Solution Structure of hMBD1 CXXC1

A thesis submitted to the

COLLEGE OF MEDICAL, VETERINARY & LIFE SCIENCES

For the Degree of DOCTOR OF PHILOSOPHY

By Ross Thomson

College of Medical, Veterinary & Life Sciences
Institute of Molecular Cell and Systems Biology
University of Glasgow

June 2011

Abstract

Methylation of CpG dinucleotides is the major epigenetic modification of mammalian DNA which results in the remodelling of transcriptionally active euchromatin to transcriptionally inactive heterochromatin. Recognition of methylated CpG by methylated DNA binding proteins, the MBD family, the Kaiso zinc finger family and the SRA domain proteins results in deacetylation and methylation of histone side chains through the recruitment of HDAC and HMT enzymes. Methylation of DNA is a heritable process ensuring Methylation dependant transcriptional repression is passed from mother to daughter cell during replication. Some of the proteins involved in this chromatin remodelling, MBD1, DNMT1, MLL, and CFP1 contain CXXC domains. hMBD1 contains 2 or 3 CXXC domains depending on the splice variant, with only the third CXXC domain shown to bind CpG dinucleotides.

This thesis describes the work done to elucidate the structure of hMBD1 CXXC1 and to investigate hMBD1 CXXC12 di-domain by NMR spectroscopy and biochemical characterisation. The hMBD1 CXXC1 & CXXC12 domains were successfully over expressed in *E. coli* and purified. Unlabelled and uniformly ^{15}N labelled proteins were produced for nuclear magnetic resonance (NMR) studies. Assignment of NMR spectra was carried out and constraints generated enabling structure determination of hMBD1 CXXC1 and to investigate the relationship between CXXC1 and CXXC2 of hMBD1.

The solution structure of hMBD1 CXXC1 determined here was compared to the previously determined solution structure of hMLL CXXC in order to investigate their differences in DNA binding. NOE data from hMBD1 CXXC1 and CXXC12 are compared in order to investigate the domain structure of CXXC12. DNA titration of hMBD1 CXXC1 showed no significant interaction with a single CpG oligo while the loop region of hMBD1 CXXC1 differs significant in both structure and surface charge suggesting the loop region to be important for DNA binding. The recorded NOE data of hMBD CXXC12 suggests the two CXXC domains form a globular rather than a linear structure

Table of Contents

List of Tables

List of Figures

Acknowledgments

Firstly I would like to thank my supervisor, Dr Brian smith, for his continued support, advice and extreme patience over the course of my studies.

I'd also like to thank Dr Yinan Fu, Steve Vance and other members of the lab for their advice, chats, support and grilling during lab meetings.

A big thank you for Dr John Parkinson at the WestCHEM NMR Facility for the use of their 600 MHz magnet also Dr Dusan Uhrin & Prof. Paul Barlow at the Edinburgh biomolecular NMR unit for use of their 800 MHz magnet when our 600 MHz magnet quenched (twice).

Finally a big thank you to my wife, children and parents for putting up with me during my studies.

"You know, what these people do is really very clever. They put little spies into the molecules and send radio signals to them, and they have to radio back what they are seeing." – Felix Bloch

I hereby declare that this thesis, and the research it describes, is entirely my own work except where explicitly stated. It has not been presented in whole, or part, for any other degree or award.

Ross Thomson
March 2011

Definitions/abbreviations

2D	two-dimensional
3D	three-dimensional
ARIA	ambiguous restraints for iterative assignment
bp	base pair
Cα	alpha proton nucleus
Cβ	beta proton nucleus
Cδ	delta proton nucleus
Cγ	gamma proton nucleus
Hζ	epsilon proton nucleu
dH_2O	distilled and deionised water
DNA	deoxyribonucleic acid
DNMT1	DNA methyl transferase 1
DTT	dithiothreitol
EDTA	ethylendiaminetetraacetic acid
FID	free induction decay
H_N	amide proton nucleus
HSQC	heteronuclear single quantum correlation
IPTG	isopropyl-b-D-galactopyranoside
MBD	methylated-DNA binding domain
NH	amide group
NMR	nuclear magnetic resonance
NOESY	nuclear Overhouser effect spectroscopy
PDB	protein data bank
RMSD	root mean square deviation
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
TOCSY	total correlation spectroscopy
UV	ultraviolet radiation

CHAPTER ONE

INTRODUCTION PART 1: EPIGENETICS

1.1 Epigenetics

Epigenetics is the study of heritable changes in gene function that occur without a change in the sequence of nuclear DNA. This form of inheritance allows the transmission of information from mother to daughter cell without the information being encoded in the nucleotide sequence of the gene, for example when a liver cell divides, the daughter cells do not start to express proteins specific to muscle cells. Methylation of a CpG dinucleotide at cytosine C5 (mCpG) is a major epigenetic gene silencing modification in vertebrate genomes. This modification recruits proteins which specifically recognise this motif. These methylated DNA binding proteins then recruit enzymes which chemically and physically alter chromatin, which induces transcriptional repression. Although most CpG motifs are methylated it should be noted that short (500-2000bp) CG-rich regions, known as CpG islands, found within 60% of human gene promoters remain non-methylated (Bird, A. 2002). While this is true for normal cells, *de novo* methylation of CpG islands occurs in various cancers, inducing silencing of tumour suppressor genes e.g. *CDH1* in breast, bladder and prostate cancer (Graff *et al*. 1995), *CDKN2A* in many epithelial cancers (Merlo *et al*. 1995), and the Rb gene in retinoblastomas (Sakai *et al*. 1991).

Control of gene expression/repression is important in cell development and differentiation to ensure that only cell-specific genes are transcribed. Some genes are constitutively transcribed in almost all cells, e.g. Glyceraldehyde 3 phosphate dehydrogenase (*GAPDH*); some genes are only transcribed in certain cell types, e.g. carnitine palmitoyltransferase I C in rat brain and testes (Price *et al*, 2002); others are only transcribed after a signaling cascade has been initiated, e.g. induction of insulin gene transcription by glucose in beta cells of the Islets of Langerhans (Leibiger *et al*. 1998). Eukaryotic DNA is packaged into a nucleoprotein complex called chromatin that is organized into two structurally distinct domains, euchromatin and heterochromatin (Heitz, 1928). Euchromatin is condensed during cell division but more open and transcriptionally active during interphase while heterochromatin is tightly packed and transcriptionally inactive throughout the cell cycle. While there is only one class of euchromatin, heterochromatin has two variants, constitutive or facultative. Constitutive heterochromatin is fixed, irreversible and located at very specific spots in the genome that consist of DNA that contains many tandem (not inverted) repeats of a short repeating unit known as satellite DNA. Facultative heterochromatin can revert to a euchromatin state for example, when a woman transmits the X-chromosome to a son; it

reverts to euchromatin from heterochromatin. Replication of these two types of chromatin occurs at different time points, with heterochromatic DNA late and euchromatic DNA early within the cell cycle (Gilbert, 2002). DNA methylation is functionally connected to these two states through histone modification. In a simplified view, euchromatic DNA contains non-methylated CpGs and the histones are often acetylated whereas the CpGs in heterochromatin DNA are methylated and the histones are deacetylated with methylation occurring at histone 3 lysine 9 ($H3K9^m$) (Cameron *et al*. 1999). De-methylation of CpG motifs allows $H3K9^m$ to be re-acetylated (Bachman *et al*. 2003) which ultimately leads to a switch from hetero- to euchromatic DNA. Thus DNA methylation is a pivotal signal for the epigenetic control of gene expression in a reversible heritable manner.

1.2 Chromatin Structure

The Basic building blocks of Chromatin are five types of proteins called histones which contain a high proportion of charged amino acids. These histones fall into two categories, the nucleosomal histones and the H1 histones. The nucleosomal group is made up of four histones which are called H2A, H2B, H3, H4 and their variants. They are small proteins 102-135aa with H3 and H4 histones being among the most highly evolutionally conserved of all proteins. The H1 histones are larger, approx 220aa, and are less conserved than the nucleosomal group. Although histones are the building blocks of chromatin, there are also architectural proteins that are comprised of acidic or non-histone chromatin proteins.

The 11 nm diameter nucleosome (Fig.1.1), the fundamental building block of chromatin, is a complex comprised of an 146 bp (base pair) length of DNA wrapped around an octomer core of histone proteins consisting of two copies each of H2A, H2B, H3, and H4 (Richmond *et al*. 1984, Dorigo *et al*. 2004).

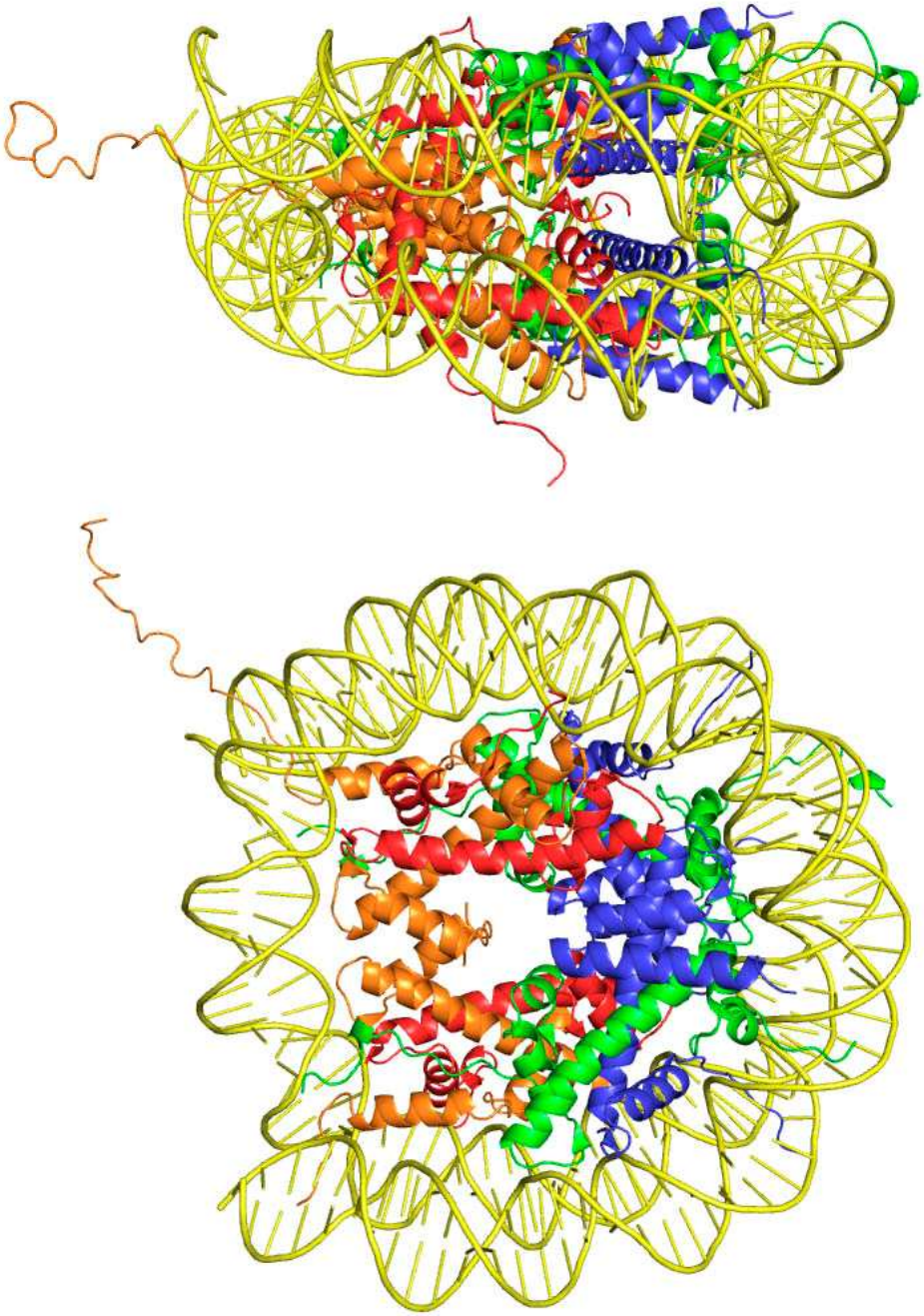


Figure 1.1 X-ray crystal structure of the nucleosome core particle solved by Luger *et al* (1997). Ribbon traces for the 146-bp DNA phosphodiester backbones (Yellow) and eight histone protein main chains (orange: H3; red: H4; green: H2A; blue: H2B). Image generated from PDB ref 1AO1 (Figure prepared using PyMol).

With a small length, 10-100bp, of DNA (linker) between each nucleosome, this type of chromatin thread can be visualized by electron microscopy which gives rise to the 'beads on a string' structure description. The H1 histone protein binds to a specific region of the nucleosome and mediates the packing of the 11nm nucleosome into a 30nm chromatin fibre. From this 30nm fibre the DNA can be condensed into a chromosome. The histones also possess N- and C- terminal tails, 14-38aa in length, that protrude from the nucleosome (Luger *et al*. 1997) and direct the formation of higher-order chromatin structures (Dorigo *et al*. 2003, Dorigo *et al*. 2004, Gordon *et al*. 2005).

As mentioned previously not all proteins associated with chromatin are histones. These non-histone architectural proteins are responsible for the chromatin regulatory mechanisms such as covalent histone modifications or ATP-dependent chromatin remodeling (Cosgrove *et al*. 2004, Smith & Peterson, 2004, Studisky *et al*. 2004). The packing of chromatin at its lower level is defined by DNA-histone and nucleosome-nucleosome interactions. Chromatin remodeling at specific loci is carried out through tagging of the histone tails (Fig 1.2a). Covalent modification of the histone tails in the form of acetylation and methylation are important for the remodeling of chromatin to activate and repress transcription respectively (Berger, 2001, Strahl and Allis, 2000). Other covalent modifications include phosphorylation, ubiquitylation, sumoylation (Gill, 2004) and ADP-ribosylation (Cohen-Armon *et al*. 2004) (fig 1.2b)

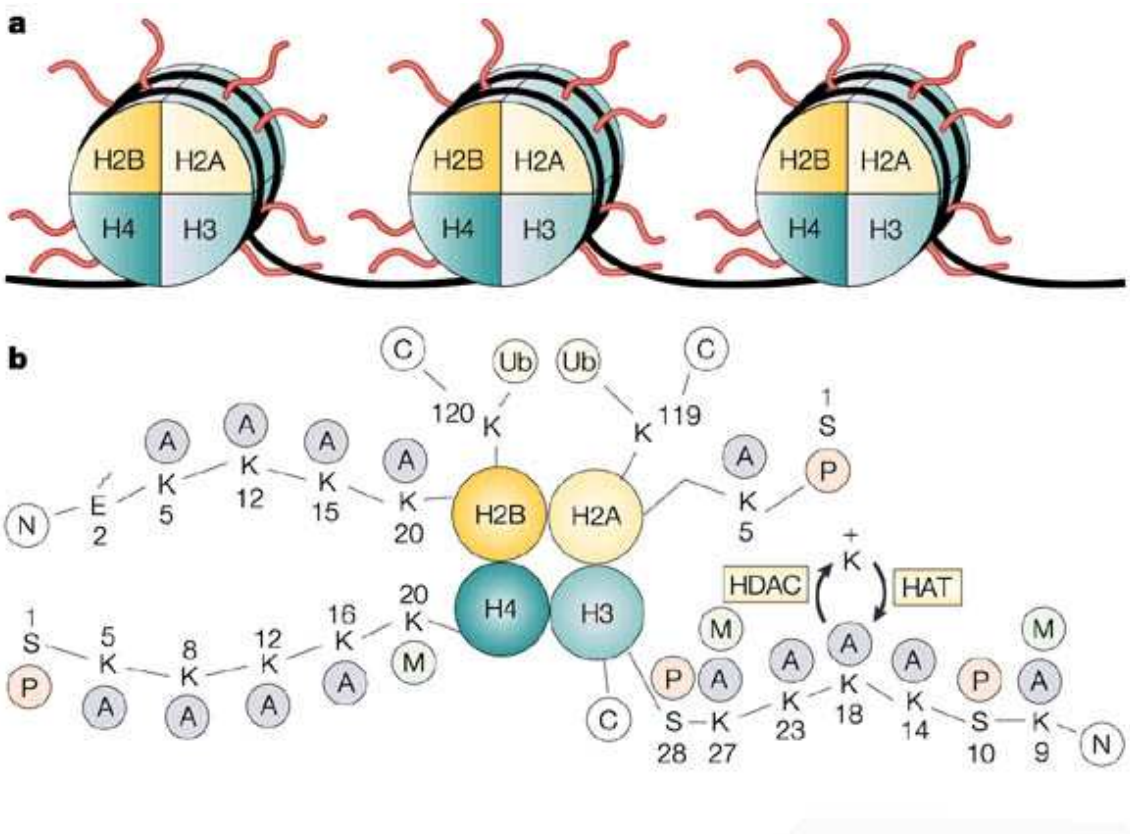


Figure 1.2 **a** Diagram of the 11 nm beads on a string chromatin conformation showing histone tails. **b** The various histone tail covalent modifications. A , acetyl ; P, phosphate; M, methyl; Ub, Ubiqutin; C, carboxy terminus; N, amino terminus. Adapted from Marks *et al* (2001)

Histone deacetylation (H3K9) by histone deacetylase enzymes (HDAC) and methylation of the same residue by histone methyl transferases (HMT) (Fuks *et al*. 2003) can be brought about by the binding of methylated DNA binding proteins to methyl-CpG motifs which recruit HDACs to chromatin (Nan X *et al*. 1998). The methylation of H3K9 allows the binding of heterochromatin protein 1 (HP1) via its chromodomain (Banister *et al*. 2001) which stabilizes the formation of heterochromatin and induces transcriptional repression. DNMT1 also binds the HDACs HDAC1 (Robertson *et* al. 2000) and HDAC2 (Rountree *et* al, 2000) via its N-terminal domain. DNMT1 binding to HDAC2 is associated with binding to another protein, DMAP1 (DNMT1 associated protein) which can mediate transcriptional repression (Rountree *et al*. 2000). It should also be noted that loss of DNMT1 from *DNMT1* knockout human cancer cells KO1 results in an increase of H3 acetylation and loss of trimethylation of lysine 9. Also HDACs and HP1 no longer interact with histone H3 and pericentrometric repetitive sequences (Espada *et al*. 2004). This suggests that DNMT1 contributes more to heterochromatin formation than just the methylation of CpG motifs.

1.3 DNA Methylation

Methylation is the major covalent modification of eukaryotic DNA and almost always found in a palindromic CpG sequence such that the 5 methyl cytosine is found symmetrically on both strands (Sinsheimer 1955). Methylated DNA was first proposed by Scarano (1971) and later by Holliday & Pugh (1975) and Riggs (1975) to be involved in cell differentiation and gene function. Later Yeivin and Razin (1993) showed that epigenetic control of gene expression in mammals involves DNA methylation associated with gene-specific methylation patterns. Methylation occurs at 70-80% of mammalian CpG sites (Ehrlich and Wang 1981) with the remaining unmethylated CpG sites being found in CpG rich islands. CpG islands have been shown to be present at promoter regions of all constitutively expressed and approximately 40% of tissue specific genes (Larsen *et al*, 1992. Zhu *et al*, 2008). However, not all CpG islands are non-methylated, some of which are involved in X-inactivation and genomic imprinting (Edwards & Ferguson-Smith, 2007., Reik, 2007)

Without the replacement of methylated cytosines, DNA replication would remove 50% of methylated CpG from DNA after two rounds of replication through passive demethylation. In order to protect the symmetrical methylation of CpG, the enzyme DNA methyl transferase (DNMT1) catalyses the covalent attachment of a methyl group to the naked cytosine in a hemi-methylated CpG (Bestor and Ingram 1983). DNMT1 is associated with proliferating cell nuclear antigen during DNA replication at the replication foci, with methylation occurring within one to two minutes of replication to restore a state of full methylation to the DNA (Chuang *et al*. 1997). It is through this process that methylated DNA induced gene repression is inherited from mother to daughter for many generations of actively dividing cells (Razin and Riggs 1980).

The reason that CpG Islands are not methylated at replication could be explained by their replication during G1/early S phase (Delgado *et al*. 1998) when DNMT1 levels are low (Szyf *et al*, 1991) and levels of the transcription factor p21 are high which could out-compete DNMT1 for binding to PCNA (Chuang *et al*. 1997). This could prevent DNMT1 being present at the replication foci of CpG islands thus preventing their methylation. This theory is strengthened by the fact that many CpG islands are also replication origins (Delgado *et al*. 1998) and, due to the proteolytic factors involved in DNA replication, could block DNMT1 access to the DNA. However, it does not explain the lack of methylated CpG islands that are located in the promoters of genes that lack

p21 recognition sites. Recently the CXXC Finger Protein 1 (CFP1), a component of the SET1 H3K4 methyltransferase complex (Skalnik *et al*, 2008), has been shown to bind CpG islands (Thompson *et al*, 2010). Methylation of H3K4 is a euchromatic mark associated with transcriptional activity (Bernstein *et al*, 2002., Santos-Rosa *et al*, 2002., Schneider *et al*, 2004). Therefore the recruitment of CFP1 to CpG islands and subsequent methylation of H3K4 may prevent methylation through steric interference.

It has been reported that methylated DNA does undergo active demethylation (Oswald *et al*, 2000) but this is not necessarily associated with replication (Kafri *et al*. 1992) while the paternal genome in fertilized mouse eggs is stripped of DNA methylation prior to zygotic fusion and subsequent replication (Mayer *et al*. 2000). This demethylation of the paternal genome is a feature of mouse, rat, human and pig zygotes but surprisingly not of rabbit and sheep zygotes (Beaujean *et al*. 2005, Beaujean *et al*. 2004a, Young & Beaujean, 2004). Interestingly the sheep paternal genome is significantly demethylated when injected into mouse oocytes (Beaujean *et al*. 2004b) suggesting it is not inherently resistant to demethylation. While DNMT1 is responsible for maintaining methylation patterns during DNA replication, *de novo* methylation at previously unmethylated sites during early embryogenesis is catalyzed by DNMT3a & DNMT3b (Okano *et al*. 1999). However, DNMT1 can be stimulated to methylate CpG dinucleotides by the addition of fully methylated DNA (Fatami *et al*, 2001, Laing *et al*, 2001). DNMT1 is essential for mouse embryogenesis as targeted knockouts of DNMT1 results in death around mid gestation (Li *et al*. 1992, Bird & Wolffe, 1999). This is also true for *Xenopus laevis* embryos which die before mid-blastula transition if DNMT1 is knocked out (Stancheva & Meehan, 2000). However, this is independent of DNMT1's methyltransferase activity (Dunican *et al*, 2008)

Aberrant DNA methylation is of particular importance in cancer (Jones & Baylin, 2002) as hypomethylation can result in the inappropriate expression of genes while hypermethylation of CpG islands can switch off genes (Jones & Baylin, 2002, Feinberg & Vogelstein, 1983). Hypermethylation of CpG islands found in tumor suppressor gene promoters is a prevalent mechanism for the transcriptional silencing of these genes (Robertson, 2001, Rountree *et al*. 2001. Esteller & Herman, 2002., Jones & Baylin, 2002.). Removal of methyltransferase activity from the human colorectal cancer cell line HCT116 by genetic (antisense/siRNA against DNMT1 RNA) or pharmacologic methods (5-aza-2'-deoxycytidine) causes extensive hypomethylation throughout the

genome, including hypermethylated regions (Robert et al. 2003, Ting et al. 2004). However, controversy surrounds the question of knocking out DNMT1 is sufficient for this observed hypomethylation event. Ting et al. (2004) suggest that DNMT1 is redundant due to a compensatory role for DNMT3b. In order to investigate this they produced a double knockout of DNMT1 & DNMT3b cells which showed greater than 95% reduction in genomic methylation and crucially these cells re-expressed the tumour suppressor gene Cdkn2a, providing a direct link between the methylation of its promoter and its transcription. These double knockout cells had a 90% reduction in growth compared to the wild type HCT116 cells. DNMT1 might also be targeted to methylate DNA at Histone H3-K9 methylated chromatin via interaction with the histone methyltransferase SUV39H1 (Fuks et al. 2003).

Females contain two copies of the X chromosome compared the male having just one. In order to compensate for this random inactivation of one copy occurs during early development (Lyon, 1961). This inactivation is associated with widespread DNA methylation (Heard et al. 1997), while transcription of genes on the inactive chromosome can be reactivated when cells are treated by inhibitors of DNA methylation (Boumil & Lee, 2001). Not all gene transcription on the inactive X chromosome is silenced as some CpG islands remain unmethylated (Heard et al. 1997).
DNA methylation plays a critical role in genomic imprinting, the phenomenon whereby a small subset of all the genes in the genome are expressed according to their parent of origin (Bartolomei & Tilghman, 1997). The control the reciprocal parental gene for almost all imprinted genes identified to date is associated with differentially methylated regions (Paulsen & Ferguson Smith, 2001).

Thus DNA methylation states play an important role in the transcriptional repression of genes, regulation of chromatin structure and the development of embryos while aberrant DNA methylation can cause disease, including cancer & developmental disorders, and death during gestation.

1.4 Methylated DNA binding proteins

MeCP2, MBD1, MBD2, MBD3 and MBD4 constitute a family of proteins which contain the methyl-CpG-binding domain (MBD). In the late 1980s a nuclear factor called MeCP1 (Methyl CpG binding Protein) was discovered. This factor could discriminate between DNA molecules that contained 12 or more methylated CpG pairs

and unmethylated DNA in band shift assays (Meehan *et al*. 1989). MeCP1 is a large multi-subunit protein complex and has been shown to be comprised of MBD3 (Zhang *et al*. 1999), the Mi2/NuRD complex and MBD2 (Ng H *et al*. 1999, Feng and Zhang 2001). More recently other methyl-CpG binding proteins have been discovered. The zinc finger proteins Kaiso (Daniel and Reynolds, 1999., Prokhortchouk *et al*, 2001), ZBTB4 and ZBTB38 (Filion *et al*, 2006) and UHRF1 & -2 that contain an SRA (SET and Ring finger-associated) domain (Unoki *et al*, 2004). The first well characterized member of the MBD family is MeCP2.

1.4.1 MeCP2

Identified in 1992, MeCP2 is a 53kDa protein capable of binding one methyl-CpG pair (Lewis *et al*. 1992) although it is suggested that, *in vitro*, a flanking run of 4 or more A/T base pairs (Klose *et* a*l*. 2005) enhances the affinity. This is backed up by analysis of the binding sites in two of its known targets, Bdnf and Dlx6, which show the methyl-CpG site flanked by an A/T greater than 4 motifs (Klose & Bird 2006). Levels of MeCP2 protein vary depending on cell type with rat brain containing the highest followed by kidney, spleen, liver with the lowest found in testes (Nan X *et al*. 1993). Although widespread, MeCP2 is not a major repressor of gene regulation (Tudor *et al*. 2002) but binds to loci not targeted by other MBDs (Klose *et al*. 2005). However, these reports used cultured neurons from embryonic or immature mice where MeCP2 levels are low. Skene *et al* (2010) determined the binding pattern in mature mouse brains which showed a different pattern, with MeCP2 tracking the methyl-CpG density of the genome. These results suggest, in mature mouse brain, that MeCP2 is a global regulator of chromatin structure.

The MBD domain, defined as the minimum required to bind methyl-CpG, comprises amino acids 78-162. This MBD domain has similar characteristics of the full length protein when binding methyl-CpG DNA in that it will bind a single methyl-CpG motif in duplex DNA and has negligible binding to hemimethylated DNA or 5-methylcytosine at non-CpG motifs (Nan X *et al*. 1993). Within mouse EB28/10 cells, MeCP2 shows punctate labeling, co-localising with heterochromatin when transiently over expressed. This localisation to the chromatin is dependant on methylation, as over expressed MeCP2 in MTase mutant (methylase negative) mouse ES cells produced diffuse nuclear staining with only 20% localizing to chromatin (Nan X *et al*. 1996). The small percentage of MeCP2 binding to chromatin is likely to be due to the lower levels of

methyl-CpG remaining (Nan X *et al*. 1996) and non specific DNA binding (Nan X *et al*. 1993). Once bound to methylated CpG, MeCP2 initiates transcriptional repression by recruiting the transcriptional co-repressor Sin3 and histone deacetylase (HDAC) complex (Nan X *et al*. 1998). Binding of the HDAC is via a transcriptional repression domain (Nan X *et al*. 1997) and is localized to amino acids 207–310 (Yu *et al*. 2000). Removal of the acetyl group from histone H3 lysine 9 (Nan X *et al*. 1998) facilitates the methylation of the same residue by a histone methyltransferase (HMT) (Fuks *et al*. 2003). Treatment with trichostatin A (an inhibitor of HDAC), however, does not fully remove transcriptional repression which suggests that MeCP2 can repress by another pathway (Yu *et al*. 2000) or interaction with other complexes that do not require HDAC to inhibit transcription.

1.4.2 MBD1

Cross *et al*. (1998) identified MBD1 (formerly PCM1) by searching the XREF database for sequences homologous to the MBD of MeCP2. MBD1 was shown to bind methyl-CpG via its MBD. MBD1 was originally thought to be part of the MeCP1 complex (Cross *et al*. 1998). However, Huck-Hui *et al*. (2000) showed that MeCP1 band shift assays were not affected when MBD1 was immunoprecipitated using a sheep antibody against MBD1 amino acids 351-556 from the extract. As well as the N-terminal MBD domain, MBD1 also contains, depending on alternative splicing, 2 or 3 CXXC domains (Fujita *et al*, 1999, Jørgensen *et al*, 2004). There are a total of 8 known isoforms of MBD1 with the number and type of CXXC domains shown in figure 1.3

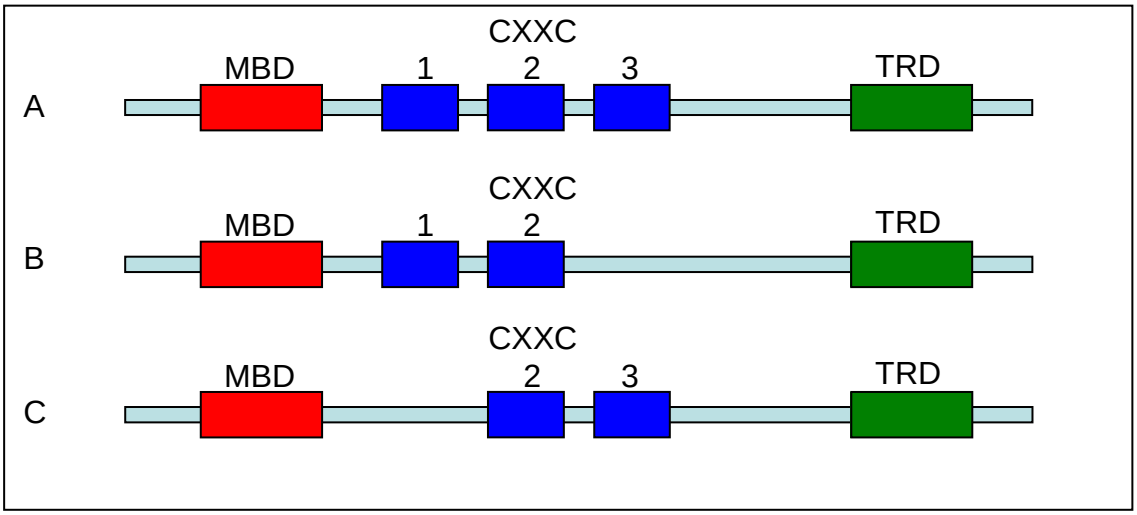


Figure 1.3 schematic representations of MBD1 splice variants with different combinations of CXXC domains. A: represents isoforms 1, 2, 6, 8; B: represents isoforms 3, 4 ,7; C: isoform 5 (formerly known as PCM1).

A subunit of CAF-1 (chromatin assembly factor 1), p150, and SUV39h1, a methyltransferase enzyme both interact with the MBD (Fujita *et al*. 2003b, Resse *et al*. 2003). It is thought that MBD1 is displaced from methyl-CpG during replication but that through its interaction with CAF-1 via PCNA, it remains at the replication fork during S-phase. Once replication has occurred, DNMT1 activity fully methylates hemimethylated CpG motifs allowing MBD1 to rebind to the methyl-CpG (Sarraf & Stancheva, 2004) (Fig. 1.4)

Like MeCP2, MBD1 contains a transcriptional repression domain which, when fused to the Gal4 DNA binding domain, can repress transcription from a Gal containing promoter. The transcriptional repression domain (TRD) is known to interact with a methyl-DNA glycosylase (Watanabe *et al*. 2003) and MCAF (MBD1-containing chromatin associated factor) (Fujita N *et al*. 2003a). Although MBD1 does not appear to interact with HDAC1 or HDAC2 (Ng H *et al*. 2000) there is variable repression sensitivity when cells are treated with the HDAC inhibitor trichostatin A (Ng H *et al*. 2000, Fujita N *et al*. 2003a) suggesting that MBD1 induced repression does not always require HDACs

Although MBD1 knock out mice have been produced, the $MBD1^{-/-}$ mice show no detectable development defects and appear physically healthy throughout their life span. However, adult knockout mice show decreased neurogenesis and impaired spatial

learning, while neural stem cells exhibited reduced neural differentiation and increased genomic instability (Zhao *et al*. 2003).

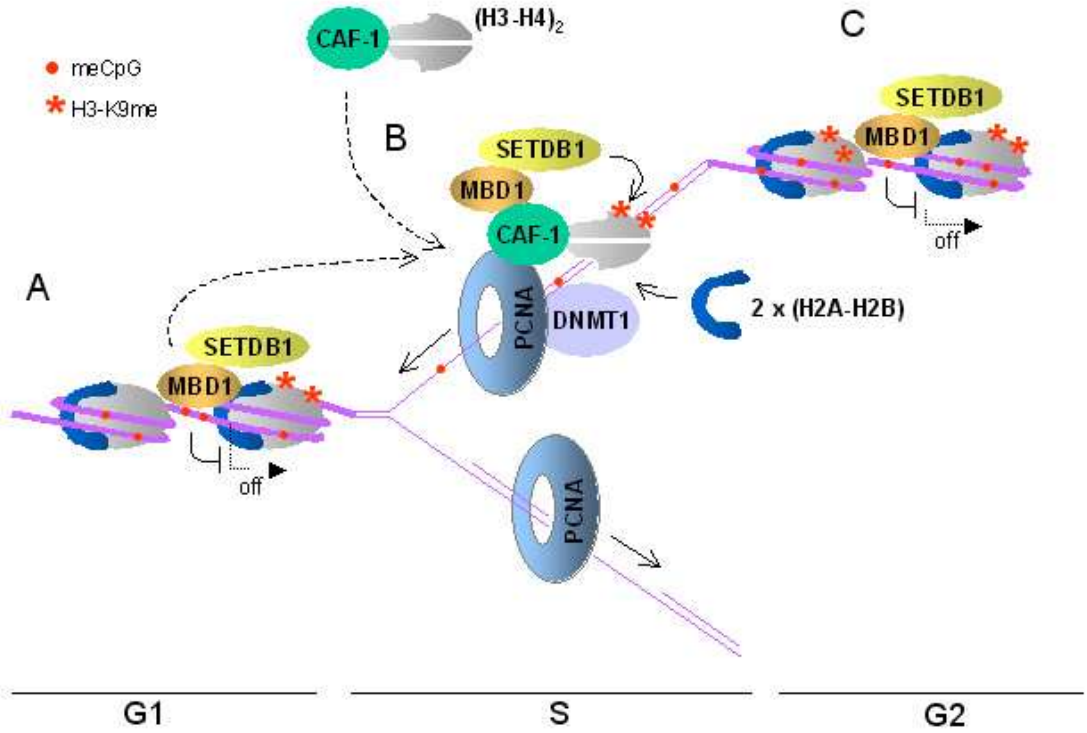


Figure 1.4. Model summarizing the function of MBD1 complexes in transcriptional repression and DNA replication-coupled maintenance of histone H3-K9 methylation. MBD1 and H3-K9 methylase SETDB1 form a stable MBD1/SETDB1 complex detectable throughout the cell cycle in G1, S, and G2 (Reproduced from Sarraf & Stancheva, 2004)

1.4.3 MBD2

MBD2 was identified by Hendrich & Bird (1998) by searching for putative MBD domains in the EST database. It has two possible isoforms depending on translation from the first (MBD2a, 43.5 kDa) or second (MBD2b, 29.kDa) ATG codon. The translation from the second ATG codon produces a protein with the MBD at the extreme N-terminus (Hendrich & Bird, 1998) lacking the glycine rich region found in MBD2a (figure 1.5).

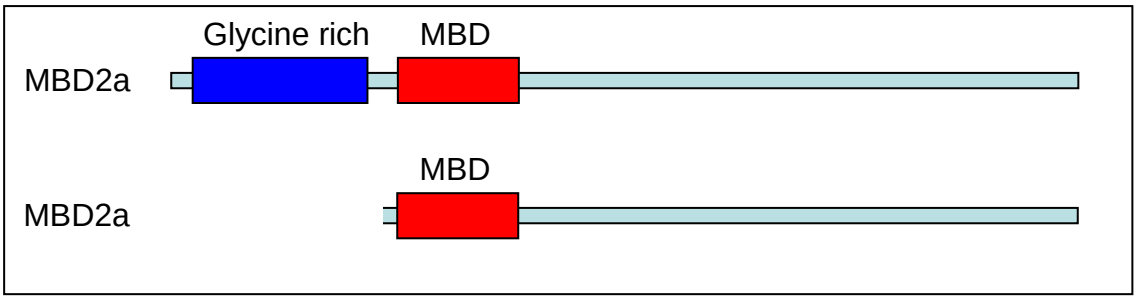


Figure 1.5. Schematic representation of the two isoforms of MBD2.

It was proposed that *MBD2b* has DNA demethylase activity (Bhattacharya *et al*. 1999) but subsequent attempts by Ng *et al*. (1999) and Wade *et al*. (1999) were unable to reproduce these results. Like MeCP2, transcriptional repression via MBD2 occurs through histone modification and it was proposed that MDB2 is part of the MeCP1 complex (Ng *et al*. 1999) which comprises MBD3 (Zhang Y *et al*. 1999) and the Mi2/NuRD complex (Feng Q and Zhang Y 2001). Recently however, Guezennec *et al*. (2006) have suggested that MBD2/Nurd and MBD3/Nurd are two distinct complexes. Evidence to further back up Guezennec's results is shown in MBD2 and MBD3 knockout mice. While *MBD3*$^{-/-}$ mice die during early embryogenesis, *MBD2*$^{-/-}$ mice are viable and fertile suggesting distinct but overlapping roles for MBD2 & MBD3. Although *MBD2*$^{-/-}$ mice are viable, they show a lack of maternal instinct that results in reduced pup weight (Hendrich *et al*. 2001). When *MBD2*$^{-/-}$ mice are crossed with *Apc*$^{min/+}$ mice, the offspring show reduced adenoma and extended life span compared to normal *Apc*$^{min/+}$ mice (Berger & Bird. 2005).

It has been shown that MBD2 interacts with the NuRD complex via a 633 aa protein called p66 (Brackertz *et al*. 2002). This interaction between p66 and MBD2 has been shown to involve two domains. In p66 aa134-238 (CR1) & aa372-633 (CR2), with the second sequence also containing a Zinc finger GATA domain. In MBD2 the domains aa27-45 and aa211-262 bind p66 in a pull down assay (Brackertz *et al*. 2002). The MeCP1 complex also contains p68 which has high amino acid similarity to p66. p68 also interacts with MBD2 but only through the highly conserved CR1 N-terminal binding site.

1.4.4 MBD3

Human MBD3 is a 265 amino acid 32.2kDa protein with an N-terminal MBD domain and a C-terminus rich in acidic residues. It is has a high sequence similarity to MBD2b

although unlike other members of the MBD family it does not bind methyl-CpG *in vitro* or localize with major satellite DNA *in vivo* (Hendrich & Bird. 1998). Similarly it does not possess DNA demethylase activity (Wade *et al*. 1999). Unlike the human form, *Xenopus laevis* xMBD3 binds methyl-CpG (Wade *et al*. 1999). *Xenopus laevis* also expresses an isoform xMBD3LF, a product of alternative spicing with a sequence of 20 amino acids inserted in the MBD domain, which cannot bind methyl-CpG (Wade *et al*. 1999). xMBD3 is expressed constitutively in the early stages of development with high levels detected in the prospective eye regions, brain, and branchial arches. Suppression of xMBD3 in early development severely effects eye formation and brain development (Iwano *et al*. 2004). Although MBD3 does not bind methyl-CpG it does co-purify with the histone remodeling complex Mi-2/NuRD (Guezennec *et al*. 2006). This coupled with the fact that *MBD3*$^{-/-}$ knockout mice are non-viable, dying during early embryogenesis (Hendrich *et al*. 2001), suggests that MBD3 plays an important role in early development like that of xMBD3. While the MBD domain of MBD3 does not bind methyl-CpG it does bind HDAC and metastasis-associated gene 2 (*MTA2*) (Saito & Ishikawa, 2002). This suggests that MBD3 has conserved its MBD because of the secondary role played by the MBD in protein-protein interactions.

1.4.5 MBD4

MBD4 was identified at the same time as MBD2 and MBD3 by Hendrich & Bird (1998). This 62.6kDa protein has its MBD domain at the N-terminus and lacks the CXXC domains of MBD1. Like MBD1, MBD2 and MECP2, MBD4 can also bind to one symmetrically methylated CpG. The MBD is most similar to that of *MeCP2* in sequence and the protein shows little homology to the other MBD family members out with the MBD domain. MBD4 does however share a low scoring match to bacterial DNA repair enzymes. (Hendrich & Bird, 1999). As well as the N-terminal MBD domain, MBD4 contains a C-terminal DNA glycosylase catalytic domain (Hendrich & Bird, 1999, Bellacosa *et al*. 2005) and it has been shown that MBD4 is a thymine and uracil glycosylase specific for G-T and G-U mismatches resulting from the deamination of 5-methylcytosine and cytosine, respectively, at CpG sites (Hendrich *et al*. 1999). MBD4 is expressed in many human tissues and MBD4-GFP constructs co-localise to the foci of heavily methylated satellite DNA (Hendrich & Bird, 1998). Results from *MBD4* deficient mice suggest that MBD4 suppresses tumorigenesis and CpG mutability *in vivo* as there is an increase in CpG→TpG mutation and accelerated tumor formation in cancer susceptible *Apc*$^{min/+}$ mice (Miller *et al*. 2002).

1.4.6 MBD domain structures

There several MBD domain structures that have been deposited in the protein data bank. The solution structure of MBD1 MBD (Ohki *et al*. 1999), MBD1 MBD in complex with DNA (Ohki *et* al, 2001), solution structure of MeCP2 MBD (Wakefield *et al*. 1999) and the crystal structure of MeCP2 MBD in complex with DNA (Ho *et* al, 2008) have been solved. The NMR derived structures for MBD1 MBD and MeCP2 MBD show similar structures consisting of a wedge shaped fold with four anti-parallel β-strands which constitute one face of the wedge. Comparison of the crystal structure of MeCP2 in complex with DNA (PDB 3C21) and the unliganded NMR structure of MeCP2 (PDB 1QK9) illustrated in figure 1.6 reveals little difference apart from the loop (L1) where most of the protein to DNA sugar-phosphate backbone contacts are made. While there is only moderate homology between MBD1 and MeCP2 there are conserved residues within the MBD of the whole family. It is likely that the MBD of MBD2, MBD3 and MBD4 have similar structures. The lack of binding to methyl-CpG by mammalian MBD3 can be attributed to two amino acid differences within the defined MBD domain, a phenylalanine at position 34 in place of a tyrosine and a histidine at position 30 in place of an arginine (MBD1 & MBD4) or a lysine (MBD2 & MeCP2) (Saito &, Ishikawa. 2002). The effect of the F34 difference has the greatest effect on mCpG binding. This is not surprising due to the interaction of the tyrosine side chain hydroxyl group hydrogen bonding to structural water groups that are involved in the protein-DNA interaction.

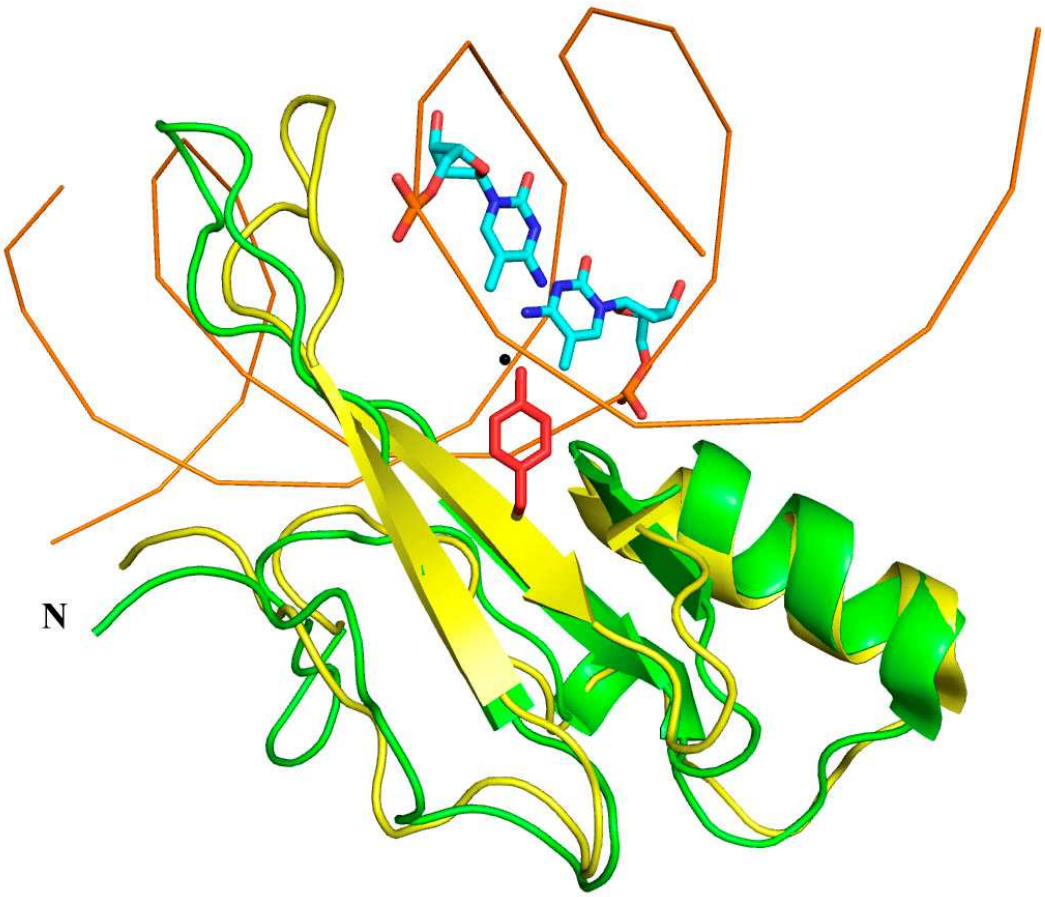


Figure 1.6 Superposition of unbound MeCP2 (green) and MeCP2 bound to DNA (yellow). Methyl-cytosine (sticks) shown as cyan on DNA shown in orange (ribbon), Tyrosine 34 shown in red (sticks) and the structural water shown as a black sphere. N, N-terminus. Image generated in pymol.

1.4.7 The Zinc finger family of methyl-CpG binding proteins

The first member of the zinc finger family of mCpG binding proteins, Kaiso, was identified by a yeast 2 hybrid screen for p120 catenin (Daniel and Reynolds, 1999) and is a component of the NCoR corepressor complex binding via its BTB domain (Yoon *et al*, 2003) Kaiso contains three tandem zinc fingers at the C-terminus that recognize two consecutive symmetrical CpG sequences. In cultured cells, Kaiso recruits the NCoR complex to the MTA2 promoter when methylated. This results in transcriptional repression with H3 K9 being deacetylated and methylated (Yoon *et al*, 2003). Although in normal cells there is presently no information about the target of Kaiso, it is essential for amphibian development. When embryonic levels of the *Xenopus* Kaiso homologue, xKaiso (Kim *et al*, 2002), are depleted by morpholino oligos, premature activation of

zygotic genes occurs before the mid-blastula transition, resulting in failure of gastrulation and death (Ruzov *et al*, 2004). Kaiso also binds to a non-methylated DNA sequence, TCCTGCNA, known as a Kaiso-Binding Sequence (KBS) (Daniel *et al*, 2002). This KBS sequence has been found in the promoter of matrilysin (Spring *et al*, 2005) and the 5' beta-globin insulator (Defossez *et al*, 2005). The proteins ZBTB4 & ZBTB38 (Filion *et al*, 2006) contain a homologous Kaiso zinc finger domain, which also contain a BTB domain but also have additional zinc fingers making them longer and more complex. Unlike Kaiso, the ZBTB proteins can bind a single methylated-CpG which may be due to sequence differences compared to Kaiso within the 3rd Kaiso like zinc finger although their *in vivo* targets are as yet unknown. ZBTB38 has been shown to recruit CtBP to the densely methylated pericentrometric heterochromatin (Sasai *et al*, 2005). CtBP is a corepressor which has also been found to interact with the CXXC domain from MLL (Mixed Linkage Leukemia) (Xia *et al*, 2003) and can induce transcriptional repression by recruiting HDACs (Subramanian & Chinnadurai, 2003). Unlike ZBTB38, ZBTB4 lacks the CtBP interaction site and the BTB domain is disrupted by a long serine-rich insertion (Perez-Torrado *et al*, 2006).

1.4.8 Methylated CpG binding by the SRA domain

UHRF1 was originally identified by Hopfner *et al* (2000) as CCAAT box binding protein before Unoki and co workers (2004) showed that UHRF1 could bind methylated CpG with higher affinity. The SET and Ring finger associated (SRA) domain of UHRF1 is responsible for recognizing a single methylated CpG and interacting with HDAC1 (Unoki *et al*, 2004). Recently Bostick *et al* (2007) showed UHRF1 has a preference for hemi-methylated DNA and that UHRF1 co-localizes with heterochromatin replication foci and interacts with DNMT1 & PCNA, suggesting a role in the transfer of heritable transcriptional repression similar to that of MBD1.

1.5 Epigenetic Disease

Many cellular processes including embryonic development, transcription, chromatin structure, X chromosome inactivation, genomic imprinting and chromosome stability are controlled by epigenetic mechanisms. Any disruption to these processes will have a profound effect on the cell and the organism. A common disruption in humans (1:10,000 births) is Rett Syndrome, a complex neurological disorder mainly affecting girls who become profoundly and multiply disabled. In a large proportion of patients this has been linked to mutations in MeCP2 (Amir *et al*, 1999). Epigenetic disruption is common in

many cancers with the tumour cells displaying global hypomethylation but having hypermethylated CpG islands in conjunction with histone modifications that repress/silence tumour suppressor genes (Fahrner *et al*. 2002, Ballestar *et al*. 2003).

1.5.1 Rett syndrome

Rett syndrome is both sporadic and familial in origin with various mutations found within the gene responsible for its aberrant activity. Amir *et al*. (1999) identified missense mutations within the highly conserved MBD region as well as missense and frame shift mutations within the TRD region while further novel mutations were identified by Wan *et al*. (1999) & Cheadle *et al*. (2000). A recent review by Matijevic *at al* (2009) noted that 67% of all *MeCP2* mutations are caused by a C>T mutation at 8CpG sites (R106, R133, T158, R168, R288, R270, R294 and R306) the most common of which is R168X.

The various mutations of *MeCP2* (nonsense, missense, and frame shift) vary the level of functionality that the protein has. Nonsense and frame shift mutations in the 5' region are likely to lead to unstable transcripts through the process of nonsense-mediated decay (Dragich *et al*. 2000) while several missense mutation show reduced binding to mCpG *in vitro* (Ballestar *et al*. 2000). Two common mutations, R106W & R133C, are located within the β-sheet (Wakefield *et al*. 1999) and show 100 fold reduction in affinity for binding methylated DNA in unassembled chromatin (Ballestar *et al*, 2000). However in native mouse heterochromatin only R106W had markedly reduced binding, whereas R133C showed wild-type-like affinity and increased repression of a non-methylated reporter construct in *Drosophila* cells (Kudo *et al*. 2001). This increased affinity for non-methylated CpG in the R133C mutant may be party due to the alteration of the DNA binding site since R133 is completely conserved between MBD family members while mutation of this arginine to an alanine in MBD1 abolishes binding to mCpG (Ohki *et al*. 1999). Another common mutation, T158M, found in the loop structure of the MBD, showed only a 2-fold reduction in the binding to methylated DNA while it retained partial affinity for heterochromatin (Kudo *et al*. 2001). Disruption of mCpG binding and the ability of MeCP2 to interact with HDACs will affect the role of MeCP2 in transcriptional repression. The brain has the highest abundance of MeCP2 compared to any other tissue (Nan *et al*. 1997) which might suggest that brain tissue is highly susceptible to abnormalities arising from aberrant gene expression. This idea is supported by work carried out in 2000 by Cheadle *et al*. who noted that patients with

missense mutations had significantly milder disease compared to those with truncated isoforms. This may be due to truncation abolishing MeCP2's ability to interact with Sin3A and recruit histone deacetylase to repress transcription through its TRD (Jones *et al*, 1998, Nan *et al*, 1998). However, since inhibition of histone deacetylase activity by the drug Trichostatin A only partially relieves transcriptional repression by the TRD suggesting the TRD can mediate repression in a histone deacetylase independent manor (Jones *et al*, 1998, Nan *et al*, 1998). This may not be the only mechanism.

1.5.2 Disease associated with other MBD proteins

MBD1, 2, 3 do not have any known disease/syndrome associated with mutations in their genes. Knockout mice have been produced for all except MBD3 for which knockout is a lethal genotype. MBD1 knockout mice show no obvious development abnormalities and appeared healthy throughout their life. However, their neural stem cells exhibited reduced neuronal differentiation, increased genomic instability and deficits in adult neurogenesis and hippocampal function (Zhao *et al*, 2003) indicating that MBD1 maybe important for normal brain function. As a consequence of knocking out MBD2, *MBD2*$^{-/-}$ mice appear to suffer post natal depression but no other physical or mental abnormalities (Hendrich *et al*. 2001). MBD4 knockout mice suffer increased (3-fold) C to G mutations particularly at mCpG sites and when crossed with the *Apc*$^{Min/+}$ mouse the resultant *Mbd4* $^{/}$, *Apc*$^{Min/+}$ litter have increased levels of intestinal neoplasia when compared to *Apc*$^{Min/+}$ (Miller *et al*. 2002).

While there are no disease syndromes attached to the MBD1-3 proteins, MBD4 has been shown to be mutated in colorectal tumours (Bader *et al*. 1999) and primary gastric cancer (Yamada *et al*. 2002). However, expression levels of these proteins have been shown to vary in cancer cell lines and tissue. Patra *et al*. (2003) compared the expression of MBD1, MBD2 and MECP2 in prostrate and BPH cancer cell lines. They found protein expression levels of MBD1 increased in the cancer cells compared with a BPH (human prostatic epithelial) cell line, with the protein levels of MBD2 & MeCP2 being repressed at the translation step, while levels of DNMT1 and HDAC1 where significantly increased. MBD1 has been shown to repress transcription of the p16 tumour suppressor gene (Hendrich *et al*. 1999) although levels of this protein where not analysed.

1.5.3 Disease associated with Hypermethylation

Hypermethylation of CpG islands occurs within the promoter regions of many critical cancer genes (Baylin *et al*. 1999, Esteller *et al*. 2001, Billard *et al*. 2002) which leads to transcriptional repression. The promoter region of the π-class glutathione *S*-transferase gene, *GSTP1*, has been shown to be hypermethylated in >90% of prostate cancers (Lee *et al*. 1997, Lin *et al*. 2001.), >70% in liver cancers (Tchou *et al*. 2000) and >30% in breast cancer (Esteller *et al*. 1998). The reduction in gene transcription has been attributed the hypermethylation event as treatment of the cells with inhibitors of DNMT1 increased both GSTP1 mRNA and protein levels (Lin *et al*. 2001, Singal *et al*. 2001). Using siRNA techniques Lin & Nelson (2003) knocked down the transcription of MBD2 (~79%), DNMT1 (~83%) & MeCP2 (~77%) in MCF-7 (breast cancer) cells and recorded an increase in transcription of GSPT1 from only the MBD2 and DNMT1 siRNA treated cells. DNMT1 and MBD2 have also been detected bound to the hypermethylated GSTP1 promoter in MCF-7 cells using ChIP assays (Lin & Nelson, 2003) but it is unknown if DNMT1 is repressing transcription although the recruitment of DNMT1 to the GSTP1 promoter is dependent on the methylation of the CpG island (Lin & Nelson, 2003). These data suggest that both MBD2 and DNMT1 are involved in aberrant transcriptional repression of GSTP1 in the MCF-7 cell line although the exact role of DNMT1 has not been described at this time.

MBD2 expression has been studied in other cell lines. Billard *et al*. (2002) studied MBD2 expression during normal and pathological growth of the human mammary gland. They found that MBD2 expression was 20-30 fold higher in benign tumours than normal tissue, but in neoplastic samples MBD2 and MECP2 were deregulated. It was also noted that MBD2 expression varied with tumour size in invasive ductal carcinomas. In contrast to the varied transcription in breast cancer, human colorectal, gastric cancerous tissues (Kanai *et al*. 1999) and the peripheral blood lymphocytes in bladder cancer patients (Zhu *et al*. 2004) show a decrease in transcription suggesting there could be a protective role for MBD2 in these diseases.

DNMT1 has also been shown to be over expressed in various cancers compared to normal tissue (Sun *et al*. 1997, Kanai *et al*. 2001, Saito *et al*. 2001), and that over expression correlated significantly with detection of methylated CpG islands (Kanai *et al*. 2001). However, DNMT1 is expressed mainly during S-phase and because tumour tissue is presumed to contain a greater proportion of dividing cells than normal tissue is,

the increase in DNMT1 expression could be due to an increase in dividing cells or an increase in each cell.

MBD2 and DNMT1 have been implicated as key players in various gastric cancers and are therefore targets for anticancer drugs. MBD2 appears to be the more attractive target as knocking out MBD2 does not appear to have any distinct or severe side effects unless you are a nursing mouse, and while inhibition of DNMT1 can restore almost wild type growth in cell culture it is not without side effects. The interaction between p66 and MBD2 within the MeCP1 complex is an obvious target for drug design.

1.6 Non-methylated CpG and CXXC domains

The CXXC domain is characterized by a cystine rich CXXCXXC repeat where X is any amino acid. CXXC domains are found in chromatin binding proteins MBD1 (2 or 3 CXXC domains depending on splice variant), DNA methyltransferase 1 (DNMT1) (Bestor 1992), CpG-binding protein (CGBP) (Lee *et al*. 2001) and the mixed-linage leukaemia gene (MLL) (Tkachuk *et al*, 1992). The CXXC domain in DNMT1 as been shown to bind zinc (Bestor 1992) while the CXXC domain of CGBP has been shown require zinc to bind non-methylated CpGs in vitro (Lee *et al*. 2001). The 3rd CXXC domain in MBD1 has also been shown to bind non-methylated CpG motifs and cause the repression of non-methylated reporter genes via the transcriptional repression domain while CXXC1 & -2 do not (Jorgensen *et al*. 2004). The CXXC domains (1 to 3) from MBD1, HRX/MLL, CGBP and DNMT1 show sequence similarity mainly through 8 cystine residues that form the two zinc clusters (fig 1.7) with CXXC-3 showing further homology to CGBP and other non-methylated DNA binding proteins (Jorgensen *et al*. 2004). The CXXC domain sequences shown in figure 1.7 also contain a glycine residue succeeding the first cysteine in each cluster (pink) However, MBD1 CXXC1 and CXXC2 lack the sequence KFGG (and other conserved residues D 31 & Q43) compared to the other CXXC domains.

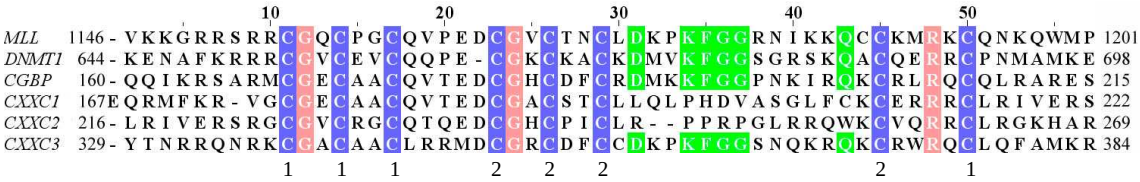


Figure 1.7 Alignment of CXXC domains from hMLL (NP_005924), hDNMT1 (NP_001370.1), hCGBP NP_055408), MBD1 CXXC1, CXXC2, CXXC3 (NP_056671). Sequences were aligned using Jalview (Clamp *et al*, 2004). Cysteines involved in forming the zinc clusters are shaded blue and numbered according to the cluster they participate in. Residues conserved between domains than bind non-methylated CpG are shaded green and non cysteine common residues between all sequences are coloured pink

As noted above (p6) methylation occurs at 70-80% of mammalian CpG sites (Ehrlich and Wang 1981) with the remaining unmethylated CpG sites being found in CpG rich islands usually found in the promoter regions of house keeping genes (Bird 1986). It is therefore likely that CXXC domains have a role in binding to CpG islands. Recently Thomson *et al* (2010) showed that CXXC finger protein 1 (*CFP1*), part of the Setd1 H3K4 methyltransferase complex (Lee *et al*, 2007), localize to CpG islands and appears to be required for H3K4 methylation, which is normally associated with transcriptionally active promoters. In acute myeloid leukemia the MLL protein is fused with *LCX* (leukemia-associated protein with a CXXC domain) (Ono *et al*, 2002). This fusion protein is essential for target recognition, transactivation, and myeloid transformation by the MLL oncoprotein (Atton *et al*, 2004). With a mutated CXXC domain from MLL which is unable to bind non-methylated CpG, the MLL/LCX fusion protein is unable to induce transcription from target genes. The direct functions of the CXXC domains from MBD1 are currently unknown. It has been shown that the CXXC domain from DNMT1 is essential for enzymatic activity and that single point mutations in the CXXC domain disrupt the DNA binding ability but do not interfere with DNMT1 targeting to replication foci (Fatemi *et al*, 2001., Pradhan *et al*, 2008). CXXC domains also have been shown to interact with various proteins involved in chromatin remodeling. Xia *et al* (2003) investigated the protein-protein interactions of the MLL/LCX CXXC domain. Using GST-pull down assays and co-immunoprecipitation they were able to show that the CXXC domain interacts with CtBP (C-terminal binding protein) and hPc2 (human polycomb 2 protein). CtBP can meditate transcriptional repression by recruiting HDACs (Subramanian & Chinnadurai, 2003) while hPc2 is a member of the PRC1 complex (shao *et al*, 1999). MLL/LCX CXXC recruitment of PRC1 mediates transcription

repression independent of HDAC activity (Xia *et al*, 2003). Fuks *et al* (2000) also reported the DNMT1 CXXC domain could mediate HDAC-independent transcriptional repression using the Gal4 reporter system. The MBD1 CXXC3 domain is similar to the MLL & DNMT1 CXXC domain in sequence and it can bind non-methyl CpG DNA (Jorgensen *et al*. 2004). With this similarity in mind, Sakamoto *et al* (2007) used a yeast two hybrid screen to identify proteins that interact with the three CXXC domains from MBD1. They were able to show the MBD1 CXXC3 also interacted with hPc2 while CXXC12 interacted with amino acids 250-337 of Ring1b. Ring1b, like hPc2, is a major component of the Polycomb group (PcG) multiprotein PRC1 complex (Min *et al*, 2003). This suggests that although the first two CXXC domains from MBD1 do not bind CpG DNA they have a role in transcriptional repression through the interaction of the PRC1 complex.

The solution structure of the MLL CXXC domain (Allen *et al*, 2006), MLL in complex with DNA (Cierpicki *et al*, 2009), xDNMT1 (Thomson & Smith, unpublished) and the crystal structure of DNMT1 CXXC in complex with DNA (Song *et* al, 2010) have been solved. The structure of MLL CXXC domain forms a crescent like structure containing two zinc atoms. Each zinc atom is coordinated by 4 cystine residues (figure 1.8). These two zinc atoms are required for structural integrity of the domain since mutation of any of the conserved cysteine residues results in an unfolded protein (Allen *et al*, 2006, Cierpicki *et al*, 2009)

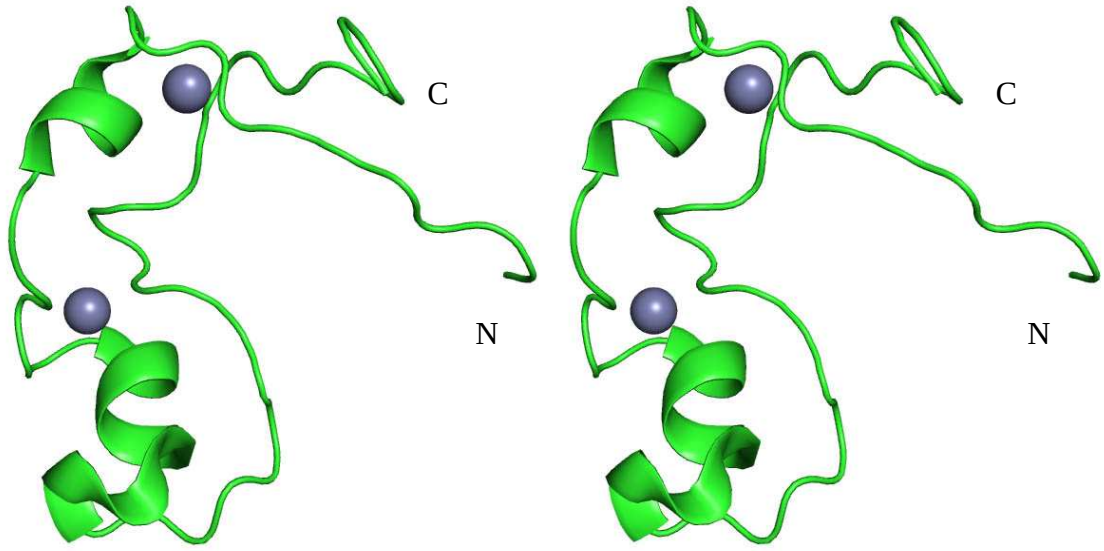


Figure 1.8 Stereo view of the solution structure of the MLL CXXC domain in cartoon form (Cierpicki *et al*, 2009). Zinc atoms shown as gray spheres, N, N-terminus; C, C-terminus. PDB 2KKF (Figure prepared using PyMol).

Shown in figure 1.9, the MLL CXXC domain binds the CpG motif in the major groove though formation of hydrogen bonds from the N4-amine groups of C118 and C106 and the backbone oxygen of K1185 and K1186 respectively. The side chain of K1186 forms hydrogen bonds to G119 while the side chain amide of Q1187 hydrogen bonds to G107 (Cierpicki *et al*, 2009). Cierpicki *et al* (2009) also discovered weak NOEs between the side chains of R1150, S1152 and L1197 and the minor groove. Mutation of R1150 and L1197 to alanine resulted in a 5 and 4 fold reduction in DNA binding affinity respectively.

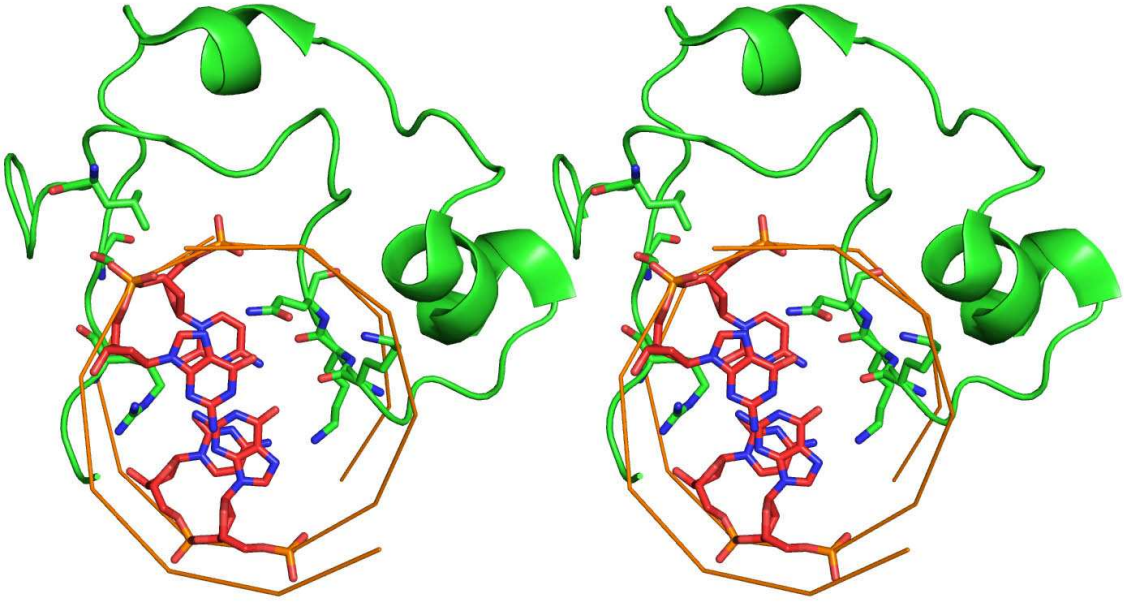


Figure 1.9 Stereo view of the solution structure of the *MLL* CXXC domain in complex with DNA in cartoon form showing the CpG palindrome in red, the DNA backbone in orange and the residues involved in DNA binding N, N-terminus; C, C-terminus. PDB 2KKF (Cierpicki *et al*, 2009) (Figure prepared using PyMol).

CHAPTER TWO

INTRODUCTION PART 2: PROTEIN NMR

2.1 Nuclear Magnetic resonance

In this section the fundamental theory underlying the experiments that I applied in investigating MBD1 CXXC1 and MBD1 CXXC12 are outlined. I have written this chapter from the point of view of a biologist using NMR as a technique used to answer questions about a proteins function. As such I hope that a scientist without any experience of NMR should be able to follow and understand.

2.1.1 Basic Theory of one dimensional NMR

Nuclear magnetic resonance (NMR), first described by Felix Bloch and Edward Mills Purcell in 1946, is the phenomenon of atoms absorbing radio-frequency electromagnetic radiation (RF) under the influence of a magnetic field. For protein NMR the detection of this absorption requires the use of isotopes of carbon, nitrogen and hydrogen which have a spin 1/2. Spin is a quantum mechanical property attributed to protons, neutrons and electrons, measured in multiples of ½ and can be positive or negative. In protein NMR, isotopes of carbon, nitrogen and hydrogen that have a nuclear net spin +/- ½. Nuclear net spin ½ occurs when the isotopes have an unpaired proton e.g. ^{1}H, ^{13}C, ^{15}N are used. It is a quantum mechanical requirement that nuclei with a spin ½ be in one of two states (low or high energy state) when placed in a magnetic field (figure 2.1). When no external magnetic field is applied the two states are of equal energy. The energies of the spin states are split with the application of a magnetic field.

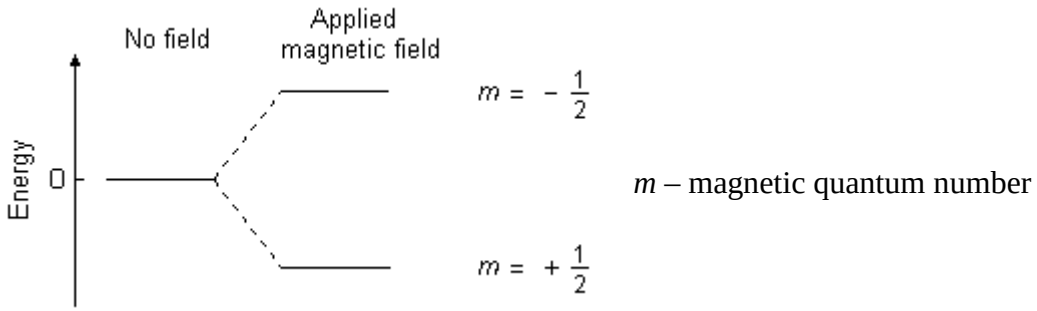


Figure 2.1 Figure showing the effect on the potential energy of the spin states on a nucleus when an external magnetic field is applied.

At equilibrium jumping from one spin state to the other is a relatively infrequent occurrence. However, application of RF energy equal to the difference between the two states will generate a greater amount of jumping between states. The frequency of the RF energy required for nuclei to absorb radiation is known as the Larmor frequency. The Larmor frequency is equal to the frequency of the precessional orbit of the nucleus in the applied magnetic field. After a pulse of RF energy, the spin states of the majority of nuclei will under go relaxation to their original state. This relaxation is known as a Free Induction Decay (FID) The energy required to flip the states depends on the strength of the magnetic field they are placed in, the type of nucleus e.g. ^{1}H or ^{13}C, and the chemical environment of the nucleus e.g. the methyl and hydroxyl protons of methanol absorb at different frequencies and the amide protons of two tryptophans in a protein are likely to absorb at different frequencies because they are in different chemical environments. This effect is known as chemical shift and is a major source of information in NMR spectra for assignment of atom types since protons attached to Cα, amide N and side chain heavy atoms have characteristic chemical shifts.

2.2 Homonuclear ^{1}H NMR

In protein NMR there are various experiments that can be performed in order to elucidate a three dimensional structure. In essence they fall into 2 categories: through bond, where the magnetization is transferred from atom to atom through the intervening covalent bonds; and through space, where the magnetization is transferred between atoms that are close together (< 5Å apart) but not necessarily covalently linked. The experiments are performed on samples that are either unlabeled or labeled with ^{13}C and/or ^{15}N. NMR structures of small proteins <10kDa can normally be elucidated with unlabelled samples using homonuclear NMR, since naturally abundant hydrogen has a spin ½ (Wuthrich, 1986).

2.1.1 Water suppression

The NMR experiments carried out for this thesis were performed in aqueous solutions. The concentration of ^{1}H nuclei in water is approximately 110 M compared to 0.5-1.0 mM for the protein molecules. This results in ~10^5 greater equilibrium magnetization of water ^{1}H spins. Without suppressing the signal from the water, detection of the protein signal without spectrum distortions is impossible. By replacing H_2O with D_20 (99.999%) as the solvent, it is possible to reduce the H_2O resonance signal a factor of 10^5. However, signals from exchangeable, e.g. amide ^{1}H, nuclei will be reduced or

absent from the spectra. Since scalar coupling and dipolar interactions between backbone amide ^{1}HN spins with amide ^{15}N and ^{1}Hα spins are required for backbone assignment certain NMR experiments must be carried out in an H_2O solvent.

The simplest method of water suppression is presaturation. A long low powered pulse which is applied scrambles the water signal around the applied Z-axis. The net effect of this pulse causes the water signals from different parts of the sample to cancel each other out before the main sample pulse sequence is initiated. Water suppression by presaturation is a simple and effective method. However, $^{1}H^{\alpha}$ spins close to the water signal can be saturated by the RF field and saturation transfer may partially saturate exchangeable ^{1}H spins. A more effective water suppression technique is excitation sculpting (Hwang and Shaka, 1995). This method of water suppression is executed before detection of the FID and requires pulse field gradients. A gradient pulse is first applied to scramble the homogeneity of magnetization in the sample. A selective 180º pulsed on water followed by a non selective 180º pulse turns the water through 360º. Since the initial magnetization on non-water protons has been turned though 180º a second gradient pulse refocuses the non-water signals but not the water. A variation of this excitation sculpting is known as a 3-9-19 watergate. This type of water suppression is more suited to 2D and 3D experiments due to the use of selective pulses that do not excite water rather than the water selective and non-selective pulse in the Hwang and Shaka (1995) method. The use of the non-water selective pulses shortens the time of the water suppression element in the pulse sequence, reducing the length of time the sample has to relax.

2.2.2 1D homonuclear NMR

A 1D ^{1}H NMR experiment consists of two parts: preparation and acquisition. In the simplest case, shown in figure 2.2, preparation is a 90˚ pulse applied along the x-axis which transfers the magnetization from the z-axis into the transverse plane to the negative y-axis. The receiver coil records the signal emitted by the nuclei as they precess at their Larmor frequency (figure 2.3)

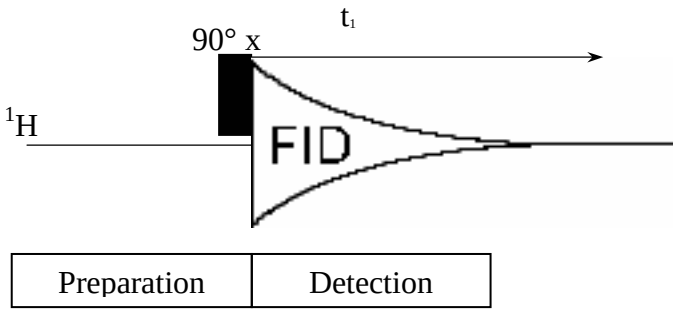


Figure 2.2 **S**chematic diagram of a simple 1D NMR experiment

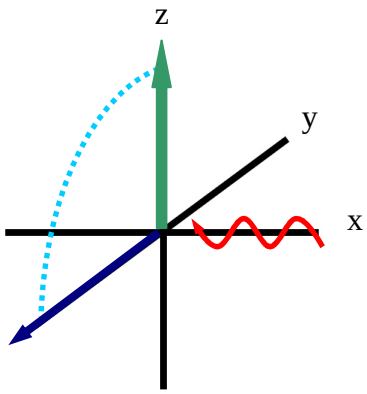


Figure 2.3 Effect of applying an external RF field (red) of sufficient strength along the x axis to induce rotation of the equilibrium magnetization (green) magnetization into the transverse plane to –y (dark blue).

The 1D ^{1}H NMR experiment records the RF emission from the nuclei of ^{1}H atoms only and the FID, a concentration of sinusoidal waves (time domain) is Fourier transformed to provide a frequency domain spectrum (figure 2.4).

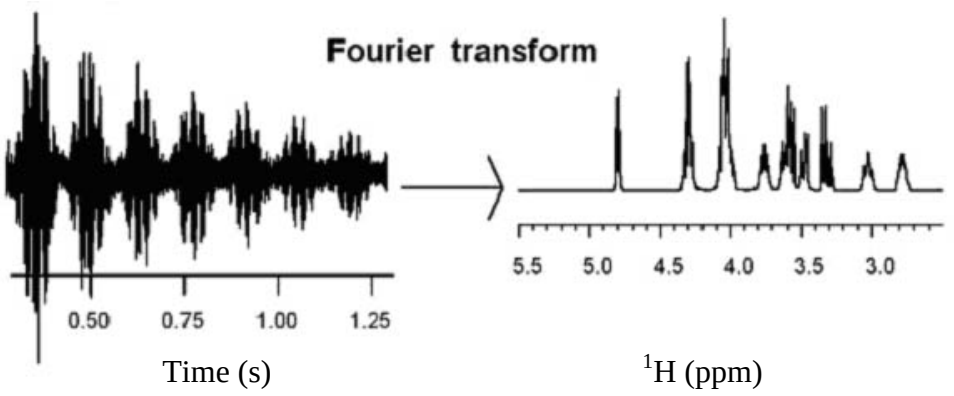


Figure 2.4 Fourier transformation of a time domain signal to a frequency domain spectrum

This 1D spectrum can be used to provide information as to whether the protein is structured. The x-axis represents the chemical shift, a relative frequency scale expressed in parts per million compared to the reference compound tetramethylsilane and figure 2.5 relates the peaks to different chemical environments.

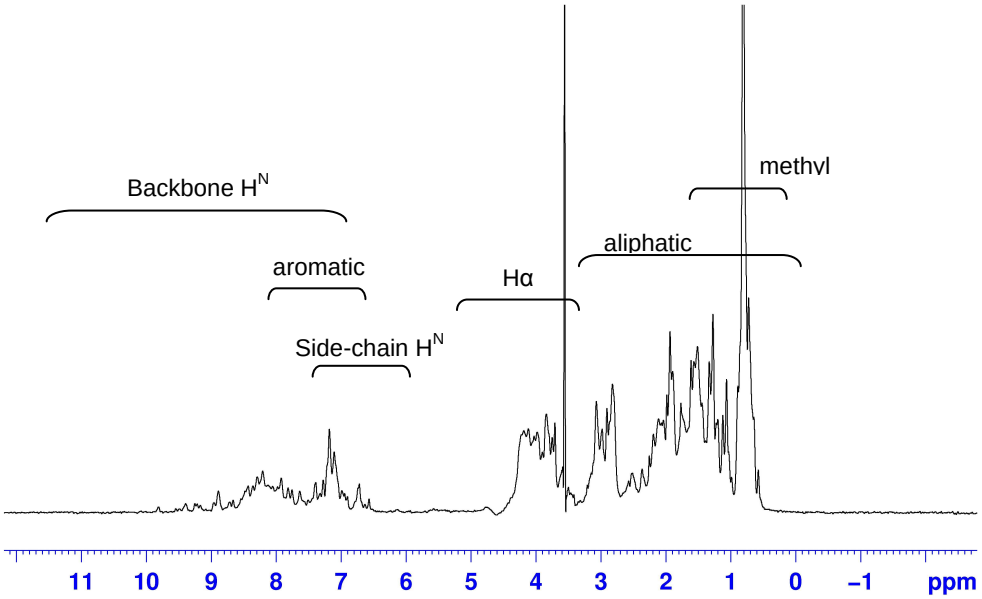


Figure 2.5 1D ^{1}H spectra of unlabelled hMBD1 CXXC1 (10mM deuterated tris(hydroxymethyl) aminomethane, 250mM NaCl, pH 6.5) relating chemical shift (in ppm) to different chemical environment.

The 1D ^{1}H experiment can be a powerful diagnostic tool since the sharpness or broadness of the peaks relate to how structured the protein/peptide is. Very sharp methyl peaks coupled with lack of dispersion (where peaks cannot be distinguished one from another) of the backbone amides clustered around 7-8.5ppm can indicate unstructured protein. A structured protein will show good dispersion of peaks (Figure 2.5) in the methyl and backbone amide regions. Broader than expected line widths can be indicative of protein aggregation and/or conformational flexibility.

2.2.3 2D Homonuclear 1H NMR

The 1D ^{1}H experiment is limited in the information it can provide and it is not usually possible to assign individual peaks to specific amino acid residues due to overlap. 2D homonuclear NMR experiments utilize magnetization transfer between hydrogen nuclei that appear as crosspeaks in the spectra. This correlation of hydrogen nuclei is the basis of amino acid assignment and distance restraint measurement. The simplest 2D homonuclear experiment (COSY - Correlation Spectroscopy) takes the preparation and

detection from a 1D experiment (figure 2.2) and adds an indirect evolution time (T_1) and mixing sequence (figure 2.6).

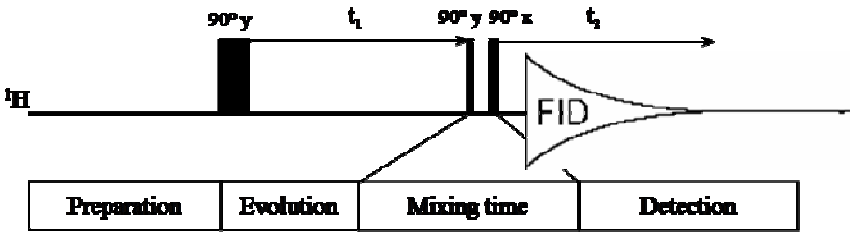


Figure 2.6 Anatomy of a simple 2D homonuclear experiment (COSY).

After preparation, the spins precess for a given time, T_1, where the magnetization is labelled with the chemical shift of a proton. The mixing time transfers some of magnetization to another proton before detection. During detection, the magnetization is labelled with the chemical shift of the correlated proton. Magnetization can be transferred during the mixing time by scalar coupling for through bond Total Correlation Spectroscopy/Correlation Spectroscopy (TOCSY/COSY) or dipolar interactions for through space Nuclear Overhauser Effect Spectroscopy (NOESY) experiments. The COSY experiment only transfers magnetization through up to 3 bonds due to the weakness of the scalar coupling over greater number of bonds. Figure 2.7 illustrates the connectivities that give rise to cross peaks in TOCSY & NOESY spectra. A 2D homonuclear experiment produces a spectrum with two frequency axes. Signals for each proton detected are recorded along a diagonal, splitting the spectrum into two halves.

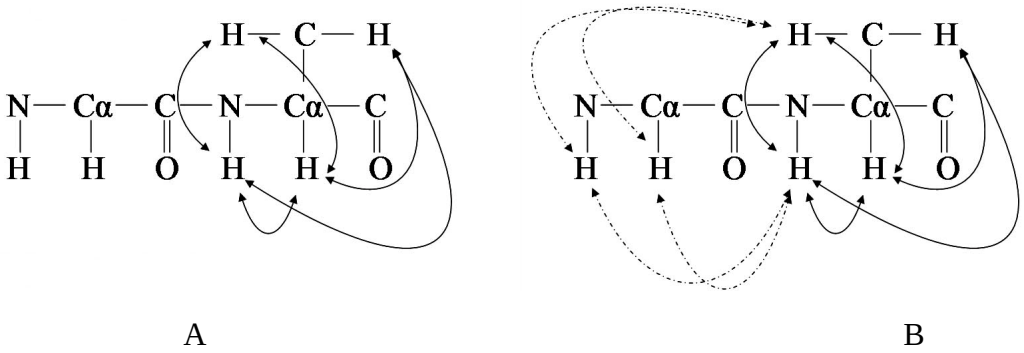


Figure 2.7 Connectivities that can give rise to cross peaks in A) ^{1}H-TOCSY and B) ^{1}H-NOESY. Solid arrows represent intra-residuer transfer, dotted lines represent inter-residue transfer.

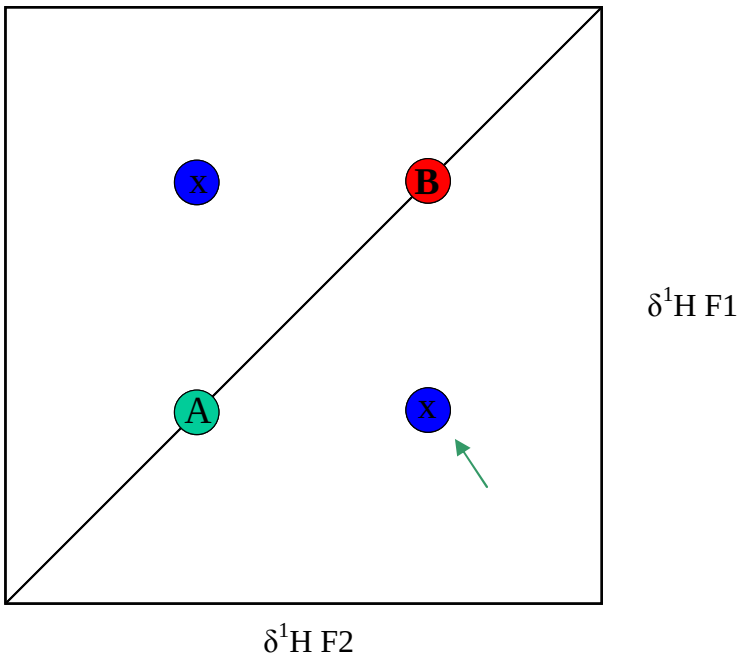


Figure 2.8 Schematic of a 2D homonuclear spectrum. Protons A & B are correlated by the cross peaks X. Cross peaks either side of the diagonal result from magnetization originating from each proton. The green arrow indicates a cross peak arising from magnetization which originated on proton A that was transferred to B.

This diagonal results from magnetization which has not been transferred during the mixing time and therefore remains on the same nucleus. Signals not on the diagonal are the result of two protons exchanging magnetization during the mixing time (figure 2.8). The cross peaks that correlate a proton to another proton provide information for assignment of each cross peak in the spectrum.

2.3 Protein Assignment using ^{1}H homonuclear data

There are typically three 2D homonuclear experiments, COSY (Correlation Spectroscopy), TOCSY (Total Corralation Spectroscopy) and NOESY (Nuclear Overhouser Effect Spectroscopy) that can be used for resonance assignment and extraction of distance information. These experiments are used to assign spin systems, each of which is a group of resonances representing the chemical shifts of each residues protons (amide, alpha and side chain). As with the 1D ^{1}H spectrum (figure 2.5) specific regions within the 2D ^{1}H spectrum correlate transfer between different proton groups. Figure 2.9 shows the layout of a 2D ^{1}H spectrum relating chemical shift coordinates to proton-proton magnetization transfer.

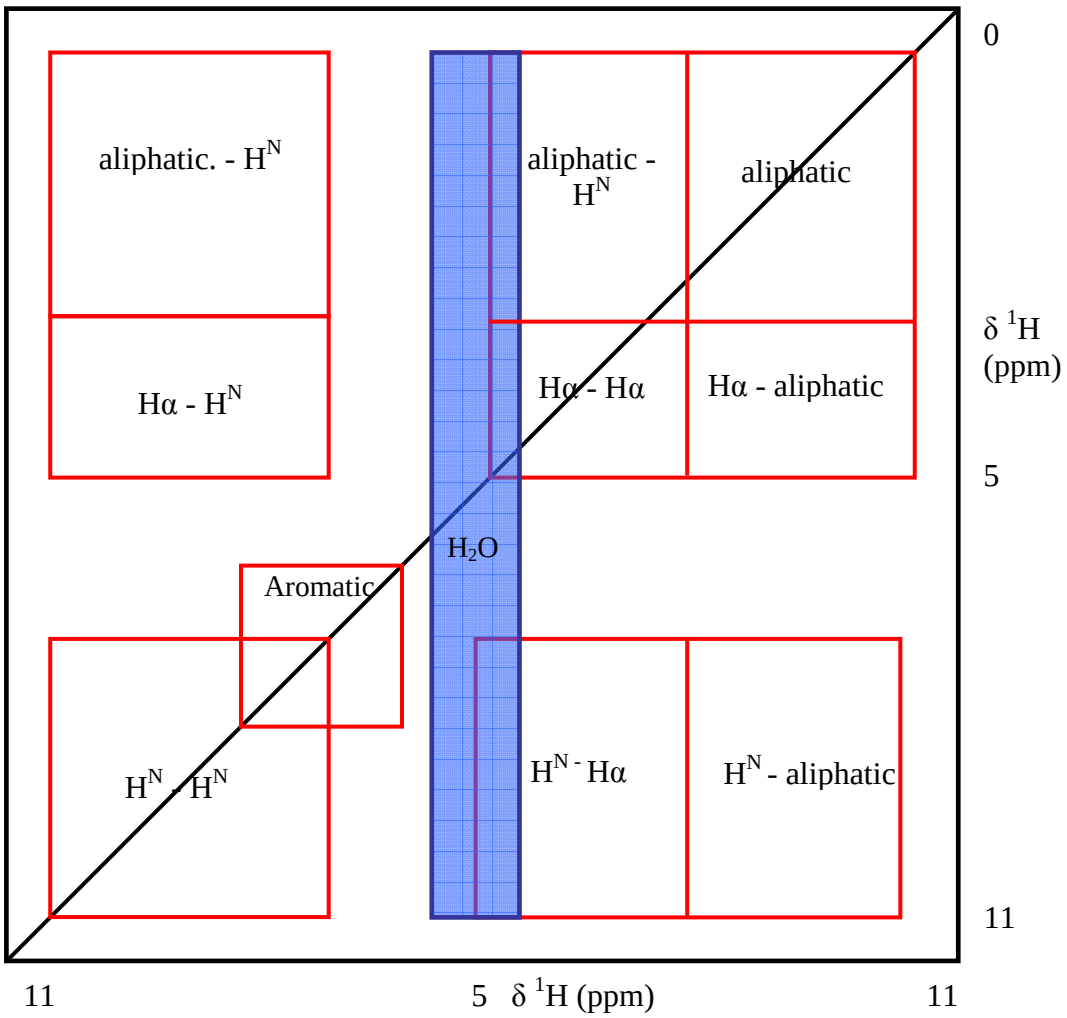


Figure 2.9 Layout of magnetization transfer for a 2D ^{1}H experiment in H_2O

2.3.1 Correlation spectroscopy

The COSY experiment transfers the magnetization by scalar couplings and as such only crosspeaks between protons that are up to 3 bonds apart in the spectrum since couplings over 3 bonds are too small to be visible. The COSY experiment is important for discriminating β & γ protons for example in glutamate where the chemical shift ranges of the β & γ protons overlap. There would be a COSY cross peak correlating the Hα to the Hβ but not to the Hγ. Every amino acid has specific chemical shift ranges for protons with in each residue.

The TOCSY (total correlation spectroscopy) experiment correlates every proton in an amino acid residue (figure 2.10) Magnetization is transferred through the entire spin system (amino acid) by successive scalar couplings. As both the TOCSY and COSY experiments depend on scalar coupling, they can be used for identifying the chemical shift values for the resonances within each spin system.

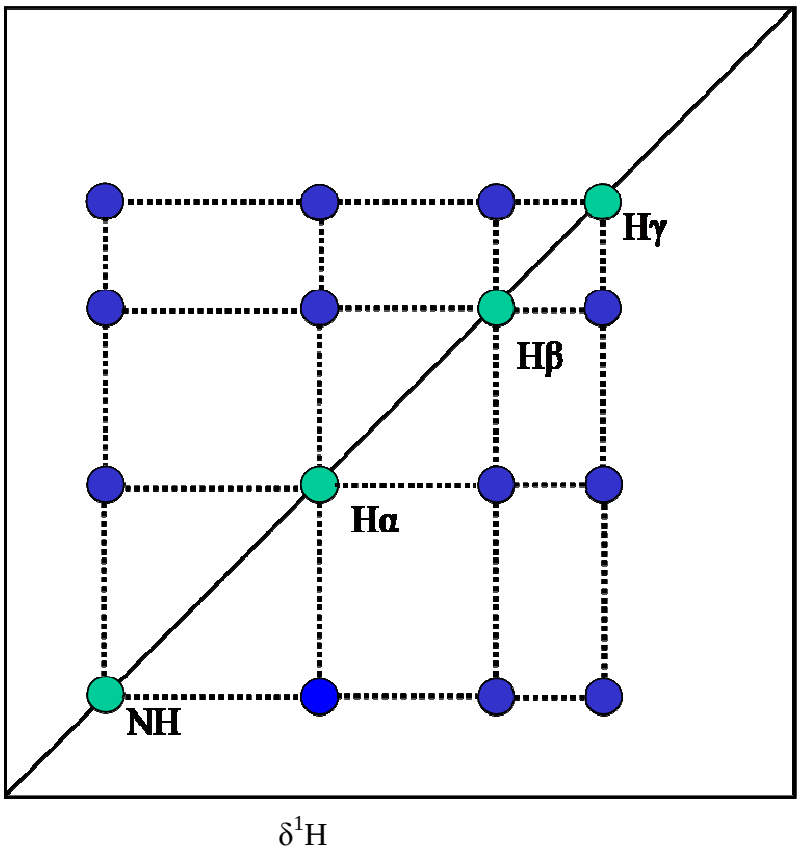


Figure 2.10 Diagram the pattern of crosspeaks for an arbitrary spin system in a TOCSY experiment. Green circles – autocorrelation crosspeaks for the nuclei of the spin system, blue circles – cross peaks correlating nuclei.

However, this characteristic pattern (figure 2.11) is shared between certain amino acids e.g. cysteine, aspartic acid, phenylalanine, histidine, asparagine, tryptophan and tyrosine have the same pattern. Therefore while it is possible to assign cross peaks to a particular spin system it is not possible to assign the spin system to a specific amino acid on the basis of the pattern of chemical shifts alone. It is not possible to sequentially link or calculate structure based solely on spin system assignments in COSY/TOCSY spectra. Assignment of spin systems to sequence specific amino acids requires the spin systems to be linked together (see chapter 2.3.3).

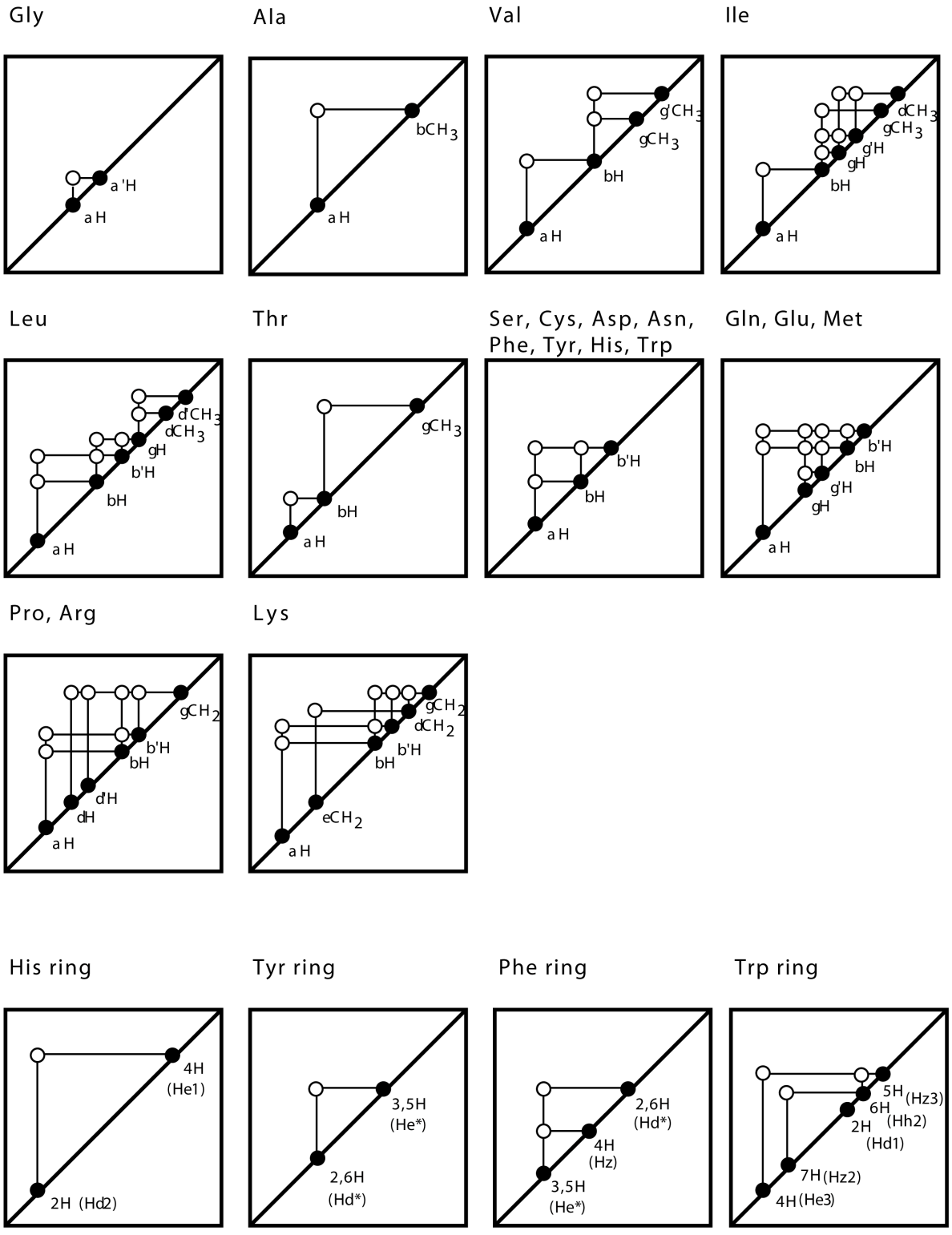


Figure 2.11 Characteristic COSY spectra for each of the 10 classes of amino acid residiues and aromatic side chains (reproduced from Smith, 1994)

2.3.2 Nuclear Overhauser Effect Spectroscopy

For structure determination by proton NMR, information about the distance between atoms is crucial. The NOESY (Nuclear Overhauser Effect Spectroscopy) experiment relies on dipolar interactions of spins to transfer the magnetization, through space, from one nucleus to another. This through space magnetization transfer allows distance information to be obtained for both intra- and inter-residue interactions between protons that are typically less than 5Å apart. The strength of correlation between two nuclei depends on the distance between them. The closer the nuclei are to each other the more intense the cross peak in the spectrum. The NOESY spectrum not only contains cross peaks for intra-residue correlation, which are at the same chemical shift coordinates as the COSY/TOCSY correlations, but also inter-residue correlations between nuclei close in space. This information is required for linking spin systems together in order to assign them sequence specifically.

2.3.3 Sequential Assignment

Since correlations between nuclei in the same spin system are at the same chemical shift in both NOESY and TOSCY/COSY spectra and the Hα to $H_{N\ i+1}$ distance is often short there is typically an NOE cross peak correlating the Hα from one amino acid and the H_N the next in sequence. Figure 2.12 shows an example of backbone sequential linking using the $H\alpha_i$ to $H_{N\ i+1}$ inter-residue NOE. It is not possible to rely solely on the $H\alpha_i$ to $H_{N\ i+1}$ crosspeaks to sequentially link spin systems since there is not always a strong or well resolved $H\alpha_i$ to $H_{N\ i+1}$ cross peak. For example, residues in an alpha helix have weak $H\alpha_i$ to $H_{N\ i+1}$ but strong H_N to $H_{N\ i+1}$ NOEs. Using the known protein sequence it is possible to link spin systems using NOE crosspeaks correlating H_N and Hβ protons and H_N and H_N resonances of i to i_{+1} to make the sequential assignment less ambiguous. Another problem arises when trying to sequentially link a proline residue since proline lacks an amide proton. Proline residues can normally be sequentially linked using the correlation of the proline Hδ and the Hα of the residue before. Once spin systems have been linked sequentially, assignment of the side chain chemical shifts can be completed using the COSY experiment to help distinguish side chain protons.

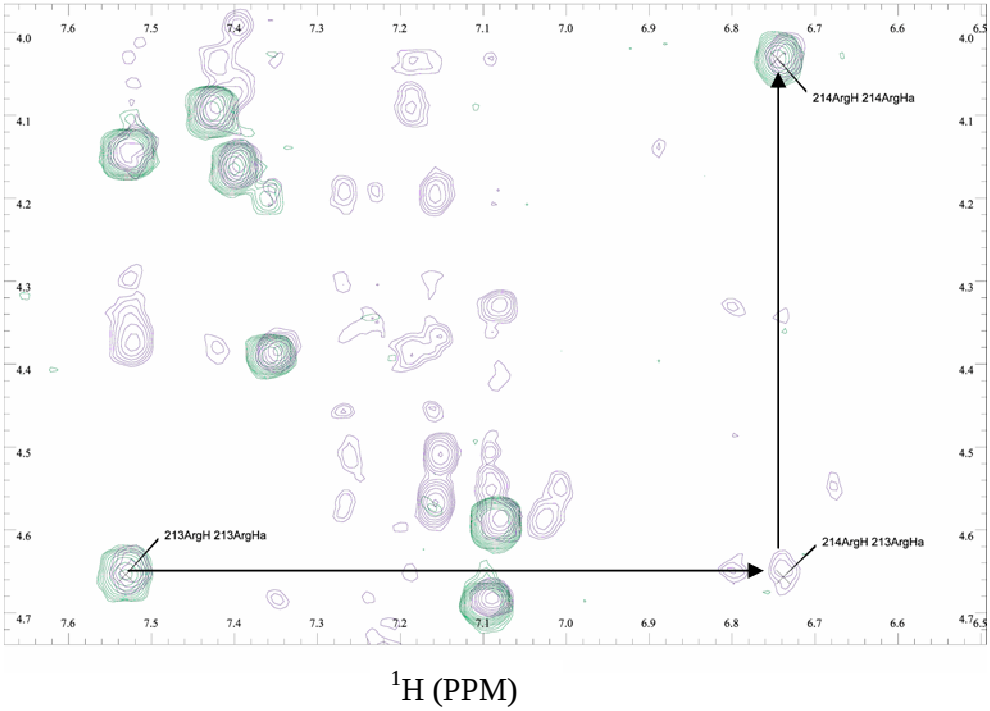


Figure 2.12 Example of sequentially linking backbone resonances of hMBD1 CXXC1 using 2D homonuclear TOCSY (green) & NOESY (blue) spectra by identifying the inter-residue 213Arg Hα to 214Arg H_N crosspeak.

2.3 Heteronuclear NMR

Atom specific assignment of resonances observed in homonuclear experiments are dependant on well resolved spectra. This can normally be carried out for small globular proteins that produce spectra with well dispersed chemical shifts for the protons in each of the residues. Spectra become less well resolved as proteins increase in size due to:

1) The increase in number of resonances due to the increase in number of protein residues while the chemical shift range over which signals are spread does not change. This increases the probability of overlap of cross peaks making assignments more difficult.

2) The decreased rotational correlation time of the protein causing faster relaxation, thus increasing overlap due to from broader linewidths.

The use of ^{15}N labelled protein allows NMR experiments to be carried out which add a third (nitrogen) dimension to the spectra. Separation overlapped 2D crosspeaks possible because the spin systems can have different nitrogen chemical shifts (see chapter 4 p81 for a specific example). When proteins are labelled with both ^{13}C and ^{15}N, NMR experiments can be recorded that transfer magnetization across the peptide bond connecting two spin systems though bonds meaning you don't have to use NOEs to get sequential assignment to work.

2.3.1 2D ^{15}N-HSQC

The ^{15}N heteronuclear single quantum correlation (HSQC) experiment correlates the nitrogen atom of an $NH_{(x)}$ group to the attached proton. Thus, the HSQC spectrum contains a cross peak for each backbone amide group except proline. Also visible in an HSQC spectrum are the NH_2 side chain groups of Asn & Gln, the aromatic NH groups of Trp & and more rarely His and the $N_\varepsilon H_\varepsilon$ of Arg (figure 2.13).

The pulse sequence for a basic HSQC experiment is illustrated in figure 2.14. The INEPT (Insensitive Nuclei Enhanced by Polarization Transfer) pulse sequence is used to transfer magnetization from the proton to the nitrogen atom. A 180º pulse mid evolution inverts the proton scalar coupling so that F_1 (nitrogen dimension) lineshape does not contain contributions from ^{1}H scalar coupling. A reverse INEPT sequence transfers the magnetization back to the proton and detection occurs. Decoupling the nitrogen channel removes the contribution of the nitrogen scalar coupling to the F_2 line shape. There is no diagonal in the HSQC spectrum because different nuclei are observed during T_1 and T_2

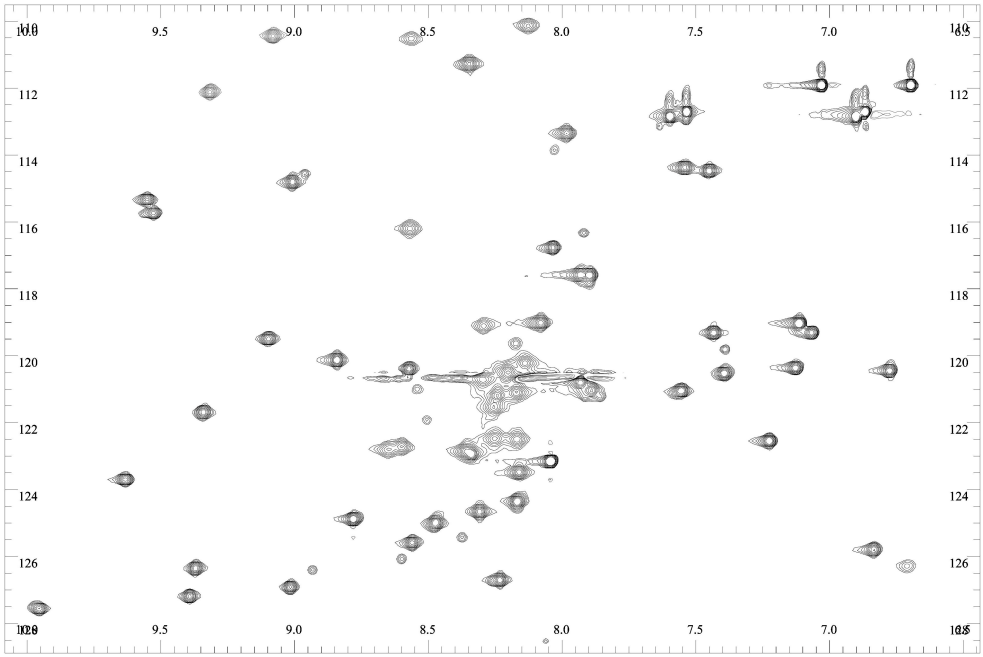


Figure 2.13 ^{15}N HSQC of hMBD1 CXXC1, (Bruker Avance 800 MHz with cryoprobe, 10mM deuterated tris(hydroxymethyl)aminomethane,, 250mM NaCl, pH 7.5, 293K).

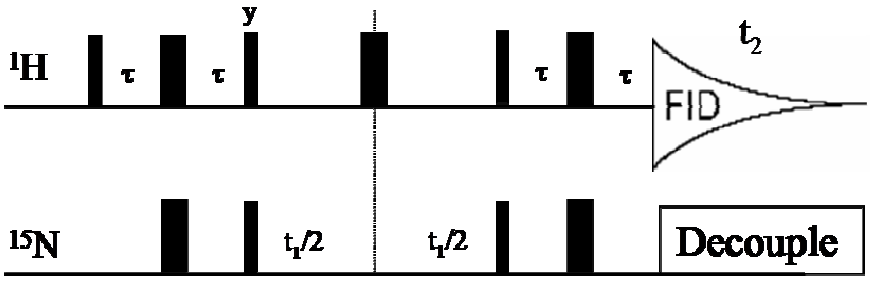


Figure 2.14 Pulse sequence for a basic HSQC experiment. Thin bars are 90º & thick bars are 180º pulses. Pulses are applied with x-phases unless indicated above the bar

Although it is not possible to assign cross peaks to specific residues using an HSQC alone, the spectrum provides a unique footprint for each protein. This footprint can change depending on buffer conditions and binding to ligands. This is due to a change in the nuclear shielding that the atoms are experiencing, which affects the chemical shift at which their resonances are observed at. This is a useful technique for investigating which residues, if assignments are available, interact with an added compound. Another important technique that requires use of the HSQC experiment is hydrogen-deuterium exchange for investigating solvent accessibility and hydrogen bonding. In solution, the amide hydrogen in the peptide bonds exchange with the solvent. When D_2O is used as a solvent in place of H_2O, the amide hydrogen exchange with deuterons. Since deuterons cannot be detected in a ^{15}N HSQC experiment those cross peaks that survive are considered to originate from slowly exchanging H_N which result from amide protons either involved in hydrogen bonds, or burred within a hydrophobic core.

2.3.2 3D ^{15}N heteronuclear NMR experiments

Ambiguity of sequential resonance assignment in overlapped 2D spectra can be generally overcome using 3D heteronuclear versions of 2D homonuclear experiments. These experiments combine a 2D homonuclear TOCSY or NOESY pulse sequence and an HSQC pulse sequence. For an HSQC-NOESY experiment the pulse sequence (Figure 2.15) starts with a homonuclear NOESY with a 90º pulse to switch magnetization to the transverse plane, a variable evolution time t_1 to provide chemical shift information in the F1 domain then another 90º pulse to move the magnetization back to the z-axis with magnetization transfer through dipolar coupling during the mixing time T_m. The remainder of the pulse sequence is identical to the HSQC experiment described above.

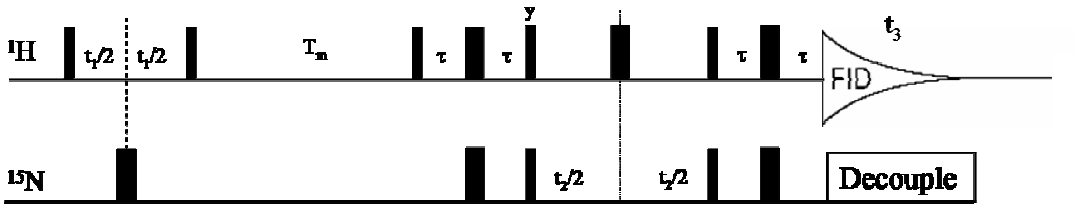


Figure 2.15 Pulse sequence for a 3D ^{1}H-^{15}N HSQC-NOESY experiment. Thin bars are 90º & thick bars are 180º pulses. Pulses are applied with x-phases unless indicated above the bar.

The advantage of introducing the 3rd dimension in a ^{15}N HSQC-NOESY or –TOCSY experiment is that since only cross peaks linked to an NH are seen in each slice, overlapped peaks can normally be separated by their ^{15}N chemical shift (figure 2.16).

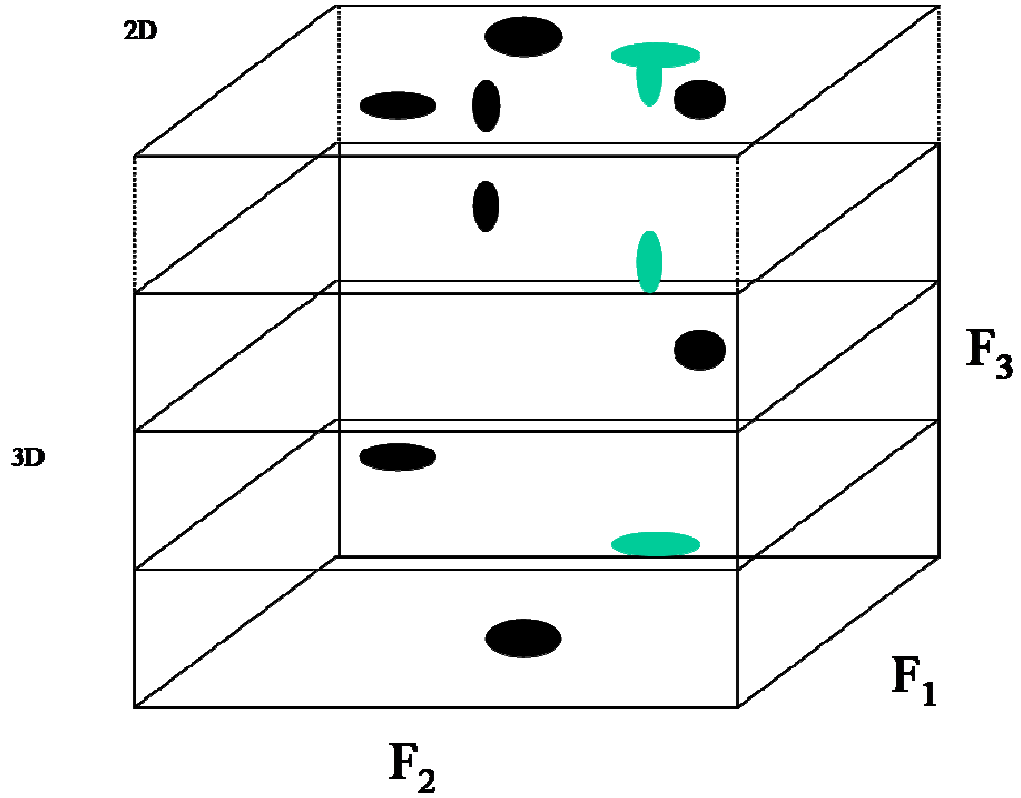


Figure 2.16 Illustration showing the relationship between a 2D homonuclear spectrum and a 3D heteronuclear spectrum. The overlapped cross peaks in the 2D spectrum (green) are separated by the introduction of the nitrogen dimension (F3).

2.4 Data Processing

Before resonance assignment can begin, the recorded NMR data has to be processed from the time domain to the frequency domain. Normally the direct Fourier transformation (DFT) of the recorded NMR signal does not yield an optimal spectrum. In order to maximize the information that can be extracted from the spectrum a number of digital signal processing techniques can be applied before and after Fourier transformation.

2.4.1 Pre-Fourier transformation data processing

A DFT of the FID recorded by the NMR spectrometer produces a frequency domain signal with a Lorentzian line shape. The shape of the line can be improved to give greater resolution and sensitivity by applying functions to the FID before DFT. The Lorentzian line shape is produced from an FID that decays exponentially. Therefore, if the decay of the FID is altered to decay with a different time dependence the resolution and/or single to noise can be adjusted. Multiplying the FID by exponential function will deemphasize the later parts of the FID where the signal from the protein is weakest. The Lorentzian-to-Gaussian is a common function applied to the FID. The Lorentzian lines are converted to Gaussian lines by multiplying by an increasing exponential followed by a Gaussian function. Figure 2.17 shows the effect of applying the Gaussian function to the FID.

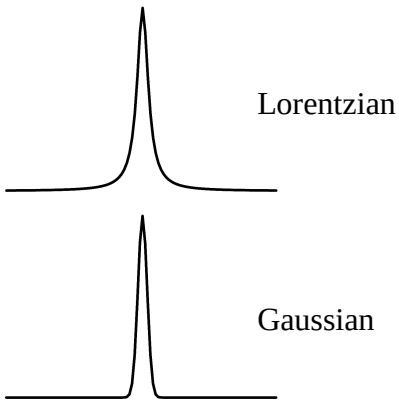


Figure 2.17 Diagram of the effect on line shape for a Lorentzian-to-Gaussion transformation. Line width at half height are the same. (Adapted from James Keeler lecture notes University of Cambridge, http://www-keeler.ch.cam.ac.uk/lectures/understanding/chapter_4.pdf)

A function similar to the Gaussian is the sine bell function. There is also a squared sine bell function which is more concentrated around the maximum. Figure 2.18A illustrates the difference between sine bell and squared sine bell and figure 2.18B illustrates the flexibility of the (squared) sine bell function. The squared sine bell function is good to use because it goes smoothly to zero which produces fewer truncation artifacts compared to the sine bell function.

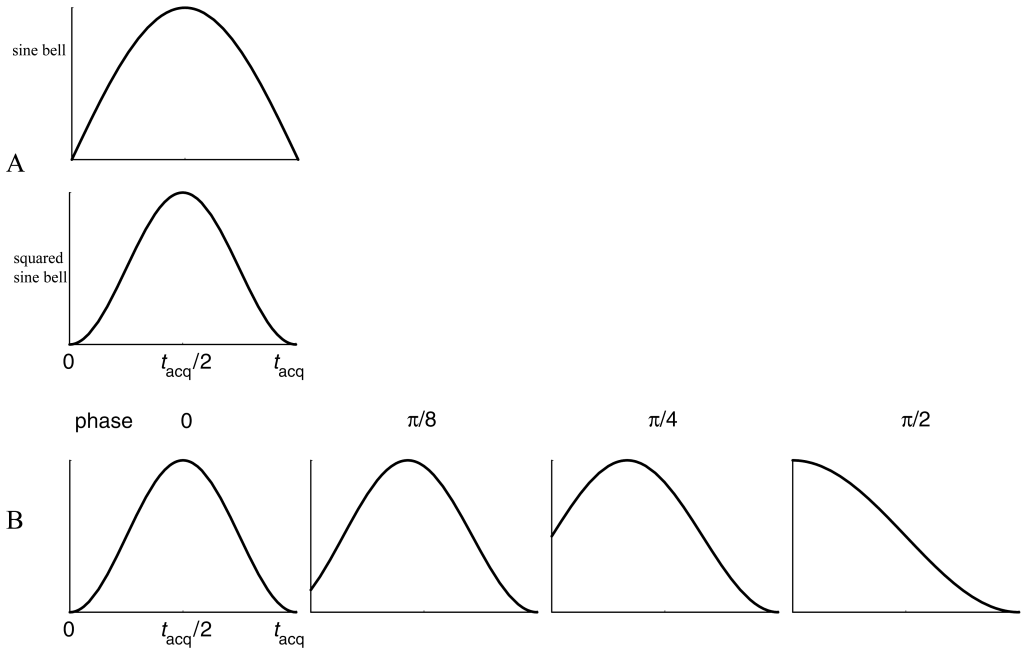


Figure 2.18 comparison of sine bell & squared sine bell curves (A) Flexibility of the squared sine bell function by applying a phase shift (B). (Adapted from James Keeler lecture notes, University of Cambridge, http://www-keeler.ch.cam.ac.uk/lectures/understanding/chapter_4.pdf)

For each NMR experiment a sweep width is set which is the range over which the data is collected (set in ppm or Hz). To obtain a high digital resolution it is necessary to record a large number of data points in this sweep width (Hz/data points). The noise in an FID remains more or less constant whereas the signal sinusoids decay over time. Therefore, at some point noise contributes more to the FID than the nuclear resonances. It is possible to record a smaller number of points and then extrapolate the FID by process called zero-filling if the signal has decayed by the end point. Zero-filling simply adds zeros to the data sequence. If the signal has not decayed when recording stops the FID is truncated and zero-filling leads to oscillations around the base of the peak called sinc-wiggles (the name comes from the peak shape as it is related to a sinc function). (figure 2.19).

These sinc-wiggles can cause problems during resonance assignment later on.

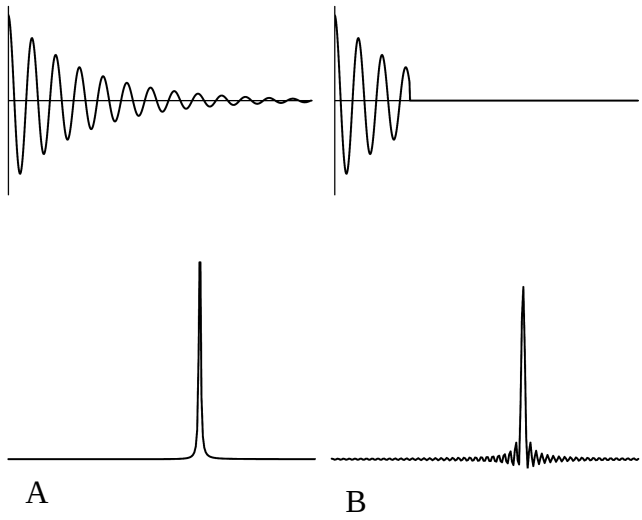


Figure 2.19 Figure illustrating the effects of truncation of the FID on the corresponding spectrum. For spectrum A the FID was allowed to decay to almost zero where as spectrum B shows sinc wiggle artifacts caused by truncation of the FID. (Adapted from James Keeler lecture notes, University of Cambridge, http://www-keeler.ch.cam.ac.uk/lectures/understanding/chapter_4.pdf)

2.4.2 Post-Fourier transformation data processing

After DFT the spectrum normally displays a phase error. This is because the time-domain signal is a complex signal as the sum of sine and cosine functions. The DFT of this complex time domain results in a frequency domain comprising of real part (absorption) and an imaginary part (dispersion), shown in figure 2.20, which are out of phase. This results in their full width of the peak at half maximum no longer being the same for each part.

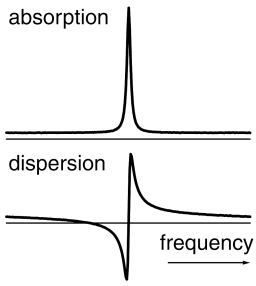


Figure 2.20 illustration showing the real (absorption) and imaginary part (dispersion) of a Lorentzian.

Applying zero-order (frequency independent) correction will correct phasing at the chosen pivot point. To correct remaining phase errors away from the pivot point, first order (frequency dependent) correction is applied. A curved or rolling baseline makes it more difficult to pick peaks in a spectrum and is also a source of error in quantification. Baseline distortions are mainly caused by the corruption of the first few data points of an FID. Baseline correction methods fall into two categories, time domain correction methods that reconstruct the corrupted data points in the FID and frequency domain correction methods that construct baseline curves to fit the spectra directly then subtract these baseline curves to remove the distortion. After adjusting the processing parameters to produce a well resolved spectrum, the data can be loaded into various programs for resonance assignment and production of NMR distance constraints (See chapter 4)

2.5 Alternative to DFT processing

Processing using Fourier transformation is a fast way of producing a frequency domain spectrum from a recorded FID. The DFT, as shown above, introduces oscillation artifacts when the FID is truncated. If strong window functions are applied to reduce these artifacts there is a reduction in resolution. In the case of 3D NMR experiments where the indirect detected dimensions are truncated due to time constraints, various methods have been developed to overcome this problem. For example, maximum entropy reconstruction (Sibisi *et al*, 1984), linear prediction (Barkhuijsen *et al*, 1985), maximum likelihood (Hoffman *et al*, 1989) and Multi-Dimensional Decomposition (Orekhov *et al*, 2003) . Processing of the 3D data in this thesis was carried out using the maximum entropy reconstruction method. The maximum entropy reconstruction method

creates a trial frequency domain spectrum then inverse FT back to an FID to compare with the FID recorded on the spectrometer. The software then adjusts the simulated frequency domain spectrum until the inverse FT matches the recorded FID. The inverse FT will never exactly match the recorded FID since the simulated frequency domain is devoid of noise. Noise can be introduced to the simulated frequency domain spectrum so that the inverse FT will match the recorded FID more closely. Using maximum entropy reconstruction on truncated data will normally produce a frequency domain spectrum with improved resolution compared to the Fourier transformation spectrum. However, maximum entropy reconstruction does not produce spectra that are as resolved and distortion free as the Fourier transformation of an untruncated FID.

2.6 important considerations for recording NMR spectra

Recording NMR spectra is not as simple as putting a protein sample in a tube and placing it in the NMR machine. Various characteristics of the protein have to be taken into account in order to prolong the life of the sample enough to record the required data. Not only are conditions that suit the protein important but also conditions that result in good spectra.

2.6.1 Temperature, pH and Salt

Protein samples should be kept at the lowest temperature possible in order to prolong the life of the sample. However, low temperatures are generally not suited to recording NMR spectra due to slower molecular tumbling, which leads to broader lines and lower sensitivity. However, high temperatures can also increase the rate of exchange with water and broaden the line widths due to increased dynamics and will ultimately cause the denaturation of the protein. Therefore a temperature must be chosen to give the best overall compromise of resolution and sample life. Figure 2.21 shows the effect of temperature on the xDNMT1 CXXC domain in 5K steps from 283 K to 303K. As we increase temperature, the resolution increases as lines narrow around the 8.75 to 9.5 ppm region. The spectrum at 293K appears to present the best overall compromise conditions for adequate resolution.

The optimum pH for protein NMR is pH 4.0-4.5 since amide proton exchange with water is slowest at this pH. As you move away from this range, amide exchange increases which can result in fewer amide protons being visible. This can be seen in the ^{15}N spectra of hMBD1 CXXC1 where NH of histidine 200 is not visible (p). However not all proteins are stable at low pH, or their catalytic activity may be altered due to

altered protonation states of the active site residues.

Sodium chloride is an important constituent of many protein chemistry buffers as it can prevent the protein from precipitating. This prevention of precipitation may be due to solvent counter-ions migrating towards the protein resulting in a layer which repulses the co-ions left in solution. However, as you increase NaCl concentration the conductivity of the solution also increases. The increase in conductivity not only affects the effective radio frequency power felt by the sample but also by the receiver coils of the spectrometer used to record the RF signal. Thus sensitivity is reduced. The reduction in effective RF power can be partially corrected by increasing pulse width, but this results in pulses that are further from ideal with various consequences for the quality of the resulting spectrum.

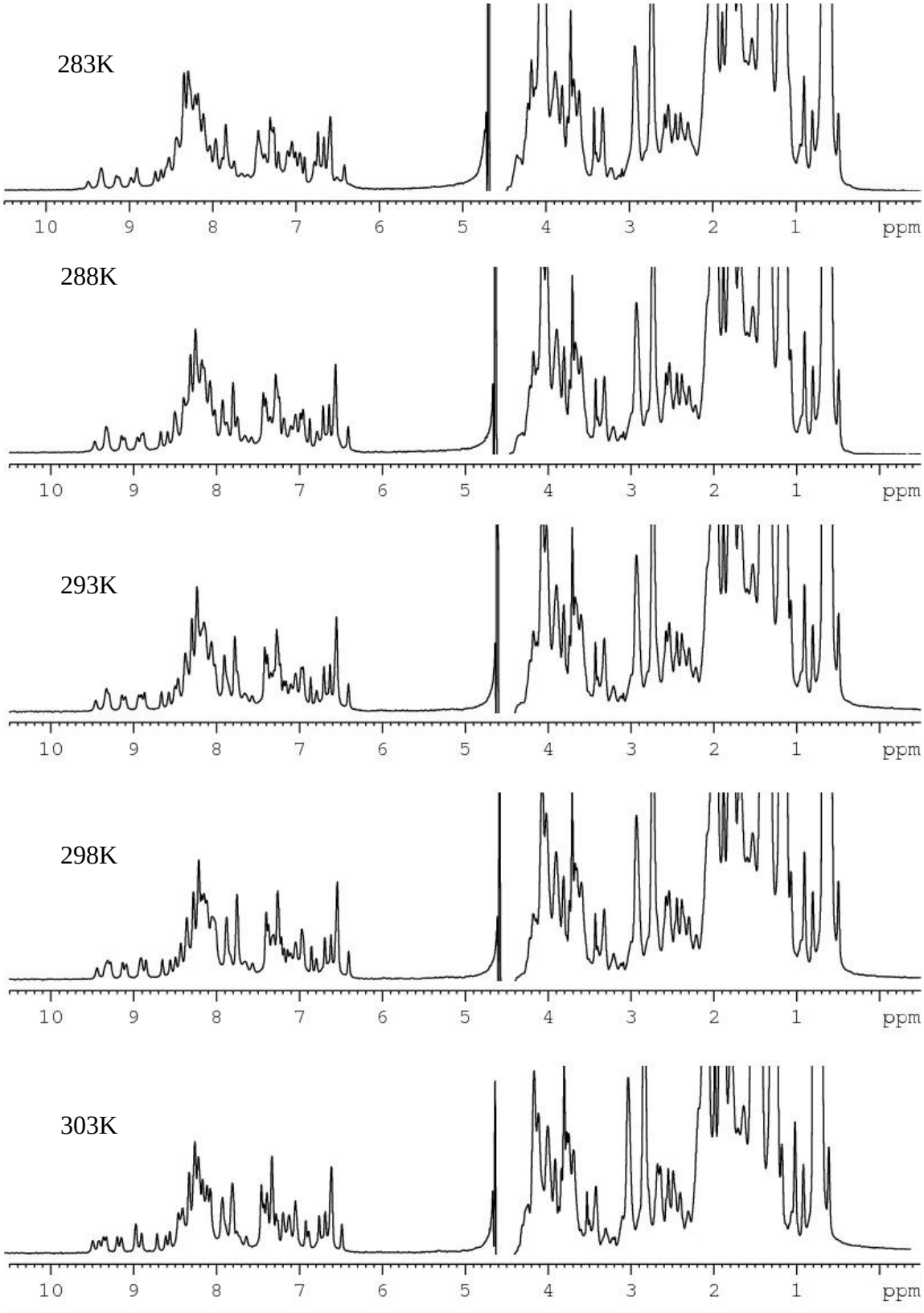


Figure 2.21 1D ^{1}H spectra of unlabelled xDNMT1 CXXCb (10mM deuterated tris (hydroxymethyl)aminomethane, 250mM NaCl, pH 7.0) at various temperatures. Water suppression using the method of Hwang and Shaka (1995)

2.6.2 Protein size

In solution NMR, amide ^{1}H from any unstructured regions of a protein will appear in the 7-8.5 ppm region of a 1D ^{1}H spectrum, resulting in a decrease in peak dispersion and increase in the overlap of the cross peaks in the spectra. Shortening of the domain at the N- and C- termini can have a pronounced effect on the spectra recorded since the termini of protein domains are often unstructured. In the case of xDNMT1 CXXC domain, shortening the construct from both termini had a dramatic effect on the resolution of the recorded spectra. Figure 2.22 shows the differences between the three constructs

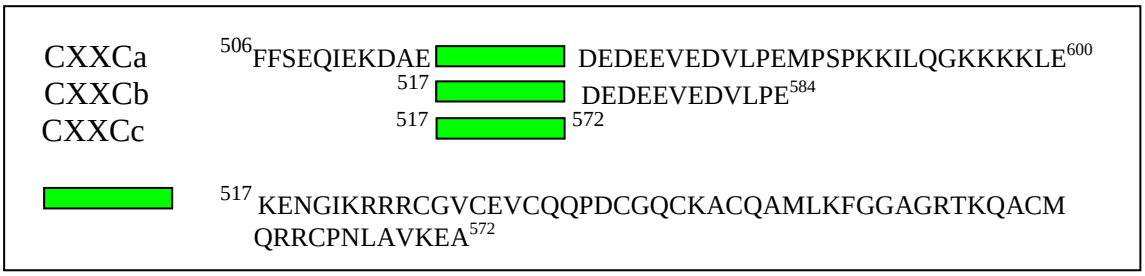


Figure 2.22 Schematic showing the amino acid sequences of the three xDNMT1 CXXC constructs. The green box represents xDNMT1 CXXCc whose sequence is shown below

NMR spectra for the three xDNMT1 CXXC constructs shown above are shown in figure 2.23. Both CXXCb and CXXCc show less peak overlap in the 7-8.5 ppm region when compared to CXXCa, with the 8.5 – 9.5 ppm region becoming better resolved as the size of the construct becomes smaller. The peak around 0.6 ppm also becomes more intense relative to other peaks in the aliphatic region as the domain is shortened (this peak disappears when the protein is unstructured e.g. upon addition of EDTA to the sample).

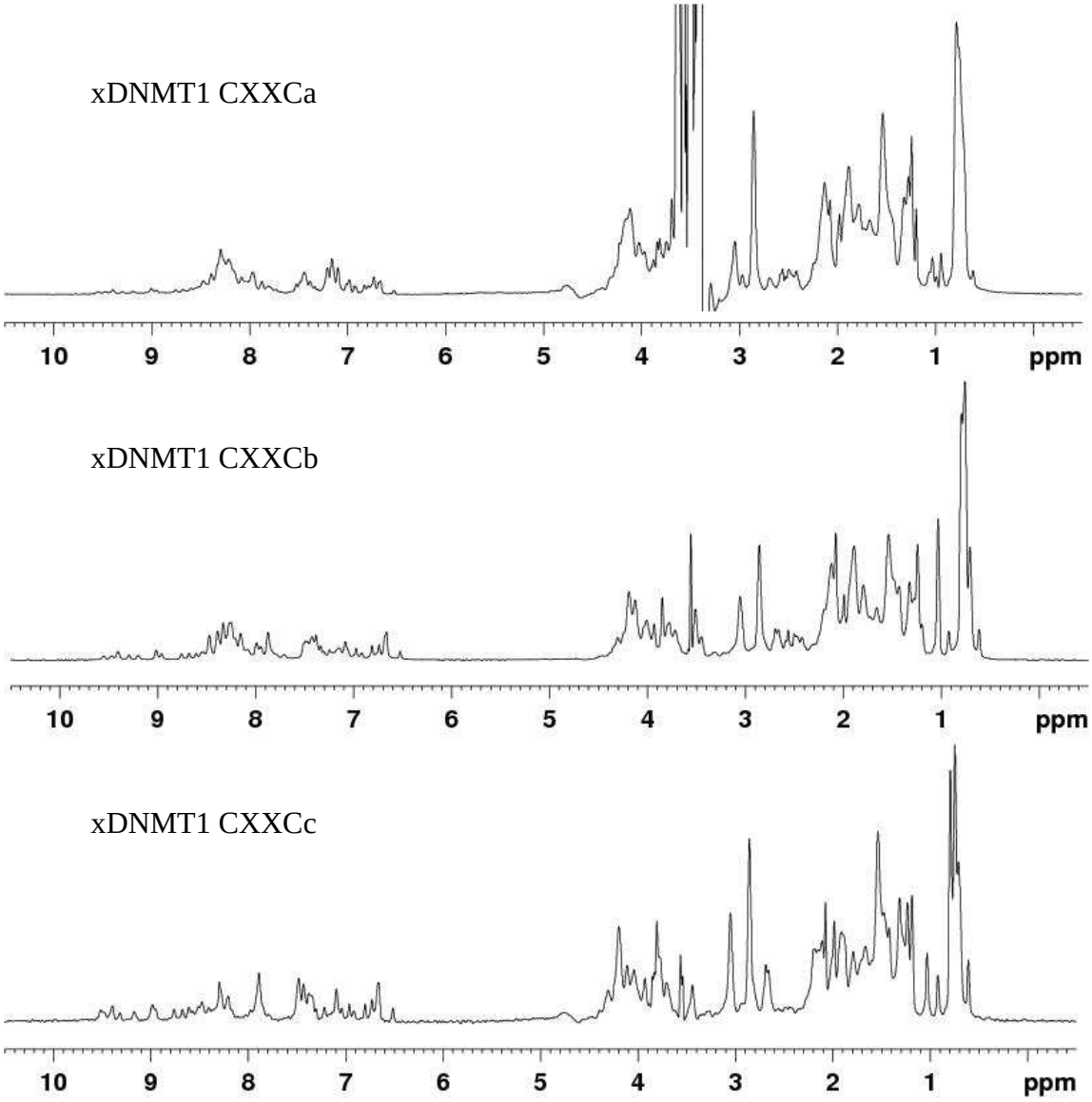


Figure 2.23 1D ^{1}H spectra of unlabelled various xDNMT1 CXXC contructs (10mM deuterated tris (hydroxymethyl) aminomethane, 250mM NaCl, pH 7.0) Water suppression using the method of Hwang and Shaka (1995).

CHAPTER THREE

SAMPLE PREPARATION & NMR SPECTROSCOPY of hMBD1 CXXC1

3.1 Cloning of hMBD1 CXXC1 Domain

This section describes the strategy used to design hMBD1 CXXC1 constructs, to express and purify the hMBD1 CXXC1. The NMR experiments carried out are also described.

3.1.1 Bacterial transformations

All bacteria used were chemically competent. Heat shock was therefore used to transform the DNA into the cells. The bacteria were removed from –80˚C and immediately placed on ice to defrost before 1 µl (mini prep DNA) or 1.5 µl (DNA ligation) was added and gently stirred with the pipette tip. The bacteria were incubated on ice for 2 minutes before being heat shocked at 42˚C for 30 seconds and then placed back on ice for a further 5 minutes. 300 µl 37˚C pre-warmed SOC media was added before the bacteria were incubated at 37˚C & 200 rpm for 1 hour. For plasmid transformations 50, or for ligations 350 µl of culture was spread onto pre-dried LB-agar containing the required selection antibiotics. The culture was allowed to dry into the LB-agar (approximately 15 minutes) before being placed in a 37˚C incubator overnight. Single colonies could then be picked the next day for inoculation into LB for plasmid purification or protein expression.

3.1.2 Agarose Gel Electrophoresis

1 gram of multiple purpose agarose (Roche, Cat No. 1388991) was put into a 250 ml conical flask containing 100 ml TBE buffer. The agarose was dissolved by heating in a 900 W microwave for approx 90 s. Ethidium bromide was added to the agarose to a final concentration of 10 µg/ml once it had cooled to approximately 50˚C. It was then poured into the casting tray of the electrophoresis apparatus (BioRad, Cat No. 164-0310) containing a sample comb and allowed to solidify at room temperature. Once the gel had set the comb was removed carefully and TBE buffer added to the tank in order to cover the gel by approx 5-6 mm. DNA samples were mixed with 6x loading buffer (Promega, Cat No. G1881) then pipetted into the wells. NEB 2-log ladder (Cat. No. N3200S) was added for size referencing before electrophoresis was carried out under conditions of constant voltage at 100V for as long as it was required to separate each specific sample (30-60 minutes). Migration of the DNA was detected by fluorescence of ethidium bromide bound to DNA under UV illumination

3.1.3 Considerations when choosing domain size to express

When studying a protein in solution, non-structured areas of a protein are more flexible

than the structured domains. The flexible unstructured regions will tend to have no long range NOEs and when present at the N & C termini of a protein will not contribute any information to the overall structure of the protein. Flexible N- & C- termini will also give rise to additional poorly resolved cross peaks in the NMR spectra making assignment more difficult. Thus, it is important to express a domain with minimal N- & C-terminal overhangs if possible.

3.1.4 PCR of hMBD1 CXXC1 Domain

Using the sequence identified by Cross *et al* (1997) a BLAST search revealed an IMAGE clone that contained the cDNA required to amplify the hMBD1 CXXC1 domain. The clone containing the DNA sequence of interest, GENBANK accession number CF552871, was purchased from Geneservice Ltd (clone ID 30529682). The cDNA had been cloned into pCMV SPORT (ampicillin^{R}) which was transformed into chemically competent DH5α *E.coli* cells (Invitrogen, Cat No.18265-017). A stock of plasmid was prepared by inoculating a single colony into 10 ml of Lauria Broth (LB) + 100 µg LB and grown overnight at 37ºC in an orbital incubator at 200 rpm. Plasmids were purified by alkali-lysis using a GenElute^{TM} plasmid mini prep kit (Promega, Cat No PLN-350). PCR primers were designed based on the solution structure of the MLL CXXC domain (Allen *et al*, 2006) in order to minimize flexible non-structured N- & C-termini. Allen *et al* assigned and calculated the structure for residues V1146 to K1214 of the MLL CXXC domain but found that only residues R1150-P1201 adopted a well-defined tertiary structure. Based on their structure primers were designed to amplify the region from MBD1 encoding amino acids 166-222, to give the least amount of flexible N- & C- termini (figure 3.1). The gene fragment was cloned into a pGEX-6P1 (GE healthcare, Cat No. 27-4597-01) expression vector containing an N-terminal PreScission protease site (Figure 3.2). PreScission protease is a fusion protein of glutathione S-transferase and human rhinovirus type 14 3C protease (Habig *et al*, 1974) allowing a one step cleavage/protein elution step when using a GST-resin for purification.

```
                 167                                                  222
MBD1 CXXC1         EQRMFKRVGCGECAACQVTEDCGACSTCL--LQLPHDVASGLFCKCERRRCLRIVERS
MLL CXXC     PQEPPVKKGRRSRRCGQCPGCQVPEDCGVCTNCLDKPKFGGRNIKKQCCKMRKCQNLQWMPSKAYLQKQAKAVK
             1146                                                                1214
```

Figure 3.1 Amino acid alignment of CXXC domains from MLL and hMBD1 comparing the sequence used by Allen et al and the domain cloned for this study

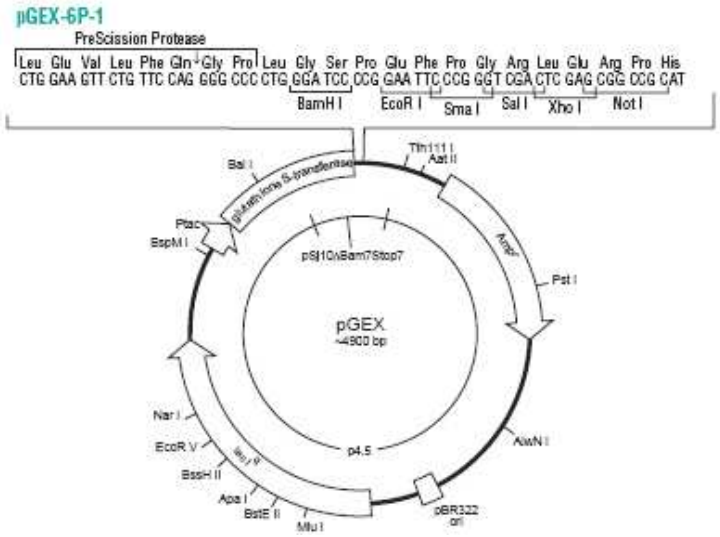


Figure 3.2 Plasmid map and multiple cloning site of pGEX-6P1 showing the PreScission protease cleavage site (GST handbook, GE healthcare).

The PCR primers (figure 3.3) included restriction sites to clone the PCR product into pGEX-6P1 so that the CXXC domain would be expressed in frame with the GST tag.

hMBD1 CXXC1 fwd 5' G**GGATCC**GAGCAGAGAATGTTTAAG 3'
hMBD1 CXXC1 rev 5' **CTCGAG**<u>TCA</u>GCTCCTTTCCACAATC 3'

Figure 3.3 Primers for the PCR amplification of hMBD1 CXXC1 domain. The restriction sites used (fwd BamH1, rev Xho 1) are shown in bold with the stop codon underlined

PCR primers were ordered from Sigma Genosys and diluted to a stock concentration of 100 µM. Amplification was carried out on a Techne TC-512 thermo cycler using PfuTurbo® DNA polymerase (Stratagene, Cat No. 600250) with the IMAGE clone 30529682 as a template. The DNA template was diluted 1:100 from a plasmid mini prep stock. PCR reaction mix was as follows:

1 µl (0.5µg) DNA template
1 µl of each primer (50µM stock)
1 µl of dNTP mix (NEB, Cat No N0447S)
0.5 µl (1.25 units) Pfu Turbo DNA polymerase
5 µl 5x Pfu Turbo DNA polymerase buffer
41.5 µl ultrapure water

As PfuTurbo is not a hotstart enzyme, it was added to the reaction after the initial 5 minutes at 95ºC followed by

1min 95ºC
1min 60ºC } 35 cycles
1min 72ºC

With a final extension of 10 mins at 72ºC producing a product 183bp long.

To purify the DNA, the sample was subjected to agarose gel electrophoresis. The PCR reaction was mixed with loading buffer (Promega, Cat No. G1881) and run on a 1% agarose gel (Roche, CAT No 1388991) prestained with 5µl ethidium bromide (stock 10mg/ml) alongside 2-log DNA marker (NEB, Cat No N3200S) (fig 3.4). The DNA was visualized by detection of ethidium bromide fluorescence upon exposure to ultra violet light and excised from the gel. The DNA was purified from the gel slice using a QIAquick gel extraction kit (Qiagen, Cat No 28704) according to the manufacturer's instructions and eluted in 30µl.

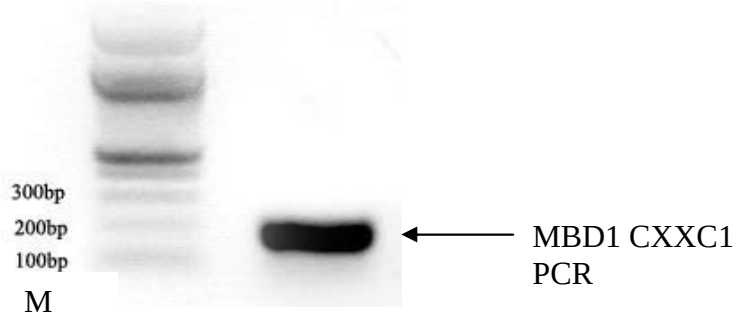


Figure 3.4 Agarose gel electrophoresis negative image of the MBD1 CXXC1 PCR product. M - 2-log DNA marker

3.1.5 Cloning of PCR product into pGEM-T easy

In order to clone the cDNA fragment into the pGEM-T easy vector (Promega, Cat No 1360) adenine overhangs (A-tail) had to be created since Pfu Turbo is a proof reading enzyme and removes any overhangs. Taq polymerase (NEB, Cat No M0273S) was used in the following reaction to add the A-tail to the PCR fragment. The reaction was incubated for 15 mins at 72ºC.

1 µl (5 units) Taq DNA polymerase
1 µl dATP 4mM stock (Promega, Cat No. U1201)
1 µl 10x Taq polymerase buffer
7 µl gel purified DNA

The A-tailed DNA was then cloned into pGEM-T easy according to the manufacturer's instructions and transformed into chemically competent DH5α cells. Positive clones were identified using blue/white selection (see appendix), inoculated into 10 ml LB + 100 µg/ml ampicillin and grown overnight at 37ºC in an orbital incubator at 200 rpm. Plasmids were purified by alkali-lysis as described above. Cloning into pGEM-T easy allowed the digestion of the construct with BamH1 (NEB, Cat No R0136S) & Xho1 (Promega, Cat No R6161) to release the MBD1 CXXC1 fragment with sticky ends enabling ligation into pGEX-6P1. pGEX-6P1 was digested with the same enzymes; the fragments were then separated by agarose gel electrophoresis and then excised from the gel (Fig. 3.5). The restriction digest which included the following

34µl DNA from standard mini prep elution
4µl NEB BamH1 buffer
1µl BSA (10mg/ml stock concentration)
1µl of BamH1 (20 units) & Xho1 (10 units) enzymes

was incubated for 90 minutes at 37ºC

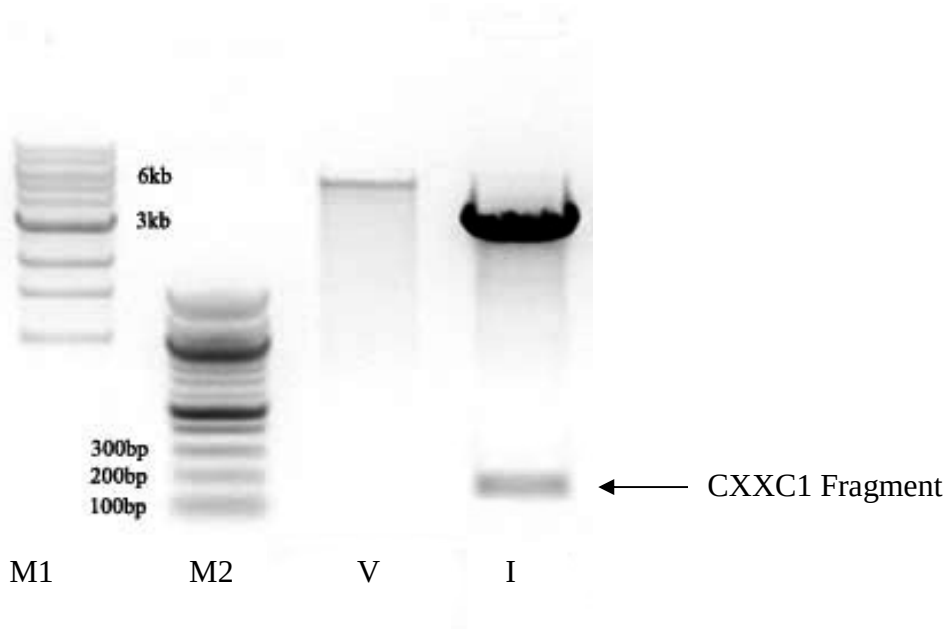
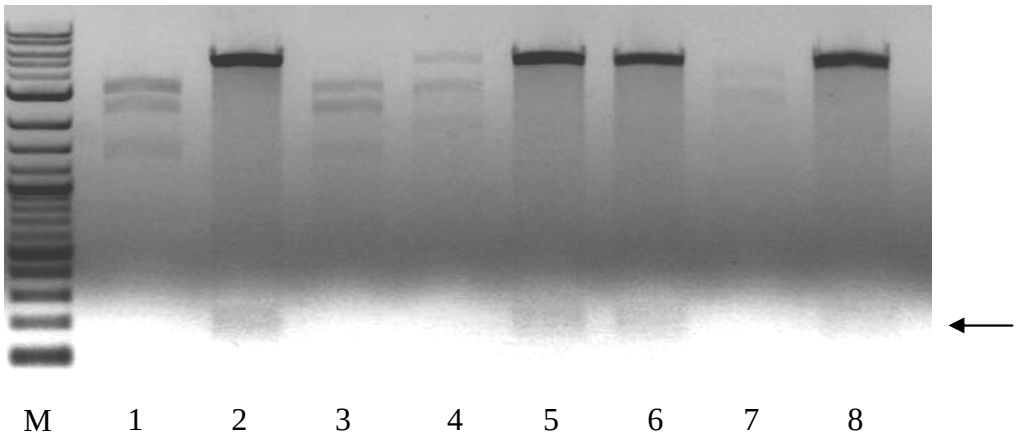


Figure 3.5 Agarose gel electrophoresis of the digested pGEX-6P1 vector and pGEM-T CXXC1 construct. M1 1KB ladder NEB, M2 100bp ladder NEB V – pGEX-6P1, I – pGEM-T CXXC1

3.1.6 Cloning of MBD1 CXXC1 into pGEX-6P1

The MBD1 CXXC1 insert and pGEX-6P1 were ligated using NEB T4 Ligase (Cat No M202S) according to the manufacturer's instructions, transformed into DH5α cells and spread onto LB agar plates containing 100 µg/ml ampicillin and grown overnight at 37ºC. Colonies were inoculated into 10 ml LB + 100 µg/ml ampicillin and grown overnight at 37ºC in an orbital incubator at 200rpm. Plasmids were purified from the bacteria using alkali-lysis as described above and digested with restriction enzymes BamH1 & Xho1 to identify clones containing an insert of the expected size (fig 3.6). Correct clones were sent to Cogenics for sequencing. Clones with the correct sequence were used for protein expression



M 1 2 3 4 5 6 7 8

Figure 3.6 Agarose gel electrophoresis of a selection of 8 putative pGEX6P1-hMBD1CXXC1 clones digested with BamH1 and Xho1. The arrow indicates the expected position of the CXXC1 fragment. Clones 2, 5, 6 & 8 show the expected fragment pattern. M- NEB 2-log ladder.

3.2 Protein Sample preparation

The pGEX-6P1 system utilizes a GST (Glutathione S-Transferase) tag for purification of the target protein on a Glutathione sepharose^{TM} 4 fastflow resin (GE healthcare, Cat No. 17-5132-01). Cleavage of the GST tag from the target protein was achieved using PreScission Protease (GE Healthcare, Cat No. 27-0843-01).

3.2.1 Expression of the hMBD1 CXXC1 domain

Based on previous experience with the CXXC domain from xDNMT1 and the publication of the MLL CXXC structure (Allen *et al*, 2006) pGEX-6P1 CXXC1 was transformed into chemically competent Tuner^{TM} DE3 cells genotype F^- *ompT hsdS*$_B$ (r_B^- m_B^-) *gal dcm lacY1* (Novagen, Cat No 70623-4), spread onto LB plates containing 100 µg/ml ampicillin and grown overnight at 37ºC. Single colonies were inoculated into 50 ml of LB containing 100 µg/ml ampicillin in a 250 ml conical flask then incubated overnight.

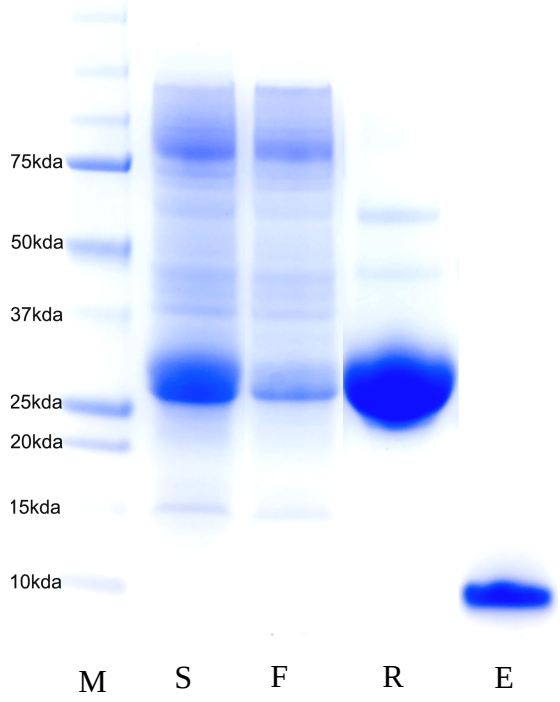
The 50 ml cultures were pelleted by centrifugation at 4300 g for 5 mins at 5000 rpm. The supernatant was removed and the pellet resuspended in 5 ml LB which was used to inoculate 500 ml of LB containing 100 µg/ml ampicillin in a 2 L conical flask. The bacteria were grown at 37ºC in an orbital shaker, 200 rpm, to an OD_{600} between 0.6-0.8 before being cooled to 4ºC. Protein expression was induced by the addition of 150 µl of IPTG (1 M stock) to give a final concentration of 300 µM. The bacteria were placed back in the orbital incubator to express overnight at 13ºC. For isotopically labelled

protein, the bacteria were grown in 500 ml of M9 minimal media rather than LB.

3.2.2 Purification of hMBD1 CXXC1 domain

A typical preparation of 4-5 mg of protein required 2 L of bacteria grown in 4x500 ml lots. The bacteria were pelleted by centrifugation for 10 minutes at 4300 g. The supernatant was removed and the bacteria resuspended in phosphate buffered saline solution (PBS) pH 7.3 (10 ml per 1 L of culture) then frozen at -20ºC. To lyse the cells, they were first defrosted and warmed to room temperature (RT) before 2 ml of 10x BugBusterTM (Novagen, Cat No 70921-5) and 125 units of benzonase® nuclease (Novagen, Cat No 70746-3) were added. The sample was incubated on a roller mixer for 20 mins at room temperature (RT). Soluble and insoluble material was separated by centrifugation for 25 mins at 19872 g. An empty econo-pac column (Bio-Rad, Cat No 732-1010) was filled with a 3 ml bed volume of FastFlowtm 4 glutathione sepharose and equilibrated by washing the resin with 5 bed volumes of PBS pH 7.3. The supernatant was passed through the resin twice which was then washed with 5 bed volumes of PBS pH 7.3. To remove the tag, 80 units of dialysed PreScission was mixed with 3 ml of PreScission protease buffer minus EDTA. To remove EDTA from the PreScission protease it was dialysed against 1L 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM DTT using a Thermo Scientific Slide-A-Lyzer Dialysis Cassette 7000MWCO (Cat. No. 66373). Glycerol was added to the dialysed sample to a final concentration of 20%. The PreScission protease was aliquoted and stored at –20˚CThis was then added to the column and the resin gently resuspended before being incubated at 4ºC overnight. The cleaved MBD1 CXXC1 domain was eluted from the column by applying 20 ml of elution buffer (see appendix). Samples of the soluble lysis fraction, the flow through, the resin and the elution were analysed by SDS-PAGE (Figure 3.7). The samples were mixed with 4X NuPAGE LDS sample buffer (Invitrogen, Cat No. NP0007) prior to being loaded into pre-cast NuPAGE 4~12% bis-tris gels (Invitrogen, Cat No. NP0321BOX). The samples were heated to 85 ^{0}C for 10 min with 90 mM β-Mercaptoethanol before loading. Protein markers (BioRad, Cat No. 161-0373) were diluted with two volumes of water and one volume of NuPAGE LDS sample buffer and 10 µl was loaded. The gel was run at 200 V constant voltage for 35 minutes. The protein bands were visualized after staining the gel with approximately 100 ml of coomassie stain for 5 to15 min followed by destaining in 100 ml of destaining buffer overnight at room temperature

Using the conditions described above for expression and purification of the protein resulted in an approximate yield of 4-5 mg of purified soluble protein when measured by Bradford assay which could be concentrated and buffer exchanged (chapter 3.2.3).



M – Biorad prestained molecular weight marker
S – 20 µl soluble lysis fraction
F – 20 µl flow through fraction
R – 20 µl of resin after washing
E – 10 µl of eluted MBD1 CXXC1 protein

Figure 3.7 SDS-PAGE analysis of the expression and purification of MBD1 CXXC1 domain expressed in Tuner^{TM} DE3 cells. The image has been edited to place lanes F, R and E together since they were spaced out on the original gel.

3.2.3 Sample concentration and buffer exchange

For NMR spectra to be recorded within a sensible time, the protein concentration within the NMR sample must be at least 0.5 mM. Since experiments can take place at temperatures which make the protein more liable to go off (precipitation, protease digestion, unfolding) the shorter the experimental time taken the better. More concentrated samples produce better signal to noise enabling this. Using a vivaspin20 5kDa molecular weight cut off spin concentrator (Sartorius Stedim, Cat No VS2012) it is possible to both buffer exchange and concentrate the sample to the appropriate volume (500 µl). The concentrator was washed with water to remove the glycerol preservative before addition of the eluted protein. The sample was concentrated to 500 µl by centrifugation at 3056 g and 10ºC. NMR buffer was added slowly to a total volume of 20 ml and the process repeated twice achieving a 1600 fold dilution of the original buffer. The purification of MBD1 CXXC1 resulted in NMR samples of between 1-1.25mM

The concentrated sample was centrifuged at 13000 rpm for 5mins in a bench top microcentrifuge to remove any particles before 30 µl of D_2O was added. D_2O provides a deuterium lock signal for the NMR spectrometer. The sample was made up to 600 µl with ultrapure water before being gently pipetted into an NMR tube (Wilmad 535-pp or Novell S607)

3.3 NMR spectroscopy

Data was recorded on Bruker AVANCE 600 MHz & 800MHz NMR spectrometers, equipped with 5 mm TCI cryoprobes. NMR experiments were carried out using 1 mM protein samples except for DNA titration (100 µM). The experiments carried out are shown in table 3.1. 1D and 2D NOESY, TOSCY experiments unlabelled protein was used. For COSY (and another 2D NOESY) experiment, NMR buffer was made with D_20 in place H_20. 3D TOCSY-HSQC and NOESY-HSQC, ^{15}N labelled protein was expressed using M9 minimal media supplemented with ^{15}N labelled ammonium chloride. All experiments were carried out at a temperature of 293 K.

3.3.1 DNA Titration of hMBD1 CXXC1

A Palindromic DNA oligonucleotide of 12 bp length (12mer -GCTTA**CG**TAAGC) was purchased from Sigma Genosys (reverse phase HPLC purified). The oligo was resuspended in nuclease free water (promega) to a concentration of 84 uM per ml. Annealing of the oligo was performed by heating the sample to 95˚C then allowing it to cool slowly to room temperature. The 12mer was then aliquoted into 600 µl (50 µM) samples before being lyophilized. This allowed the DNA to be mixed with the NMR sample without decreasing the concentration of the NMR sample. An HSQC was carried out at protein:DNA concentration ratios 1:0, 1:0.5, 1:1, 1:3 to investigate the DNA binding ability of the hMBD1 CXXC domain

	Dimension 1			Dimension 2			Dimension 3					
Experiment	Nuc	TD	SW	Nuc	TD	SW	Nuc	TD	SW	RD ms	NS	WS
100ms NOSEY	^{1}H	2048	8503	^{1}H	2048	7001				1.0	32	ES
200ms NOSEY	^{1}H	2048	8503	^{1}H	2048	7001				1.0	32	ES
100ms NOSEY*	^{1}H	2048	9615	^{1}H	1024	9596				1.0	32	ES
60 ms D2O NOSEY	^{1}H	2048	8503	^{1}H	2048	7000				1.0	16	PS
60 ms TOCSY	^{1}H	2048	8503	^{1}H	2048	7001				1.0	32	ES
COSY	^{1}H	2048	8503	^{1}H	2048	7001				1.0	32	WG
D2O COSY	^{1}H	2048	8503	^{1}H	2048	6999				1.0	24	
HSQC*	^{1}H	2048	12820	^{15}N	128	1702				1.0	8	WG
HSQC + DNA	^{1}H	2048	9615	^{15}N	128	1277				1.0	32	WG
HSQC-NOSEY*	^{1}H	1024	11160	^{1}H	512	11160	^{15}N	128	1702	1.0	8	WG
HSQC-TOCSY*	^{1}H	1024	11160	^{1}H	256	11160	^{15}N	128	1702	1.0	8	WG

Table 3.1 Acquisition parameters of NMR experiments (Bruker AVANCE 600 MHz unless * then Bruker AVANCE 800 MHz) used for resonance assignment and structural restraints collection. Nuc: Nucleus, TD: time domain points (complex), SW: sweep width in Hz, RD: relaxation delay, NS: number of scans, WS: water suppression, WG: water gate, ES: excitation sculpting, PS: presaturation, ms: milliseconds,

CHAPTER FOUR

CHEMICAL SHIFT ASSIGNMENT AND STRUCTURE CALCULATION OF hMBD1 CXXC1

4.1 NMR protein structure calculation

This chapter shows how the chemical shifts of the resonances of hMBD1 CXXC were assigned using CcpNMR analysis software (Vranken *et al*, 2005), how distance restraints were derived from NOESY spectra and how ARIA (Ambiguous Restraints for Iterative Assignment) was used for structure calculations.

4.2 Is the protein is structured?

NMR experiments used to generate spectra for chemical shift assignment and to calculate NOE distance restraints takes significant time to complete. Before such experiments are undertaken the sample must be checked to evaluate whether it is structured. For an unlabeled sample, the observation of particular features in a 1D ^{1}H spectrum is normally sufficient to suggest that a sample is structured. The 1D ^{1}H spectrum of unlabeled hMBD1 CXXC1 (figure 4.1) suggests that the protein sample is folded due to the overall dispersion of chemical shifts in the methyl (0-2ppm) and amide (8.5-10ppm) regions and the ^{1}H line widths. The 1D ^{1}H experiment is relatively short, typically 8 seconds for a 0.5mM hMBD1 CXXC1 sample, therefore it is ideal to run, to be sure the sample has not unfolded or precipitated, between the longer 2D TOCSY, NOESY, COSY experiments.

For larger proteins, >10kDa, the 1D ^{1}H NMR experiment is not ideal due to potential overlap of chemical shifts and broader line widths. In these situations isotopically labeled proteins are required due to increasing overlap in homonuclear 2-D spectra.

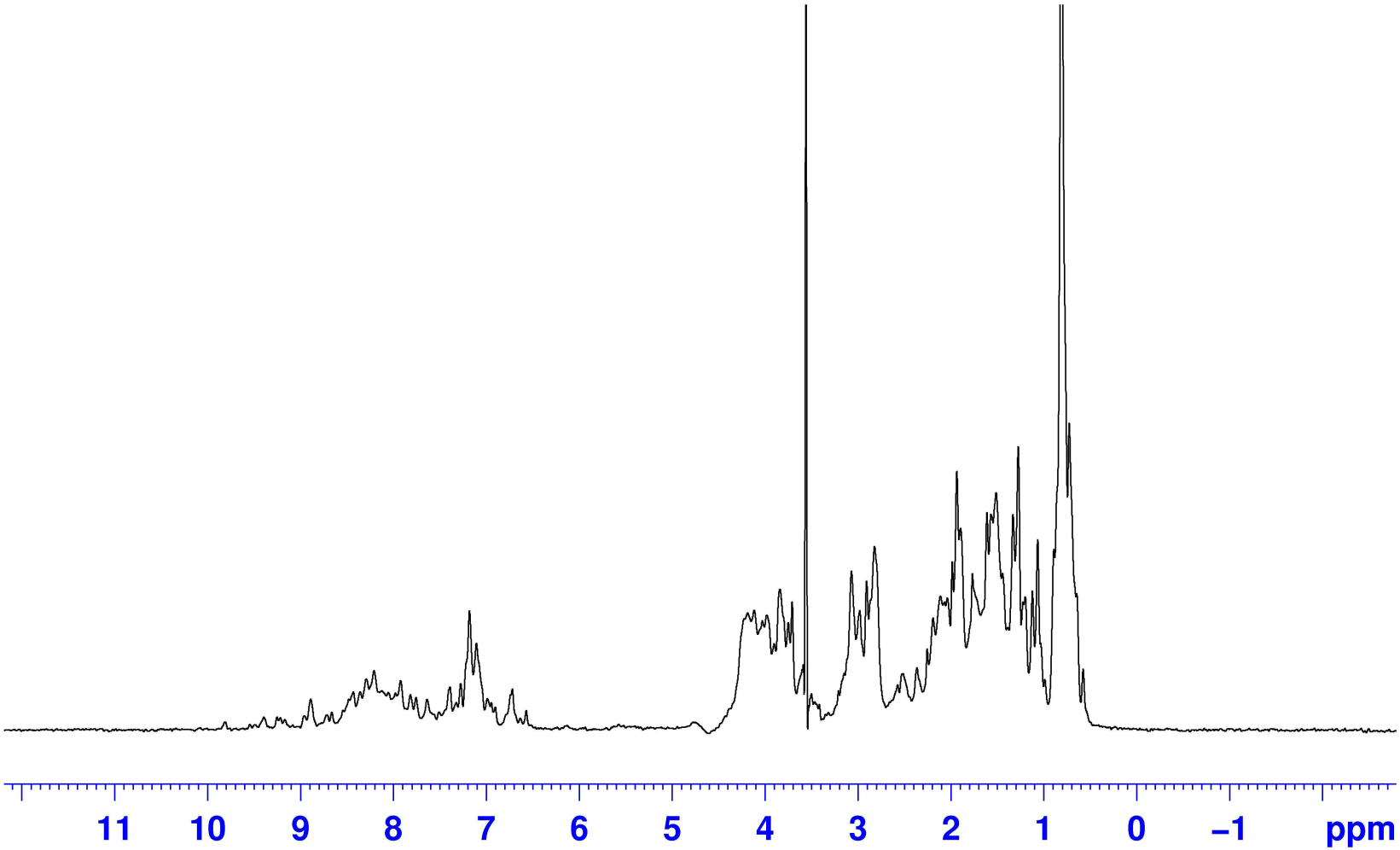


Figure 4.1 1D ^{1}H spectrum of unlabeled hMBD1 CXXC1 (10mM dTris, 250mM NaCl, pH 7.5, 293K). Water suppression using the method of Hwang and Shaka (1995).

4.3 Sequence specific resonance assignment of hMBD1 CXXC1

The relatively small size of the CXXC domain lends itself to structural determination though homonuclear NMR, thus not incurring the high cost associated with ^{13}C & ^{15}N labeling or problems of achieving sufficient quantity of protein from growth in M9 minimal media. Sequence specific resonance assignment of hMBD1 CXXC1 was attempted using 2D NOESY & TOCSY spectra according to the method described by Wutrich *et al*, and although significant stretches of amino acids could be assigned, crucially the remaining resonances could not be assigned due to features of the spectrum as set out below.

4.3.1 Overlapping regions make sequential assignment difficult

The TOCSY and NOESY spectra for hMBD1 CXXC1 are mainly well dispersed except for a part for the Hα-H_N correlation region where there is significant overlap (figure 4.2). This region, centered around 4.3, 8.3 ppm proved problematic for sequential assignment of the backbone H_N to Hα atoms. An example of the difficultly in assigning crosspeaks in overlapped spectra is shown in figure 4.3. Figure 4.3 shows the TOCSY Hα-H_N and aliphatic-H_N crosspeaks for arginines 169 & 173. Arginines 169 & 173 share the same H_N chemical shift which makes it difficult to distinguish which cross peak belongs to which amino acid. It was also not possible to assign sidechain cross peaks to either of these arginines.

Not only did it prove difficult to assign certain amino acids using only 2D spectra but it was also difficult to sequentially assign spin systems. As previously described in chapter two (p49) side chains have characteristic crosspeak pattern and chemical shift patterns which allow spin systems to be assigned to an amino acid type. This allows spins systems to be labelled as e.g. a cysteine or glycine without assigning it in sequence. For example, the amino acid sequence of hMBD1 CXXC1 contains only one histidine. Therefore it was possible to identify which crosspeaks belonged to histidine 200. However, using only 2D homonuclear NOESY & TOCSY spectra is was not possible to assign the adjacent residues 199 and 201.

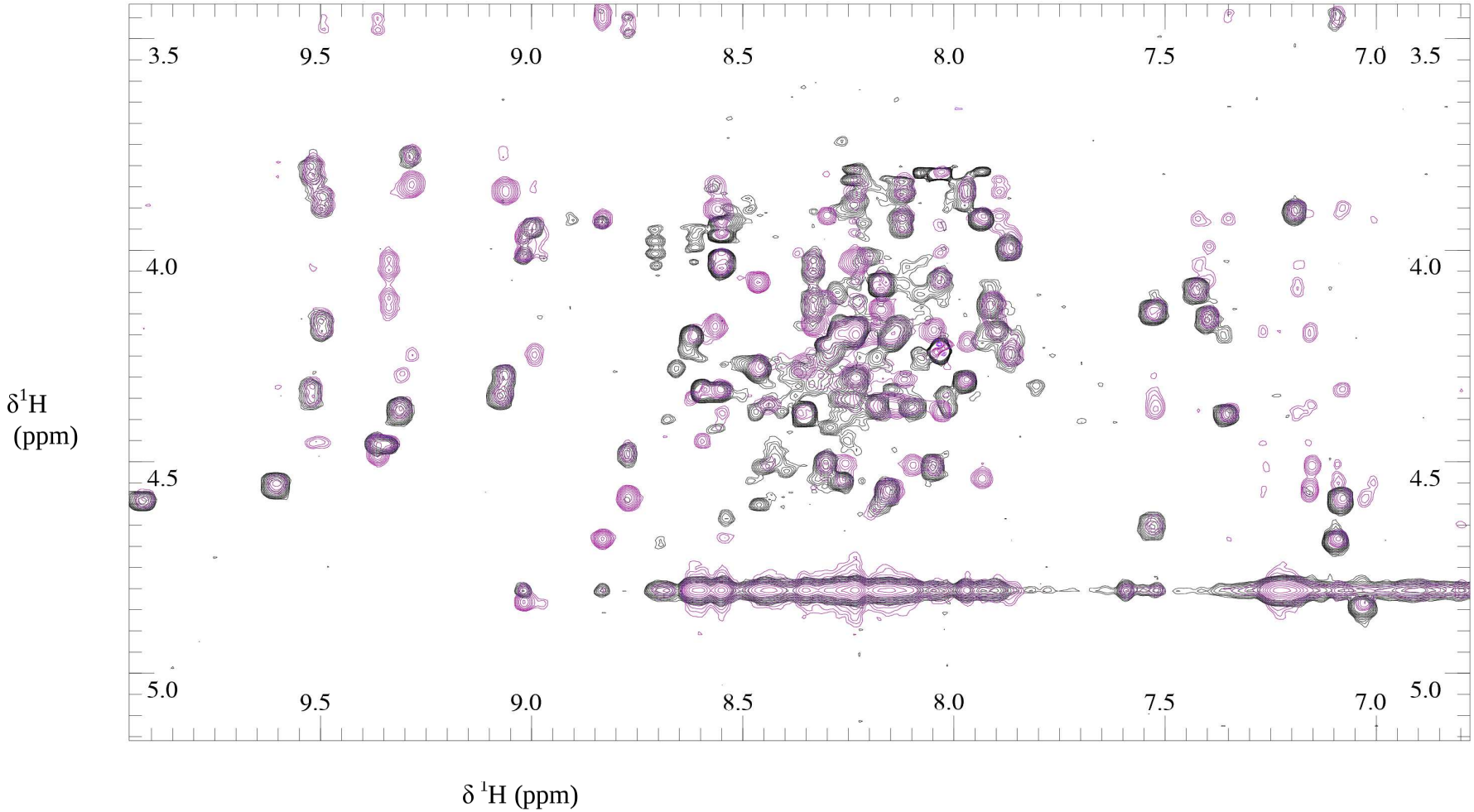
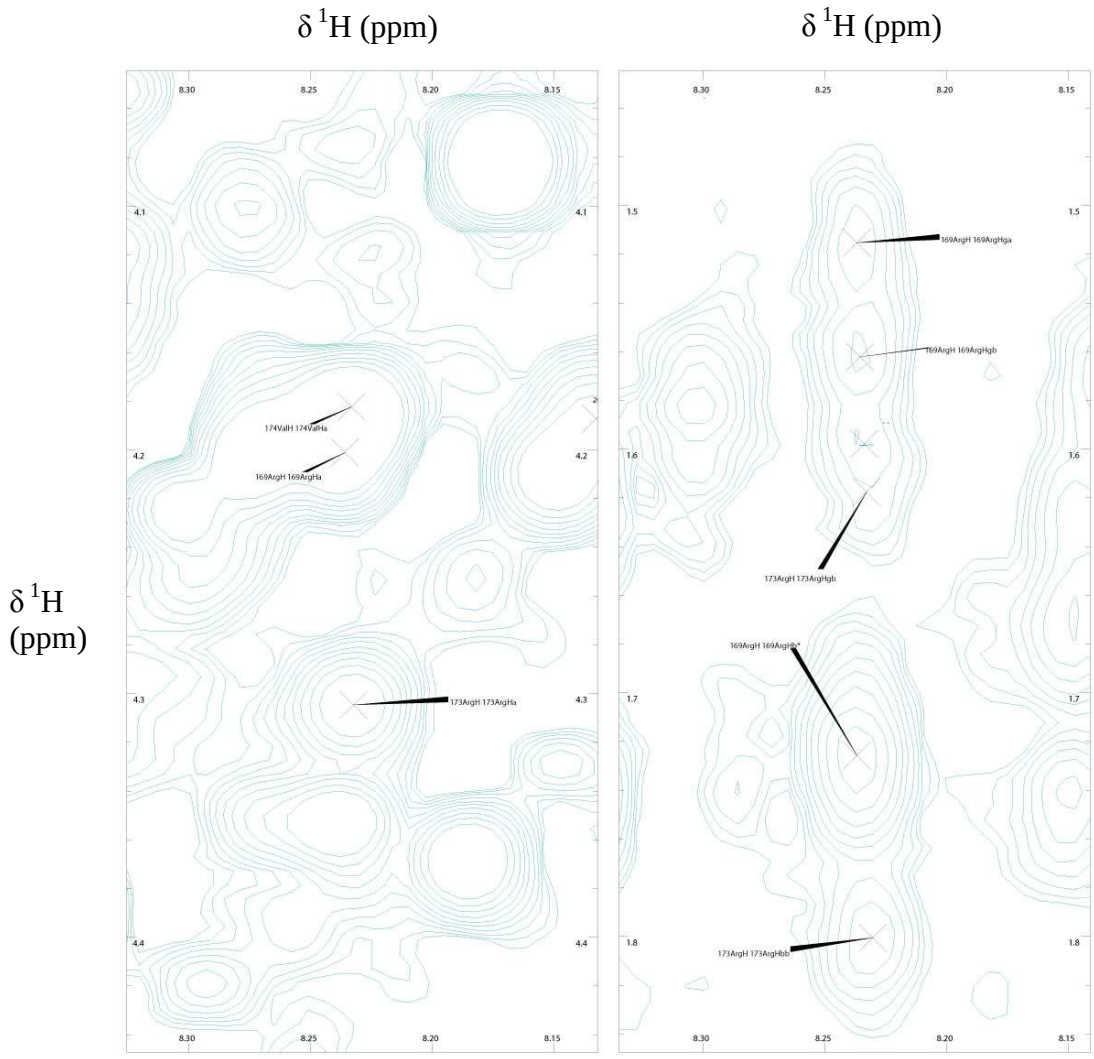


Figure 4.2 Hα-H_N region of the homonuclear 2D TOCSY (black) and 2D NOESY (purple) spectra of hMBD1 CXXC1. Note overlapped region is between 8-8.5pmm (H_N) and 4-4.5ppm (Hα)



a) Hα- H_N chemical shift region

b) Aliphatic-H_N chemical shift region

Figure 4.3 TOCSY strips for the partially assigned arginine 169 & 173 residues. Annotated with assignments made using ^{15}N labelled protein

4.3.2 Assignment of phenylalanine side chains

Using D_2O rather than H_2O as the NMR sample solvent means the H_N are replaced by a deuteron due to proton exchange with the solvent. The D_N no longer produces a ^1H NMR signal and therefore there are no cross peaks in the H_N region of the spectrum. However the cross peaks from aromatic sidechains remain visible (figure 4.4). Of the 3 types of aromatic protons in phenylalanine, only Hδ is expected to give strong NOESY cross peaks to Hα or Hβ, with Hε and Hζ showing weaker signals. hMBD1 CXXC1 only contains two phenylalanines. Thus, with the sequence specific assignment of P207 from C208 this allowed for the rapid assignment of P207 sidechain. The remaining crosspeaks were therefore where assigned to P171.

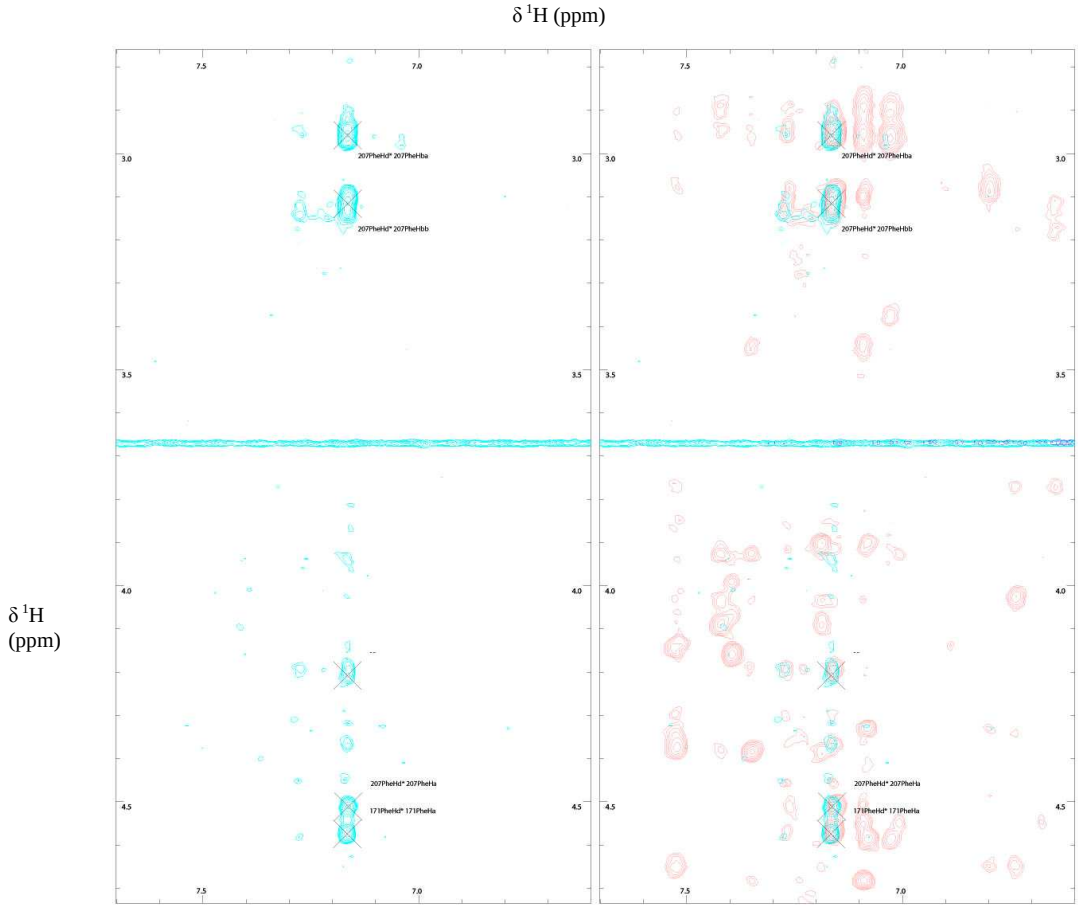


Figure 4.4 Strips for the NOESY of hMBD1 CXXC1 recorded in D_20 showing cross peaks of 171F & 207F. A, D_2O NOESY only (aqua) B, D_2O NOESY overlaid with the NOESY experiment carried out in H_2O (red).

Because complete resonance assignment could not be achieved using solely homonuclear data, ^{15}N hMBD1 CXXC1 was prepared and 3D HSQC-TOCSY & HSQC-NOESY was recorded

4.3.3 Assignment using 3D spectra

Assignment of cross peaks in 2D spectra is a very labour intensive process. Any one horizontal or vertical at a particular chemical shift can have cross peaks belonging to more than one amino acid as figure 4.3 shows. By moving through the nitrogen dimension in 3D HSQC-TOCSY & HSQC-NOESY spectra we can distinguish spin systems based on their nitrogen chemical shift. For example, R169 and R173 that have very similar H_N chemical shift, their spin systems could be more completely assigned

using 3D HSQC-TOCSY & HSQC-NOESY since they have different nitrogen chemical shifts (figure 4.5).

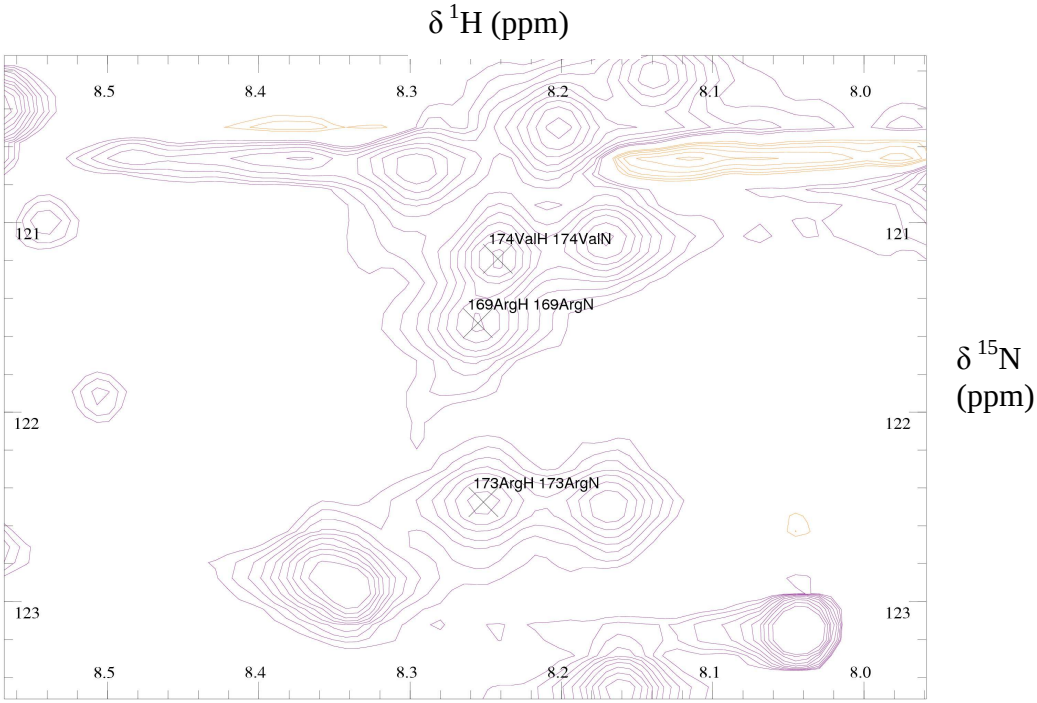


Figure 4.5 ^{15}N HSQC of hMBD1 CXXC1 showing different nitrogen chemical shifts for R169 and R173

Not only do isotopically labeled samples allow the use of 3D experiments to help resolve overlapped regions but it also allows the use of analysis's (Vranken *et al*, 2005) semi automated linking of spin systems. When assigning cross peaks to a particular amino acid or spin system we look for specific correlations in the NOESY spectrum to help qualify the assignments e.g. H_N to Hα NOE cross peaks for i to i+1, NOE cross peaks between sidechain and H_N for i to i+1. The software can automatically look for possible matches. For example, starting from C176 (figure 4.6), analysis will look for NOE cross peaks assigned to the H_N of other spin systems which match the chemical shifts of the assigned C176 TOCSY cross peaks. As the residue following C176 is G177 we can look for cross peaks at a particular chemical shift amongst the matching spin systems. For a glycine we would expect to see 2 TOCSY cross peaks correlating both Hα protons to the H_N proton. Once the next amino acid is identified the software can then be used to search for the next spin system to match to it.

Assignment of H-H cross peaks in the HSQC-TOCSY spectrum allows the assignment of the N-H cross peak in the ^{15}N-HSQC. The complete assignment of the hMBD1 CXXC1 ^{15}N-HSQC is shown in figure 4.7 which shows well dispersed peaks for most amides as well as nitrogen bound proteins in the side chains of glutamines 168, 183, 197 and arginines 213 & 214. Amino acids G162, M170, and H200 are unassigned in the HSQC. Although there are still overlapped peaks in the 8.25ppm ^{1}H region there is sufficient dispersion in the 15N chemical shift around 120ppm to resolve the cross peaks. Also present are cross peaks which can be assigned to a minor form of the protein. The minor form of the protein may be due to proteolysis, loss of a zinc ion or an incorrectly folded version. Although there are N-H correlations from the minor form there are no visible cross peaks in the ^{15}N-HSQC-TOCSY or ^{15}N-HSQC-NOESY spectra.

Using NMR experiments listed in table 3.1 (p75) a near complete assignment of hMBD1 CXXC1 was possible. Resonances that could not be assigned to cross peaks are listed in table 4.1. These amino acids are found in unstructured regions of the hMBD1 CXXC1 domain.

Amino acids	Unassigned protons
R 169	Nε Hε
F 171	γ
R 173	Nε Hε
H 200	Nδ1 Nδ2
R 212	Nε Hε
R 214	Nε Hε
R 221	Nε Hε

Table 4.1 list of protons and their amino acids that could not be assigned

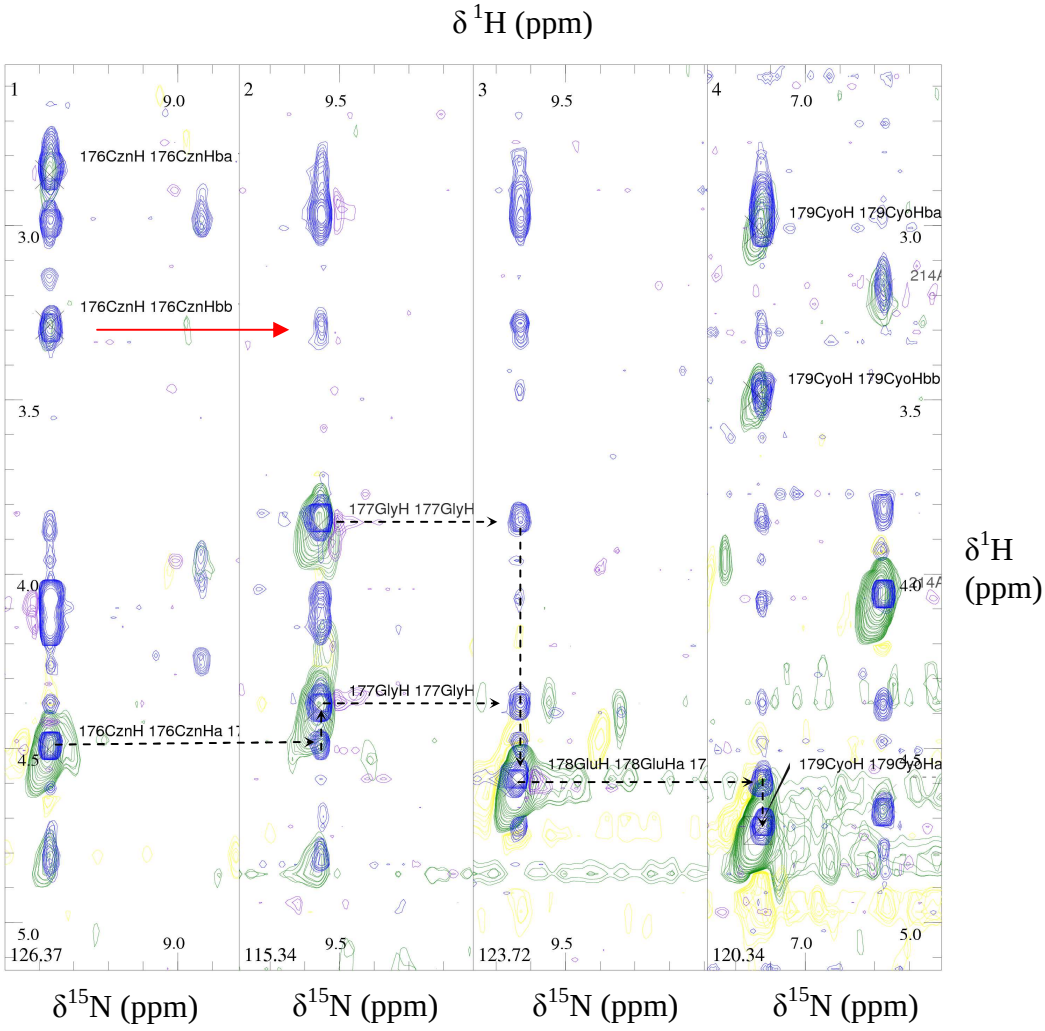


Figure 4.6 Strips from 3D NOESY (blue) & TOCSY (green) spectra showing sequential assignment of amino acids C176 to C179 in hMBD1 CXXC1. The dashed line shows the sequential links involving Hα, while the red arrows links NOE cross peaks correlating C176 Hβb to G176 H_N and G177 H_N

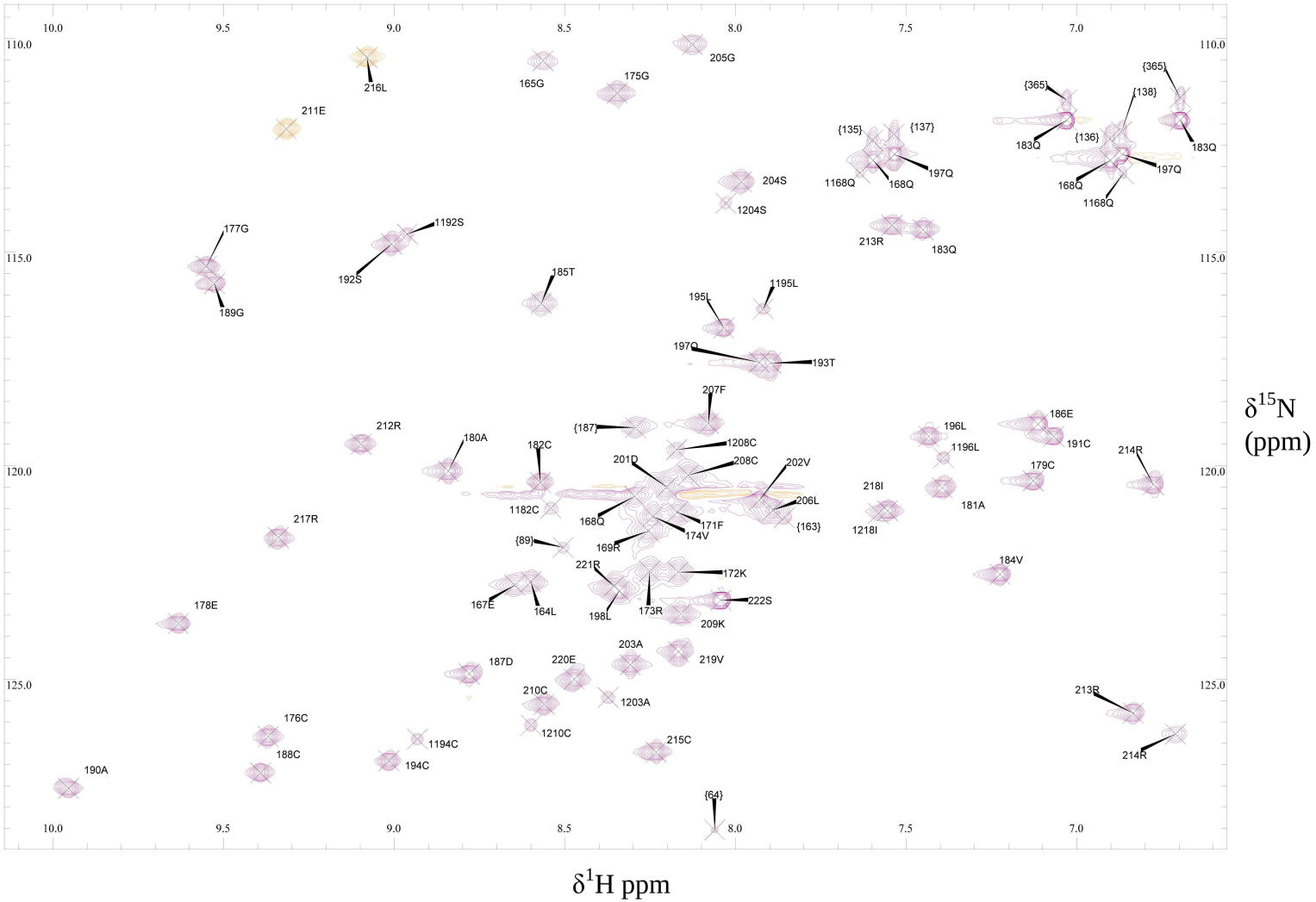


Figure 4.7 15N-HSQC of hMBD1 CXXC1 showing assignments for 56 out 61 amino acids. Minor form chemical shift cross peaks are also shown e.g. 1194C is the minor form of residue C194, crosspeaks with {n} are unassigned spin systems not part of the sequential assignment of the

4.4 Structure calculation of hMBD1 CXXC1 using ARIA

Since the majority of restraints for NMR structure calculations are provided by NOESY derived distance restraints, the cross peaks within the spectrum have to be assigned to a particular resonance. However, due to the large population of protons within a protein many protons may share the same chemical shift so there may be several assignment possibilities for each NOE cross peak. Manual assignment is a time consuming process which may not provide an accurate structure due to mis-assignment. In order to speed up the process and reduce the risk of mis-assignment, structure calculations were carried out using the program ARIA 2.2 (**A**mbiguous **R**estraints for **I**terative **A**ssignment) (Rieping *et al*, 2007). ARIA does not perform the structure calculations itself but drives the calculation of structures through the use of ambiguous distance restraints (unassigned crosspeaks which may have more than one contributing resonance) in an iterative structure calculation scheme. Structures are calculated using the program CNS (Crystallography & NMR System) (Brunger *et al*, 1998), ARIA then analyzes the conformers obtained in order to update the restraints and obtain a set of improved conformers. During distance restraint generation, non-degenerate prochiral chemical shifts e.g. methylene protons or the methyl groups of valine & leucine are arbitrarily stereospecifically assigned. This can result in errors if the wrong member of a prochiral pair is chosen for a particular restraint. To work around this, ARIA applies a floating chiral assignment testing both alternatives during structure calculation. The conformation with the lowest energy is accepted for each prochiral centre.

Distance restraints derived from NOESY cross peaks can be generated in two ways: restraints can be generated by ARIA itself or restraints can be generated in Analysis and imported into ARIA. Both methods were tried for structure calculations as ARIA cannot use spin diffusion correction with restraints generated in Analysis.

4.4.1 Structure calculation using constraints generated by Analysis

Ambiguous distance constraints were generated in Analysis using the shift match restraint function. This function generates a constraint list based on user defined input data. Since the structure of the hMLL CXXC (*Allen et al*, 2006) domain was known it was hypothesized that certain secondary structures present would also be seen within the structure of the hMBD1 CXXC1 domain. Amino acids 1170-1173 of hMLL CXXC form an alpha helix incorporating two cysteines from the second CXXCXXC cluster. Since the corresponding residues in hMBD1 CXXC1, 187-190, had strong NH^{i} to NH^{i+1}

and weak NH^i to $H\alpha^{i+1}$ NOESY resonances suggesting it was likely to be an alpha helix. Since the average distance of N-$H1^i$ to N-H^{i+1} in an alpha helix is 2.8Å, the average NOE intensity of N-Hi to N-Hi+1 cross peaks for amino acids 191-195 should approximately correspond to a distance of 2.8Å. Reference NOE intensities for two 2D NOESY and one 3D HSQC-NOESY experiments were calculated. Parameters used for generating constraints are set out in Table 4.2. To calculate the distance constraints for unambiguous cross peaks the same parameters were used.

Ref. Distance	2.8Å		
Distance function	Intensity^ –1/6		
Intensity type	Height		
Max ppm tolerance	F1 0.05	F2 0.05	F3 0.4
Min ppm tolerance	F1 0.05	F2 0.05	F3 0.4

Table 4.2 Parameters used for generating distance constraints using CcpNMR. Analysis software. F1 – direct proton dimension, F2 – indirect proton dimension, F3 – Nitrogen Dimension

The problem with this method of distant constraint generation is that it is not currently possible to subsequently use ARIA's spin diffusion correction algorithms to recalibrate the restraints based on the calculated structure. For structures calculated from these restraints NOE energies failed to dip below 180 kJ mol^{-1} and it was not possible to obtain structures without violations below 0.5Å. Persistent violations above 0.5 Å included intra-residue Valine H_N to methyl groups and Histidine H_N to β hydrogens. Currently ARIA 2.2 does not support spin diffusion correction (chapter 4.4.4 p87) when importing distance constraints generated in Analysis.

4.4.2 Structure calculation using constraints generated by ARIA

In order to use spin diffusion correction during structure calculation, the assigned chemical shift list and crosspeaks were imported from Analysis into ARIA for each of the NOESY spectra. The tolerances for direct, indirect proton and heteronuclear dimensions were set to 0.025, 0.025 and 0.25 respectively. These values differ from the tolerances used in analysis due to the way they are used. In Analysis the tolerance is the total distance across the crosspeak where as in ARIA the tolerance is set from the centre of the cross peak. Therefore if the same tolerances used in Analysis for generating

constraint lists were to be used in ARIA, ARIA would produce significantly more constraints. This would likely result in the generation of poor structure ensembles.

Spin diffusion correction requires the input spectrometer frequency, mixing time and molecule correlation time into the ARIA project for each spectrum. Molecule correlation time was worked out using stokes law. The hydrodynamic radius was estimated from molecular weight.

4.4.3 Additional restraints used by ARIA during structure calculations

In addition to the distance restraint information generated from the NOESY spectra, restraints specifically for the co-ordination of each of the zinc ions by 4 cysteine residues were included during structure calculations. In order to use zinc co-ordination restraint information, one of the cysteines involved in each cluster was defined as residue type CZN (cysteine + zinc ion) and the remaining 3 cysteines CYS in Analysis. This involved the addition of residue definition files and changing the residues from CYS to either CZN or CYS of the protein sequence located in the molecule description. To proceed with the calculation, the ARIA CNS protocol file was modified by the addition of a code patch. This patch, provided by the ARIA programmers, defined the bond lengths between the sulfur atoms of the cysteines and the zinc ion as well as the bond angles (table 4.3). Structure calculations without the zinc co-ordination restraints were also carried out to verify which cystines clustered together.

Atoms	Bond Length	Bond Angle
S Zn	2.3 Å	
Cα – Cβ - S		114.3558
Cβ - S - Zn		109.5000
Hβ - Cβ - S		107.9185
S – Zn - S		109.5000

Table 4.3 Additional restraints used by ARIA for zinc co-ordination by four cysteine residues. S, sulfur; Zn, zinc.

Restraints generated from residual dipolar couplings, J couplings, dihedral angles, and disulphide bridges, can also be used during structure calculations by ARIA in order to build a more refined structure that can be produced by NOESY distance restraints alone.

4.4.4 Spin diffusion

The choice of mixing time in a NOESY experiment is governed by a number of facts. A NOESY experiment with a long mixing time has the advantage of no zero-quantum peaks (artifacts present in the NOESY spectra from J-coupled peaks, such as ortho-protons on a ring). Long mixing times allow for multiple magnetization transfers which distort cross peak intensity to distance relationship. This multiple magnetization transfer, known as spin diffusion, can affect cross peaks between pairs of protons both close and far apart in space. Protons close in space can be affected by back transfer of magnetization $I_1 \rightarrow I_2 \rightarrow I_1$ reducing the intensity of NOE cross peak that would normally result. Protons far apart in space can be affected by an indirect transfer of magnetization $I_1 \rightarrow I_2 \rightarrow I_3$, increasing the intensity of the cross peak that would normally result or create a cross peak that would not normally exist in the absence of spin diffusion. Therefore, the mixing time needs to be long enough to let the NOE build up and allow zero-quantum effects to decay but short enough to minimise the chance of spin diffusion. As discussed in section 2.2.1 (p44) the NOESY experiment utilises dipolar interactions between two spins $\leq$5 Å apart. The intensity of these cross peaks is used for calculating distance restraints. If left uncorrected, spin diffusion leads to inaccurate distance restraints between protons leading to overly constrained and incorrect protein structures. Spin diffusion correction in ARIA (Linges *et al*, 2004) is based on the calculation of a theoretical transfer matrix from the set of structures produced in each iteration. The theoretical intensity values are then used to calibrate the experimental intensities and to correct the distance target. The calibrated intensities can be then used to estimate the error.

4.4.5 Iterative structure calculation scheme

ARIA uses an iterative strategy for structure calculations (table 4.4). For the structure calculation of hMBD1 CXXC1, unambiguous (manually assigned resonances) and ambiguous restraints were used from the start, iteration 0. When structures have been calculated, a user defined subset from the ensemble is chosen on the basis of their potential energy. These structures are used to filter the restraints (partial assignment filter) based on their contribution to the intensity of the crosspeak. Restraints which are above the threshold, a percentage of the total contribution, are removed from the restraint list. Restraints that are not filtered out are used to recalibrate the spectra based on their average distance chosen from the subset of structures. The restraint list is checked again for violations using the new intensity - distance matrix. In the next

iteration structures are calculated based on the new restraint list created at the end of the previous iteration. Each iteration subsequently reduces the partial assignment filter threshold and violation tolerance. This process is repeated for a user defined number of iterations. In order to improve the recalibration of restraints, spin diffusion correction was included from iteration 3 using a distance cut off of 6 Å. The cut off is used to filter the calculated intensities used for correction, by including only those corresponding to a distance smaller than the distance cutoff value. Assignment statistics are carried out on a user defined number of structures in each iteration which are selected on the basis of having lowest total energy. The number of structures chosen to be used depends on the number of structures calculated in each iteration and the range of total energies of the ensemble. The greater the number of structures calculated the greater number can be used for analysis. In the initial round of structure calculations where assignment mistakes and artifact peaks are likely to be present there is no advantage in generating a large number of structures during each iteration. Structure calculation can be a time consuming process depending on the number of processors that can be accessed. With this in mind the total number of structures in each iteration generated in the initial round of structure calculations was set to 20.

Iteration	Restraints	Partial Assignment Filter	Spin diffusion	Violation Tolerance Å	No. of Structures	No of structures used for analysis
0	Original restraints	1.0	No	10000	20	7
1	Checked restraints from round 0	0.999	No	5.0	20	7
2	Checked restraints from round 1	0.99	No	3.0	20	7
3	Checked restraints from round 2	0.99	Yes	1.0	20	7
4	Checked restraints from round 3	0.99	Yes	1.0	20	7
5	Checked restraints from round 4	0.96	Yes	1.0	20	7
6	Checked restraints from round 5	0.93	Yes	0.3	50	7
7	Checked restraints from round 6	0.9	Yes	0.3	100	7
8	Checked restraints from round 7	0.9	Yes	0.3	200	20

Table 4.4 Iterative strategy used for hMBD1 CXXC1 structure calculation using ARIA.

For speed, structures are typically calculated “in vacuum” which can result artifacts as a result of the simplified treatment of non-bonded forces and missing solvent contacts. ARIA provides the option of refining structures in a shell of water molecules for a defined number of structures from the final iteration with a full molecular dynamics force field incorporating electrostatics (Linge *et al*, 2003). This refinement helps to avoid unrealistic side chain packing and unsatisfied hydrogen bond donors or receptors. To ensure there are no systematic differences that could influence validation results, the force fields used for water refinement (file PARALLHDG 5.3) are consistent with the force fields used for structure calculation and validation.

4.4.6 Analysis and report files

At the end of each iteration, ARIA generates various output files which report the analysis ARIA carried out for each iteration.

1. *noe_restraints.unambig & noe_restraints.ambig*: These files contain all the unambiguous and ambiguous restraints, noting the reference cross peak, the restraint bound, the distance found in the ensemble and the result of the violation analysis
2. *noe_restraints.violations*: lists just the violated restraints containing the same detail as for 1. above
3. noe_restraints.assignments: provides information on wither the assignments stem from fully, partially or unassigned cross peaks.
4. *noe_restraists.merged*: lists all restraints discarded by the merging procedure.
5. *report*: summarises analysis of the restraint lists and the structure ensemble.

In addition to these reports, ARIA can export the last iteration of structures (and water refined structures) to Analysis, a peak list of the assignments (for each spectrum) and a list of violated restraints which were not used for structure calculation in the final iteration. Ambiguous assignment of resonances by ARIA is not infallible since noise and artifact peaks can be assigned if they match the chemical shift data. ARIA can also mis-assign ambiguous peaks if any manual assignments are incorrect. This is due to manual assignment forcing ARIA to satisfy the distant restraint calculated from the NOE crosspeak driving structure calculations in a specific way. After each round of structure calculation violated peaks were inspected and either manually assigned or removed from the peak list. Noise and artifact peaks are removed while mis-assigned peaks were

corrected. Mis-assignment can be a problem if the tolerances for automatic assignment are set too low (see chapter 4.4.6) which can for example allow intra-residue cross peaks to be assigned as inter-residue.

Once calculations started to yield structures that have low energies then the number of structures calculated in iteration 8 was increased. An increase in the number of structures calculated will increase the number of low energy structures within the ensemble. For example, if 20 structures are calculated they may not all be statistically similar even though they fit the experimental data. However, if 200 structures are calculated, then 25 structures of the ensemble maybe of significant similarity. Once NOE energies had reached a sufficiently low level, < 100 kJ mol^{-1}, water refinement was introduced to the calculation. Since the water refinement step introduces new constraints, the overall energies of the ensemble tend to increase. Therefore there is no advantage in introducing this step in earlier round of calculations when the structures are not near converging. When calculation produced structures of significantly low NOE energies <50 kJ mol^{-1} and no more improvement could be made using the current restraints, analysis was carried out to verify the precision and accuracy of the structures. Using NMR to generate a unique three dimensional structure is not possible because multiple structures can be calculated that are consistent with restraints derived from experimental data. This is compounded by less defined regions of protein structure which are dynamic in solution. Thus an ensemble of structures with the lowest energies can be produced to show a meaningful representation of the structure consistent with the experimental data.

.

CHAPTER FIVE

STRUCTURE AND FUNCTIONAL ANALYSIS OF HMBD1 CXXC1

5.1 Overview

This chapter reports the structure of hMBD1 CXXC1 and discusses its implications in relation to the other known CXXC structures. The CXXC domains that have been shown to bind non-methylated CpG contain a conserved amino acid sequence ^{361}KFGG^{364} and other conserved residues (D358 & Q370) as found in hMBD1 CXXC3. hMBD1 contains, depending on alternative splicing, 2 (55kDa) or 3 (66kDa) CXXC domains (Fujita *et al*, 1999, Jørgensen *et al*, 2004) with only the third CXXC domain capable of binding non-methylated CpG (Fujita *et* al, 2000 & Jorgensen *et al*, 2004). Recently, the solution structure of the MLL CXXC domain in complex with DNA (Cierpicki *et al*, 2009) and the crystal structure of DNMT1 CXXC (Song *et* al, 2010) have been solved revealing the amino acids responsible for DNA contact. Comparison of these structures to hMBD1 CXXC1 reveals why hMBD1 CXXC is unable to bind non-methylated CpG.

5.1.1 The solution structure of hMBD1 CXXC1

The hMBD1 CXXC1 ensemble, comprising 20 lowest energy of 25 water refined structures (figure 5.1), shows a well defined backbone region for residues 173-196 & 208-218, with the N- & C- termini and residues 197-207 poorly defined. RMSD was calculated for each residue using CcpNMR Analysis. Analysis uses singular value decomposition (SVD) to calculate an optimum rotation for two coordinate sets with co-located centroids. Each atom is weighted, both for the SVD and centroid calculation. There are multiple rounds of superimposition to get the ensemble from pairwise model comparison and to better refine the weights. Initially the weights come from the atomic masses, but in the later stages of the alignment the weights come from the atomic RMSDs calculated in the earlier round. Accordingly, the dissimilar parts have

Proportionately little influence on the final ensemble. Figure 5.2 shows the per residue RMSD for the hMBD1 CXXC1 ensemble and the number of inter-residue NOEs from the ^{15}N NOSEY. Only the ^{15}N NOSEY was used because the peak list imported from ARIA has less ambiguouity than the 2D ^{1}H NOESY giving a truer representation of NOEs used. Comparison of the number of inter-residue NOEs and RMSD of each residue reveal that the defined structured regions of the ensemble are characterized by a high number of inter-residue NOEs per residue while the N- and C- termini have very few inter-residue NOEs that consist mainly of short range, sequential correlations

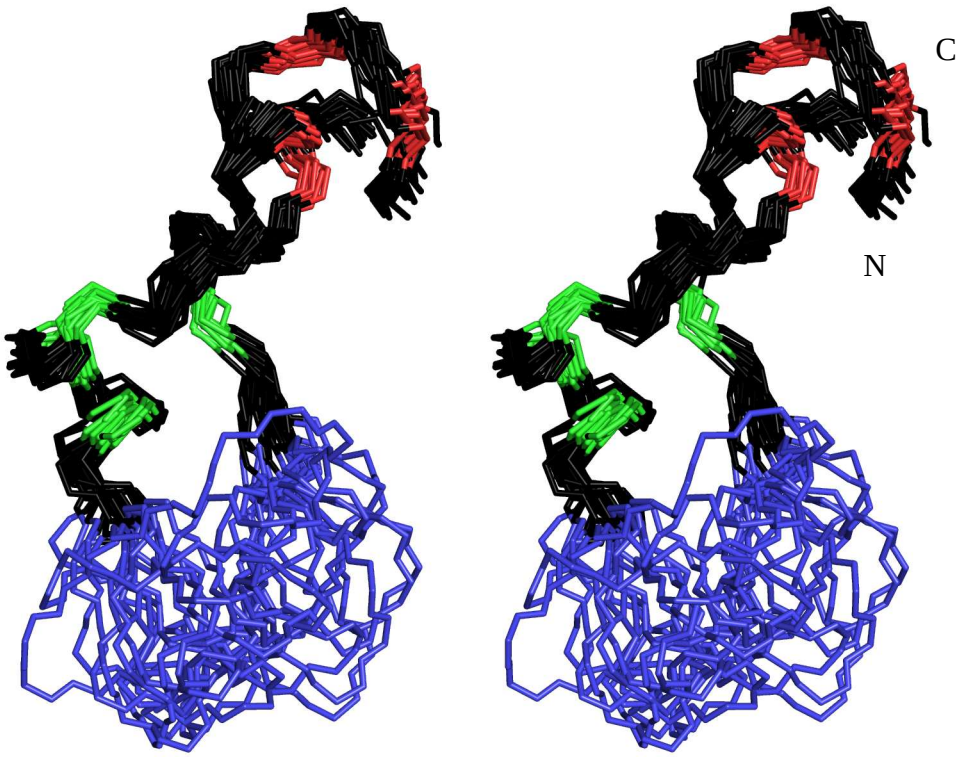


Figure 5.1 A stereo view of the ensemble of 20 lowest energy of 25 final hMBD1 CXXC1 solution structures superposed on the Cα of residues in well-defined regions (residues 175 to 196 & 208 to 217). The less well defined loop (residues 197 to 207) is blue with the two cysteine clusters coloured red and green (Figure prepared using PyMol)

.

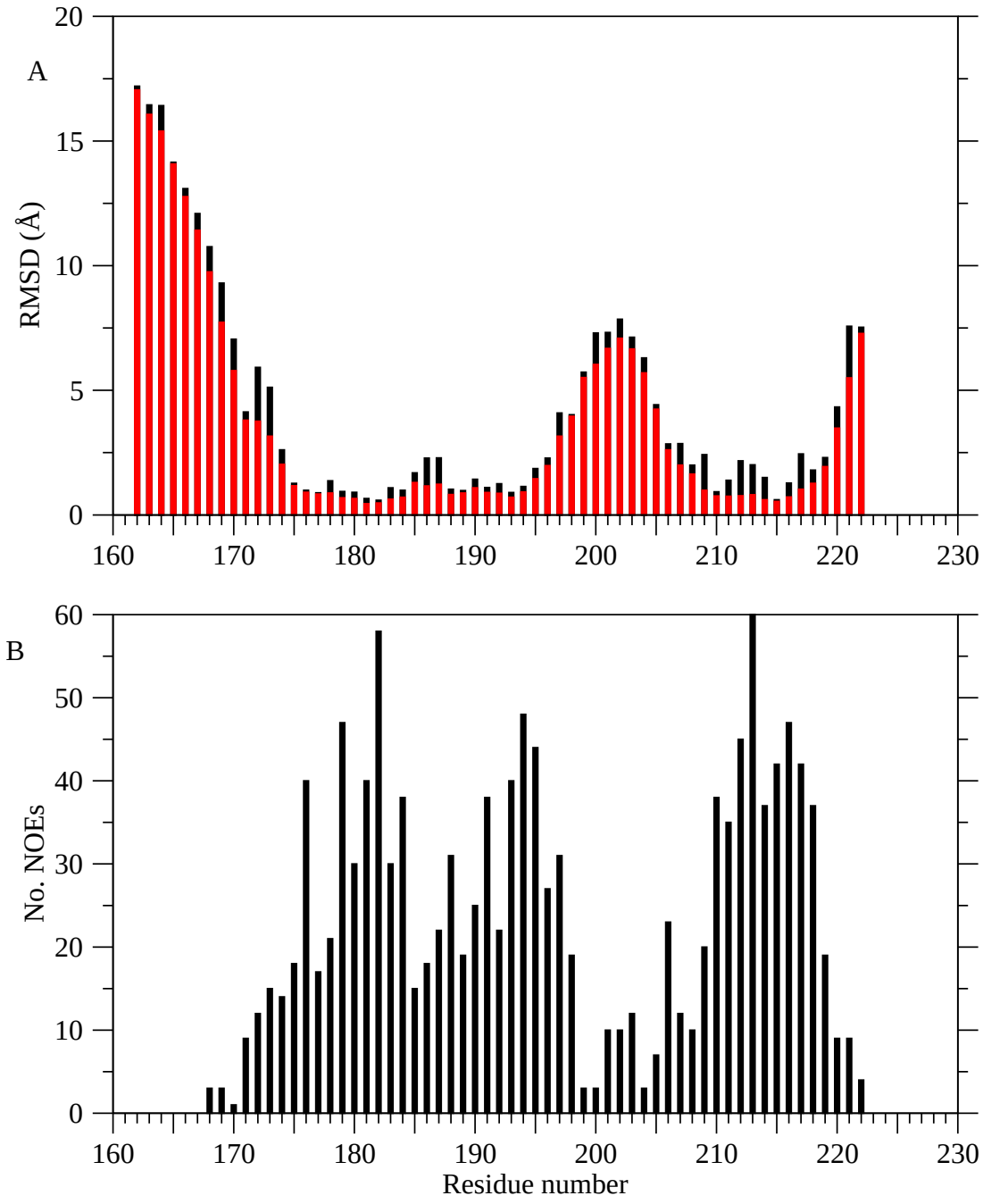


Figure 5.2 A. Per residue backbone and all atom RMSD for the final 25 hMBD1 CXXC1 structures. Red bars, backbone; black bars, all atom. RMSD calculated using CcpNMR Analysis 2.1.5. B. Number of per residue NOE assignments from ^{15}N-NOESY generated by ARIA for structure calculations in iteration 8.

5.1.2 Precision of the ensemble of hMBD1 CXXC1 structures

To be confidant that the ensemble of structures is an accurate representation of the experimental data the similarity of the structures was assessed by calculation of the Root Mean Squared Deviation (RMSD) of the atomic coordinates of each structure. It is possible to calculate the RMSD by superposition on a selected structure (biased) or by global minimization of all fits (complete intensity). However, by using an unbiased mean the calculation of the RMSD is not forced on one structure which may not be the most accurate representation of the ensemble. To calculate the unbiased mean structure, the unweighted mean program UWMN (Hartshorn and Caves, University of York) was used. UWMN creates a matrix containing the average distances between atoms in the ensemble of structures. Since this matrix cannot typically be projected into cartesian space perfectly, the matrix is projected into multi-dimensional space and orientated so that when it is projected back into three dimensions the mean structure is calculated with the least loss of structural information. The differences between the mean and the structures are listed as well as the average RMSD to the unbiased mean of the ensemble which is listed in table 5.1

	RMSD to the unbiased mean structure (Å)
Heavy atoms	1.143
Backbone atoms	0.854

Table 5.1 RMSD of the final structures in the ensemble calculated to an unbiased mean. RMSD calculated for residues 175-196 & 208-217.

5.1.3 Geometric analysis of the ensemble of hMBD1 CXXC1 structures

The geometric quality of an NMR structure can be validated by the distribution of the backbone (φ and ψ) side chain (χ_1) torsion angles (Morris *et al*, 1992). The φ angle is the dihedral angle between the H_N and Cα, the ψ angle is the dihedral angle between the Cα and carboxylic carbon. The geometric quality of a structure can be analysed the program using Procheck-NMR (Laskowski *et al*, 1996). Procheck-NMR assesses the stereochemical quality of structures by comparing them with the known stereochemical properties of well refined, high resolution X-ray structures. Therefore, Procheck

assesses how normal, or how unusual, the geometry of the residues in a given protein structure is. This analysis is reliant on the reference structures having normal geometric quality. The summary of Ramachandran statistics for the residues (excluding all prolines and glycine residues) of the ensemble of hMBD1 CXXC1 structures is shown in table 5.2 and the plots shown in appendix B. Greater than 90% of the residues fall within most favorable and additionally allowed regions. This is indicative of a structure with a good geometric quality. The contribution of unfavorable φ and ψ angles (Appendix B, figure b2a-c) in the unstructured N- (163-174) & C- (217-222) termini as well as the poorly defined loop region (197-207) results in higher than average number of residues in the disallowed region. This is due to a lack of restraints to define these regions of the domain. If the unstructured regions are removed from the Ramachandran analysis, the percentage of residues in the disallowed regions is significantly reduced.

Regions of Ramachandran plot	% residues
Residues in most favoured regions	61.3% (77.7%)
Residues in additional allowed regions	32.3% (21.0 %)
Residues in generously allowed regions	3.3% (0.7%)
Residues in disallowed regions	3.1% (0.7%)

Table 5.2 Summary of the Ramachandran statistics for the ensemble of 25 hMBD1 CXXC1 structures. Percentage in brackets indicates statistics for well defined regions only (residues 175-196:208-216)

The geometric quality of a structure can also be analysed by the distribution of side chain torsion angles χ_1. This is the angle of N-Cα-Cβ-Xγ1 about the Cα-Cβ bond. The distribution of side chains in hMBD1 CXXC are relatively well defined with no residue lying in unfavourable regions (appendix B2d-e).

Table 5.3 lists the statistics of experimental restraints used for calculating the structure of hMBD1 CXXC1. Total number of NOEs represents NOEs merged from the three NOESY spectra. The constraint violations and R.M.S deviation data refers to 20 lowest energy structures from 25 water refined structures calculated.

NMR distance restraints	
Total NOE	1327
Ambiguous (main contributors)	593
Intra-residue	204
Inter-residue	
Sequential (i-j=1)	156
Short range (i-j= >1 <4)	118
Long range (i-j>5)	115
Unambiguous	734
Intra-residue	334
Inter-residue	
Sequential (i-j=1)	211
Short range (i-j= >1 <4)	102
Long range (i-j>5)	87
Constraint Violations per structure	
>0.5 Å (+/- SD)	0.10 (+/- 0.06)
>0.3 Å (+/- SD)	1.35(+/- 0.16)
>0.1 Å (+/- SD)	19.05 (+/- 0.81)
R.M.S. deviations from the ideal geometry	
Bond length (Å)	0.0038 (+/- 0.000023)
Bond Angle (º)	0.600 (+/- 0.05)
Improper angles (º)	1.476 (+/- 0.04)
Dihedrals (º)	42.17 (+/- 0.137)

Table 5.3 Experimental restraint statistics for hMBD1 CXXC1

5.2 Structural features of the hMBD1 CXXC1 zinc binding motifs

The solution structure of MBD1 CXXC1 adopts a crescent shape incorporating two zinc ions (figure 5.3). Four cysteines provide the ligands for the coordination of each zinc ion, three from each CXXCXXC motif. The main chain loops back 180º after the second CXXCXXC motif to providing the fourth cysteine in each case. The overall structure of the domain is governed by the presence of the two zinc clusters. The zinc clusters have similar structures to each other with a backbone RMSD of 0.86 Å over the CXXCXXC sequence (calculated using Superpose. Maiti *et al*, 2004) and form a small helical turn between the first and second cysteines followed by a single turn of alpha helix, that includes residues 180A-183Q cluster 1 (figure 5.4a) and 192S-196L cluster 2 (figure 5.4b). However, the two CXXCXXC motifs appear to differ with respect to their side chain χ_1 dihedral angles (Figure 5.3c). This is consistent over the ensemble. The χ_1 dihedral angles could be determined from J couplings determined from an HNHB experiment:

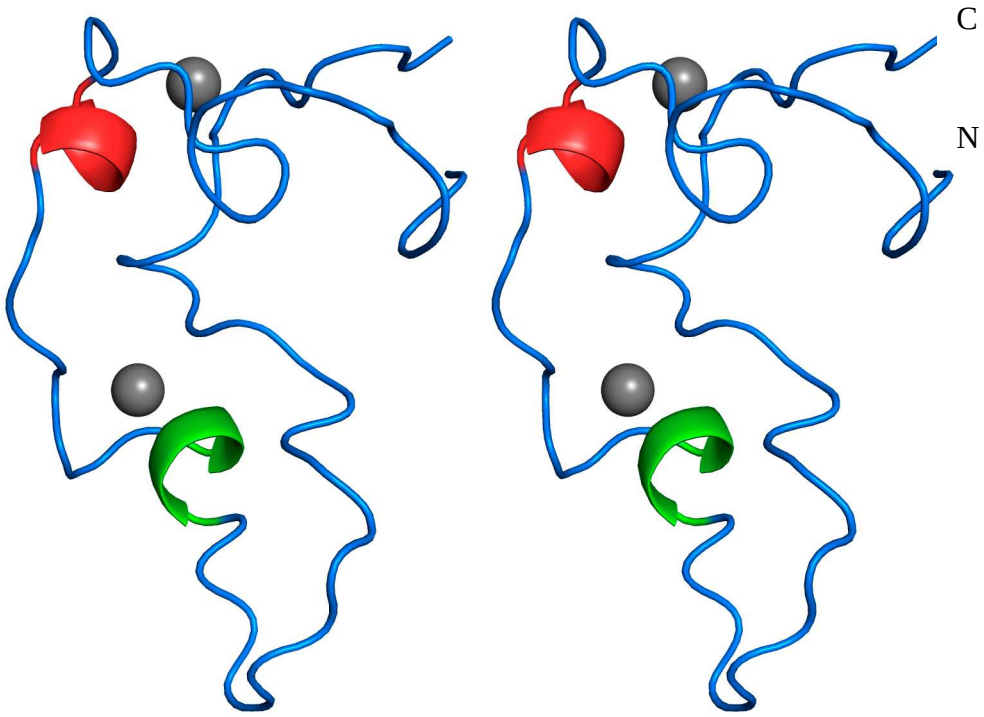


Figure 5.3 A stereo view of the representative structure of the closest to the mean structure of the 25 calculated hMBD1 CXXC1 solution structures. The zinc atoms are shown as gray spheres; alpha helix are shown in red and green; N, N-terminus; C, C-terminus (Figure prepared using PyMol).

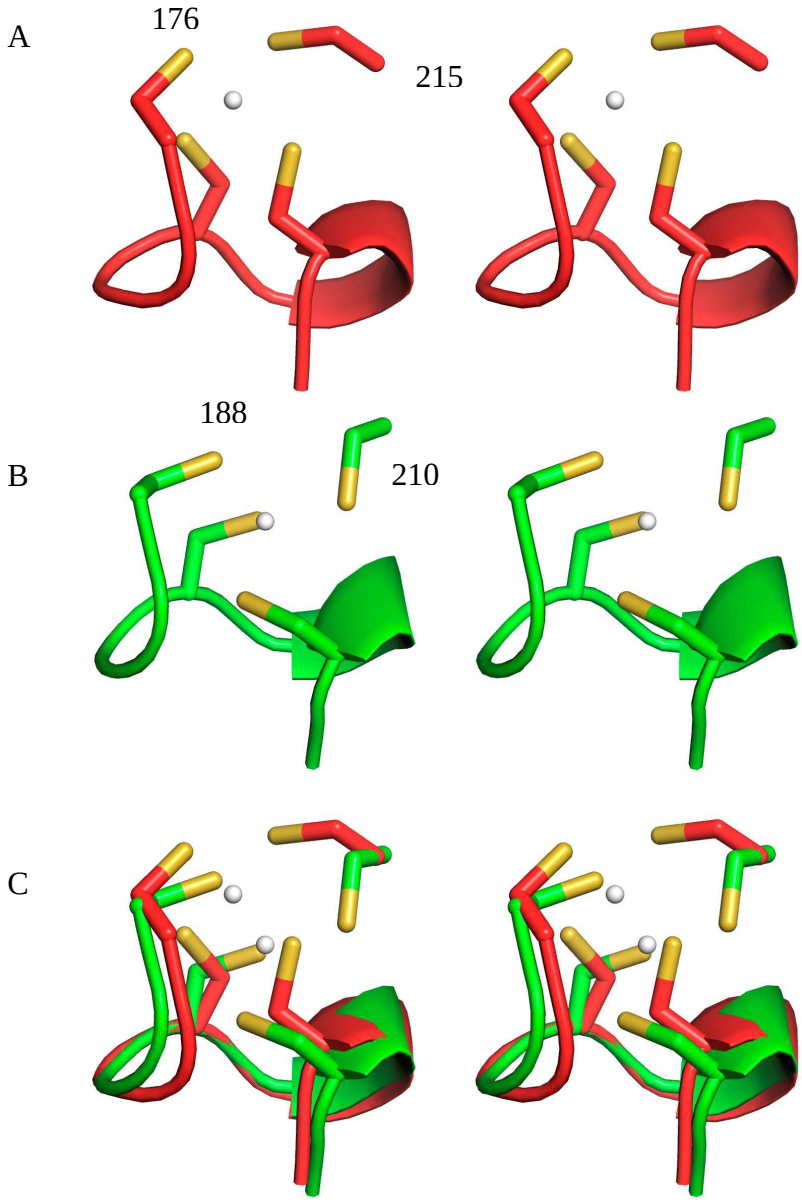


Figure 5.4 Stereo representation of the two cysteine zinc clusters from hMBD1 CXXC1 (A) First CXXCXXC motif, residues 176-183 & 215 (B) Second CXXCXXC, motif residues 188-194 & 210 (C) superposition of the two CXXCXXC motifs. Sulphur atoms, yellow; zinc atoms, white (Figure prepared using PyMol).

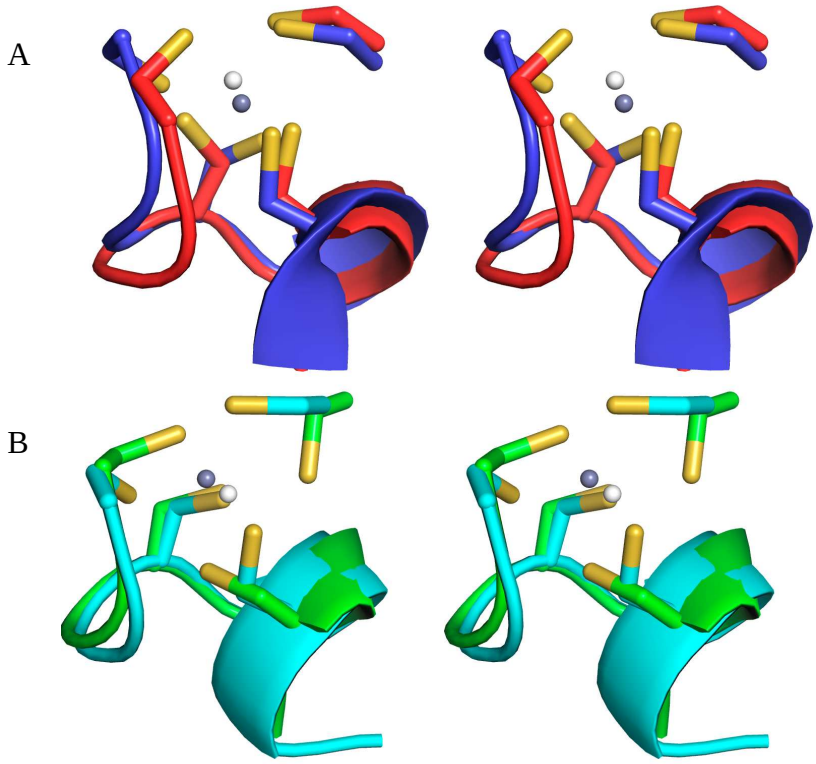


Figure 5.5 Stereo representation of the structural alignment of (A) hMBD1 CXXC1 residues 176-183 & 215 (red) and hMLL CXXC residues 1155-1162 & 1194 (blue). (B) hMBD1 CXXC1 residues 188-194 & 210 (green) and the hMLL CXXC residues 1167-1175 & 1189 (cyan). White spheres, hMBD1 CXXC zinc atom; gray spheres, hMLL CXXC zinc atom. (Figure prepared using PyMol)

The zinc clusters from hMBD1 CXXC1 have a fold which is very similar to the zinc clusters from hMLL CXXC (Figure 5.5). Backbone RMSDs of 1.12 Å and 0.46 Å were calculated using Superpose (Maiti *et al*, 2004) for the first and second CXXCXXC motifs respectively. A search of the structural data base http://www.ebi.ac.uk/pdbe-site/pdbemotif/ for proteins that co-ordinate zinc by four cysteine residues revealed 62 structures. However, nine of these structures use two separated CXXC motifs to coordinate the zinc atom rather than the CXXCXXC motif found in hMBD1 CXXC1 and MLL CXXC. Alignment of the CXXC motifs found in the pdbe database to residues 179-182, 182-185, 188-191 and 191-194 hMBD1 CXXC1 reveals a highly similar backbone conformation with a pairwise RMSD 0.77 Å (±0.33). Three structures

where found with a CXXCXXC motif: horse liver alcohol dehydrogenase (PDB - 1N8K) (Rubach & Plapp, 2003); *E.coli* RecQ helicase (PDB – 1OYW) (Bernstein *et al*, 2003) and the delta subunit of the clamp loader complex of E. coli DNA polymerase III (PDB – 1A5T) (Guenther *et al*, 1997). The horse liver dehydrogenase has the fourth cysteine C-terminal to the CXXCXXC motif like hMBD1 CXXC1 and unlike the RecQ helicase and delta subunit of the clamp loader complex of E. coli DNA polymerase III in which it is N- terminal to the CXXCXXC motif. Table 5.4 shows the pairwise RMSD for the CXXCXXC motifs of hMBD1 CXXC1 and the other CXXCXXC motifs described above. Figure 5.7 shows the superposition of the CXXCXXC motifs from horse liver alcohol dehydrogenase, *E.coli* RecQ helicase, delta subunit of the clamp loader complex of E. coli DNA polymerase III and the second CXXCXXC motif from hMBD1 CXXC1

	1N8K 97-103	1A5T 59-65	1OYW 397-403
hMBD1 176-182	0.85 Å	0.98 Å	0.95 Å
hMBD1 188-194	0.39 Å	0.49 Å	0.51 Å

Table 5.4 Pairwise backbone RMSD of CXXCXXC motifs compared to the CXXCXXC motifs from hMBD1 CXXC1 domain. 1N8K, horse liver alcohol dehydrogenase; 1A5T, delta subunit of the clamp loader complex of E. coli DNA polymerase III; 1OYW, *E.coli* RecQ helicase.

A common feature of all the CXXCXXC motifs of known structure is a glycine residue following the first cysteine. Preference for glycine at this position may be due to the requirement for a positive phi angle following the first cysteine. The CXXCXXC motif has a characteristic helical turn between the 1st and 2nd cysteines, similar to that of the single CXXC motif, before an alpha helix incorporating the 3rd cysteine (figure 5.6). Mutations of one of the zinc coordinating cysteines will destabilise the structure of the MLL CXXC domain (Allen *et al*, 2006. Cierpicki *et al*, 2009), MBD1 CXXC3 (Stancheva *et al*, 2010) and CGBP (Skalnik *et al*, 2001) suggesting the CXXCXXC motifs are absolutely required for maintaining the structure of the domain.

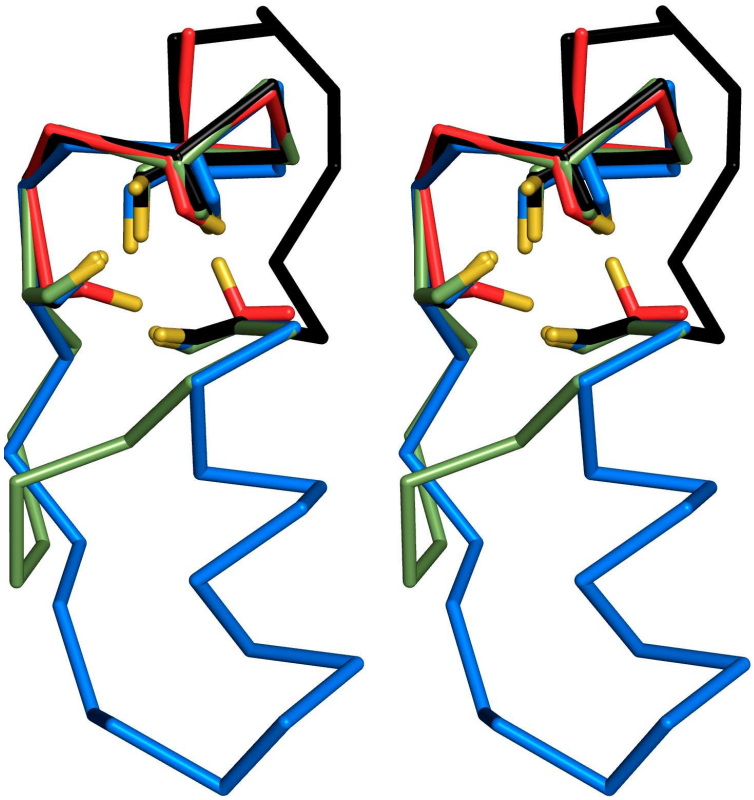


Figure 5.7 Stereo view of the superposition of CXXCXXC motifs from hMBD1 CXXC1 (red), horse liver dehydrogenase (black), *E.coli* RecQ helicase (blue), and the delta subunit of the clamp loader complex of E. coli DNA polymerase III (green). Cystine side chains shown as sticks, sulphur atoms in yellow.

5.3 hMBD1 CXXC1 does not bind CpG DNA

Previous DNA binding experiments (EMSA) have shown hMBD1 CXXC1 not to bind CpG DNA. However, DNA titration by NMR would show weak binding that might not be observed by EMSA. Although chemical shift perturbation analysis by ^{15}N HSQC does not require the structure to be known, the assignment of spectrum crosspeaks is required to know which residues are affected by the titration of DNA. When DNA is titrated into a protein sample a change in chemical shift will be recorded for any amino acid that is affected by the binding of the protein to DNA. This change in chemical shift of a residue is due to an environmental change that may be due to direct interaction with the DNA or a resultant conformational change of the protein.

5.3.1 ^{15}N HSQC DNA titration of hMBD1 CXXC1 reveals no protein DNA interaction

The titration of hMBD1 CXXC1 with DNA measured by ^{15}N HSQC is shown in figure 5.8. There is little or no chemical shift perturbation with 3 fold excess DNA for the assigned resonances suggesting that hMBD1 CXXC1's Kd for DNA is significantly above 150 μM. However, small chemical shift perturbations are seen for the unassigned spin system 435 and the minor form A1203. Resonance A1203 was assigned as the minor form of A203 as they share the same Hα and Hβ chemical shifts. It is likely that resonance 435 is also from a minor form and since perturbation is only visible at the 3 fold excess DNA concentration, this suggests non-specific DNA binding. The minor form resonance A1203 is also only visible at the 3 fold excess DNA concentration. This is likely due to unfolding or degradation of the protein during the course of the titration. Intensity differences are seen for glycine 165 and the side chain Nε of arginine 214. These highly solvent exchangeable resonances appear to be very sensitive to sample conditions since their intensities vary considerably from sample to sample. It is therefore likely that the effect seen here is not DNA-specific, but rather reflects small differences in ionic strength. The lack of chemical shift perturbation that suggests no protein DNA interaction is consistent with the findings of Fujita *et al* (2000) and Jorgensen *et al* (2004) who observed no DNA shift in electrophoresis mobility shift assay experiments.

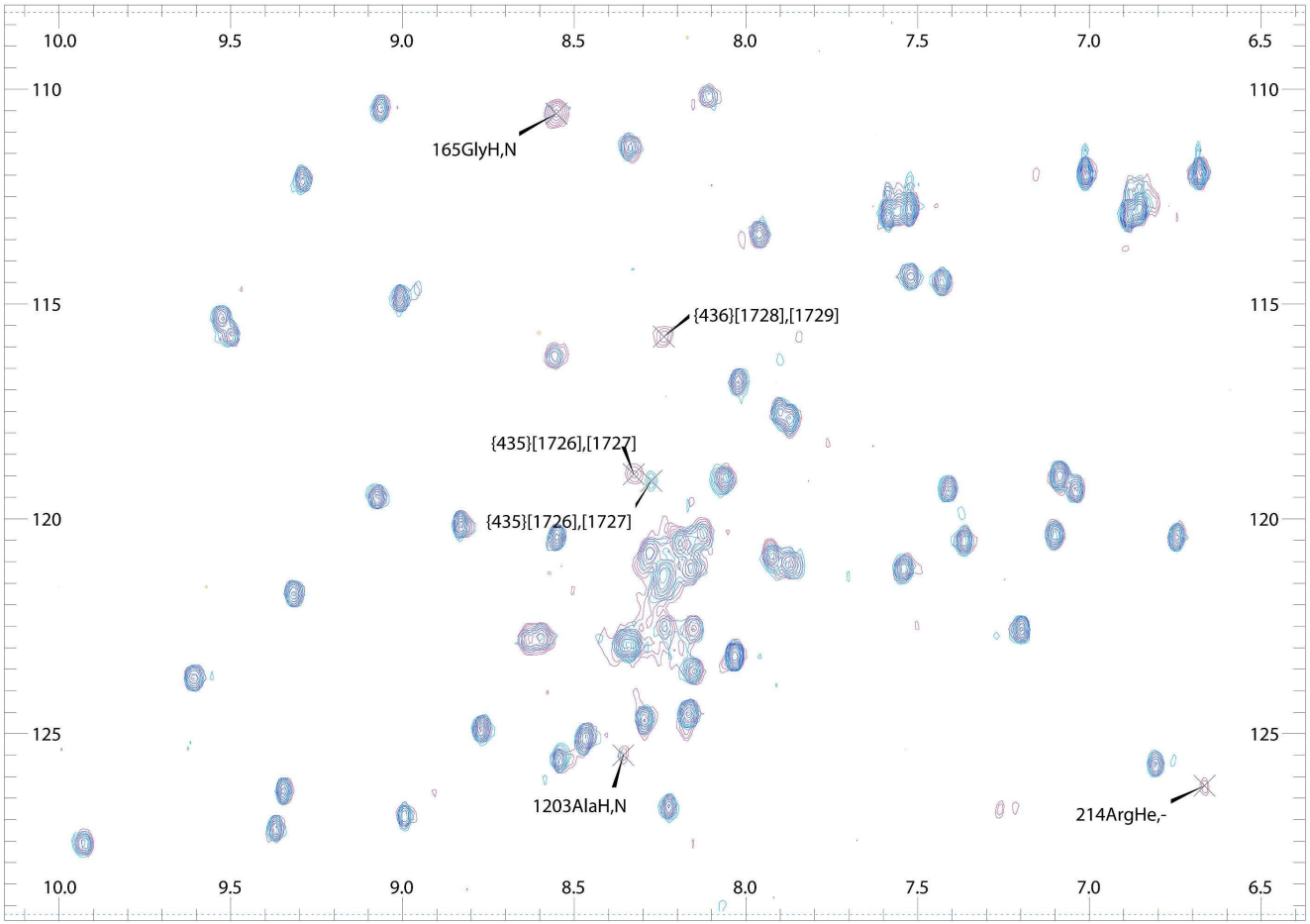


Figure 5.8 ^{15}N HSQC of hMBD1 CXXC1 without (blue) or with (red) DNA. 50 µM protein, 10 mM dTris, 250 mM NaCl, pH 7.5, 293K, final titration ratio 1:3 protein:DNA. Resonances which change are marked. Assignments in {} are from a minor form.

5.3.2 Why does hMBD1 CXXC1 not bind DNA?

The MLL CXXC (Lee *et al*, 2001), MBD1 CXXC3 (Bird *et al*, 2004) , CGBP CXXC (Skalnik *et al*, 2001) and DNMT1 CXXC (Meehan *et* al, 2008., Pradhan *et al*, 2008.) have been shown to bind CpG DNA and therefore a comparison of the known DNA binding CXXC structures with that of hMBD1 CXXC1 should provide an insight as to why hMBD1 CXXC1 does not bind CpG DNA. The solution structure of MLL CXXC (Cierpicki *et al*, 2008) and the X-ray crystal structure of DNMT1 CXXC (Song *et* al, 2010) in complex with CpG containing DNA reveal that the positively charged face of the domain binds to the target DNA major groove.

The MLL CXXC domain binds the CpG motif in the major groove though formation of hydrogen bonds to the CpG. Using the terminology of Cierpicki *et al* (2009), the N4-amine groups of the negative and positive stand cytosines hydrogen bond to the backbone oxygen of K1185 and K1186 respectively. The side chain of K1186 forms hydrogen bonds to the negative strand guanine while the side chain amide of Q1187 hydrogen bonds to positive strand guanine (Cierpicki *et al*, 2009). Cierpicki *et al* (2009) also discovered weak NOEs between the side chains of R1150, S1152 and L1197 and the DNA minor groove and mutation of R1150 and L1197 to alanine resulted in a 5 and 4 fold reduction in DNA binding affinity respectively. DNA binding domains are generally evolved to contain positively charged amino acids though which they contact the negatively charged DNA phosphate backbone (Jones *et al*, 2003) while the shape of the DNA binding face is often complementary to the DNA duplex surface (Tsuchiya *et al*, 2004). hMLL CXXC has made electrostatic interactions to the DNA and residues R1154, K1176, K1178, K1185, K1190, R1192, K1193.

Table 5.5 show the corresponding residues from hMBD1 CXXC1 and xDNMT1 CXXC. hMBD1 CXXC1 differs from hMLL CXXC in many of the residues required for CpG DNA binding, not only in the residues that hydrogen bond to the CpG but also in residues involved in electrostatic interaction. In addition to electrostatic interactions and hydrogen bonds, repulsive forces between the protein and DNA may reduce or ultimately abolish DNA binding. Mutation of residue 1188 from a cysteine to an aspartate in hMLL CXXC is sufficient enough to abolish detectable DNA binding as mentioned by NMR.

hMLL CXXC	hMBD1 CXXC1	xDNMT1 CXXC
R1150	R161	E516
S1152	A163	N518
R1154	N165	I520
K1176	Q197	M546
K1178	P199	K548
K1185	L206	T555
K1186	F207	K556
Q1187	C208	Q557
K1190	E211	M560
R1192	R213	R562
K1193	R214	563
L1197	I218	L567

Table 5.5 List of residues involved in electrostatic interaction with DNA from hMLL CXXC1 (Cierpicki *et al*, 2009) and the corresponding residues in hMBD1 CXXC1 and xDNMT1 CXXC. Residues mutated to alanine that abrogate DNA binding shown in bold, mutated residues that still bind DNA shown in blue.

A surface charge plot of hMBD1 CXXC1 was produced using the Adaptive Poisson-Boltzmann Solver (APBS) plugin in PyMol reveals a positive charge surrounding the first CXXCXXC motif while the second CXXCXXC motif and loop region has an overall negative charge. This contrasts with the positive charged face of hMLL CXXC domain (figure 5.9). In conclusion hMBD1 CXXC does not bind DNA because lacks the residues that are important for hydrogen bonding and has a slightly negative charged face compared to the strongly positively charged face of hMLL CXXC.

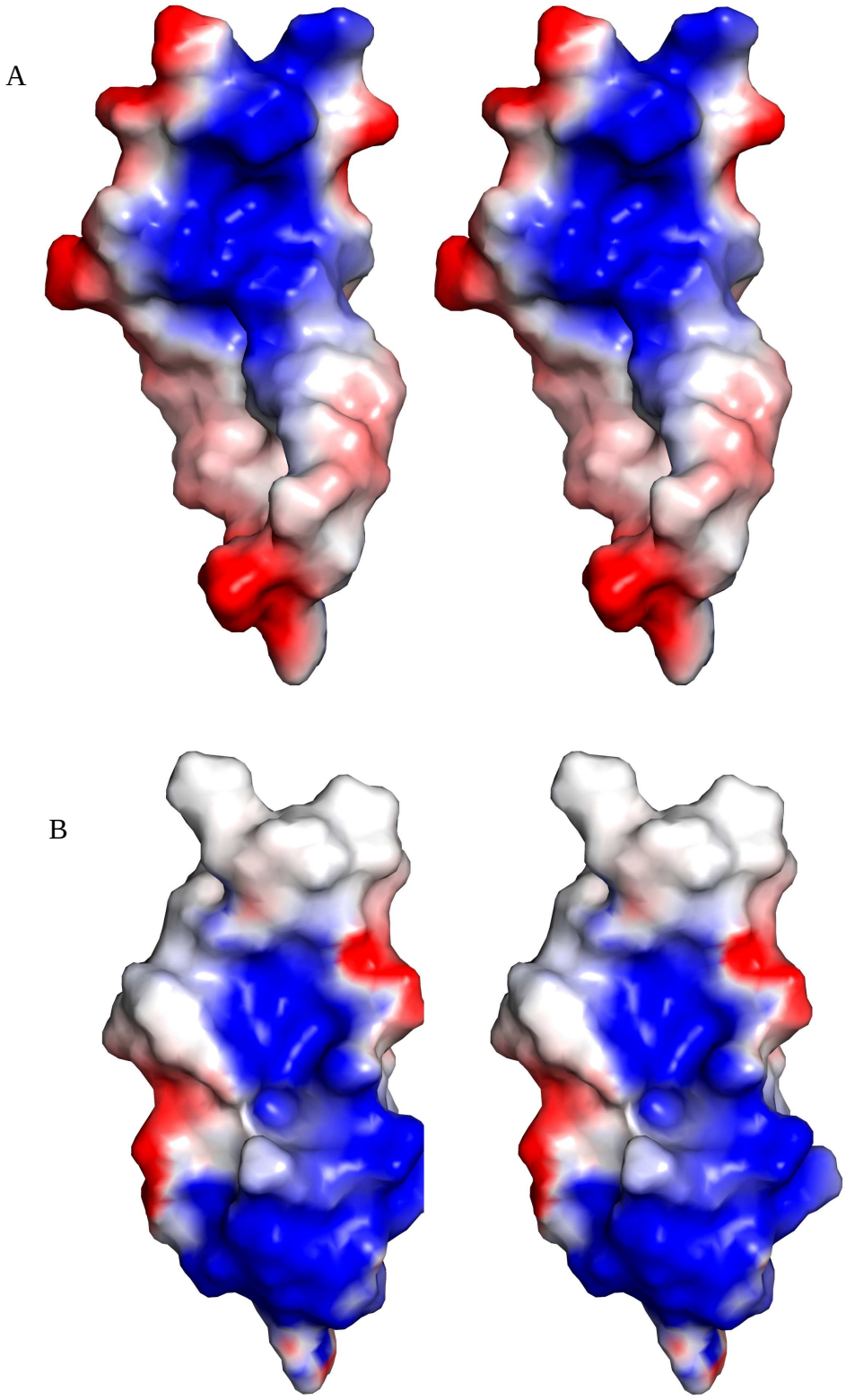


Figure 5.9 Stereo representation of the electrostatic surface potential of the hMBD1 CXXC1 (A) and hMLL CXXC (B). Coloured using a linear colour ramp from -75.0 KT hMLL (red) to +75hMLL (blue). (Figure prepared using the APBS plugin in PyMol)

5.4 Comparison of the loop region of hMBD1 CXXC1 and hMLL CXXC

Currently the RCSB protein data bank contains three CXXC domain structures which are those of the hMLL CXXC domain (Allen *et al*, 2006., Cierpicki *et al*, 2008) and the DNMT1 CXXC (Song *et* al, 2010). Figure 5.10 illustrates the structural superposition of hMBD1 CXXC1 with hMLL CXXC (PDB ID 2J2S) that has a pair wise RMSD of 1.83 Å over the well defined regions of hMBD1 CXXC (residues 175-196 & 208-217). The loop region residues 197-207 is not well defined in the hMBD1 CXXC1 ensemble due to a lack of long range NOEs, whereas the loop region in hMLL CXXC is well defined in the NMR ensemble.

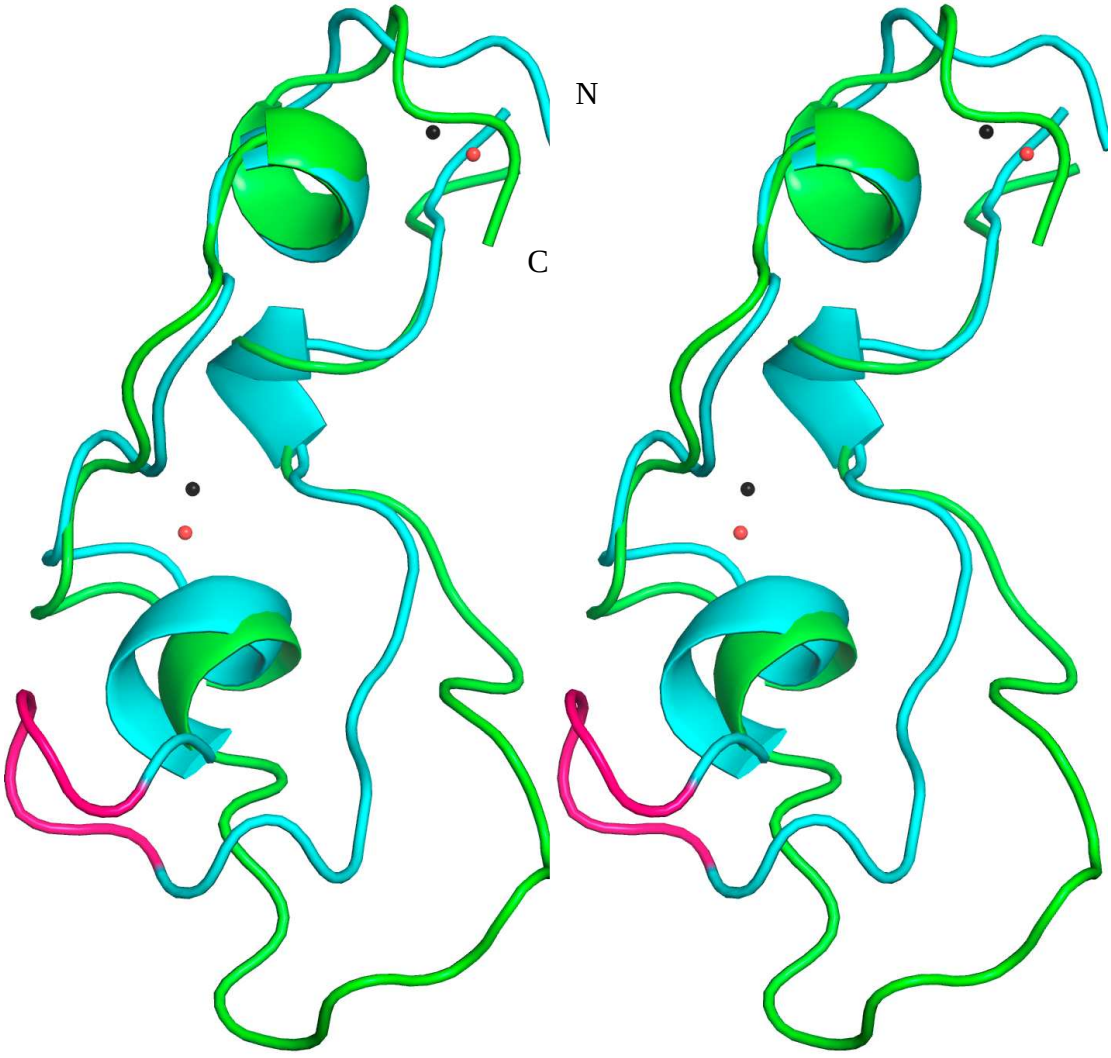


Figure 5.10 Structural superposition of hMBD1 CXXC1 R169-S221 (blue) and hMLL CXXC S1152-W1196 (green), PKFGG loop (pink); red spheres hMBD1 CXXC zinc; black spheres hMLL CXXC zinc. N, N-terminal; C, C-terminal (Figure prepared using PyMol)

A motif conserved in CpG DNA binding CXXC domains but not present in hMBD1 CXXC1 (or hMBD1 CXXC2) is the KFGG motif (figure 5.11). In hMBD1 CXXC1 the corresponding sequence is PHDV which has a significant effect on the structure of the loop region since compared with the phenylalanine the histidine is less hydrophobic and partially positively charged it is not able to be tightly packed into the main chain. The kink in the backbone (figure 5.12) due to the presence of a proline preceding the H200 does not allow the hydrophobic side chain to pack up against β protons of cysteine 194, where as the proline in hMLL CXXC is two residues before the phenylalaine. In DNMT1 CXXC the KFGG sequence is preceded by a lysine rather than the proline in the MLL CXXC domain.

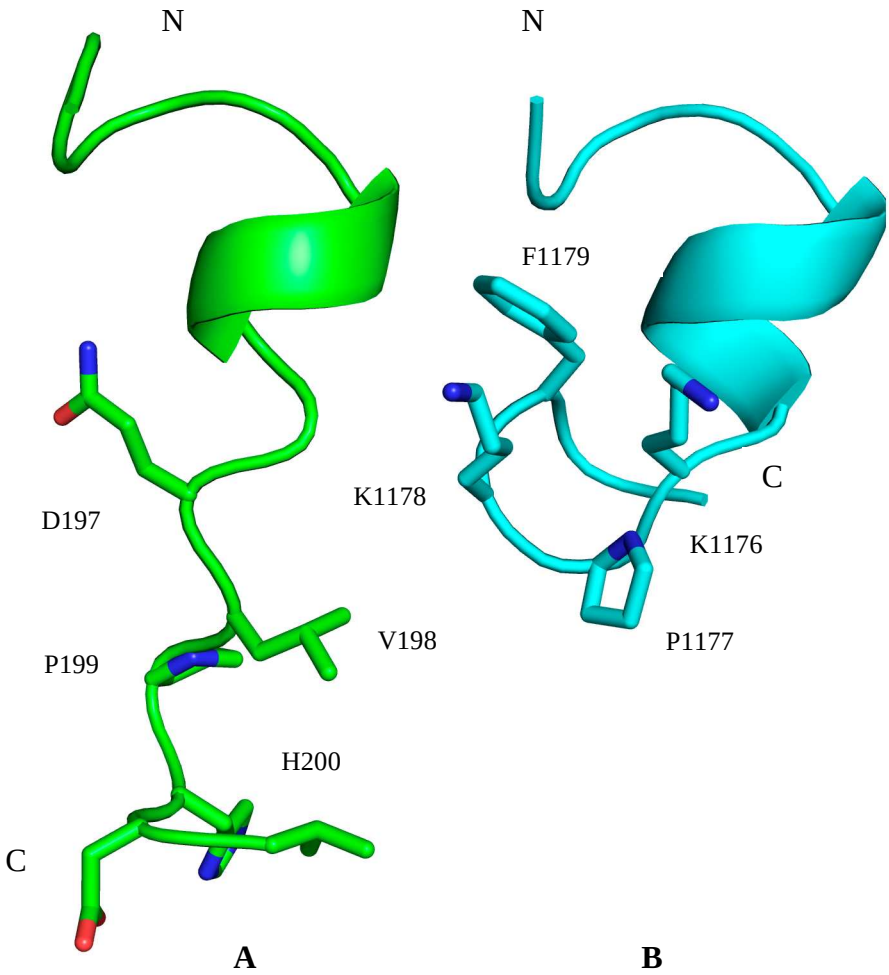


Figure 5.11 Comparison of second CXXCXXC motif & N-terminus of the loop region of hMBD1 CXXC1 C188 to V202 (A) and hMLL CXXC C1176 to G1181 (B) residues. N, N-terminal; C, C-terminal (Figure prepared using PyMol) QLPHD

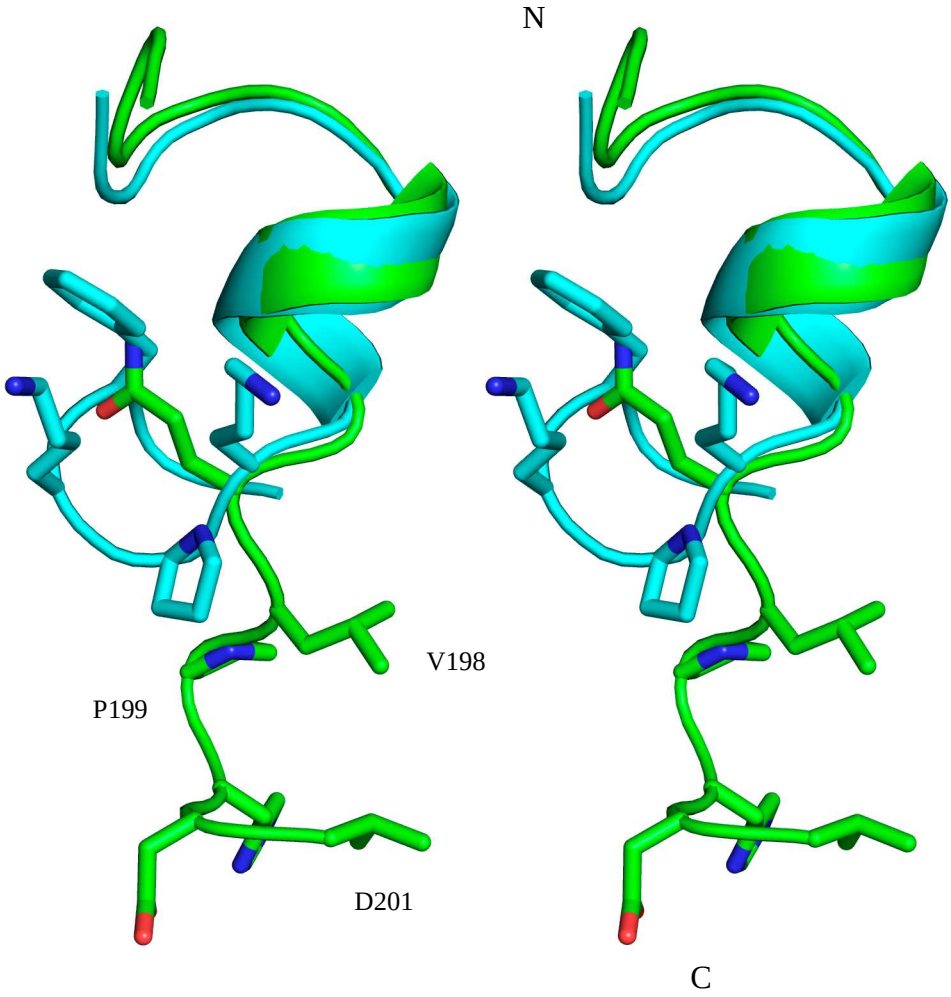


Figure 5.12 Stereo representation of the structural superposition of hMBD1 CXXC1 C188-V202 and hMLL CXXC C1176-G1181. N, N-terminal; C, C-terminal (Figure prepared using PyMol)

The solution structure ensemble of xDNMT1 CXXC (Thomson & Smith, unpublished) has a less well defined loop region than hMLL CXXC, therefore proline 1177 maybe required for loop stability. Mutation of hMLL CXXC residues K1178 or F1179 to alanine significantly reduced the ability to bind a 12bp single CpG DNA in an EMSA (Allen *et al*, 2006). These mutations of hMLL CXXC indicate that the precise, and ordered structure of this loop be required for efficient DNA binding.

5.5 Conclusion

The solution structure of hMBD1 CXXC1 presented in this thesis shows both similarities and differences to the known structures of hMLL CXXC (Allen *et al*, 2006) and DNMT1 CXXC (Song *et* al, 2010., Thomson & Smith, unpublished). The two zinc coordinating cysteine clusters of the CXXCXXC domain form a common structure shared with both hMBD1 and DNMT1 CXXC. This fold is highly likely to be common

to other CXXCXXC domains.

However, hMBD1 CXXC1 lacks the KFGG motif which is present in the non-methyl CpG binding CXXCXXC domains. The KFGG motif forms a small helix after the second CXXCXXC motif which appears to stabilise the DNA binding loop. Mutations in this region have been shown to reduce DNA binding (Allen *et al*, 2006). hMBD1 CXXC1 also differs from other CXXCXXC domains by lacking residues important for DNA binding. These differences not only change the surface charge of the loop thus not attracting the DNA through charge-charge interaction but lack the side chains to hydrogen bond to the DNA bases. The function of hMBD1 CXXC1 is currently unknown although it has been shown to bind residues 250-337 of Ring1b (Min *et al*, 2003). Ring1b is a component of the Polycomb group (PcG) multiprotein PRC1 complex (Min *et al*, 2003) which is required to maintain the transcriptionally repressive state of many genes. Since hMBD1 CXXC3 does not interact with Ring1b that suggests the difference in the loop region of hMBD1 CXXC1 may be involved in the protein interaction with ring1b.

CHAPTER SIX

NMR STUDY OF hMBD1 CXXC12

6.1 Overview

The first two MBD1 CXXC domains lack the consensus sequence DXXKFGG and have been shown not to bind CpG DNA (Jorgensen *et al*. 2004). The function of this pair of domains is still unknown although it has been shown to bind AA 250-337 of Ring1b (Sakamoto *et al*, 2007) suggesting a role in recruitment of the polycomb repressor complex 1 (PRC1) to DNA for maintenance of transcriptional repression. hMBD1 CXXC12 has also been shown to interact with the HMT SETDB1 and CAF150 in association with the MBD and TRD (Sarraf & Stancheva, 2004)

With the structure of the first MBD1 CXXC domain known, I wish to ask wither the presence of the second MBD1 CXXC domain affects the 3D structure of hMBD1 CXXC1? Should the two CXXC domains interact, we would expect to see chemical shift perturbation for residues whose environment has been affected by the presence of the second CXXC domain. Changes in the extreme C-terminal region of CXXC1 are also likely since this area should be more structured in the presence of CXXC2. There may also be chemical shift changes for resonances assigned to the loop or well defined regions of CXXC1 if the CXXC2 domains interact more extensively.

6.2 Cloning, expression of hMBD1 CXXC12

Using the previously described pGEX 6P1 expression and purification system, PCR amplification of DNA containing the CXXC1 & CXXC2 of hMBD1 was performed using the primers shown in table 6.1. The resultant fragment was cloned into pGEM T-easy vector and subcloned into pGEX 6P1 as previously described in chapter 2. The PCR (Figure 6.1), cloning into pGEM T-easy (Figure 6.2) and subcloning into pGEX 6P1 (Figure 6.3) are shown below. Plasmids were sequenced to make sure no errors were present.

hMBD1 CXXC1 fwd	5’ G**GGATCC**GAGCAGAGAATGTTTAAG 3’
hMBD1 CXXC2 rev	5’ C**CTCGAG**TTCAGCGGGCATGTTTA 3’

Table 6.1 Primers used for the PCR amplification of hMBD1 CXXC12. Restriction enzyme sites shown in bold & stop codon underlined.

The amino acid sequence of the expressed hMBD1 CXXC12 construct is shown in table 6.2 which has a predicted molecular weight of 11.8 kDa. Based on the structured region of hMBD1 CXXC1 it is likely the two domains are separated by a small linker sequence.

GSEQRMFKRVGCGECAACQVTEDCGACSTCLLQLPHDVASGLFCKCERRRCL RIVERSRGCGVCRGCQTQEDCGHCPICLRPPRPGLRRQWKCVQRRCLRGKHAR

Table 6.2 Amino acid sequence of the cloned hMBD1 CXXC12 domain. The structured domain of CXXC1 is shown in red, the CXXC2 sequence shown in blue with the putative linker sequence underlined.

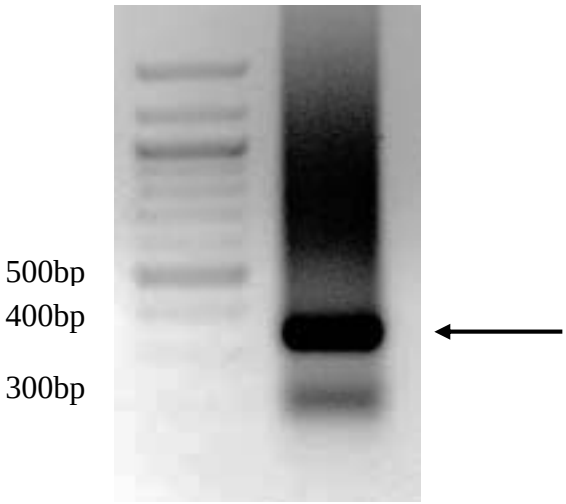


Figure 6.1 Agarose gel electrophoresis image hMBD1 CXXC12 PCR product. M – 100bp DNA marker NEB (N3231), 1 – hMBD1 CXXC12 arrow indicated band of interest

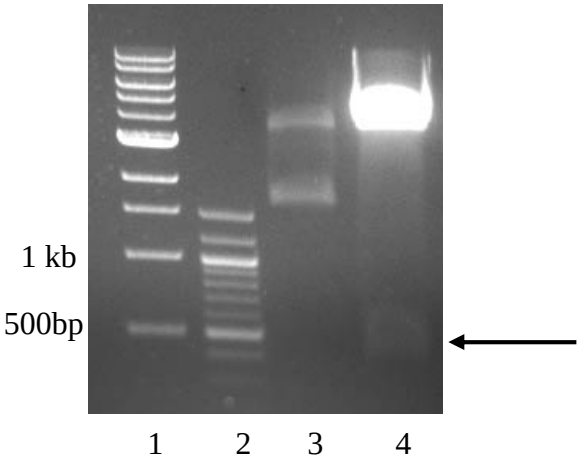


Figure 6.2 Agarose gel electrophoresis of a selection of 2 putative pGEM-T hMBD1 CXXC12 clones digested with BamH1 and Xho1. 1 - 1KB ladder NEB, 2 - 100bp ladder NEB, 3-4 putative pGEM-T hMBD1 CXXC12 clones. Arrow indicates band of correct size for CXXC12 insert. Lane 4 show the expected fragment pattern

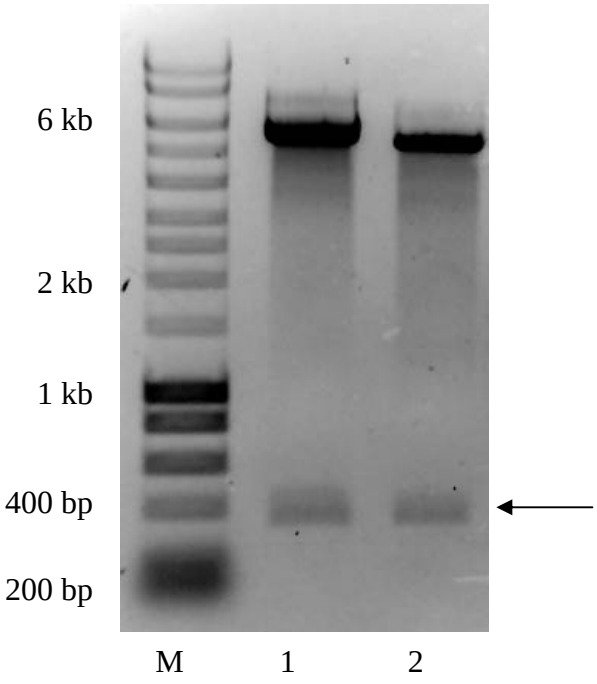


Figure 6.3 Agarose gel electrophoresis of a selection of 2 putative pGEX6P1-hMBD1 CXXC12 clones digested with BamH1 and Xho1. The arrow indicates the expected position of the CXXC12 fragments. Clones 1 & 2 show the expected fragment pattern. M- Bioline hyperladder 1

6.3 Purification of hMBD1 CXXC12

Purification of hMBD1 CXXC12 was carried out using the method described in chapter 2. Figure 6.4 illustrates the purification of hMBD1 CXXC12. The preparation of hMBD1 CXXC12 was hampered by precipitation of the purified protein during overnight cleavage of the tag. The protein sample suffered further precipitation during buffer exchange and concentration thus making it difficult to produce sufficient quantity for isotopic labeling in an economic way.

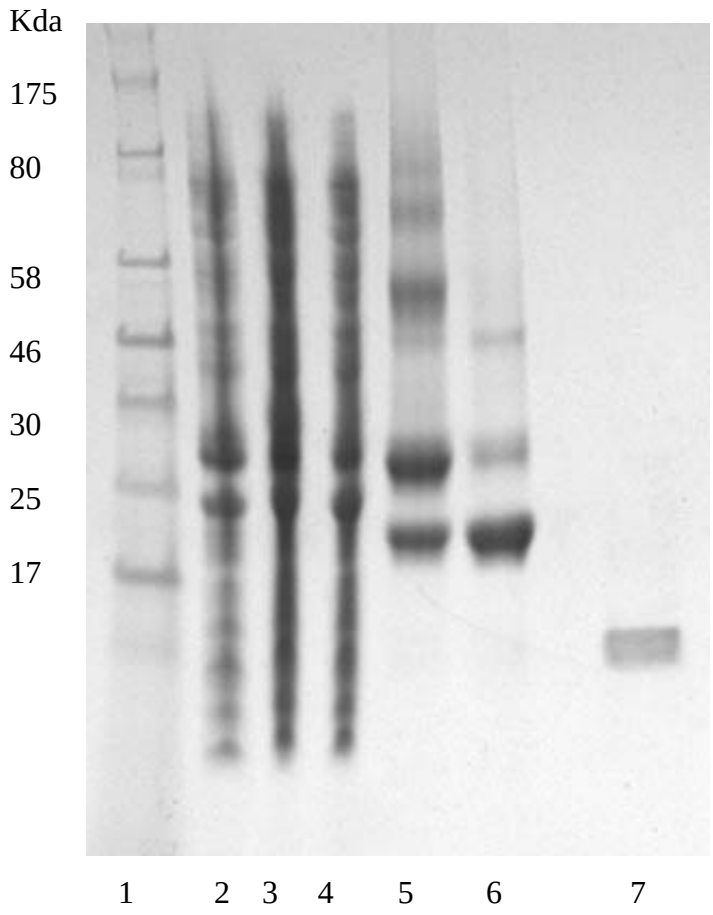


Figure 6.4 SDS-PAGE analysis of the expression and purification of hMBD1 CXXC12 domain expressed in Tuner^{TM} DE3 cells. 1 – NEB prestained protein marker (cat. No. P7708S) , 2 – 5µl whole cell fraction, 3 – 5µl of soluble fraction, 4 – 5µl of flow through, 5 – 20µl of resin before Pre-Scission protease treatment, 6 – 20 µl resin after Pre-Scission protease treatment, 7– 20µl elution of hMBD1 CXXC12 domain.

6.4 NMR studies of hMBD1 CXXC12

During long NMR experiments hMBD1 CXXC12 would degrade & precipitate. This coupled with the lack of economic viability to express isotopically labelled protein meant that at this time it was only possible to record homonuclear experiments. Shown in table 6.3 are the NMR experimental parameters for the 2D ^{1}H NOSEY experiment7

	F1			F 2					
Experiment	Nuc	TD	SW	Nuc	TD	SW	RDms	NS	WS
100ms NOSEY	^{1}H	2048	8503	^{1}H	1700	7001	1.0	32	ES

Table 6.3 Acquisition parameter of NMR experiment on hMBD1 CXXC12 (Bruker AVANCE 600 MHz) used for resonance assignment and structural restraints collection. Nuc: Nucleus, TD: time domain points (complex), SW: sweep width in Hz, RD: relaxation delay, NS: number of scans, WS: water suppression, ES: excitation sculpting, ms: milliseconds,

6.4.1 Is hMBD1 CXXC12 structured?

To asses the quality of the CXXC12 sample, a 1D ^{1}H experiment was carried out to investigate if the purified hMBD1 CXXC12 protein was structured. The recorded 1D ^{1}H spectrum of hMBD1 CXXC12 is shown in figure 6.5

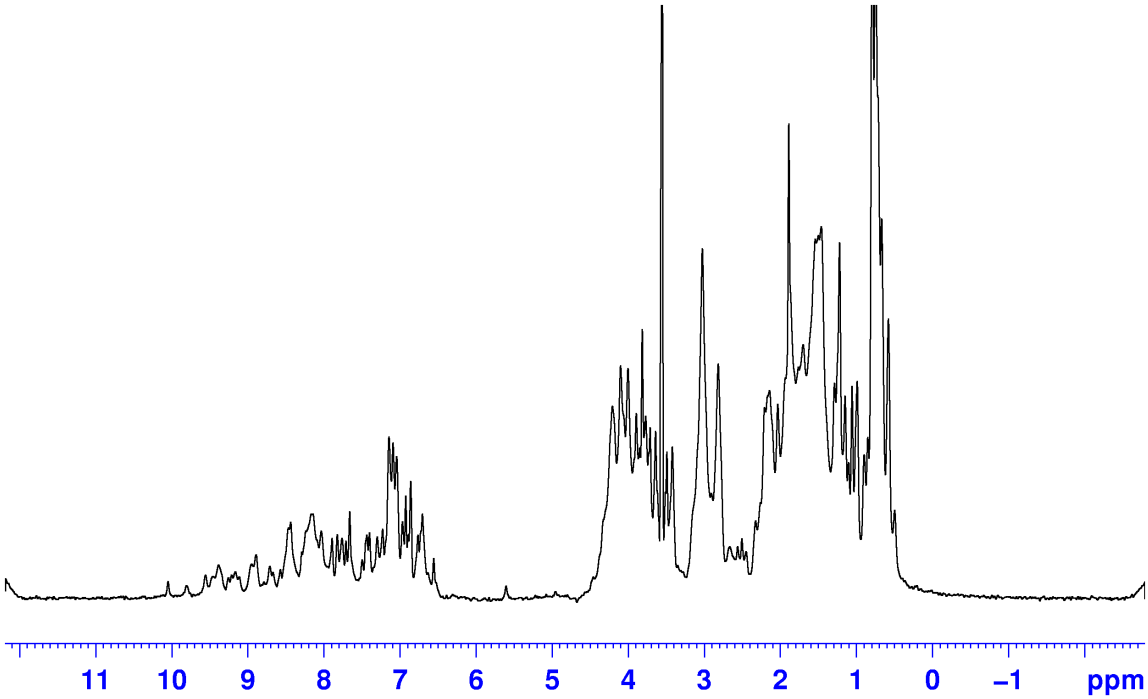


Figure 6.5 1D ^{1}H spectrum of unlabeled hMBD1 CXXC12 (10mM dTris, 250mM NaCl, pH 7.5, 293K). Water suppression using the method of Hwang and Shaka (1995)

The 1D ^{1}H spectrum of unlabeled hMBD1 CXXC12 suggests that the protein sample is folded due to the overall dispersion of chemical shifts and the ^{1}H line widths. The region of the spectrum around 10 ppm shows peaks clearly not present in the 1D ^{1}H spectrum of CXXC1 while the peak around 0.5 ppm is still present. Also the 6.5-8.5 ppm region for CXXC12 is not as intense compared to the same region in CXXC1 suggesting that fewer residues are in unstructured highly dynamic conformations.

6.4.2 Is CXXC1's structure affected by the presence of CXXC2?

A 2D homonuclear NOESY experiment was carried out in order to compare the spectrum with the 2D NOESY of hMBD1 CXXC1. Figure 6.6 shows the Hα-H_N region of hMBD1 CXXC12 & CXXC1 with the overlaid H_N-H_N region in figure 6.7. Since the sample precipitated during the course of the experiment further 2D spectra that would present a more complete analysis could not be recorded.

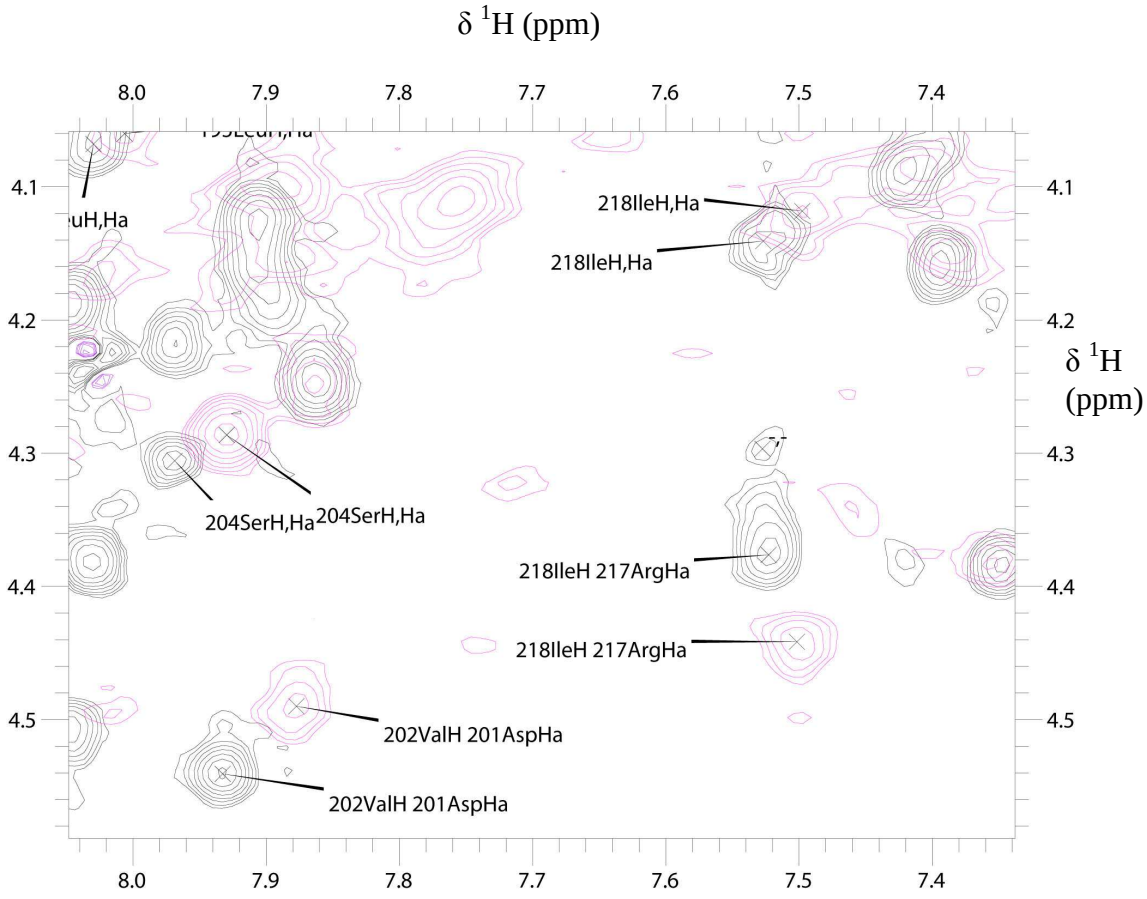


Figure 6.6 H_N-Hα region of the 2D homonuclear spectra of hMBD1 CXXC1 (black) and hMBD1 CXXC12 (pink) overlaid (10mM dTris, 250mM NaCl, pH 7.5, 293K). Examples of assigned crosspeaks which have changed chemical shifts are indicated.

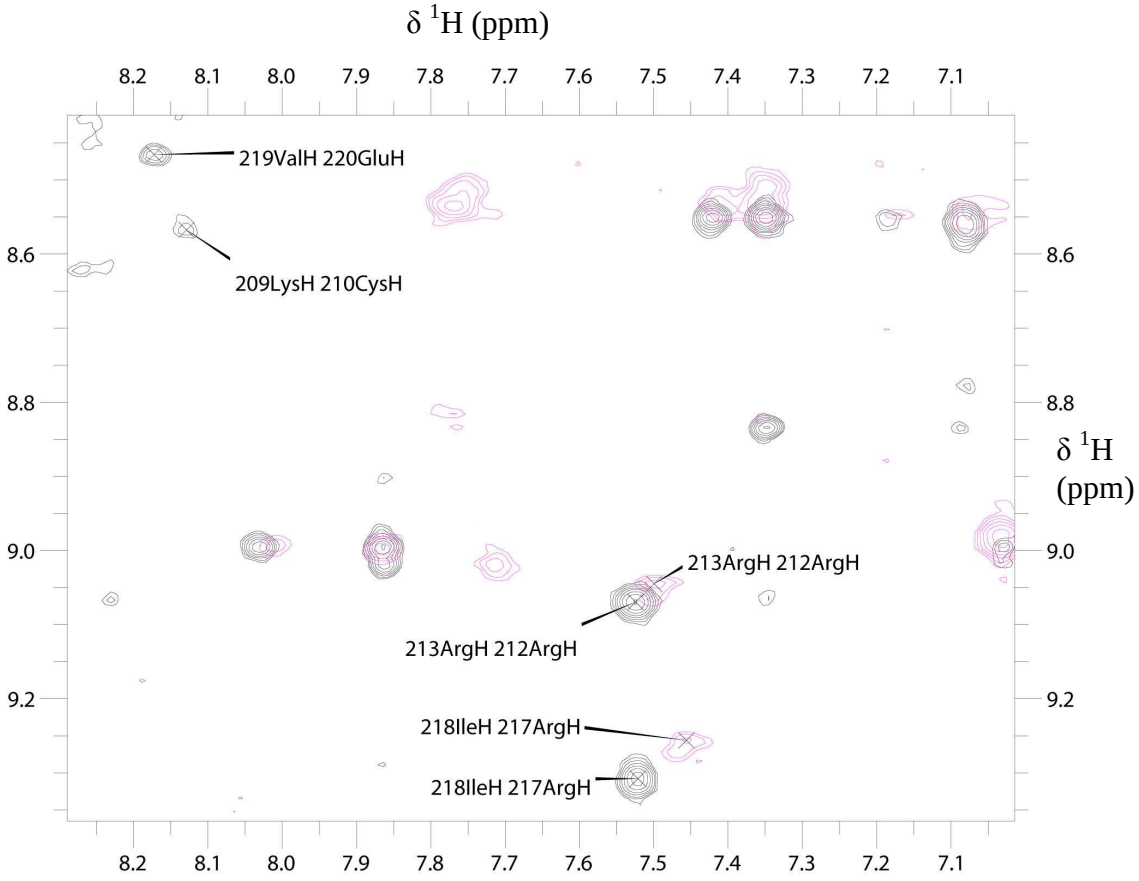


Figure 6.7 H_N-H_N region of the 2D homonuclear spectra of hMBD1 CXXC1 (black) and hMBD1 CXXC12 (pink) overlaid (10mM dTris, 250mM NaCl, pH 7.5, 293K). Examples of assigned crosspeaks which have changed chemical shifts are indicated.

However, from a comparison of the CXXC12 & CXXC1 spectra it is clear that certain CXXC1 amino acids have altered chemical shift in CXXC12. The most notable changes are seen in the dynamic loop region of CXXC1with amino acids S204 and P207, while others such as C179, A190 & T193 show no chemical shift change. These chemical shift changes indicate a change in environment of the loop region suggesting that it is possibly in close proximity to the second CXXC domain. In the loop region there are changes between the two constructs with the Hβ protons of C208 no longer correlated to the Hγ of T193 in CXXC12 while previously strong cross peaks have reduced intensity e.g. E211 γ protons correlation to C191 β protons. These NOE changes in the loop region suggest the second CXXC domain is changing the shape of the loop region. Residues C179, A190, & T193 are found within the zinc clusters and seem to be unaffected by the presence of the second CXXC domain. Residues at the extreme C-terminus, amino acids 219-222, are unstructured when hMBD1 CXXC1 is expressed on

its own. However, in hMBD1 CXXC12, their NOE cross peaks shift or disappear. Certain clearly identifiable crosspeaks in the CXXC1 2D NOESY such as the Hγ-H_N of E220 and the sequential H_N - H_N & H_N -Hα cross peaks for residues V219-R222 are absent or significantly shifted in the CXXC12 spectrum. The lack of these NOE crosspeaks means that the structure of this region has changed in the CXXC12 construct.

6.5 Conclusions

Initial NMR experiments have shown that the expressed hMBD1 CXXC12 domains are structured in solution. However, to carry out 3D NMR experiments to facilitate assignment and eventually structure calculations, protein purification conditions must be further optimized. Changes to pH are limited with the use of PreScission protease due to a narrow functional pH range, but NaCl concentration can be adjusted up to 1 M. Therefore, it would be possible to significantly alter the salt concentration of the purification and NMR buffer in an attempt to prevent precipitation of the protein.

Although the NOE data recorded is not sufficient to fully assign and calculate the structure of hMBD1 CXXC12, it is possible to suggest the relative orientation of the two CXXC domains. There are significant NOE differences in the loop region (201N-209K) in hMBD1 CXXC12 to suggest there is a change in environment. This may be due to the close proximity of the second CXXC domain. In hMBD1 CXXC1 residues 216L-222S are poorly constrained resulting in a flexible free end. Therefore, it is unsurprising that there are chemical shift changed in this region since in hMBD1 CXXC12 they form a small interconnecting bridge between the two domains. Due to the observed NOE changes it is likely that the two domains are folded together in one structure rather than two independent domains. In order to verify this theory, recording ^{15}N relaxation data for both constructs would provide information as to the relative dynamics of the two domains domain. If the structure of CXXC12 was as suggested by the homonuclear NOE data then the ^{15}N relaxation parameters recorded for each of the two domains would be quite similar and the T1/T2 ratios for CXXC12 should be indicative of an ~12kDa globular species rather than independently tumbling ~6kDa species. Should the precipitation problem with hMBD1 CXXC12 be overcome it will be possible to isotopically label the protein and carry out 3D NMR experiments. It will also be possible to investigate the protein-protein interactions with amino acid 250-337 domain of ring1b (Sakamoto *et al*, 2007) with MBD1 CXXC12 once the domain is isotopically labelled.

Appendix A Laboratory buffers and methods

Bacterial Growth Media

Appropriate selection antibiotics were added in all media prior to use.

LB Medium	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% NaCl (w/v)
5XM9 stock	3.4% (w/v) Na_2HPO_4, 1.5% KH_2PO_4, 0.25% (w/v) NaCl, autoclaved prior to use
M9 minimal medium	for 500 ml; 100 ml 5XM9, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.3% D-glucose (w/v), 8 mM $(NH_4)_2SO_4$, 0.004% thiamine (w/v)
Labelled medium	For ^{15}N labelling, replace 8 mM $(NH_4)_2SO_4$ with 16 mM $^{15}NH_4Cl$ (Isotec, 11186AE).

GST purification buffers

Lysis/binding buffer	10× PBS: 1.4 M NaCl, 27 mM KCl, 100 mM Na_2HPO_4, 18 mM KH_2PO_4, pH 7.3. Use at 1x
PreScission buffer	50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM DTT
Elution buffer	300 mM NaCl, 2.7 mM KCL, 10 mM Na_2HPO_4, 1.8mM KH_2PO_4, 1 mM DTT, pH 7.3

Bradford assay

The protein concentration was measured using Bradford assay kit brought from Pierce (Product No. 23236). The measurement was performed according to the standard protocol in the manufacturer's instructions.

SDS-PAGE

Coomassie Stain	45% methanol, 45% dH_2O, 10% acetic acid, 0.25% (w/v) coomassie R250
Destain Buffer	5% methanol, 10% acetic acid, 85% dH_2O

Appendix B Ramachandran plots, $\chi_1$torsion angle distributions and average secondary structure for the final ensemble of hMBD1 CXXC structures. Structural analysis by Procheck-NMR

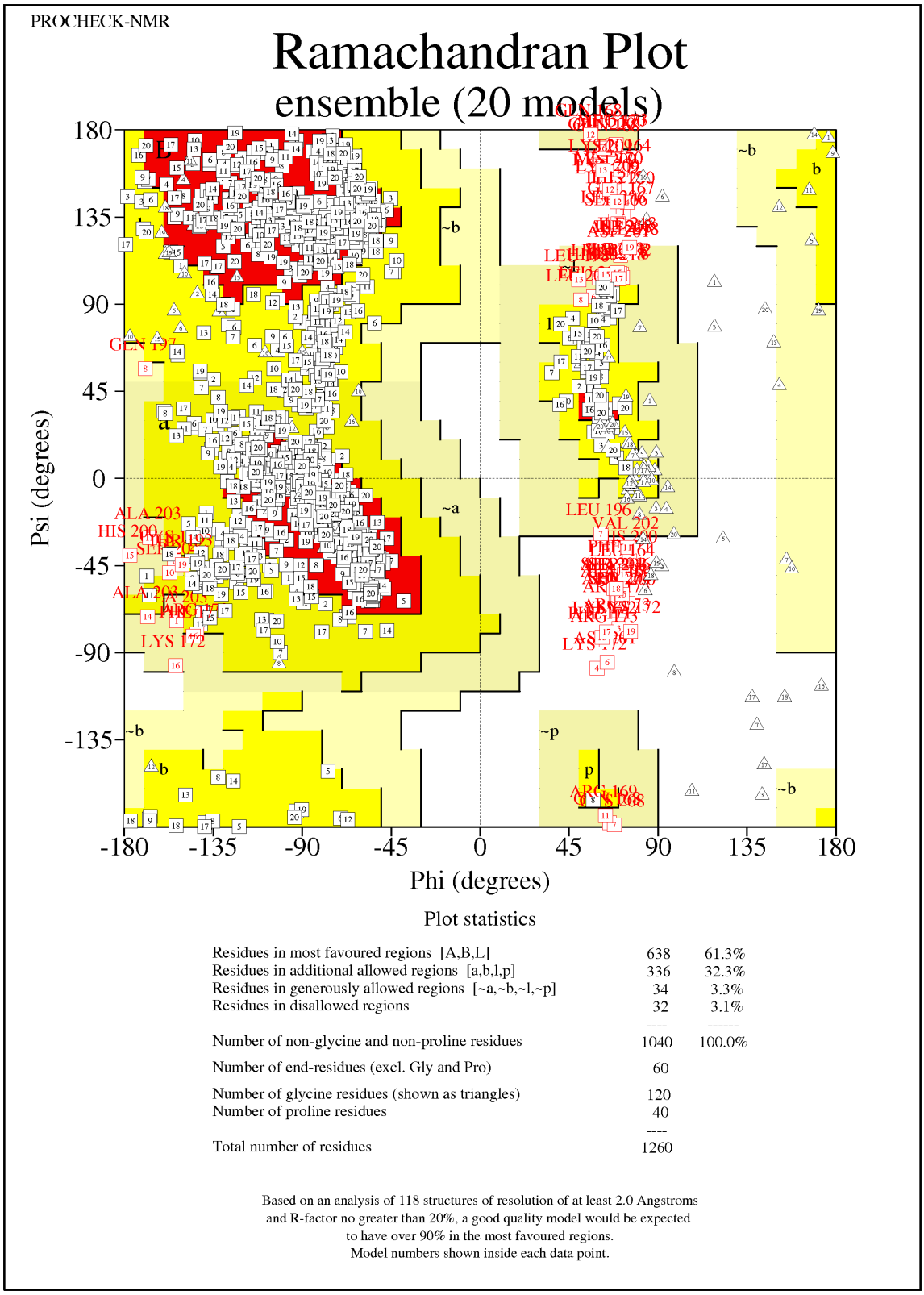


Figure B.1 Graphical summary of Ramachandran statistics for the final ensemble of 25 hMBD1 CXXC structures.

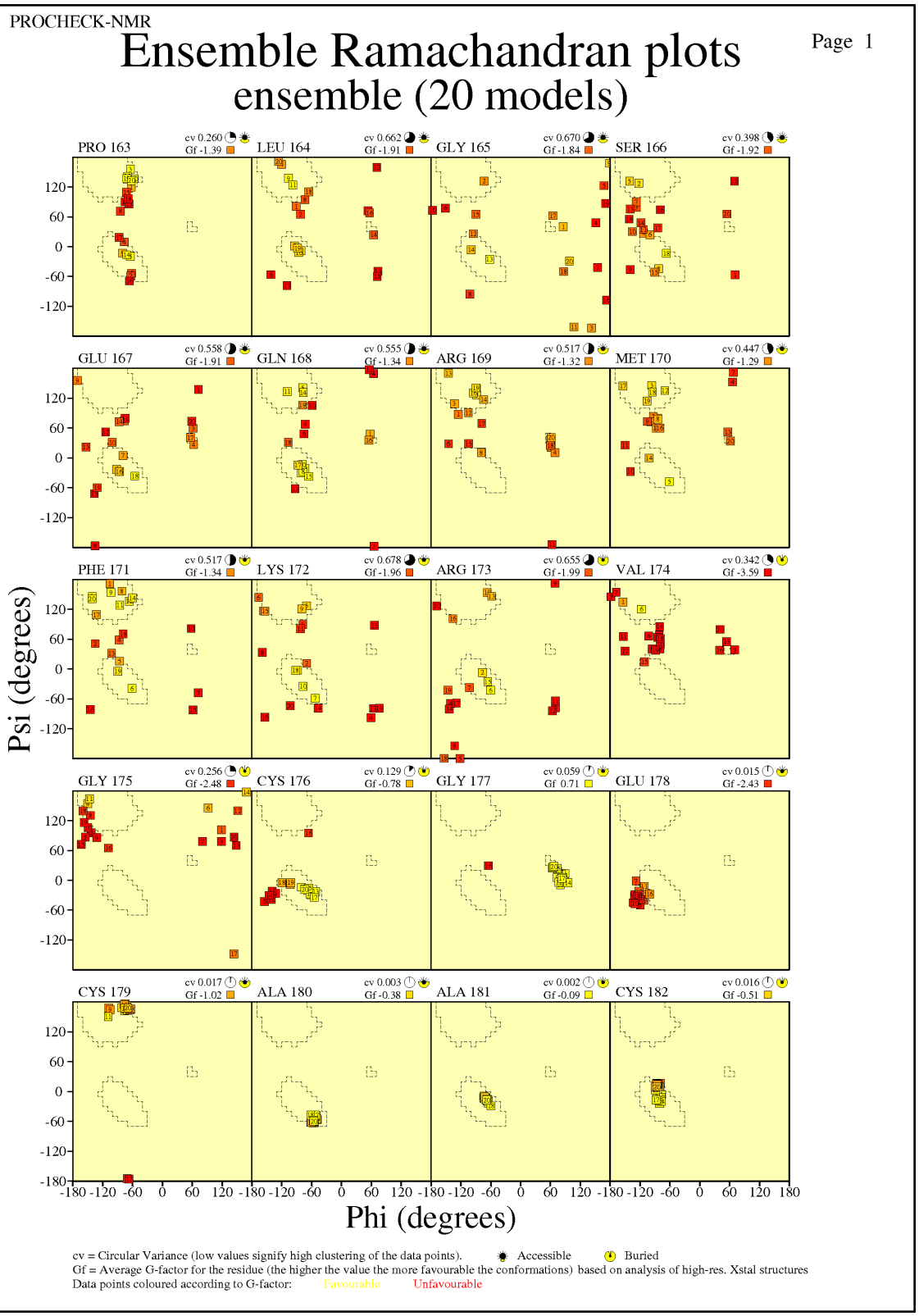


Figure B.2 Per-residue Ramachandran plots for the final ensemble of 20 hMBD1 CXXC1 structures. Yellow boxes indicated favourable and red boxes indicate unfavorable φ, ψ dihedral angle combinations.

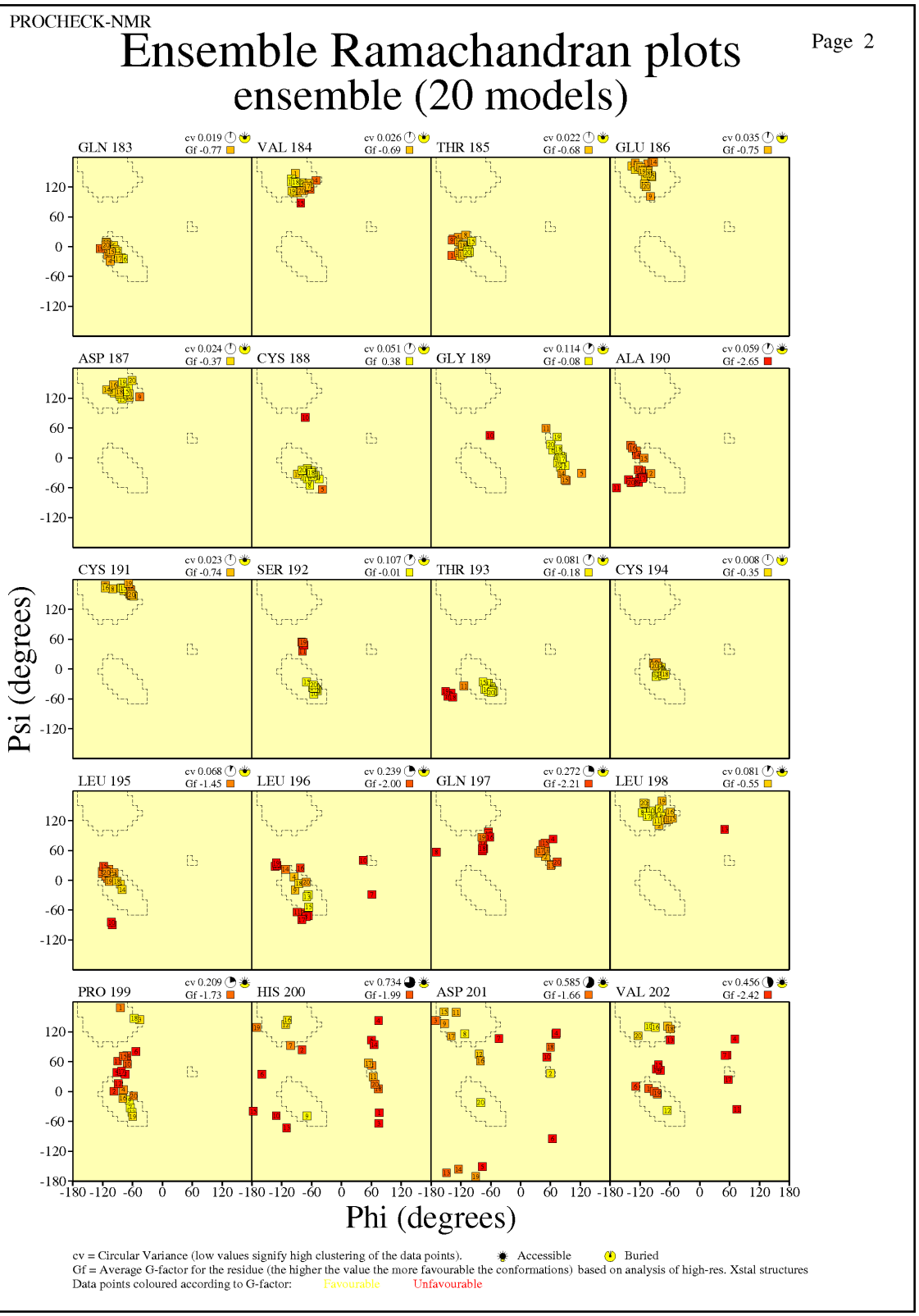


Figure B.2b Per-residue Ramachandran plots for the final ensemble of 20 hMBD1 CXXC1 structures. Yellow boxes indicate favourable and red boxes indicates unfavorable φ, ψ dihedral angle combinations respectively.

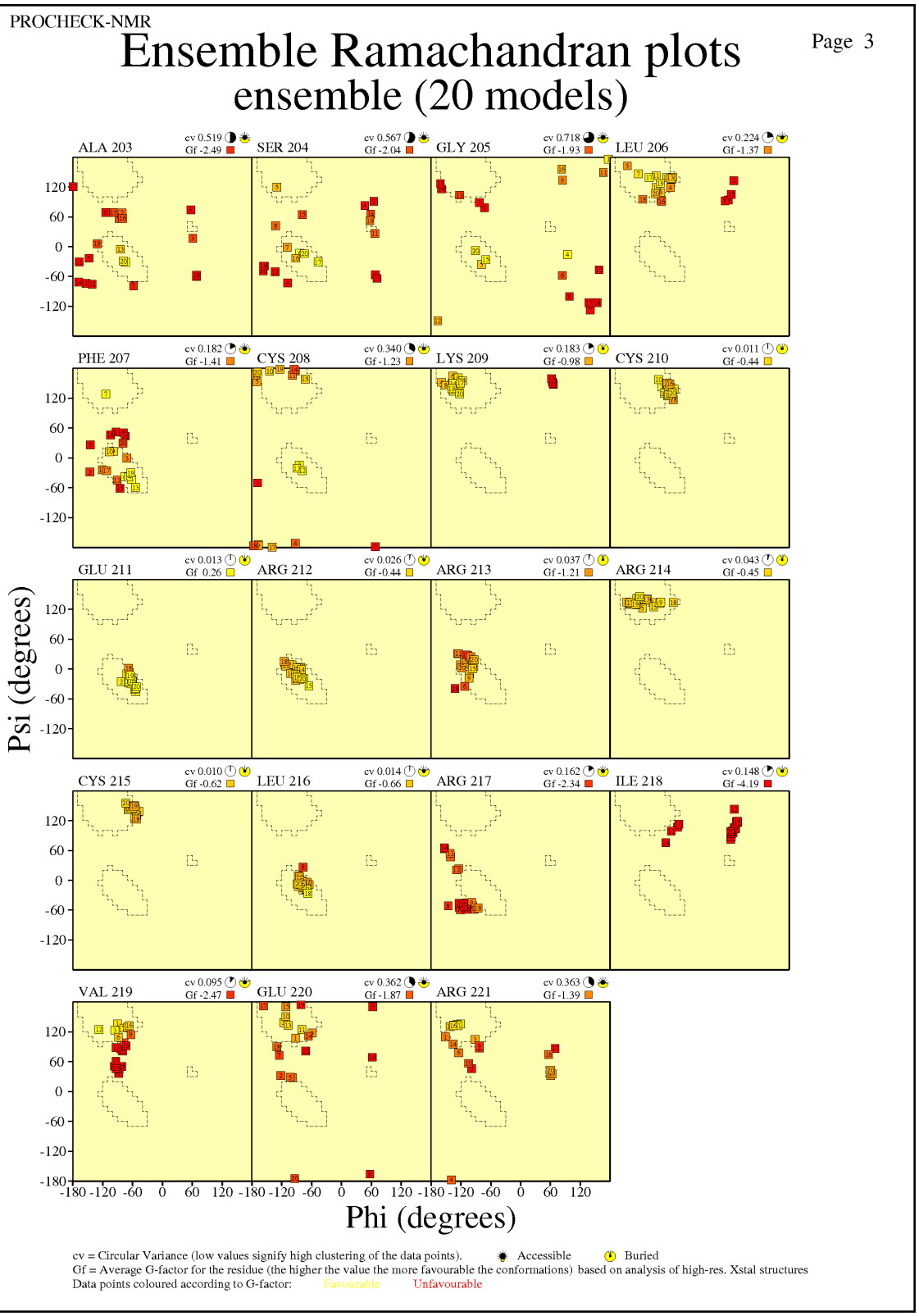


Figure B.2c Per-residue Ramachandran plots for the final ensemble of 20 hMBD1 CXXC1 structures. Yellow boxes indicate favourable and red boxes indicates unfavorable φ, ψ dihedral angle combinations respectively.

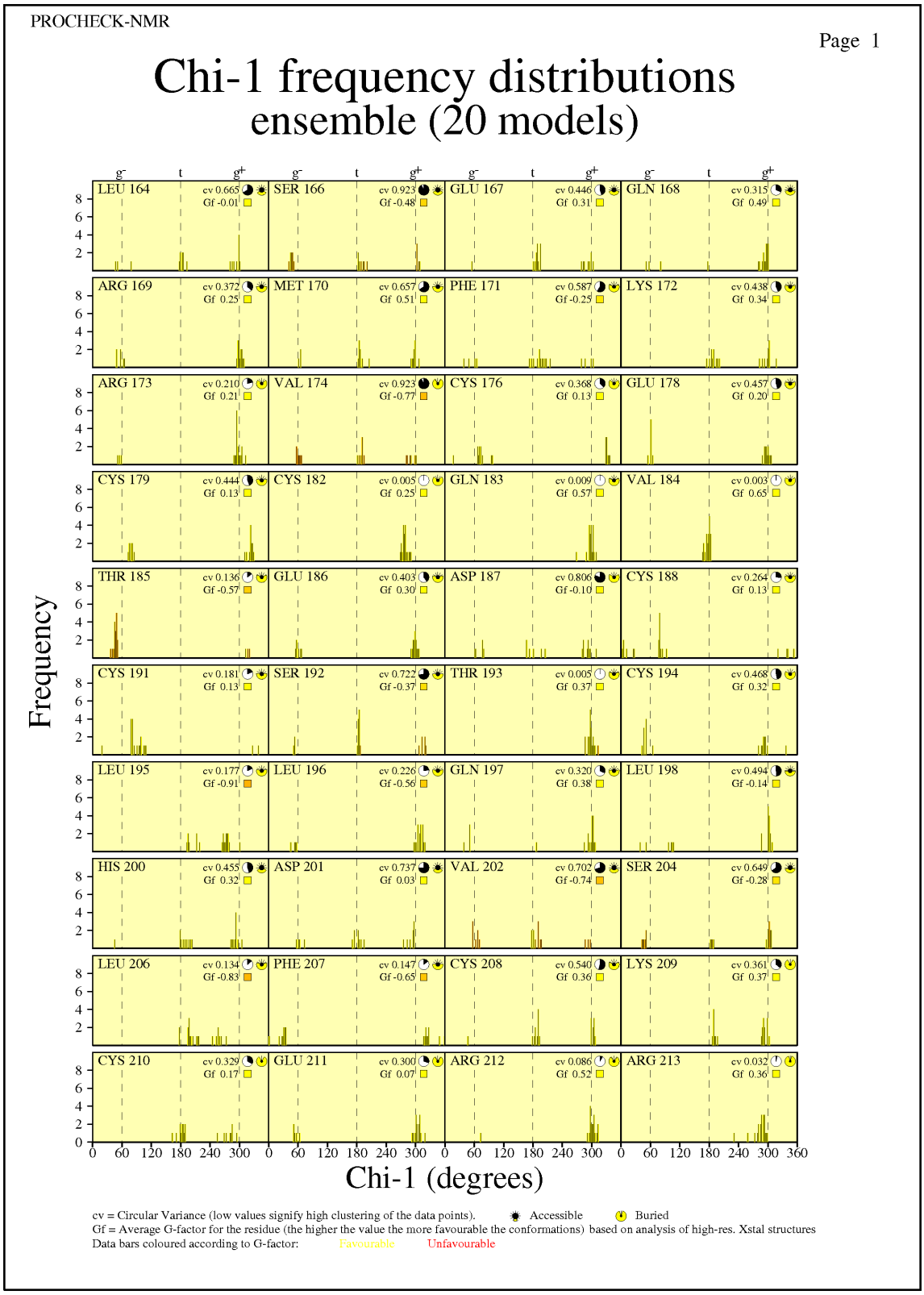


Figure B.2d Per-residue X1 frequency plots for the final ensemble of 20 hMBD1 CXXC1 structures. Yellow boxes indicated favorable and red boxes indicates unfavourable φ, ψ dihedral angle combinations.

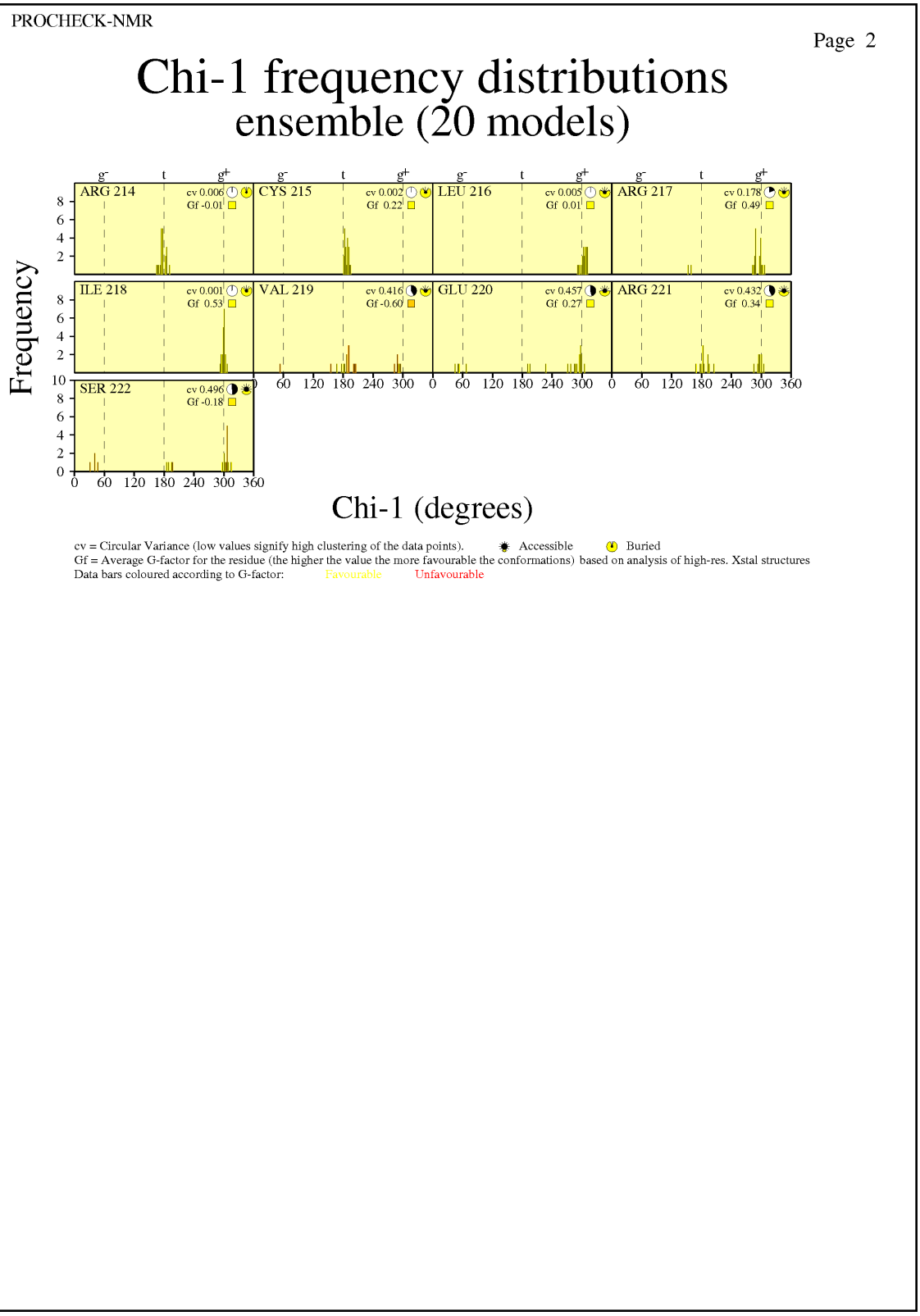


Figure B.2e Per-residue X1 frequency plots for the final ensemble of 20 hMBD1 CXXC1 structures. Yellow boxes indicated favorable and red boxes indicates unfavourable φ, ψ dihedral angle combinations.

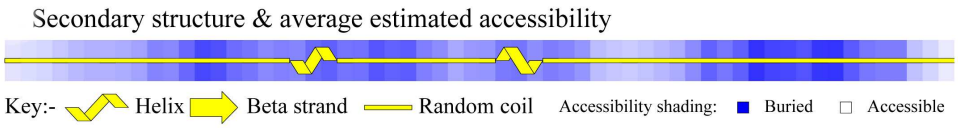


Figure B.2d Average secondary structure and accessibility plot for the final ensemble of 25 hMBD1 CXXC1 structures

Appendix C Chemical shift assignments of hMBD1 CXXC1. Chemical shifts are given in p.p.m, no chemical shift indicates unassigned or unobserved resonances

Residue	H	N	NE	NZ	NH	HA	HB	HG	HD	HE	HZ
163Pro	-	-	-	-	-	4.46	1.92,2.30	1.99,1.99	3.56,3.56		
164Leu	8.6	122.72	-	-	-	4.34	1.57,1.59	1.66	0.87,0.91	-	-
165Gly	8.56	110.52	-	-	-	3.95,4.05	-	-	-	-	-
166Ser	8.24	123.15	-	-	-	4.35	3.82,3.89	-	-	-	-
167Glu	8.62	122.8	-	-	-	4.2	1.95,2.02	2.26,2.20	-	-	-
168Gln	8.28	120.76	112.84	-	-	4.2	1.96,2.07	2.34,2.34	-	7.59,6.90	-
169Arg	8.24	121.54	-	-	-	4.21	1.73,1.73	1.51,1.57	3.14,3.14	-	-
170Met	8.19	-	-	-	-	4.38	1.91,1.92	2.47,2.39	-	-	-
171Phe	8.17	121.09	-	-	-	4.59	3.12,2.96	-	7.19	7.3	-
172Lys	8.15	122.49	-	-	-	4.21	1.75,1.67	1.37,1.31	1.63,1.63	2.93,2.94	-
173Arg	8.24	122.47	-	-	-	4.31	1.81,1.75	1.57,1.62	3.15,3.15	-	-
174Val	8.24	121.19	-	-	-	4.2	2.09	0.94,0.94	-	-	-
175Gly	8.34	111.29	-	-	-	4.06,4.14	-	-	-	-	-
176Cys	9.36	126.38	-	-	-	4.48	2.80,3.25	-	-	-	-
177Gly	9.55	115.33	-	-	-	3.83,4.35	-	-	-	-	-
178Glu	9.62	123.72	-	-	-	4.57	1.93,2.06	2.15,2.22	-	-	-
179Cys	7.12	120.34	-	-	-	4.7	2.98,3.46	-	-	-	-
180Ala	8.84	120.15	-	-	-	3.95	1.42	-	-	-	-
181Ala	7.39	120.51	-	-	-	4.4	1.28	-	-	-	-
182Cys	8.57	120.38	-	-	-	4.05	2.92,2.95	-	-	-	-
183Gln	7.44	114.45	111.92	-	-	4.11	1.88,1.87	2.34,2.25	-	7.02,6.69	-
184Val	7.21	122.55	-	-	-	3.92	2.25	1.19,1.02	-	-	-

Appendix C Chemical shift assignments of hMBD1 CXXC1. Chemical shifts are given in p.p.m, no chemical shift indicates unassigned or unobserved resonances

Residue	H	N	NE	NZ	NH	HA	HB	HG	HD	HE	HZ
185Thr	8.57	116.22	-	-	-	4.37	4.42	1.19	-	-	-
186Glu	7.1	119.02	-	-	-	4.6	1.96,1.77	1.99,2.16	-	-	-
187Asp	8.78	124.9	-	-	-	4.5	2.42,3.47	-	-	-	-
188Cys	9.39	127.21	-	-	-	4.47	3.29,2.84	-	-	-	-
189Gly	9.52	115.74	-	-	-	4.19,3.89	-	-	-	-	-
190Ala	9.95	127.56	-	-	-	4.61	1.35	-	-	-	-
191Cys	7.06	119.33	-	-	-	4.85	3.40,2.97	-	-	-	-
192Ser	9.01	114.78	-	-	-	4.81	3.99,4.03	-	-	-	-
193Thr	7.89	117.58	-	-	-	4.01	4.26	1.04	-	-	-
194Cys	9.01	126.91	-	-	-	3.96	2.98,2.92	-	-	-	-
195Leu	8.03	116.75	-	-	-	4.08	1.63,1.62	1.46	0.70,0.80	-	-
196Leu	7.42	119.34	-	-	-	4.17	1.87,1.87	1.5	0.84,0.92	-	-
197Gln	7.92	117.57	112.73	-	-	4.15	2.08,2.02	2.31,2.32	-	6.86,7.53	-
198Leu	8.32	122.95	-	-	-	4.52	1.60,1.41	1.64	0.90,0.91	-	-
199Pro	-	-	-	-	-	4.39	1.86,2.30	2.02,2.02	3.84,3.55	-	-
200His	8.45	-	-	-	-	4.51	3.10,3.10	-	8.16	7.11	-
201Asp	8.22	120.5	-	-	-	4.55	2.65,2.65	-	-	-	-
202Val	7.93	120.82	-	-	-	3.93	2.08	0.92,0.92	-	-	-
203Ala	8.31	124.63	-	-	-	4.23	1.39	-	-	-	-
204Ser	7.98	113.35	-	-	-	4.32	3.89,3.86	-	-	-	-
205Gly	8.13	110.16	-	-	-	3.87,3.96	-	-	-	-	-
206Leu	7.89	121.02	-	-	-	4.2	1.44,1.43	1.34	0.74,0.81	-	7.23
207Phe	8.07	119.02	-	-	-	4.53	2.97,3.12	-	7.18	7.29	-

Appendix C Chemical shift assignments of hMBD1 CXXC1. Chemical shifts are given in p.p.m, no chemical shift indicates unassigned or unobserved resonances

Residue	H	N	NE	NZ	NH	HA	HB	HG	HD	HE	HZ
208Cys	8.13	120.23	-	-	-	4.38	2.71,2.78	-	-	-	-
209Lys	8.15	123.48	-	-	-	4.19	1.58,1.65	1.36,1.37	1.58,1.58	2.90,2.90	-
210Cys	8.56	125.58	-	-	-	3.87	2.91,2.30	-	-	-	-
211Glu	9.31	112.13	-	-	-	3.8	2.27,2.10	2.48,2.70	-	-	-
212Arg	9.09	119.51	-	-	-	4.35	2.06,2.11	1.83,1.83	3.29,3.24	-	-
213Arg	7.54	114.39	83.8	-	-	4.66	1.42,2.38	1.91,1.29	3.11,3.11	6.83	-
214Arg	6.77	120.46	84.52	-	-	4.04	1.63,1.77	1.51,1.84	3.15,3.19	6.67	-
215Cys	8.23	126.72	-	-	-	3.89	2.98,2.76	-	-	-	-
216Leu	9.08	110.46	-	-	-	4.31	1.78,1.84	2.11	0.98,1.15	-	-
217Arg	9.33	121.68	-	-	-	4.38	1.95,1.81	1.69,1.58	3.31,3.28	-	-
218Ile	7.54	121.04	-	-	-	4.15	1.85	1.12,1.50,0.85	0.82	-	-
219Val	8.17	124.39	-	-	-	4.1	2.02	0.89,0.90	-	-	-
220Glu	8.47	124.99	-	-	-	4.29	1.92,2.01	2.20,2.25	-	-	-
221Arg	8.35	122.82	-	-	-	4.39	1.75,1.88	1.62,1.62	3.17,3.17	-	-
222Ser	8.04	123.13	-	-	-	4.24	3.83,3.84	-	-	-	

References

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

Ayton, P.M., Chen, E.H., Cleary, M.L. (2004). Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein, Molecular and Cellular biology **24**, 10470–10478

Bachman, K.E., Park, B.H., Rhee, I., Rajagopalan, H., Herman, J.G., Baylin, S.B., Kinzler, K.W., Vogelstein, B. (2003). Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. Cancer Cell. **3**, 89-95.

Bader, S., Walker, M., Hendrich, B., Bird, A., Bird, C., Hooper, M., Wyllie, A. (1999). Somatic frameshift mutations in the MBD4 gene of sporadic colon cancers with mismatch repair deficiency. Oncogene **18**, 8044-7.

Ballestar, E., Paz, M.F., Valle, L., Wei, S., Fraga, M.F., Espada, J., Cigudosa, J.C., Huang, T.H., Bartolomei, M.S., Tilghman, S.M. (1997). Genomic imprinting in mammals. Annu Rev Genet **31**, 493-525.

Ballestar, E., Yusufzai, T.M., and Wolffe, A.P. (2000). Effects of Rett syndrome mutations of the methyl-CpG binding domain of the transcriptional repressor MeCP2 on selectivity for association with methylated DNA. Biochemistry **39**, 7100–7106

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature **410**, 120-124.

Barkhuijsen, H., de Beer, R., Bov´ee, W. M. M. J., van Ormondt, D. (1985). Retrieval of frequencies, amplitudes, damping factors and phases from time-domain signals using a linear least-squares procedure J. Magn. Reson. **61**, 465–481

Baylin, S.B., Herman, J.G., Graff, J.R., Vertino, P.M., Issa, J.P. (1998). Alterations in DNA methylation: a fundamental aspect of neoplasia. Adv. Cancer Res. **72**, 141-196.

Beaujean, N., Hartshorne, G., Cavilla, J., Taylor, J., Gardner, J., Wilmut, I., Meehan, R., Young, L. (2004). Non-conservation of mammalian preimplantation methylation dynamics. Curr Biol. **14**, R266-267.

Beaujean, N., Taylor, J.E., McGarry, M., Gardner, J.O., Wilmut, I., Loi, P., Ptak, G., Galli, C., Lazzari, G., Bird, A., Young, L.E., Meehan, R.R. (2004). The effect of interspecific oocytes on demethylation of sperm DNA. Proc Natl Acad Sci U S A. **101**, 7636-7640.

Beaujean, N., Martin, C., Debey, P., Renard, J.P. (2005). Reprogramming and epigenesist. Med Sci (Paris) **21**, 412-421.

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

Berger, S.L. (2001). An embarrassment of niches: the many covalent modifications of histones in transcriptional regulation. Oncogene **20**, 3007–3013.

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

Bernstein, D.A., Zittel, M.C., Keck, J.L. (2003). High-resolution structure of the E. coli RecQ helicase catalytic core. EMBO J **22**, 4910–4921

Bestor, T.H. (1992). Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. EMBO J. **11**, 2611-2617.

Bestor, TH (2000). The DNA methyltransferases of mammals.. Hum Mol Genet. **9**, 2395-2402

Bestor, T.H., Ingram, V.M.. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc Natl Acad Sci U S A. **80**, 5559-5563.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature **397**, 579-583.

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

Bird, A.P. & Wolffe, A.P. (1999). Methylation-induced repression--belts, braces, and chromatin. Cell **99**, 451-454.

Bird,A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature, **321**, 209–213

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. **16**, 6-21.

Bernstein, B.E., Humphrey E.L., Erlich R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., Schreiber, S.T. . (2002) Methylation of histone H3 Lys4 in coding regions of active genes. Proc. Natl. Acad. Sci. U. S. A. **99**, 8695–8700

Boumil, R.M. & Lee, J.T. (2001). Forty years of decoding the silence in X-chromosome inactivation. Hum Mol Genet. **10**, 2225-2232.

Bowen, N.J., Palmer, M.B., Wade, P.A. (2004). Chromosomal regulation by MeCP2: structural and enzymatic considerations. Cell Mol Life Sci. **61**, 2163-2167.

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

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N.S., Read, R.J., Rice, L.M.,

Simonson, T., and Warren, G.L. (1998). Crystallography and NMR system (CNS): A new software system for macromolecular structure determination, *Acta Cryst.* **D54**, 905-921

Cameron, E.E., Bachman, K.E., Myohanen, S., Herman, J.G., Baylin, S.B. (1999). Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. Nat Genet. **21**, 103-107.

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

Cheadle, J. P., Gill, H., Fleming, N., Maynard, J., Kerr, A., Leonard, H., Krawczak, M., Cooper, D. N., Lynch, S., Thomas, N., Hughes, H., Hulten, M., Ravine, D., Sampson, J. R., Clarke, A. (2000). Long-read sequence analysis of the MECP2 gene in Rett syndrome patients: correlation of disease severity with mutation type and location. Hum. Molec. Genet. **9**, 1119-1129.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science **277**, 1996-2000.

Cohen-Armon, M., Visochek, L., Katzoff, A., Levitan, D., Susswein, A.J., Klein, R., Valbrun, M., Schwartz, J.H. (2004). Long-term memory requires polyADP-ribosylation. Science **304**, 1820-1822.

Cosgrove MS, Boeke JD & Wolberger C (2004). Regulated nucleosome mobility and the histone code. Nat Struct Mol Biol **11**, 1037–1043

Daniel, J.M., Reynold,s A.B. (1999). The catenin p120(ctn) interacts with Kaiso, a novel BTB/POZ domain zinc finger transcription factor. Mol Cell Biol. **19**, 3614-3623.

Delgado, S., Gomez, M., Bird, A., Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. EMBO J. **17**, 2426-2435.

Dennis. K., Fan, T., Geiman, T., Yan, Q., Muegge, K. (2001). Lsh, a member of the SNF2 family, is required for genome-wide methylation. Genes Dev. **15**, 2940–2944.

Dorigo, B., Schalch, T., Bystricky, K., Richmond, T.J. (2003). Chromatin fiber folding: requirement for the histone H4 N-terminal tail. J Mol Biol **327**, 85–96.

Dorigo, B., Schalch, T., Kulangara, A., Duda, S., Schroeder, R.R., Richmond, T.J. (2004). Nucleosome arrays reveal the two-start organization of the chromatin fiber. Science **306**, 1571–1573.

Dragich, J., Houwink-Manville, I., Schanen, C. (2000). Rett syndrome: a surprising result of mutation in MECP2. Hum. Molec. Genet. **9**, 2365-2375.

Dunican DS, Ruzov A, Hackett JA, Meehan RR. (2008) xDnmt1 regulates transcriptional silencing in pre-MBT Xenopus embryos independently of its catalytic function. Development **135**, 1295-1302.

Edwards, C.A. & Ferguson-Smith, A.C. (2007) Mechanisms regulating imprinted genes in clusters. Curr. Opin. Cell Biol. **19**, 281–289.

Ehrlich, M., & Wang, R.Y. (1981). 5-Methylcytosine in eukaryotic DNA. Science **212**, 1350-1357

Espada, J., Ballestar, E., Fraga, M.F. (2004). Human DNA methyltransferase 1 is required for maintenance of the histone H3 modification pattern. J Biol Chem **279**, 37175–37184

Esteller, M. (2003). Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. EMBO J. **22**, 6335-6345.

Esteller, M., Corn, P.G., Baylin, S.B., Herman, J.G. (2001). A gene hypermethylation profile of human cancer. Cancer Res. **61**, 3225-3229.

Esteller, M., Corn, P.G., Urena, J.M., Gabrielson, E., Baylin, S.B., Herman, J.G. (1998). Inactivation of glutathione S-transferase P1 gene by promoter hypermethylation in human neoplasia. Cancer Res. **58**, 4515-4518.

Esteller, M. & Herman, J.G. (2002). Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J. Pathol. **196**, 1–7.

Fahrner, J.A., Eguchi, S., Herman, J.G., Baylin, S.B. (2002). Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. Cancer Res. **62**, 7213-8.

Fatemi, M., Hermann, A., Pradhan, S., Jeltsch, A. (2001). The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. J Mol Biol. **309**, 1189-1199.

Feinberg, A.P. & Vogelstein, B. (1983). B.Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature **301**, 89-92.

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

Frank, D., Keshet, I., Shani, M., Levine, A., Razin, A., Cedar, H. (1991). Demethylation of CpG islands in embryonic cells. Nature **351**, 239-241.

Fuks, F., Burgers, W.A., Brehm, A., Hughes-Davies, L., Kouzarides, T. (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. Nature Genetics **24**, 88-91.

Fuks, F., Hurd, P.J., Deplus, R., Kouzarides, T. (2003). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res. **31**, 2305-2312

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem. **278**, 4035-4040.

Fujita, N., Shimotake, N., Ohki, I., Chiba, T., Saya, H., Shirakawa, M., Nakao, M, (2000). Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. Mol Cell Biol. **20**, 5107-5118.

Fujita, N., Takebayashi, S., Okumura, K., Kudo, S., Chiba, T., Saya, H., Nakao, M. (1999). Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms. Mol Cell Biol. **19**, 6415-6426

Fujita, N., S. Watanabe, T. Ichimura, Y. Ohkuma, T. Chiba, H. Saya, and M. Nakao. (2003). MCAF mediates MBD1-dependent transcriptional repression. Mol. Cell. Biol. **23**, 2834-2843

Fujita, N., S. Watanabe, T. Ichimura, S. Tsuruzoe, Y. Shinkai, M. Tachibana, T. Chiba, and M. Nakao. (2003). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J. Biol. Chem. **278**, 24132-24138

Filion, G.J., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., Defossez, P.A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription.. Mol Cell Biol. **26**, 169-181

Gilbert, D..M. (2002). Replication timing and transcriptional control: beyond cause and effect. Curr Opin Cell Biol. **14**, 377-383

Gordon, F., Luger, K., Hansen, J.C. (2005). The core histone N-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. J Biol Chem. **280**, 33701-33706.

Geinman, T..M., Robertson, K..D. (2002). Chromatin Remodeling, Histone Modifications, and DNA Methylation – How Does it All Fit Together? Journal of Cellular Biochemistry **87**, 117-125

Gill, G. (2004). SUMO and ubiquitin in the nucleus: different functions, similar mechanisms? Genes Dev. **18**, 2046-2059.

Gruenbaum, Y., Szyf, M., Cedar, H., Razin, A. (1983). Methylation of replicating and post-replicated mouse L-cell DNA. Proc Natl Acad Sci U S A. **80**, 4919-4921.

Graff, J.R., Herman, J.G., Lapidus, R.G., Chopra, H., Xu, R., Jarrard, D.F., Isaacs , W.B., Pitha, P.M., Davidson, N.E., Baylin, S.B. (1995). E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas. Cancer Res. **55**, 5195-5199.

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

Hendrich, B., Abbott, H., McQueen, H., Chambers, D., Cross, S., Bird, A. (1999). Genomic structure and chromosomal mapping of the murine and human MBD1, MBD2, MBD3, and MBD4 genes. Mamm. Genome **10**, 906–912

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

Heard, E., Clerc, P., Avner, P. (1997). X-chromosome inactivation in mammals. Annu Rev Genet. **31**, 571-610

Heitz, E. 1928. Das Heterochromatin der Moose. I. Jahrb Wiss Bot **69,** 762-818

Ho, K.L., McNae, I.W., Schmiedeberg, L., Klose, R.J., Bird, A.P., Walkinshaw, M.D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG.. Mol Cell. **29**, 525-531.

Hoffman, R.E., Kumar, A., Bishop, K.D., Borer, P.N., Levy, G.C. (1989) Application of the maximum likelihood method to a large 2D NMR spectrum using a parallel computer; Journal of magnetic resonance. **83**, 586-594

Holbert, M.A. & Marmorstein, R. (2005). Structure and activity of enzymes that remove histone modifications. Curr Opin Struct Biol. **15**, 673-80

Holliday, R. & Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science. **187**, 226-232

Hopfner, R., Mousli, M., Jeltsch, J.M., Voulgaris, A., Lutz, Y., Marin, C., Bellocq, J.P., Oudet, P., Bronner, C. (2000). ICBP90, a novel human CCAAT binding protein, involved in the regulation of topoisomerase IIalpha expression.. Cancer Res. **60**, 121-128

Hwang, T. L. & Shaka, A. J. (1995). Water Suppression that works excitation sculpting using arbitrary wave-forms and pulse-field gradients. Journal of Magnetic Resonance **112**, 275-279.

Iwano, H., Nakamura, M., Tajima, S. (2004). Xenopus MBD3 plays a crucial role in an early stage of development. Dev Biol. **268**, 416-428.

Jeddeloh, J.A., Stokes, T.L., Richards, E.J. (1999). Maintenance of genomic methylation requires a SWI2/SNF2-like protein. Nature Genet. **22**, 94–97.

Jones, P.A. & Baylin, S.B. (2002). The fundamental role of epigenetic events in cancer. Nat Rev Genet. **3**, 415-428.

Jones, P. A., & Laird, P. W. (1999). Cancer epigenetics comes of age. Nature Genetics **21**, 163-167.

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

Jones, S., Shanahan, H.P., Berman, H.M., Thornton, J.M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res. **31**, 7189-7198.

Jorgensen HF, Ben-Porath I, Bird AP. (2004). Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol Cell Biol. **24**, 3387-3395.

Kanai, Y., Ushijima, S., Kondo, Y., Nakanishi, Y., Hirohashi, S. (2001). DNA methyltransferase expression and DNA methylation of CpG islands and peri-centromeric satellite regions in human colorectal and stomach cancers. Int J Cancer **91**, 205-212

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

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev. **6**, 705-14

Kim, S.W., Fang, X., Ji, H., Paulson, A.F., Daniel, J.M., Ciesiolka, M., van Roy, F., McCrea, P.D. (2002) Isolation and characterization of xKaiso, a transcriptional repressor that associates with the catenin Xp120(ctn) in Xenopus laevis.. J Biol Chem. **277**, 8202-8208

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

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

Kudo, S., Nomura, Y., Segawa, M., Fujita, N., Nakao, M., Dragich, J., Schanen, C., Tamura, M. (2001) Functional analyses of MeCP2 mutations associated with Rett syndrome using transient expression systems. Brain Dev. Suppl **1**, 165-73.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H. (1992) CpG islands as gene markers in the human genome Genomics. **4,** 1095-107.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., Thornton, J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR **8**, 477-86.

Lee, J. H., Tate, C. M., You, J. S. & Skalnik, D. G. (2007) Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J. Biol. Chem.* **282**, 13419–13428

Lee JH, Voo KS, Skalnik DG. (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. J Biol Chem. **276**, 44669-44676

Lee, W. H., Isaacs, W. B., Bova, G. S., Nelson, W. G. (1997) CG island methylation changes near the GSTP1 gene in prostatic carcinoma cells detected using the polymerase chain reaction: a new prostate cancer biomarker. Cancer Epidemiol. Biomark. Prev. **6**, 443-450.

Leibiger, B., Moede, T., Schwarz, T., Brown, G.R., Kohler, M., Leibiger, I.B., Berggren, P.O. (1998) Short-term regulation of insulin gene transcription by glucose. Proc Natl Acad Sci U S A. **95**, 9307-9312

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

Lin, X., Nelson, W.G. (2003). 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.

Le Guezennec, X., Vermeulen, M., Brinkman, A.B., Hoeijmakers, W.A., Cohen, A., Lasonder, E., Stunnenberg, H.G. (2006). MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. Mol Cell Biol. **26**, 843-851

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

Lin, X., Tascilar, M., Lee, W. H., Vles, W. J., Lee, B. H., Veeraswamy, R., Asgari, K., Freije, D., van Rees, B., Gage, W. R., Bova, G. S., Isaacs, W. B., Brooks, J. D., DeWeese, T. L., De Marzo, A. M., Nelson, W. G. (2001). GSTP1 CpG island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells. Am. J. Pathol., **159**, 1815-1826.

Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. (2003) Refinement of protein structures in explicit solvent Proteins **50**, 496-506.

Luger, K. Mader, AW., Richmond, R.K., Sargent, D.F., Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature **389**, 251-260.

Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature **190**, 372-373.

Matijevic, T., Knezevic, J., Slavica, M., Pavelic, J. (2009). Rett syndrome: from the gene to the disease.. Eur Neurol. **1** 3-10.

Marks, P., Rifkind, R.A., Richon, V.M., Breslow, R., Miller, T., Kelly, W.K. (2001). Histone deacetylases and cancer: causes and therapies. Nature reviews Cancer **1**, 194-202

McBryant, S.J., Adams, V.H., Hansen, J.C. (2006). Chromatin architectural proteins. Chromosome Res. **14**, 39-51.

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

Meehan, R.R., Lewis, J.D., Bird, A.P. (1992). Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic Acids Res. **20**, 5085-5092.

Merlo, A., Herman, J.G., Mao, L., Lee, D.J., Gabrielson, E., Burger, P.C., Baylin, S.B., Sidransky, D. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med. **1**, 686-692.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature **403**, 501-502.

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., Bird, A (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science **297**, 403-405.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins **12**, 345-364.

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

.

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

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

Nan, X., Tate, P., Li, E., Bird, A. (1996). DNA methylation specifies chromosomal localization of MeCP2. Mol Cell Biol. **16**, 414-421

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

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

Nan, X., Campoy, F.J., Bird, A. (1997) .MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell **88**, 471-81.

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

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

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

Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., Hayashi, Y. (2002). LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). Cancer Res. **62**, 4075-4080.

Orekhov, V. Y., Ibraghimov, I., Billeter, M. (2003). Optimizing resolution in ultidimensional NMR by three-way decomposition. J. Biomol. NMR **27**, 165-173

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Curr Biol. **10**, 475-478

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. EMBO Rep. **10**, 1235-1241.

Patra SK, Patra A, Zhao H, Carroll P, Dahiya R. (2003). Methyl-CpG-DNA binding proteins in human prostate cancer: expression of CXXC sequence containing MBD1 and repression of MBD2 and MeCP2.Biochem Biophys Res Commun. **302**, 759-766.

Patra, S.K., Patra, A., Zhao, H., Dahiya, R. (2002). DNA methyltransferase and demethylase in human prostate cancer. Mol Carcinog. **33**, 163-171

Paulsen, M., Ferguson-Smith, A.C., (2001). DNA methylation in genomic imprinting, development, and disease. J Pathol. **195**, 97-110.

Perez-Torrado, R., Yamada, D., Defossez, P.A. (2006). Born to bind: the BTB protein-protein interaction domain.. Bioessays. 2006 Dec;28(12):1194-202

Pradhan, M., Estève, P.O., Chin, H.G., Samaranayk,e M., Kim, G.D., Pradhan, S. (2008). CXXC domain of human DNMT1 is essential for enzymatic activity. Biochemistry. **47**, 10000-10009.

Price N, van der Leij F, Jackson V, Corstorphine C, Thomson R, Sorensen A, Zammit V.(2002). A novel brain-expressed protein related to carnitine palmitoyltransferase I. Genomics **80**, 433-442

Prokhorchuk, A.V., Aĭtkhozhina, D.S., Sablina, A.A., Ruzov, A.S., Prokhorchuk, E.B. (2001). KAISO a new member of the BTB/POZ family specifically binds to methylated DNA sequences. Genetika. **37**, 737-744

Razin, A., Kafri, T. (1994). DNA methylation from embryo to adult. Prog Nucleic Acid Res Mol Biol. **48**, 53-81.

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

Record, M.T., deHaseth, P.L., Lohman, T.M. (1977). Interpretation of monovalent and divalent cation effects on the lac repressor-operator interaction.. Biochemistry **16**, 4791-4796.

Reese, B. E., K. E. Bachman, S. B. Baylin, and M. R. Rountree. (2003). The methyl-CpG binding protein MBD1 interacts with the p150 subunit of chromatin assembly factor 1. Mol. Cell. Biol. **23**, 3226-3236

Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature **447**, 425–432.

Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics **23**, 381-382

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genetics **14**, 9-25.

Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D., Klug, A. (1984) Structure of the nucleosome core particle at 7 A resolution. Nature **311**, 532-537.

Robert, M..F., Morin, S., Beaulieu, N., Gauthier, F., Chute, I.C., Barsalou, A., MacLeod, A.R. (2003). DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat Genet. **33**, 61-65.

Robertson, K.D. (2001). DNA methylation, methyltransferases, and cancer. Oncogene **20**, 3139–3155.

Rountree, M.R., Bachman, K.E., Herman, J.G. & Baylin, S.B. (2001). DNA methylation, chromatin inheritance, and cancer. Oncogene **20**, 3156–3165.

Ruzov, A., Dunican, D.S., Prokhortchouk, A., Pennings, S., Stancheva, I., Prokhortchouk, E., Meehan, R.R. (2004). Kaiso is a genome-wide repressor of transcription that is essential for amphibian development.. Development **131**, 6185-6194

Scarano, E. (1971). The control of gene function in cell differentiation and in embryogenesis. Adv Cytopharmacol. **1**, 13-24

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

Saito, Y., Kanai, Y., Sakamoto, M., Saito, H., Ishii, H., Hirohashi, S. (2001). Expression of mRNA for DNA methyltransferases and methyl-CpG-binding proteins and DNA methylation status on CpG islands and pericentromeric satellite regions during human hepatocarcinogenesis. Hepatology **33**, 561-568

Sakai, T., Toguchida, .J, Ohtani, N., Yandell, D.W., Rapaport, J.M., Dryja, T.P. (1991). Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. Am J Hum Genet. **48**, 880-888.

Sakamoto, Y., Watanabe, S., Ichimura, T., Kawasuji, M., Koseki, H., Baba, H., Nakao, M. J (2007). Overlapping roles of the methylated DNA-binding protein MBD1 and polycomb group proteins in transcriptional repression of HOXA genes and heterochromatin foci formation Biol Chem. **282**, 16391-16400.

Santos-Rosa, H., Schneide,r R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., Kouzarides, T. (2002) Active genes are tri-methylated at K4 of histone H3. Nature. 2002 **419**, 407-11

Sarraf, S.A., Stancheva, I. (2004). 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.

Schneider, R., Bannister, A..J, Myers, F.A., Thorne, A.W., Crane-Robinson, C., Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol. 2004 **1**,73-7

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., Kingston, R.E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell. **98**, 37-46.

Sibisi, S., Skilling, J., Brereton, R.G., Laue, E. D. Staunton, J. (1984). Maximum entropy signal processing in practical NMR spectroscopy Nature **311**, 446 – 447.

Singal, R., van Wert, J., Bashambu, M. (2001). Cytosine methylation represses glutathione S-transferase P1 (GSTP1) gene expression in human prostate cancer cells. Cancer Res. **61**, 4820-4826 .

Sinsheimer, RL. (1955). The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. J Biol Chem. **215**, 579-583

Smith, C.L., Peterson, C.L. (2004). ATP-dependent chromatin remodeling. Curr Top Dev Biol **65**, 115–148.

Song, J., Rechkoblit, O., Bestor, T.H., Patel, D.J. (2010). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science.

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

Stancheva, I. and Meehan, R. R. (2000). Transient depletion of xDnmt1 leads to premature gene activation in Xenopus embryos. Genes Dev. **14**, 313 -327

Studisky, VM., Walter, W., Kireeva, M., Kashlev, M., Felsenfeld, G. (2004). Chromatin remodeling by RNA polymerases. Trends Biochem Sci. **29**, 127–135

Subramanian, T. & Chinnadurai, G. (2003). Association of class I histone deacetylases with transcriptional corepressor CtBP FEBS Letters **540**, 255-258.

Sun, L., Hui, A.M., Kanai, Y., Sakamoto, M., Hirohashi, S. (1997). Increased DNA methyltransferase expression is associated with an early stage of human hepatocarcinogenesis. Jpn J Cancer Res. **88**, 1165-1170

Szyf, M., Gruenbaum, Y., Urieli-Shova,l S., Razin, A. (1982). Studies on the biological role of DNA methylation: V. The pattern of E.coli DNA methylation. Nucleic Acids Res. **10**, 7247-59.

Ting, A.H., Jair, K.W., Suzuki, H., Yen, R.W., Baylin, S.B., Schuebe,l K.E. (2004). Mammalian DNA methyltransferase 1: inspiration for new directions. Cell Cycle **3**, 1024-1026.

Tate, P., Skarnes, W., Bird, A. (1996) The methyl-CpG binding protein MeCP2 is essential for embryonic development in the mouse. Nat Genet. **12**, 205-208.

Tchou, J. C., Lin, X., Freije, D., Isaacs, W. B., Brooks, J. D., Rashid, A., De Marzo, A. M., Kanai, Y., Hirohashi, S., Nelson, W. G. (2000). GSTP1 CpG island DNA hypermethylation in hepatocellular carcinomas. Int. J. Oncol., **16**, 663-676.

Tudor, M., Akbarian, S., Chen, R.Z., Jaenisch, R. (2002). Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. Proc Natl Acad Sci U S A. **99**, 15536-15541.

Tsuchiya, Y., Kinoshita, K., Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces.. Proteins **55**, 885-594.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., Skalnik, D.G. (2000). Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol Cell Biol. **20**, 2108-2121

Wan, M., Lee, S. S. J., Zhang, X.. Houwink-Manville, I.. Song, H.-R., Amir, R. E., Budden, S., Naidu, S., Pereira, J. L. P., Lo, I. F. M. Zoghbi, H. Y., Schanen, N. C., (1999). Francke, U. Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. Am. J. Hum. Genet. **65**, 1520-1529.

Wade, P.A. (2001). Methyl CpG-binding proteins and transcriptional repression. Bioessays. **23**, 1131-1137.

Wade, P.A., Gegonne, A., Jones, P.L., Ballestar, E., Aubry, F., Wolffe, A.P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. Nat Genet. **23**, 62-66.

Watanabe, S., T. Ichimura, N. Fujita, S. Tsuruzoe, I. Ohki, M. Shirakawa, M. Kawasuji, and M. Nakao. (2003). Methylated DNA-binding domain 1 and methylpurine-DNA glycosylase link transcriptional repression and DNA repair in chromatin. Proc. Natl. Acad. Sci. USA **100**, 12859-12864.

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

Xia, Z.B., Anderson, M., Diaz, MO., Zeleznik-Le, NJ. (2003). MLL repression domain interacts with histone deacetylases, the polycomb group proteins HPC2 and BMI-1, and the corepressor C-terminal-binding protein Proc Natl Acad Sci U S A. **100**, 8342-8347.

Yeivin, A., Razin, A. (1993). Gene methylation patterns and expression. EXS. **64**, 523-568

Yamada, T., Koyama, T., Ohwada, S., Tago, K., Sakamoto, I., Yoshimura, S., Hamada, K., Takeyoshi, I., Morishita, Y. (2002). Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability.Cancer Lett. **181**, 115-20.

Yoon, H.G., Chan, D.W., Reynolds, A.B., Qin, J., Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Mol Cell. **12**, 723-34

Young, L.E., Beaujean, N. (2004). DNA methylation in the preimplantation embryo: the differing stories of the mouse and sheep. Anim Reprod Sci. **82**, 61-78.

Yu, F., Thiesen, J., Stratling, W.H. (2000). Histone deacetylase-independent transcriptional repression by methyl-CpG-binding protein 2. Nucleic Acids Res. **28**, 2201-2206.

Zhu, J., He, F., Hu, S., Yu, J. (2008) On the nature of human housekeeping genes. Trends Genetics. **24**, 480-484

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

Zhao, X., Ueba, T., Christie, B.R., Barkho, B., McConnell, M.J., Nakashima, K., Lein, E.S., Eadie, B.D., Willhoite, A.R., Muotri, A.R., Summers, R.G., Chun, J., Lee, K.F., Gage, F.H. (2003). Mice lacking methyl-CpG binding protein 1 have deficits in adult neurogenesis and hippocampal function. Proc Natl Acad Sci U S A. **100**, 6777-6782.

Zhang, Y., Ng, H.H, (1999). 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 Dev. **13**, 1924-1935.