.200 1.000

assi ((resid 83 and name HB#)) ((resid 85 and name HN)) 4.000 2.200 1.000

assi ((resid 85 and name HA)) ((resid 85 and name HN)) 3.500 1.700 0.300

assi ((resid 85 and name HN)) ((resid 85 and name HB#)) 4.000 2.200 1.000

assi ((resid 85 and name HB#)) ((resid 148 and name HN)) 4.000 2.200 1.000

assi ((resid 86 and name HN)) ((resid 85 and name HB#)) 4.000 2.200 1.000

assi ((resid 86 and name HN)) ((resid 85 and name HA)) 3.500 1.700 0.300

assi (resid 86 and name HG*) ((resid 86 and name HN)) 3.500 1.700 0.300

assi ((resid 86 and name HN)) ((resid 86 and name HA)) 4.000 2.200 1.000

INCOMPLETE

!^{13}C NOE

assi (resid 80 and name HG2#) ((resid 80 and name HB)) 4.000 2.200 1.000

assi (resid 80 and name HG2#) ((resid 80 and name HA)) 4.000 2.200 1.000

assi ((resid 80 and name HA)) ((resid 81 and name HN)) 4.000 2.200 1.000

assi ((resid 82 and name HB#)) ((resid 82 and name HA)) 4.000 2.200 1.000

assi ((resid 83 and name HN)) ((resid 82 and name HA)) 4.000 2.200 1.000

Table 3.9 continued…

```
assi   (( resid 84   and name  HN  ))  (( resid 83   and name  HA  ))     4.000     2.200     1.000
assi   (( resid 85   and name  HA  ))  (( resid 85   and name  HB# ))     3.500     1.700     0.300
assi   (( resid 85   and name  HB# ))  (( resid 148  and name  HN  ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 86   and name  HB  ))     3.500     1.700     0.300
assi   (  resid 86   and name  HG*)  (( resid 86   and name  HA  ))     4.000     2.200     1.000
assi   (( resid 86   and name  HN  ))  (( resid 86   and name  HB  ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 86   and name  HB  ))     3.500     1.700     0.300
assi   (( resid 86   and name  HA  ))  (( resid 86   and name  HB  ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 86   and name  HA  ))     4.000     2.200     1.000
assi   (( resid 86   and name  HA  ))  (( resid 87   and name  HD# ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 87   and name  HD# ))     4.000     2.200     1.000
assi   (( resid 86   and name  HA  ))  (( resid 87   and name  HD# ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 87   and name  HD# ))     4.000     2.200     1.000
assi   (( resid 86   and name  HA  ))  (( resid 87   and name  HD# ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 90   and name  HB# ))     4.000     2.200     1.000
assi   (  resid 86   and name  HG*)  (( resid 91   and name  HA  ))     4.000     2.200     1.000
assi   (( resid 88   and name  HN  ))  (( resid 82   and name  HB# ))     4.000     2.200     1.000
assi   (( resid 88   and name  HA  ))  (( resid 82   and name  HB# ))     4.000     2.200     1.000
assi   (( resid 88   and name  HA  ))  (( resid 88   and name  HN  ))     4.000     2.200     1.000
assi   (( resid 89   and name  HN  ))  (( resid 88   and name  HA  ))     4.000     2.200     1.000
assi   (( resid 89   and name  HA# ))  (( resid 113  and name  HG# ))     4.000     2.200     1.000
```

INCOMPLETE

3.6 Structure calculation using XPLOR-NIH

Xplor-NIH is a package used for NMR structure calculations using experimental NMR data constraints and known geometric data. It minimizes target function that depends on NMR constraints and covalent geometry using molecular dynamics in Cartesian and torsion angle space. Restrained molecular dynamics (rMD) often involves simulated annealing method to explore the conformational space efficiently. In addition to the theoretical potential energy function, penalty functions are added that select against conformations that do not agree with experimental data. A conformation with the lowest intrinsic energy is determined that is also consistent with the experimental data. NMR structural restraints, NOE, H-bond, torsion angles in the correct format are used as input parameters. Initial random coil coordinates and geometric protein structure files (PSF), which contain information on protein sequence, molecular bonds, bond angles etcetera, are needed. The PDB file of MeCP2 was mutated and used as initial structure file for MBD4. The 3D structures are simulated at higher temperature (2000K) using molecular dynamics taking forces of the covalent geometry restraints (bond length, bond angles, dihedral angles, improper torsions and van der Waals) and experimental restraints (NOE and J coupling); NOEs are given stronger force constants than torsion angle restraints. The system is then cooled down to 300K with temperature steps of 50K. Energy minimizations in 100 such cycles are simulated. At high temperatures, the system is able to occupy high energy regions of conformational space and therefore can pass over high energy barriers. This avoids local minima and global energy minimum is most likely achieved. In such a way, ensembles of low energy structures consistent with the input data are generated. For a high quality structure, more than 10 distance restraints per residue are used. NOE and torsion angle violations are filtered and

removed if needed. More NOE cross-peaks can be assigned based on the initial structure and used for refining the structure further. Refined 3D structure has ensemble of structures with minimum violations of input restraints and minimum root-mean-square deviation between members of the ensemble. A solution structure ensemble of MBD4 was calculated and for residues 88 to 140 (avoiding flexible terminal regions) the backbone RMSD of 0.95Å and overall RMSD of 1.42Å was seen (Figure 2.4). More iterative assignments of NOEs will result in further refinement of the protein structure.

3.7 References

1. F. Delaglio *et al.*, NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* **6**, 277-293 (1995).

2. W. F. Vranken *et al.*, The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins.* **59**, 687-696 (2005).

3. Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR.* **44**, 213-223 (2009).

4. C. D. Schwieters, J. J. Kuszewski, N. Tjandra, G. M. Clore, The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65-73 (2003).

5. M. Fatemi, P. A. Wade, MBD family proteins: reading the epigenetic code. *J. Cell. Sci.* **119**, 3033-3037 (2006).

4.0 Summary

The studies presented in this thesis *'Structural basis of DNA binding complexes'* encompass two major aspects of DNA methylation as an epigenetic mark; one that revolves around a crucial coiled-coil interaction central to MBD2-mediated silencing and the other that describes dynamic behavior of MBD4 in recognizing DNA methylation mark.

Chapter one: '*Unique features of the anti-parallel, heterodimeric coiled-coil interaction between methyl-cytosine binding domain 2 (MBD2) homologues and p66α dictate high affinity binding'* describes the coiled-coil interaction between two proteins (MBD2 and p66α) that is critical for DNA methylation dependent gene silencing. Here, we characterized this unique interaction and determined the parameters that drive heterodimeric specificity and high affinity binding. By comparing MBD2 homologues and their mutation products, we showed a direct correlation between helical content of the coiled-coil domains in isolation and binding affinity for p66α. Importance of maintaining specific ionic interactions and having complementary electrostatic surface potentials was demonstrated. This study would help our understanding of coiled-coil interactions, which form a common interaction motif in eukaryotic proteins. Possessing the knowledge of affinity determinants will facilitate development of small peptide based drugs which target such interactions in nature. Additional studies can be performed for improving the potency by the cell penetrating peptide by making it more helical, cell-penetrable and stable against proteolytic degradation.

Chapter two: 'Dynamic behavior *of MBD4 in methylated DNA recognition'* included studies on MBD4 as a methyl DNA binding protein and its preference for DNA methylation mark. Here, we demonstrated that MBD4 can have diverse substrate binding abilities but it

prefers mCpG and TpG DNA over hydroxymethylated and unmethylated DNA. Further, with the help of chemical exchange studies we showed that MBD4 exchanges slowly on two inter-molecular DNA binding sites; but on two intra-molecular sites, MBD4 exhibits fast exchange. This suggested a facilitated diffusion rather than free 3D diffusion mechanism. Introducing more bases or a defect between two intra-molecular sites did not affect the fast exchange rate indicating that MBD4 may not always stay in contact with the DNA. This suggested that the hopping mechanism of facilitated diffusion was utilized by MBD4. This is the first time we demonstrated how MBD domains localize themselves on their cognate DNA binding sites. This study can be expanded further to analyze the effects of degree of separation of specific DNA binding sites and/or more prominent DNA defects, thereby providing new insights into DNA-protein recognition.

Using NMR spectroscopy to elucidate a 3D structure is the focus of chapter three: *'Solving the solution structure of MBD domain of MBD4 on methylated DNA by NMR'.* This chapter should help with basic workflow used in NMR structure determination.

I hope this dissertation project summarizing my work with DNA binding complexes provide some useful insights into understanding the complex field of epigenetic regulation.

VITA

Ninad M Walavalkar was born in Dombivali, India on December 27, 1984, to the proud parents Maheshwar and Uma Walavalkar. Ninad completed his high school education from DNC multipurpose high school and Junior college education in sciences from R.V. Nerurkar Junior college. He earned a Bachelor of science degree with honors in the field of Life sciences and biochemistry from St. Xavier's college, Mumbai, India. In 2007 Ninad earned a masters degree in the field of Life sciences with biological macromolecule as specialization from University of Mumbai, India. Then he joined integrative life sciences PhD program at Virginia Commonwealth University and started his thesis work under the guidance of Dr. David Williams in Massey Cancer Center. Ninad can be contacted at 401 College Street, rm 111F, Richmond VA, 23298.

Publications

Walavalkar NM, Gordon N, Williams DC Jr.; *Unique features of the anti-parallel, heterodimeric coiled-coil interaction between methyl-cytosine binding domain 2 (MBD2) homologues and GATA zinc finger domain containing 2A (GATAD2A/p66α)*. J Biol Chem. 2013 Feb 1;288(5):3419-27

Gnanapragasam MN, Scarsdale JN, Amaya ML, Webb HD, Desai MA, **Walavalkar NM**, Wang SZ, Zu Zhu S, Ginder GD, Williams DC Jr.; *p66Alpha-MBD2 coiled-coil interaction and recruitment of Mi-2 are critical for globin gene silencing by the MBD2-NuRD complex.* Proc Natl Acad Sci U S A. 2011 May 3;108(18):7487-92