3. Results

3.1 MBD protein family

3.1.1 Objective

Bioinformatics approaches to find gene products with certain properties have become fashionable since the publication of the complete human genome in 2001 (Lander *et al*., 2001, Venter *et al*., 2001) and its free availability to researchers. This is particularly true for families of proteins. A protein family is a group of proteins that share at least a part of their amino acid sequence – usually a domain. Protein domains are defined as independent and often globular folding units within a three-dimensional protein structure. Many databases have been created that group proteins according to the domains they contain (see Table 20).

Database	Internet address
Pfam	http://www.sanger.ac.uk/Pfam
Smart	http://smart.embl-heidelberg.de
Prosite	http://www.expasy.org/prosite
Systers	http://systers.molgen.mpg.de
ProDom	http://prodes.toulouse.inra.fr/prodom/current/html/home.php
InterPro	http://www.ebi.ac.uk/interpro
Blocks	http://blocks.fhcrc.org
PROTOMAP	http://protomap.cornell.edu
SBASE	http://hydra.icgeb.trieste.it/sbase
TIGRFAM	http://www.tigr.org/TIGRFAMs

Table 20. Protein domain and protein family databases. Of these databases, Pfam, Prosite, and Smart proved to be very useful for this study.

To this end, the aa sequence of known domains are defined as motifs using sophisticated mathematical algorithms such as Hidden Markov Models (HMMs). Proteins are grouped into a protein family if they comprise a certain motif, that is to say, contain an aa sequence similar (each database has its own criteria for these similarities) to a motif in the database. This implies, that one protein can belong to several protein families if the corresponding motifs are present in its sequence. Most of these databases use an alignment of polypeptide sequences known to represent a certain motif in order to create a standard profile that is then applied to

screen other databases for the occurrence of a similar sequence.
In this project, the MBD motif from the Pfam database was used to look for additional proteins that could potentially bind methylated CpGs.

3.1.2 Human MBD proteins

The MBD of human MECP2 served as query sequence to search for new members of the MBD protein family. Initial standard BLAST searches of the NCBI, Celera, and SwissProt databases resulted only in five MBD proteins (MECP2, MBD1, MBD2, MBD3, and MBD4) which had previously been described and studied intensively.
However, the search of protein domain family databases (CDD, Pfam, Prosite and Smart) revealed six additional proteins, i.e. BAZ2A/TIP5, BAZ2B, CLLD8/SETDB2, SETDB1, KIAA1461, and KIAA1887 with similarities to the MBD of MECP2. Due to the results of this study, KIAA1461 has been named MBD5 and KIAA1887 is now called MBD6. The Pfam, Smart, and Prosite database use HMMs to detect motifs in amino acid sequences. This method is more sensitive than standard sequence similarity searches such as BLAST and FASTA, which explains why our BLAST searches did not detect these additional proteins. An MBD has been described in CLLD8, SETDB1, and BAZ2A/TIP5 earlier (Mabuchi *et al*., 2001, Schultz *et al*., 2002, Strohner *et al*., 2001) without including these proteins into the MBD protein family.
Nine of the eleven MBD-containing protein sequences could also be detected by screening the Sequence Similarity DataBase (SSDB) (Kanehisa *et al*., 2002) at GenomeNet. The cDNAs for KIAA1461/MBD5 and KIAA1887/MBD6 were not found since the KEGG database underlying the SSDB contains only confirmed but not predicted protein sequences. Table 21 summarizes all proteins identified, their domains, the position of the MBDs and the search method by which the protein was found.
The MBD amino acid sequences of the five previously published proteins as well as the MBD sequence of the six newly described human MBD protein family members were aligned to analyze the conservation of amino acids across the different proteins (Fig. 9).

Accession No.[a)]	Name	Search method /database[b)]	Domains[c)]	MBD position[d)]
P51608	MECP2	a, b, c, d, e, f, g, ac, bc, cc		96-149
Q9UIS9	MBD1	a, b, c, d, e, f, g, ac, bc, cc	zf-CXXC	7-59
Q9UBB5	MBD2	a, b, c, d, e, f, g, ac, bc, cc		151-204
O95983	MBD3	a, b, c, d, e, f, g, ac, bc, cc		8-60
Q9Z2D7	MBD4	a, b, c, d, e, f, g, ac, bc, cc	HhH-GPD	82-135
Q9UIF9	BAZ2A/TIP5	e, f, g	AT-hook, DDT, PHD, bromodomain	526-577
Q9UIF8	BAZ2B	e, f, g	DDT, PHP, bromodomain	549-600
Q15047	SETDB1	e, f, g	SET	597-653
Q96T68	SETDB2	e, f, g	SET	162-216
Q9P267	MBD5	e, f	PWWP	21-79
Q96Q00	MBD6	e, f		304-456

Table 21 Summary of the human MBD polypeptides. [a)] Accession numbers according to the SwissProt database. [b)] a = BLASTP, b = TBLASTN, c =TBLASTX, d = PSI-BLAST, e = Pfam, f = Ncbi, g= SSDB, ac = BLASTP in Celera database, bc = TBLASTN in Celera database, cc = TBLASTX in Celera database. [c)] domain nomenclature according to the Pfam database, all sequences contain an MBD in addition. [d)] position in amino acid sequence.

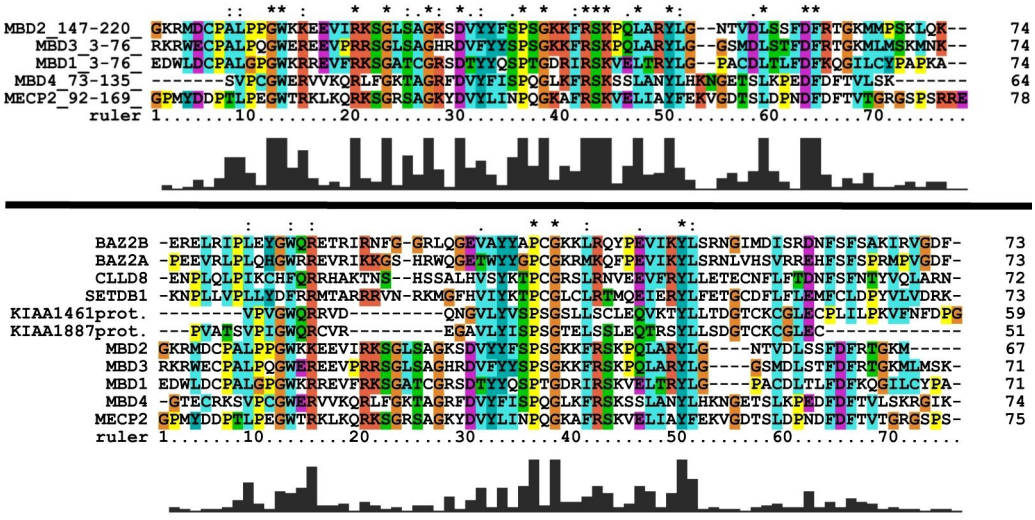

Fig. 9. Alignment of the human methyl-CpG-binding proteins according to Swissprot. ClustalX alignment of the original human MBDs (above) and all MBDs (below). Columns are colored by conservation and property (Thompson *et al.*, 1997). Residue conservation above each column indicates: "*" completely conserved; ":" favored substitutions; "." weakly favored substitutions. A quality graph is depicted below each alignment.

The analyses of all 11 polypeptides implicate a small number of highly conserved and apparently essential amino acids within the MBD domain, especially the proline at position 36 (P36) as well as glycine 38 (G38) and tyrosine 50 (Y50). Five positions with conservative substitutions can be found (indicated by ":").

Fig. 10 depicts a sequence logo derived from the alignment of all eleven human MBD sequences.

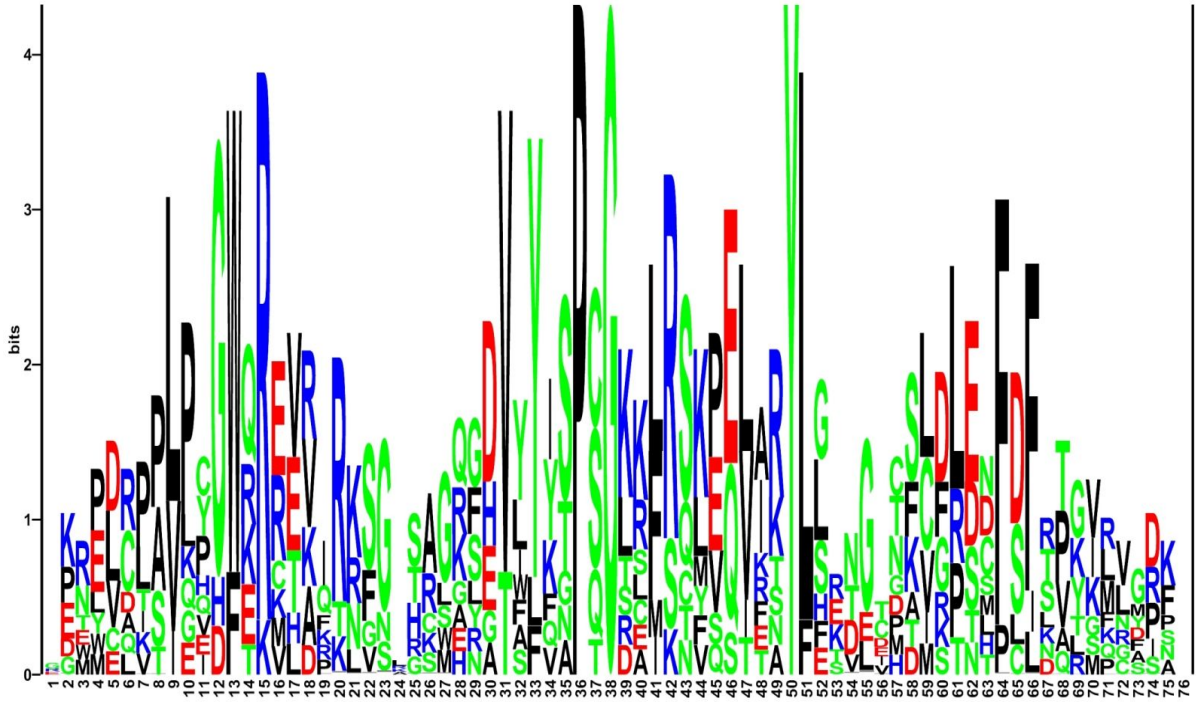

Fig. 10. Sequence logo of the eleven human MBD sequences. The height of the letters relative to the other aa, corresponds to the frequency of the amino acid at its position. The total height of each stack stands for the information present at this position, measured in bits. Top letters represent the consensus sequence. Grey bars indicate gaps in some of the aligned sequences.

A phylogenetic tree of the MBD amino acid sequences of all eleven polypeptides was computed using the ClustalW software at GenomeNet and is shown in Fig. 11.

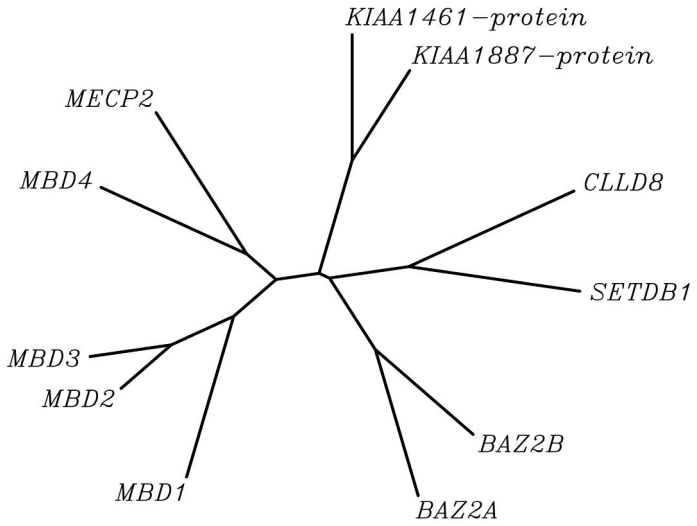

Fig. 11. Unrooted dendrogram depicting methyl-CpG-binding domains of the human protein family. A tree representation shows the similarity between the human MBD proteins. The left branch clusters the original 5 MBD proteins. Branch lengths are proportional to the amount of inferred evolutionary change.

Four major MBD subsets are indicated in Fig. 11. The MBDs of the originally described proteins (MBD1, MBD2, MBD3, MBD4, and MECP2) are found as one group besides a second (BAZ2A/TIP5, BAZ2B) and a third subset (CLLL8/SETDB2 and SETDB1) which are joined by a very short branch. KIAA1461/MBD5 and KIAA1887/MBD6 appear in a fourth branch. MBDs of the original five proteins are more similar to each other than to the novel ones, which explains why BLAST analyses with the MBD of MECP2 as query failed to identify the second, third, or fourth class.

3.1.3 Domain analysis

An analysis of the amino acid sequences revealed that the MBD was the only domain shared by all eleven sequences.

The MBD of MECP2, MBD1, MBD2, MBD4, and BAZ2A/TIP5 mediates binding to DNA, in case of MECP2, MBD1, and MBD2 preferentially to methylated CpG (Ng *et al*., 2000, Strohner *et al*., 2001, Lewis *et al*., 1992, Hendrich and Bird, 1998, Bird, 2002). MBD4 has a special role acting as a DNA repair enzyme as described in the introduction.

In case of human MBD3 and SETDB1, the MBD has been shown to mediate protein-protein interactions (Schultz *et al*., 2002, Saito and Ishikawa, 2002). *Xenopus* MBD3 is exceptional in its binding to methylated CpGs which can be explained by the difference of an amino acid residue within the MBD (Lys30) important for DNA binding (Saito and Ishikawa, 2002). It remains to be determined whether the MBDs of BAZ2B, CLLD8/SETDB2, KIAA1461/MBD5, and KIAA1887/MBD6 mediate DNA binding or protein-protein

interactions.

Additional domains found in seven of the eleven polypeptides indicate that they are associated with chromatin and function in epigenetic mechanisms of gene regulation. Some of the proteins are already known to be involved in transcriptional repression and the domains of the remainder strongly suggest a comparable function.

MECP2 recruits the Sin3A co-repressor complex and MBD2 the Mi-2/NuRD co-repressor complex, which itself contains MBD3. Both complexes contain HDACs, and MBD1 is also associated with HDAC activity although the identity of the deacetylase remains unknown (Ng *et al*., 2000). Within the C-terminal part of MECP2 a histidine and proline-rich region is present which is conserved in certain neural-specific transcription factors (Vacca *et al*., 2001).

BAZ2A/TIP5 is part of the nucleolar remodeling complex (NoRC) which represses rDNA transcription by recruiting histone methyltransferases, HDACs, and DNA methyltransferases (Santoro *et al*., 2002).

BAZ2B has a domain structure similar to BAZ2A/TIP5. Both contain a DDT domain (DNA binding homeobox and Different Transcription factors) and a tandem PHD-bromodomain. The PHD domain is a C4HC3 zinc-finger-like motif and the bromodomain consists of 110 amino acids and is found in many chromatin-associated proteins that can interact specifically with acetylated lysine. Tandem PHD-bromodomains have been found in several transcriptional co-repressors (Schultz *et al*., 2002). The DDT domain is exclusively associated with nuclear domains in other proteins and was found in different transcription and chromatin remodeling factors (Doerks *et al*., 2001). An AT-hook motif (which allows binding to the minor groove of AT-rich DNA regions) was found in BAZ2A/TIP5 but not in BAZ2B.

SETDB1 is a H3-K9 histone methyltransferase (Schultz *et al*., 2002). Its mouse homologue, ESET, has furthermore been shown to interact with the mSin3A/B co-repressor complex (Yang *et al*., 2002). The SET domain is a signature motif for lysine-specific histone methyltransferases (Wang *et al*., 2001, Tachibana *et al*., 2001). This domain is also present in CLLD8/SETDB2 to which no function has yet been assigned. A recent study showed the formation of an S-phase-specific complex including SETDB1, MBD1, and CAF-1 that facilitates methylation of H3-K9 during replication-coupled chromatin assembly (Sarraf and Stancheva, 2004) .

The predicted protein sequence of KIAA1461/MBD5 harbors a PWWP motif (Stec *et al*., 1998) named after the conserved amino acids Pro-Trp-Trp-Pro. It was first described in the

WHSC1 protein, encoded by a gene within the Wolf-Hirschhorn syndrome critical region. The PWWP domain of Dnmt3b, a DNA methyltransferase, has recently been shown to bind to DNA (Qiu *et al*., 2002). A common feature of PWWP containing proteins is the presence of additional domains known to be associated with chromatin (Qiu *et al*., 2002). Furthermore, KIAA1461/MBD5 has been shown to interact with the KIAA1549 protein in a yeast-two-hybrid experiment (http://www.kazusa.or.jp/huge/ppi). This interaction partner has not been studied in detail so far, but contains a serine-rich stretch as well as a helix-turn-helix motif (PS00622, LuxR family) according to Prosite. Helix-turn-helix motifs can be found in many transcription regulation proteins. Interestingly, the *KIAA1461/MBD5* gene lies in a region that was found to be deleted in a mentally retarded patient (Koolen *et al*., 2004).
In the predicted protein sequence of KIAA1887/MBD6 only a proline-rich extension but no protein motif as such could be found.
The co-existence of MBDs and domains involved in histone modification (e.g. the SET domain and the bromodomain) in SETDB1, SETDB2, BAZ2A, and BAZ2B indicates a potential connection between the recognition of methylated DNA and histone modifications. This link is especially tempting for SETDB1, which has been shown to methylate histones (Schultz *et al*., 2002), while its mouse homologue ESET binds to mSinA co-repressor complex (Yang *et al*., 2002) that is also bound by MECP2 (Nan *et al.*, 1998).

3.1.4 MBD proteins in other species

In the mouse homologues were found for all human MBD proteins. Sequence identity scores range from 63.8% to 94.0% (Tab. 22) indicating a conserved function in both species. Human and mouse MBD1, MBD2, MBD3, MBD4 and MECP2 are curated orthologues (Hendrich and Bird, 1998, Hendrich *et al*., 1999, Quaderi *et al*., 1994, D'Esposito *et al*., 1996).

Human protein	Mouse protein	% identity
MECP2	MECP2 (541 aa)	93.4% in 542 aa
MBD1	MBD1 (713 aa)	66.8% in 698 aa
MBD2	MBD2 (454 aa)	93.2% in 545 aa
MBD3	MBD3 (362 aa)	85.1% in 355 aa
MBD4	MBD4 (631 aa)	63.8% in 647 aa
BAZ2A	BAZ2A (1972 aa)	80.1% in 2039 aa
BAZ2B	ENSMUSP00000028367 (2065 aa)	82.6% in 2027 aa
SETDB1	ESET (1457 aa)	91.0% in 1465 aa
SETDB2	ENSMUSP00000022552 (701 aa)	65.2% in 715 aa
MBD5	ENSMUSP00000036847 (1518 aa)	94.0% in 1521 aa
MBD6	ENSMUSP00000026476 (101 aa)	93.5% in 216 aa

Table 22. Mouse homologues to human MBD proteins. Comparison of the human MBD proteins and their mouse homologues. Identity scores were calculated with the LALIGN program (http://www.ch.embnet.org/software/LALIGN_form.html). The numbers of aa residues are indicated. [a)] The number of aligned amino acids can exceed the residue number of the sequences due to gaps inserted by the algorithm.

Homology searches in the Ensembl database revealed the following murine homologues of BAZ2B, CLLD8, *KIAA1461*/*MBD5* protein, and *KIAA1887*/*MBD6* protein. The mouse homologue of BAZ2B, the Ensembl protein ENSMUSP00000028367 (gene ENSMUSG00000026987), shows a 82.6% amino acid sequence identity. The predicted mouse gene ENSMUSG00000021980 coding for Ensembl protein ENSMUSP00000022552 has a 65.2% amino acid sequence identity to the human CLLD8/SETDB2. It was furthermore found, that the predicted gene ENSMUSG00000036792 coding for the Ensembl protein ENSMUSP00000036847 is a mouse homologue of human KIAA1461/MBD5 protein with a 94.0% amino acid sequence identity. For the KIAA1887/MBD6 protein, the mouse gene sequence ENSMUSG00000025409 (Ensembl protein ENSMUSP00000026476) with 93.5% sequence identity was present in the database. However, the latter database entry consists of only 216 amino acids.

The existence of corresponding translated proteins remains to be determined for all four mouse genes.

DNA methylation as a mechanism of gene expression regulation exists also in plants. In the database searches plant MBD proteins were detected as well. The Pfam database contains polypeptides from *Arabidopsis thaliana* (thale cress) and *Triticum aestivum* (bread wheat).

BLAST analyses revealed additional proteins in *Zea Mays* (maize), *Hordeum vulgare* (two-rowed barley) and *Lycopersicon esculentum* (tomato). Entries for MBD containing proteins from plants over *C. elegans* (nematode) to mouse and human are present in Pfam.

3.1.5 Expression patterns of MBD genes

Expression analyses had been carried out previously for all genes of the mouse/human MBD family except for *KIAA1461/MBD5* and *KIAA1887/MBD6* (only the abundance of *KIAA1887/MBD6* ESTs in different tissues had been reported (Nagase *et al*., 2001a)). The results of published Northern blot experiments are summarized in Tab. 22. Since expression levels of *MBD4* were too low to be detected by Northern blots, only results of RT-PCR studies in three tissues are shown. However the presence of *MBD4* EST sequences from numerous tissues points to a ubiquitous expression (Unigene).

Tissue	MECP2[a)]	MBD1[b)]	MBD2[b)]	MBD3[b)]	MBD4[b)]	BAZ2A[d)]	BAZ2B[d)]	SETDB1[e)]	SETDB2[c)]	MBD5[f)]	MBD6[g)]
Brain	+	+	+	+	+	+	(+)	+	(+)	(+)	(+)
Heart	+	+	+	+		+	+	+	+	+	+
Kidney	+	+	+	+		+	-	+	+	(+)	+
Liver	+	(+)	(+)	+		(+)	-	+	+	(+)	+
Lung	+	+	+	+		+	-	+	+	(+)	(+)
Skeletal muscle	+	+	+	+		+	+	+	(+)	+	+
Spleen		+	+	+		+	-	+	+		+
Testis		+	+	+	+	+	+	+	+		+
ES cells		-	(+)	+	+						
Placenta	+					+	+	+	(+)	+	+
Pancreas	+					+	+	+	+	+	+

Table 23. Expression patterns of the mouse/human MBD genes. The table shows the expression of the eleven MBD genes in major tissues as detected by Northern blot. Empty spaces denote lacking information for that tissue. (+) indicates very low or doubtful expression, - no expression. [a)] (D'Esposito *et al*., 1996, Coy *et al*., 1999). [b)] (Hendrich and Bird, 1998) For these genes, comprehensive expression data was only available from mouse. [c)] (Mabuchi *et al*., 2001) Expression in additional tissues has been reported. [d)] (Jones *et al*., 2000) Expression in additional tissues has been reported. [e)] (Nomura *et al*., 1994) *SETDB1* was originally called *KIAA0067*. [f)] Northern blot results of this study. [g)] (Nagase *et al*., 2001a) and Northern blot results of this study.

Northern blot analyses for *KIAA1461/MBD5* and *KIAA1887/MBD6* were performed during

this thesis. Strong signals, representing a transcript of ~8 kb were detected for KIAA1461/MBD5 in skeletal muscle, heart, pancreas, kidney and placenta. A faint band could be detected in brain, lung and liver. For KIAA1887/MBD6 a strong band corresponding to a transcript of ~5 kb was present in heart, kidney, liver, skeletal muscle, placenta, and pancreas, and weaker signals could be seen for brain and lung tissue (Fig. 12).

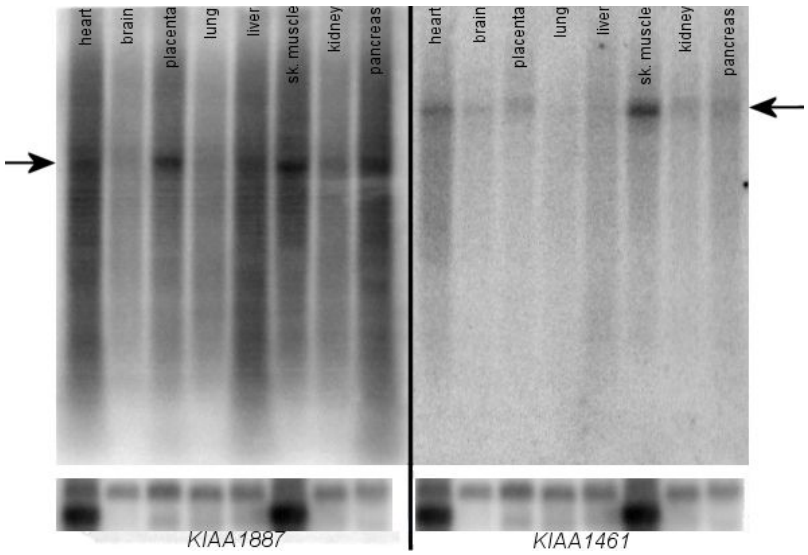

Fig. 12. Northern blots for *KIAA1461/MBD5* and *KIAA1887/MBD6*. Human multiple tissue Northern blot showing the expression of the KIAA1461/MBD5 and KIAA1887/MBD6 genes. The calculated sizes of the transcripts are ~5 kb for *KIAA1887/MBD6* and ~8 kb for *KIAA1461/MBD5,* indicated by arrows. A β-actin probe was hybridized as loading control, shown at the bottom.

Taken together, *MBD1, MBD2, MBD3,* and *MECP2* as well as *SETDB1, CLLL8/SETDB2, BAZ2A, KIAA1461/MBD5,* and *KIAA1887/MBD6* show a broad tissue distribution. In contrast the expression of *BAZ2B* is more restricted according to Northern blot results (Table 23) (Jones *et al*., 2000). It is noteworthy, that *CLLL8/SETDB2, BAZ2A*, *KIAA1461/MBD5,* and *KIAA1887/MBD6* are expressed at very low levels in brain.

3.2 Search for MECP2 paralogues

3.2.1 Objective

The function of MECP2 is not only dependent on the MBD, but also on other domains of the protein and on its overall structure (Klose and Bird, 2004). As yet, only the 3D-structure of the MBD has been resolved while the exact shape of the entire protein remains to be elucidated.

A bioinformatics approach was therefore used to find human proteins with a global similarity to MECP2. Such paralogue proteins are thought to arise through gene duplication and can

give clues about the tertiary structure and about related functions of the protein of interest.

3.2.2 Proteins with an overall sequence similarity to MECP2

To avoid false positives in this search, MECP2 was first checked for repetitive sequences. A dotplot analysis revealed three such regions with low sequence complexity (aa 20 - aa 90, aa 160 - aa 200, and aa 340 - aa 400, see also Fig. 13). These regions were masked and BLASTP searches against the NCBI protein sequence database were performed.

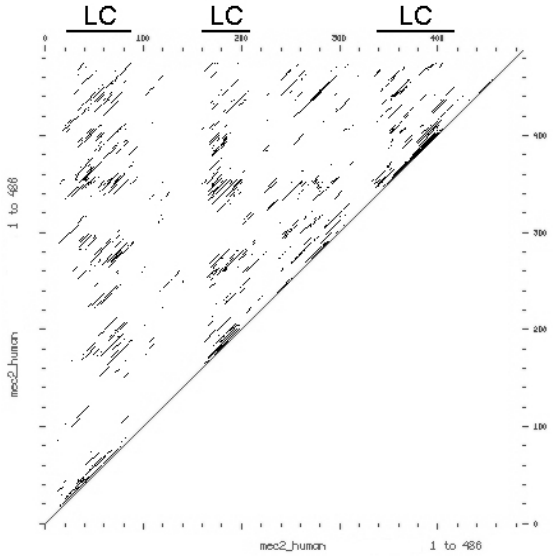

Fig. 13. DotPlot output of a comparison of MECP2 to itself. Every dot in the graph represents a similarity. Clusters of dots that do not lie on the diagonal indicate potential repetitive sequences. Such regions with low complexity are indicated by LC.

The BLASTP searches were followed by database queries of the sequence similarity database (SSDB) at GenomeNet and the BLink database at NCBI using the search terms hsa:4204 and P51608 respectively.

The BLink search revealed a total of 30 matches in the *Homo sapiens* proteome, but due to redundancy there were only 12 unique hits, four of which were MBD1, MBD2, MBD3 and MBD4. Since similarities to the MBD protein family members were mainly due to consensus in the MBD domains and therefore not of interest for overall structural similarity, they were disregarded. The remaining eight proteins are summarized in Table 24. The SSDB query with a SW-score cut-off at 200 yielded 18 non-redundant hits, including MBD1, MBD2 and MBD4.

The list resulting from the BLink search (8 proteins) and the SSDB query (15 proteins) was compared to the 100 best hits of the protein Blast with the masked MEC2_HUMAN as query sequence. Only one protein gave a positive result with all three search strategies.

Gene symbol	Protein name	SSDB	BLink	Blast
NEFH	Neurofilament heavy polypeptide	x	x	x
MLLT2/AF4	MLL/AF1p fusion - Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*)	x	x	
FOXG1B	Forkhead box G1B, Oncogene QIN		x	
TRDN	Triadin		x	
PRG4	Proteoglycan	x		
SRRM1	Serine/arginine repetitive matrix 1	x		
SRRM2	Serine/arginine repetitive matrix 2	x		
MDC1	Mediator of DNA damage checkpoint 1, KIAA0170 gene product	x		
NOLC1	Nucleolar and coiled-body phosphoprotein 1	x		
BAT2	HLA-B associated transcript 2	x		
IRS2	Insulin receptor substrate 2	x		
SRCAP	Snf2-related CBP activator protein	x		
CRK7	CDC2 related protein kinase 7	x		
AIM1	Absent in melanoma 1	x		
	Hypothetical protein BC001584	x		
ANK2	Ankyrin 2, neuronal	x		
ZNF469	Zinc finger protein 469	x		

Table 24. Results of the database queries. Crosses indicate positive results in the searches. SSDB = Sequence similarity database, BLink = Blast Link database, Blast = Blast search with masked MEC2_HUMAN as query. Only NEFH could be found in all 3 search methods.

The human neurofilament NEFH (Mattei *et al*., 1988) belongs to the intermediate filament family and has been associated with amyotrophic lateral sclerosis (ALS). Overexpression of *Nefh* causes motor neuron degeneration in transgenic mice (Collard *et al*., 1995) and deletions in the C-terminal part of *NEFH* can be found in ALS patients (Figlewicz *et al*., 1994), supporting the link between NEFH and ALS. The alignment with MECP2 shows a similarity in the C-terminal part of NEFH as depicted in Fig. 14 and Annex 6.1.

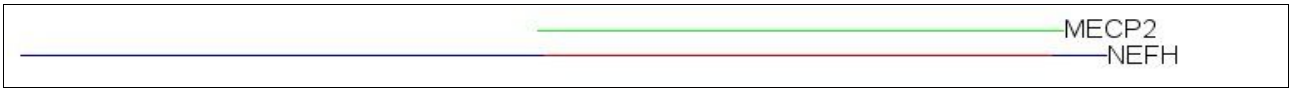

Fig. 14. Alignment of MECP2 (486 aa) and NEFH (1020 aa) derived form the SSDB. The overlap of 494 aa includes gaps and has an overall similarity of 0.235 (with 1 being a perfect similarity). Rather than having an overlap in only one domain, the whole MECP2 protein has a similarity to the C-terminal part of NEFH.

The product of the *MLLT2* gene, also called AF4 for "ALL1-fused gene from chromosome 4"

is a serine- and proline-rich putative transcription factor with a glutamine-rich carboxy terminus that results from translocation of ALL1 in acute leukemias. Fig. 15 represents an alignment of MLLT2 and MECP2.

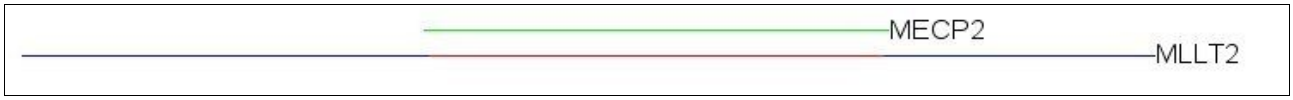

Fig. 15. Alignment of MECP2 (486 aa) and MLLT2 (2311 aa) derived from the SSDB. The overlap of 528 aa includes gaps and has an overall similarity of 0.225 (with 1 being a perfect similarity).

FOXG1B was described as the *QIN* gene product and is a transcription factor with a forkhead domain (Li and Vogt, 1993). It controls the fate of Cajal-Retzius cells (Hanashima *et al*., 2004).

Triadin (TRDN) is a protein thought to be involved in the calcium release from muscle contraction (Brandt *et al*., 1993). It is also the founder of the poorly understood triadin protein family (for review see Marty, 2004). Proteoglycan 4 (PRG4) also known as SZP and CACP can act as megakaryocyte-stimulating factor (Merberg *et al*., 1993). It is expressed by chondrocytes in bovine articular cartilage (Schumacher *et al*., 1994). *PRG4* is thought to be the disease gene for camptodactyly-arthropathy-coxavara-pericarditis (CACP) syndrome and has an isoform known as HAPO for hemangiopoietin. Finally, it is a growth factor acting on hematopoietic and endothelial progenitor cells (Liu *et al*., 2004).

SRRM1 is a co-activator of pre-mRNA splicing. It binds different splicing factors, associates with the nuclear matrix, and is involved in mRNA export (Wagner *et al*., 2003). SRRM2 is also involved in splicing, but was shown to be less important than SRRM1 for this process *in vitro* (Blencowe *et al*., 2000).

MDC1, also termed NFDB1, works with H2AX to promote the recruitment of repair proteins to DNA breaks and it controls damage-induced cell-cycle arrest checkpoints (Stewart *et al*., 2003).

NOLC1, formerly known as p130 or Nopp140, is a putative snoRNP assembly factor (Yang *et al*., 2000).

Little is known about BAT2, but Lehner *et al*. in 2004 proposed a role in the regulation of pre-mRNA splicing.

IRS2 links cell surface receptors to the intracellular insulin/IGF signaling cascade and has an important role in both peripheral insulin response and pancreatic beta-cell growth and

function. Dysregulation of Irs2 signaling in mice causes the failure of compensatory hyperinsulinemia during peripheral insulin resistance (for a review see Lee and White, 2004). In humans it has been associated with type II diabetes (Mammarella *et al*., 2000), severe obesity, and glucose intolerance (Lautier *et al*., 2003) .

SRCAP is an ATPase binding to the CREB-binding protein (CBP) and a transcription activator (Johnston *et al*., 1999). Furthermore, it is also a co-activator for the androgen receptor and can enhance glucocorticoid receptor-mediated transcription (Monroy *et al*., 2003).

CRK7, also termed CrkRS, is a protein kinase found in nuclear SC35 speckles (Ko *et al*., 2001).

A single-nucleotide polymorphism (SNP) in *AIM1* has been associated with differences in skin color in humans (Fukamachi *et al.*, 2001).

ANK2 is a spectrin-binding protein that is required for localization of inositol 1,4,5-trisphosphate receptor and ryanodine receptor in neonatal cardiomyocytes (Mohler *et al*., 2004). It was furthermore linked to long QT syndrome 4 (sick sinus syndrome with bradycardia) (Mohler *et al*., 2004).

ZNF469 is a zinc finger protein that was first described as the *KIAA1858* gene product (Nagase *et al*., 2001b).

None of the proteins could be found in the protein structure database PDB, which means that their 3D structure has not been determined as yet. Sequence comparison of MECP2 with proteins for which structural data is available, could have helped to get an idea of the 3D structure of MECP2 itself.

3.3 MECP2 target genes

3.3.1 Objective

Even though *MECP2* has been identified as disease gene for Rett syndrome in 1999 (Amir *et al*., 1999), the disease is still poorly understood. Especially knowledge on the exact function of MECP2, i.e. the genes regulated by MECP2 as a transcriptional repressor is missing. In addition, it is unknown whether MECP2 has functions other than transcriptional regulation.

A mouse model created by the Bird group (Guy *et al*., 2001) was used to find genes regulated

by MECP2 in the brain. The approach consisted of a microarray hybridization of total brain RNA from a *Mecp2*$^{-/y}$ mouse versus total brain RNA from a healthy litter mate control. The hybridization results were verified by northern blot and real-time PCR performed by our collaborators in Edinburgh. Subsequently, chromatin immunoprecipitation was carried out to prove direct binding of MECP2 to the genomic region of genes that might be involved in the disease.

Four microarray studies on Rett syndrome have been published since the start of this thesis. Two studies used cultured cells, i.e. primary fibroblasts (Traynor *et al*., 2002) and lymphoblastoid cells (Ballestar *et al*., 2005). Colantuoni and colleagues isolated RNA from post mortem human brain tissue (Colantuoni *et al*., 2001) and Tudor and colleagues isolated RNA from whole brains of MECP2 mutant mice (Tudor *et al*., 2002).

Apart from the study by Tudor and colleagues, the biological material used in these experiments does not seem to be well suited to find the neuronal target genes of MECP2 responsible for the phenotype (see 4.3.6 DNA microarray studies and chromatin immunoprecipitation).

3.3.2 Genes differentially expressed in *Mecp2*-null mice – a microarray study

Gene expression levels in the brains of a symptomatic *Mecp2*$^{-/y}$ mouse were compared to the expression levels of a *wt* litter mate control brain. Isolated RNA was sent by our collaborators from Edinburgh. The RNA was fluorescently labeled with Cy3 and Cy5 in a reverse transcription reaction and co-hybridized to a microarray with 13,627 cDNA clones.

After scanning, image analysis and quality control the data was subjected to ANOVA. From the resulting clones, those with an intensity ratio of 2.00 or higher were considered relevant. A total of 17 clones were found to be either up- or down-regulated at least 2-fold in the brain RNA of the mutant mouse. Due to redundancy, the 17 clones correspond to 11 transcripts of which three were down- and eight were up-regulated (Table 25).

Fold change	Up / down regulated	Gene description	UniGene cluster	Accession number
3.4412	+	*Sgk1 Serum/glucocorticoid regulated kinase*	Mm.28405	AA273540
3.2992	+	*Sgk1 Serum/glucocorticoid regulated kinase*	Mm.28405	AI527833
2.6313	+	*Fkbp5 FK506 binding protein 5 (51 kDa)*	Mm.276405	BC015260
2.5815	+	*Fkbp5 FK506 binding protein 5 (51 kDa)*	Mm.276405	BC015260
2.3051	+	*Fkbp5 FK506 binding protein 5 (51 kDa)*	Mm.276405	BC015260
2.4131	+	*Cirbp Cold inducible RNA binding protein*	Mm.17898	NM_007705
2.254	+	*Cirbp Cold inducible RNA binding protein*	Mm.17898	NM_007705
2.1947	+	*Cirbp Cold inducible RNA binding protein*	Mm.17898	NM_007705
2.2748	+	*Sult1a1 Sulfotransferase family 1A, phenol-preferring, member 1*	Mm.17339	AB029487
2.2651	-	*Sorcin*	Mm.96211	AK008970
2.2386	-	*Hsp105 Heat shock protein, 105 kDa*	Mm.270681	AA105012
2.0727	-	*Hsp105 Heat shock protein, 105 kDa*	Mm.270681	BC018378
2.2012	+	*Pomc1 Pro-opiomelanocortin-alpha*	Mm.277996	BC061215
2.1584	+	*Scya17 Small inducible cytokine subfamily A17*	Mm.41988	NM_011332
2.1477	-	*Gja12-pending Gap junction membrane channel protein alpha 12*	Mm.40016	AW742272
2.1375	+	*S3-12-pending Plasma membrane associated protein, S3-12*	Mm.347924	NM_020568
2.0019	+	*RIKEN cDNA 4930546H06 gene*	Mm.227456	AK016052

Table 25. Gene expression changes in the brain of an *Mecp2*$^{-/y}$ animal compared to a *wt* litter mate control detected by microarray hybridizations. Transcripts with more than 2-fold expression difference are shown. Some genes are represented by more than one differential clone on the array and therefore listed several times. The identity of all 17 clones was confirmed by sequencing of the spotted DNA. Up-regulation (+) and down-regulation (-) in the *Mecp2*$^{-/y}$ sample are indicated.

At least 5 of the 11 differentially expressed genes (*Fkbp5*, *Sgk*, *Pomc*, *Sult1A1*, and *Hsp105*) are known to be regulated by the stress hormones, glucocorticoids. Stress provokes release of corticotropin-releasing hormone (CRH) by the hypothalamus, which stimulates synthesis of adrenocorticotropic hormone (ACTH). ACTH in turn causes the adrenal cortex to produce circulating glucocorticoids (cortisol in humans, corticosterone in rodents) which bind to glucocorticoid receptors and coordinate the transcriptional response (Reichardt and Schütz, 1998). Negative feedback by glucocorticoids on the hypothalamus and pituitary ensures that the stress response is transient under normal conditions.

The POMC polypeptide is the precursor of ACTH. The POMC gene was found to be higher

expressed in the brain of the *Mecp2*$^{-/y}$ mouse as compared to the brain of its *wt* litter mate. *Sgk* and *Sult1A1* are reportedly induced by glucocorticoids (Lang and Cohen, 2001, Duanmu *et al*., 2001) and they were found to be up-regulated in the brain of the *Mecp2*$^{-/y}$ animal. Another up-regulated glucocorticoid-inducible gene, *Fkbp5*, encodes a peptidyl-prolyl cis-trans-isomerase associated with glucocorticoid receptor complexes (Yoshida *et al*., 2002). *Hsp105*, also called *Hsp110*, was down-regulated in the brain of the *Mecp2*-deficient animals (Fig. 29 A). Its expression is inhibited by the glucocorticoid dexamethasone (Wadekar *et al*., 2001). *Cirp* (the cold-inducible RNA binding protein gene) is induced by low temperature or low oxygen tension (Wellmann *et al*., 2004) but is not known to be induced by glucocorticoids.

3.3.3 Localization of MECP2, FKBP5, and SGK in mouse brain

Next, the immunolocalization of FKBP5, SGK (=SGK1) and MECP2 proteins in selected brain regions of adult female mice were examined. The results revealed a significant co-localization of FKBP5 with cells that also synthesize MECP2 (Fig. 16 A, B). A few MECP2-positive cells showed low or indetectable levels of FKBP5 staining (e.g. cortical cells in the position of Cajal-Retzius cells; Fig. 16 A). Many cells positive for SGK and MECP2 were found, though fewer than positive for FKBP5 and MECP2 (Fig. 16 C-E). In addition, cells producing either MECP2 or SGK are present in the brain (Fig. 16 D, E). As these animals had not been stressed or glucocorticoid-treated, the data suggest that presence of MECP2 in a neuronal cell is compatible with FKBP5 and SGK synthesis as proposed in Fig. 16. Therefore, MECP2 does not act as a transcriptional silencer on *Fkbp5* of *Sgk* in the brain.

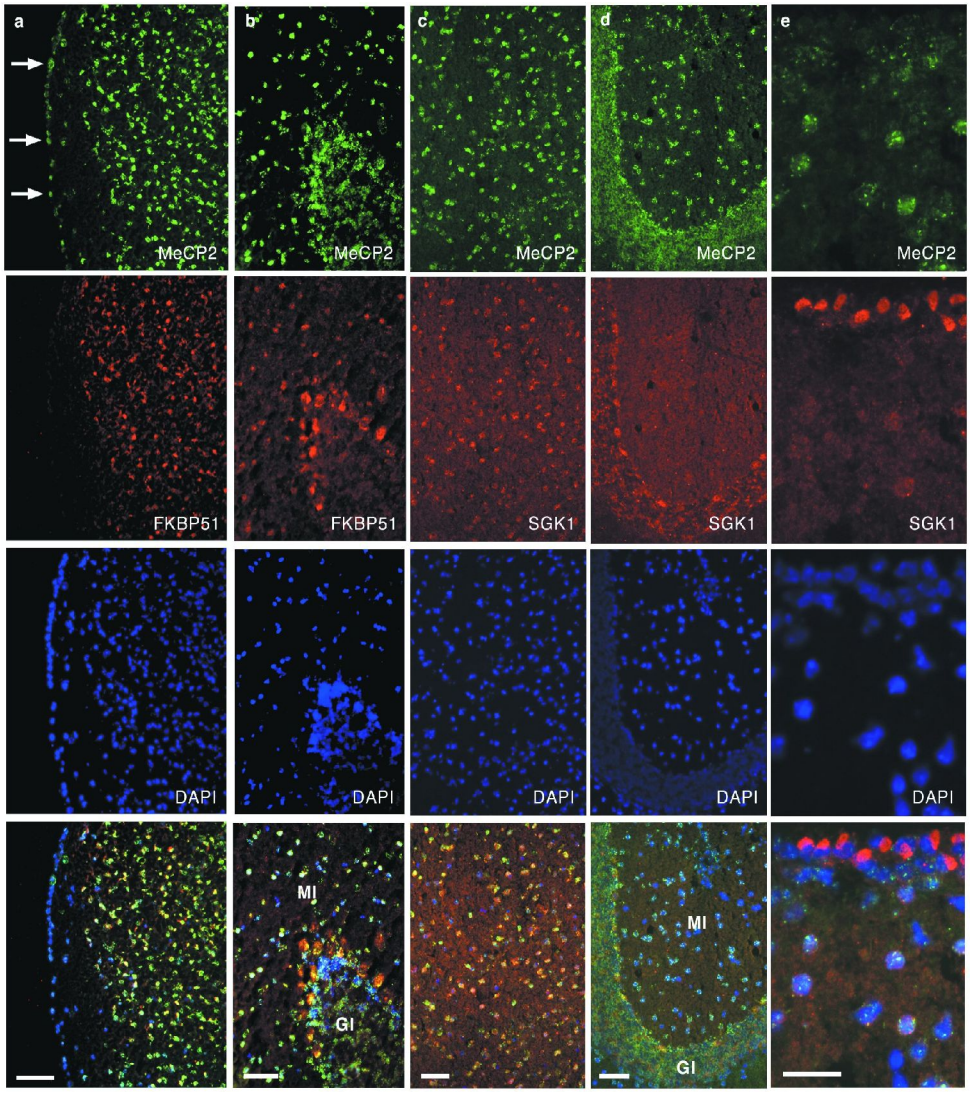

Fig. 16. Expression of MECP2, FKBP5, and SGK (=SGK1) in the brain of adult female mice. Immunolocalization of MECP2 (green signals, FITC-coupled secondary antibody), FKBP5 (red signals in A and B, Cy3-coupled secondary antibody), and SGK (red signals in C-E, Cy3-coupled secondary antibody) on cryostat sections of mouse brain. Panels show the temporal cortex (A), the cerebellum (B, D), the region immediately above the corpus callosum (C), and the lateral ventricle (E). Blue signals: DAPI-stained nuclei. Gl: granular cell layer. Ml: molecular cell layer. A clear cellular overlap of the staining with anti-MECP2 and anti-FKBP5 antibodies can be seen for the majority of cells in all brain regions. There are few MECP2-positive cells with an absent or weak FKBP5 expression: the most superficial cortical layer is FKBP5-negative (arrows in A) and some cells with weak FKBP5 signals can be seen in the molecular and granular cell layer of the cerebellum. Many cells co-express MECP2 and SGK in the brain, for example cells above the corpus callosum (C) and cells in the granular cell layer of the cerebellum (D). Distinct MECP2-positive cells in the molecular layer show only a weak homogeneous staining with the SGK antibody (D). The strongest SGK signals were detected in cells lining the ventricles (top cell layer in E). These cells show absent or very weak MECP2-staining in contrast to cells at the subluminal site. Scale bars in A equal 100 µm, in B, C, D 50 µm, and in E 25 µm.

3.3.4 Establishment of chromatin immunoprecipitation (ChIP)

To test the direct interaction of MECP2 and *Fkbp5*, the chromatin immunoprecipitation method was established and applied in this work. ChIP is the technique of choice to study the *in vivo* binding of DNA-associated proteins.

Even though laborious optimization of many parameters is required to obtain proper results with ChIP (for a review, see Das *et al*., 2004), the technique has been widely used in recent years. Among the conditions that have to be adjusted are the time and intensity of fixation, the purification of the input material (i.e. purification of cells, or even isolation of nuclei or chromatin) the shearing of the chromatin, the washing of the immunoprecipitated complexes, and the final detection.

For the experiments carried out in this thesis, fixation with 1% formaldehyde in the medium for 10 minutes at 37°C proved to be the most efficient way to bind the proteins of interest to the DNA. Using these settings, the MECP2 polypeptide was properly coupled to the DNA while most of the epitopes recognized by the polyclonal anti-MECP2 antibody were still accessible. Differences between batches of anti-MECP2 antibodies were however observed. Chromatin immunoprecipitation with antibodies for histones was generally more reliable and resulted in more precipitated DNA. This is probably due to the higher abundance of the histone proteins as compared to MECP2 or GR in the chromatin regions studied.

The shearing parameters for the fragmentation of the chromatin had to be optimized for the two types of biological material used (primary neurons and grounded brain tissue). Different sonication times and intensities were tested, using a Branson sonifier. Optimal results were achieved with 5 pulses of 30 seconds at 100 % duty and output level 5. The sonication was carried out in 2 ml buffer containing sodium dodecyl sulfate (SDS). The samples were kept in ice cold salt water to prevent the disruption of the DNA-protein complexes by heat as well as to hinder the formation of foam during the shearing procedure.

The described settings resulted in a smear of DNA ranging from 200 bp to about 1000 bp. These fragment sizes allowed a proper amplification after immunoprecipitation and isolation of the DNA.

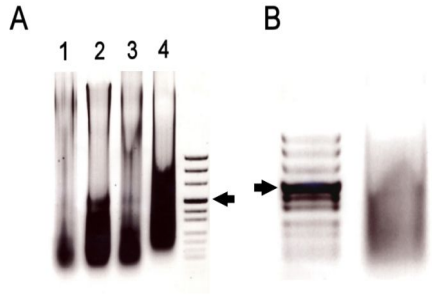

Fig. 17. DNA fragment distribution obtained by sonication of chromatin. (A) Lanes 1 to 4 depict the smears obtained by sonication with 2, 3, 4 and 5 pulses of 30 sec at output level 5, respectively. Line 4 represents the optimal DNA fragment size distribution for amplification with specific primers after ChIP. (B) DNA smear resulting from sonication (5 pulses of 30 sec, output level 5) of chromatin from brain tissue. Arrowheads indicate the pUC mix marker band at 501 bp.

3.3.5 MECP2 binding sites in the genomic regions of *Fkbp5*

Since the only DNA motif necessary for MECP2 binding is the dinucleotide m^5CpG (Nan *et al.*, 1993), it is not trivial to predict binding sites from the genomic sequence alone. Therefore, a comparative genomics approach was used to find genomic fragments conserved between the mouse *Fkbp5* and the human *FKBP5 (*see methods section 2.2.1.1). This data was compared to binding sites of known transcription factors which were mapped to the *Fkbp5* genomic region. To do so, all motifs of the Transfac database were blasted against the *Fkbp5* region of in the mouse genome. Finally, the CpG content of the sequence was considered, since CpG islands are often found in promoters of genes, but are usually unmethylated and hence not a promising binding region for MECP2. Primer pairs were designed covering regions of about 1 kb at the two most promising target sites (region 1 with primer pairs 1_0 to 1_4 and region 2 with primer pairs 2_1 to 2_4, see Fig. 18).

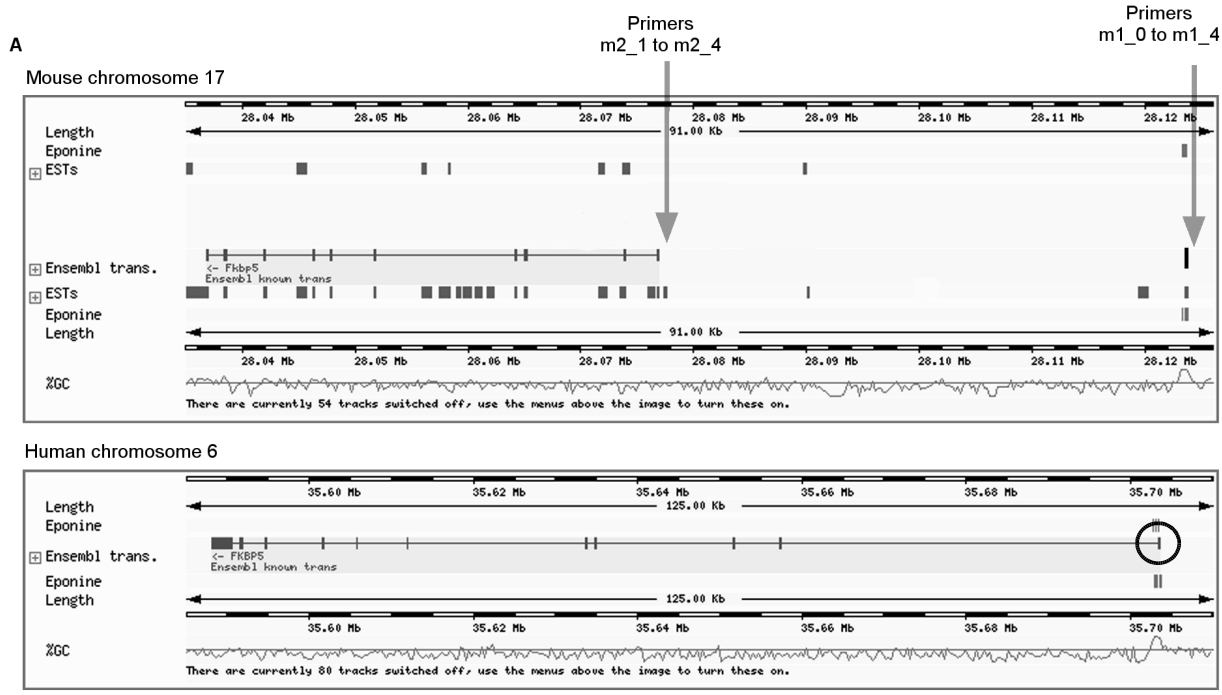

Fig. 18. Comparison of human *FKBP5* and mouse *Fkbp5* genomic regions. Ensembl representation of sections of human chromosome 6 and mouse chromosome 17 depicts the known transcripts, 5' to 3', right to left. Eponine denotes regions predicted to contain transcription start sites and promoters. The % GC curves show the percentage of GC base pairs in the sequence. Arrows indicate the regions covered with primers to find MECP2 binding sites. The black circle depicts the human exon 1 that was missing in the mouse transcript.

In the Ensembl database (release 24.33.1) the mouse *Fkbp5* transcript was missing the first exon present in the human transcript (Fig. 18). Thus, an RT-PCR experiment was performed with one primer corresponding to a sequence in the fourth exon and the other primer corresponding to a sequence in the first and second exon of the predicted transcript. The size of the PCR band confirmed the existence of a mouse transcript that contains an exon (similar to the first exon of *FKBP5* in the human genome) upstream of the transcription start in the Ensembl database (Fig. 19).

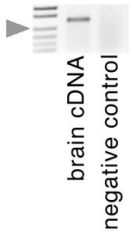

Fig. 19. PCR using mouse brain cDNA with primers specific for mouse *Fkbp5*. The first exon of *Fkbp5* is not part of the Ensembl known transcript (ENSMUST00000062167), but was present as EST in the database. RT-PCR revealed a cDNA fragment that corresponds to a sequence from exon 1 to exon 4 in brain tissue. Negative control: no reverse transcriptase was added to the reaction. Arrowhead: 242 bp DNA marker.

To confirm the binding of MECP2 to the two regions 1 and 2 of *Fkbp5*, ChIP was performed with whole mouse brain tissue (Fig. 20). Specific interactions could be detected in both regions in *wt* brains. As a negative control, *Mecp2*-null mouse brain tissue was used. In this case, no PCR products were obtained. This argues for the specificity of the antibody and against a cross-reaction with a protein other than MECP2. An antibody against acetylated histone H3 was used as positive control.

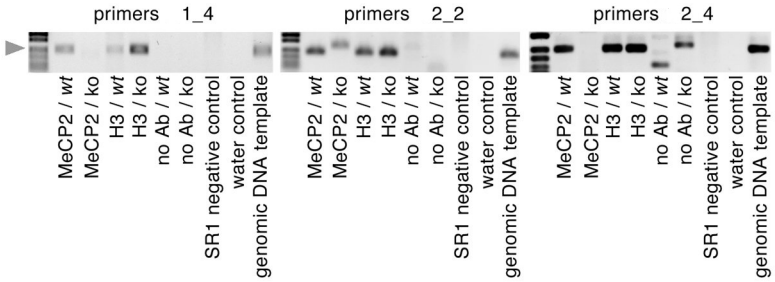

Fig. 20. Chromatin immunoprecipitation shows binding of MECP2 to the genomic region of *Fkbp5*. ChIP with antibodies against MECP2 and acetylated histone H3 using total brain tissue. MECP2 binds to region 1_4 and region 2 (2_2 and 2_4) of the *Fkbp5* gene in *wt* brain. Arrowhead: 242 bp DNA marker. SR1 negative control: no template was used in the SR1 PCR reaction. Water control: H_2O instead of template was used in the PCR with specific primers.

Reliable binding of MECP2 to the *Fkbp5* genomic regions 1_4, 2_2, and 2_4 could be shown in this thesis, as well as to the regions fkp and fkp1 detected by our collaborators in Edinburgh. The relative position of the primers can be seen in Fig. 21. The regions 1_4, 2_1 and 2_2 are poor in CpG content while the regions fkp1 and 2_4 contain many CpGs (Fig. 21). 1_4 and fkp1 flank a CpG island.

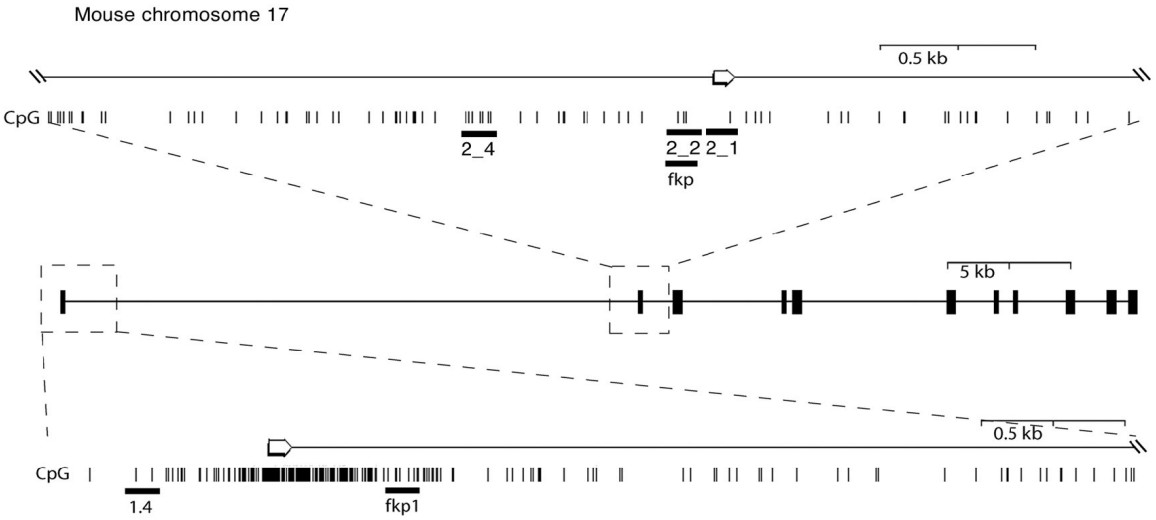

Fig. 21. Schematic representation of the mouse *Fkbp5* gene. Positions of primer pairs used for the amplification of MECP2 antibody precipitated *Fkbp5* genomic DNA fragments are indicated by black horizontal bars. The positions of CpGs present in the region are visualized by vertical bars.

MECP2 only binds to methylated DNA. To determine whether the potential binding sites of MECP2 determined by ChIP are methylated, bisulfite sequencing was performed. Genomic DNA from *wt* mouse brain was treated with sodium bisulfite, amplified with primers specific for the genomic regions 1_4, 2_1, 2_2, and 2_4 after bisulfite conversion, cloned into a vector, and sequenced. Since sodium bisulfite deaminates unmethylated cytosine residues which results in an uracil residue, only m^5C still produced a cytidine in the sequencing output. The CpGs in the regions studied were mainly methylated, supporting the idea that these are binding sites of MECP2 (Fig. 22).

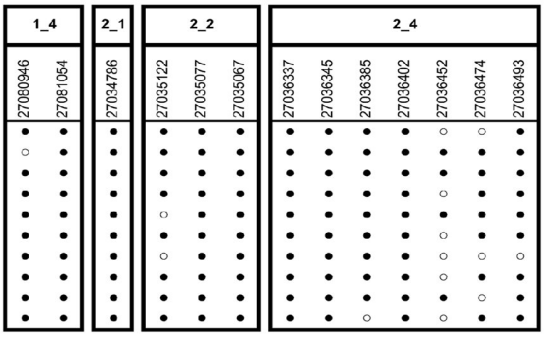

Fig. 22. Methylation status of CpGs in *Fkbp5* genomic DNA determined by bisulfite sequencing. The four genomic regions, 1_4 and 2_1, 2_2 (= fkp), and 2_4 were analyzed. DNA was derived from the brain of a *wt* mouse. For each region, 10 clones were sequenced. Numbers correspond to the location on Ensembl mouse genome chromosome 17 (release 24.33.1) Methylated CpGs are depicted by filled circles; unmethylated CpGs by open ones.

3.3.6 MECP2 and the glucocorticoid receptor compete for binding to the GRE_2 locus

In response to glucocorticoids, the glucocorticoid receptor has been shown to bind to human *FKBP5* and induce its expression (U *et al*., 2004). So far, no studies on GR binding sites in the mouse *Fkbp5* genomic region were done.

Human and mouse *Fkbp5* genomic sequences were compared to identify the region in the mouse genome corresponding to the GR binding site in the human genome. Such a region, however, could not be detected in the mouse *Fkbp5* gene. Therefore, the Transfac database was used to search for all regions in the *Fkbp5* gene that show similarities to the GR binding site consensus sequence stored in the database. Six such loci, called glucocorticoid response elements (GRE), could be identified in this thesis (Fig. 23).

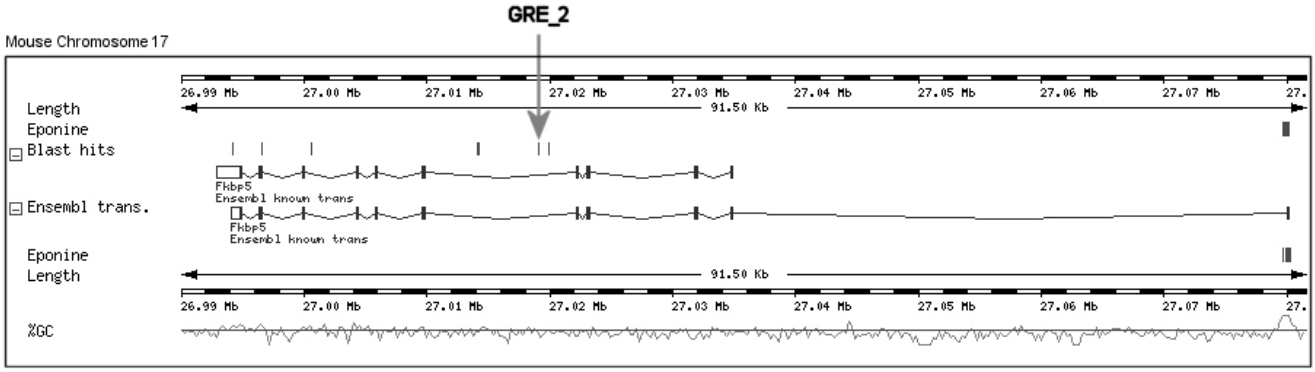

Fig. 23. *Fkbp5* intron/exon structure and the localization of GR response elements. Two Ensembl transcripts (starting at the two promoters, respectively) are represented by black squares (exons), 5' to 3', right to left. Eponine denotes regions predicted to contain transcription start sites and promoters. The % GC curves shows the percentage of GC base pairs in the sequence. Glucocorticoid receptor binding sites are indicated as blast hits and the binding site common to GR and MECP2 (GRE_2) is marked by a grey arrow.

In order to test whether MECP2 binds to the same genomic regions as GR, chromatin immunoprecipitation experiments were performed with antibodies against MECP2 and GR. Primary neurons were treated with an synthetic glucocorticoid (dexamethasone), with a glucocorticoid inhibitor (RU-486), with both, dexamethasone and RU-486, or with ethanol as solvent control.

Fig. 24. MECP2 and GR both bind to the GRE_2 genomic region. Chromatin immunoprecipitation results reveal MECP2 binding to the GRE_2 region under normal conditions (Ethanol) and in the presence of both, RU-486 (RU) and dexamethasone (Dex). In contrast, GR only binds when ethanol or dexamethasone are present, but not if RU-486 is added to the medium. Reactions without antibody (No Ab) served as negative controls and genomic DNA as positive control.

Fig. 24 illustrates, that under normal conditions MECP2 and GR bind to the GRE_2 region. MECP2 binding is abolished in the presence of the glucocorticoid dexamethasone. In contrast, GR binding can be reversed by addition of RU-486 which restores MECP2 binding. These results suggest a model in which MECP2 and GR compete for binding to the GRE_2 region and the repressor function of MECP2 is fine-tuned by glucocorticoids. Addition of RU-486 alone led to dissociation of MECP2 as well as GR from the GRE_2 region. This suggests, that excessive amounts of RU-486 disturb the gene expression regulation of *Fkbp5* in that neither the repressor MECP2 nor the activator GR binds to GRE_2 under this condition.