The Structure of the Methyl-Binding Domain of MBD2, and Studies of the MBD:DNA interactions

A thesis presented for the degree of

Doctor of Philosophy

at the University of Edinburgh

Kirsten L. Lillie

Structural Biology Group

Institute of Structural and Molecular Biology

University of Edinburgh

June 2006

Abstract

DNA methylation is a powerful epigenetic control mechanism, essential for the regulation of transcriptional processes. In mammals, methylation occurs mainly at CpG dinucleotides and the modified sequence is recognised by members of the methylated DNA-binding domain (MBD) family. Methylated DNA binding domain 2 (MBD2) is a 44 kDa protein with the ability to distinguish between methylated and unmethylated DNA. It is associated with histone deacetylases and nucleosome remodelling complexes *in vivo*, and functions as a transcriptional repressor. MBD2 expression levels correlate with certain types of cancer and MBD2 may contribute to tumour development, by inappropriate silencing of tumour suppressor genes. MBD2 is a link between DNA methylation, chromatin structure and certain disease mechanisms. Understanding how methylated DNA is recognised by MBD2, will help establish its specific role as a methylation interpreter.

This thesis describes the structure determination of the 72 residue MBD from mouse MBD2, using solution NMR spectroscopy. The MBD has a characteristic wedge-shaped fold, with one face made up of β-sheets, and the other of an α-helix plus a hairpin loop. Loop L1, between two long beta-strands, is much more rigid in this structure compared to the equivalent region in other members of the MBD family. It appears to be a pre-formed DNA binding loop and overlays well with the structure of the MBD (from MBD1):DNA complex, solved by another group. NMR was subsequently used to follow the titration of the MBD from MBD2, with a 12 base-pair DNA oligomer, symmetrically methylated at a central CpG pair. Large chemical shift changes in the ^{15}N-HSQC spectrum were observed for L1 loop residues, and for an arginine residue at the base of the α-helix. This corresponds to the DNA binding interface in the MBD1 MBD:DNA complex structure.

Surface plasmon resonance was used to compare the affinities of three different MBDs for DNA oligomers in various methylation states. The MBD2 MBD showed weaker binding to symmetrically methylated DNA, when compared to MeCP2 and MBD4 MBDs. In general, there was no difference in affinity for any of the methylated oligomers, and no construct bound unmethylated DNA. There is some evidence that oligomer length affects binding, although this was not studied in detail.

Dedication

To Brian,

One down and many more to go. For you that is - not for me!

Acknowledgements

Thankyou to the following people:

My unofficial first supervisor, Brian Smith, for just about everything.

Dušan Uhrín and Juraj Bella for fixing pulse programs, endless help setting up NMR experiments and general chat about Slovakia and Scottish holidays.

My official first supervisor, Paul Barlow, for some keen observations, useful insights and a good reference; Everyone else in the Barlow and Uhrín groups: Thanks for your advice, whether I asked for it or not!

Sharon Kelly for training, and help with SPR; Everyone in Brian's group in Glasgow, particularly Nichola and Krystyna.

All of the Swann level 3 crowd, in particular Russell Hamilton (for making everything work when it was beyond me), and Andy Hinton (thanks for your magic perl script and magic tractor joke).

Adrian Bird and everyone in the Bird lab, particularly Donald, Skiri, Jennifer and both Robs.

Finally, I'd like to thank the huge support team of my parents, Russell, Ran, and everyone at Viewforth church for looking after me: physically, emotionally and spiritually.

This work was funded by the Medical Research Council.

Abbreviations

Å	Angstroms
APC^{Min}	Adenomatous polyposis coli germ-line mutation
AUC	Analytical ultracentrifugation
BLAST	Basic local alignment search tool
CM5	Carboxymethyldextran
CNS	Crystallography and NMR system
COSY	Correlated spectroscopy
DDM1	Decrease in DNA methylation 1
DNMT	DNA methyltransferase
DTT	Dithiothreitol
EDC	1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide
EDTA	Ethylene-diamine-tetraacetic-acid
EMSA	Electromobility shift assay
EST	Expressed sequence tag
FID	Free induction decay
FPLC	Fast protein liquid chromatography
FT	Fourier transformation
GSTP1	π-class glutathione S-transferase
HAT	Histone acetyl transferase
HDAC	Histone deacetylase
HEPES	N-2-hydroxyethyl-piperazine-N-2-ethanesulfonic acid
HKMT	Histone lysine (K) methyltransferase
HSQC	Heteronuclear single quantum correlation
ISWI	Imitation switch
LB	Luria-Bertani medium
MBD	Methyl-CpG-binding domain
MBD2	Methylated DNA binding domain protein 2
MD	Molecular dynamics
MES	2-(N-morpholino)-ethanesulphonic acid
MBDin	MBD2 interacting protein
MeCP2	Methyl-CpG-binding protein 2
MEM	Maximum entropy method
MT	Methyl transferase
MTA1	Metastases associated protein 1
MIZF	MBD2-interacting zinc finger protein
meCpG	Methylated-CpG
NaPi	Sodium phosphate
NCP	Nuclear core particle
NEB	New England Biolabs
NER	Nucleotide excision repair
NHS	N-hydroxysuccinimide
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NOESY	NOE spectroscopy

Abbreviations (continued)

NuRD	Nucleosome remodelling/HDAC complex
Ni-NTA	Nickel-nitrilotriacetic acid
PAH	Paired amphipathic helix
PDB	Protein Data Bank
PRMT	Protein arginine (R) methyltransferases
PEG	Polyethylene glycol
R-CEMSA	Reverse capillary electrophoretic mobiliy shift assay
RDC	Residual dipolar coupling
RF	Radiofrequency
RU	Resonance units
SA[a]	Streptavidin
SA[b]	Simmulated annealing
SDS-PAGE	Sodium dodecylsulphate-polyacrylamide gel electrophoresis
SNR	Signal-to-noise ratio
SPR	Surface plasmon resonance
SWI/SNF	Switch/sucrose non-fermenting
TIR	Total internal reflection
TOCSY	Total correlated spectroscopy
TRD	Transcriptional repression domain
TSB	Trypticase soy broth
XPA	Xeroderma Pigmentosum group A protein
XAB1	XPA binding protein 1

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Chromatin

In eukaryotic organisms, chromosomal DNA is packaged into chromatin. Chromatin is a complex array of helical DNA and many associated proteins, but the essential building blocks are repeating nucleoprotein units, called nucleosomes. DNA is wound around proteins called histones, which are small, basic proteins with a higher than average lysine and arginine content. Individual histones have a characteristic 'histone fold' secondary structure, with protruding, charged N-terminal tails. Two each, of four different histones: H2A, H2B, H3 and H4 form an octamer core particle, while histone H1 is known as the linker histone [1]. In 1997, the first high-resolution structure of the nucleosome core particle was determined: 146bp of DNA wrapped around the histone octamer [2](figure 1.1).

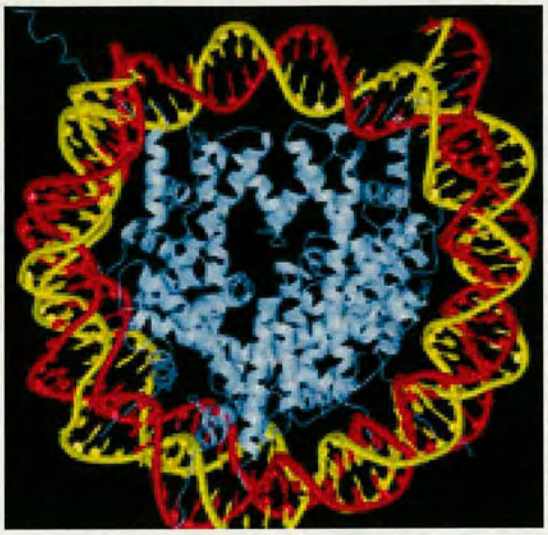

Figure 1.1: The structure of the nucleosome core particle solved to a resolution of 2.8A. 146bp of DNA (shown in red and yellow) is wrapped around the histone octamer (shown in light blue). Taken from [3].

1.1.1 Dynamic chromatin

Tightly-packed DNA, however, must also be made accessible for the nuclear processes which are essential for life, such as replication and transcription. So how does the cell manage to effectively package several centimetres of DNA into a a few μm, while still retaining its functionality? Chromatin is dynamic and consists of more and less densely packed regions, called heterochromatin and euchromatin, respectively [4]. Euchromatin is commonly thought of as the "open" state, where the DNA sequence is available to be read or recognised or modified. Heterochromatin can be divided into two subclasses: Constitutive heterochromatin is found around centromeres and telomeres of all somatic cell types, packaging DNA with a low gene content. That is to say, there are similar regions in all cells characterised by densely packed chromatin, which makes highly repetitive, non-coding DNA inaccessible.

Facultative heterochromatin is cell-type dependent, packaging genes that are not required for the function of a particular cell or tissue type. This inactive state will be passed down the cell lineage.

Over the years, researchers have established what the features of the different types of chromatin are, especially with respect to epigenetic markers. Epigenetics refers to heritable changes in phenotype, that are not coded for in the DNA sequence, but may still be passed on to successive generations by another mechanism [5]. Examples of epigenetic heritability include maintenance of DNA methylation states after replication and preservation of chromatin architecture after cell division.

1.1.2 Histone modification

Although histones were once thought to have a purely structural role, they are now known to play an active part in regulating biological processes. The N-terminal tails of histone molecules stick outwards from the surface of the nucleosomes and are the sites of many post-translational modifications including acetylation, methylation, phosphorylation and many others (for review see [6]). The distinct patterns of such modifications, led in 2001 to the "histone code" hypothesis, which suggests that signals encoded on histone tails are read by other proteins to regulate transcriptional states [7].

1.1.3 Histone acetylation and chromatin remodelling

The most extensively-studied histone tail modification, is acetylation. Histone acetyl transferases (HATs) and histone deacetylases (HDACs) are the enzymes

responsible for catalysing the reversible reaction. Generally speaking, acetylation is associated with a more open chromatin state and transcriptional activation. Conversely, HDACs are commonly linked to gene silencing and a closed, heterochromatic structure.

Acetylation may have a direct influence on chromatin structure by neutralising charged residues in the histone tails, thus weakening the DNA-nucleosome interaction [7]. However, it has also been shown that acetylated histones can recruit transcriptional activators [8], and that HATs are often part of large protein complexes, targeted to promoters [9, 10].

HDACs are also commonly found as part of multi-subunit complexes, but these are transcriptionally repressive. A number of different pathways exist, all featuring the recruitment of HDACs to DNA-binding proteins by intermediate factors [11]. The Mi2/NuRD (Nucleosome Remodelling/HDAC) and Sin3/HDAC complexes share a common core comprising HDAC1/2 and histone binding proteins RbAP46/48 [12]. These particular complexes will be described in more detail later. Other complexes lacking RbAP46/48 [13], or containing only one HDAC unit have also been identified [14], so there seem to be a wealth of possiblities for specific deacetylation-signalled events. Some HDAC complexes even display specificity for certain histone tails [15].

Chromatin remodelling complexes function in two main ways: either the location of the nucleosome on a particular DNA sequence, or the state of the nucleosome itself, may be altered [16]. The histone octamer can be slid along the DNA and histone variants (eg. H2A.Z) may be incorporated into the nucleosome to change its structure and accessibility. The most important group of enzymes

involved in remodeling is the Snf2-like, ATPase/helicase family, of which the best characterised subfamilies are SWI/SNF (Switch/Sucrose non-fermenting) and ISWI (imitation switch). Both have been shown to activate and repress transcription, as part of larger remodeling complexes such as α/BAF and NURF [17, 18]. Mi2/NuRD is part of the CHD1 subfamily. It is unable to alter nucleosomal spacing [19], but as part of the MeCP1 complex (together with MBD2 and p66/p88) can preferentially remodel and deacetylate methylated nucleosomes [20]. Thus, nucleosome remodeling and histone deacetylation are linked to gene silencing by methylation.

1.2 DNA methylation

DNA methylation occurs at the C5-position of cytosine and functions as a powerful epigenetic control mechanism. Methylation at CpG dinucleotides is the most abundant covalent modification of the vertebrate genome, but there is a lower than expected frequency of CpGs in mammals. This is apart from in small regions of about 1 kilo base-pairs (bp), called CpG islands. These islands have a high proportion (50%) of the total unmethylated genomic CpGs, and are often located in or around a gene's promoter or first exon. A strong correlation has been found between methylation of CpG islands and repression of downstream genes, implicating methylation in the maintenance of silent chromatin [21].

What is the mechanism by which CpG methylation inhibits transcription? Several hypotheses have been proposed, including direct and indirect mechanisms. It was thought that cytosine methylation could exclude the binding of

transcription factors and so directly silence genes, which is true in a limited number of cases. However, it is now generally accepted that meCpG recruits specific repressors to mediate its effects [1].

Studies in *Xenopus* oocytes have shown that repression of a methylated template coincides with nucleosome array formation [22] demonstrating the link between DNA methylation and chromatin remodelling. Chen *et al.* showed that in plants, the HDAC inhibitor, Trichostatin A, can reactivate silent rRNA genes, in a similar manner to 5-aza-dC, the DNA methylation inhibitor [23]. Thus, DNA methylation and histone acetylation appear to act synergistically.

1.2.1 The DNMT family

Methylation patterns are established and maintained by DNA methyltransferases (DNMTs). In eukaryotes, five members of the DNMT family are known.

DNMT1 is the most abundant methyltransferase in somatic cells and was the first to be identified [24]. It has a preference for hemimethylated DNA over unmethylated DNA, and is targeted to replication foci [25]. Therefore, it is thought mainly to be responsible for maintaining methylation patterns after DNA replication [26, 27]. In contrast, DNMT3a and 3b have an equal preference for unmethylated and hemimethylated DNA, and are known as *de novo* methyltransferases [28]. However, the categorisation of the DNMTs into *de novo* and maintenance groups is becoming less rigid. Overexpression of DNMT1 in cancer cells results in endogenous CpG island methylation [29], and disrupting the DNMT1 gene in somatic cells results in only a 20% decrease in genomic methylation levels [30]. Recently, a role for DNMT1 as a positive effector of *de novo* methylation has

been described [31].

In mice, targeted mutation of Dnmt1 has an embryonic lethal effect [32]; Dnmt3b knockouts are not viable and Dnmt3a knockouts die at around one month [33]; so the DNMT genes do appear to have some unique functions.

The DNMT2 protein was first thought to be catalytically inactive, but has since been found to methylate DNA at a very low level *in vitro* [34]. In addition, *Drosophila* and mouse Dnmt2 proteins are both able to methylate cotransfected plasmid DNA in transient transfection studies [35]. However, a biological niche for DNMT2 has not been found.

A protein with some homology to DNMT3a and 3b, is the most recent family member to be identified. DNMT3L seems to regulate the initiation of maternal methylation imprints [36], but is not catalytically active *in vitro* [37]. A study of the interplay within the mammalian DNMT family, has shown an interaction between DNMT3L and the *de novo* methylatransferases Dnmt3a/b [38], while the mouse Dnmt3L and Dnmt3a/b proteins are found to associate *in vitro* [39]. It has been suggested that Dnmt3L recruits an active methylase to maternal imprinting targets, and that it can also function as a HDAC1-dependent transcriptional repressor [40].

1.3 Introduction to the MBD family

Once the methylation patterns are in place, how are they recognised to produce a change in transcriptional state?

Methyl-CpG-binding Protein 1 (MeCP1) and MeCP2 were first described as

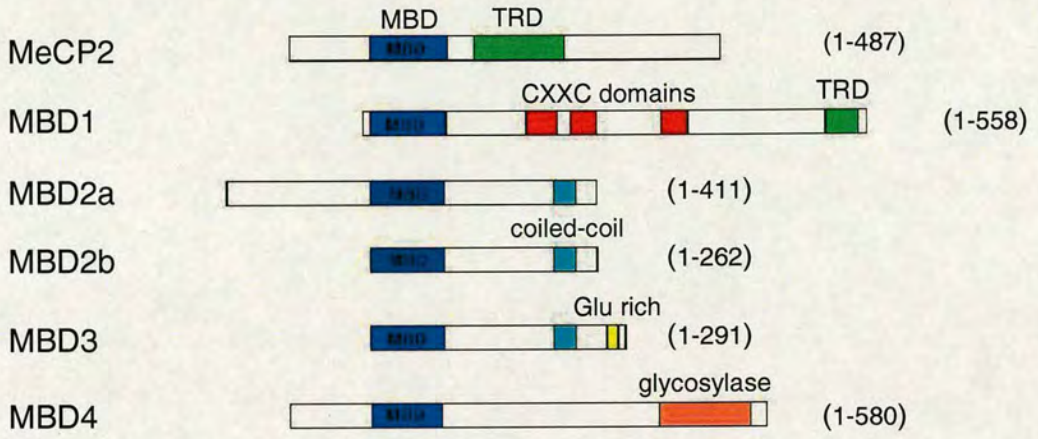

Figure 1.2: Schematic diagram of the five core members of the MBD family. Various domains are indicated: MBD in blue; TRD in green; putative coiled coil in cyan; Glutamate-rich region in yellow; CXXC domains in red; and the glycosylase unique to MBD4 in orange.

activities that had a specific affinity for methylated DNA. MeCP1 was later shown to be a complex, while MeCP2 is a 55kDa protein. A deletion analysis of MeCP2 identified the minimal region necessary for methylated-CpG (meCpG) binding, the methyl-CpG-binding domain (MBD). The MBD is considered a novel DNA-binding motif and is about 70 residues long. In order to find other MBDs in mammals, the expressed sequence tag (EST) database was screened to yield novel proteins MBD2, 3 and 4. Together with MeCP2 and previously discovered MBD1, these five proteins make up the core MBD family [41].

The MBD is the only region of sequence similarity common to all family members, and they do not resemble each other outside this domain. The exception is MBD2 and MBD3, which are 93% similar and 72% identical over their entire length. Aligning the MBDs shows the similarity of MBD1-3 in particular, whilst MeCP2 and MBD4 occupy a separate subclass. In spite of the general differences, an intron is found at a conserved position in all the MBD genes. Also, several

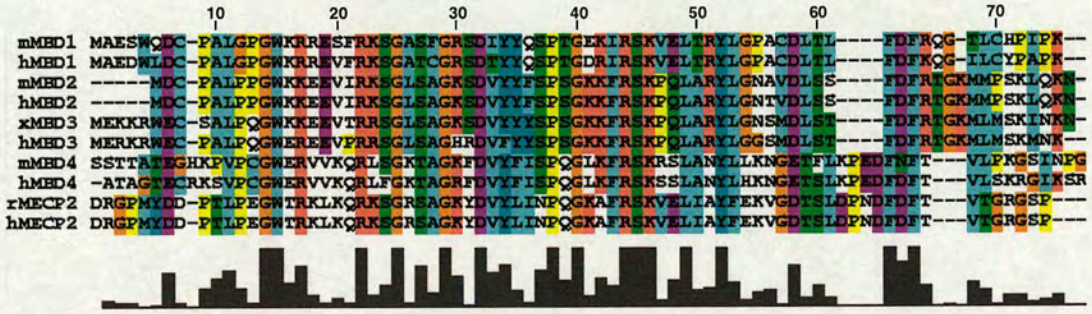

Figure 1.3: A multiple sequence alignment of the MBDs of various MBD family members, numbered as for mouse MBD1. Amino acids are coloured by residue type and conservation, which is also summarised by the black *quality curve*. The alignment was done with MUSCLE [42], and the plot created with ClustalX [43].

family members have a transcriptional repression domain (TRD), highlighting another common *in vivo* function. A brief description of each family member is given below.

1.3.1 MeCP2

MeCP2 has been very well characterised owing to its link with Rett syndrome [44]. It can bind one meCpG [45], distinguishing it from the MeCP1 complex, which requires 12 meCpGs for binding [46]. It has a large TRD and functions as a transcriptional repressor through its association with the Sin3a/HDAC complex [47, 48]. However, this is not a stable interaction and MeCP2 is likely to act through other repressive pathways [49]. A recent study has shown that MeCP2 displays some sequence selectivity in DNA binding, requiring AT bases adjacent to the meCpG [50].

1.3.2 MBD1

MBD1 is the largest member of the family (70kDa), and is known as a methylation-dependent transcriptional repressor, due to its C-terminal TRD. As well as the MBD, it has a variable number of CXXC motifs, depending on alternative splicing. One of these motifs has been shown to bind unmethylated CpG sites *in vivo*. Although it was originally identified as the MBD activity of MeCP1 [51], this has now been shown not to be the case [52]. MBD1 forms a stable complex with SETDB1, a histone H3:K9 methyltransferase, and is responsible for its recruitment to chromatin assembly factor CAF1 [53].

1.3.3 MBD2

MBD2 exists as two isoforms, produced by initiation at one of two methionine codons. MBD2a is 43.5kDa protein, characterised by repeated glycine and arginine residues at the N-terminus, which is probably unstructured. MBD2b (29.1kDa) lacks the first 140 amino acids of MBD2a, and starts at the MBD. MBD2 has an overlapping MBD and TRD, and is associated with HDAC1 in mammalian cells. It acts as a transcriptional repressor and is the true MBD component of the MeCP1 complex [52]. Reported demethylase activity for MBD2 [54] has not been reproduced in other research groups [55, 52].

1.3.4 MBD3

MBD3 is about 70% identical to MBD2. However, the mammalian protein differs at two amino acids within the MBD making it incapable of binding methylated DNA. In contrast, MBD3 from *Xenopus laevis* does bind meCpG. The precise

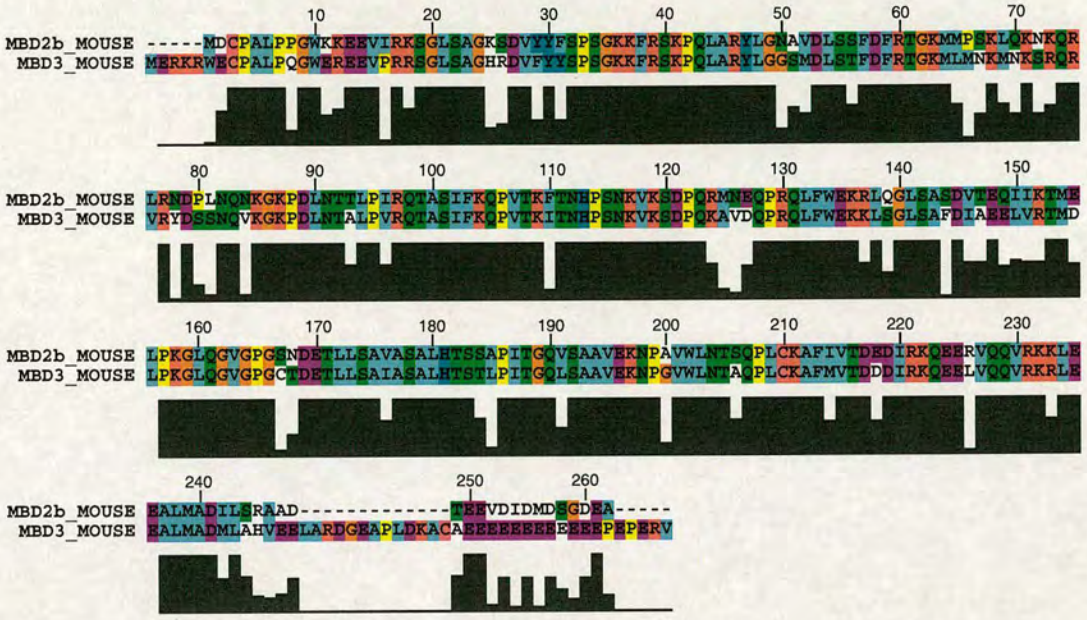

Figure 1.4: Alignment of the MBD2b and MBD3 proteins from mouse. The numbering is as for MBD2b and the key is the same as for figure 1.3.

nature of the differences and the implications for mCpG recognition will be discussed later. MBD3 has many splice forms, the most common of which is a 32kDa variant.

1.3.5 MBD4

MBD4 has a function distinct from the other family members. In addition to the MBD it has a second DNA binding domain at the C-terminus. The G:T mismatch-specific DNA glycosylase domain is responsible for the ability of MBD4 to efficiently recognise and remove T from a G:T mismatch, and U from a G:U mismatch [56]. MBD4 is similar in function to the thymine-DNA glycosylase (TDG) enzyme, despite a complete lack of sequence homology. The crystal structure of the glycosylase domain has been solved, and shows that it belongs to the helix-hairpin-helix DNA glycosylase superfamily [57]. MBD4's MBD prefers 5-methyl

CpG paired with TpG, compared to symmetrically methylated CpGs. However, the MBD does not influence the specificity of the glycosylase activity for G : T mismatches [58, 59]. As the G:T mismatch is the expected product of deaminating one side of meCpG, MBD4 has been designated a repair enzyme specific for methylated DNA. MBD4 has also been shown to interact with MLH1, a protein implicated in mismatch repair.

1.3.6 More mammalian MBDs

Methyl-CpG-Binding domains have also been found in 6 additional human proteins. A BLAST search of protein domain family databases (eg. Pfam, Smart) using the MBD of MeCP2, revealed two novel proteins and 4 known ones (including SETDB1). Most of these proteins have been linked to regulation of gene expression, but their ability to bind methylated CpGs is unknown [60].

1.3.7 Kaiso

Kaiso is a methyl-CpG binding protein with a difference. Rather than an MBD, it uses a zinc-finger motif to recognise its preferred sequence of two consecutive mCpGs [61]. It also s as a methylation-dependent transcriptional repressor.

1.4 MBDs in other organisms

1.4.1 MBDs in plants

In plants, epigenetic control is mediated by cytosine methylation at both CpG and CpNpG sequences. To date, 12 putative MBD proteins have been identified

in *Arabidopsis thaliana*, of which 3 have been shown to bind specifically to symmetrically methylated CpGs. These also happen to be those with MBD motifs most similar to the human forms. Of these, AtMBD7 is unique in that it contains two MBDs, and AtMBD6 is associated with HDAC activity [62]. Work by a second group has revealed subtle differences in the specificity of AtMBD5 and 6: While MBD5 could bind a symetrically or hemimethylated probe, MBD6 could only form a complex with symmetrically methylated DNA [63]. None of the proteins bound methylated CpNpG sequences in either study [62, 63]. Recent efforts have shown the interaction of *At*MBDs with the chromatin remodelling factor DDM1 (decrease in DNA methylation 1) [64]. It has been proposed that the DDM1-*At*MBD interaction may constitute a plant complex, similar in function to mammalian MeCP1.

1.4.2 Nematode MBDs

Little is known in general, about DNA methylation in nematodes, but the *C. elegans* genome is completely free of methylation. A dnmt2 gene was found in *P. pacificus*, and the protein product is 30-40% identical to DNMT2 from *D. melanogaster* and mouse. However, there is no evidence for DNA methylation in this species either, and dnmt2 morpholinos, or depletion of dnmt2 by RNAi, produced no discernable phenotype. An mbd-2 gene is present in *C. elegans*, *C. briggsae* and *P. pacificus*, albeit with different sequences and gene structure. All the nematode MBD2 sequences analysed lacked a typical MBD, which might be expected due to the lack of methylation in their genomes. However, MBD2 RNAi phenotypes are distinctive. The most severe of these included paralysis and early

death in *C. briggsae* [65]. This could indicate a role for MBD2 that is unrelated to DNA methylation.

1.4.3 MBD2/3 in insects

Drosophila melanogaster has no detectable 5mC in the adult genome, but embryonic DNA is methylated at very low levels [66]. A *Drosophila* protein homologous to the mammalian MBDs has been found, which is 56% similar to MBD2 and 73% similar to MBD3. It exists in two forms: dMBD2/3 and a shorter splice variant, dMBD2/3Δ. cDNA for MBD2/3 and MBD2/3Δ has also been isolated from the cricket (*Acheta domesticus*), which does have a methylated genome. The cricket proteins are more closely related in sequence to the mammalian MBDs than to dMBD2/3, and aMBD2/3 specifically binds methylated DNA [67]. Two studies in *Drosophila* showed no detectable methyl-CpG binding by MBD2/3 [67, 68], but in a third study, weak binding was observed [69]. It seems unlikely that dMBD2/3 would have this capability. There is both a 20 amino acid insertion, and a deletion in its MBD which would significantly disrupt the domain fold. The flexible loop, L1, would be hugely extended, and the α helix deleted - both are essential structural features of the MBD. It has been proposed that the role of dMBD2/3 is in developmental regulation, where it functions as a corepressor to silence specific genes [68]. The long-form appears in the cytoplasm in early embryogenesis. It then enters the nucleus at a stage which coincides with increased genomic activation and peak-level DNA methylation. Later still, MBD2/3Δ is the predominant form [70]. dMBD2/3 can associate with dHDAC1 and dMi-2, so it could be a component of *Drosophila*NuRD-like complex [67].

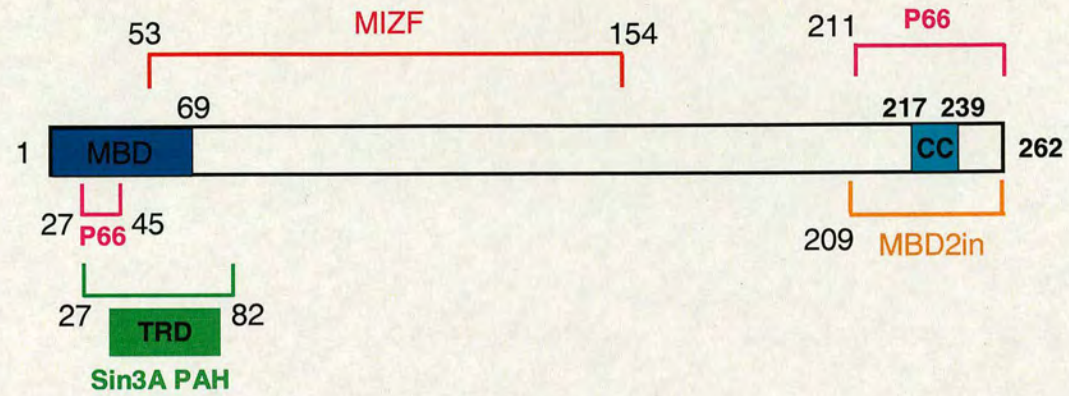

Figure 1.5: Diagram to show the regions of MBD2b that interact with various proteins and protein-complexes

1.5 MBD2 interactions

The focus of this thesis is MBD2, so it is important to look in more detail at the specific interactions and functions of this protein. A number of different MBD2 binding partners have been discovered, and are described below (see figure 1.5).

1.5.1 MBD2-MBD3 complex

MBD3 has been reported to form a complex with MBD2. Since the MBD2-MBD3 complex was shown to have a preference for hemimethylated DNA, it is thought to be targeted to replication foci, to establish transcriptionally repressive chromatin. MBD2 and MBD3 were both shown to homodimerise, and to form heterodimers together. In the case of MBD2b, this most likely involves both N- and C-terminal domains, in particular the MBD and the predicted coiled-coil [71]. This interaction suggests a potential role for MBD3 in maintaining transcriptionally repressed chromatin.

1.5.2 The Sin3 and NuRD complexes

Sin3 is a large multidomain protein found in yeast, *Drosophila* and higher vertebrates. It has four, highly conserved imperfect repeats, which are predicted to fold into paired amphipathic helix (PAH) domains, connected by flexible linkers. Sin3 is thought to function as the foundation of large repressive complexes, with multiple components. Small variations in the subtype and stoichiometry of these components tailor Sin3 to its various functions in different species. MeCP2 and MBD2 have both been shown to recruit the Sin3 complex; MBD2 binds to the PAH3 of Sin3 variant Sin3A, through a short region (aa. 27-82) which overlaps the MBD, (see figure 1.5). Because this region was also the minimum needed for efficient transcriptional repression, it is termed the TRD [55]. MeCP2 has a C-terminal TRD [47], which is also capable of mediating a repressive effect through Sin3A [48, 72]. The Nucleosome Remodelling/HDAC (NuRD) complex, also known as Mi-2, is a 2MDa complex found in mammals, *Xenopus* and probably also in plants. The NuRD and Sin3 complexes share a common core of HDACs 1 and 2, and histone binding proteins RbAp48 and 46. Each complex also contains three other polypeptides. These are Mi-2, MTA2 and MBD3 in NuRD; and Sin3A, and its associated proteins SAP30 and SAP18 in the Sin3 complex. The ATP-dependent nucleosome-remodelling activity of NuRD is due to Mi-2, which belongs to the CHD subfamily of SWI2/SNF2 DNA helicases/ATPases. Another feature of NuRD is metastases associated protein 1 (MTA1) or the similar protein MTA2. However, MTA1 is of an unknown biochemical function.

Both MBD2a and MBD2b interact with NuRD, although the former is a more efficient association. The mammalian NuRD complex does not bind to

methylated DNA on its own, but MBD2a can tether NuRD to methylated DNA [73]. MBD3 copurifies with NuRD, and has been shown to interact with MBD2. It is thought that the MBD2-MBD3 interaction may be the mechanism by which NuRD is tethered to methylated DNA.

1.5.3 MeCP1 and p66

MBD2 has been identified as a component of the MeCP1 complex [52], which contains Mi2/NuRD and two additional polypeptides, p66 and p68. However, it is not known which of the MeCP1 components binds directly to MBD2. Two mammalian proteins, homologous to the p66 component of the *Xenopus* NuRD complex, were identified as MBD2b binding partners. One of these, $hp66\alpha$, is orthologous to the p66 component of Mi-2/NuRD, whilst $hp66\beta$ is identical to p66 of MeCP1. MBD2 and hp66α colocalise in the nucleus, where hp66α functions as a potent transcriptional repressor [74]. There are two regions on MBD2 which specifically bind $hp66\alpha$: A region from amino acids 27 to 45, termed ID1, and a region from 211 to the very C-terminus (termed ID2), see diagram 1.5. p66/p88 have a GATA-type zinc-finger domain and two highly conserved regions. CR1 is involved in the association with the MECP1 complex, while CR2 is needed for localising p66/p68 and MBD3 to specific loci [75].

Both p66α and β also bind MBD3 [74].

1.5.4 MIZF

The transcriptionally repressive effect of MBD2 is enhanced by MIZF, a 60kDa novel zinc finger protein. This occurs in an HDAC-dependent pathway, although

MIZF can't associate directly with HDACs. The interaction between MBD2 and MIZF was identified in a yeast two-hybrid screen, then confirmed by IP experiments and and *in vitro* binding studies. Although MIZF has 7 zinc finger domains similar to the C2H2 type (found in numerous TFs [76] and KAISO [77]), only fingers 4 to 7 are necessary for MBD2 binding. This lies in the region between amino acids 53-154 [78] (see figure 1.5).

MIZF acts as a negative regulator of Rb expression, and so may inhibit myogenic differentiation [79]. Could MBD2 be involved in this process? MIZF binds to a specific sequence in the gene's promoter, requiring zinc fingers 1-4 and 7 [80]. However, this does not happen in a methylation dependent manner, despite the presence of at least one CpG. The consensus MIZF binding sequence is also found upstream of the GSTP1, p14/ARF and p16/INK4a genes, which are all selectively associated with MBD2 [81, 82]. Therefore, cooperation between the two proteins to control gene silencing is a possibility.

1.5.5 MBDin

In contrast, a 374a.a. protein called MBDin relieves MBD2-mediated transcriptional repression in transient transfection assays. The C-terminal 53 a.a.s of MBD2 are sufficient for this interaction [83]. MBDin was found to be identical to XAB1 (XPA binding protein 1), a cytoplasmic GTPase and the interaction partner of the DNA repair enzyme, Xeroderma Pigmentosum group A protein (XPA) [84]. In co-IP experiments, XPA was detected with anti-MBD2 antibodies, but only in the presence of MBDin, so perhaps it acts as a bridge between the two? Although MBDin/XAB1 is mainly cytoplasmic, the protein contains a potential

nuclear export signal (NES) and has been detected in the nucleus, co-localising with MBD2 to major satellite DNA. It is able to re-activate MBD2-mediated repression of promoters in this densely-methylated region, without altering methylation patterns. This could constitute a link between DNA methylation and nucleotide excision repair (NER) systems, similar to the MBD4-MLH1 interaction for mismatch repair [85].

MBDin/XAB1 does not appear to act via a mechanism like that of the displacement of MBD2 by GATA-3 in IL-4 induction [86], where there is direct competition between the two proteins. The presence of MBD2 actually improves gene reactivation by MBDin/XAB1, so perhaps MBD2 is recruited by MBDin, to block an otherwise repressive pathway.

1.6 MBDs and disease

There are a number of different associations between MBD proteins and human disease. The most intensively studied is the mutation of MeCP2 and Rett syndrome, a neurodevelopmental disorder. Mutations are mainly found in the MBD, TRD and C-terminus, and are thought to interfere with methylation-dependent transcriptional silencing. MeCP2 is now also thought to regulate neuronal gene expression and the formation of new synaptic connections.

The link between MBD mutations and autism has been investigated [87]. Apart from the occurance of MeCP2 polymorphisms in some autistic patients, no clear link between the MBD family and autism could be found.

1.6.1 MBDs and cancer

A relatively new area of investigation is into the connection between the MBDs and human cancer. Tumour development in humans is associated with a change in methylation pattern: DNA of cancerous cells exhibits local hypermethylation at CpG islands, whereas the rest of the genome is globally hypomethylated [88].

It is not clear which comes first. Is methylation of CpG islands the cause of transcriptional silencing or do genes become methylated after being silenced by another process? In embryonic cells, gene silencing precedes *de novo* methylation, so DNA methylation is thought to maintain the inactive transcriptional state rather than establish it (discussed in [89]).

However, the role of MBDs in cancer progression is unclear due to conflicting evidence about their effects. This is illustrated by the volume of knowledge about MBD2.

MBD2 has been shown to associate with the aberrantly methylated, regulatory ends of the p14 and p16 tumour-suppressor genes. It is thought to exert a repressive effect through a histone deacetylase-dependent pathway [81], and so silence this locus in colon carcinomas. Could these tumour-supressor genes be reactivated by the deletion of MBD2?

Transcriptional silencing of the π-class glutathione S-transferase, GSTP1, is mediated by MBD2. This gene is commonly inactivated by somatic CpG island hypermethylation, in prostate, liver and breast cancers [90].

MBD2 has been implicated in the failure of mice with adenomatous polyposis coli germ-line mutations (APC^{Min}), to survive (see figure 1.6. APC^{Min} mice are prone to developing colorectal tumours, but MBD2 nulls hardly get any tumours

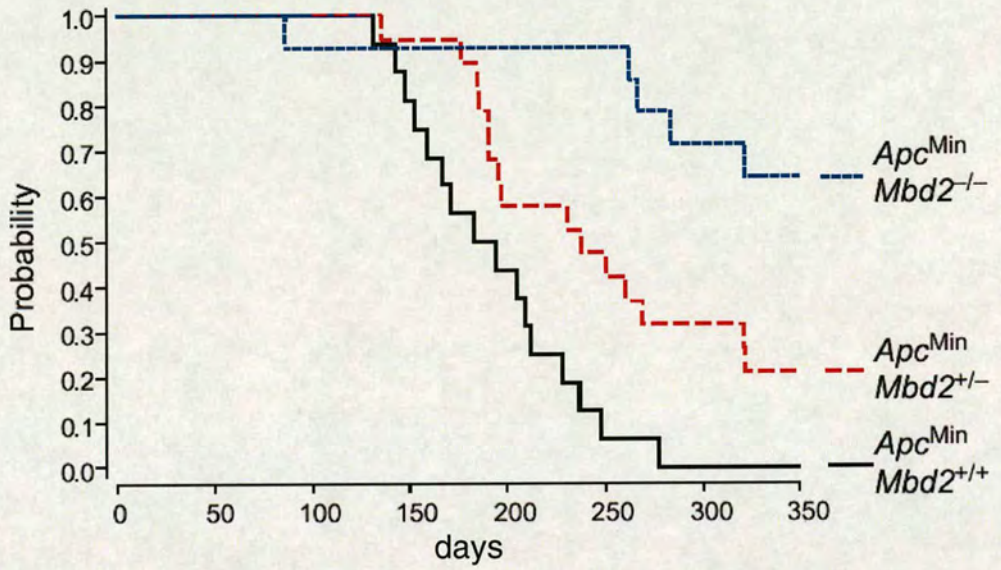

Figure 1.6: A survival curve of APC^{Min} mice, for three different Mbd2 genotypes. Figure taken from [91].

---

and show much better survival rates. The heterozygotes show a survival pattern halfway between that of the wild type and nulls, so the concentration of MBD2 appears to be proportional the effect [91].

MBD2 deficiency also does not accelerate lymphomagenesis in p53-null mice. In contrast, when DNMT1 is deleted, there is increased tumorigenesis and genomic instability [92]. Methylation interpreters are therefore better potential candidates for drug targets. However, there is evidence to suggest that deregulation of MBD2 itself occurs in some cancers. Breast carcinomas can be divided into two groups based on their expression level of MBD2; MBD2 mRNA levels are significantly associated with tumour size in invasive ductal carcinomas [93]; and a reduction in MBD2 mRNA expression was found in colorectal and gastric cancerous tissues [94]. Finally, a significantly reduced risk of developing bladder cancer is associated with high levels of MBD2 expression, demonstrating a protec-

tive role for MBD2 [95]. These observations will only be reconciled when we have a better grasp of the underlying mechanisms of cancer epigenetics. One recent study has found that oxidative DNA damage in the methylated MBD recognition sequence, can reduce MeCP2 binding by an order of magnitude [96]. By lowering the ability of MeCP2 to distinguish between methylated and non-methylated sites, DNA damage might result in expression of normally silent genes. The same group has also investigated the effect of halogenation at the 5 position of cytosine, which enhances binding of the MeCP2 MBD to DNA.[97]. These halogens, produced by inflammatory processes, may then act as false methylation initiation signals. By recruiting DNMTs and other machinery, the aberrant methylation patterns so characteristic of cancer, might be produced.

1.7 MBD structure

1.7.1 Comparison of MBD folds

The MBDs of both MeCP2 and MBD1 have been solved by NMR [98, 99]. Although the MeCP2 construct used for structure determination was larger, the ordered portion corresponds to 69 MBD residues that overlap with the MBD1 (1-75) structure. The domain from both proteins has the same basic wedge-shaped α/β-sandwich fold, shown in figure 1.7. One face of the wedge is made up of a twisted β-sheet and is extended by a flexible loop. The other side is an α-helical region in MeCP2 and an α-helix plus hairpin loop in MBD1.

The fold has a well-defined hydrophobic core, and several well-conserved hydrophobic residues such as Leu11, Trp15, Leu49, Phe62 and Phe64 (MBD1 num-

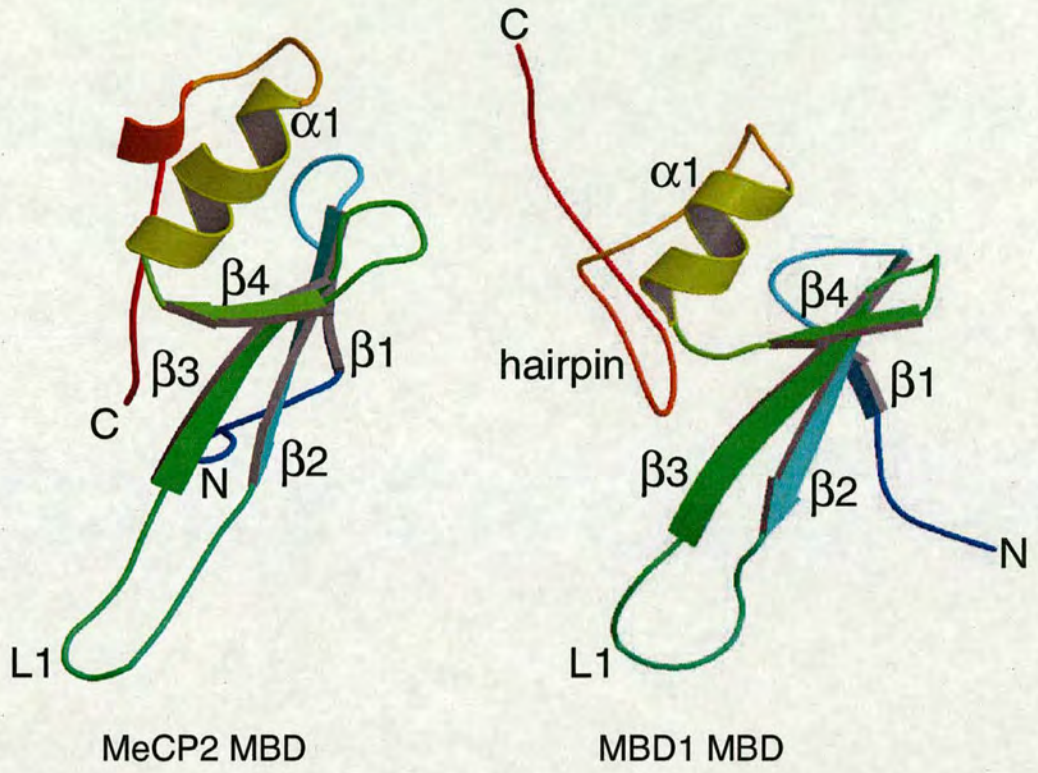

Figure 1.7: A comparison of the MeCP2 and MBD1 MBDs in solution, showing a common fold. A cartoon of each MBD in a similar orientation is shown. Secondary structure elements are numbered and displayed as arrows (β-strands) or ribbons (α-helices) and the long loop is denoted L1. MeCP2 has a longer α-helix and lacks the C-terminal hairpin loop.

bering) are common to both structures. Tyr34 and Tyr52 are also highly conserved and have solvent exposed aromatic rings in both structures. Several basic residues are found grouped together on one side of the molecule; a surface made up of Arg30, Arg42, Arg44, Lys46 and Lys65; and Arg22 and Lys23 in the L1 loop. In both structures, the flexible L1 loop has few long-range NOEs. This is perhaps surprising as its residues are well-conserved amongst MBD family members. Heteronuclear NOE values are smaller in the loop residues compared to the well-defined regions on either side.

In the MBD1 structure, the hairpin loop is held together by a main-chain

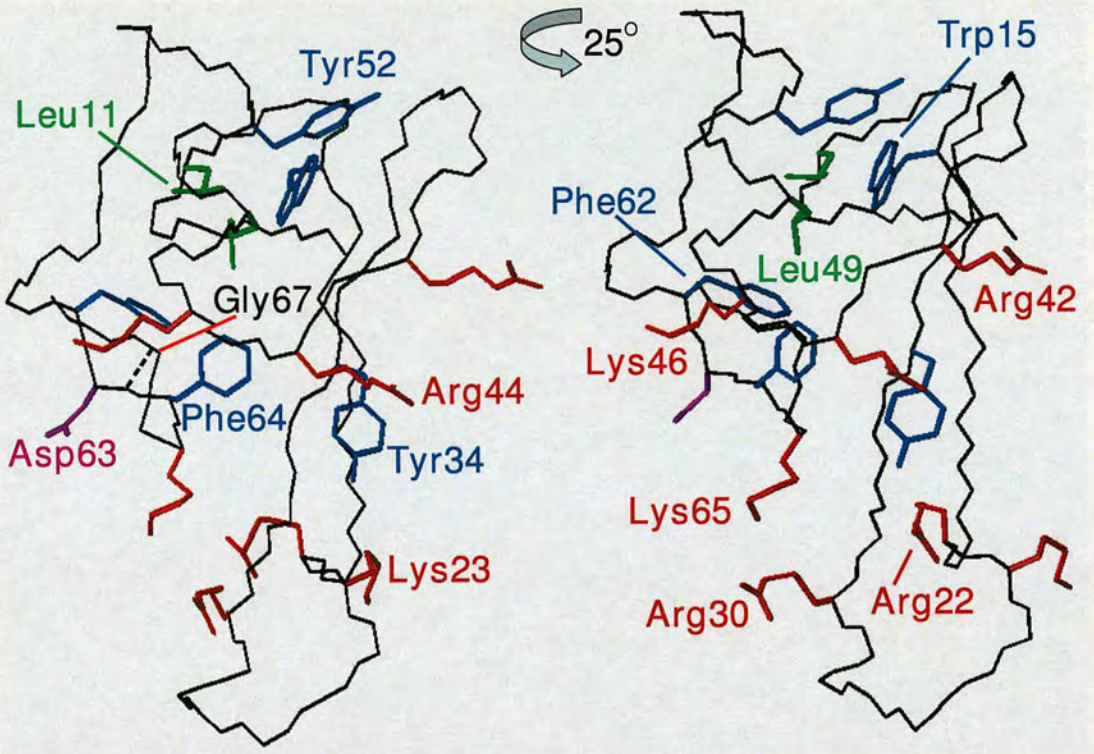

Figure 1.8: The structure of the MBD from MBD1, with the backbone shown in black. The side chains of certain residues are coloured: arginines and lysines in red; aromatics in blue; aliphatics in green and Asp63 in purple. The position of Gly67 is indicated with a red line, and the hydrogen bond between it and Asp63, is shown by a black dotted line. The molecule has been rotated by 25^o between each depiction.

---

hydrogen bond between Gly67 HN and Asp63 carbonyl. The two C-terminal phenylalanines, Phe62 and Phe64, are important for forming contacts with the rest of the molecule to stabilise the orientation of the hairpin. Substitution of Phe64 with Ala in MBD1 disrupted folding and DNA-binding activity. Deletion of residues 157-162 in MeCP2 (which corresponds to 64-69 in MBD1), causes a complete loss of methyl-CpG-binding ability. This hairpin is not present in the MeCP2 structure, but the C-terminus still plays an important role in fold definition.

1.7.2 MBD:DNA complex

The structure of the MBD1 MBD in complex with methylated DNA has also been solved by NMR [100] (see figure 1.9). A construct comprising residues 1-75, was combined in an equimolar ratio with a 12bp symmetrically-methylated oligonucleotide. In the complex, the MBD has a similar fold to the free domain, and most of the structure is unaltered by DNA binding. Major structural changes do occur in loop L1 however, which becomes much more rigid and adopts a hairpin-like shape. L1 heteronuclear NOE values increase, so as to be comparable with those found in the more permanently structured regions. This difference is due to seven out of nine residues in the loop, which form mainly hydrophobic contacts and hydrogen bonds with one strand of the DNA duplex. The second strand is bound by the N-terminal end of the α-helix, and loop L2, which is inserted into the major groove. Base contacts are made by the side-chains of L2 residues, whilst the helical residues interact with the backbone.

1.7.3 MeCpG recognition

In the MBD1:DNA structure, specific recognition of the meCpGs is due to a 'hydrophobic' patch made up of 5 residues. One methyl group is recognised by the side chains of Val20, Tyr34 and Arg22, while the other is contacted by Arg44 and Ser45 side chains (see figure 1.10). With the exception of Val20, mutation of these residues, which are highly conserved throughout the MBD family, results in reduced or abolished DNA binding [100].

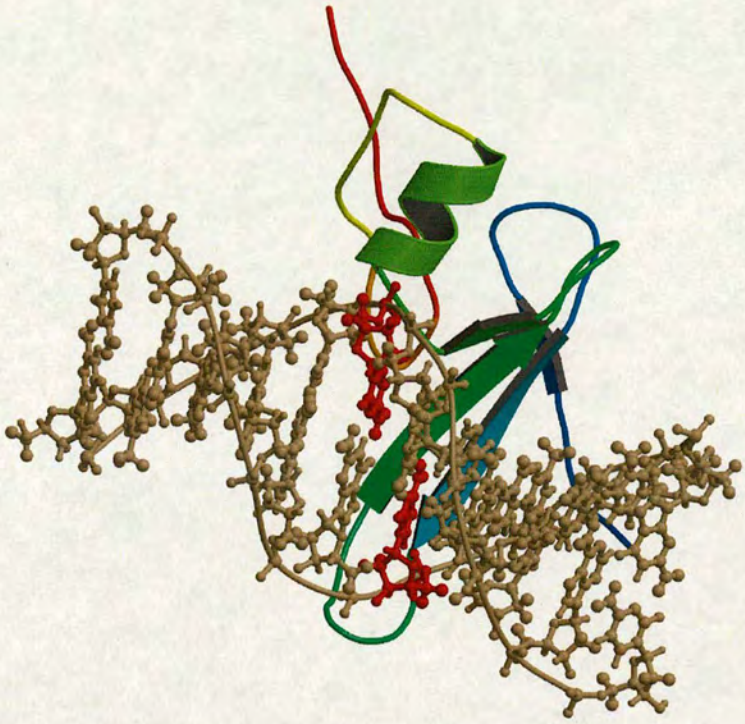

Figure 1.9: The MBD1 MBD in complex with methylated DNA. The figure was produced from coordinates in the Protein Data Bank (PDB) [101], code 1IG4. The protein is shown as a cartoon, in a similar orientation to those in figure 1.7. The DNA is in ball and stick form with the methylated cytosines highlighted in red.

1.8 Project aims

The aim of this project is to better understand the particular role of MBD2 in methylated DNA binding. Although the MBD sequence is well conserved throughout the family, the proteins have distinctive functions. Many of those are due to regions of the protein outside the MBD, and will not be addressed here. However, it remains to be discovered if there are differences in the methylated DNA binding *mechanism* of the various family members. Previous structural work on the family has almost exclusively employed NMR. Apart from the MBDs of MBD1 and MeCP2 mentioned earlier, the matrix attachment region-binding

domain of chicken MeCP2, which has an MBD-like fold [102], was also solved by NMR. The exception is the glycosylase domain of MBD4 [57], which is a crystal structure, but it has a completely different fold. The free MBD is likely to be too flexible to form crystals, but due to the more rigid structure formed upon DNA binding, it may be possible to crystallise the MBD:DNA complex. The areas of investigation in this project are summarised by the following objectives:

- To determine the structure of the MBD from MBD2 by NMR.
- To study the DNA-binding interface of MBD2 using NMR or X-ray crystallography, or both.
- To investigate the ability of MBDs to discriminate between DNA with a single, central CpG, methylated to different degrees.

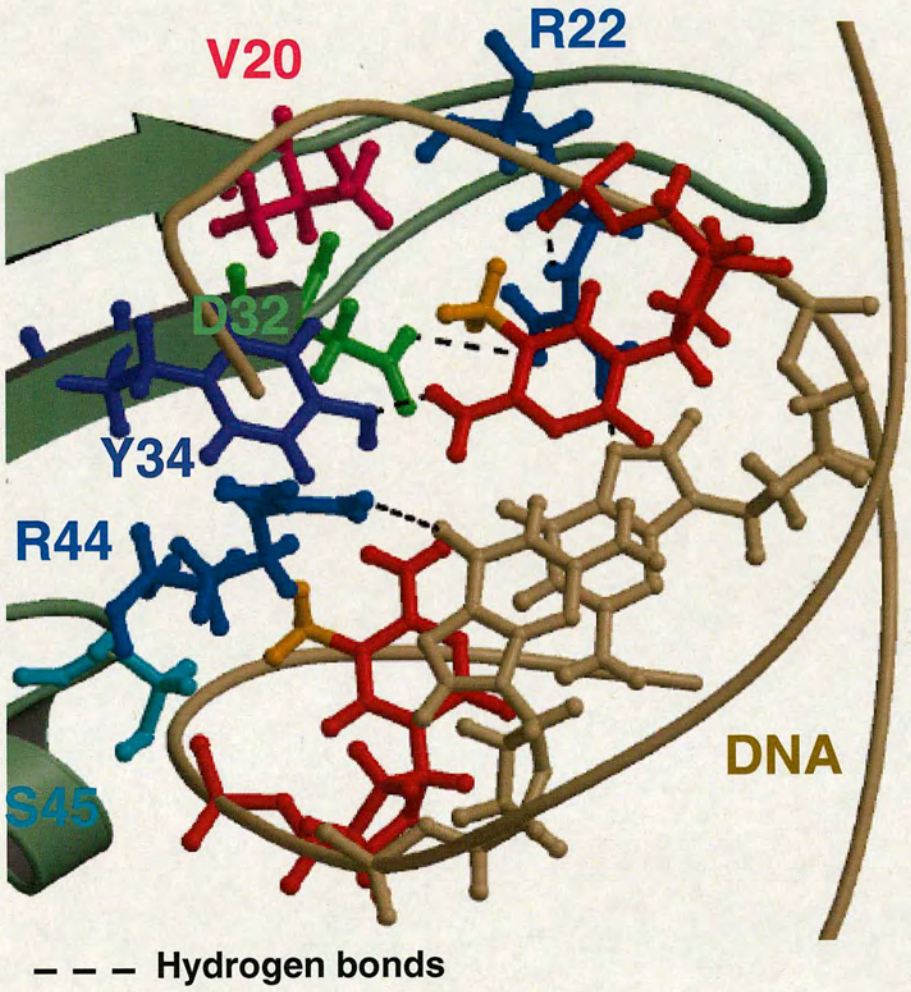

Figure 1.10: The MBD-DNA interface, with the DNA in beige, apart from cytosines in red. Methyl groups are shown in orange. The protein backbone is shown in dark green. The 5 key residues involved in meCpG recognition are coloured and labelled accordingly. Asp32 (bright green) forms an important salt-bridge hydrogen bond to direct Arg22 [100].

Chapter 2

Sample Preparation

2.1 Introduction

This chapter describes the production of protein and protein:DNA samples for a variety of investigative techniques. The method for producing the MBD2 MBD for NMR is described in detail, as it is the basis for the majority of the work done. Variations on this method were used to produce other samples.

2.2 Plasmid description

DNA corresponding to amino acids 153-224 of mouse MBD2a, (1-72 of MBD2b) had previously been cloned into the pET6H vector [103], by Andrew Free at The University of Edinburgh. The expressed construct has a total of 80 residues and includes an N-terminal 6 histidine purification tag and two additional non-native residues (MHHHHHHA). The theoretical mass of the construct is 9.1kDa as calculated using the EXPASY website [104]. The MeCP2 construct was derived

from DNA corresponding to residues 76-167 from rat MeCP2, and cloned into the pET6H vector [103] by Andrew Free. The expressed construct has a total of 100 residues and includes an N-terminal 6 histidine purification tag and two additional non-native residues (MHHHHHHA). DNA corresponding to residues 63-141 of mouse MBD4, including a C75S mutation, was cloned into the pET30b vector (Novagen) by Brian Smith and Julia Boyd at The University of Edinburgh. The expressed construct has a total of 125 residues and includes an N-terminal 6 histidine purification tag and 46 additional residues from the pET30 sequence.

MBD	native residues	vector	total residues	Mw (kDa)	pI	ε_{280nm}
mouse MBD2	1-72	pET6H	80	9.1	9.9	9530
rat MeCP2	76-167	pET6H	100	11.4	9.8	10810
mouse MBD4	63-141	pET30b	125	13.8	9.3	8250

Table 2.1: A table of the protein contructs used, along with theoretical values for molecular weight (Mw), isoelectric point (pI) and extinction coefficient at 280nm (ε_{280nm}), calculated using the EXPASY[104] tool *ProtParam*.

DNA for all constructs was provided by Adrian Bird's group at the Wellcome Centre for Cell Biology, The University of Edinburgh. A summary is given in table 2.1.

2.3 Transformation

Constructs were provided as minipreps (plasmid DNA in water or buffer), stored at -20°C. *E. coli* BL21 (DE3) pLysS cells (Novagen) were made competent using either calcium chloride [105], or TSB methods (see Appendix A). The cells were then transformed with 1μl of a 1:10 dilution of plasmid miniprep (approx. 0.5pg

DNA). After a final incubation step (37°C for 1 hour), the cells were plated out on agar with appropriate antibiotic selection. Chloramphenicol selects for pLysS cells [106], kanamycin for pET30b and ampicillin for pET6H plasmids.

After incubation, single colonies were picked from which to grow overnight cultures for protein expression. Fresh stocks of plasmid were also prepared from overnight cultures, using the Spin Miniprep kit (Qiagen). Instructions supplied with the kit were followed; plasmid DNA was eluted from the preparation columns with deionised, autoclaved water.

Fresh (previous day) transformations were always used for protein expression as older transformations were suspected to be linked to sample instability.

2.4 Protein expression

Protein samples were made from 0.5-2L of culture, depending on expression levels. Each 500ml of Luria-Bertani (LB) growth medium was inoculated 1:50 with an overnight culture of *E. coli* cells, using appropriate antibiotic selection. The incubation temperature was 37°C and cells were shaken at 200rpm. Cell density was measured by the absorbance at 600nm (A_{600}) on a UV/visible range spectrophotometer (Lambda 20, Perkin Elmer). Once the A_{600} reached between 0.5 and 0.6, (when cells are typically in log phase growth) protein expression was induced by the addition of 1mM isopropyl-β-D-thiogalactopyranoside (IPTG, Sigma) [107].

Cell density was monitored regularly during the protein expression phase as an indicator of cell health. The expression time was typically 5 hours, and not more than 6 hours, to give the maximum, high-quality yield. Cells were harvested

immediately if a decrease in A_{600} was measured, but this was not usually observed during growth times of less than 6 hours.

Cells were harvested by centrifugation (Beckman Coulter) at 15,000g for 15mins. The supernatant was discarded and the cell pellet transferred to Falcon tubes before being spun again in a benchtop centrifuge at 3,000rpm for 10mins, to compact the pellet as much as possible. The supernatant was again removed and the pellet weighed. At all stages after harvesting, the cells were kept cold, at 4°C. A typical yield produced 4.5g cells/L and the pellet may be stored frozen at -20°C for several months.

2.4.1 Isotopically labelled samples for NMR

Isotopically labelled samples were produced as described in section 2.4 above except LB media is replaced by M9 minimal media. Minimal media must be supplemented with iron, zinc, and sources of carbon and nitrogen. This allows samples to be labelled with stable, NMR-active isotopes. Either $^{15}NH_4Cl$ (and unlabelled glucose) can be added to make a ^{15}N-labelled sample, or both $^{15}NH_4Cl$ and ^{13}C-labelled glucose to make a double-labelled sample. Recipes and experimental details are shown in Appendix A. Subsequent harvesting and purification steps are the same for all samples.

2.4.2 Cell lysis

Harvested cell pellets were resuspended in Bugbuster detergent (Novagen) to disrupt the cell membranes, causing the cells to lyse. Benzonase nuclease (Novagen) was added to digest nucleic acids, which made resuspension easier. The cell lysate

was kept cold and Complete Mini EDTA-free protease inhibitors (Roche) were added - both help to prevent the degradation of the protein. Resuspension was carried out for 20-30mins on a rotating wheel or rollers, until the lysate was homogeneous and free of large clumps of cells.

The lysate was then transferred to centrifuge tubes and spun at 50,000g for 30mins. Protein expressed from all three MBD constructs, is predominantly found in the soluble cell fraction, with a negligible amount in the insoluble fraction. Due to this fact, the supernatant was removed and the pellet discarded.

2.4.3 SDS-PAGE

Proteins were separated and analysed using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). SDS is an anionic detergent, molecules of which wrap around, denature and confer a negative charge on a protein, in proportion to its length. Separation occurs in an electric field, produced by applying a current to the polyacrylamide gel, a porous matrix. Thus, in SDS-PAGE, proteins become negatively charged rods, which migrate through the gel towards the anode. Migration distance is inversely proportional to the logarithm of mass of the protein [108]. MBD proteins were separated on 15% polyacrylamide gels, run at a constant 200 Volts for 45 minutes, or until the dye front had run off the gel. Proteins were visualised by staining with Coomassie blue (for protocols see [109]). Typical gels are shown in figures 2.1 and 2.6.

2.5 Purification

Due to the requirement for high levels of purity in an NMR sample, the MBD purification strategy is a three-step process. It was originally devised by Dr. Andrew Free and optimised in our group.

2.5.1 Nickel-affinity purification

The N-terminal, 6 histidine tag (present in each construct), was used to affinity-purify the protein by nickel chelation [110]. 2-3cm nickel-nitrilotriacetic acid resin (Ni-NTA, Qiagen) was poured in Econopak columns (BioRad) for this purpose. Immobilised nickel in the NTA column resin coordinates the histidines, while proteins without multiple histidines wash through. The supernatant from cell lysis was made up to 1mM β-mercaptoethanol (to prevent oxidation and protein aggregation) and 10mM imidazole (to minimise non-specific binding). It was then applied to the Ni-NTA column. EDTA-free protease inhibitors were used earlier because EDTA, a divalent metal cation chelator, will compete with the histidine tag for nickel sites in the resin.

The column was then washed with binding buffer (50mM NaPi pH 8.0, 300mM NaCl), followed by increasing concentrations of imidazole in binding buffer. Imidazole competes with the histidine side chain for Ni-chelation until the histidine-tagged protein is eventually eluted (in this case at a concentration of 250mM imidazole). Following elution from the Ni-NTA column, dithiothreitol (DTT) was added to the protein solution to a final concentration of 5mM. The sample is 80% pure at this stage, but is still contaminated, for example by high molecular-

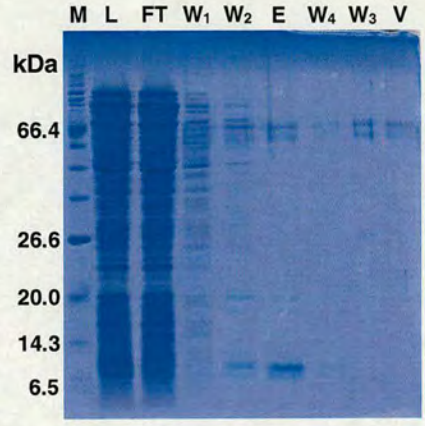

Figure 2.1: An SDS-PAGE gel showing MBD2 MBD protein present in each fraction collected from the Ni-NTA purification column. Fractions are: lysate (L), flow-through (FT), 4 washes (W), eluate (E) and Vivaspin concentrator flow-through (V). Broad range protein markers (NEB), are denoted by M, and selected molecular weight standards are labelled. The need for further purification steps is clear due to contamination of the eluate by several bands of high molecular weight (at around 66.4 kDa).

---

weight *E. coli* proteins with a high histidine content (see figure 2.1).

2.5.2 Ion-exchange chromatography

Ion-exchange chromatography [111] was carried out using a fast protein liquid chromatography (FPLC) system (Amersham Akta), controlled by an associated computer and UNICORN (v.2.01) software. A cation-exchange column, Fractogel $(SO_4)^{2-}$ (Merck), in 50mM Tris pH 8.0, was used to purify the MBD. Buffer-exchange into a solution of lower salt concentration was not necessary due to the high affinity of the MBD for the Fractogel matrix at this pH (See pI values in table 2.1 above). The protein was eluted from the column using a salt gradient from 0-1.5M NaCl, achieved by gradually combining 50mM Tris, 2M NaCl pH 8.0 with the running buffer. A typical elution profile is shown in figure 2.2. MBD2 MBD was eluted at a NaCl concentration of 450mM. All the fractions

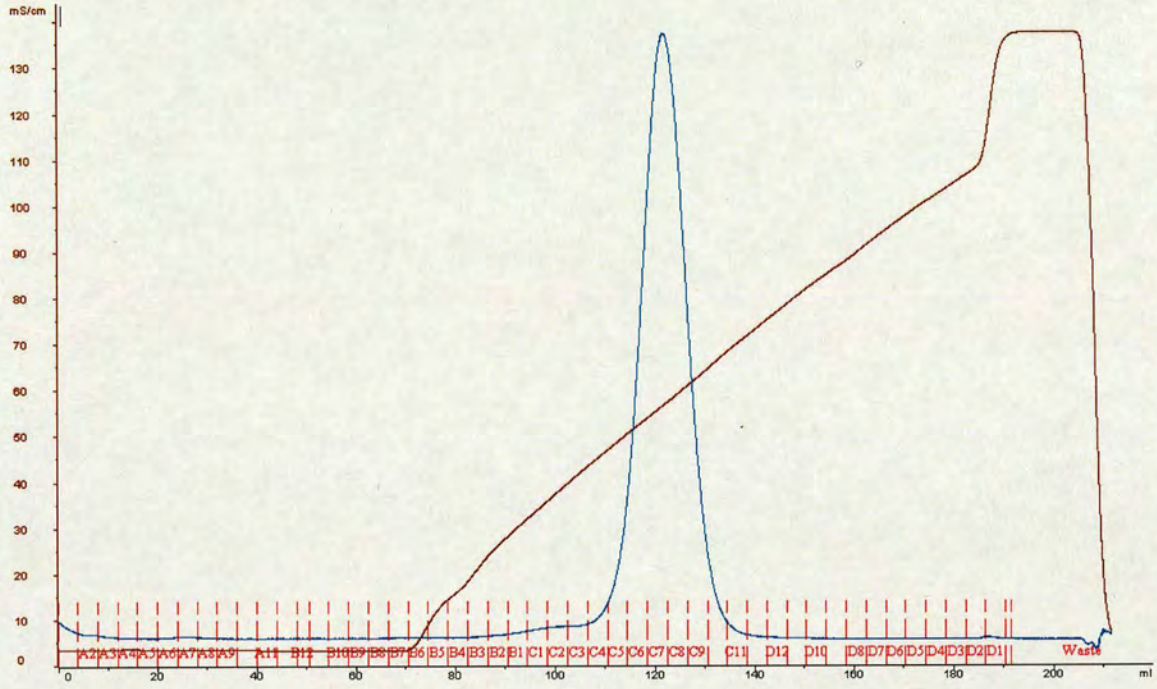

Figure 2.2: An ion-exchange trace of the MBD2 MBD (in blue), as determined by absorbance at 280nm. The change in conductivity, due to the linear (0-1.5M) NaCl gradient, followed by a 2M NaCl wash, is shown in brown. 4ml fractions were collected, the divisions of which are shown in red.

contributing to the elution peak were combined, and then concentrated using 20ml Vivaspin concentrators with a 5kDa Molecular Weight cut-off (VivaScience). These were spun at 4,000rpm in a bench-top centrifuge, until the required volume was obtained. DTT was then added to a concentration of 5mM.

2.5.3 Purification by gel filtration

The protein was further purified by size exclusion [111], using a Superdex 75 gel filtration column (Pharmacia). This may be done in the buffer of choice for the next stage, for example 50mM NaPi, 50mM NaCl pH 7.0 (filtered and degassed) for NMR samples. The Superdex column can handle a maximum volume of 200μl

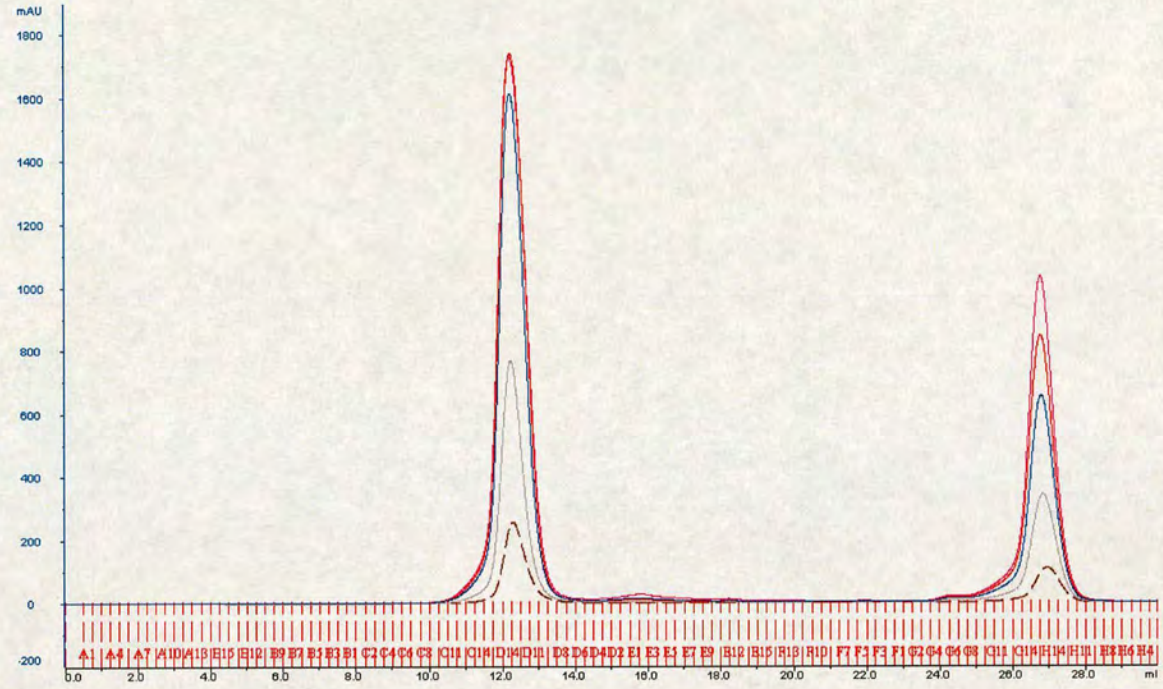

Figure 2.3: A gel filtration trace of the MBD2 MBD, as determined by absorbance at 280nm. Overlaid in different colours are successive runs of the same sample. Run 1, magenta; 2, red; 3, blue; 4, grey; 5, brown. 0.25ml fractions were collected, shown as red divisions along the x-axis.

and care was taken to avoid over-loading the column with too much protein, resulting in sample loss. Typically, 1ml of 2mM protein was purified by FPLC in up to five runs. An elution profile for the MBD2 MBD is shown in figure 2.3.

2.6 Final sample preparation

The final stage in sample preparation was to combine the peak fractions of the multiple gel filtration runs. These were concentrated using a 6ml Vivaspin tube (VivaScience) spun at 4,000rpm in a bench-top centrifuge, until the desired volume ($\sim 200\mu l$) was reached. An estimate of the sample concentration may be obtained by measuring its absorbance at 280nm (A_{280}). The theoretical extinc-

tion coefficient (see table 2.1) can then be used to calculate the concentration according to Beer's Law. It was most efficient to do this immediately before the final concentration step, even though a tiny amount of protein may have been lost in the concentration process. Yields for the MBD2 MBD construct after the ion-exchange step were 15mg/L in LB media and 30mg/L in M9 minimal media.

2.7 NMR samples

To achieve the greater level of purity required for NMR samples, only the central gel filtration fractions were combined (e.g. fractions C15 to D10 in figure 2.3). Vivaspin columns for the final concentration step were thoroughly washed with water and then NaPi NMR buffer, to remove any trace of glycerol from their membranes. If it was suspected that the sample had become contaminated with small molecules such as glycerol, the Vivaspin columns were thoroughly rinsed to achieve at least a 1000-fold dilution of any contaminants. After concentration, the protein sample was removed from the Vivaspin column and spun at 13,000 rpm for 10 minutes in a benchtop microfuge with controllable temperature (Kendro). This was done immediately before putting the protein in an NMR tube (Wilmad-Labglass), to produce a particulate-free sample. The cap was then placed on the tube and sealed with Nescofilm (Karlan) to prevent evaporation during NMR experiments. Typically, from 1L M9 culture of pET6H MBD2 MBD, a 600μl 1.4mM NMR sample can be obtained.

2.7.1 NMR sample in H_2O

Samples in H_2O, both labelled and unlabelled, were made up to a volume of 550-600μl[1], including 10% D_2O (Sigma). The sample also contained 10mM deuterated DTT and 1mM deuterated EDTA.

2.7.2 NMR sample in D_2O

A ^1H sample in D_2O was made by several rounds of buffer-exchange after the final concentration step outlined in section 2.6. The protein solution was first exchanged into the 99% deuterated NaPi buffer (50mM NaPi, 50mM NaCl, pH 7.0). This was done twice so as to achieve a 500-fold dilution of any H_2O present. The process of buffer exchange was repeated with 99.9% NaPi buffer,to achieve as high a deuteration level as possible. A description of how the deuterated NaPi buffer was made is given in Appendix A.

2.8 Crystallography sample

To make protein for crystallography, the NaPi buffer used in the gel filtration step (described in section 2.5.3) was replaced with 50mM MES (2-morpholinoethanesulphonic acid), 50mM NaCl pH 6.5. It is advisable to avoid phosphate buffers since these invariably crystallise, especially when combined with metal additives. The protein was concentrated, and the concentration of the sample calculated as described above (Section 2.6). 12 bp duplex DNA, with a symmetrically methylated central CpG pair, (sequence: 5'-GCTTAmCGTAAGC-

[1]This volume in an NMR tube corresponds to the minimum depth required by the spectrometer probe, for reliable measurements.

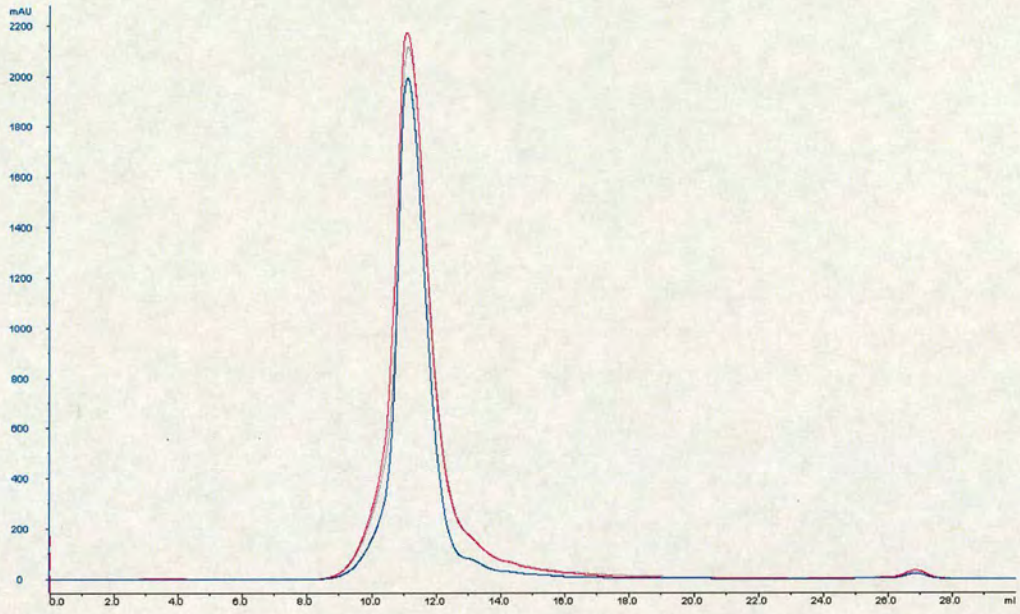

Figure 2.4: A gel filtration trace of the MBD2 MBD:DNA complex, as determined by absorbance at 280nm. Overlaid in different colours are successive runs of the same sample. Run 1, grey; 2, magenta; 3, blue. 0.5ml fractions were collected (not shown).

3', Oswell), was added to the purified protein in a near-1:1 ratio. A slight excess of protein was maintained to reduce the likelihood of DNA-crystal formation. The complex was then purified by gel filtration and the trace is shown in figure 2.4.

A comparison of gel filtration traces shows a shorter elution time for the MBD:DNA complex, indicative of its larger size (see figure 2.5). The central peak fractions were collected and concentrated to make a stock solution of the complex. This was used for crystallisation trials.

2.8.1 Crystal trials of the MBD2 MBD:DNA complex

A sample of the MBD2 MBD:DNA complex was made as described in section 2.8. Crystallisation was carried out using the vapour diffusion or 'hanging drop'

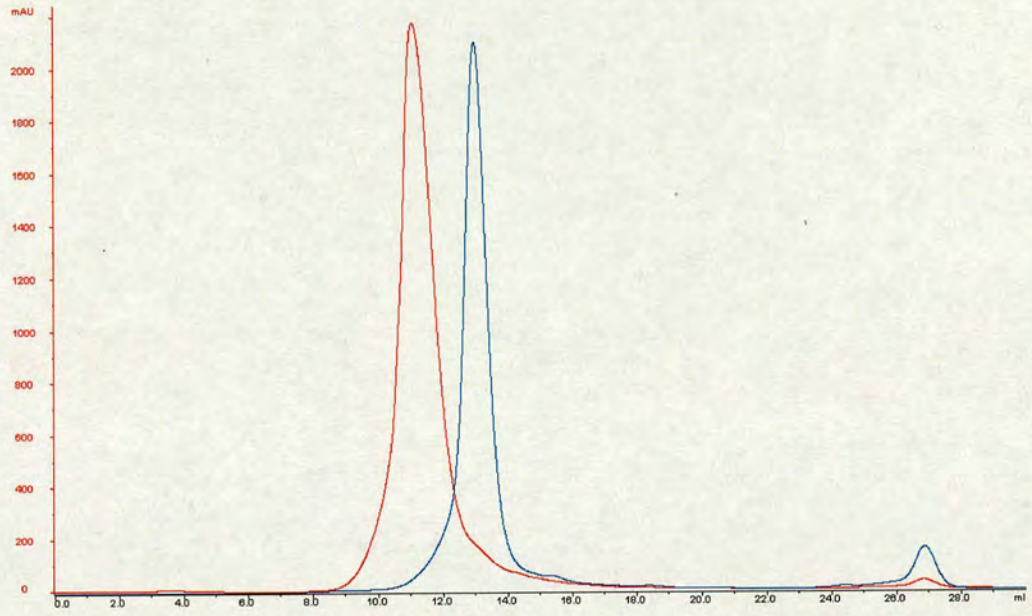

Figure 2.5: A comparison of the gel filtration traces of free MBD2 MBD (blue) and the MBD2 MBD:DNA complex (red), as determined by absorbance at 280nm. The complex is larger, and so takes longer to elute from the Superdex 75 column.

method, in 24-well plates. A range of buffers and precipitants make up the 'well', while tiny drops of protein are suspended above the well, in a sealed system. Crystals form as water diffuses out of the drop, eventually creating optimal growth conditions. The initial conditions were taken from previous work on the crystallisation of DNA binding proteins [112]. A broad-range screen yielded one potential candidate for optimisation. After several months at 18^{o}C, crystals were produced with 0.36mM MBD:DNA complex, in 100mM MES pH 6.5, and 18% 6kDa polyethylene glycol (PEG 6K). The 0.05m hexagon-shaped crystal was mounted in a cryo-loop and frozen in cryo-protectant (well solution plus 20% glycerol). In spite of attempts using both home and synchrotron (at the SRS Daresbury) X-ray sources, no diffraction pattern was observed. Attempts were made to improve crystal size and quality with additives such as divalent metal

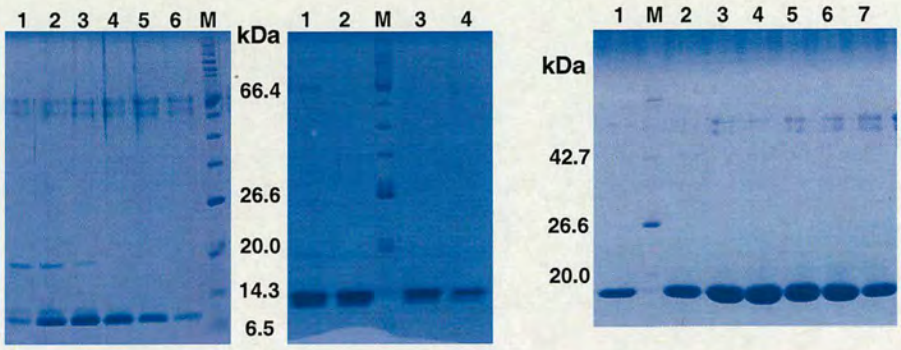

Figure 2.6: SDS-PAGE gels showing the MBD protein present in selected fractions after the ion-exchange purification step. For MeCP2 and MBD4, the fractions were judged to be pure enough for SPR experiments. For MBD2, all the numbered fractions were pooled and further purified using a gel filtration column (see text). Broad range protein markers (NEB), are denoted by M, and selected molecular weight standards are labelled.

cations and techniques such as microseeding, but with no success. Trials using a specialised broad-range DNA complex screen (Hampton Research) also failed to produce any single crystals.

2.8.2 SPR samples

Protein samples of MBD2, MeCP2 and MBD4 MBDs were made for surface plasmon resonance (SPR) experiments. MeCP2 and MBD4 were purified by ion-exchange as a final step, since the level of purification required is lower. MBD2 was further purified by size-exclusion column (see figure 2.6). Although specialised buffers were used for SPR (see Chapter 3), protein purification was carried out in the buffers described above. Later on, dilution of the sample to achieve the required protein concentration, reduced the purification buffer concentration by at least 1000-fold.

2.9 Analysis

Electrospray mass spectrometry [113] on the expressed MBD from MBD2 yielded a molecular mass of 9097 Da, which agrees well with the predicted value. There is no evidence that any of the MBDs dimerise, either by NMR (see section 4.5.5) or by analytical ultracentrifugation (AUC) [114], (data not shown).

2.10 Sample behaviour

The main problems encountered with the MBD2 MBD, were due to sample instability. Originally, the purified protein was known to precipitate at concentrations greater than 1mM, if kept below 283K for even a short time. However, at 298K the sample only remained in a folded state for a maximum of 4-5 weeks, during which time the sample quality declined (rapidly after week 3). This is not long enough to collect all the NMR experiments for structure determination, so multiple samples were used. The MBD as an isolated domain may lack additional contacts with other residues, necessary for long-term stability.

2.11 Conclusion

Expression levels of all constructs were excellent. In particular, the MBD2 MBD sample for NMR could be made from as little as 1L of culture. Expression in minimal media was even higher than in LB media. Problems with MBD2 MBD precipitation and unfolding at low and high temperatures respectively, were common. No causes have been determined for either observation, and no solutions,

particularly for variable sample stability were found. All experiments were therefore limited by the sample lifetime of 3 weeks.

Chapter 3

Biophysical Studies

3.1 Introduction

Previous work on investigating the affinity of different MBDs for methylated DNA has been done using gel shift assays [41, 115]. These were done with oligomers of at least 20 base pairs in length and usually with multiple methylation sites. In contrast, this investigation uses Surface Plasmon Resonance (SPR), a technique that allows accurate determination of equilibrium constants (gel shifts are non-equilibrium systems). The affinity of various MBDs for a single, central methylated CpG in a short oligomer, was tested. The aim was to determine if differences between the MBD family members are apparent in the MBD and methyl recognition mechanism itself. It is also possible that the proteins rely on a common binding mode, with other regions of the proteins responsible for specific binding *in vivo*. SPR is used to measure the affinity of MBDs from three proteins for different oligomers. A comparison is made of binding to unmethylated, methylated, hemimethylated and mismatched DNA.

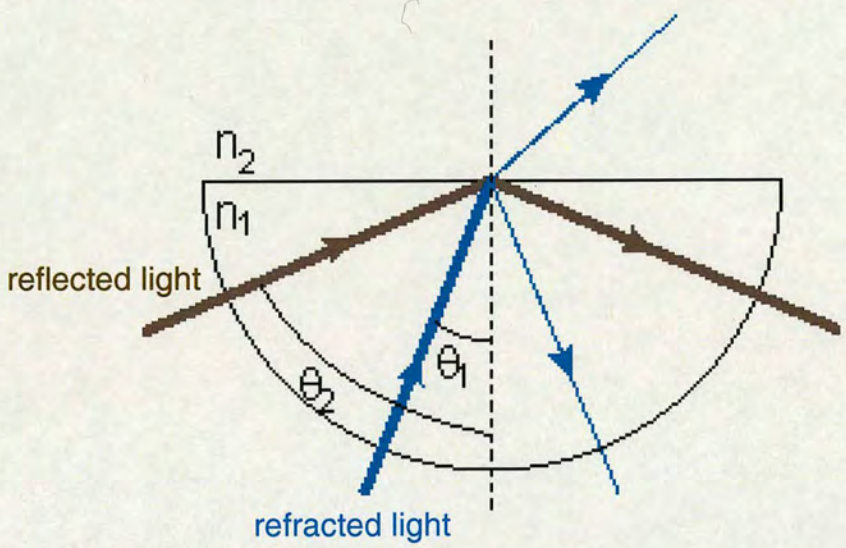

Figure 3.1: A diagram showing total internal reflection (TIR) when the angle of incident light is greater than θ, the critical angle.

3.2 Surface plasmon resonance

3.2.1 Introduction to SPR

SPR relies on a physical phenomenon called total internal reflection (TIR) [116]. When light travels from a dense medium to a less-dense medium (such as when passing through glass to air), it is refracted at the interface of the two media. Some of the light is refracted and some reflected back into the denser medium, as shown by the blue ray in figure 3.1. However, at a certain angle of incidence, θ, the light is refracted so as to be parallel to the plane of the interface. This angle is called the critical angle, and values at, or above the critical angle produce TIR. In this case, no light is refracted into the less-dense medium. Figure 3.1 shows two angles of incidence: $\theta 1$ is below, and $\theta 2$ above the critical angle. So the black ray is an example of TIR.

Providing the less-dense medium can absorb light at the appropriate wave-

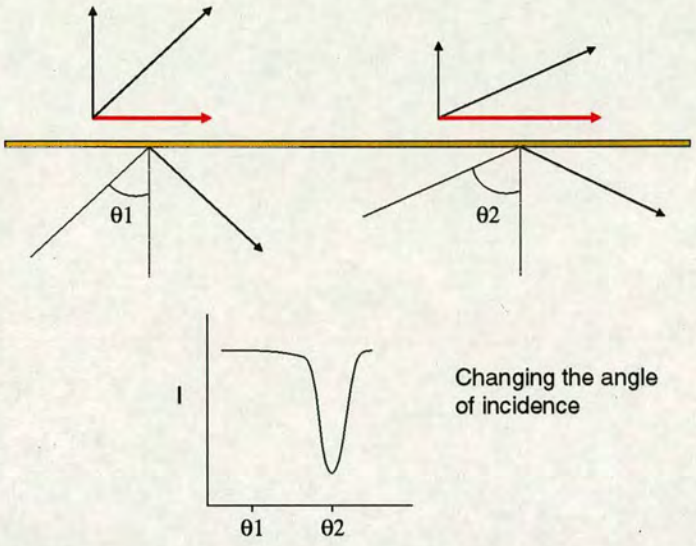

Figure 3.2: A diagram to illustrate how the vector component in the direction of the plane interface may be altered by changing the angle of incidence.

length, the reflected beam produces an electrical field intensity, called an evanescent field wave, at the media interface. If this field wave is allowed to contact a conducting layer, such as a gold strip, this leads to excitation and oscillation of the metal's free electrons. Then, electromagnetic waves called surface plasmons, propagate at the interface. The evanescent field wave is enhanced as the surface plasmons propagate on the metal, and will penetrate the less-dense medium.

3.2.2 Resonance conditions

Plasmons are produced in the metal layer, only when the energy and momentum of the incident light vector is the same as that of the surface plasmons. There are two ways to alter the incident light so as to arise at this situation: One is to change the wavelength, so changing the energy and momentum of a photon. The other is to change the angle of incidence so that the proportion of energy in the vector component parallel to the surface is altered. This is shown in figure 3.2.

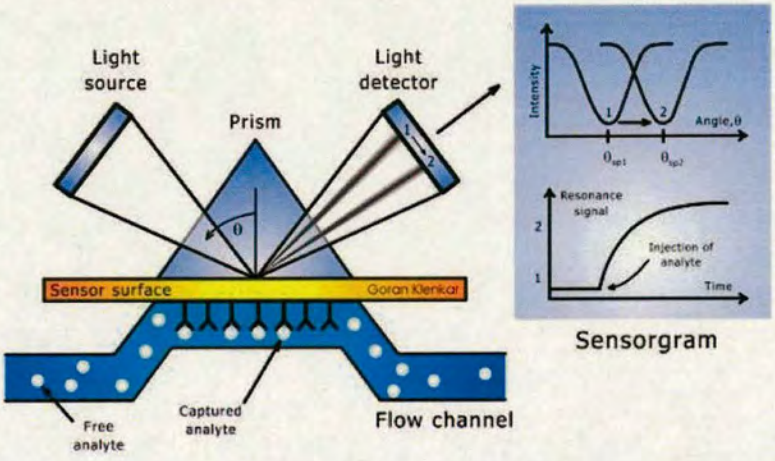

Figure 3.3: Analyte-ligand binding alters the refractive index at the chip's surface. The angle of incident light required for SPR changes, and it is this change which is displayed in real-time in a sensorgram. Figure based on that in [117], page 26.

As the energy of the light is transferred to plasmons, the intensity of the reflected light is reduced. In the case where the surfaces and wavelength are kept constant, a dip in the intensity of the reflected light will occur at the angle required for SPR.

Although the surface plasmons are present only in the gold strip, the energy and momentum of their wave-vector, is affected by the density of the medium immediately adjacent to, (within one wavelength of) the metal strip. So small changes in the medium can be measured by the corresponding change in the angle of incidence required for SPR. This principle is exploited in BIACORE to measure the change in refractive index, caused by a biomolecular interaction on the chip's surface. In practise, rather than individual angles of θ, the BIACORE system uses a wedge-shaped beam of light covering a range of angles of incidence. A detector measures the intensity of the corresponding angles of reflection [117].

3.2.3 The BIACORE system

The BIACORE instrument brings together an optical measurement unit, removable sensor chip, and automatic sample processing, all controlled by an associated computer. The sensor chip is made of glass coated with a thin gold layer, and a matrix designed for biomolecule attachment. The chip is a transducer which converts the change in refractive index caused by biomolecular interactions at it's surface into the SPR signal, which is read by the optical unit. The SPR change is measured in Resonance Units (RU), where 1000 RU represents a change in angle of 0.1degrees. This has been shown to correspond to a change in surface protein concentration of 1 ng/mm^2. The BIACORE 2000 instrument, used in this study, has a detection limit of 10 RU.

Two sensor chips were used in this study: The CM5 has a carboxymethyldextran matrix layer, and the SA chip, which has in addition, streptavidin protein bound to the activated dextran. Biotinylated molecules may be directly immobilised on the SA chip, as biotin has an extremely high affinity for Avidin and Avidin-derived proteins. The CM5 chip was modified by NeutrAvidin capture (using the amine coupling reaction) [118], effectively converting the CM5 to a SA chip.

The BIACORE system was used to measure the interaction between various MBD proteins and DNA oligonucleotides. The change in refractive index is measured as the analyte is allowed to flow over a sensor chip-bound ligand. Since the proteins are inexpensive to make, compared to purchasing DNA, these were used as the analyte, while the DNA was immobilised on the sensor chip via a biotin linker. The DNA is thus the ligand, and the MBDs the analyte, in BIACORE

terminology. Each chip has four flow cells, which may be handled independently (for instance when attaching different ligands), or together in series. The latter allows referencing of interactions in cells 2 to 4, with a blank cell 1.

3.3 Experimental methods

3.3.1 Preparation of the CM5 chip surface

A new CM5 chip was introduced into the Biacore 2000 machine using the automated docking procedure. All four flow cells were primed with running buffer (50mM HEPES, 150mM NaCl, 3mM EDTA, 0.001% non-ionic surfactant P20 or Tween 20) and normalised by injection of 40% glycerol solution, as detailed in the Biacore manual [118]. To activate the dextran matrix, 60μl of a 1:1 mixture of EDC/NHS (0.4M 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide and 0.1M N-hydroxysuccinimide) was injected at a flow rate of 10μl/min. Following this injection, the flow rate was increased to 50μl/min for one minute, to wash out any left-over EDC/NHS solution. NeutrAvidin solution (Biacore) diluted 1:6 in running buffer at pH 4.5 was then injected. A few microlitres were injected to test the response, followed by two injections of 20-30μl solution. The total change was 2500-3000 RU. Following 3 consecutive conditioning washes (40μl 50mM NaOH, 1M NaCl, at 40μl/min) there were no further changes in response. Individual flow cells were then washed with 70μl 1M Ethanolamine-HCl at 10μl/min, to deactivate any left-over reactive ester groups, followed by a wash with running buffer at 100μl/min for one minute. After this stage, the CM5 chip had effectively been converted to a SA chip, and the subsequent procedure was the same

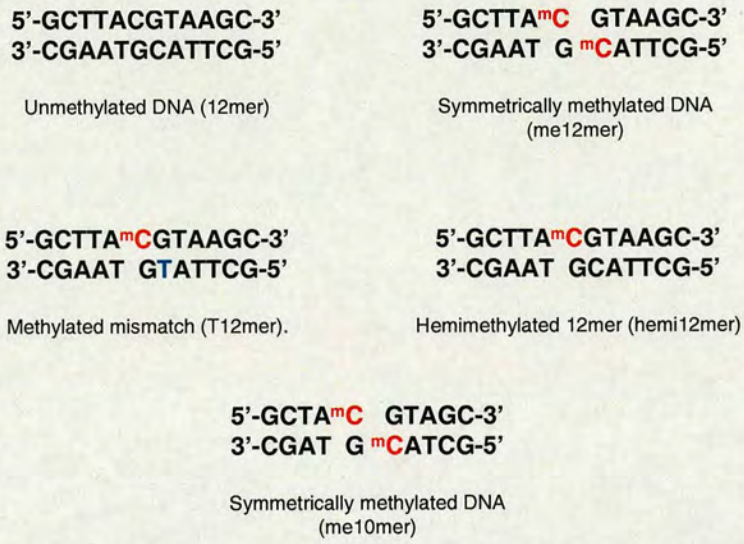

Figure 3.4: The five oligonucleotides used as ligands in the SPR study. Methylated cytosines are shown in red, the mismatched thymine in blue. Short-forms of the names (in brackets) are used in the text.

for both chips.

3.3.2 Preparation and immobilisation of oligonucleotides

Double-stranded oligonucleotides, biotinylated at one end, were produced by combining a ratio of 1:10, 5'-biotinylated:non-biotinylated single-stranded oligos (Oswell). The seqences were as shown in figure 3.4. This was done to produce a 50μg/ml stock solution (based on the biotinylated oligo concentration), which was heated to 45^oC for 5 minutes and allowed to cool slowly. This decreased the likelihood of producing any double-stranded DNA molecules containing two biotinylated strands. 50- to 5000-fold dilutions of the 50μg/ml oligo stock solutions were made into running buffer. The different oligos were injected at low flow rates (5-10μl/min), over individual flow cells, so as to produce a change in response of around 200-250 RU. The capture of hemimethylated 12mer is shown in the

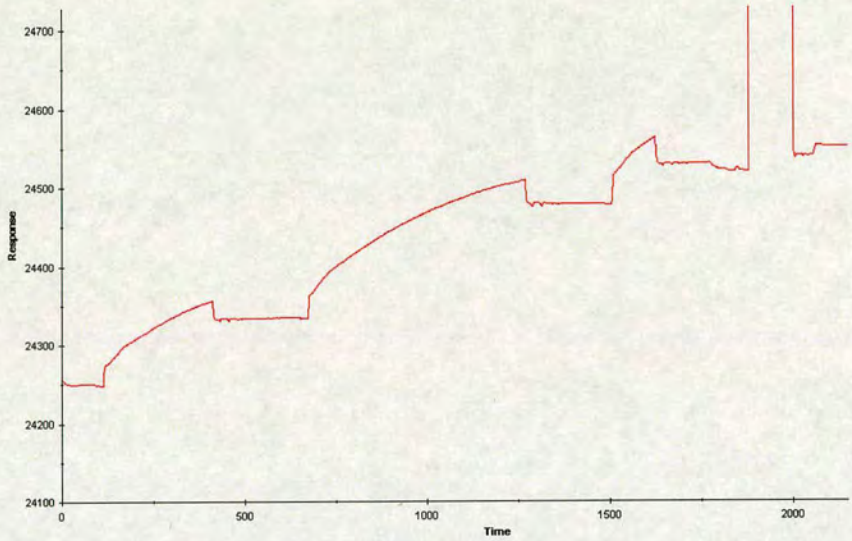

Figure 3.5: T12mer immobilisation on a SA chip. Three injections of 50ng/ml T12mer are shown: 50μl, 100μl, then 20μl, followed by a 2 minute 2M NaCl wash. The total change in response was 300 RU.

sensorgram in figure 3.5. Following immobilisation, the oligos were distributed as follows: Chip 1 (a converted CM5 chip) contained 12mer, T12mer and me10mer; Chip 2 (CM5) contained hemi12mer; and Chip 3 (SA chip), me12mer.

Chip	Chip type	Flow cell	oligomer	Amount (RU)
1	modified CM5	2	12mer	300
1	modified CM5	3	T12mer	300
1	modified CM5	4	me10mer	275
2	modified CM5	3	hemi12mer	215
3	SA	3	me12mer	245

Table 3.1: Summary of transducer chips used in the SPR experiments. Each chip contains 4 flow cells which were handled in series during the kinetics experiments. The distribution of the different oligos and the amount immobilised in each case is shown.

3.3.3 Preparation of proteins

Proteins studied were the MBDs of MBD2, MBD4 and MeCP2. Sample production was as described in Chapter 2. Protein concentrations were determined by

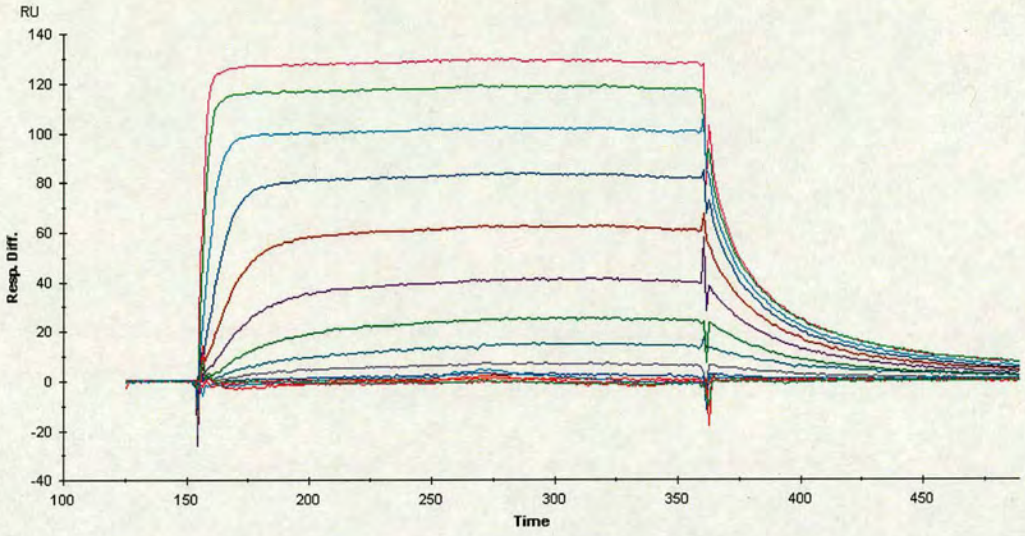

Figure 3.6: An overlay of the binding curves for MBD2 and me12mer, at the full range of concentrations used (8uM to 2nM, in a 2-fold series).

by a 4 minute dissociation phase. The "KINJECT" protocol, which resets the pumps, was used so that a continuous injection was performed. This minimises disturbances to the binding curves caused by mechanical noise from the instrument. Regeneration conditions of 100μl 2M NaCl at 100μl/min, were required to re-establish the baseline after each injection. For each protein, the full range of dilutions was injected over each chip twice, so that the binding to all 5 oligos was tested in duplicate. In each of the 3 chips, a blank reference surface in cell 1 was used to measure the bulk contribution of the protein solutions. This response was then subtracted from the other cells to give the change due to protein:DNA interaction alone. Data are recorded in the form of a sensorgram, which displays the change in response against time. A typical example is shown in figure 3.6.

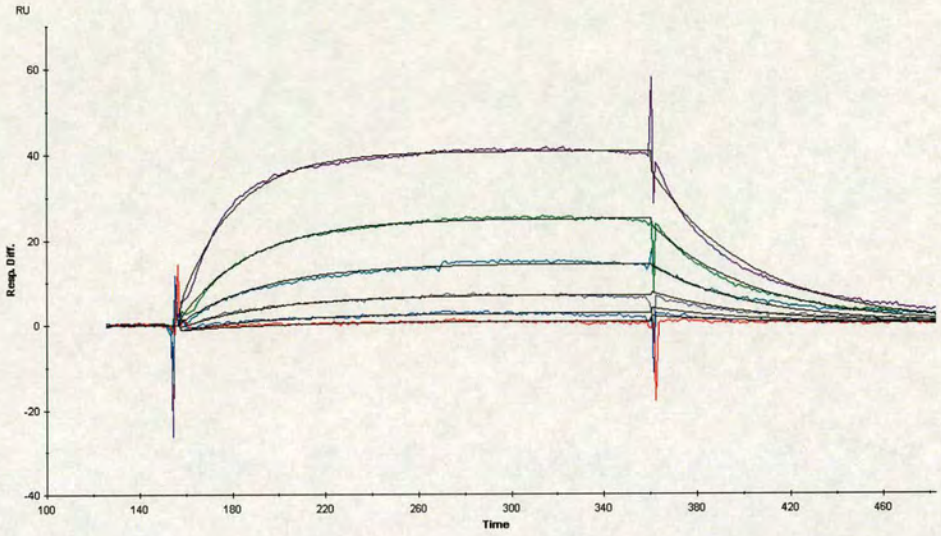

Figure 3.7: An example of curve-fitting (MBD2 and me12mer)

3.4 Data analysis

3.4.1 Curve-fitting

BIAevaluation software [118] was used for curve-fitting of kinetic data according to the supplied guidelines. MBD:DNA interactions are predicted to be 1:1, so the standard Langmuir model was used for all experiments. Simultaneous k_a/k_d fitting was carried out on a variable subset of the protein dilutions. Certain concentrations were excluded, either because the response curves were too low and below the limit of detection, or too high, with an association phase too fast to be fitted in relation to the rest of the series. For each data set, values are given for the association and disassociation rate constants, k_a and k_d; the equilibrium constants K_A and K_D; and the value of χ^2. χ^2 is a standard statistical measure of the goodness of fit, and should ideally be about the same as the noise level (2 RU), and certainly not more than 10% of the maximum response (Rmax).

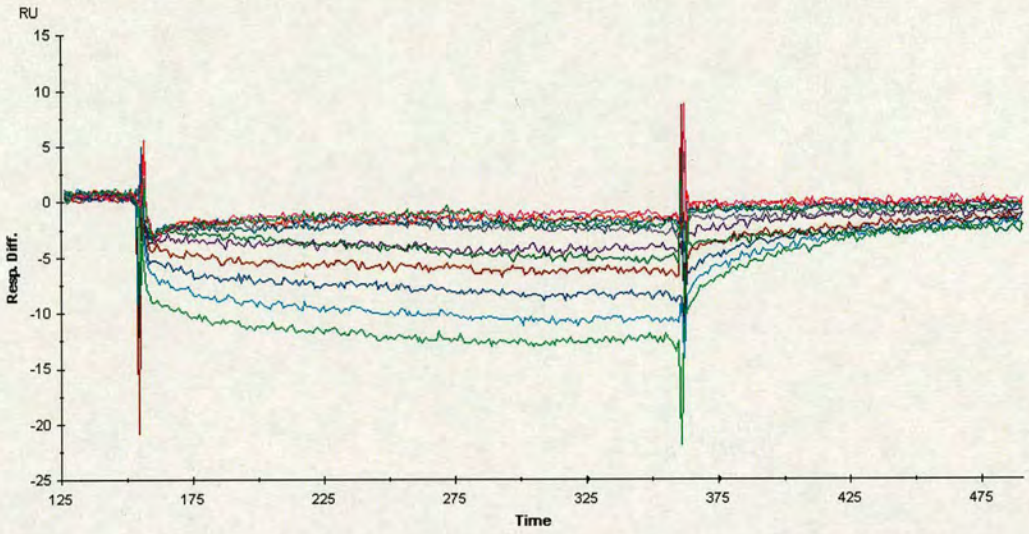

Figure 3.8: The absence of MBD2 MBD binding to 12mer, as shown by SPR binding curves.

3.5 Results

3.5.1 Comparison of MBD binding to 12mer and me12mer

There was no detectable binding between any of the MBDs and the 12mer. Sensorgrams were typically as shown for MBD2 in figure 3.8, with near-baseline response, even for the highest protein concentrations. This agrees well with previous work on the full-length proteins [41] and the MBD of MeCP2 [115] as seen by Electromobility Shift Assay (EMSA). Although preliminary studies did show a positive response on the sensorgram, this was always a typical "non-interaction" curve, as seen for MBD4 in figure 3.9 below. The completely vertical association and dissociation phases demonstrate the absence of an on- and off-rate. When these experiments were repeated, results were as for figure 3.8, and it is not clear why initially there was a response greater than the bulk interaction.

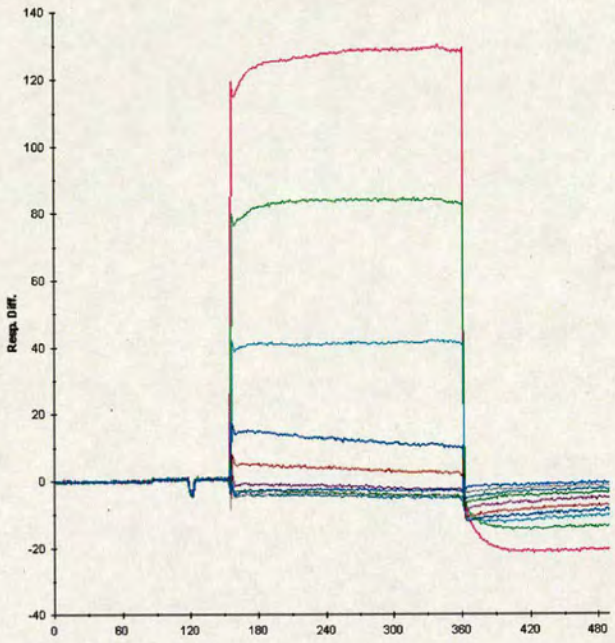

Figure 3.9: Interaction between MBD4 and 12mer

As expected, all 3 MBDs bound to the me12mer. The MBD of MBD2 bound tightly (average $K_D = 350$nM), but bound more weakly than the MBD of MeCP2. Does this reflect the ability of full-length MeCP2 to bind a single symmetrically methylated CpG pair, whereas the MeCP1 complex (of which MBD2 is a component) prefers more densely methylated sites? It may also reflect the dependence of MBD:DNA binding on elements of secondary structure outside the MBD. The MeCP2 construct used (amino acids 76-167) is 16 residues longer at the N-terminus than the MBD2 MBD. MBD4 MBD binding appears to be equally as tight as MeCP2, but the reported values are undoubtedly less reliable, due to problems with curve-fitting. In contrast to MBD2 and MeCP2, it was difficult to reproduce the fits for MBD4 and me12mer. Values obtained from these experiments are shown in table 3.2 below.

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	8.0x10^4	0.023	3.5x10^6	2.9x10^{-7}	0.38
$MBD2_2$	7.3x10^4	0.024	3.1x10^6	3.2x10^{-7}	0.28
$MeCP2_1$	1.5x10^6	0.040	3.6x10^7	2.75x10^{-8}	1.04
$MeCP2_2$	1.4x10^6	0.045	3.1x10^7	3.2x10^{-8}	0.66
$MBD4_1$	7.6x10^5	0.054	1.4x10^7	7.2x10^{-8}	0.87
$MBD4_2$	2.0x10^6	0.066	3.0x10^7	3.4x10^{-8}	0.42

Table 3.2: Analysis of binding to me12mer.

3.5.2 Effect of length on methylated DNA binding

MBD4 MBD did not bind to the me10mer, and MeCP2 MBD binding was extremely weak if present at all. A reliable analysis was not possible for these two proteins. Although the data-fit is relatively poor, MBD2 MBD does bind the shorter oligo, with a similar affinity to the me12mer. This is as expected from the structure of the MBD1 MBD in complex with a methylated 12mer DNA, where it was shown that DNA:protein contacts were confined to four central base pairs [100]. The MBDs of MBD1 and MBD2 are at least 50% identical and 70% similar. However, this also suggests that the DNA contact region in MeCP2 and MBD4 is larger than that of MBD2. This could be a feature of the subclass of MBDs containing MeCP2 andMBD4, which have greater similarity in their MBDs compared to those of MBD1-3 [41].

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	2.4x10^4	0.02	1.2x10^6	8.2x10^{-7}	10.50
$MBD2_2$	2.1x10^4	0.02	1.0x10^6	9.9x10^{-7}	7.85

Table 3.3: Analysis of binding to me10mer.

3.5.3 Comparison of me12mer, hemi12mer and T12mer binding

An interaction was observed between all proteins and each of the oligos that contain one or more methylated cytosines. MBD2 has the weakest MBD:DNA interaction, about an order of magnitude lower than those of MeCP2 and MBD4, which have similar $K_D s$. MBD2 MBD has a moderate preference for the me12mer and hemi12mer over the T12mer (see table 3.4), which suggests that a protein:base contact may be disrupted by the replacement of the amine group with oxygen. However, there is nothing in the MBD1 MBD:DNA structure to indicate that such a contact exists. MeCP2 on the other hand, moderately prefers the hemi12mer over the other two oligomers.

Oligo	$K_A\ (M^{-1})$	$K_D\ (M)$
me12mer	3.3x10^6	3.1x10^{-7}
hemi12mer	3.1x10^6	3.2x10^{-7}
T12mer	2.3x10^6	4.4x10^{-7}

Table 3.4: Average values of equilibrium constants for MBD2 and the various oligos.

Oligo	$K_A\ (M^{-1})$	$K_D\ (M)$
me12mer	3.4x10^7	3.0x10^{-8}
hemi12mer	7.2x10^7	1.4x10^{-8}
T12mer	3.1x10^7	3.3x10^{-8}

Table 3.5: Average values of equilibrium constants for MeCP2 and the various oligos.

An average value for MBD4 and me12mer is given, but the experiments were not as reproducible as for MBD2 and MeCP2. One experiment gave very similar

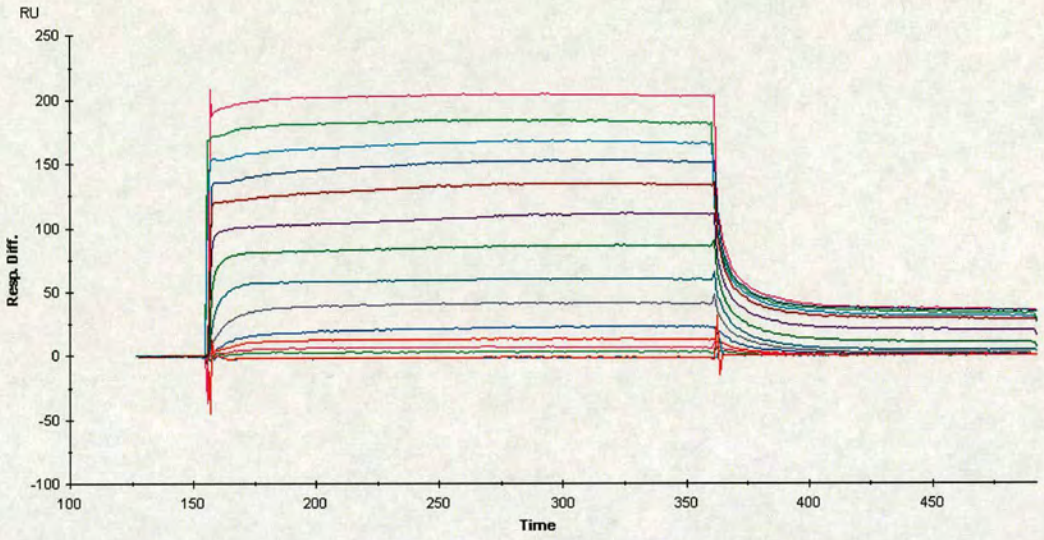

Figure 3.10: The MBD4 and me12mer binding curve.

values to those for the hemi12mer and T12mer, so it's possible that MBD4 has a similar affinity for all 3 oligos. These results are slightly dubious however, as the interaction may be too fast to measure accurately at this protein concentration.

For the T12mer, although the equilibrium constants for MBD4 and MeCP2 are similar, MBD4 has a much faster on and off rate. A faster turnover could perhaps be expected due to it's function as a mismatch repair enzyme [56].

Oligo	$K_A\ (M^{-1})$	$K_D\ (M)$
me12mer	2.2x10^7	5.3x10^{-8}
hemi12mer	3.0x10^7	3.4x10^{-8}
T12mer	3.0x10^7	3.0x10^{-8}

Table 3.6: Average values of equilibrium constants for MBD4 and the various oligos.

3.6 Discussion

In this study, the MBD2 MBD was found to have the weakest affinity for a methylated oligo. Previous attempts to quantify the MBD:DNA interaction have all employed non-equilibrium systems and much longer oligos, sometimes with multiple methylation sites. By Reverse Capillary Electrophoretic Mobiliy Shift Assay (R-CEMSA), mouse MBD2b was found to have a $R_{1/2}$ value of 1-8nM for methylated oligos (45-50 bp in length). This is compared to a value of 120-192nM for MeCP2. Of all the MBDs tested, MBD2b was also found to have the greatest capacity to differentiate between methylated and unmethylated oligos [119]. The apparent discrepancies between the SPR and R-CEMSA data can best be explained by a model in which the MBD is the foundation of DNA-binding, but not the discriminative domain. Specificity may be conferred by regions of the protein outside the MBD.

By SPR, there is no significant difference in the affinity of a particular MBD, for any of the methylated oligos. MBD2, however, is able to resist being competed off a single mCpG, by a 100-fold excess of hemimethylated oligo, whereas MBD4 (and MBD1) could not [120]. This is peculiar due to the presence of multiple CpGs in the competitor - MBD2 is known to have a preference for densely methylated regions [41].

It has also been observed that by R-CEMSA, all the MBDs bind non-specifically to unmethylated DNA [119]. In gel-retardation assays, however, the MBD proteins form a complex with methylated DNA, but not with an unmethylated version of the same probe [120]. It is not clear if what has been observed in this study with an unmethylated oligo, is non-specific binding or non-interaction.

While preliminary studies seemed to produce a greater than 'bulk' interaction, these results were not reproducible.

3.6.1 Result significance

Estimate of error in this SPR study is given by χ^2, a deviation from idealised fit. All χ^2 values were comparable to the noise level and below 10% of the maximum response, Rmax. The vast majority were in fact below 1% of Rmax, making the estimates of equilibrium constants as reliable as possible by this method. It was however, not possible to translate χ^2 into an error in nM, to allow better comparison of the binding constants.

The MeCP2 construct is 20 residues longer than the MBD2 construct. This could have prevented it from binding the me10mer, due to steric hinderance with the chip matrix. The MBD4 construct is also slightly longer than MBD2, by a few residues, but it seems unlikely that this could have the same effect.

3.6.2 Future investigations

Given that there is so much conficting evidence about the affinity of the different MBD proteins for methylated DNA, a further SPR study would be invaluable. The most interesting question concerns the relative affinity of the MBD (domains) for DNA, compared to the full-length proteins in which they are found. Sequence selectivity may also play a role, as MeCP2 has been shown to prefer DNA which has a methylated CpG flanked by an AT rich region [50]. The sequence of the oligos used in this study may have resulted in increased MeCP2 affinity (see sequences in figure 3.4) compared with previous studies where adjacent ATs were

not present. An extensive study examining length, sequence and protein construct dependence of meCpG recognition, is required.

Chapter 4

NMR

4.1 Introduction

Nuclear Magnetic Resonance (NMR) is a technique that is mainly used to gain information about molecules in solution. It has become extremely useful for work on macromolecules due to advances in technology and understanding over the past 25 years. NMR exploits an intrinsic property of certain atomic nuclei, called spin.[1]. The spin state of a nucleus in an applied, external magnetic field, can assume one of two possible orientations. The difference between the total number of spins in each state (the equilibrium population difference), is the basis for the NMR experiment.

NMR experiments involve applying a series of radiofrequency (RF) pulses, described in a pulse program, to a sample in a magnetic field. The type of pulse program determines what information is obtained from the sample, but in each case the pulses act to disturb the population difference. As the system returns

[1]Only nuclei with a non-zero spin quantum number are *NMR active* and give NMR spectra. Examples include ^{1}H, ^{15}N, ^{13}C and ^{31}P.

to equilibrium, the NMR signal is recorded, which is then digitised, transformed and processed to generate a spectrum for analysis.

4.1.1 Protein structure determination by NMR

NMR spectra of macromolecules such as proteins are extremely complicated, due to the large number of nuclei in similar, but non-identical chemical environments. However, multidimensional heteronuclear techniques (explained below) resolve these signals to give information about the identity of each nucleus and its relative location in space. This information is used to create a list of interatomic distances in a molecule, which provides an input for structure calculations. Bond-angle information can be included by determining through-bond coupling constants, and the structure may be further refined using residual dipolar couplings (RDCs)[2]. A computational approach is used to find possible configurations of atoms, which match the information collected in the NMR experiments: NMR is thus an indirect technique for determining protein structure. A schematic diagram to illustrate the process is shown in figure 4.1.

4.2 NMR experiments

NMR experiments were collected on Bruker Avance 800 MHz, and 600 MHz spectrometers, both with triple resonance (TXI) probes with triple axis gradients. Experiments were run using Bruker Xwinnmr 3.6 software, and standard Bruker pulse programs, modified by Dr. Dušan Uhrín. The different types of experiment

[2]RDCs depend upon the relative orientation of bond vectors with respect to a common axis. They are useful because they provide information about long-range order [121].

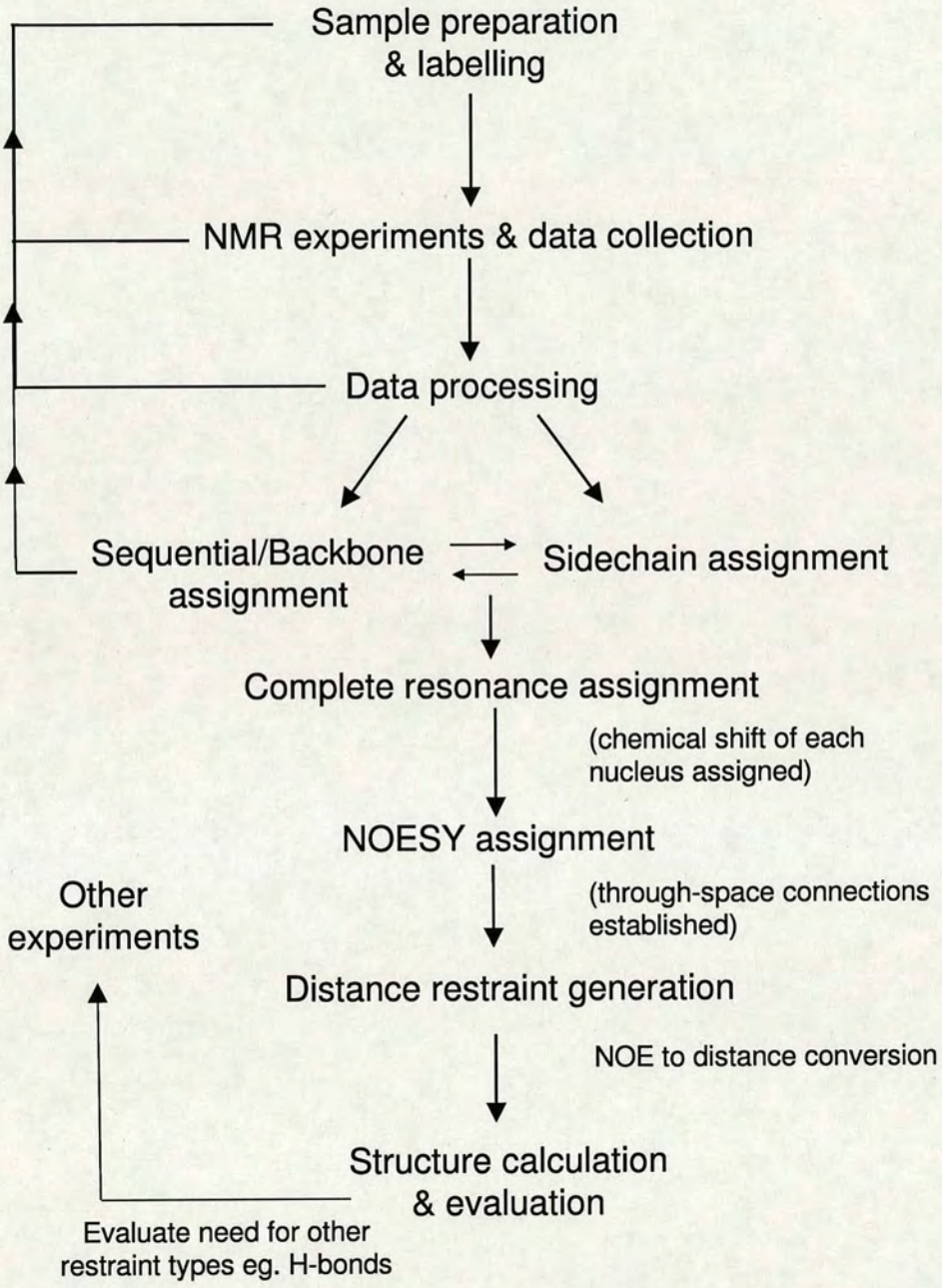

Figure 4.1: A schematic diagram to illustrate the process of protein structure determination by NMR. Each stage is described more fully in this and the following chapter.

and their use, are outlined below.

4.2.1 Homonuclear experiments

The most useful, and fortunately also the most abundant, nucleus in protein NMR is the proton (^{1}H). Homonuclear experiments collect information from protons only, but this is not as limited as it might seem. Protons are affected by their local chemical environment which is apparent by a specific *shift* of their position in the NMR spectrum. Chemical shifts arise due to local variations in the magnetic field experienced by an individual nucleus. In general, a nucleus is shielded against the effects of the external magnetic field by its orbiting electrons, and is deshielded if those same electrons are drawn away (for example by an electronegative neighbouring atom). Shielding corresponds to a decrease, and deshielding to an increase in the frequency at which a particular nucleus resonates.

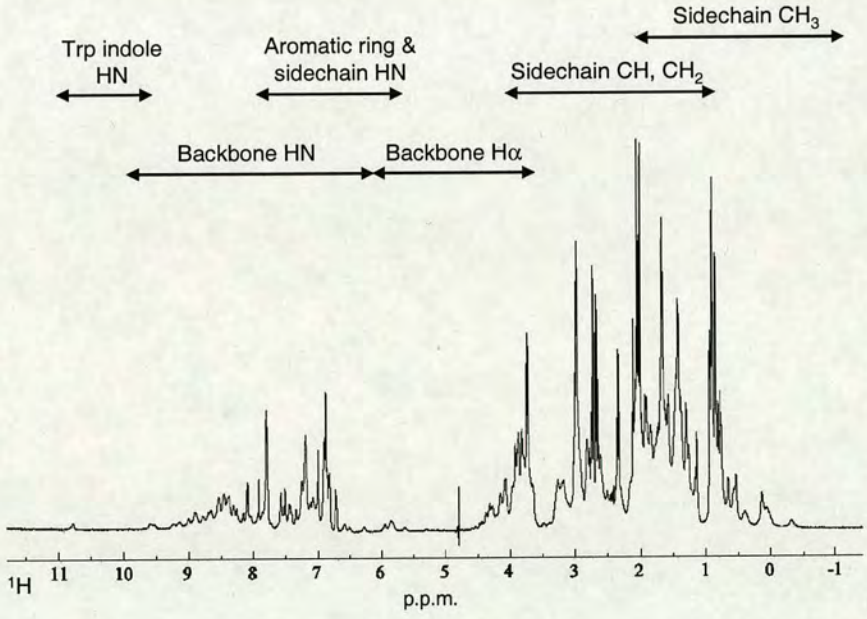

Figure 4.2: A 1D proton spectrum of an unlabelled sample of the MBD2 MBD. Regions of the spectrum are labelled to show where particular chemical types are found.

4.2.2 1D ^{1}H experiment

The one dimensional ^{1}H experiment produces a spectrum displaying the relative frequency, expressed in parts per million (p.p.m.), of every proton in the sample. Although there is a large variation in the local chemical environment experienced by each proton, particular chemical types are found in certain areas of the spectrum (see figure 4.2). This spectrum is very overlapped and cannot be used for resonance assignment, since there is no way of identfying individual atoms. However, it is useful for examining the general quality of the sample, as small-molecule contaminants and protein unfolding can be detected by a 1D experiment.

4.2.3 Multidimensional experiments

Due to the complexity of 1D protein NMR spectra, it is necessary to spread signals out over two or more dimensions. In a multidimensional experiment, magnetisation is transferred between nuclei at the end of an *evolution time* of variable length. The resulting spectrum is a function of multiple frequency variables, and gives information about nuclei that are coupled, either through bonds (scalar, or J-coupling) or through space (dipolar coupling). Heteronuclei (nuclei other than ^{1}H) can also be studied in these experiments. This is discussed in section 4.2.5.

4.2.4 2D homonuclear experiments

Several different types of 2D homonuclear experiment were used in this project, and an overlay of two experiments is shown in figure 4.3. In a 2D correlated spectroscopy (COSY) experiment, magnetisation is transferred between protons that are no more than three bonds apart. On-diagonal peaks represent magneti-

sation which has not transferred to a second nucleus, whereas so-called crosspeaks appear at the chemical shifts of paired nuclei. COSY crosspeaks are useful for identifying aromatic ring protons and protons in longer sidechains. The total correlated spectroscopy (TOCSY) experiment is similar to COSY, except that the scalar coupling is not limited to three bonds, but may correlate all protons in a spin-system. An example of a spin system is a single amino acid residue, although overlap often prevents assignment of all the atoms. This is a particular problem with Hδ and Hϵ, in lysine and arginine sidechains.

The NOE spectroscopy (NOESY) experiment correlates all protons within a certain distance of each other. This is possible due to the nuclear Overhauser effect (NOE). The *mixing-time* must be carefully chosen to allow long enough for magnetisation transfer to take place whilst still retaining signal intensity, before it decays due to spin diffusion[3].The size of an NOE crosspeak is approximately inversely proportional to the sixth power of the distance between the nuclei. Therefore, NOESY crosspeaks are useful for both identifying individual atoms, and determining their relative location in space. This property forms the basis of structure determination by NMR.

In this project, 2D homonuclear experiments were run in D_2O to prevent a strong water signal obscuring part of the spectrum. The H_α region (3.5 - 6 p.p.m.) is particularly vulnerable to this effect, due to water protons resonating at around 4.8 p.p.m. The problem with samples in D_2O, is that labile protons undergo chemical exchange with the solvent, and so become invisible in a ^1H experiment. Fortunately, the purpose of these experiments is to identify stable

[3]A leakage of magnetisation to surrounding nuclei

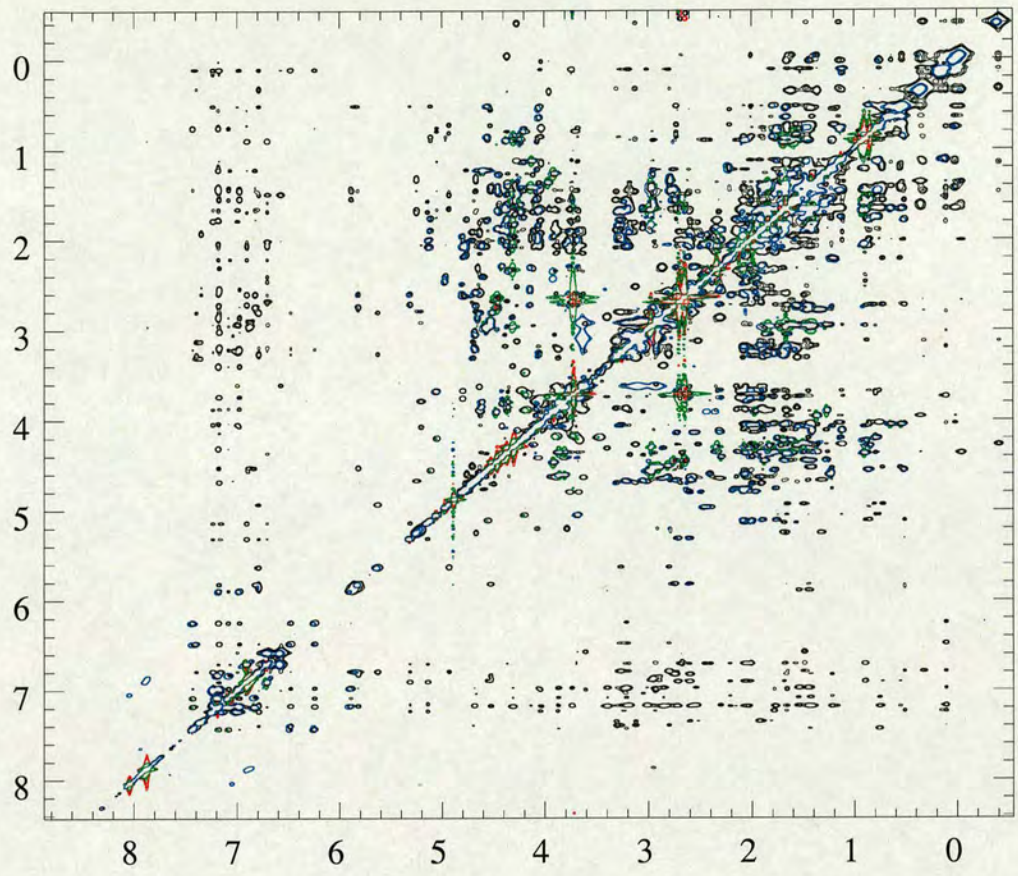

Figure 4.3: An overlay of 2D homonuclear spectra of the MBD2 MBD in D_2O, at 288K. Positive and negative contours are shown in black and red (NOESY), and blue and green (TOCSY) respectively. Both axes display ^1H shifts in p.p.m.

aliphatic and aromatic, rather than amide protons, for which other experiments are used.

4.2.5 Multidimensional heteronuclear experiments

In protein NMR, ^1H is not the only magnetically active nucleus. The heteronuclei ^{15}N and ^{13}C can be used, although they have a low natural abundance. This means that samples have to be isotopically enriched for one or both isotopes (see Chapter 2). They also have much smaller gyromagnetic ratios than ^1H,

which results in low detection sensitivity[4]. This latter problem is minimised by having heteronuclei as intermediates in the magnetisation transfer process, while ^{1}H remains the nucleus from which magnetisation originates and on which signal detection takes place.

4.2.6 The HSQC experiment

A very important 2D experiment is the heteronuclear single quantum correlation (HSQC) experiment. In the ^{15}N-HSQC, nitrogen atoms are correlated with their directly attached protons (if they have them). Each HSQC peak corresponds to one incidence of an NH, which are predominantly found in the protein backbone. Side-chain NHs, found in arginine, histidine and tryptophan, and NH_2 groups of arginine, asparagine snd glutamine may also be seen. There is no diagonal in an HSQC spectrum, as different nucleus types with very different frequency ranges, are observed and correlated. A similar experiment, the ^{13}C-HSQC, correlates all carbon-attached protons with their corresponding carbon atom.

4.2.7 3D ^{15}N experiments

3D ^{15}N experiments are a combination of a 2D homonuclear experiment and the ^{15}N HSQC. The purpose is to resolve signals that were overlapped in the 2D experiment, by correlating them with a third heteronuclear dimension. In the TOCSY version, scalar couplings, from an NH to protons in the same residue are observed. The ^{15}N NOESY is used to correlate NH with all the protons close in space. Together, these experiments allow sequential assignment of the protein

[4]The NMR frequency, ν, of a nucleus in a fixed magnetic field, is directly proportional to its gyromagnetic ratio, γ.

sequence (described in section 4.5.2), and the NOESY is also used to derive interatomic distances (see Chapter 5). A schematic diagram of the 3D experiment is shown in figure 4.4.

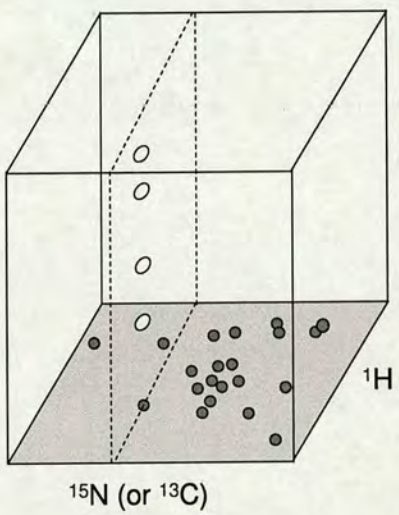

Figure 4.4: A schematic representation of 3D experiments with an HSQC step. The ^{15}N HSQC face of the cube is shown in grey, and a ^{1}H-^{1}H slice with four crosspeaks (in white) is indicated. The slice is through a ^{15}N HSQC peak and so is called a ^{15}*N plane* of the experiment. ^{15}N planes are used to identify amino acid types (from the shifts of the sidechain nuclei) and establish sequential connectivities. ^{13}C can also be used as the third resolving dimension, for instance in the HCCH-TOCSY experiment (see section 4.2.8).

4.2.8 HCCH-TOCSY and ^{13}C NOESY

In the HCCH-TOCSY experiment, magnetisation is first transferred from a proton to its directly attached carbon. This is followed by a fast TOCSY transfer to neighbouring carbons, and finally to *their* directly attached protons, from which detection takes place. It is an extremely useful experiment, since every ^{1}H-^{13}C pair is correlated with every other proton (and hence its attached carbon), in the same residue. A complete assignment of each spin system can be achieved by

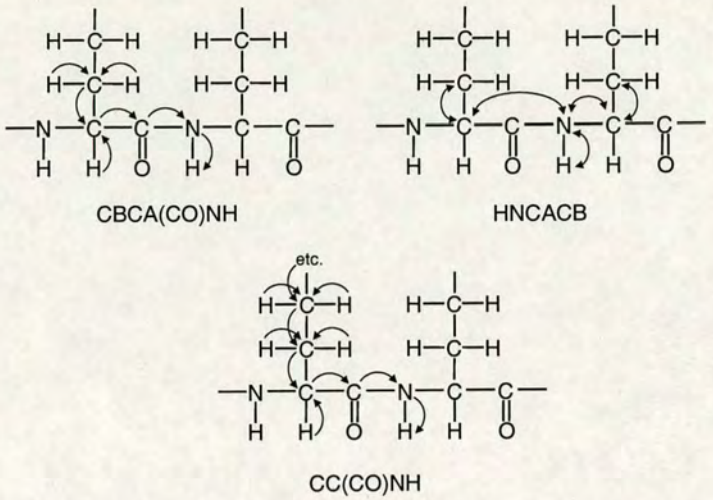

Figure 4.5: Magnetisation pathways of three triple resonance experiments, which give rise to the experiment names.

comparing crosspeaks in related ^{13}C planes. This information cannot, however, be directly linked to the corresponding backbone amide shift, since the experiment lacks a ^{15}N dimension. Instead, by working out likely residue types from crosspeak patterns, and by comparing chemical shift values obtained in other experiments, each spin system is assigned to a particular residue. The ^{13}C NOESY is analogous to the HCCH-TOCSY experiment, except that the correlations are not limited to through-bond couplings. The spectrum strip belonging to each ^{1}H-^{13}C pair will contain a NOESY crosspeak for every dipolar coupled proton.

4.2.9 Triple resonance experiments

Triple resonance experiments are used for identifying sidechain ^{13}C shifts, correlating these with backbone amide shifts and thus sequentially assigning the protein chain. Each of the experiments is named after the magnetisation pathway it employs to probe the protein molecule; undetected nuclei are shown in

Experiment	Solvent	Mixing time	Processing	T (K)	Ref.
2D ^{15}N HSQC	H_2O	-	FT	288	[122][123]
2D ^{13}C HSQC	H_2O	-	FT	288	[124]
2D ^{1}H NOESY	D_2O	100 ms	FT	288	[125]
2D ^{1}H TOCSY	D_2O	60 ms	FT	288	[125]
2D ^{1}H COSY	D_2O	-	FT	288	[125]
3D ^{15}N NOESY	H_2O	100 ms	MEM	288	[126][123]
3D HCCH-TOCSY	H_2O	15 ms	MEM	288	[127]
3D ^{13}C NOESY	H_2O	100 ms	MEM	288	[128]
3D CBCA(CO)NH	H_2O	-	MEM	298	[129][130]
3D HNCACB	H_2O	-	MEM	298	[129][131]
3D CC(CO)NH	H_2O	-	MEM	298	[132][133]
3D ^{15}N NOESY	H_2O	100 ms	MEM	298	[126][123]
3D ^{15}N TOCSY	H_2O	35 & 60 ms	MEM	298	[123]

Table 4.1: A summary of experiments collected on the MBD2 MBD. Solvent described as H_2O is actually 90% H_2O, 10% D_2O; T is temperature in Kelvin. For a general introduction to the various NMR experiments, see [125].

brackets. The CBCA(CO)NH experiment correlates each NH with the C_α and C_β carbons of the previous residue. HNCACB, an *out-and-back* experiment, correlates each NH with C_α and C_β from both its own spin system and that of the preceding residue. CC(CO)NH is designed to correlate NH with the C_α and all other ^{13}C sidechain resonances of the previous residue. The magnetisation pathways are shown in figure 4.5. A description of how these experiments were used for sequential assignment is given in section 4.5.2.

4.3 Data acquisition

In one model of NMR, magnetisation is described as a vector, rotating about the field direction (Z axis) at an angular velocity proportional to the field strength. In an NMR experiment, the total field is not stationary and also rotates about the Z axis. By rotating at the same freqency as the field, so that the field appears

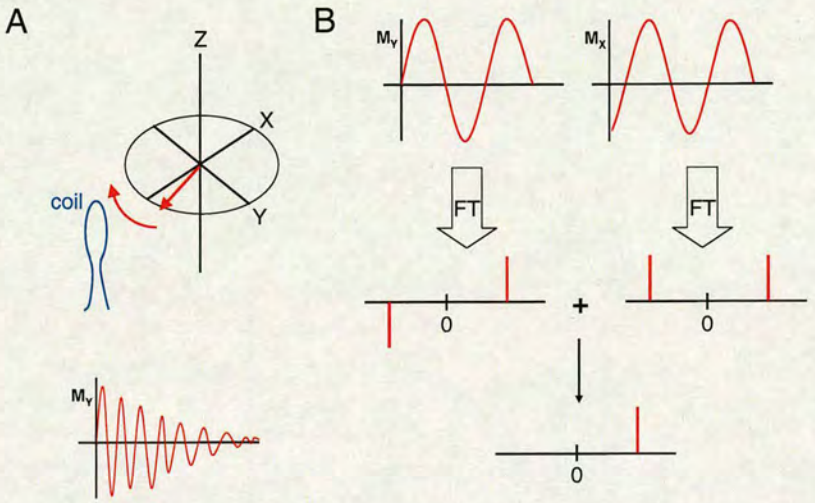

Figure 4.6: Detection of the NMR signal. A:The magnetisation induces a decaying sine wave in the receiver coil. B:Quadrature detection allows the sign of the frequency to be determined, by summing frequency-domain data from orthogonal axes.

to be stationary, we can more easily visualise the behaviour of the magnetisation vector[5]. At the end of an NMR experiment, the magnetisation components, are flipped into the XY plane where they induce a current in the receiver coil, located on the X-axis. This gives rise to a decaying sine wave, called a free induction decay (FID), the NMR signal. The sine wave is a result of plotting the current induced in the coil as a function of time, and it decays as the spins gradually dephase (due to relaxation).

In reality, it is not possible to detect the direction of rotation of the vector using only one receiver. This is because there are two possible Fourier transforms of the sine wave. A second detector for the imaginary component, allows positive and negative vector rotation to be distinguished, by summing real and imaginary components. This is known as quadrature detection. Since the transmitter frequency can then be placed at the centre of the spectrum, and the digital filters

[5]Magnetisation is thus observed in the *rotating-frame* of reference.

set to a narrower range, less noise is folded back into the spectrum, improving the signal-to-noise ratio (SNR).

4.3.1 Data improvement by FID manipulation

Due to experimental time limitations, a number of processes are used to gain maximum quality and resolution from data that are otherwise limited.

Zero-filling is the process of increasing the digital resolution of a spectrum by adding null data to the FID. Resolution is proportional to the density of sampling the data and the interval between sample points is inversely proportional to the acquisition time. So, by artificially increasing the size of the data array, enhanced digital resolution is possible without increasing the acquisition time (or decreasing the sweepwidth). This is only valid if the FID does actually decay to zero, and in reality there isn't time to wait until this eventually happens. Instead, the FID is multiplied by a window function to cause the FID to decay smoothly to zero - this is called apodisation. There are different types of window function, which apply a weighting to different parts of the FID. Gaussian and sinebell squared are common functions, which serve to improve resolution and maintain a good signal-to-noise ratio (SNR), without introducing excessive line-broadening.

4.4 Data processing

4.4.1 Processing methods

The FID is a composite of many signals, of varying amplitude and frequency. The individual signals may be extracted from the FID using several methods,

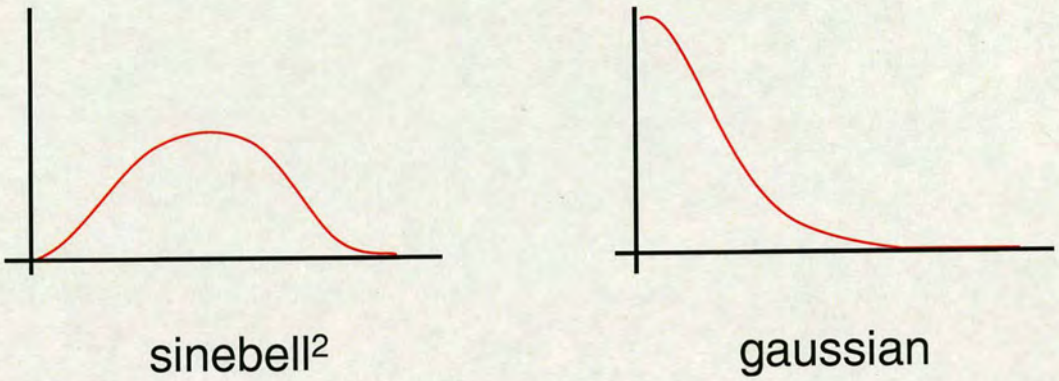

Figure 4.7: Examples of window functions used in this project. The FID is multiplied by a window function before Fourier Transformation, in order to make the signal decay smoothly to zero. Zero-filling may then be used to improve the digital resolution of the spectrum.

two of which were used in this project. Fourier Transformation (FT) converts the time-domain FID to a freqency-domain spectrum with both real and imaginary components. Subsequent phasing of the data reduces the contribution of the imaginary part, so that only the real (pure absorption Lorentzian) lines are seen in the spectrum. The Maximum Entropy Method (MEM) works by reconstructing FIDs from simulated frequency domain spectra, by inverse FT. Because there are numerous spectra which could be transformed in this way to generate the same FID, the method selects the simplest solution, that with the maximum entropy.

4.4.2 FT processing

The AZARA suite of programs [134] was used for all NMR data processing. A script used to process a 2D experiment (a ^{15}N HSQC), by the FT method, is shown in Appendix B. First, the directly detected (^{1}H) dimension was processed (with the program *process*), and the 1D spectrum of the first transient plane was viewed with *plot2*. After appropriate phase correction, the second (^{15}N)

dimension is processed and the resulting 2D spectrum viewed. After further phase and baseline corrections, the spectrum is referenced and an appropriate contour threshold level selected.

4.4.3 Maximum entropy method processing

All 3D experiments in this project were processed with the MEM in the indirectly detected dimensions. This is advantageous since the digital resolution in indirect dimensions is often limited by the time available for data acquisition. Data are first processed by the FT method, then commands for MEM processing are added into the script for the relevant dimensions (see example in Appendix B). Spectral noise is estimated by analysis of 2D planes of each dimension, using *plot2*. Individual planes are processed and the result compared with FT processed data. The aim is to improve the resolution of the spectrum, without loss of information. Processing the whole spectrum with MEM is very computationally intensive, which is why it is first optimised on a number of 2D planes. After processing the entire spectrum, the log files are checked, to see if all the 2D planes have converged. A program written to do this is shown in Appendix B. MEM processing often completely failed on planes dominated by water signal.

4.4.4 Contour generation and display

The AZARA [134] program *contours* is used to generate contour levels, both positive and negative starting from the contour threshold level used in *plot2*. A *contours* input script is shown in Appendix B. The output is read, along with the experimental parameter file, by ANSIG, an assignment program [135]. ANSIG

allows easy navigation of multiple spectra, keeps track of assignments, and has a number of associated macros. These simplify the assignment process, and subsequent data manipulations such as crosspeak integration.

4.5 Assignment

4.5.1 Overview of assignment strategy

The experiments described in the previous section, were used to carry out the process of assignment. This involves assigning each particular resonance within the spectra, to a particular atom in the molecule. Firstly, in protein NMR, a limited set of (mainly backbone) assignments are used to determine sequential connectivity. Next, all remaining atoms in the simpler sidechains are assigned. Finally, aromatics and longer sidechains are assigned. Once all the assignments have been made, these are used to assign NOESY crosspeaks, which represent through-space connections. From these, distance restraints are derived for structure calculation.

4.5.2 Sequential assignment

Sequential connectivities may be determined using either 3D ^{15}N data, ^{13}C data or both. Starting from a ^{15}N HSQC crosspeak, the corresponding ^{15}N plane of the 3D data (see figure 4.4) is examined. A ^{15}N-^{1}H-^{1}H TOCSY experiment provides information about atoms within a residue, and the corresponding NOESY experiment about atoms that are close in space, including *inter*-residue proximities. Therefore, NOESY crosspeaks matching those in the TOCSY experiment,

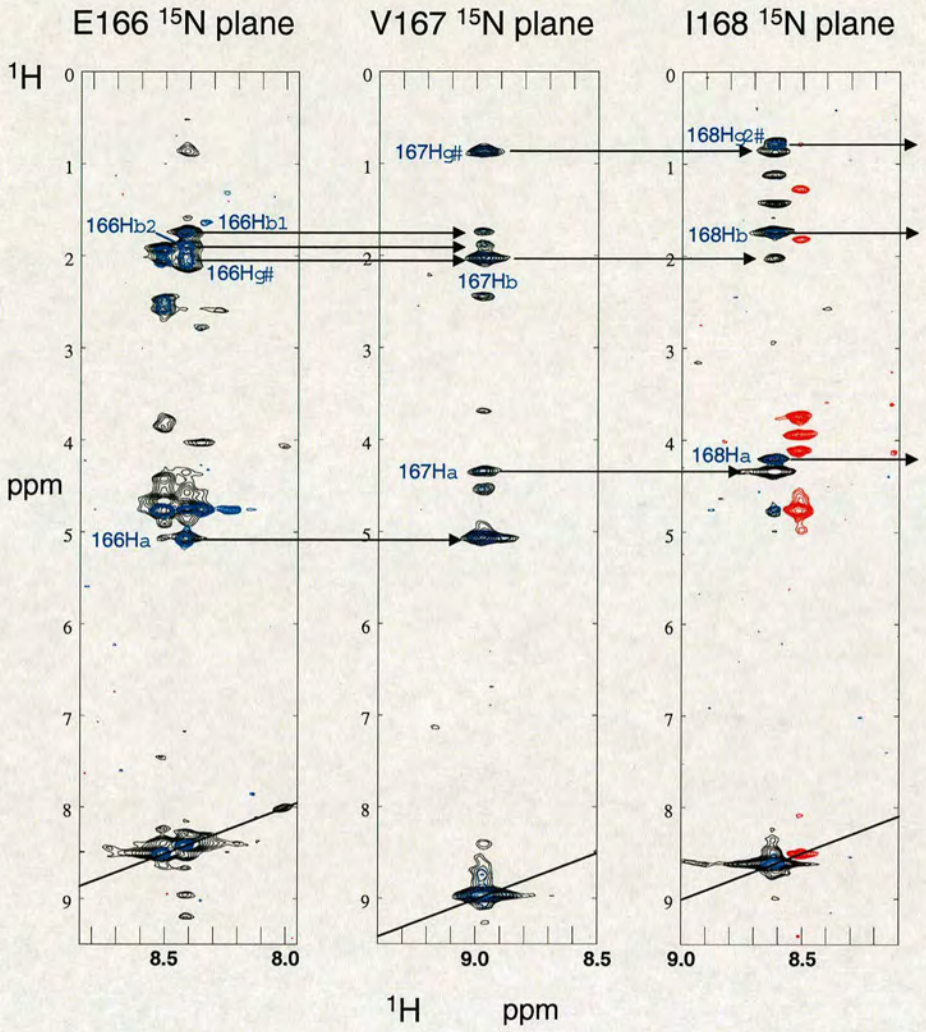

Figure 4.8: Three ^{15}N slices of ^{15}N-NOESY (black and red) and ^{15}N-TOCSY (blue and green) spectra, showing how the data are used for sequential assignment. Some assignments shown were only determined in conjunction with other experiments.

---

will be present in the ^{15}N plane of the next residue. This procedure may also be done in reverse: by searching for NOESY-matching TOCSY crosspeaks, in the ^{15}N plane belonging to the previous residue. Thus it is possible to string together the amino acid chain (see figure 4.8), until a proline is reached or the information is limited. Proline inevitably breaks up this linking process as it lacks an NH. Matching peaks may also be found in a number of different ^{15}N planes, so connections must be worked out by a process of elimination. For MBD2 residues

172, 186 and 207-215, no proton chemical shift information was obtained using the 3D ^{15}N experiments.

To complement the procedure described above, it is possible to use a similar principle to sequentially assign a protein chain using ^{13}C resonances (see figure 4.9). The experiments required are CBCA(CO)NH and HNCACB (see figure 4.5 and earlier description). CBCA(CO)NH relates the NH resonance of residue i, to the Cα and Cβ of residue i-1. HNCACB correlates Cα and Cβ to NH of the same residue, i. Sequential residues in the chain can therefore be connected by finding CBCA(CO)NH crosspeaks that match HNCACB peaks at another NH resonance, and vice-versa. Cα and β crosspeaks in the HNCACB experiment have the opposite sign to each other, so this identifies which shift is which. In both cases, comparison of ^{1}H and ^{13}C shifts with a database of average chemical shifts, and information from the CC(CO)NH experiment, give clues to the identity of the residue. Although several residues have similar patterns of crosspeaks (eg. aromatic residues), chains of linked residues may be identified by comparison with the amino acid sequence.

4.5.3 Side-chain assignment

Carbon side-chain assignments, other than Cα and Cβ are initially obtained from the CCCONH experiment, whilst some proton resonances are found in the 3D ^{15}N-NOESY and -TOCSY experiments. Detailed, systematic assignment however, is obtained by analysis of the HCCH-TOCSY experiment. This correlates all carbon-attached protons within a residue, and each C-H connection is tracked using a ^{13}C HSQC (or a 2D, ^{13}C-^{1}H, projection of the ^{13}C NOESY). As well as

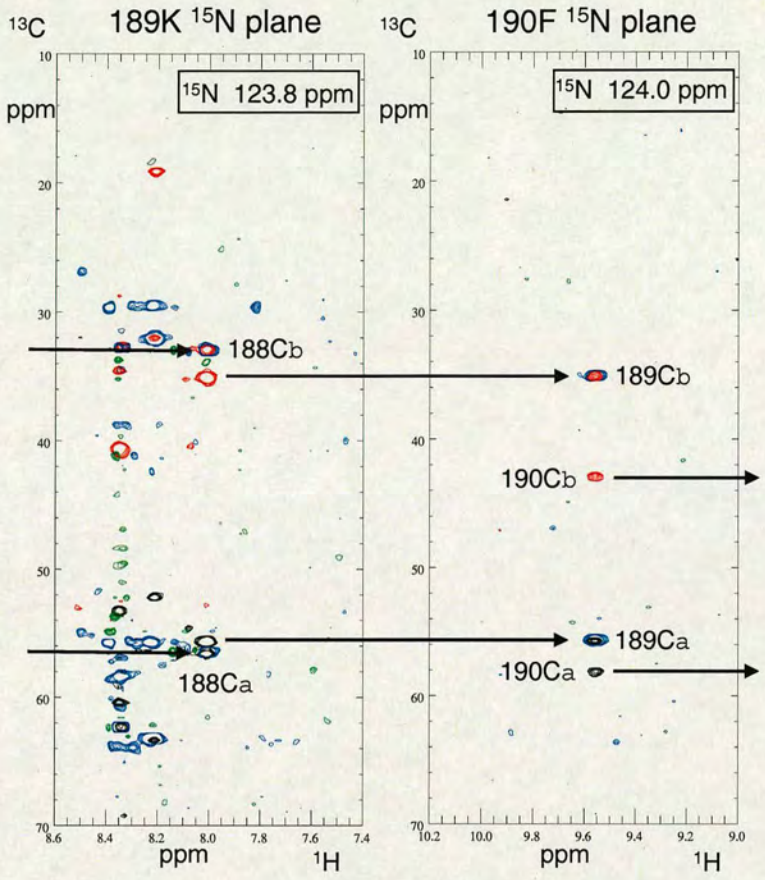

Figure 4.9: Two ^{15}N slices of triple resonance spectra, showing how the data are used for sequential assignment of the protein chain. CBCA(CO)NH contours are shown in blue (positive) and green (negative, folded), whilst HNCACB contours are in black and red respectively. Some slices (such as the F190) are easy to interpret, whilst K189 has an NH resonance in an overlapped region of the spectrum.

---

identifying previously unassigned atoms, the HCCH-TOCSY experiment discriminates between protons attached to the same carbon atom, and those attached to different carbons; a detail not found in ^{15}N experiments. Other experiments used to cross-check assignments are homonuclear NOESY, TOCSY and COSY experiments in D_2O. These are useful, particularly for identifying aromatic proton resonances and for Hαs close to the water signal.

Temperature differences

The first set of experiments collected on the MBD2 MBD were done at 298K, whilst subsequently it was discovered that the stability of the protein improved at 288K (see Chapter 2). Although all experiments to provide restraints for structure calculation were obtained at the same temperature (288K), some assignment experiments were carried out at 298K. The temperature difference of 10K caused a variable change in the chemical shifts found in backbone assignment experiments (^{13}C experiments and one set of ^{15}N experiments). Proton shifts are more sensitive to local environment changes induced by temperature, due to the effect on H-bonding and water exchange rates. This problem was resolved by transfering proton assignments from a ^{15}N NOESY at 298K, to a ^{15}N NOESY experiment at 288K.

4.5.4 NOESY assignment

NOESY experiments help to establish sequential connectivities, identify residues and determine proline isomer types. Once the assignment process was completed, additional crosspeaks, corresponding to through-space connections, were picked in the ^{13}C- and ^{15}N-NOESY experiments. This was done by carefully considering all possible contributions to a particular crosspeak. For each possible contribution, the symmetry-related peak (in another spectrum, or another part of the same spectrum) was sought. If more than one pair of atoms could be contributing to the crosspeak, or there was any doubt about the authenticity of the assignment, the crosspeak was left unassigned. Such crosspeaks are "ambiguous" and are handled in a different way to crosspeaks whose assignment is unambiguous. It

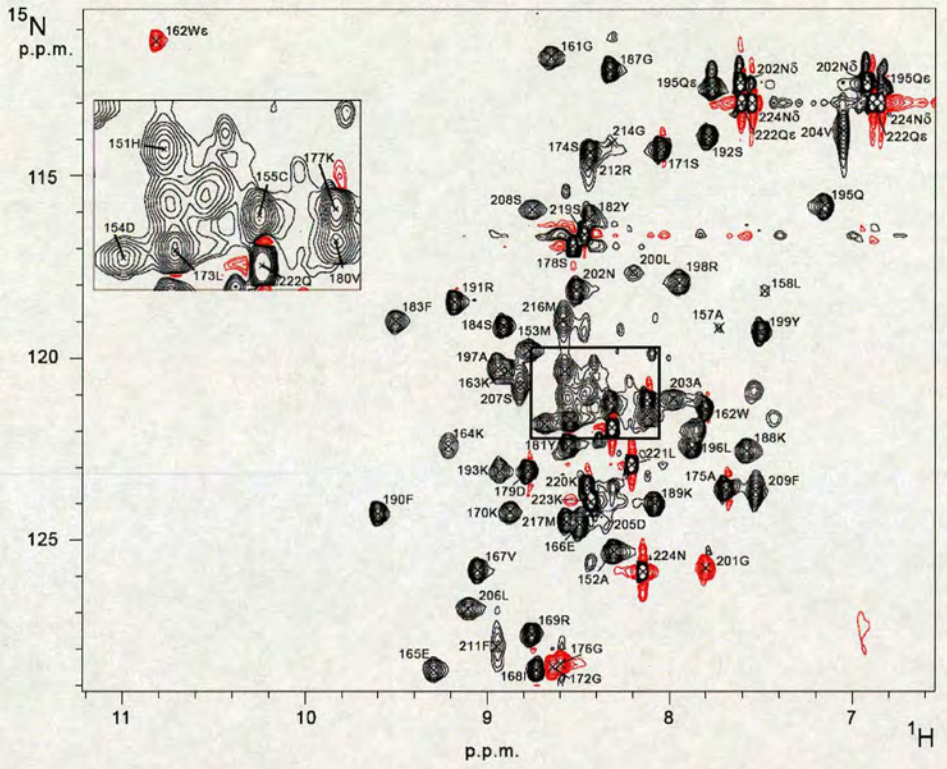

Figure 4.10: A summary of the assignment of the MBD2 MBD at 288K. Each NH crosspeak is labelled with the single letter amino acid code and sequence number. Positive contours are shown in black, negative in red.

is crucial to be careful about assignment of NOESY crosspeaks as mistakes will bias the structure calculation. This is explained in more detail in Chapter 5.

4.5.5 Summary of MBD assignment

Assignment of the free MBD at 288K was carried out using the experiments summarised in table 4.1, and is 86% complete (excluding the His tag). An assigned ^{15}N HSQC is shown in figure 4.10 and the full chemical shift list is found in Appendix C. These data were used for structure determination, as described in Chapter 5. The MBD2 MBD appears to be monomeric in solution, as the number of peaks in the ^{15}N HSQC correspond approximately to the number of amino acid residues. The possibility of a perfectly symmetrical homodimer can

be excluded, as there is insufficient line-broadening in the spectrum associated with the slower tumbling of a larger molecule. There is evidence for minor forms of some amino acids however, most likely caused by proline isomerisation leading to localised chemical shift changes. As the peaks were generally very weak, there was insufficient data to link alternative amino acids forms and so include them in the assignment. For some residues, the alternate form predominated at 298K making transferring of assignments and comparison with the complex difficult.

4.6 Studies of the MBD:DNA complex

Although NMR and crystallographic techniques were employed to investigate the interaction between the MBD and methylated DNA, attempts to crystallise the MBD:DNA complex were unsuccessful (see Chapter 2). NMR is thus the sole method used in this project to probe the structure of the MBD.

4.6.1 MBD:DNA complex preparation for NMR

A ^{15}N, ^{13}C-labelled protein sample was titrated with a 12 bp, symmetrical DNA duplex, with a methylated central CpG pair (sequence: 5'-GCTTAmCGTAAGC-3', Oswell). High quality, purified and dried DNA oligomer was made up to a concentration of 1mM, with dH_2O. Ten aliquots of 200μl were lyophilised ready for titration into the MBD sample. ^{15}N HSQC spectra were used to monitor the titration. After each spectrum was collected, the sample was removed from the magnet and carefully pipetted into an eppendorf with lyophilised oligomer. In this way, each titration step increased the DNA concentration by 0.2mM. After

mixing the protein sample with a new aliquot of DNA, the tube was centrifuged for 5 minutes at 13,000rpm and then replaced back into the NMR tube. The sample was put back in the magnet, re-locked and shimmed before the next HSQC spectrum was recorded. All parameters were kept identical so that changes due to the increased DNA concentration could be clearly observed. In particular, the receiver gain was set at a lower than optimal level and the temperature monitored. Each time, the sample was allowed at least 15 minutes to equilibrate at temperature, inside the magnet. An intermediate exchange regime was expected, due to previous experience with the MBD:DNA complexes of other proteins [136]. Since the frequency difference (in Hz) between two states depends on the field strength, the titration was carried out at 600MHz rather than 800MHz. This increases the likelihood of observing faster exchange, and thus being able to see peaks of the MBD:DNA complex, in the ^{15}N-HSQC spectrum. The titration was continued as long as changes in the HSQC spectrum were observed. Figures 4.11 and 4.12 show the HSQC spectra of the titration.

4.6.2 Effect of DNA titration

In the early stages of the DNA titration, almost all the ^{15}N HSQC peaks disappear. After the addition of five aliquots (1mM DNA), all that remains are peaks from extreme C-terminal residues and the amino ends of Gln and Asn sidechains. This indicates that the system is in intermediate exchange. The signals in the HSQC are extremely broad, due to a coalescence of the two resonance states resulting from complexed and free protein. The protein is in excess, with a concentration in the range 1.4-1.6mM, as calculated during the purification process.

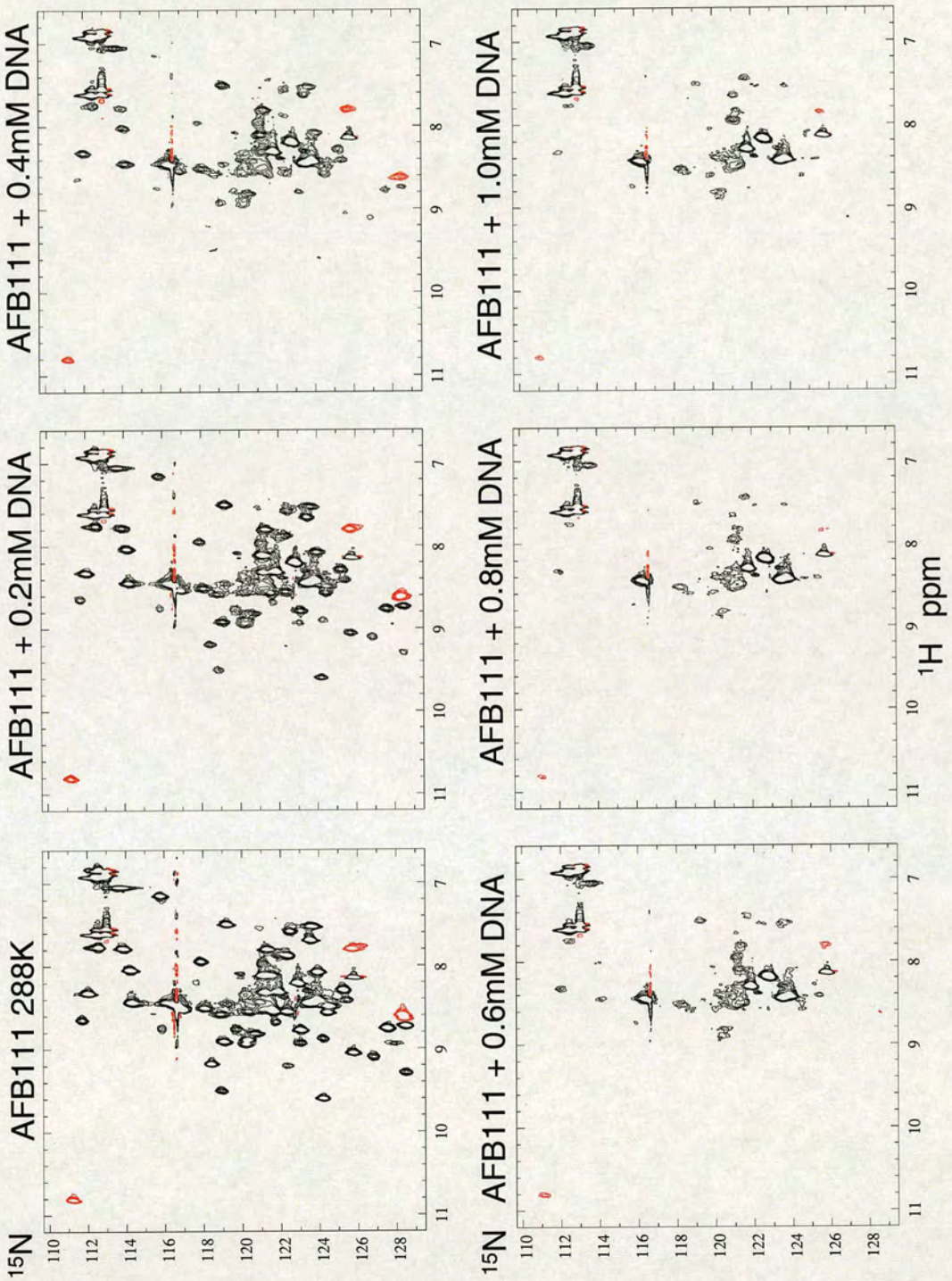

Figure 4.11: Titration of the MBD2 MBD with double-stranded, methylated DNA (sequence: 5'-GCTTAmCGTAAGC-3'), monitored using the ^{15}N HSQC experiment. Part I.

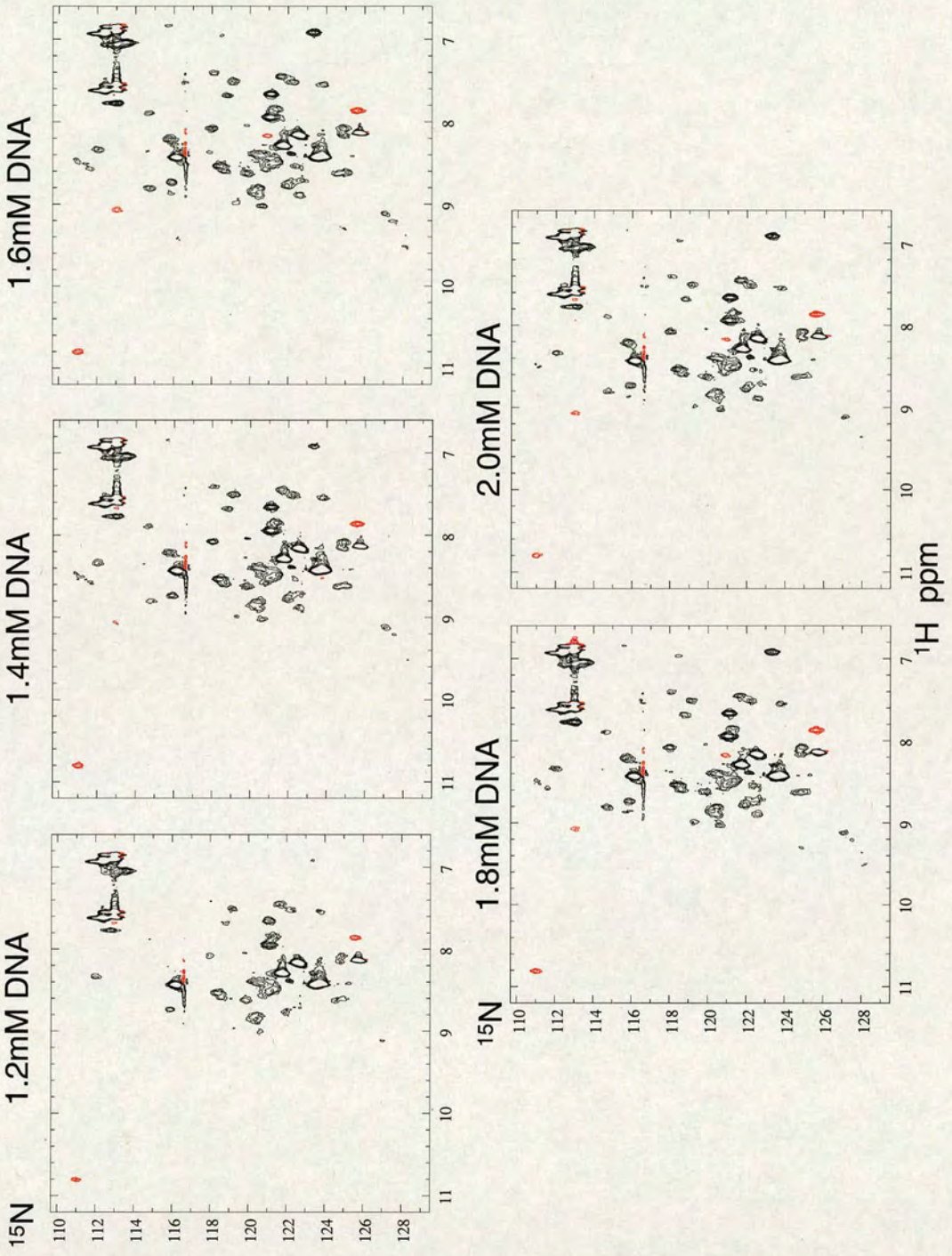

Figure 4.12: .Titration of the MBD2 MBD with double-stranded, methylated DNA (sequence: 5'-GCTTAmCGTAAGC-3'), monitored using the ^{15}N HSQC experiment. Part II.

Adding more DNA to the sample, restored many of the peaks to new positions in the spectrum. The complexed form of the protein has become the dominant resonance state. Although the spectrum is poor, and the peaks are weak due to exchange broadening, it is possible to see that some dramatic shifts have taken place. During the titration process, no further change was observed after adding the final aliquot of DNA (2mM). Later inspection of the spectra revealed that the receiver gain had been set too high for the experiment at 1.8mM DNA. There is in fact little difference between 1.6mM and 1.8mM. A slight excess of DNA is therefore present in the final sample.

4.6.3 Changing the temperature

In spite of having an at, or near equimolar MBD:DNA ratio, the HSQC spectrum of the complex is still very poor and intermediate exchange is still in evidence. The type of exchange regime in NMR depends on the size of the exchange rate compared to the frequency difference of the two states.

$$k = \frac{\pi \delta \nu}{\sqrt{2}} \tag{4.1}$$

Equation 4.1 describes the relationship between the exchange rate, k and the frequency difference ($\delta\nu$) in intermediate exchange. If k is greater than the right hand side of equation 4.1, the system will be in fast exchange.

To alter the exchange regime from intermediate towards fast, the temperature was increased in four steps from 288K to 298K. Figure 4.13 shows the effect this had on the HSQC spectrum. The quality of the spectrum improves steadily, until

at 298K, a good spectrum of the complex is obtained. The spectrum is better dispersed than that of the free protein, which indicates a more structured protein backbone in the complex.

4.6.4 Sequential assignment of complex

In order to assign the ^{15}N HSQC of the MBD:DNA complex, ^{13}C backbone experiments, CBCA(CO)NH and HNCACB, described previously, were collected at 298K on a 600MHz spectrometer. From this pair of experiments, sequential assignment was completed, with 80% of Cα and Cβ atoms assigned. This process was made easier by knowledge gained from the assignment of the free protein. The improved quality of the connections in the backbone assignment suggested that the whole molecule had become more rigid in the complex. The C-terminus was much more ordered, but backbone amide assignments for residues 210-213, could still not be found (see figure 4.14).A summary of experiments collected on the MBD:DNA complex, is shown in table 4.2.

Experiment	Solvent	Processing	T (K)	Ref.
2D ^{15}N HSQC	H_2O	FT	288-298	[122][123]
3D CBCA(CO)NH	H_2O	MEM	298	[129][130]
3D HNCACB	H_2O	MEM	298	[129][131]

Table 4.2: A summary of experiments collected on the MBD2 MBD in complex with methylated DNA. Solvent described as H_2O is actually 90% H_2O. ^{15}N HSQCs were collected at a range of temperatures; T is temperature in Kelvin.

4.6.5 Assignment of the MBD at 298K

In order for a comparison to be made between the free and bound states of the MBD, a limited assignment of the free protein at 298K was carried out. This

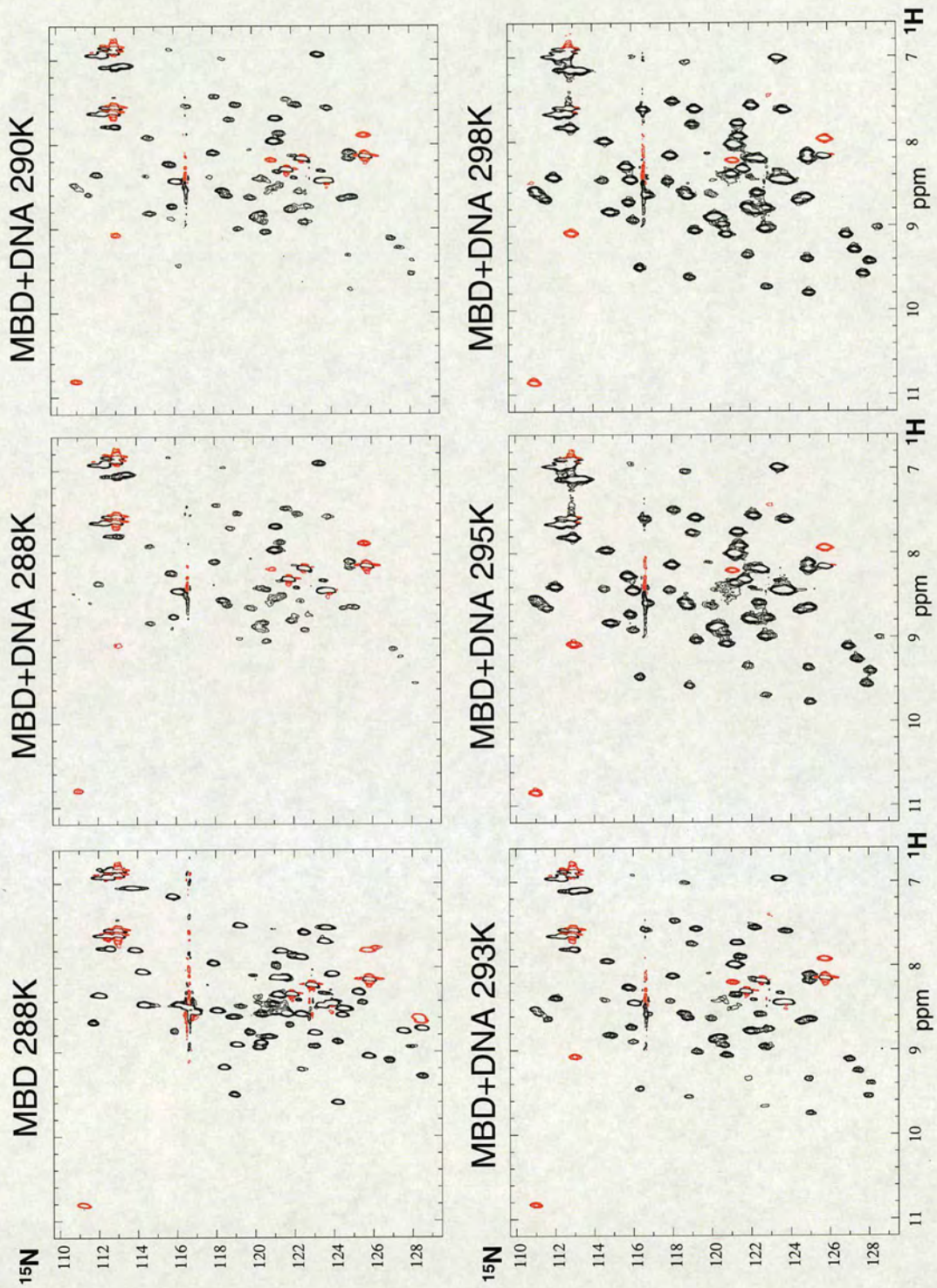

Figure 4.13: Effect of changing temperature on the ^{15}N HSQC of the MBD2 MBD:DNA complex. Free, unbound, MBD2 MBD at 288K is also shown, for comparison.

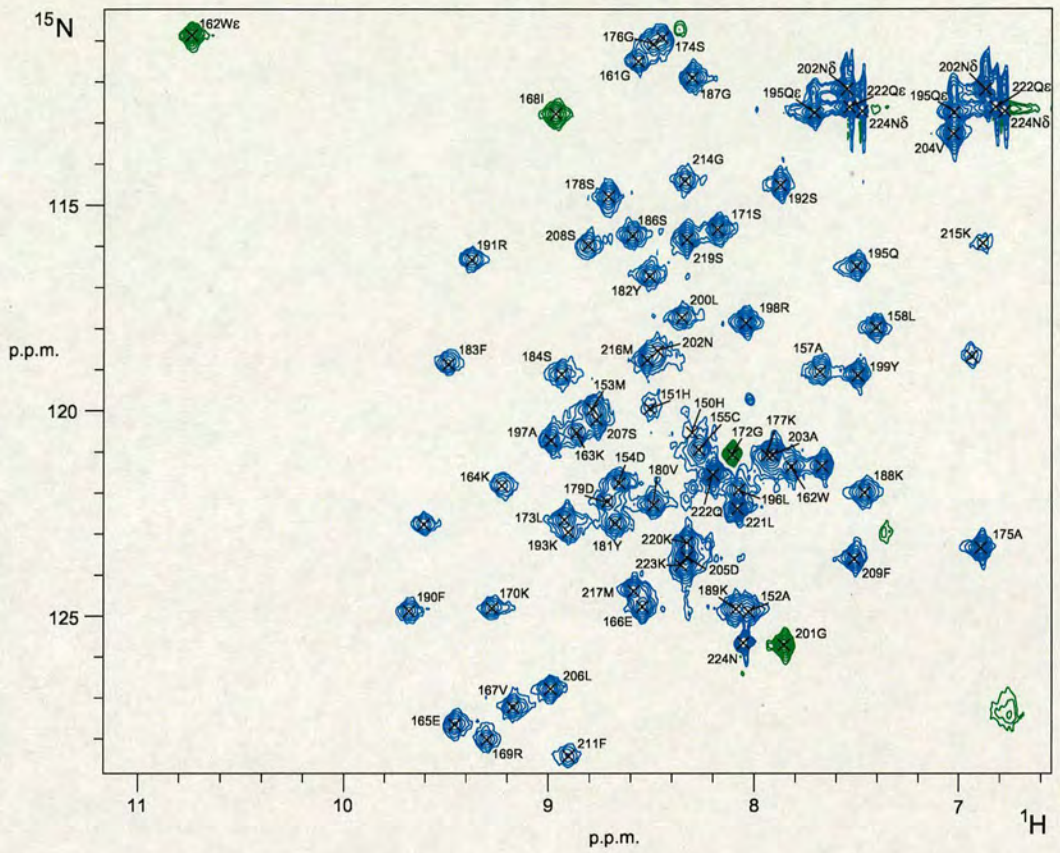

Figure 4.14: A summary of the assignment of the MBD2 MBD:DNA complex at 298K. Each NH crosspeak is labelled with the single letter amino acid code and sequence number. Positive contours are shown in blue, negative (folded) in green.

was done using a combination of the backbone experiments collected at 298K (see table 4.1) and information from the completed assignment at 288K. It was possible to assign the ^{15}N-HSQC, apart from a few residues in the disordered termini (150-151, 210, and 213-215) and Lys170. This residue is only observed at 298K, in the bound state. Figure 4.15 shows the assigned spectrum.

4.6.6 Comparison of NH chemical shifts

Using data of the free MBD collected at 298K, it was possible to compare chemical shifts of free and DNA-bound protein. A graph of backbone amide chemical

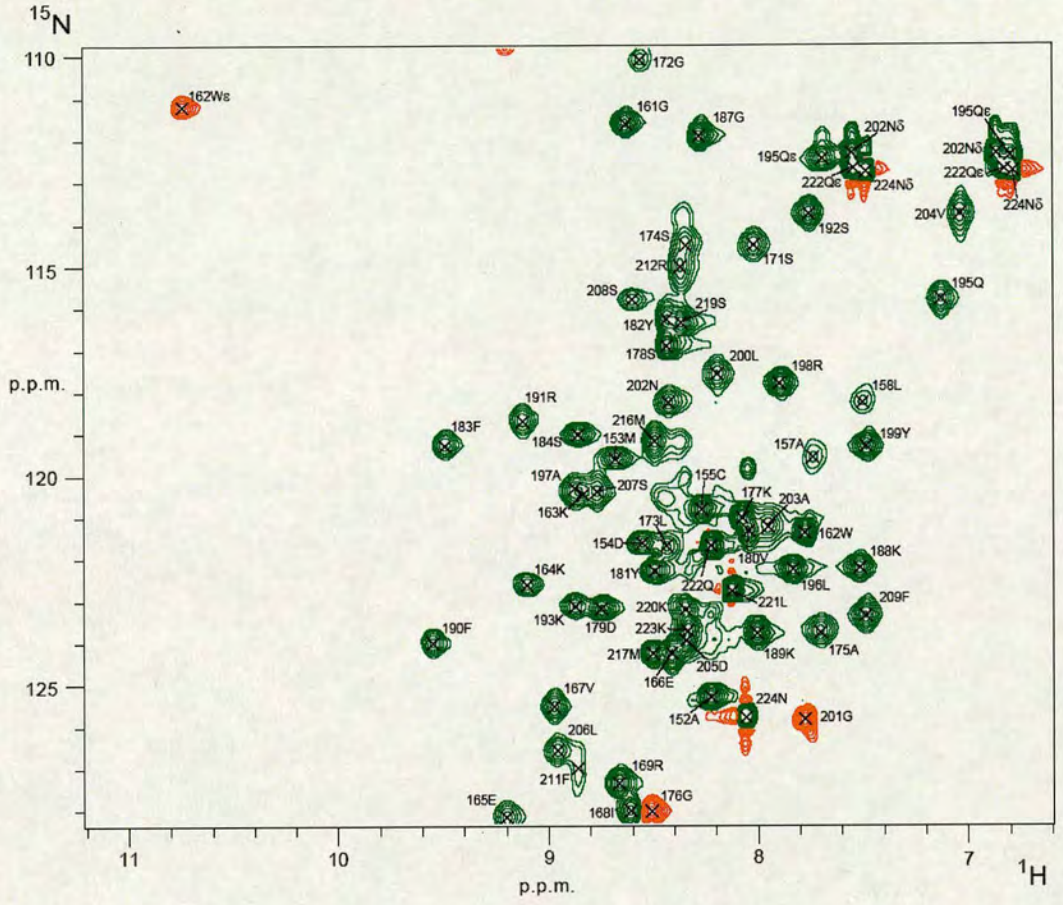

Figure 4.15: A summary of the assignment of the MBD2 MBD at 298K. Positive contours are shown in green, negative (folded) in orange. Each NH crosspeak is labelled with the single letter amino acid code and sequence number. These assignments were used for comparing chemical shift changes induced upon DNA binding.

---

shift changes induced upon DNA binding is shown in figure 4.16, part A for MBD1 MBD, and figure 4.16, part B for MBD2. A similar pattern of changes is seen for both proteins, and they also compare well with data for the MBD from MeCP2 [98]. The most affected residues are in the loop, L1, and the α-helix plus the preceding loop. In MBD2, large chemical shift changes are also likely to occur in the C-terminal region corresponding to the hairpin loop in MBD1, but are not seen due to incomplete assignment of these residues. In MBD1, these three regions form a continuous surface, later shown to be the

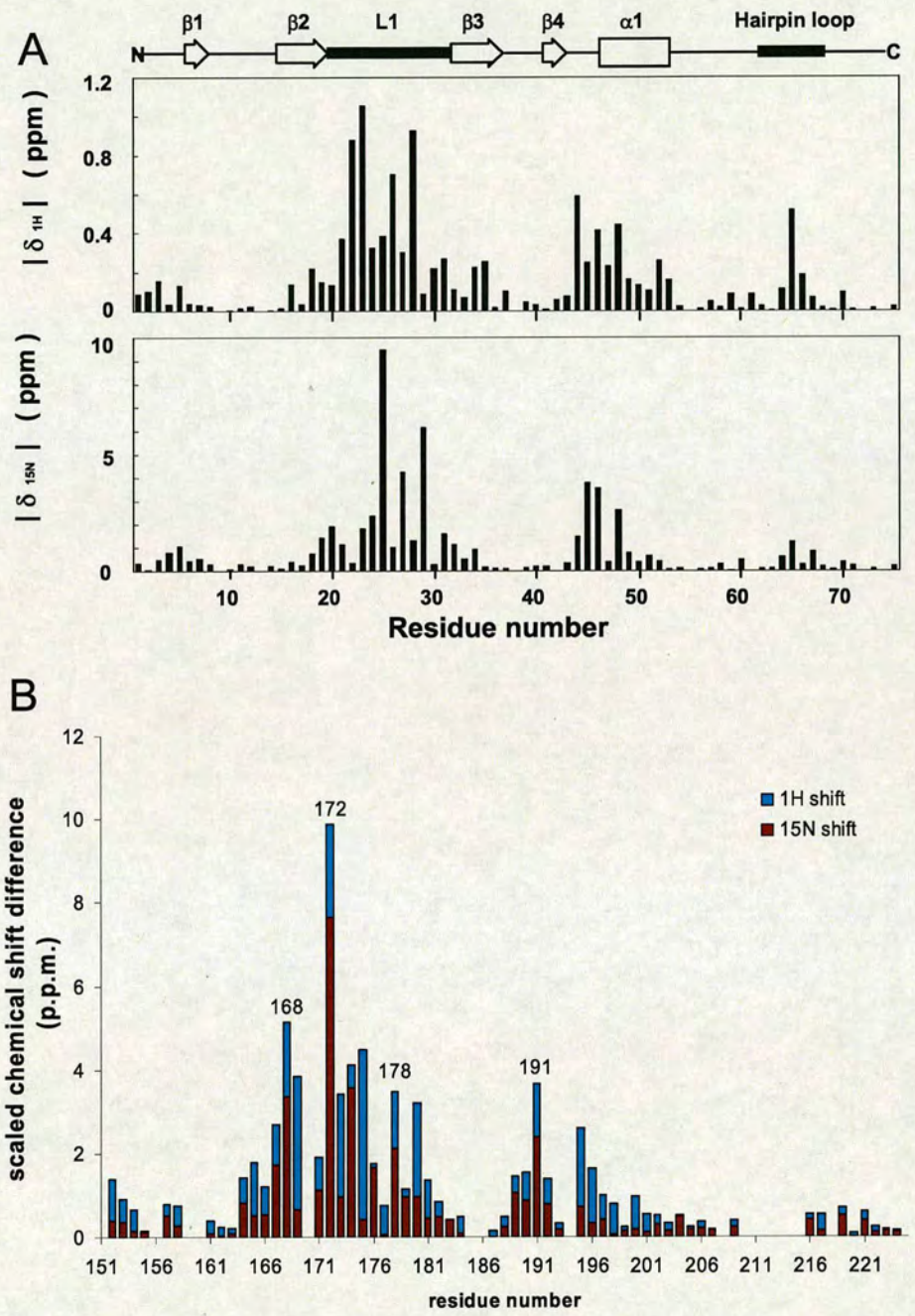

Figure 4.16: A:Graph to show chemical shift perturbation upon DNA binding to the MBD of MBD1. ^{15}N shifts are in the lower panel and ^{1}H shifts in the upper panel. Taken from [99] with the authors' permission. B:Graph to show chemical shift perturbation upon DNA binding to the MBD of MBD2. Stacked shifts are shown. ^{15}N shifts are in maroon and ^{1}H shifts, scaled up by a factor of five, are in blue. Prolines, and residues for which amide assignments are missing, are shown as zero.

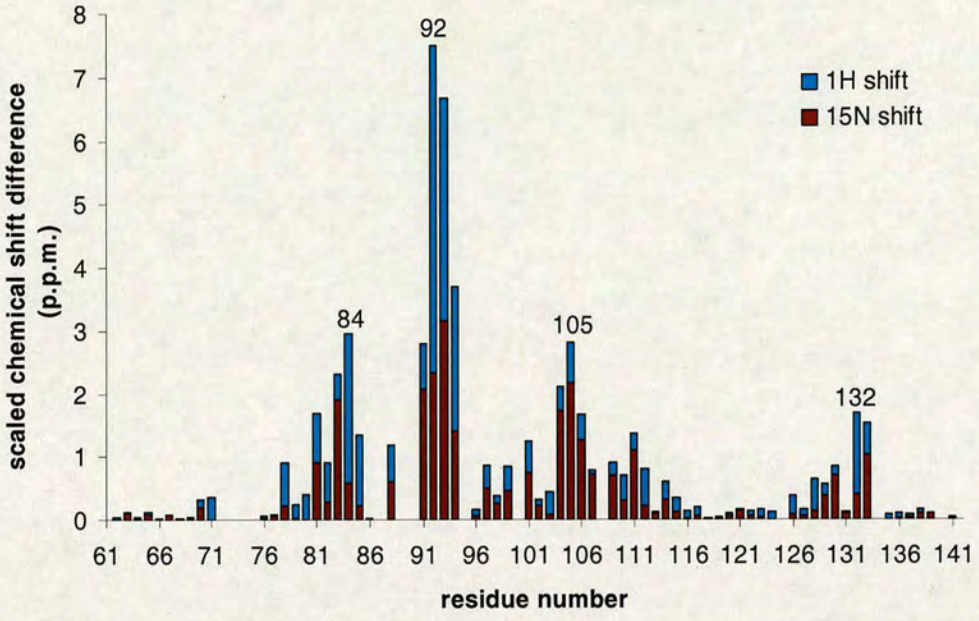

Figure 4.17: Graph to show chemical shift perturbation upon DNA binding to the MBD of MBD4. Stacked shifts are shown. ^{15}N shifts are in maroon and ^{1}H shifts, scaled up by a factor of five, are in blue.

DNA-binding interface [100]. Smaller changes were observed on MBD1 MBD binding to an identical but unmethylated probe, and the system exhibited a fast exchange regime. The equivalent experiment was not done with MBD2, but in surface plasmon resonance studies, for all MBDs tested, binding to unmethylated DNA could not be detected (see Chapter 3).

MBD4 NH shift changes

A similar titration was done with the MBD from MBD4, and a single-stranded, 8-mer, hairpin oligonucleotide (sequence: 5'-GAGmCGTCT-hexaethylene glycol-AGATGCTC-3', Oswell). The construct used was the same one described in Chapter 2 (see table 2.1) except the N-terminal tag had been cleaved off leaving only residues 61-141 and three, non-native residues (AMA) at the N-terminus. This work was done by Brian Smith. The backbone amide chemical shift changes

are shown in figure 4.17 for comparison. Although the same general trend is observed as for the other MBDs, some missing assignments make the pattern less obvious. The largest changes are observed around residues 84 and 92, which are at the beginning and end of the L1 loop equivalent region. The middle of this loop showed the greatest change in MBD1 and MBD2, but Gly87 is unassigned in MBD4, making a comparison impossible. The completely conserved region around Phe105 also showed a large change in chemical shift upon DNA binding. The equivalent residue to Arg106, (Arg44 in MBD1) forms a major groove contact from the base of the α-helix. It is also one of five residues which make up the hydrophobic, methyl-recognition patch [100]. Residues 124-127 are an insert, present in MBD4 and MeCP2, but not in the other family members. There is no evidence for any role in DNA binding for these residues, as their resonance positions are unchanged.

4.6.7 ^{13}C shifts

A difference in backbone carbon shifts, may indicate a change in the basic secondary structure [137]. For MBD2, Cα shifts remain virtually the same in all but a few residues. Shifts above 0.5 p.p.m. are only observed for 7 residues in the L1 equivalent region, and for Arg191. This is consistent with the model of DNA-binding in which only a small part of the MBD changes shape. The pattern for Cβ shifts indicates that residues 170-181 and 191-192, are the most affected by DNA binding.

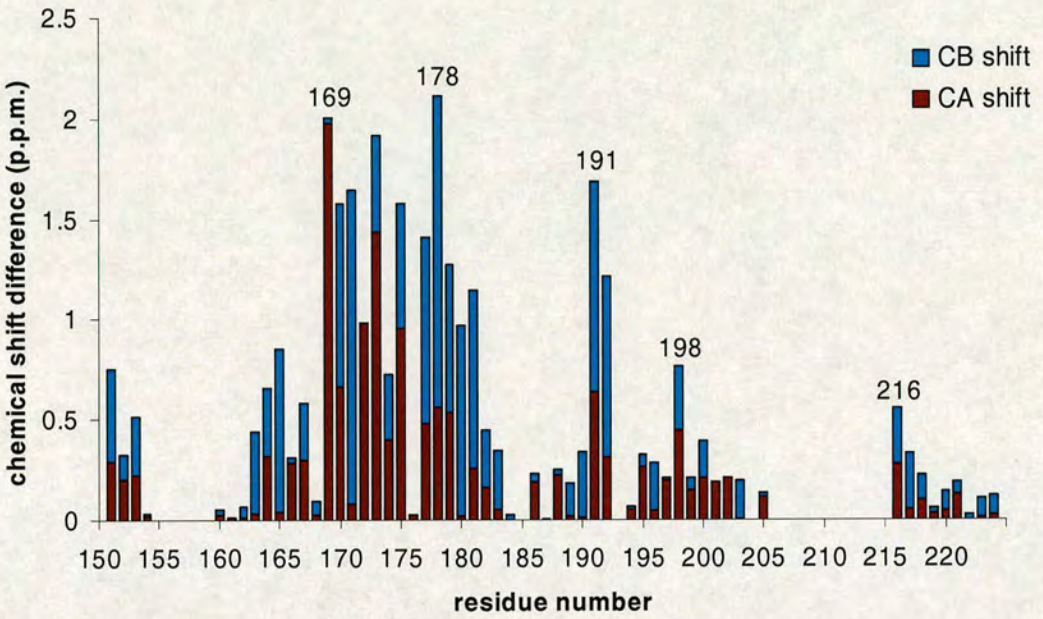

Figure 4.18: Graph to show Cα and Cβ chemical shift perturbation upon DNA binding to the MBD of MBD2. Contributions from each resonance are stacked, coloured and labelled accordingly. Incomplete assignment of spectra made comparison of some residues impossible (shown as zero).

4.6.8 X-filtered experiments

X-filtered experiments for measuring inter-molecular NOEs [138], were collected on the sample of the MBD:DNA complex, but none were observed. It is not clear why this was the case, as 91 intermolecular NOEs were identified in the structure of the MBD1:DNA complex [100].

4.7 Chapter conclusion

At 288K, sufficient NMR experiments were collected to enable 86% of native MBD2 MBD atoms to be assigned. The ^{15}N-HSQC of the free protein is of reasonable quality and most residues were easily identified using either ^{15}N-, or ^{13}C-backbone experiments. The data suggested a typical MBD fold, with struc-

tured regions of the correct length matching β-sheets, interspersed with less well-ordered loops. Residues 194-201 had strong Hα-NH crosspeaks both within a residue and between residues 3-4 apart, consistent with an α-helical structure. The C-terminus is the most disordered region, but this is due to its flexibility, shown by broadened ^{15}N line widths (see particularly residues 207S, 209F, 211F and 216M in figure 4.10). Proline isomerisation is likely to be responsible for alternate, minor forms of some residues, although only the major form could be assigned and sequentially linked. For instance, Ala157 and Leu158 in the sequence PALP, are very weak but identifiable. Assignments of the MBD2 MBD were used for structure determination of the free protein, described in Chapter 5.

Titration of the MBD2 MBD with a methylated duplex DNA produced a complex in an intermediate exchange regime. Increasing the temperature by 10K, to 298K, vastly improved the quality of the HSQC spectrum. It was then possible to collect backbone experiments and complete a backbone assignment of the MBD:DNA complex. A similar procedure had previously been carried out on the MBD4 MBD. From NMR studies of the complexes, a mode of DNA binding very similar to that of the MBD1 MBD is expected for both MBDs. The DNA-binding surface appears to be well conserved, and the domain does not undergo major structural changes on DNA binding, in either MBD2 or MBD4. The only exception is the L1 loop region and a residue immediately preceding the α-helix, R191 (R106 and neighbouring residues in MBD4). Few conclusions can be drawn about the function of the C-terminal hairpin loop due to incomplete data for MBD2, but it appears to be inflexible in MBD4. Figure 4.19 maps the largest chemical shift changes due to DNA binding, onto the MBD2 MBD ^{15}N-HSQC.

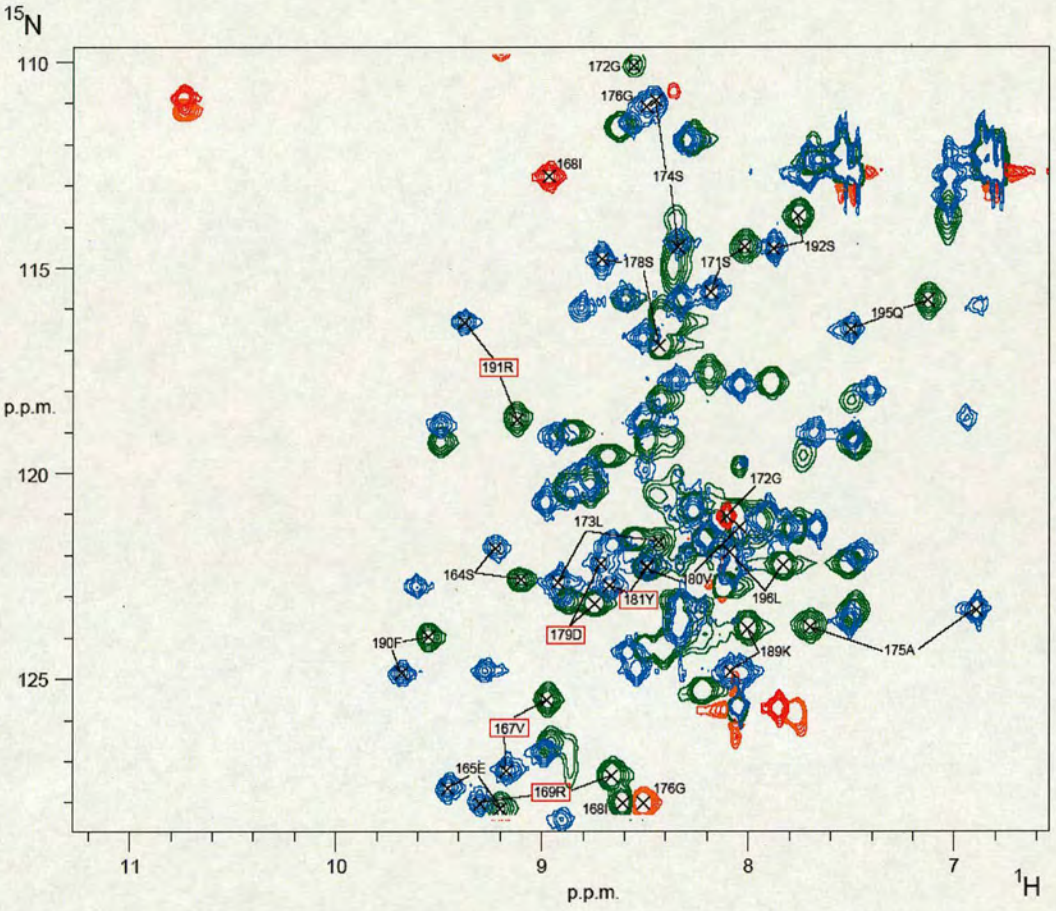

Figure 4.19: The largest shift changes induced upon DNA binding are labelled. The five hydrophobic patch residues, reportedly responsible for methyl-specificity are boxed in red. Each NH crosspeak is labelled with the single letter amino acid code and sequence number. Contours of the complex are coloured blue (positive) and red (negative, folded), and those of the free MBD, are shown in green and orange.

Chapter 5

Structure Calculation

5.1 Introduction

This chapter describes how the information obtained from NMR experiments (explained in Chapter 4) was used to determine the structure of the MBD from MBD2. The final, calculated structure is discussed and compared to that of other MBDs.

5.2 Distance restraint generation

The applications of the NOE experiment have been described previously (see Chapter 4). In short, the intensity of the NOE between two nuclei is inversely related to the sixth power of the distance between them. Using ANSIG [135], the NOE crosspeaks in both NOESY experiments were integrated to give a measure of their intensity. This information was then compiled in a list, containing the chemical shifts, assignments and relative intensities of all marked (ie. recognised,

but not necessarily assigned) peaks. The crosspeak file was then subjected to a number of manipulations in order to extract and sort this information. There are several objectives of this sorting process. The first is to select nuclei with an unacceptably high chemical shift range (e.g. >0.05 p.p.m. in ^{1}H) and check the relevant assignments and peak positioning in ANSIG. There were very few of these so-called "offenders" in this project. All were eliminated after a single re-examination of the spectra. Secondly, a table of *connected* nuclei was created. This summarises chemical shift information for protons, together with that for their attached heavy atoms. Finally, the table of connected nuclei described above, was correlated with the list of NOESY crosspeaks, from both ^{15}N and ^{13}C spectra. This was done separately for each spectrum, thus producing two files of crosspeaks with their matching assignments. Some crosspeaks match the chemical shift of more than one nucleus , in which case they are ambiguous (see section 5.2.2). Alternatively, some NOE crosspeaks could not be matched to any known (assigned) nucleus and chemical shift. These unmatchable peaks were usually either erroneously marked noise, or due to incomplete assignment in a poorly-defined region of the protein.

NOE intensities were later classified and converted to a distance in Angstroms (Å), according to the categories listed in table 5.1.

NOE category	Relative intensity boundary	Interproton distance (Å)
Strong	> 3.0	< 2.7
Medium	> 1.5	< 3.3
Weak	> 0.3	< 5.0
Very weak	> 0.0	< 6.0

Table 5.1: A table of NOE distance categories and relative intensities, which are converted to interproton distances using the *connect* script in AZARA[134].

5.2.1 NOE to distance conversion

The calibration of NOE intensity to distance restraint depends on the individual system being studied. In general, the validity of using a particular calibration method may be determined by analysing the distribution of NOE violations in the resulting structures. Violation "clustering" in certain residue types or particular secondary structure elements, or at distance boundaries, may indicate a poor relationship between the NOE calibration and the actual inter-nuclear distances.

The low number of violations in the MBD2 MBD structures was taken as confirmation that the NOE calibration used was at least a conservative one.

5.2.2 Ambiguous distance restraints

Crosspeaks are unambiguous if they only have one possible assignment in each dimension. Ambiguous assignments are made when two or more nuclei pairings may be contributing to the crosspeak intensity. This ambiguity is common in proteins due to highly overlapped proton resonances. Erroneously assigning a crosspeak as unambiguous, can lead to incorrect and incomplete restraints. The distance restraint between two nuclei i and j, may be too short, as the contribution from a second NOE (i to k) is wrongly attributed to the first pairing. In addition, the distance restraint between nuclei i and k is lost, buried under the i-j NOE. This is illustrated in figure 5.1. A limitation of ANSIG is that each dimension of a crosspeak is either unambiguous or wholly ambiguous. It is not possible to describe a finite list of contributors to any one NOE, so that all resonances with similar chemical shifts are included as possibilities. This has the effect of weakening the restraints, but subsequent filtering and checking processes

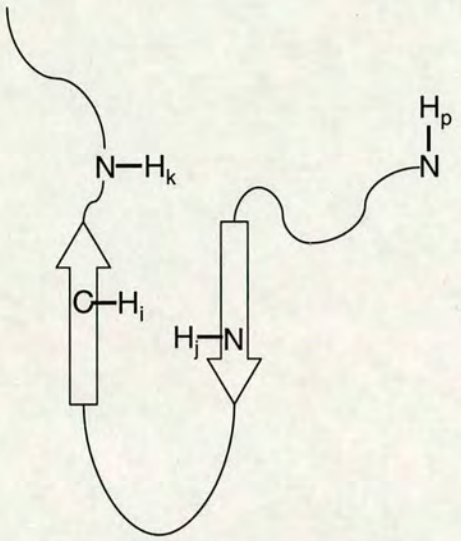

Figure 5.1: The nature of ambiguous restraints. In this example, all HN protons have similar chemical shifts. The NOEs between H_i and H_j and H_i and H_k will be overlapped, so the assignment must be left ambiguous. In addition, a contribution from H_p will be included in the calculation, even though it is too far away for this to be genuine. Filtering (in later refinement steps) will probably discount H_p as a contributor.

minimise spurious contributions (see 5.4.2). One problem with this approach, is that an ambiguous restraint has the potential to consistently restrain a pair of atoms in an under-restrained part of the structure, and consequently violate other (true) restraints. In other assignment software such as CCPNAnalysis [139], it is possible to assign a discrete number of atoms to an ambiguous crosspeak.

5.3 Theory of structure calculation

The aim of the structure calculation protocol, is to combine empirically derived stereochemistry and experimental data, to generate structures which satify all the restraints. The structure calculation method is molecular dynamics in Cartesian space, using simulated annealing. As it is not possible to explore all conforma-

tions of a molecule, the method provides an efficient minimisation strategy by monitoring the internal energy of the system. The conformations which satisfy the restraints with the lowest possible energy, are the best solutions to the problem, and have the highest probability of representing the true structure of the molecule.

5.3.1 Simulated annealing

The concept of simulated annealing (SA) is based on the way in which liquids freeze [140]. The annealing process starts with a high temperature "melt", which is slowly cooled so as to maintain thermodynamic equilibrium across the system. As cooling progresses, the system becomes more ordered, gradually approaching a theoretical "frozen" state. If the starting temperature is too low or the cooling process carried out too rapidly, then the system may become trapped in a region of local minimum energy. This must be avoided as the aim is to arrive at the global minimum energy state, and so solve the particular problem.

In the original protocol, the SA algorithm compared each state of a system to a neighbouring state, which differed in a small, defined way from the initial condition (see [141] for generalised example). The probability of the system moving to the neighbouring state was weighted according to the internal energy of each state; the system tended to move towards lower energy. Important parameters to be defined were the neighbourhood function, the transition probabilities and the annealing schedule. The SA protocol used in this thesis, however, is more straight-forward, and simply involves the slow removal of kinetic energy. This is achieved by lowering the effective temperature or by increasing the force con-

stants on the restraints. In all cases, the annealing schedule is limited by the computational time available; an unavoidable limitation on finding the global minimum.

5.3.2 Molecular dynamics

Molecular Dynamics (MD) is a computer simulation technique used to sample the potential conformations of a system, for a given set of restraints [142] and internal energy. Since the potential energy is the sum of all empirical and experimental energy terms, the contribution of each of these terms to the total can be varied. The MD protocol is thus designed to reliably represent the changes in the conformation of a molecule over time. Experimental distance restraints derived from NOESY spectra are incorporated into the structure calculation along with information from a forcefield, (the non-experimentally derived restraints). The forcefield defines aspects of the covalent structure of a molecule, such as bond-lengths, chirality, and non-bonded interactions. The parallhdg5.3 forcefield was used in this project [143].

5.4 Crystallography and NMR system

The Crystallography and NMR System (CNS) programs [144], were used to carry out the structure calculation protocol. CNS brings together distance restraints, the forcefield, and a molecular topology file which defines the protein sequence and chemical bonds. There are three main stages to the calculation process. The first involves random structure generation in Cartesian space. A minimisation

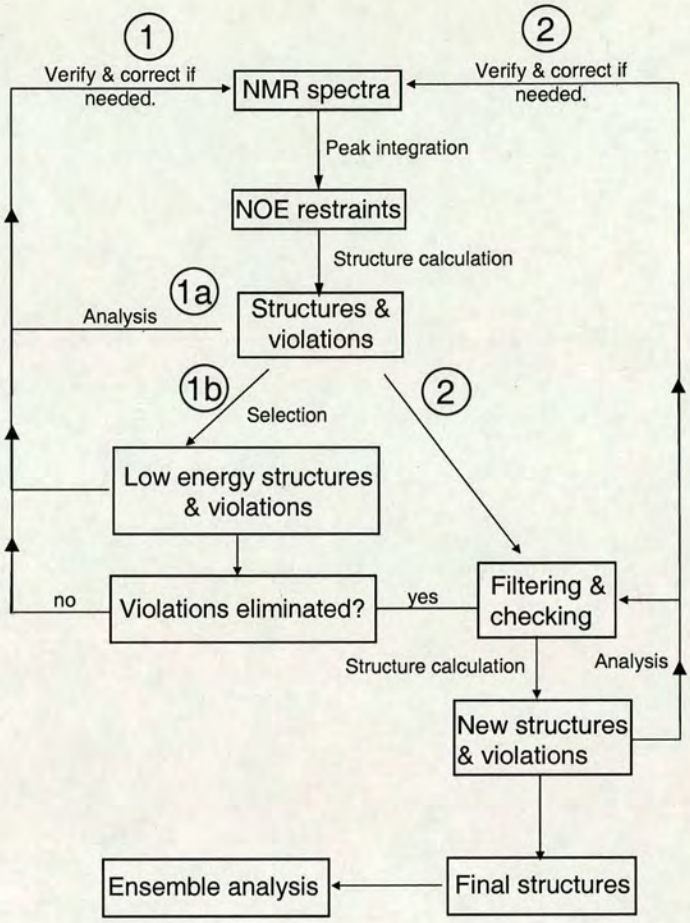

Figure 5.2: A flow diagram to illustrate the iterative structure calculation process. Path 1a was followed in the first round, subsequent rounds followed route 1b. Once significant violations had been removed, the refining stages in path 2 were used to create more precise structures

process is carried out on the van der Waals potential energy term, for heavy atoms only. In the second stage, these random structures are regularised by enforcing the correct local geometry on each individual residue. Chirality and planarity (improper) terms are introduced and the system undergoes two SA phases. The aim is to remove one possible cause of the system becoming trapped in local energy minima. Local energy barriers may be overcome by allowing random swapping of prochiral groups. The final stage is a refinement process in which the whole system is cooled, and undergoes two further SA stages. Prochiral

groups such as methylene protons are swapped to give the lowest contribution to NOE energy.

Structures from all three stages are written as Protein Data Bank (PDB) files [101], but generally only the refined structures are analysed in the next stage.

5.4.1 Structure analysis

The structure calculation process is iterative and involves several rounds of generating structures, analysing them, making changes to the restraints and repeating the calculations. An overview is given in figure 5.2. The first round of structure calculation was atypical. Only 10 refined structures were generated in order to quickly identify major violations. Violations can be thought of as unsatisfied restraints, in this case, inter-nuclear distances in the structure which do not fit the intensity of their NOEs. Deviations from experimental distance values of more than 0.5Å, were classified as a violation. A common reason for violations is misassignment of NOESY crosspeaks, due to typographical or judgement errors, and poor crosspeak placement, giving incorrect intensities. The first round analysis quickly identified some violations of this type, which were easily eliminated. See path 1a in figure 5.2.

From round 2 onwards, the refined structures were ranked according to their total energy and a subset of structures chosen to be analysed (path 1b, figure 5.2. The subset is made up of structures whose total energy is below a defined threshhold. The definition is made for each set, where a sudden increase in energy is observed. Typically, each structure calculation round generated 30-50 structures, of which 80% formed the low energy subset. These were analysed as before, taking

into account additional reasons for the occurrence of violations. Violations may arise due to overlapping crosspeaks where it's difficult to determine how much of the intensity is due to a particular nuclei pair. Violations may also result from ambiguous crosspeaks, due to inclusion of all possible contributors (see section on Ambiguous NOEs). If false restraints are generated in an otherwise relatively unrestrained part of the molecule, then true contributions may be violated, in the attempt to satisfy the erroneous restraints. After several rounds of identifying and removing violations, the next step is to refine the structures.

5.4.2 Refining the structure

Two processes, filtering and checking are employed to refine the structures. Filtering is a method of limiting the ambiguity in ambiguously assigned crosspeaks. By comparing the restraints with low energy structures, the contributors to each NOE are ranked according to their percentage contribution. The percentages are added up until a defined total is reached, after which, any remaining contributions are discarded. Another round of calculations is then performed with these filtered restraints, and the process repeated. The filtering level started at 99% (in which the final 1% of contributors was discarded) and was decreased in subsequent rounds. The MBD2 MBD structures were filtered three times at 99%, then at 98%, and finally at 95%. Checking removes duplicate restraints that arise from symmetry related crosspeaks. The less intense NOE is discarded in each case. New structures were generated from filtered and checked NOEs, and then a subset was used as the comparison structures in subsequent rounds (see path 2 in figure 5.2).

Description	Value
Total NOEs	1960
Total Unambiguous NOEs	1329
Intraresidue:	380
Interresidue:	
sequential	390
short range	211
long range	348
Average no. violations >0.5Å:	1
Max. no. of violations per structure:	3
Max. distance contraint violation:	0.9Å
Deviations from idealised geometry:	
Bond lenghts:	0.047Å(0.0025 r.m.s.d.)
Bond angles:	0.41Å(0.019 r.m.s.d.)
Impropers:	0.31Å(0.026 s.d.)
PROCHECK Ramachandran plot stats (151-204):	
Most favoured	64.7%
Additionally Allowed	32.2%
Generously Allowed	2.5%
Disallowed	0.5%
R.m.s.d from the mean (residues 151-204)	
Backbone atoms	0.63Å
All heavy atoms	0.95Å
R.m.s.d from the mean (residues 151-168 & 178-204)	
Backbone atoms	0.49Å
All heavy atoms	0.84Å

Table 5.2: A table showing the statistics relating to the final structures of the MBD from MBD2. All values relate to the low energy ensemble, which comprises 37/45 final calculated structures, but in some cases flexible residues have been excluded from the statistics.

5.5 Results of calculations

1960 NOEs were used to solve the structure of the MBD from MBD2 (residues 153-224). This excludes two non-native residues and five of six N-terminal histidines, for which no NMR assignments could be found. The final round of structure calculation produced 45 structures, of which 37 had similar energies and were included in the final ensemble of structures (see figure 5.4). An energy plot showing both total energy and NOE energy contributions, is shown in fig-

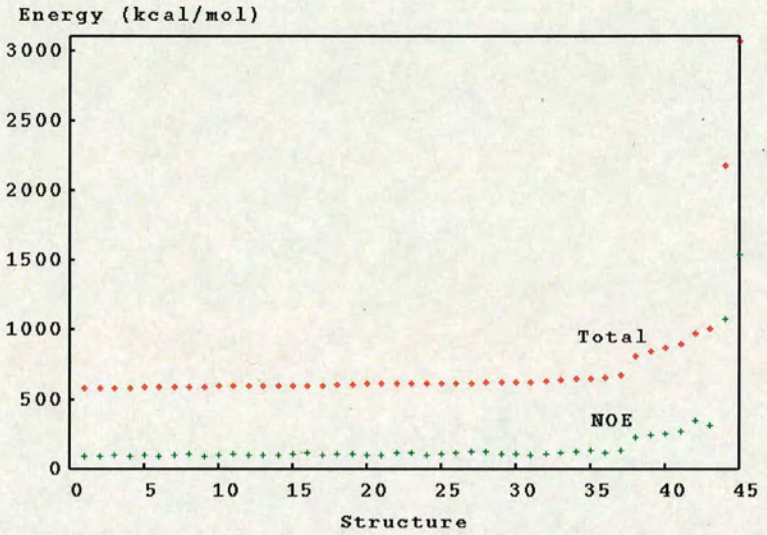

Figure 5.3: Plot of Total energy (in red) and NOE energy (in green) for each of the 45 final structures, ordered according to the total energy. The low energy ensemble is made up of structures 1-37.

---

ure 5.3. There are 37 NOE violations above 0.5 in the 37 structure ensemble, and a maximum of 3 violations in any one structure. The convergence level is high, (greater than 80%), which indicates a high level of precision in the structure. Full statistics for the final ensemble are shown in table 5.1. PROCHECK [145] statistics were calculated using residues 151-204 (and an additional r.m.s.d calculated exclusing loop L1). The 20 C-terminal amino acids were excluded as ^{15}N spectral linewidths suggested that there is genuine flexibility in this region. Broadened ^{15}N-HSQC peaks were also observed for L1 loop residues (see Chapter 4), which fits with the definition of L1 as flexible, from previously determined MBD structures.

Figure 5.4: An ensemble of the 37 lowest energy structures calculated using CNS. All assigned amino acids (151-224) are included here, and the ensemble has been rotated between each depiction, as indicated. The backbone of residues shown in black was used for fitting purposes in MOLMOL [146], whilst blue regions were excluded due to their flexibility.

5.5.1 MBD2 MBD fold

The MBD2 MBD has the typical wedge-shaped, α/β-sandwich fold, common to all the MBD structures solved so far. The secondary structure elements are identical to those of the MBD1 MBD, with a four-stranded, twisted β-sheet (β1, residues 153-156; β2, 162-168; β3, 178-184; β4, 188-191) and a long loop (168-178) on one side. A short helix (194-201) and C-terminal hairpin loop (207-217) make up the other face. Although much of the backbone is well-defined, and similar to the MBD1 structure, there are some notable differences in the loops. The C-terminal 20 residues are poorly ordered in the MBD2 structure, due to incomplete assignment in this region. In spite of this, a β-hairpin structure, like that of MBD1, is identifiable in an ensemble of the lowest 10 energy structures (see figure 5.5). The long loop, L1, between β-strands 2 and 3, appears to be considerably less flexible in the MBD2 MBD. It is two residues shorter, due to

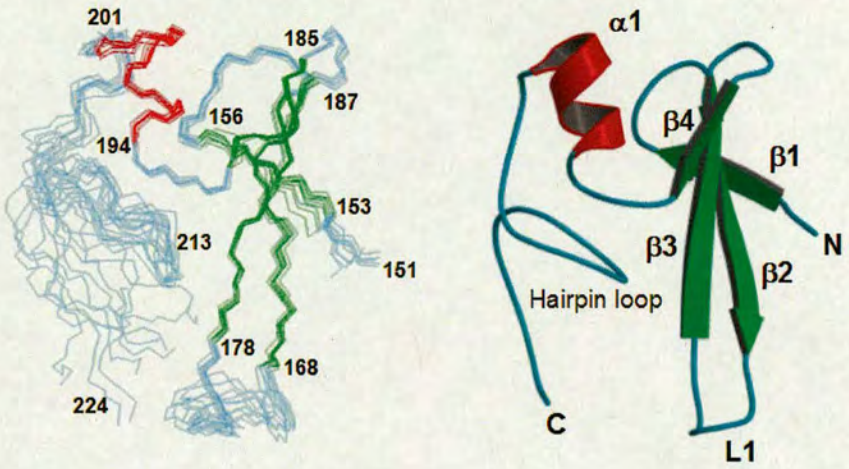

Figure 5.5: Backbone superposition of the 10 lowest energy structures of the MBD2 MBD (made using MOLMOL [146], labelled with amino acid numbers, and a cartoon representation (made using MolScript [147]) showing the secondary structure elements. The superposition was done as for figure 5.4.

---

the extension of each flanking β-strand, and has a ladle shape.

The MBD has a well-defined hydrophobic core, made up of Leu158, Trp162, Val180, Tyr182, Phe190, Leu196, and Leu200. Other C-terminal residues, Leu206 and Phe211 may also form part of the core, although they are not so well-defined in this structure. The MBD2 MBD has three solvent-exposed, aromatic amino acids. Tyr181 is completely conserved throughout the mammalian MBD family, apart from in MBD3, and the position occupied by Phe183 is hydrophobic in all family members apart from MBD1. Tyr199 is completely conserved and may form a lid on the core whilst being largely solvent-exposed. Another conserved group of amino acids form a basic surface on one side of the molecule. Arg 169, Lys170, Lys177, Lys189, Arg191, K193 and Arg212 are mostly clustered between helix α1 and loop L1, extending along the linker between α1 and β4. This is the region in which DNA-binding occurred in MBD1 [100]. Figure 5.6 shows the distribution of amino acids making up the hydrophobic core and positively

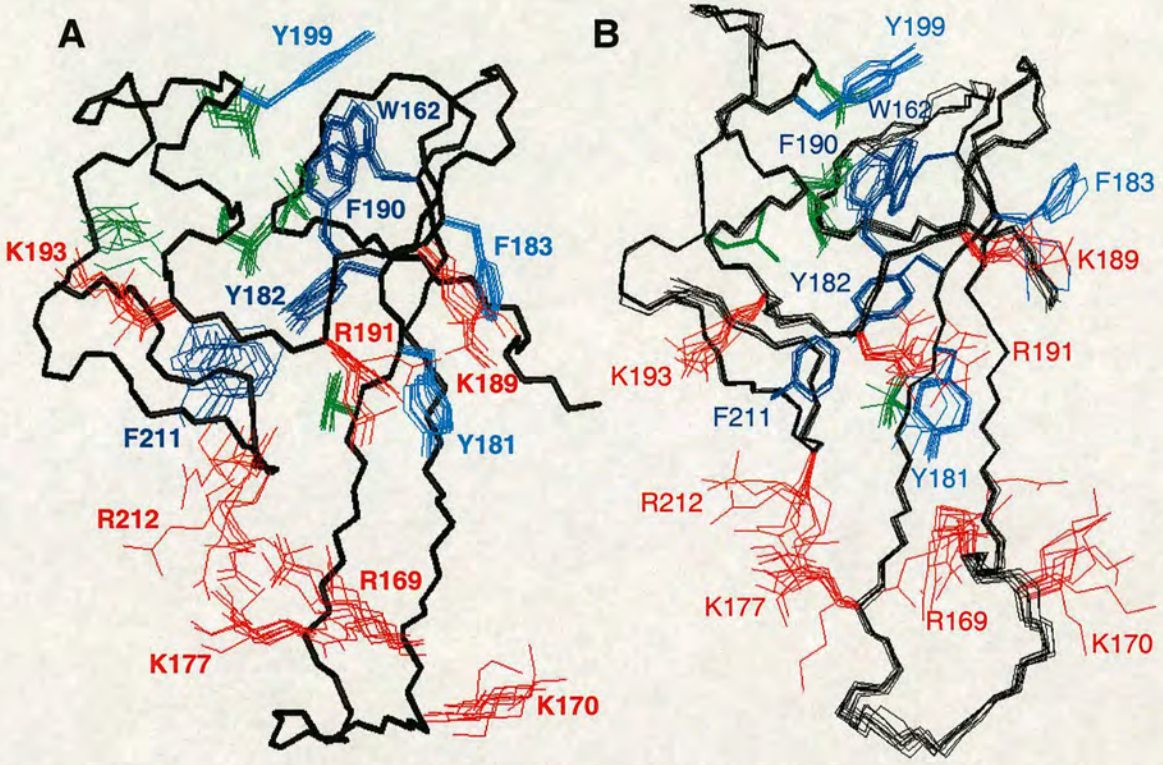

Figure 5.6: A: Overlay of the sidechains from the 10 lowest energy structures of the MBD2 MBD. A single backbone (in black) is shown and the disordered, C-terminal residues (218-224) have been omitted for clarity. B: An ensemble of 10 modelled structures of the MBD2 MBD, based on the MBD1 structure. The backbone (black) is shown for all modelled residues (151-216). For both images: Hydrophobic core residues are shown in dark blue (aromatics) and green (aliphatics); solvent-exposed aromatic residues are in light blue; positively charged residues forming the basic surface are coloured red.

charged DNA-binding face.

5.6 Comparison with predicted model

Due to the high level of sequence similarity between MBDs, a model of the MBD2 MBD was created, using the MBD1 MBD structure as a template. The alignment used for model building was the same as that shown in Chapter 1, figure 1.3. 10 models were generated using MODELLER 8v1 [148, 149], and the ensemble is

shown in figure 5.6.

The model of the MBD2 MBD, has the same overal fold as the calculated NMR structure, except that it is more extended. This is due to a relatively elongated helix (α1), and less rigid loop (L1), in the MODELLER structure. The orientation of core sidechains agrees well between predicted and calculated models, although some of the C-terminal residues are better defined in the former. This is because of the higher degree of precision for MBD1 in this region. The orientation of the Trp162 sidechain differs in the two models. This appears to be a genuine difference between MBD1 and MBD2, but why this should be (in one of the most conserved, core residues) is not clear.

5.6.1 DNA-binding loop

The more rigid structure of the loop in MBD2, is attributable to many ambiguous, short-range NOEs between loop residues. These, together with a small number of long-range NOEs (2 unambiguous, 9 ambiguous), from L1 to residues in the hairpin, are also responsible for the ladle-shape. In figure 5.7, the MBD2 MBD structure has been superimposed onto the MBD1 structure from the MBD:DNA complex [100]. Free MBD2 has a DNA-binding face which more closely resembles that of MBD1 in complex, than free MBD1.

5.7 Conclusion

The structure of the MBD from MBD2, was solved with good internal consistency, and the fold agrees well with other published MBD structures. A comparison

with models reveals that some differences in the loop, L1, were not predicted based on sequence similarity alone. Although highly similar to MBD1, the MBD2 MBD appears to have a pre-formed DNA-binding loop, which may account for differences in the efficiency of methyl-CpG recognition, and DNA-binding.

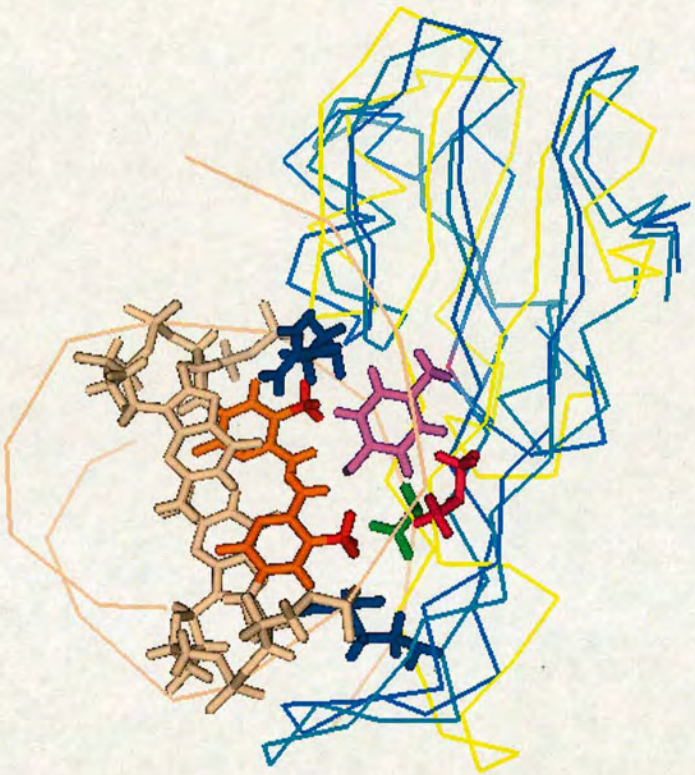

Figure 5.7: Overlay of the backbone of three MBD structures: MBD1 MBD (cyan) from the MBD:DNA complex, free MBD2 MBD (blue), and a model of the MBD2 MBD, based on the free MBD1 structure (yellow). The DNA backbone is a beige line; meCpGs and the five residues from MBD1 that form the hydrophobic meDNA recognition patch, are shown as sticks. Cytosines are in orange, with methyl groups coloured red; Val20, pink; Arg22 and Arg44, blue; Asp 32, green and Tyr34, magenta. The loop, L1, forms a major part of the DNA binding interface. In MBD2, it's conformation more closely resembles that of the loop in the MBD1 MBD:DNA complex, than that in the free MBD1 protein.

Chapter 6

Discussion

6.1 Structural work

6.1.1 MBD structure

NMR was used to solve the structure of the free MBD from MBD2. Structure statistics, generated using PROCHECK [145] and by analysis of the NOE restraints, indicate good quality with a low backbone r.m.s.d. of 0.45Å. Ramachandran plot statistics for structured residues (151-204, excluding the C-terminus), include 99.5% of residues in favoured and allowed regions. The overall fold of the domain is similar to three previously solved MBDs or MBD-like domains [100, 102, 98], but most closely resembles the MBD of MBD1, with which it has the highest sequence similarity.

A notable difference is found in the long loop, L1, which adopts a smaller range of conformations and appears to be more rigid than the equivalent region in MeCP2 and MBD1. The structure of the free MBD overlays well with the

structure of the MBD1 MBD in the bound state with DNA (see figure 5.7), so it is possible that MBD2 has a pre-formed DNA-binding loop. However, residues 170-177, around the long loop, gave unusually sharp NMR signals, resembling those observed at both termini. This suggests that the loop is in fact dynamic. At the relatively low temperature (288K) used in these experiments however, NOEs from loop residues are seen. These include a large number (40 per residue) of short-range ambiguous NOEs between loop residues, and a few (9 ambiguous, 2 unambiguous) NOEs between L1 and the C-terminal hairpin loop. In contrast, NMR data used to solve the MBD1 MBD structure was collected at 303K, and for the MeCP2 MBD, at 291K. At these temperatures, particularly for MBD1, which was determined at 15K higher temperature, the faster motion of the loop may have resulted in NOE averaging, effectively eliminating restraints in this region. How then, is it possible that a limited set of loop conformations *are* observed in the MBD2 MBD structure? Due to the relationship between NOE intensity and internuclear distance[1], a significant NOE may result from a conformation that is present only a small proportion of the time. If that NOE produces a restraint that can only be satisfied on one face of the protein, however ambiguously, the subset of structures will be biased towards that side. In the MBDs, there is an absence of structured residues on the back face of loop L1, with which to form NOEs. This could prevent effective sampling of a whole range of alternate loop conformations. The presence of 2 unambiguous restraints between L1 and the hairpin indicates that this conformation is a likely, but not necessarily the only, possibility.

[1]to a first approximation, NOEs are proportional to r^{-6}, where r is the internuclear distance

6.1.2 MBD:DNA complex

Titration of methylated DNA into a labelled NMR sample allowed analysis of perturbations of the backbone NH chemical shifts. The observed effects were as expected from previous work on MeCP2 and MBD1. Although the L1 loop region in the MBD2 MBD may be more structured than in either of the two other proteins, large chemical shifts were still observed for almost all loop residues. This indicates that the loop conformation is still significantly altered by methylated DNA binding. Evidence of reduced flexibility was also found in the sequential assignment of the MBD:DNA complex. CBCA(CO)NH and HNCACB connections were much improved compared to the free domain. Figure 6.1 maps the 20 residues with the largest chemical shift changes onto the solved structure of the MBD. Almost all the residues are found in the DNA-binding interface proposed in the MBD1:DNA structure [100], including the five amino acids which form the methyl-recognising hydrophobic patch. Five residues (Lys 164, Glu 165, Phe 190, Gln 195 and Leu 196) are in regions that might reasonably be expected to undergo significant chemical shift changes, as neighbouring residues make contact with the DNA. The largest changes were observed for certain residues at the base of the loop: Ile 168, Gly 172, Ser 174 and Ala 175. Similar patterns of altered chemical shift were observed for both MBD2 and MBD4 MBDs, and these agree well with published data for MBD1. There is no evidence to suggest that the actual mode of methyl recognition varies between proteins.

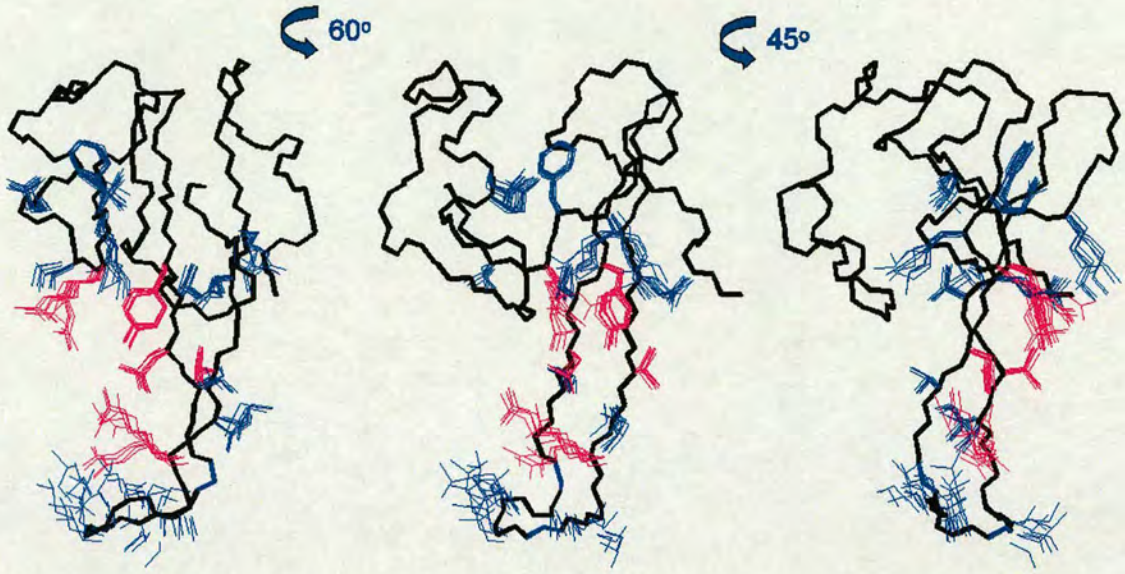

Figure 6.1: The largest changes in chemical shift, induced upon DNA binding, mapped onto the solved MBD structure. These include the five residues that form the hydrophobic, methyl-recognition patch (sidechains in magenta) as well as 15 other residues (sidechains in blue). Rotation between each depiction is as indicated.

6.2 Methylated DNA binding

Surface Plasmon Resonance (SPR) studies in this thesis provide the first reliable binding contants obtained for a MBD:DNA interaction. This is possible because the Biacore system allows determination of the rate constants to a high level of accuracy and reproducibility. Previous attempts to quantify the affinities have been by band shift experiments, using longer oligos with multiple methylation sites. These did not give information about any differences in recognition of a single meCpG.

MeCP2 and MBD4 MBDs have a higher affinity for methylated DNA, when compared to MBD2. None of the proteins bound unmethylated DNA, and each MBD had a similar affinity for all three methylated oligos. This suggests that the

integrity of a *symmetrically* methylated CpG is unimportant for the MBD:DNA interaction.

A recent publication reports the use of molecular dynamics simulations, starting from the MBD1:DNA structure, to investigate the finer details of methyl recognition [150]. In that study, the second methyl group was found to contribute an almost negligible amount to the total binding energy, which supports the conclusion that there is no difference in MBD affinity for symmetrically methylated, hemimethylated and CpG:TpG mismatched oligonucleotides by SPR.

The study found that the overal mode of DNA binding remained unchanged, but that one conserved residue in the DNA-binding interface, Asp 32 (Asp 179 in MBD2), could form a direct cytidine base contact, where before it had not. Two forms of direct contact were observed: In the minor form (for about 1 ns of the simulation), Asp 32 bridges the two methylated cytosines with its carboxylate group. This is stabilised by a hydrogen bond between one of the carboxylate oxygen atoms and the nearby hydroxyl group of tyrosine 34 (Tyr 181 in MBD2). In the major conformation, two carboxylate hydrogen bonds are made with the guanidinium group of Arg 22 (Arg 169 in MBD2) and one with N4 of an unmethylated, flanking cytosine.

These data suggest that there is a reasonable degree of flexibility within the recognition mechanism, and that these motions are not detectable on the NMR timescale (milliseconds). Such precise adjustments indicate that the methyl recognition mechanism may be more complex than previously thought.

6.2.1 Length and sequence dependence of MBD binding

By SPR it was observed that none of the MBDs bound a methylated 10-mer. The apparent dependence, of MBD:DNA binding, on the length of the oligonucleotide, will require further investigation. It could simply be due to steric hinderance with the streptavidin chip matrix.

A recent study found that the MeCP2 MBD specifically recognises an AT-rich region at least 4 b.p. away from the methylated CpGs [50]. The same study also found that while MBD2 was able to compensate for morpholino depleted MeCP2 at over half its binding sites, the reverse was not true. MeCP2 was only able to bind to 3 out of 25 sites previously occupied by MBD2. MeCP2 has been found to possess an AT hook *C-terminal* to the MBD, so it was thought that this domain might be responsible for the observations. However, a deletion analysis found that the MBD plus a 13 residue *N-terminal* extension, was the minimum required for AT-specific recognition. It is difficult to reconcile this information with the methyl-recognition mechanism that is currently understood. The domain itself is not large enough to have an additional DNA contact region, and the 13 residue extension is on the opposite side of the domain, away from the DNA-binding interface. It is perhaps just possible that an unstructured N-terminal extension may reach to an AT rich region, but this seems unlikely. AT and meCpG recognition therefore constitute two steps in the binding mode.

The reason why MeCP2 cannot compensate for MBD2, may be due to other domains outside the MBD, which stabilise binding to certain DNA sequences. Various different interaction partners, particular to each protein, may also have this effect. Either of these possibilities could explain a preference for specific

promoter regions.

6.3 Future structural investigations

6.3.1 MBD2 MBD:DNA complex

The solution structure of the MBD2 MBD:DNA complex could be solved by NMR, but sample stability is the main barrier to collecting adequate data. An attempt has been made to design more stable constructs, by using data from limited proteolysis. In order to produce these however, a large scale screen of MBD2b for suitable fragments will be necessary. An alternative is to produce mutants, or carry out chemical modifications that stabilse the MBD fold. Alkylation of the single N-terminal cysteine was achieved but had no effect on the lifetime of the sample.

Essential to solving the structure of the MBD:DNA complex, is optimisation of X-filtered experiments [138, 151]. These are used to measure intermolecular NOEs, for example between ^{15}N, ^{13}C-labelled protein and unlabelled DNA. NOEs may be measured from ^{13}C-^{1}H to all ^{12}C- and ^{15}N-attached protons, but not to those that are ^{13}C-attached. Since the protein is labelled, then observed NOEs to ^{12}C-^{1}H, must all be intermolecular and from the DNA. Intermolecular NOEs define individual interatomic contacts in much more detail compared to the more general information obtained from chemical shift perturbation experiments. A common problem for weak complexes, however, is that intermolecular NOEs are averaged out across the ensemble of structures, so that they effectively dissappear from the spectrum. If there is highly variable local geometry at the binding

surface, as suggested by molecular dynamics simulations [150], this could account for the difficulties experienced.

It would also be interesting to compare the structure to that of the MBD in complex with unmethylated DNA, if such a complex will form under conditions required for NMR. Titration of the MBD1 MBD with an unmethylated oligomer did not appear to produce a complex [99], but the proteins may be sufficiently different, that the experiment is worth repeating with MBD2.

6.3.2 Loop flexibility

The structure and function of the loop L1, is worthy of further investigation, to establish if it is a cause of observed differences in MBD:DNA affinity. Both local and global protein flexibility can be investigated using the relaxation properties of backbone amides. ^{15}N-relaxation experiments could be used to find out how flexible the L1 loop is, on a pico- to millisecond timescale. Using the ratio of the longitudinal to transverse relaxation times[2], backbone dynamics of the MBD2 MBD may be obtained and compared to those of other MBDs.

6.3.3 Prospects for crystallography

Attempts to crystallise the MBD2 MBD:DNA complex were unsuccessful. So far, no crystal structures of MBD-like folds have been published, so NMR is therefore the technique of choice for investigating this interaction. Ongoing work to crystallise the MeCP2 MBD:DNA complex has attempted to exploit the recently discovered sequence specificity of MeCP2 [50] in choosing the DNA oligomer.

[2]T1/T2 ratio

Co-crystallisation of the MeCP2 MBD:DNA complex has been achieved, but resolution below 4$\mathring{A}$ remains elusive [152]. Although the crystallographic approach has been unfruitful, it is still attractive because of the potential to discover the details of methyl recognition involving the termini of arginine side chains. It is not always possible to assign or even detect these in NMR experiments. On the other hand, if the mechanism involves a series of recognition steps and rearrangement of the MBD on the DNA, then a crystal structure will not necessarily provide information that distinguishes between MBDs. Work on the full-length proteins is likely to be hampered by unstructured regions connecting the various different motifs. These make the MBDs too flexible for crystallisation (see figure 6.2).

6.3.4 MBD2:protein interactions

Possibilities for future projects include solving the structure of selected regions of MBD2, in complex with various binding partners. One region of MBD2, 211-262, (363-414) specifically binds to hp66α [74], and is predicted to fold into a helical region. It has a high degree of similarity to a coiled-coil and probable protein-protein interaction domain [153] (see figure 6.2), and may provide the foundation of a discrete region of MBD2. This makes it a likely target for co-crystallisation and NMR studies of the complex with p66.

It may also be possible to understand how the repressive function of MBD2 is modulated by its interactions with certain proteins. MBDin reactivate promoters that have been silenced by MBD2 [83], and acts antagonistically to MIZF, which enhances MBD2-mediated repression [78]. Each partner binds to discrete regions of MBD2: MBDin to the C-terminus, residues 209-262 (361-414), and MIZF to a

region which overlaps with the C-terminal end of the MBD, residues 53-154 (205-306). It will, however, be difficult to evaluate these complexes in the context of the full-length protein structure (see section 6.3.3), which may be crucial to understanding the fine-tuning of the MBD2 gene-silencing mechanism.

It is hoped that dissecting these interactions will reveal much more about the various roles of MBD2, in effecting gene silencing due to altered chromatin architecture.

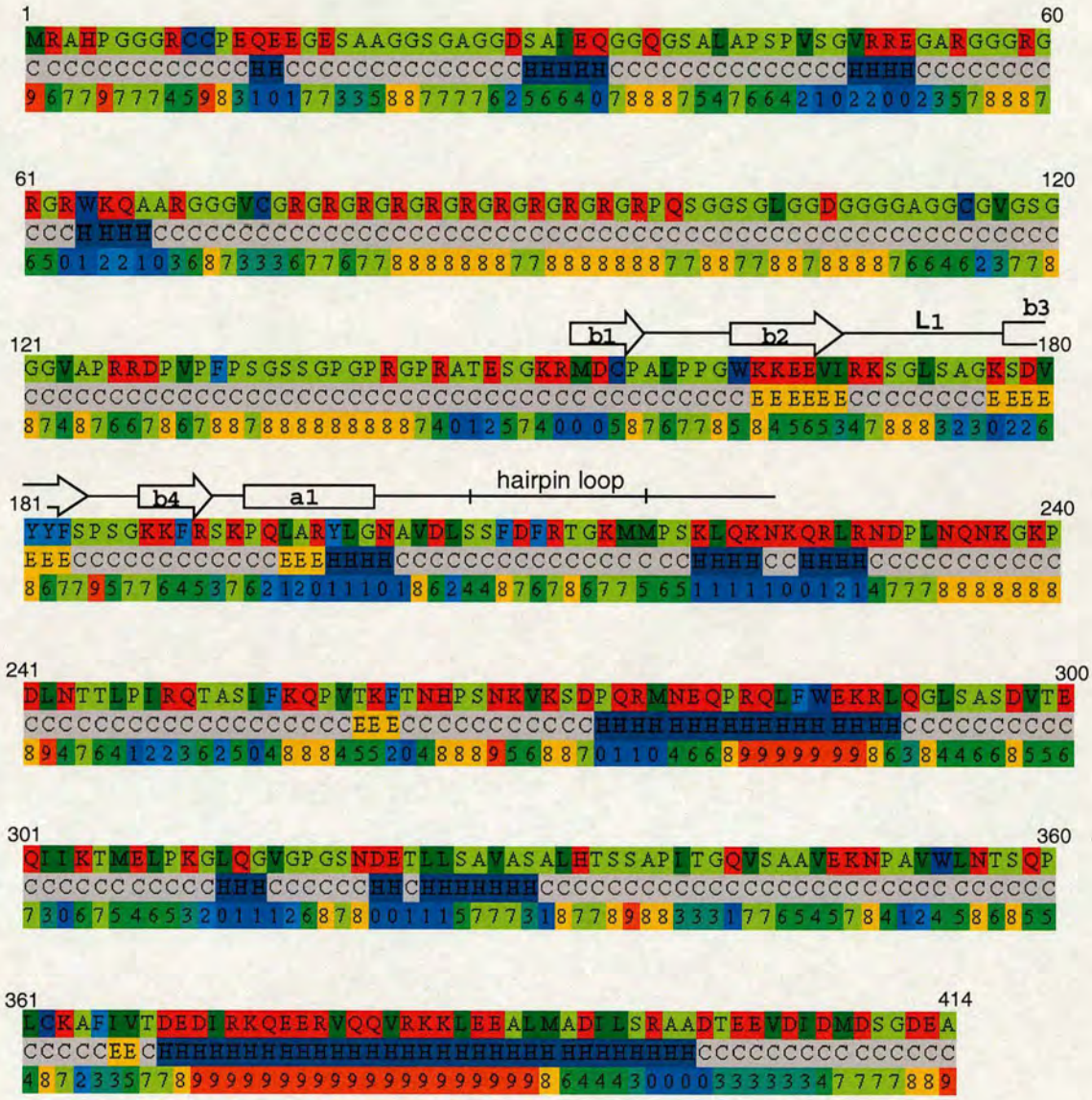

Figure 6.2: *Ab initio* structure prediction for full-length MBD2a, using the structure prediction tool, 3D-PSSM [153]. The amino acid sequence is shown on the top line, and regions predicted to be a defined secondary structure, are shown as follows: Alpha helices are denoted by H (blue), beta-strands by E (orange), and random coil, by C (grey). The confidence value of the prediction is shown as a number between 0 (lowest) and 9 (highest), on the line below the prediction. The secondary structure of the MBD, as determined in this project, is indicated in black above the corresponding portion of the sequence.

Appendix A

Laboratory Manual

A.1 Production of competent *E. coli*

A.1.1 The $CaCl_2$ method

1. Inoculate 50ml LB, 1:50 from an overnight culture of *E. coli*, and grow to $A_{600} \approx 0.5$.
2. Spin down at 3000 r.p.m. for 10 minutes.
3. Resuspend the cells in 20ml ice cold, 50mM $CaCl_2$, and leave on ice for 20 minutes, before spinning down again.
4. Resuspend the cells in 2ml ice cold, 50mM $CaCl_2$, and leave on ice for 2-24 hours maximum.

A.2 Production of labelled samples for NMR

To make 1L of 5x M9 minimal media:

64g $Na_2HPO_4.7H_2O$
15g KH_2PO_4
2.5g NaCl
Make up to 1L with deionised H_2O

Per 500ml flask of M9 minimal media:

100ml 5x M9
1ml 1M $MgSO_4$
50μl 1M $CaCl_2$
200μl 5mM $FeCl_3$
0.5g $^{15}NH_4Cl$ or 0.6g $(^{15}NH_4)_2SO_4$
1.5g glucose, ^{13}C-labelled or unlabelled as required
Make up to 500ml with deionised H_2O

A.3 Production of deuterated NaPi buffer

50ml of a buffer solution of 50mM NaPi, 50mM Nacl pH 7.0, was dessicated on a rotary evaporator overnight (full vacuum, 30°C). The buffer powder was then made up to 50ml using 99% D_2O (Sigma). The protein solution was concentrated and then buffer-exchanged into the 99% deuterated NaPi buffer using a 6ml vivaspin tube with a 5K cut-off (Vivascience). This was done twice so as to achieve a 500-fold dilution of any H_2O present. 2ml of 99.9% NaPi buffer was made by dissolving the required amount of dessicated buffer in 99.9% D_2O (Sigma). The process of buffer exchange was repeated with 99.9% deuterated NaPi buffer, to give the final 600μl sample for NMR.

Appendix B

Scripts

B.1 Processing scripts

The following are examples of scripts used to process multidimensional NMR data with AZARA.

B.1.1 FT processing of 2D ^{15}N HSQC

```
script_com 1
  complex
  gaussian_sw 20 1.3 0
  zerofill 1
  fft
  phase 0 25785
  phase -149 0
  reduce
  upper 1024
end_script

script_com 2
  mask_ppmm
  complex
!  conjugate
  sinebell2 70
  zerofill 1
  fft
  phase 90 -180
  reduce
end_script
```

B.1.2 MEM processing of 3D HCCH-TOCSY

```
input ser.par
output m.spc

script_com 1
  complex
  avance 12 16
  gaussian_sw 20 1.3 0
  zerofill 1
  fft
  avance_phase
  phase -148 0
  reduce
end_script

script_com 2
  mask_ppmm
  complex
 ! sinebell2 90
 ! zerofill 1
 ! fft
 ! phase 0 0
 ! reduce
end_script

script_com 3
  mask_ppmm
  complex
  conjugate
 ! sinebell2 90
 ! zerofill 1
 ! fft
 ! phase 90 -180
 ! reduce
end_script

maxent2_com 2 3

 iter 100
 noise 4500.0
log m.log
 rate 0.5

 dim 1
 npts 512
```

```
 complex
 phase 0 0

 dim 2
 npts 128
 complex
 phase 90 -180
end_maxent
```

B.1.3 Perl script to analyse MEM plane convergence

This script checks the convergence status of each plane, as reported in the m.log file. If a plane has not reached convergence ("status = 0"), then the following plane number is printed, allowing easy identification of non-converged planes.

```
#!/bin/perl
open (IN,"m.log") || die "File bogus:$!";
$i=0;
while (<IN>) {
                $Buff[$i]=$_;
                if(($i!=0)&&($_=~ m/^working/)){
                        if($Buff[($i-2)] !~ m/status = 0/)
                  {
                      print "$Buff[$i]";
                  }
                undef @Buff;
                 $i = 0;
                                                }
        $i++;

             }
if($Buff[($i-2)] !~ m/status = 0/)
                 {
                     print "final plane not converged either!\n";
                 }
close (IN);
```

B.2 Contour generation scripts

The following script generates contours which can be read by ANSIG. One contour description file is needed per 'view' of the data, so that for 3D data, two separate contours files are required.

```
input spc.par.ref
output cnt_complex

dims 1 2
levels  300000          -300000
levels  600000          -600000
levels  1200000         -1200000
levels  2400000         -2400000
levels  4800000         -4800000
levels  9600000         -9600000
levels  19200000        -19200000
levels  38400000        -38400000
levels  76800000        -76800000
levels  153600000       -153600000
```

Appendix C

Chemical Shift Tables

C.1 MBD2 MBD chemical shifts

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
His	151	HA	4.6644	CA	55.7340
His	151	HB1	3.0489	CB	30.4725
His	151	HB2	3.2099	CB	30.4725
His	151	HD2	7.0291	CD2	120.1173
His	151	HE1	-99	CE1	-99
His	151	HN	8.5769	N	120.3832
His	151	HD1	-99	ND1	-99
His	151	HE2	-99	NE2	-99
Ala	152	HA	4.5818	CA	52.0910
Ala	152	HB1	1.2345	CB	19.8585
Ala	152	HB2	1.2345	CB	19.8585
Ala	152	HB3	1.2345	CB	19.8585
Ala	152	HN	8.2994	N	125.3200
Met	153	HA	4.7723	CA	54.2948
Met	153	HB1	2.0536	CB	35.4305
Met	153	HB2	2.1889	CB	35.4305
Met	153	HE1	-99	CE	-99
Met	153	HE2	-99	CE	-99
Met	153	HE3	-99	CE	-99
Met	153	HG1	2.6213	CG	31.7143
Met	153	HG2	2.6213	CG	31.7143
Met	153	HN	8.7694	N	119.7930
Asp	154	HA	4.6671	CA	55.7302
Asp	154	HB1	2.5891	CB	41.3627
Asp	154	HB2	2.7336	CB	41.3627
Asp	154	HN	8.6777	N	121.8010
Cys	155	HA	5.1931	CA	56.0106
Cys	155	HB1	2.8017	CB	28.0069
Cys	155	HB2	2.9171	CB	28.0069
Cys	155	HN	8.3187	N	121.2535
Pro	156	HA	4.5923	CA	64.6403
Pro	156	HB1	2.0585	CB	32.2005
Pro	156	HB2	2.4092	CB	32.2005
Pro	156	HD1	3.7910	CD	51.5238
Pro	156	HD2	4.0580	CD	51.5238
Pro	156	HG1	2.0500	CG	27.0280
Pro	156	HG2	2.0500	CG	27.0280
Ala	157	HA	4.2660	CA	53.0040
Ala	157	HB1	1.4870	CB	20.0947
Ala	157	HB2	1.4870	CB	20.0947

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Ala	157	HB3	1.4870	CB	20.0947
Ala	157	HN	7.4833	N	118.1702
Leu	158	HA	4.2684	CA	52.0250
Leu	158	HB1	0.8960	CB	40.8548
Leu	158	HB2	1.4622	CB	40.8548
Leu	158	HD11	-0.3871	CD1	21.3382
Leu	158	HD12	-0.3871	CD1	21.3382
Leu	158	HD13	-0.3871	CD1	21.3382
Leu	158	HD21	0.3389	CD2	26.3615
Leu	158	HD22	0.3389	CD2	26.3615
Leu	158	HD23	0.3389	CD2	26.3615
Leu	158	HG	1.1632	CG	26.8223
Leu	158	HN	7.7237	N	119.1377
Pro	159	HA	4.6396	CA	62.2260
Pro	159	HB1	1.9338	CB	29.8250
Pro	159	HB2	2.6029	CB	29.8250
Pro	159	HD1	2.5239	CD	50.0594
Pro	159	HD2	3.1235	CD	50.0594
Pro	159	HG1	1.8094	CG	27.7590
Pro	159	HG2	1.8094	CG	27.7590
Pro	160	HA	4.3554	CA	64.6262
Pro	160	HB1	1.8825	CB	31.9757
Pro	160	HB2	2.3555	CB	31.9757
Pro	160	HD1	3.6324	CD	50.2838
Pro	160	HD2	3.8944	CD	50.2838
Pro	160	HG1	2.0670	CG	27.9922
Pro	160	HG2	2.1601	CG	27.9922
Gly	161	HA1	3.6729	CA	45.3992
Gly	161	HA2	4.2314	CA	45.3992
Gly	161	HN	8.6522	N	111.7470
Trp	162	HA	4.6972	CA	59.2150
Trp	162	HB1	3.1265	CB	30.4974
Trp	162	HB2	3.3011	CB	30.4974
Trp	162	HD1	7.3423	CD1	127.5670
Trp	162	HE3	7.3966	CE3	120.4127
Trp	162	HH2	6.2474	CH2	124.0740
Trp	162	HZ2	7.4370	CZ2	113.6818
Trp	162	HZ3	6.4834	CZ3	121.8910
Trp	162	HN	7.8025	N	121.3900
Trp	162	HE1	10.8199	NE1	111.2979
Lys	163	HA	5.2340	CA	54.7264
Lys	163	HB1	1.6673	CB	37.9311
Lys	163	HB2	1.7043	CB	37.9311
Lys	163	HD1	1.3796	CD	29.2495
Lys	163	HD2	1.4655	CD	29.2495
Lys	163	HE1	2.7930	CE	42.3500
Lys	163	HE2	2.7930	CE	42.3500
Lys	163	HG1	1.3286	CG	25.0675
Lys	163	HG2	1.3286	CG	25.0675
Lys	163	HN	8.8908	N	120.3555
Lys	163	HZ1	-99	NZ	-99
Lys	163	HZ2	-99	NZ	-99
Lys	163	HZ3	-99	NZ	-99
Lys	164	HA	4.9291	CA	54.7059
Lys	164	HB1	1.5036	CB	36.5572
Lys	164	HB2	1.6521	CB	36.5572
Lys	164	HD1	1.4467	CD	29.0057
Lys	164	HD2	1.5667	CD	29.0057
Lys	164	HE1	2.7331	CE	41.8890
Lys	164	HE2	2.7331	CE	41.8890
Lys	164	HG1	0.7141	CG	24.3275

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Lys	164	HG2	0.8136	CG	24.3275
Lys	164	HN	9.2145	N	122.4179
Lys	164	HZ1	-99	NZ	-99
Lys	164	HZ2	-99	NZ	-99
Lys	164	HZ3	-99	NZ	-99
Glu	165	HA	4.7586	CA	55.2610
Glu	165	HB1	1.7616	CB	34.9053
Glu	165	HB2	2.0299	CB	34.9053
Glu	165	HG1	2.0957	CG	36.6602
Glu	165	HG2	2.1562	CG	36.6602
Glu	165	HN	9.2935	N	128.5810
Glu	166	HA	5.1023	CA	54.7432
Glu	166	HB1	1.7581	CB	31.9631
Glu	166	HB2	1.8942	CB	31.9631
Glu	166	HG1	1.9999	CG	36.5735
Glu	166	HG2	2.0776	CG	36.5735
Glu	166	HN	8.4923	N	124.5695
Val	167	HA	4.3662	CA	61.1991
Val	167	HB	2.0508	CB	34.4603
Val	167	HG11	0.8957	CG1	21.0730
Val	167	HG12	0.8957	CG1	21.0730
Val	167	HG13	0.8957	CG1	21.0730
Val	167	HG21	0.8957	CG2	21.0730
Val	167	HG22	0.8957	CG2	21.0730
Val	167	HG23	0.8957	CG2	21.0730
Val	167	HN	9.0551	N	125.8223
Ile	168	HA	4.2007	CA	60.9413
Ile	168	HB	1.7506	CB	38.3219
Ile	168	HD11	0.7650	CD1	12.5736
Ile	168	HD12	0.7650	CD1	12.5736
Ile	168	HD13	0.7650	CD1	12.5736
Ile	168	HG11	1.1312	CG1	27.9815
Ile	168	HG12	1.4552	CG1	27.9815
Ile	168	HG21	0.7987	CG2	17.3899
Ile	168	HG22	0.7987	CG2	17.3899
Ile	168	HG23	0.7987	CG2	17.3899
Ile	168	HN	8.7292	N	128.5775
Arg	169	HA	4.3894	CA	55.9981
Arg	169	HB1	1.8422	CB	30.7510
Arg	169	HB2	1.8422	CB	30.7510
Arg	169	HD1	3.0567	CD	43.3530
Arg	169	HD2	3.1818	CD	43.3530
Arg	169	HG1	1.6507	CG	27.3398
Arg	169	HG2	1.6507	CG	27.3398
Arg	169	HN	8.7541	N	127.5865
Arg	169	HE	-99	NE	-99
Arg	169	HH11	-99	NH1	-99
Arg	169	HH12	-99	NH1	-99
Arg	169	HH21	-99	NH2	-99
Arg	169	HH22	-99	NH2	-99
Lys	170	HA	4.3582	CA	56.8631
Lys	170	HB1	1.7653	CB	33.4596
Lys	170	HB2	1.9007	CB	33.4596
Lys	170	HD1	1.6617	CD	28.9216
Lys	170	HD2	1.6617	CD	28.9216
Lys	170	HE1	2.9724	CE	42.0726
Lys	170	HE2	2.9724	CE	42.0726
Lys	170	HG1	1.4029	CG	24.9132
Lys	170	HG2	1.4942	CG	24.9132
Lys	170	HN	8.8765	N	124.2243

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Lys	170	HZ1	-99	NZ	-99
Lys	170	HZ2	-99	NZ	-99
Lys	170	HZ3	-99	NZ	-99
Ser	171	HA	4.6061	CA	57.5645
Ser	171	HB1	3.8050	CB	64.7327
Ser	171	HB2	3.8516	CB	64.7327
Ser	171	HN	8.0482	N	114.3088
Gly	172	HA1	3.8955	CA	45.1314
Gly	172	HA2	4.1807	CA	45.1314
Gly	172	HN	8.6394	N	128.4930
Leu	173	HA	4.2365	CA	56.4877
Leu	173	HB1	1.6110	CB	42.1219
Leu	173	HB2	1.6968	CB	42.1219
Leu	173	HD11	0.8835	CD1	23.3074
Leu	173	HD12	0.8835	CD1	23.3074
Leu	173	HD13	0.8835	CD1	23.3074
Leu	173	HD21	0.9293	CD2	25.0078
Leu	173	HD22	0.9293	CD2	25.0078
Leu	173	HD23	0.9293	CD2	25.0078
Leu	173	HG	1.6761	CG	27.1256
Leu	173	HN	8.5457	N	121.7170
Ser	174	HA	4.4692	CA	57.7268
Ser	174	HB1	3.9135	CB	63.2041
Ser	174	HB2	3.9135	CB	63.2041
Ser	174	HN	8.4364	N	114.3615
Ala	175	HA	3.9146	CA	53.2730
Ala	175	HB1	1.2697	CB	18.5068
Ala	175	HB2	1.2697	CB	18.5068
Ala	175	HB3	1.2697	CB	18.5068
Ala	175	HN	7.6921	N	123.6214
Gly	176	HA1	3.7237	CA	45.1180
Gly	176	HA2	4.1730	CA	45.1180
Gly	176	HN	8.5960	N	128.4090
Lys	177	HA	4.4562	CA	55.9843
Lys	177	HB1	1.8675	CB	33.6823
Lys	177	HB2	1.8675	CB	33.6823
Lys	177	HD1	1.6644	CD	29.2508
Lys	177	HD2	1.6644	CD	29.2508
Lys	177	HE1	2.9863	CE	42.2345
Lys	177	HE2	2.9863	CE	42.2345
Lys	177	HG1	1.4201	CG	25.0517
Lys	177	HG2	1.4267	CG	25.0517
Lys	177	HN	8.1182	N	121.1616
Lys	177	HZ1	-99	NZ	-99
Lys	177	HZ2	-99	NZ	-99
Lys	177	HZ3	-99	NZ	-99
Ser	178	HA	5.0381	CA	57.3042
Ser	178	HB1	3.6908	CB	65.0975
Ser	178	HB2	3.6908	CB	65.0975
Ser	178	HN	8.5325	N	116.9395
Asp	179	HA	4.9703	CA	53.7873
Asp	179	HB1	2.4281	CB	44.3409
Asp	179	HB2	2.4534	CB	44.3409
Asp	179	HN	8.7731	N	123.1056
Val	180	HA	4.5545	CA	61.4522
Val	180	HB	1.7865	CB	34.1932
Val	180	HG11	0.5099	CG1	21.5936
Val	180	HG12	0.5099	CG1	21.5936
Val	180	HG13	0.5099	CG1	21.5936
Val	180	HG21	0.8477	CG2	21.8086
Val	180	HG22	0.8477	CG2	21.8086

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Val	180	HG23	0.8477	CG2	21.8086
Val	180	HN	8.1156	N	121.5825
Tyr	181	HA	5.3088	CA	54.9236
Tyr	181	HB1	2.6036	CB	42.4050
Tyr	181	HB2	2.7151	CB	42.4050
Tyr	181	HD1	6.9016	CD1	133.2053
Tyr	181	HD2	6.9016	CD2	133.2053
Tyr	181	HE1	6.7064	CE1	118.1469
Tyr	181	HE2	6.7064	CE2	118.1469
Tyr	181	HN	8.5556	N	122.4151
Tyr	182	HA	5.6294	CA	56.7410
Tyr	182	HB1	2.8043	CB	43.1248
Tyr	182	HB2	3.2652	CB	43.1248
Tyr	182	HD1	6.7909	CD1	132.8766
Tyr	182	HD2	6.7909	CD2	132.8766
Tyr	182	HE1	5.8431	CE1	117.1750
Tyr	182	HE2	5.8431	CE2	117.1750
Tyr	182	HN	8.4401	N	116.1650
Phe	183	HA	5.8210	CA	55.5980
Phe	183	HB1	2.6072	CB	42.8580
Phe	183	HB2	2.7491	CB	42.8580
Phe	183	HD1	6.8329	CD1	131.2890
Phe	183	HD2	6.8329	CD2	131.2890
Phe	183	HE1	7.1909	CE1	131.2906
Phe	183	HE2	7.1909	CE2	131.2906
Phe	183	HZ	7.0812	CZ	129.2930
Phe	183	HN	9.5078	N	118.9930
Ser	184	HA	1.9046	CA	55.7342
Ser	184	HB1	3.2160	CB	62.6288
Ser	184	HB2	3.8713	CB	62.6288
Ser	184	HN	8.9092	N	119.1280
Pro	185	HA	4.1151	CA	65.3240
Pro	185	HB1	1.8971	CB	31.2974
Pro	185	HB2	2.3197	CB	31.2974
Pro	185	HD1	1.5622	CD	48.9806
Pro	185	HD2	3.3169	CD	48.9806
Pro	185	HG1	1.6860	CG	27.7541
Pro	185	HG2	2.0900	CG	27.7541
Ser	186	HA	4.4210	CA	58.1136
Ser	186	HB1	3.8296	CB	63.4440
Ser	186	HB2	4.0720	CB	63.4440
Ser	186	HN	-99	N	-99
Gly	187	HA1	3.4727	CA	45.3334
Gly	187	HA2	4.2013	CA	45.3334
Gly	187	HN	8.3280	N	112.1221
Lys	188	HA	4.0781	CA	56.5146
Lys	188	HB1	1.4509	CB	32.9613
Lys	188	HB2	1.5486	CB	32.9613
Lys	188	HD1	1.4226	CD	29.0072
Lys	188	HD2	1.5503	CD	29.0072
Lys	188	HE1	2.7587	CE	42.1728
Lys	188	HE2	2.9568	CE	42.1728
Lys	188	HG1	0.9118	CG	25.1375
Lys	188	HG2	1.2175	CG	25.1375
Lys	188	HN	7.5818	N	122.5470
Lys	188	HZ1	-99	NZ	-99
Lys	188	HZ2	-99	NZ	-99
Lys	188	HZ3	-99	NZ	-99
Lys	189	HA	4.5237	CA	55.7323
Lys	189	HB1	1.4310	CB	35.1999
Lys	189	HB2	1.4310	CB	35.1999

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Lys	189	HD1	1.1944	CD	29.7534
Lys	189	HD2	1.2146	CD	29.7534
Lys	189	HE1	2.2298	CE	41.3606
Lys	189	HE2	2.2298	CE	41.3606
Lys	189	HG1	0.5641	CG	25.0340
Lys	189	HG2	0.7263	CG	25.0340
Lys	189	HN	8.0756	N	124.0375
Lys	189	HZ1	-99	NZ	-99
Lys	189	HZ2	-99	NZ	-99
Lys	189	HZ3	-99	NZ	-99
Phe	190	HA	4.6884	CA	58.0154
Phe	190	HB1	3.0134	CB	42.8819
Phe	190	HB2	3.1861	CB	42.8819
Phe	190	HD1	7.1748	CD1	131.5333
Phe	190	HD2	7.1748	CD2	131.5333
Phe	190	HE1	6.9820	CE1	131.5205
Phe	190	HE2	6.9820	CE2	131.5205
Phe	190	HZ	5.9041	CZ	129.4970
Phe	190	HN	9.6003	N	124.2634
Arg	191	HA	5.1408	CA	55.5985
Arg	191	HB1	1.9961	CB	32.9257
Arg	191	HB2	2.0789	CB	32.9257
Arg	191	HD1	3.2216	CD	43.3637
Arg	191	HD2	3.2216	CD	43.3637
Arg	191	HG1	1.5672	CG	27.3634
Arg	191	HG2	1.6837	CG	27.3634
Arg	191	HN	9.1822	N	118.4686
Arg	191	HE	-99	NE	-99
Arg	191	HH11	-99	NH1	-99
Arg	191	HH12	-99	NH1	-99
Arg	191	HH21	-99	NH2	-99
Arg	191	HH22	-99	NH2	-99
Ser	192	HA	3.8897	CA	56.7478
Ser	192	HB1	3.7786	CB	66.3836
Ser	192	HB2	3.8735	CB	66.3836
Ser	192	HN	7.7842	N	113.9360
Lys	193	HA	4.0517	CA	61.4454
Lys	193	HB1	1.9137	CB	30.1441
Lys	193	HB2	2.0734	CB	30.1441
Lys	193	HD1	1.6591	CD	30.2809
Lys	193	HD2	1.6591	CD	30.2809
Lys	193	HE1	3.1807	CE	43.3613
Lys	193	HE2	3.1807	CE	43.3613
Lys	193	HG1	1.4352	CG	26.7834
Lys	193	HG2	1.4352	CG	26.7834
Lys	193	HN	8.9354	N	123.0943
Lys	193	HZ1	-99	NZ	-99
Lys	193	HZ2	-99	NZ	-99
Lys	193	HZ3	-99	NZ	-99
Pro	194	HA	4.3574	CA	66.1382
Pro	194	HB1	1.7874	CB	31.0173
Pro	194	HB2	2.3422	CB	31.0173
Pro	194	HD1	3.6975	CD	49.8194
Pro	194	HD2	3.7455	CD	49.8194
Pro	194	HG1	1.9231	CG	28.5886
Pro	194	HG2	2.1379	CG	28.5886
Gln	195	HA	4.1426	CA	58.9625
Gln	195	HB1	2.3044	CB	29.2231
Gln	195	HB2	2.5618	CB	29.2231
Gln	195	HG1	2.5362	CG	33.6748
Gln	195	HG2	2.7205	CG	33.6748

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Gln	195	HN	7.1506	N	115.8271
Gln	195	HE21	6.8275	NE2	112.6522
Gln	195	HE22	7.7664	NE2	112.6522
Leu	196	HA	3.7396	CA	58.0679
Leu	196	HB1	1.5407	CB	43.6004
Leu	196	HB2	1.7966	CB	43.6004
Leu	196	HD11	0.1129	CD1	25.5584
Leu	196	HD12	0.1129	CD1	25.5584
Leu	196	HD13	0.1129	CD1	25.5584
Leu	196	HD21	0.1503	CD2	-99
Leu	196	HD22	0.1503	CD2	-99
Leu	196	HD23	0.1503	CD2	-99
Leu	196	HG	1.1158	CG	26.6625
Leu	196	HN	7.8744	N	122.4750
Ala	197	HA	3.9554	CA	54.5540
Ala	197	HB1	1.3627	CB	18.3397
Ala	197	HB2	1.3627	CB	18.3397
Ala	197	HB3	1.3627	CB	18.3397
Ala	197	HN	8.9436	N	120.2710
Arg	198	HA	4.0634	CA	58.7473
Arg	198	HB1	1.9788	CB	30.3463
Arg	198	HB2	1.9831	CB	30.3463
Arg	198	HD1	3.2491	CD	43.5481
Arg	198	HD2	3.2956	CD	43.5481
Arg	198	HG1	1.6967	CG	27.7650
Arg	198	HG2	1.8396	CG	27.7650
Arg	198	HN	7.9443	N	117.9353
Arg	198	HE	-99	NE	-99
Arg	198	HH11	-99	NH1	-99
Arg	198	HH12	-99	NH1	-99
Arg	198	HH21	-99	NH2	-99
Arg	198	HH22	-99	NH2	-99
Tyr	199	HA	4.3200	CA	61.1987
Tyr	199	HB1	2.8519	CB	40.3927
Tyr	199	HB2	3.3992	CB	40.3927
Tyr	199	HD1	7.1804	CD1	133.4910
Tyr	199	HD2	7.1804	CD2	133.4910
Tyr	199	HE1	7.0829	CE1	118.3859
Tyr	199	HE2	7.0829	CE2	118.3859
Tyr	199	HN	7.4993	N	119.2890
Leu	200	HA	4.0354	CA	56.2165
Leu	200	HB1	1.1551	CB	42.8975
Leu	200	HB2	1.6517	CB	42.8975
Leu	200	HD11	0.0298	CD1	25.7551
Leu	200	HD12	0.0298	CD1	25.7551
Leu	200	HD13	0.0298	CD1	25.7551
Leu	200	HD21	0.7593	CD2	23.4558
Leu	200	HD22	0.7593	CD2	23.4558
Leu	200	HD23	0.7593	CD2	23.4558
Leu	200	HG	1.4399	CG	27.5311
Leu	200	HN	8.1998	N	117.6693
Gly	201	HA1	3.7476	CA	45.8535
Gly	201	HA2	3.9738	CA	45.8535
Gly	201	HN	7.7981	N	125.7620
Asn	202	HA	4.5078	CA	53.7890
Asn	202	HB1	2.8064	CB	38.1427
Asn	202	HB2	2.8064	CB	38.1427
Asn	202	HN	8.5099	N	118.1341
Asn	202	HD21	6.9203	ND2	112.4569
Asn	202	HD22	7.6186	ND2	112.4569
Ala	203	HA	4.1326	CA	54.0093

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Ala	203	HB1	1.4149	CB	19.3647
Ala	203	HB2	1.4149	CB	19.3647
Ala	203	HB3	1.4149	CB	19.3647
Ala	203	HN	7.9741	N	121.1421
Val	204	HA	4.0569	CA	60.2407
Val	204	HB	1.6617	CB	34.4857
Val	204	HG11	0.5401	CG1	21.3147
Val	204	HG12	0.5401	CG1	21.3147
Val	204	HG13	0.5401	CG1	21.3147
Val	204	HG21	0.6438	CG2	21.0770
Val	204	HG22	0.6438	CG2	21.0770
Val	204	HG23	0.6438	CG2	21.0770
Val	204	HN	7.0464	N	113.7798
Asp	205	HA	4.6284	CA	53.2282
Asp	205	HB1	2.5805	CB	40.6151
Asp	205	HB2	2.8140	CB	40.6151
Asp	205	HN	8.4176	N	124.0635
Leu	206	HA	4.3709	CA	53.5287
Leu	206	HB1	1.6119	CB	41.4044
Leu	206	HB2	1.6119	CB	41.4044
Leu	206	HD11	-0.0303	CD1	22.2246
Leu	206	HD12	-0.0303	CD1	22.2246
Leu	206	HD13	-0.0303	CD1	22.2246
Leu	206	HD21	0.3633	CD2	25.7720
Leu	206	HD22	0.3633	CD2	25.7720
Leu	206	HD23	0.3633	CD2	25.7720
Leu	206	HG	1.3819	CG	25.9936
Leu	206	HN	9.1064	N	126.8948
Ser	207	HA	4.0672	CA	63.1760
Ser	207	HB1	3.8828	CB	62.3925
Ser	207	HB2	3.9468	CB	62.3925
Ser	207	HN	8.8169	N	120.7305
Ser	208	HA	-99	CA	-99
Ser	208	HB1	-99	CB	-99
Ser	208	HB2	-99	CB	-99
Ser	208	HN	8.7583	N	115.9256
Phe	209	HA	4.3680	CA	53.5240
Phe	209	HB1	1.6122	CB	41.2111
Phe	209	HB2	2.6800	CB	41.2111
Phe	209	HD1	-99	CD1	-99
Phe	209	HD2	-99	CD2	-99
Phe	209	HE1	-99	CE1	-99
Phe	209	HE2	-99	CE2	-99
Phe	209	HZ	-99	CZ	-99
Phe	209	HN	7.5276	N	123.6590
Asp	210	HA	4.6580	CA	52.8045
Asp	210	HB1	2.3230	CB	41.8540
Asp	210	HB2	2.8718	CB	41.8540
Asp	210	HN	-99	N	-99
Phe	211	HA	3.6177	CA	61.6238
Phe	211	HB1	2.9406	CB	41.3836
Phe	211	HB2	3.2426	CB	41.3836
Phe	211	HD1	7.0113	CD1	131.7530
Phe	211	HD2	7.0113	CD2	131.7530
Phe	211	HE1	7.2400	CE1	131.7851
Phe	211	HE2	7.2400	CE2	131.7851
Phe	211	HZ	-99	CZ	-99
Phe	211	HN	8.9423	N	127.9309
Arg	212	HA	4.0367	CA	-99
Arg	212	HB1	1.9272	CB	29.8990
Arg	212	HB2	2.0018	CB	29.8990

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Arg	212	HD1	3.2650	CD	42.8715
Arg	212	HD2	3.2650	CD	42.8715
Arg	212	HG1	1.6600	CG	27.7567
Arg	212	HG2	1.7789	CG	27.7567
Arg	212	HN	8.4421	N	114.7382
Arg	212	HE	7.0000	NE	-99
Arg	212	HH11	-99	NH1	-99
Arg	212	HH12	-99	NH1	-99
Arg	212	HH21	-99	NH2	-99
Arg	212	HH22	-99	NH2	-99
Thr	213	HA	4.2737	CA	61.7790
Thr	213	HB	4.1275	CB	70.8670
Thr	213	HG21	1.1258	CG2	53.2810
Thr	213	HG22	1.1258	CG2	53.2810
Thr	213	HG23	1.1258	CG2	53.2810
Thr	213	HN	-99	N	-99
Gly	214	HA1	3.7067	CA	47.4420
Gly	214	HA2	3.8297	CA	47.4420
Gly	214	HN	8.3153	N	114.1023
Lys	215	HA	4.4440	CA	54.2600
Lys	215	HB1	1.4826	CB	35.4457
Lys	215	HB2	1.6980	CB	35.4457
Lys	215	HD1	1.5823	CD	29.1283
Lys	215	HD2	1.6223	CD	29.1283
Lys	215	HE1	2.9564	CE	42.1463
Lys	215	HE2	2.9564	CE	42.1463
Lys	215	HG1	1.2821	CG	24.8049
Lys	215	HG2	1.2821	CG	24.8049
Lys	215	HN	-99	N	-99
Lys	215	HZ1	-99	NZ	-99
Lys	215	HZ2	-99	NZ	-99
Lys	215	HZ3	-99	NZ	-99
Met	216	HA	4.6937	CA	54.0994
Met	216	HB1	1.9172	CB	31.9149
Met	216	HB2	2.0003	CB	31.9149
Met	216	HE1	-99	CE	-99
Met	216	HE2	-99	CE	-99
Met	216	HE3	-99	CE	-99
Met	216	HG1	2.6168	CG	32.7207
Met	216	HG2	2.6906	CG	32.7207
Met	216	HN	8.5844	N	118.9710
Met	217	HA	4.7957	CA	53.0304
Met	217	HB1	1.9552	CB	32.2100
Met	217	HB2	2.0739	CB	32.2100
Met	217	HE1	-99	CE	-99
Met	217	HE2	-99	CE	-99
Met	217	HE3	-99	CE	-99
Met	217	HG1	2.4938	CG	32.2076
Met	217	HG2	2.5902	CG	32.2076
Met	217	HN	8.5628	N	124.5046
Pro	218	HA	4.4149	CA	63.1740
Pro	218	HB1	1.8936	CB	32.1998
Pro	218	HB2	2.2936	CB	32.1998
Pro	218	HD1	3.6678	CD	50.6137
Pro	218	HD2	3.7657	CD	50.6137
Pro	218	HG1	2.0071	CG	27.4993
Pro	218	HG2	2.0071	CG	27.4993
Ser	219	HA	4.3589	CA	58.4715
Ser	219	HB1	3.8203	CB	63.9173
Ser	219	HB2	3.8556	CB	63.9173
Ser	219	HN	8.4666	N	116.6190

Residue	Residue no.	^{1}H	δ (ppm)	Attached heteroatom	δ (ppm)
Lys	220	HA	4.2860	CA	56.3410
Lys	220	HB1	1.7220	CB	33.1140
Lys	220	HB2	1.8340	CB	33.1140
Lys	220	HD1	1.6490	CD	29.1500
Lys	220	HD2	1.6490	CD	29.1500
Lys	220	HE1	-99	CE	-99
Lys	220	HE2	-99	CE	-99
Lys	220	HG1	1.4100	CG	24.6880
Lys	220	HG2	1.4100	CG	24.6880
Lys	220	HN	8.4492	N	123.4920
Lys	220	HZ1	-99	NZ	-99
Lys	220	HZ2	-99	NZ	-99
Lys	220	HZ3	-99	NZ	-99
Leu	221	HA	4.2947	CA	55.1633
Leu	221	HB1	1.5701	CB	42.3704
Leu	221	HB2	1.5701	CB	42.3704
Leu	221	HD11	0.8397	CD1	23.4814
Leu	221	HD12	0.8397	CD1	23.4814
Leu	221	HD13	0.8397	CD1	23.4814
Leu	221	HD21	0.8994	CD2	24.8087
Leu	221	HD22	0.8994	CD2	24.8087
Leu	221	HD23	0.8994	CD2	24.8087
Leu	221	HG	1.5698	CG	27.0150
Leu	221	HN	8.2052	N	122.9260
Gln	222	HA	4.3022	CA	55.5958
Gln	222	HB1	1.9636	CB	29.5072
Gln	222	HB2	2.0708	CB	29.5072
Gln	222	HG1	2.3301	CG	33.7073
Gln	222	HG2	2.3301	CG	33.7073
Gln	222	HN	8.3162	N	121.9090
Gln	222	HE21	6.8920	NE2	113.0222
Gln	222	HE22	7.6102	NE2	113.0222
Lys	223	HA	4.3022	CA	56.2820
Lys	223	HB1	1.7223	CB	33.1155
Lys	223	HB2	1.8333	CB	33.1155
Lys	223	HD1	1.6585	CD	29.1421
Lys	223	HD2	1.6585	CD	29.1421
Lys	223	HE1	2.9672	CE	42.1227
Lys	223	HE2	2.9672	CE	42.1227
Lys	223	HG1	1.4015	CG	24.6904
Lys	223	HG2	1.4253	CG	24.6904
Lys	223	HN	8.4287	N	123.9314
Lys	223	HZ1	-99	NZ	-99
Lys	223	HZ2	-99	NZ	-99
Lys	223	HZ3	-99	NZ	-99
Asn	224	HA	4.4540	CA	54.7660
Asn	224	HB1	2.6476	CB	40.3766
Asn	224	HB2	2.7524	CB	40.3766
Asn	224	HN	8.1439	N	125.8620
Asn	224	HD21	6.8440	ND2	113.0265
Asn	224	HD22	7.5483	ND2	113.0265

Table C.1: Table of mean chemical shifts of the MBD2 MBD, based on all spectra. "-99" indicates that the atom could not be detected in the data and is therefore unassigned.

Appendix D

SPR Data

A complete list of all the SPR data for systems in which binding constants could be measured.

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	8.0x10^4	0.023	3.5x10^6	2.9x10^{-7}	0.38
$MBD2_2$	7.3x10^4	0.024	3.1x10^6	3.2x10^{-7}	0.28
$MeCP2_1$	1.5x10^6	0.040	3.6x10^7	2.75x10^{-8}	1.04
$MeCP2_2$	1.4x10^6	0.045	3.1x10^7	3.2x10^{-8}	0.66
$MBD4_1$	7.6x10^5	0.054	1.4x10^7	7.2x10^{-8}	0.87
$MBD4_2$	2.0x10^6	0.066	3.0x10^7	3.4x10^{-8}	0.42

Table D.1: Analysis of binding to me12mer.

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	4.7x10^4	0.015	3.1x10^6	3.2x10^{-7}	0.63
$MBD2_2$	5.0x10^4	0.016	3.1x10^6	3.2x10^{-7}	0.38
$MeCP2_1$	9.4x10^5	0.013	7.3x10^7	1.4x10^{-8}	2.45
$MeCP2_2$	9.4x10^5	0.013	7.1x10^7	1.4x10^{-8}	1.96
$MBD4_1$	1.5x10^6	0.05	3.0x10^7	3.4x10^{-8}	0.53
$MBD4_2$	1.5x10^6	0.05	2.8x10^7	3.5x10^{-8}	0.44

Table D.2: Analysis of binding to hemi12mer.

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	3.0x10^4	0.013	2.4x10^6	4.2x10^{-7}	0.35
$MBD2_2$	2.9x10^4	0.013	2.2x10^6	4.6x10^{-7}	0.33
$MeCP2_1$	5.1x10^5	0.016	3.2x10^7	3.1x10^{-8}	3.1
$MeCP2_2$	4.6x10^5	0.016	2.9x10^7	3.4x10^{-8}	2.8
$MBD4_1$	1.3x10^6	0.038	3.4x10^7	3.0x10^{-8}	1.1
$MBD4_2$	1.0x10^6	0.040	2.5x10^7	4.0x10^{-8}	0.8

Table D.3: Analysis of binding to T12mer.

MBD	$ka\ (M^{-1}s^{-1})$	$kd\ (s^{-1})$	$K_A\ (M^{-1})$	$K_D\ (M)$	χ^2
$MBD2_1$	2.4x10^4	0.02	1.2x10^6	8.2x10^{-7}	10.50
$MBD2_2$	2.1x10^4	0.02	1.0x10^6	9.9x10^{-7}	7.85

Table D.4: Analysis of binding to me10mer.

Bibliography

1. Wolffe, A. P. *Chromatin*, chapter 6, 23–24. Academic Press, third edition (1998).

2. Luger, K., Maeder, A. W., Richmond, R. K. & Sargent, D. F. Crystal structure of the nucleosome core particle at 2.8Åresolution. *Nature* **389**, 2512–260 (1997).

3. http://www.bio-itworld.com/archive/100902/banking.html.

4. Voet, D. & Voet, J. G. *Biochemistry*, chapter 33, 1123–1133. Wiley, John & Sons, inc., second edition (1995).

5. Wolffe, A. P. & Matzke, M. A. Epigenetics: regulation through repression. *Science* **286**, 481–486 (1999).

6. Margueron, R., Trojer, P., & Reinberg, D. The key to development: interpreting the histone code? *Current opinion in genetics & development* **15**, 163–176 (2005).

7. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

8. Agalioti, T., Chen, G. & Thanos, D. Deciphering the transcriptional histone acetylation code for a human gene. *Cell* **111**, 381–392 (2002).

9. Roth, S. Y., Denu, J. M. & Allis, C. D. Histone acetyltransferases. *Annual Review of Biochemistry* **70**, 81–120 (2001).

10. Carrozza, M. J., Utley, R. T., Workman, J. L & Cote, J. The diverse functions of histone acetyltransferase complexes. *Trends in Genetics* **19**, 321–329 (2003).

11. Marks, P. A. Miller, T. . R. V. M. Histone deacetylases. *Current Opinion in Pharmacology* **3**, 344–351 (2003).

12. Knoepfler, P. S. & Eisenman, R. N. SIN meets NuRD and other tails of repression. *Cell* **99**, 447–450 (1999).

13. You, A., Tong, J. K., Grozinger, C. M. & Schreiber, S. L. CoREST is an integral component of the CoREST-human histone deacetylase complex. *PNAS* **98**, 1454–1458 (2001).

14. Li, J., Wang, J., Nawaz, Z., Liu, J. M., Qin, J. & Wong, J. Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. *EMBO Journal* **19**, 4342–4350 (2000).

15. Vermeulen, M., Carrozza, M. J., Lasonder, E., Workman, J. L., Logie, C. & Stunnenberg, H. G. In vitro targeting reveals intrinsic histone tail specificity of the Sin3/histone deacetylase and N-CoR/SMRT corepressor complexes. *Molecular & Cellular Biology* **24**, 2364–2372 (2004).

16. Langst, G. & Becker, p. B. Nucleosome remodeling: one mechanism, many phenomena? *Biochimica et Biophysica Acta* **1677**, 58–63 (2004).

17. Mohrmann, L. & Verrijzer, C. P. Composition and functional specificity of SWI2/SNF2 class chromatin remodeling complexes. *Biochimica et Biophysica Acta* **1681**, 59–73 (2005).

18. Barak, O., Lazzaro, M. A., Lane, W. S., Speicher, D. W., Picketts, D. J. & Shiekhattar, R. Isolation of human NURF: a regulator of Engrailed gene expression. *EMBO Journal* **22**, 6089–6100 (2003).

19. Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S. & Reinberg, D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex con-

taining histone deacetylases and nucleosome remodelling activities. *Cell* **95**, 279–289 (1998).

20. Feng, Q. & Zhang, Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes & development* **15**, 827–832 (2001).

21. Razin, A. & Riggs, A. D. DNA methylation and gene function. *Science* **210**, 604–610 (1980).

22. Kass, S. U., Landsberger, N., and Wolffe, A. P. DNA methylation directs a time-dependent repression of transcription initiation. *Curr Biol* **7**, 157–165 (1997).

23. Chen, Z. J. and Pikaard, C. S. Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica. *Proc Natl Acad Sci U S A* **94**, 3442–3447 (1997).

24. Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. *Journal of Molecular Biology* **203**, 971–983 (1988).

25. Leonhardt, H., Page, A. W., Weier, H. U., and Bestor, T. H. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71**, 865–873 (1992).

26. Pradhan, S., Talbot, D., Sha, M., Benner, J., Hornstra, L., Li, E., Jaenisch, R., and Roberts, R. J. Baculovirus-mediated expression and characterization of the full-length murine DNA methyltransferase. *Nucleic Acids Research* **25**, 4666–4673 (1997).

27. Pradhan, S., Bacolla, A., Wells, R. D., and Roberts, R. J. Recombinant human DNA (cytosine-5) methyltransferase I. Expression, purification, and comparison of de novo and maintenance methylation. *Journal of Biological Chemistry* **274**, 33002–33010 (1999).

28. Okano, M., Xie, S., and Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics* **19**, 219–220 (1998).

29. Vertino, P. M., Yen, R. W., Gao, J., and Baylin, S. B. De novo methylation of CpG island sequences in human fibroblasts over expressing DNA (cytosine-5-)-methyltransferase. *Mol Cell Biol* **16**, 4555–4565 (1996).

30. Rhee, I., Jair, K.-W., Yen, R.-W. C., Lengauer, C., Herman, J. G., Kinzler, K. W., Vogelstein, B., Baylin, S. B., and Schuebel, K. E. CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature* **404**, 1003–1007 (2000).

31. Fatemi, M., Hermann, A., Gowher, H. & Jeltsch, A. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *European Journal of Biochemistry* **269**, 4981–4984 (2002).

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

33. Okano, M., Bell, D. W., Haber, D. A., and Li, W. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).

34. Hermann, A., Schmitt, S., and Jeltsch, A. The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. *Journal of Biological Chemistry* **278**, 31717–31721 (2003).

35. Tang, L.-Y., Reddy, M. N., Rasheva, V., Lee, T.-L., Lin, M.-J., Hung, M.-S., and Shen, C.-K. J. The eukaryotic DNMT2 genes encode a new class of cytosine-5 DNA methyltransferases. *Journal of Biological Chemistry* **278**, 33613–33616 (2003).

36. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B., and Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**, 2536–2539 (2001).

37. Aapola, U., Shibuya, K., Scott, H. S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S., Krohn, K., Antonarakis, S. E., Shimizu, N., Kudoh, J., and Peterson, P. Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics* **65**, 293–298 (2000).

38. Margot, J. B., Ehrenhofer-Murray, A. E., and Leonhardt, H. Interactions within the mammalian DNA methyltransferase family. *BMC Molecular Biology* **4**, 7 (2003).

39. Hata, K., Okano, M., Lei, H., and Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983–1993 (2002).

40. Deplus, R., Brenner, C., Burgers, W. A., Putmans, P., Kouzarides, T., de Launoit, Y., and Fuks, F. Dnmt3L is a transcriptional repressor that recruits histone deacetylase. *Nucleic Acids Research* **30**, 3831–3838 (2002).

41. Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and Cellular Biology* **18**, 6538–6547 (1998).

42. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

43. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497–3500 (2003).

44. Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185–188 (1999).

45. Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905–914 (1992).

46. Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L., and Bird, A. P. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499–507 (1989).

47. Nan, X., Campoy, F. J., and Bird, A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88**, 471–481 (1997).

48. Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).

49. Klose, R. J. and Bird, A. P. MeCP2 behaves as an elongated monomer that does not stably associate with the Sin3a chromatin remodeling complex. *J Biol Chem* **279**, 46490–46496 (2004).

50. Klose, R. J., Sarraf, S. A., Schmiedeberg, L., McDermott, S. M., Stancheva, I., and Bird, A. P. DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell* **19**, 667–678 (2005).

51. Cross, S. H., Meehan, R. R., Nan, X., and Bird, A. A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nat Genet* **16**, 256–259 (1997).

52. Ng, H. H., Zhang, Y., Hendrich, B., Johnson, C. A., Turner, B. M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet* **23**, 58–61 (1999).

53. Sarraf, S. A. and Stancheva, I. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Mol Cell* **15**, 595–605 (2004).

54. Bhattacharya, S., Ramchanctani, S., Cervoni, N., and Szyf, M. A mammalian protein with specific demethylase activity for mCpG. *Nature* **397**, 579–583 (1999).

55. Boeke, J., Ammerpohl, O., Kegel, S., Moehren, U., and Renkawitz, R. The minimal repression domain of MBD2b overlaps with the methyl-CpG-binding domain and binds directly to Sin3A. *J Biol Chem* **275**, 34963–34967 (2000).

56. Petronzelli, F., Riccio, A., Markham, G. D., Seeholzer, S. H., Stoerker, J., Genuardi, M., Yeung, A. T., Matsumoto, Y., and Bellacosa, A. Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. *J Biol Chem* **275**, 32422–32429 (2000).

57. Wu, P., Qiu, C., Sohail, A., Zhang, X., Bhagw at, A. S., and Cheng, X. 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).

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

59. Petronzelli, F., Riccio, A., Markham, G. D., Seeholzer, S. H., Genuardi, M., Karbowski, M., Yeung, A. T., Matsumoto, Y., and Bellacosa, A. Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. *J Cell Physiol* **185**, 473–480 (2000).

60. Roloff, T. C., Ropers, H. H. & Nuber, U. A. Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics* **4**, 1 (2003).

61. Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* **15**, 1613–1618 (2001).

62. Zemach, A. and Grafi, G. Characterization of Arabidopsis thaliana methyl-CpG-binding domain (MBD) proteins. *Plant J* **34**(5), 565–572 (2003).

63. Scebba, F., Bernacchia, G., De Bastiani, M., Evangelista, M., Cantoni, R. M., Cella, R., Locci, M. T., and Pitto, L. Arabidopsis MBD proteins show different binding specificities and nuclear localization. *Plant Mol Biol* **53**, 715–731 (2003).

64. Zemach, A., Li, Y., Wayburn, B., Ben-Meir, H., Kiss, V., Avivi, Y., Kalchenko, V., Jacobsen, S. E., and Grafi, G. DDM1 binds Arabidopsis methyl-CpG binding domain proteins and affects their subnuclear localization. *Plant Cell* **17**(5), 1549–1558 (2005).

65. Gutierrez, A. and Sommer, R. J. Evolution of dnmt-2 and mbd-2-like genes in the free-living nematodes Pristionchus pacificus, Caenorhabditis elegans and Caenorhabditis briggsae. *Nucleic Acids Res* **32**, 6388–6396 (2004).

66. Lyko, F., Ramsahoye, B. H., and Jaenisch, R. DNA methylation in Drosophila melanogaster. *Nature* **408**, 538–540 (2000).

67. Tweedie, S., Ng, H. H., Barlow, A. L., Turner, B. M., Hendrich, B., and Bird, A. Vestiges of a DNA methylation system in Drosophila melanogaster? *Nat Genet* **23**, 389–390 (1999).

68. Ballestar, E., Pile, L. A., Wassarman, D. A., Wolffe, A. P., and Wade, P. A. A Drosophila MBD family member is a transcriptional corepressor associated with specific genes. *Eur J Biochem* **268**, 5397–5406 (2001).

69. Roder, K., Hung, M. S., Lee, T. L., Lin, T. Y., Xiao, H., Isobe, K. I., Juang, J. L., and Shen, C. J. Transcriptional repression by Drosophila methyl-CpG-binding proteins. *Mol Cell Biol* **20**, 7401–7409 (2000).

70. Marhold, J., Zbylut, M., Lankenau, D.-H., Li, M., Gerlich, D., Ballestar, E., Mechler, B. M., and Lyko, F. Stage-specific chromosomal association of Drosophila dMBD2/3 during genome activation. *Chromosoma* **111**, 13–21 (2002).

71. Tatematsu, K. I., Yamazaki, T., and Ishikawa, F. MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. *Genes Cells* **5**, 677–688 (2000).

72. Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**, 187–191 (1998).

73. Zhang, Y., Ng, H.-H., Erdjument-Bromage, H., Tempst, P., Bird, A. & Reinberg, D. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & development* **13**, 1924–1935 (1999).

74. Brackertz, M., Boeke, J., Zhang, R., and Renkawitz, R. Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *J Biol Chem* **277**(43), 40958–40966 (2002).

75. Feng, Q., Cao, R., Xia, L., Erdjument-Bromage, H., Tempst, P. & Zhang, Y. Identification and functional characterization of the p66/p68 components of the MeCP1 complex. *Molecular and Celluar Biology* **22**, 536–546 (2002).

76. Branden, C. & Tooze, J. *Introduction to Protein Structure*, chapter 10, 176. Garland Publishing, inc., second edition (1999).

77. Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* **15**, 1613–1618 (2001).

78. Sekimata, M., Takahashi, A., Murakami-Sekimata, A., and Homma, Y. Involvement of a novel zinc finger protein, MIZF, in transcriptional repression by interacting with a methyl-CpG-binding protein, MBD2. *J Biol Chem* **276**, 42632–42638 (2001).

79. Sekimata, M. and Homma, Y. Regulation of Rb gene expression by an MBD2-interacting zinc finger protein MIZF during myogenic differentiation. *Biochem Biophys Res Commun* **325**, 653–659 (2004).

80. Sekimata, M. and Homma, Y. Sequence-specific transcriptional repression by an MBD2-interacting zinc finger protein MIZF. *Nucleic Acids Res* **32**, 590–597 (2004).

81. Magdinier, F. and Wolffe, A. P. Selective association of the methyl-CpG binding protein MBD2 with the silent p14/p16 locus in human neoplasia. *Proc Natl Acad Sci U S A* **98**, 4990–4995 (2001).

82. Bakker, J., Lin, X., and Nelson, W. G. Methyl-CpG binding domain protein 2 represses transcription from hype rmethylated pi-class glutathione S-transferase gene promoters in hepatocellular carcinoma cells. *J Biol Chem* **277** (2002).

83. Lembo, F., Pero, R., Angrisano, T., Vitiello, C., Iuliano, R., Bruni, C. B., and Chiariotti, L. MBDin, a novel MBD2-interacting protein, relieves MBD2 repression potential and reactivates transcription from methylated promoters. *Mol Cell Biol* **5**, 1656–1665 (2003).

84. Nitta, M., Saijo, M., Kodo, N., Matsuda, T., Nakatsu, Y., Tamai, H., and Tanaka, K. A novel cytoplasmic GTPase XAB1 interacts with DNA repair protein XPA. *Nucleic Acids Res* **28**, 4212–4218 (2000).

85. Bellacosa, A., Cicchillitti, L., Schepis, F., Riccio, A., Yeung, A. T., Matsumoto, Y., Golemis, E. A., Genuardi, M., and Neri, G. MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. *Proc Natl Acad Sci U S A* **96**, 3969–3974 (1999).

86. Hutchins, A. S., Mullen, A. C., Lee, H. W., Sykes, K. J., High, F. A., Hendrich, B. D., Bird, A. P., and Reiner, S. L. Gene silencing quantitatively controls the function of a developmental trans-activator. *Mol Cell* **10**, 81–91 (2002).

87. Li, H., Yamagata, T., Mori, M., Yasuhara, A., and Momoi, M. Y. Mutation analysis of methyl-CpG binding protein family genes in autistic patients. *Brain Dev* **27**, 321–325 (2005).

88. Lengauer, C. Cancer. An unstable liaison. *Science* **300**, 442–443 (2003).

89. Bird, A. DNA methylation patterns and epigenetic memory. *Genes & development* **16**, 6–21 (2002).

90. Lin, X. and Nelson, W. G. 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).

91. Sansom, O. J., Berger, J., Bishop, S. M., Hendrich, B., Bird, A. & Clarke, A. R. Deficiency of Mbd2 suppresses intestinal tumorigenesis. *Nature Genetics* **34**, 145–147 (2003).

92. Sansom, O. J., Bishop, S. M., Bird, A. & Clarke, A. R. MBD2 deficiency does not accelerate p53 mediated lymphomagenesis. *Oncogene* **24**, 2430–2432 (2005).

93. Billard, L.-M., Magdinier, F., Lenoir, G. M., Frappart, L., and Dante, R. MeCP2 and MBD2 expression during normal and pathological growth of the human mammary gland. *Oncogene* **21**, 2704–2712 (2002).

94. Kanai, Y., Ushijima, S., Nakanishi, Y., and Hirohashi, S. Reduced mRNA expression of the DNA demethylase, MBD2, in human colorectal and stomach cancers. *Biochem Biophys Res Commun* **264**, 962–966 (1999).

95. Zhu, Y., Spitz, M. R., Zhang, H., Grossman, H. B., Frazier, M. L., and Wu, X. Methyl-CpG-binding domain 2: a protective role in bladder carcinoma. *Cancer* **100**, 1853–1858 (2004).

96. Valinluck, V., Tsai, H.-H., Rogstad, D. K., Burdzy, A., Bird, A., and Sowers, L. C. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* **32**, 4100–4108 (2004).

97. Valinluck, V., Liu, P., Kang, J. I. J., Burdzy, A., and Sowers, L. C. 5-halogenated pyrimidine lesions within a CpG sequence context mimic 5-methylcytosine by enhancing the binding of the methyl-CpG-binding domain of methyl-CpG-binding protein 2 (MeCP2). *Nucleic Acids Res* **33**, 3057–3064 (2005).

98. Wakefield, R. I., Smith, B. O., Nan, X., Free, A., Soteriou, A., Uhrin, D., Bird, A. P., and Barlow, P. N. The solution structure of the domain from MeCP2 that binds to methylated DNA. *J Mol Biol* **291**, 1055–1065 (1999).

99. Ohki, I., Shimotake, N., Fujita, N., Nakao, M., and Shirakawa, M. Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. *EMBO J* **18**, 6653–6661 (1999).

100. Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105**, 487–497 (2001).

101. http://www.rcsb.org/pdb/.

102. Heitmann, B., Maurer, T., Weitzel, J. M., Stratling, W. H., Kalbitzer, H. R., and Brunner, E. Solution structure of the matrix attachment region-binding domain of chicken MeCP2. *Eur J Biochem* **270**, 3263–3270 (2003).

103. Hu, C. H., McStay, B., Jeong, S. W., and Reeder, R. H. xUBF, an RNA polymerase I transcription factor, binds crossover DNA with low sequence specificity. *Mol Cell Biol* **14**, 2871–2882 (1994).

104. http://www.expasy.org.

105. Sambrook, C. & Russell, D. W. *Molecular Cloning*, volume 1, chapter 1, 116. Cold Spring Harbor Laboratory Press, third edition (2001).

106. Sambrook, C. & Russell, D. W. *Molecular Cloning*, volume 3, chapter 15, 20. Cold Spring Harbor Laboratory Press, third edition (2001).

107. Sambrook, C. & Russell, D. W. *Molecular Cloning*, volume 3, chapter 15, 14. Cold Spring Harbor Laboratory Press, third edition (2001).

108. Stryer, L. *Biochemistry*, chapter 3, 46–48. W. H. Freeman & co., fourth edition (1995).

109. Sambrook, C. & Russell, D. W. *Molecular Cloning*, volume 3, appendix 8, 40–51. Cold Spring Harbor Laboratory Press, third edition (2001).

110. Sambrook, C. & Russell, D. W. *Molecular Cloning*, volume 3, chapter 15, 44–48. Cold Spring Harbor Laboratory Press, third edition (2001).

111. Stryer, L. *Biochemistry*, chapter 3, 49–50. W. H. Freeman & co., fourth edition (1995).

112. Aggarwal, A. Crystallization of DNA binding proteins with oligodeoxynucleotides. *Meth. Enz.* **1**, 83–90 (1990).

113. Cole, R. B. *Electrospray Ionization Mass Spectrometry.* Wiley-Interscience, (1997).

114. Stryer, L. *Biochemistry*, chapter 3, 51–52. W. H. Freeman & co., fourth edition (1995).

115. Free, A., Wakefield, R. I., Smith, B. O., Dryden, D. T., Barlow, P. N., and Bird, A. P. DNA recognition by the methyl-CpG binding domain of MeCP2. *J Biol Chem* **276**, 3353–3360 (2001).

116. Nagata, K. and Handa, H. *Real-Time Analysis of Biomolecular Interactions*, chapter 1, 13–22. Springer (2000).

117. Nagata, K. and Handa, H. *Real-Time Analysis of Biomolecular Interactions*, chapter 2, 23–30. Springer (2000).

118. BIACORE. *BIAevaluation version 3.0*, chapter 4, 23–24. BIACORE (1997).

119. Fraga, M. F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P. A., and Esteller, M. The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* **31**, 1765–1774 (2003).

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

121. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M., and Bax, A. Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol* **4**, 732–738 (1997).

122. Piotto, M., Saudek, V., and Sklenar, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR.* **2**, 661–666 (1992).

123. Sklenar, V., Piotto, M., Leppik, R., and Saudek, V. Gradient-tailored water suppression for ^{1}H - ^{15}N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson. A* **102**, 241–245 (1993).

124. Vuister, G. W. and Bax, A. Resolution enhancement and spectral editing of uniformly ^{13}C enriched proteins by homonuclear broadband ^{13}C decoupling. *J. Magn. Reson. A* **98**, 428–435 (1992).

125. Cavanagh, W. J., Fairbrother, W. J., Palmer, A.G. III and Skelton, N. J. *Protein NMR Spectroscopy: Principles and Practice.* Academic Press, (1995).

126. Jahnke, W., Baur, M., Gemmecker, G., and Kessler, H. Improved accuracy of NMR structures by a modified NOESY-HSQC experiment. *J. Magn. Reson. B* **106**, 86–88 (1995).

127. Kay, L. E., Xu, G.-Y., Singer, A. U., Muhandiram, D. R., and Forman-Kay, J. D. A gradient-enhanced HCCH-TOCSY experiment for recording

sidechain ^{1}H and ^{13}C correlations in H_2O samples of proteins. *J. Magn. Reson. B* **101**, 333–337 (1993).

128. Pascal, S. M., Muhandiram, D. R., Yamazaki, T., Forman-Kay, J. D., and Kay, L. E. Simultaneous acquisition of N-15-edited and C-13-edited NOE spectra of proteins dissolved in H_2O. *J. Magn. Reson. B* **103**, 197–201 (1994).

129. Muhandiram, D. R. and Kay, L. E. Gradient enhanced triple resonance three-dimensional NMR experiments with improved sensitivity. *J. Magn. Reson. B* **103**, 203–216 (1994).

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

131. Wittekind, M. and Mueller, L. HNCACB, a high sensitivity 3D NMR experiment to correlate amide proton and nitrogen resonances with the α-carbon and β-carbon resonances in proteins. *J. Magn. Reson. B* **101**, 201–205 (1993).

132. Lyons, B. A. and Montelione, G. T. An HCCNH triple-resonance experiment using carbon-13 isotropic mixing for correlating backbone amide and side-chain aliphatic resonances in isotopically enriched proteins. *J. Magn. Reson. B* **101**, 206–209 (1993).

133. Carlomagno, T., Maurer, M., Sattler, M., Schwendinger, M. G., Glaser, S. J., and Griesinger, C. PLUSH TACSY: Homonuclear planar TACSY with two-band selective shaped pulses applied to Cα, C' transfer and Cβ, C-aromatic correlations. *J. Biomol. NMR.* **8**, 161–170 (1996).

134. http://www.bio.cam.ac.uk/azara/.

135. Kraulis, P. J. ANSIG: A program for the assignment of protein 1H 2D NMR spectra by interactive graphics. *J Magn Reson* **84**, 627–633 (1989).

136. Smith, B. O. personal communication.

137. Wishart, D. S. and Sykes, B. D. The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J Biomol NMR* **4**, 171–180 (1994).

138. Zwahlen, C., Legault, P., Vincent, S. J. F., Greenblatt, J., Konrat, R., and Kay, L. E. Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: Application to a bacteriophage N-peptide/boxB RNA complex. *J. Am. Chem. Soc.* **119**, 6711–6721 (1997).

139. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. The CCPN data model for NMR spectroscopy: development of a software p ipeline. *Proteins* **59**, 687–696 (2005). http://www.ccpn.ac.uk/analysis/.

140. Metropolis, N., Rosenbluth, A., Rosenbluth, M., and Teller, A. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).

141. Cerny, V. Thermodynamical Approach to the traveling salesman problem: An efficient simulation algorithm. *J. Opt. Theory Appl.* **45**, 41–51 (1985).

142. Haile, J. M. *Molecular Dynamics Simulations.* Wiley, John & Sons, inc., (1992).

143. Linge, J. P. and Nilges, M. Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. *J Biomol NMR* **13**, 51–59 (1999).

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

145. Laskowski, R. A., MacArthur, M. W., and Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).

146. Koradi, R., Billeter, M., and Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**, 51–55 (1996).

147. Kraulis, P, J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**, 946–950 (1991).

148. Sali, A. and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).

149. http://salilab.org/modeller/.

150. Rauch, C., Trieb, M., Wibowo, F. R., Wellenzohn, B., Mayer, E., and Liedl, K. R. Towards an understanding of DNA recognition by the Methyl-CpG Binding Domain 1. *J. Biomol. Struc & Dyn.* **22**, 695–706 (2005).

151. Neuhaus, D. & Williamson, M. P. *The Nuclear Overhauser Effect in Structural and Conformtional Analysis*, chapter 9, 347–359. Wiley, John & Sons, inc., second edition (2000).

152. Ho, K.-L. and Walkinshaw, M. W. personal communication.

153. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520 (2000).