The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation

Marco Birke, Silke Schreiner, María-Paz García-Cuéllar, Kerstin Mahr¹, Fritz Titgemeyer¹ and Robert K. Slany*

Department of Genetics and ¹Department of Microbiology, University of Erlangen, Staudtstrasse 5, 91058 Erlangen, Germany

Received October 24, 2001; Revised and Accepted December 8, 2001

ABSTRACT

Alterations of the proto-oncogene MLL (mixed lineage leukemia) are characteristic for a high proportion of acute leukemias, especially those occurring in infants. The activation of MLL is achieved either by an internal tandem duplication of 5' MLL exons or by chromosomal translocations that create chimeric proteins with the N-terminus of MLL fused to a variety of different partner proteins. A domain of MLL with significant homology to the euakaryotic DNA methyltransferases (MT domain) has been found to be essential for the transforming potential of the oncogenic MLL derivatives. Here we demonstrate that this domain specifically recognizes DNA with unmethylated CpG sequences. In gel mobility shifts, the presence of CpG was sufficient for binding of recombinant GST–MT protein to DNA. The introduction of 5-methylCpG on one or both DNA strands precluded an efficient interaction. In surface plasmon resonance a K₅ₐ of ∼3.3 × 10⁻⁸ M was determined for the GST–MT/DNA complex formation. Site selection experiments and DNase I footprinting confirmed CpG as the target of the MT domain. Finally, this interaction was corroborated in vivo in reporter assays utilizing the DNA-binding properties of the MT domain in a hybrid MT–VP16 transactivator construct.

INTRODUCTION

The mixed lineage leukemia (MLL) gene at the genomic locus 11q23 is a frequent target of chromosomal aberrations. These events activate the proto-oncogene MLL and are associated with an especially aggressive class of leukemias. 11q23 abnormalities are most prevalent in acute lymphoblastic and acute myeloid leukemias during the infant age and in therapy-related secondary myeloid leukemias after a previous treatment with topoisomerase II inhibitors (reviewed in 1,2).

The unaltered MLL gene codes for a large 3968 amino acid protein with limited but significant homology to Drosophila trithorax (TRX) (3). In the fly TRX is the prototype member of a large genetically defined group of proteins (TRX-G) with a positive influence on transcription. TRX-G members are responsible for epigenetic gene regulation by establishing a transcriptionally active, 'open' state of chromatin that can be inherited through cell division. Thus, TRX is necessary for the continued maintenance, but not initiation, of a characteristic temporal and spatial expression pattern of the Hox-cluster genes (HOM-C) (reviewed in 4–6).

A similar function was demonstrated for MLL in mouse knockout studies. While MLL−/− embryos initially were able to establish normal Hox expression domains, they could not maintain Hox gene transcription at later stages and died around day 9.5 of embryonic development (7–9). Support for the role of MLL as a regulator of chromatin architecture comes from various biochemical studies that demonstrate a direct physical interaction of MLL with members of the chromatin remodeling machinery. A highly conserved SET domain at the very C-terminus of MLL (Fig. 1A) mediates the association of MLL with INI-1, a component of the human SWI/SNF chromatin-remodeling complex (10). Another region within the C-terminal portion of MLL was originally identified as a trans-activation domain in a heterologous fusion with the GAL4 DNA-binding domain (11,12). Recent reports show that this part of MLL binds to the histone deacetylase and nuclear co-activator CREB-binding protein (CBP) (13).

In the majority of leukemias with altered MLL the complete C-terminus is removed by the chromosomal translocation and replaced in frame by a variety of different fusion partners. The resulting fusion protein acquires novel properties and the leukemogenic potential of this type of MLL fusion has been directly demonstrated in mouse models and in in vitro transformation tests (14–21). Despite the heterogeneity of the fusion partners MLL always contributes the identical N-terminal portion to the various chimeric proteins. A second type of leukemogenic MLL alteration corroborates the importance of the MLL N-terminus. In this case, the same N-terminal moiety of MLL that is retained in the fusion proteins is duplicated in tandem (22–24). This leads to the expression of a MLL protein where the N-terminus is ‘fused’ to itself (Fig. 1A).

In a structure–function analysis (21) two domains within MLL were identified that are essential for the transforming capacity of an MLL fusion protein. The first one is comprised of three AT-hook motifs that occur close to the start of MLL.

*To whom correspondence should be addressed. Tel: +49 9131 8528527; Fax: +49 9131 8528526; Email: rslany@biologie.uni-erlangen.de
Here we report that the MT domain specifically recognizes CpG dinucleotide sequences in band shift and competition experiments. Binding is abolished if the cytosine in CpG is methylated on one or both DNA strands. Quantitative measurements with surface plasmon resonance (SPR), site selection and footprinting analysis confirmed this result. As demonstrated by reporter gene studies in transfected cells this interaction also persists in vivo.

**MATERIALS AND METHODS**

**GST-fusion protein expression and purification**

For bacterial expression of GST–MT fusion proteins the cDNA coding for the MT domain was amplified by PCR with the MT-specific primers BamMTfw (5′-caaggagctcaataaagaag-gacgctgagctgaggcgg-3′) and EcoMTrev (5′-acacgaatccacagtt-tctgtactagtcacctag-3′). The PCR product was cloned into the bacterial expression vector pGEX-3X (Amersham Pharmacia Biotech AB, Uppsala, Sweden) and subsequently electro- 

**Gel mobility shift analysis**

To analyze the DNA-binding properties of the MT domain the following synthetic double-stranded oligonucleotides were produced by annealing of the corresponding single-stranded precursors: CpG52 (5′-tacccgttgaggccactaagcttagtgacagggagctcattgcct- 

**SPR measurement (BIACore)**

The kinetic binding constants of the GST–MT–DNA complex were determined with a BIACoreX-machine (BIAcore AB, Uppsala, Sweden). The primers CpG52fw (5′-tacccgttgaggccactaagcttagtgacagggagctcattgcct- 

**Figure 1.** Structure of MLL, the MT domain and purification of GST–MT protein. (A) Graphical representation (not to scale) of MLL and its oncogenic derivatives. Conserved domains are indicated as hatched bars, the MT domain is depicted in black. The oncogenic alterations are indicated. AT-hook, AT DNA-binding motif; PhD, plant homeodomain protein–protein interaction motif; TA, transacting domain interacting with CBP; SET, conserved domain. (B) Amino acid sequence of the MT domain as used in the experiments as a fusion with GST. The CGxCxxC core is boxed, basic amino acids are shaded. (C) Coomassie-stained SDS–polyacrylamide gel showing purified GST–MT fusion protein. The CGxCxxC core is boxed, basic amino acids are shaded. (D) Coomassie-stained SDS–polyacrylamide gel showing purified GST–MT fusion protein.
or 3′-biotin-GpC52 oligonucleotides were produced by dimerization with the corresponding unlabeled primer. Oligonucleotides were immobilized separately on the flow-cell surfaces of a streptavidin (SA)-coated biosensor chip (BIACore AB) by injecting twice 5 pmol 3′-biotin-labeled ds-oligonucleotides in a total volume of 50 µl x coupling buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% poly-sorbate 20) at a flow rate of 5 µl/min followed by a wash cycle with 1× binding buffer [6 mM Tris–HCl, 6 mM MgCl2, 150 mM NaCl, 1 mM DTT, 10 ng/µl BSA, 100 ng/µl poly(dA•dT), pH 7.9]. Dilutions of GST–MT protein were prepared in 1× binding buffer as indicated and 20 µl of each dilution were injected to both flow cells at a flow rate of 5 µl/min per assay cycle. Bound protein was eluted by injecting 20 µl 1× regeneration buffer (2.5 M NaCl, pH 5.3) followed by a wash cycle with 1× binding buffer between assay cycles. Data were analyzed with the BIAevaluation software (BIACore AB) and the dissociation constant (Kd) was calculated from the ratio of the dissociation rate (kd) and association rate (ka).

**PCR-based site selection**

A pool of double-stranded 74 bp oligonucleotides containing a central randomized 26 bp sequence was produced by annealing equimolar amounts of primers R74poly(N) [5′-caggctagcagggcatctgc-(N)26-gagggaattcagtgcaactgcagc-3′] and F74 [5′-ggtcagctgcaactgcagc-3′] and subsequent elongation by Klenow polymerase. Double-stranded DNA was purified and eluted by TAE-agarose gel electrophoresis and DNA contents were determined photometrically. Complex formation was performed by incubating 500!ng GST–MT protein with 100!ng of eluted ds74 bp oligonucleotides in a total volume of 20 µl with 1× binding buffer (see Gel mobility shift analysis) at standard conditions. Complexes were precipitated with GST–agarose beads (Sigma-Aldrich, Deisenburg, Germany) for 30 min at 4°C in a rotator shaker and complexed DNA was amplified in a PCR reaction with specific primers F74 and R74 (5′-caggctagcagggcatctgc-3′). After three rounds of precipitation and amplification the PCR products were 32P-end-labeled and 250 000 c.p.m. (Cerenkov radiation) were used for a gel shift with 1 µg GST–MT protein at standard conditions. Shifted complexes were purified and eluted by native polyacrylamide gel electrophoresis followed by a second round of PCR amplification and gel shift purification. Retarded fragments were amplified once more by PCR and cloned into the vector pBluescript-SK+ (Stratagene, La Jolla, CA). Randomly picked inserts were sequenced with an ABI Prism 310 sequencing machine (PE Biosystems, Foster City, CA).

**DNase I protection assays**

A double-stranded 74 bp DNA fragment (5′-caggctagcagggcatctgcggccgttgtggatatcggttagaattcaaa-3′ and MT-2rev (5′-tgtggatatcggttagaattcaaa-3′)) and F74 (5′-ggtcagctgcaactgcagc-3′) and subsequent elongation by Klenow polymerase. Double-stranded DNA was purified and eluted by TAE-agarose gel electrophoresis and DNA contents were determined photometrically. Complex formation was performed by incubating 500!ng GST–MT protein with 100!ng of eluted ds74 bp oligonucleotides in a total volume of 20 µl with 1× binding buffer (see Gel mobility shift analysis) at standard conditions. Complexes were precipitated with GST–agarose beads (Sigma-Aldrich, Deisenburg, Germany) for 30 min at 4°C in a rotator shaker and complexed DNA was amplified in a PCR reaction with specific primers F74 and R74 (5′-caggctagcagggcatctgc-3′). After three rounds of precipitation and amplification the PCR products were 32P-end-labeled and 250 000 c.p.m. (Cerenkov radiation) were used for a gel shift with 1 µg GST–MT protein at standard conditions. Shifted complexes were purified and eluted by native polyacrylamide gel electrophoresis followed by a second round of PCR amplification and gel shift purification. Retarded fragments were amplified once more by PCR and cloned into the vector pBluescript-SK+ (Stratagene, La Jolla, CA). Randomly picked inserts were sequenced with an ABI Prism 310 sequencing machine (PE Biosystems, Foster City, CA).

**RESULTS**

The MT domain binds to unmethylated CpG in gel shift analysis

To avoid a possible experimental interference by the DNA-binding activity of the AT-hooks we chose to study the properties of the MT domain as an isolated unit. Therefore, MT protein was produced fusing the cDNA encoding amino acids 1147–1244 of MLL (GenBank accession no. AAA58669) to the gene for glutathione S-transferase in the commercial expression vector pGEX-3X. The corresponding 37 kDa GST–MT fusion protein was purified from E.coli by a one-step glutathione affinity procedure. Approximately 90% pure protein was obtained as judged by SDS–PAGE followed by Coomassie staining (Fig. 1C). All further in vitro experiments have been performed with the GST–MT fusion protein.

The presence of the CxxC motif in the MT domain prompted us to investigate a possible association of the recombinant MT
fusion protein with CpG-containing DNA. For this purpose, a double-stranded 52 bp oligonucleotide (CpG52) containing six regularly spaced CpG dinucleotides was synthesized. As a control, a second oligo (GpC52) with the same sequence and overall base composition was used with the exception that all CpG combinations were now inverted to GpC (Fig. 2A). In gel shift experiments the addition of increasing amounts of GST–MT protein produced a retarded band in a dose-dependent fashion with labeled CpG52 oligonucleotide. However, no interaction could be detected with the GpC52 probe (Fig. 2B, left). GST protein alone did not bind to any of the probes tested (data not shown). These results were confirmed in a competition experiment. Whereas the addition of a surplus of unlabeled CpG52 oligo competed efficiently with the binding of GST–MT, a 100-fold excess of GpC52 did not affect the protein–DNA interaction (Fig. 2B, right).

Cytosines in CpG dinucleotides of mammalian DNA are the frequent targets for methylating enzymes that convert the base to 5-methylcytosine (5mC). To determine the influence of cytosine methylation on the DNA-binding capacity of GST–MT, oligos identical in sequence to the CpG52 probe were used but now with 5mC replacing the C of CpG sequences either in both strands (methylated DNA) or in the top or bottom strand only (hemimethylated DNA). In gel shifts this exchange of C to 5mC drastically diminished the affinity of GST–MT for DNA. Compared with the unmethylated CpG52 probe only residual binding at high concentrations of GST–MT remained visible with either fully methylated or hemimethylated CpG52 substrate (Fig. 2C, top). This was supported by the fact that neither fully methylated nor hemimethylated DNA could efficiently compete for the binding of GST–MT to unmethylated CpG52 (Fig. 2C, bottom left).

Initial observations had indicated that the addition of ‘unspecific’ poly(dIdC) competitor could completely inhibit the association of GST–MT protein with DNA (data not shown). As poly(dIdC) resembles homopolymeric CpG stretches the influence of poly(dIdC) on GST–MT binding was studied in a competition test. In this experiment the addition of an excess of poly(dIdC) effectively competed with the complex formation between GST–MT and the CpG52 oligo whereas the same amount of poly(dA+dT) had no visible consequence (Fig. 2C, bottom right).

**Determination of the binding constants in solution by SPR**

For quantitative measurements of the specific binding affinity of the GST–MT protein, a kinetic binding study with a SPR biosensor (BIAcore) was carried out. In this experiment the kinetic association and dissociation rates of complexes formed between molecules in solution and substrates pre-bound to a solid surface are determined by measuring the reflecting index change in arbitrary units of response (RU). For this purpose biotinylated CpG52 oligonucleotide was immobilized on the streptavidin sensor chip and variable concentrations of GST–MT protein were applied to the surrounding flow chamber. An example of the corresponding response curves is given in Figure 3. From the obtained initial rates of association and dissociation a $K_D$ of $3.28 \pm 1.7 \times 10^{-8}$ M (mean ± SD, $n = 8$) could be estimated. When repeated with fully methylated CpG52 oligonucleotide a very weak binding of GST–MT was observed at the highest possible protein concentrations (data not shown). Therefore, the affinity constant of methylated DNA could not be determined with accuracy. A rough estimate indicated a $K_D$ of $>1$ mM for the interaction of GST–MT with methylated template.

**CpG is the preferred binding site of the MT domain in site selection experiments**

Site selection studies were carried out to determine if the MT domain has any binding preferences outside the minimally required CpG sequence. For this purpose, a 74 bp oligonucleotide
was synthesized that was composed of 26 random nt in its center flanked by known sequence without any CpG (N74) (Materials and Methods). Two primer molecules complementary to the flanking sequence allowed the amplification of the whole 74 bp DNA by PCR. The single-stranded N74 was converted to double-strand DNA by the Klenow fragment of *E. coli* DNA polymerase I and one of the flanking primers. The resulting DNA fragment was gel-purified, eluted and incubated in solution with GST–MT protein. Bound DNA was precipitated together with the GST–MT protein by binding to glutathione immobilized on agarose beads. After a wash step, the bound fraction of DNA was amplified by PCR. The gel-purified PCR products were subjected to two more rounds of precipitation/amplification. To increase binding stringency, the selection rounds in solution were followed by two rounds of gel shifts. To this end, the PCR product was end-labeled with [\(^{\gamma}-32P\)ATP and polynucleotide kinase, reacted with GST–MT in binding buffer and applied to a native polyacrylamide gel. The retarded band was excised, eluted and re-amplified by PCR to serve as the input for the final round of gel selection. Finally, the end products of the selection process were cloned into the pBluescript vector and sequenced. Of 19 clones analyzed all contained from three to seven CpG dinucleotides randomly distributed in the center of the selection oligonucleotide (mean 4.6 ± 1.2 CpG repeats) (Fig. 4). Outside the CpG core there was no consensus sequence discernible and the nucleotides flanking each CpG followed a random distribution. In contrast, none of 19 indiscriminately picked samples from the unselected oligo pool included more than two CpG dinucleotide sequences (mean 0.7 ± 0.7 CpG motifs). These results confirm CpG as apparently the sole binding requirement for the MT domain protein.

**GST–MT protects CpG in DNase I footprinting analysis**

Next we wanted to determine the extent of DNA that is protected from DNase I digestion by the interaction with the GST–MT protein. As substrate for footprinting a high-affinity binding clone from the site selection experiment (clone 1) was chosen. The corresponding fragment was excised from the respective pBluescript plasmid and the top or bottom strand was labeled specifically by PCR amplification utilizing a 5′ [\(^{\gamma}-32P\)ATP marked primer. Complexes of end-labeled DNA with GST–MT were subjected to DNase I treatment. The analysis of the reaction products by denaturing PAGE showed three distinct regions protected from DNase I digestion, all centered on CpG dinucleotides (Fig. 5). One area of protection was visible extending from nucleotides 19 to 30 on the top strand. The exact 5′ boundary of this region could not be exactly determined due to the proximity of the labeled 5′ end to the GST–MT-binding site, thus resulting in digestion fragments too small to be separated on the gel. On the bottom strand two more GST–MT footprints were detectable that each covered 5–6 bp (nucleotides 35–40 and 46–51). As the digestion pattern indicated, the binding of GST–MT rendered the DNA outside of the protected center of the molecule increasingly DNase I sensitive with the highest susceptibility for cleavage towards the end of the molecule. Increased band intensity after DNase I treatment could be seen for the top strand starting at nucleotide 40 and extending to the end of the fragment at nucleotide 74. The corresponding pattern indicated, the binding of GST–MT rendered the DNA outside of the protected center of the molecule increasingly DNase I sensitive with the highest susceptibility for cleavage towards the end of the molecule. However, circular permutation experiments did not show any sign of a bend or kink in GST–MT-bound DNA (data not shown).

**The MT domain associates with CpG-containing DNA in vivo**

To prove that the MT domain can bind to CpG also inside a cell, reporter gene assays were performed. The rationale of

---

**Figure 3.** $K_d$ determination for the GST–MT/DNA interaction by BIAcore experiments. The CpG52 oligonucleotide probe was immobilized on a plasmon resonance sensor chip and the kinetic association/dissociation constants were determined in a flow cell charged with the given concentrations of GST–MT protein. The curves show the graphical representation of the measurements in a typical experiment. The kinetic values represent the mean ± SD ($n = 8$).

**Figure 4.** GST–MT prefers CpG sequences in site selection experiments. Site selection experiments (Materials and Methods) were performed with oligonucleotides containing a core of 26 random nt. The sequences of 19 clones with high affinity to GST–MT are shown. The vertical lines indicate the extent of the original randomly designed sequence with flanking nucleotides given if they contribute to form a CpG dinucleotide. CpG sequences are boxed.
Three different reporter constructs depicted schematically in Figure 6 were applied: (i) a luciferase gene under the control of the minimal SV40 promoter that contains several naturally occurring CpG motifs (pGL3 promoter); (ii) a related construct with six tandem copies of the core sequence of site selection clone 1 preceding the SV40 minimal promoter and therefore increasing the CpG density in the promoter region (pGL3-CG); and (iii) a plasmid with three consensus-binding sites for the GAL4 DNA-binding domain upstream of the SV40 minimal promoter (pGL3-GAL4).

When cotransfected with pGL3 promoter into 293T cells the MT–VP16 fusion was able to increase the luciferase activity modestly but significantly by ∼2-fold. The transactivation levels were lower as might be expected because of a counter-active repressor activity within the MT domain that was revealed by expression of the MT portion alone (Fig. 6A). Despite the overall weak transactivation potency of the MT–VP16 protein an increase in CpG density of the reporter construct (pGL3-CG) lead to a correspondingly higher transactivation efficiency (Fig. 6B, left). To determine the influence of DNA methylation on the MT–VP16 activity in vivo a pGL3-CG reporter was transfected after pre-methylation of the cytosine residues in CpG sequences by SssI methylase. This pre-treatment reduced the basal promoter activity of the reporter construct to ∼10% compared with the value obtained with unmethylated plasmid. Corresponding to the in vitro-binding results, the transactivation effect of MT–VP16 was completely abolished with this methylated target (Fig. 6B, right). In contrast, when MT–VP16 was directly targeted to the promoter by the GAL4 DNA-binding domain, transactivation could still be seen on methylated as well as native CpG targets (Fig. 6C).

**DISCUSSION**

In this paper we report experiments identifying the MT homology region of MLL as a DNA-binding domain that specifically recognizes unmethylated CpG sequences. In band shift analysis and SPR studies, recombinant GST–MT protein binds with high affinity to sequences containing CpG. In site selection experiments, CpG is the preferred binding target of GST–MT and subjected to DNase I digestion. Moreover, a binding of the MT domain to CpG-containing DNA can be demonstrated also inside living cells. A conversion of cytosine to 5mC eliminates the tight association with regard to function, two more proteins use CxxC domains for DNA binding. The clearest example is the CpG-binding protein CGBP1 that also contains next to the CxxC motif a DNA binding. The most complicated situation is encountered with the methyl-binding protein MBD1 (27,28). MBD1 includes three CxxC motifs and isoforms are produced by differential splicing that encompass either two or three of these domains. MBD1 binds to methylated as well as unmethylated DNA and it acts as a
transcriptional repressor. Despite the presence of three CxxC motifs only one of these seems to be involved in binding to unmethylated DNA and all are dispensable for the association with methylated DNA. A separate methyl-binding domain in MBD1 is responsible for the interaction with methylated sequences. In summary, the CxxC motif seems to be the signature of a special structural fold that confers a CpG-binding activity. However, the overall homology amongst the various CxxC domains is only ~40% and therefore it cannot be excluded that variable sequences outside of the conserved CGxGxxC core contribute to specific functions in individual proteins.

In our experiments the overexpression of an isolated MT domain had a repressive effect on the transcription of a reporter gene. This fact was noted previously with similar templates containing GAL4 sites and constructs that fused a GAL4 DNA-binding domain to MT (12). The underlying repressive action of MT most likely is one reason why we observed only moderate transactivation levels in the reporter gene assays despite the fusion with the strong VP16 transactivator region. Although an interaction with histone deacetylase has been demonstrated to exist, the repressor mechanism of the MT domain is not clear. The MT domain has a high affinity to CpG sequences with a K_0 of 4.8 \times 10^{-8} M range. This is comparable with the values obtained in similar experiments, e.g. for the HMG1/DNA interaction with a K_0 of 4.8 \times 10^{-8} (33). Maybe an isolated MT domain can compete for binding with other DNA-binding proteins, e.g. general transcription factors and in this way cause an unspecified repressive effect.

The homology to fly Trithorax, the knockout studies and the biochemical evidence (3,6,7,10,13) strongly suggest that the function of MLL is to maintain pre-activated promoters in an active transcriptional state by changing the chromatin configuration of the respective DNA. For this purpose MLL needs to recognize active promoters. It is easily visible how a domain that mediates binding to unmethylated CpG islands and prevents the association with DNA that is blocked by methylation could assist in this process. The unique combination of an AT-hook DNA-binding motif and the CpG recognition domain in MLL would target the protein to a specific subset of response elements where AT-rich and/or cruciform DNA neighbors sequence with a high frequency of CpG. The verification of this prediction will have to await the identification of MLL target genes. However, the fact that both the AT-hooks and the MT domain are strictly conserved in the oncogenic MLL proteins strengthens the argument that it is the deregulation of authentic MLL targets that is responsible for the leukemogenic effect. Consequently, a protein that has lost a crucial domain for target finding would be unable to home in on the appropriate promoters and therefore it seems unlikely that such a truncated protein might cause leukemia.

ACKNOWLEDGEMENTS

The authors wish to thank Georg Fey for continuing support. This work was supported by grant SFB473/B10 and partially by grants SFB466/C7 and SL27/4-1 from the DFG. R.K.S. is a recipient of a Ria Freifrau-von-Fritsch Stiftung career development award.

REFERENCES


