Published online 25 July 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 15* ***5073–5084***
doi:10.1093/nar/gkm504

Structural basis for recognition of the matrix attachment region of DNA by transcription factor SATB1

Kazuhiko Yamasaki[1],*, Toshihiko Akiba[2], Tomoko Yamasaki[1] and Kazuaki Harata[2]

[1]Age Dimension Research Center and [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Received March 16, 2007; Revised and Accepted June 11, 2007

ABSTRACT

Special AT-rich sequence binding protein 1 (SATB1) regulates gene expression essential in immune T-cell maturation and switching of fetal globin species, by binding to matrix attachment regions (MARs) of DNA and inducing a local chromatin remodeling. Previously we have revealed a five-helix structure of the N-terminal CUT domain, which is essentially the folded region in the MAR-binding domain, of human SATB1 by NMR. Here we determined crystal structure of the complex of the CUT domain and a MAR DNA, in which the third helix of the CUT domain deeply enters the major groove of DNA in the B-form. Bases of 5′-CTAATA-3′ sequence are contacted by this helix, through direct and water-mediated hydrogen bonds and apolar and van der Waals contacts. Mutations at conserved base-contacting residues, Gln402 and Gly403, reduced the DNA-binding activity, which confirmed the importance of the observed interactions involving these residues. A significant number of equivalent contacts are observed also for typically four-helix POU-specific domains of POU-homologous proteins, indicating that these domains share a common framework of the DNA-binding mode, recognizing partially similar DNA sequences.

INTRODUCTION

Association of the matrix attachment regions (MARs) in the chromosomal DNA with the nuclear matrix organizes the higher order structure of the genome, forming looped structures that are likely to be equivalent to active chromatin domains in terms of transcription as well as replication (1). The MAR sequences are generally AT-rich at 70% and possess potential of DNA bending (2,3). They have regions where base pairs tend to break under an unwinding stress (base-unpairing region: BUR), centered at a sequence ATATAT, which was termed as BUR nucleation sequence (4). The tendency of base unpairing in the MAR DNA was shown to be essential in binding to the nuclear matrix and enhancing the promoter activity (5).

Special AT-rich sequence binding protein 1 (SATB1) was originally isolated as a factor that specifically binds to the BUR sequence, where mutation of the ATATAT sequence impaired the binding (6). SATB1, predominantly expressed in thymocytes, recruits histone deacetylase complex to the MAR site inside the interleukin-2 receptor α gene, in order to repress its expression (7). SATB1-null mice exhibit irregular expression of the above gene and related genes in the premature $CD4^+CD8^+$ T-cells, which caused small thymi and spleens, and death at age of 3 weeks (8). SATB1 also regulates the expression of fetal globin genes in the erythroid progenitor cells, by binding to MARs in the locus control region and the ε-globin promoter region in the β-globin cluster, and forming a complex with CREB-binding protein that possesses histone acetyltransferase activity (9). Inside the cells, SATB1 is localized at nuclei and surrounds heterochromatin, forming a cage-like network structure (10). Recently, it was shown that phosphorylation by protein kinase C alters preference for proteins to bind, i.e. histone deacetylase or histone acetyltransferase, and that acetylation by the latter impairs DNA-binding activity (11).

SATB1 is ~800 amino acids in length, possessing three DNA-binding motifs, i.e. two CUT domains and a homeodomain, in which a region of ~150 amino acids that includes the N-terminal CUT domain (CUT repeat 1, CUTr1) is reported to be the region relevant to the recognition of the MAR DNA (MAR-binding domain; MBD) (12). We have revealed by NMR that the CUTr1 region of ~80 amino acids is folded whilst the remainder of MBD is largely unfolded (13). The solution structure of CUTr1 contained five α-helices, in which the N-terminal four are arranged similarly to the four-helix structures

*To whom correspondence should be addressed: Tel: +81 29 8619473; Fax: +81 29 8612706; Email: k-yamasaki@aist.go.jp of the CUT domain of hepatocyte nuclear factor 6α and the POU-specific domains of POU-homologous proteins (14–16). Our NMR titration analysis and surface plasmon resonance (SPR) experiments using groove-specific binding drugs and methylated DNAs indicated that SATB1 possesses a DNA-binding mode similar to that of the POU-specific domain, in which third helix recognizes DNA bases from the major groove side (13), although the mechanism of specific recognition of the MAR sequence has yet to be revealed.

In this study, we determined the crystal structure of the complex of SATB1–CUTr1 and a MAR DNA, by a molecular replacement method using the NMR solution structure. The domain indeed contacted DNA from the major groove side, similarly to the POU-specific domains. The contacting manners by the two types of domains show striking similarity to each other at least in part, revealing that they share a framework of the DNA binding.

MATERIALS AND METHODS

Sample preparation

A pET15b (Novagen) expression vector of a SATB1-CUTr1 mutant fragment where four basic residues Arg-Lys-Arg-Lys were attached to the C-terminus of fragment Asn368–Leu452 was produced by PCR from a previously made expression vector of fragment Val353–Asn490 (13) with primers including the mutational sequence (5′-CAACAGGTT*CATATG*AACACAGAGGTGTCTTCCGAAATC-3′ and 5′-CA*GGATCC*TATTATTTACGTTTACGCAAGCTCCTTTCCCTTTCGTCCTGG-3′; mutational sequence in the reverse primer is underlined and recognition sequences of *Nde*I or *Bam*HI are shown in italic.). Mutations of Gln402 and Gly403 to alanine (Q402A and G403A mutations, respectively) were introduced to fragment Val353-Asn490 by two-step PCR: the first step includes two independent reactions using a forward subcloning primer (5′-CCACCCT*CATATG*GTCAGTAGATCTATGAATAAGCCTTTG-3′; the *Nde*I recognition sequence is shown in italic.) and a reverse mutational primer (5′-AAGCAAGCCCGCAGTTCTGTTAAAAGCCACACG-3′ for the Q402A mutation or 5′-TGAAAGCAAGGCCTGAGTTCTGTTAAAAGCCA-3′ for G403A mutation; mutational bases are underlined) or a forward mutational primer (5′-AACAGAACTGCGGGCTTGCTTTCAGAAATCCT-3′ for the Q402A mutation or 5′-AGAACTCAGGCCTTGCTTTCAGAAATCCTCCG-3′ for the G403A mutation; mutational bases are underlined) and a reverse subcloning primer (5′-CATTA*GGATCC*TATTAATTGTTCTCTGGTTTCCCATTCCTTTC-3′; the *Bam*HI recognition sequence is shown in italic), and the second step includes a reaction with the two subcloning primers shown above, with mixture of the products of the first reactions as the template. The mutant proteins were expressed in *Escherichia coli* cells and purified as described previously (13). After the protein was mixed with a double-stranded 12-mer DNA (5′-GCTAATATATGC-3′/5′-GCATATATTAGC-3′) at a 1:1 molar ratio, the complex was purified by a Sephadex G-75 (Amersham) column chromatography. The buffer used for the chromatography was 50 mM sodium phosphate (pH5.5). Before crystallization, sample was dissolved in 10 mM Tris–HCl buffer (pH 8.0) by dialysis, and concentrated by ultracentrifugation with Amicon Ultra device (Millipore).

Surface plasmon resonance

Experiments were carried out at 293 K using a Biacore X apparatus (BIAcore) essentially as described previously (13). The running buffer was 50 mM sodium phosphate (pH 5.5) containing 0.005% Tween-20 for the mutants based on the Val353–Asn490 fragment, while for the mutant with four basic residues attached to the C-terminus of the Asn368–Leu452 fragment, 50 mM NaCl was supplemented to the buffer in order to minimize non-specific electrostatic attraction. A total of 766 resonance unit (RU) of a double-stranded 16-mer DNA (5′-bio-CGTTTCTAATATATGC-3′/5′-GCATATATTAGAAACG-3′) (‘bio’ indicates biotinylation at the 5′ end) were immobilized on the surfaces of Sensor Chip SAs (BIAcore) in one of the two flow cells. Experiments were repeated four times in order to estimate uncertainty of the binding constant. Data were analyzed as described previously (13).

Crystal growth

The protein–DNA complex of an initial concentration of 15 mg ml^{-1} was subject to crystallization by a sitting-drop vapor diffusion method at 293 K on a 96-well Protein Crystallography Plate (Corning), with a reservoir solution of 50 mM Tris–HCl buffer (pH 8.0) containing 20% (w/v) polyethylene glycol 20000 (Wako, Japan), 10 mM $MgCl_2$ or $MgSO_4$ and 20% ethylene glycol. Rod-like crystals of 100–200 μm appear within a week.

Diffraction measurements and data processing

The crystal produced in the buffer containing $MgCl_2$ was flash-frozen in a nitrogen gas stream at 90 K and subjected to the measurement on a SMART6000 diffractometer (Bruker AXS) with Cu *K*α radiation from a M06X rotation-anode generator (MAC Science) by two-axis crystal rotation. Diffraction data were indexed and scaled by programs SMART and SAINT$^+$ (Bruker AXS), and merged and converted to structure-factor amplitudes by the SCALA (17) and TRUNCATE (18) programs from the CCP4 program suite (19).

The crystal made in the buffer containing $MgSO_4$ was flash-frozen in a nitrogen gas stream at 95 K and diffraction data were collected on beam line NW12A at PF-AR synchrotron (KEK, Tsukuba, Japan) by one-axis crystal rotation. HKL2000 (20), SCALA and TRUNCATE programs were used for the data processing. The diffraction data statistics are listed in Table 1.

Structure determination and characterization

The protein–DNA complex structures for the laboratory diffraction data were calculated by a molecular replacement method, using the NMR solution structure of SATB1-MBD (PDB entry 1YSE). Initially, the structure

Table 1. Statistics of crystallographic analysis

Crystal diffraction data		
Salt in crystal buffer	10 mM $MgCl_2$	10 mM $MgSO_4$
Wavelength (Å) (X-ray source)	1.5418 (laboratory)	1.0000 (synchrotron)
Space group	*P*1	*P*1
a, b, c (Å)	29.66, 37.00, 41.25	32.66, 38.32, 40.71
α, β, γ (degrees)	71.60, 83.92, 71.07	71.34, 82.50, 66.63
Resolution range (Å)[a]	19.64–2.00 (2.10–2.00)	22.29–1.75 (1.84–1.75)
Total observed reflections[a]	35 508 (1673)	42 451 (3223)
Unique reflections[a]	9632 (867)	14 958 (1404)
Redundancy[a]	3.7 (1.9)	2.8 (2.3)
Completeness[a]	89.8 (55.2)	86.3 (55.2)
$<I/\sigma(I)>$[a]	21.6 (3.9)	9.2 (2.0)
R_{merge} (%)[a]	5.3 (15.5)	9.4 (27.5)
Refinement statistics		
Resolution range (Å)	19.64–2.00	22.29–1.75
Reflections used (>2σ)	9613	14 958
R_{work} (%)	21.0	24.5
R_{free} (%)[b]	25.5	28.3
Protein atoms	697	697
Nucleic acid atoms	486	486
Water atoms	75	108
RMSD from the ideal geometry		
Bond lengths (Å)	0.0049	0.0048
Bond angles (degrees)	1.076	1.041
Ramachandran statistics for protein[c]		
Most favored (%)	93.6	93.6
Additionally allowed (%)	6.4	6.4

[a]Parameters were obtained by the SCALA program (17) in the CCP4 package (19). Values in parentheses are for the highest resolution shell.
[b]R_{free} is calculated for 10% of the reflections, selected randomly and excluded for the model refinements.
[c]Calculated with PROCHECK (23).

of the protein moiety in the complex was obtained from the starting template of the NMR structure, by a cross-rotation method implemented in the CNS package (21). From this protein structure and a standard B-DNA structure produced by Insight II program (Accelrys), the complex model was obtained by the cross-rotation and translational search methods, where the coordinate of the protein model was fixed. Iterative cycles of model building to fit to the electron density map and dynamical refinement were carried out by XtalView (22) and CNS, respectively. The water molecules were picked up at 1.5 σ level and within distance of 4.0 Å from the protein or DNA atoms by a program implemented in CNS, although some waters shifted slightly more distant during final refinements.

The complex structure for the synchrotron data was calculated by the molecular replacement method using the complex model from the laboratory data, and refined in the similar manner.

Root mean square deviation (RMSD) values from the idealized geometry with regard to bond lengths and angles were obtained by CNS. Ramachandran profile and secondary structure elements of the protein moiety are analyzed by Procheck (23). The statistics of the structures are listed in Table 1.

Intermolecular contacts are analyzed by in-house FORTRAN programs, XtalView (22) and Insight II (Accelrys). Hydrogen bonds are defined by distance of hydrogen donor and acceptor <3.4 Å and predictable donor-hydrogen-acceptor angle >110°. Electrostatic contacts are defined by N^+-O^- distance <5.0 Å, unless the groups form a hydrogen bond(s). Among the van der Waals interactions, contacts between apolar atoms ('apolar contacts' in this study), are defined by C-C distance <4.5 Å, unless an adjacent N or O atom is closer to the counterpart C atom. Atom pairs not involved in the above three contacts, with distance <3.9 Å are classified into other van der Waals contacts (or simply 'van der Waals contacts' in this study), unless an adjacent atom is closer to and forms a hydrogen bond to the counterpart atom. Effects of amino acid or base substitutions are evaluated by the Biopolymer module in Insight II. RMSD values between the molecules are calculated by using MOLMOL (24).

RESULTS AND DISCUSSION

Structure determination

For crystallization of protein–DNA complex, we initially intended to use a CUTr1 fragment, Asn368–Ala455, which is essentially the folded region of MBD as revealed by the NMR analysis (13), since including unfolded regions may interfere crystallization. However, this fragment is rather acidic, with a pI value of 5.8 and a net charge of −1 at neutral pH, which is unfavorable for interaction with the negatively charged DNA, causing electrostatic repulsion. Consequently we did not observe significant DNA binding for this fragment (13). Therefore, we prepared a mutant containing four basic amino acids at the C-terminus of fragment Asn368–Leu452 (Figure 1a), which possesses a pI value of 10.1 and a net charge of +3. This fragment showed DNA binding with a constant of $9.5(\pm1.2) \times 10^6\,\mathrm{M}^{-1}$, which is slightly better than that of the basic MBD fragment with a net charge of +2 [$4.8(\pm0.6) \times 10^6\,\mathrm{M}^{-1}$; (13)], even though the experiment for the former was carried out in a higher salt concentration (Figure 2). We used this mutant CUTr1 fragment for crystallization of the complex with DNA, which will be treated as SATB1-CUTr1, hereafter, unless otherwise stated.

For DNA, we used a double-strand dodecamer including a sequence from the IgH enhancer region, CTAATATAT (Figure 1b), which was included in DNAs for the gel retardation and missing nucleoside experiments (6) as well as for our SPR experiment (13). In the sequence, the ATATAT sequence is termed as the BUR nucleation sequence that is critical in the base unpairing and at least partially in the binding of SATB1 (4,6).

The 1:1 mixture of SATB1-CUTr1 and dodecamer DNA was purified by a gel filtration column and subjected to crystallization. Crystals produced with different salts were subjected to the measurements with laboratory and synchrotron X-ray sources under freezing condition by nitrogen gas stream, resulting in diffraction data of resolution up to 2.00 and 1.75 Å, respectively. The structure was determined by a molecular replacement method using the NMR structure of SATB1-MBD (13)

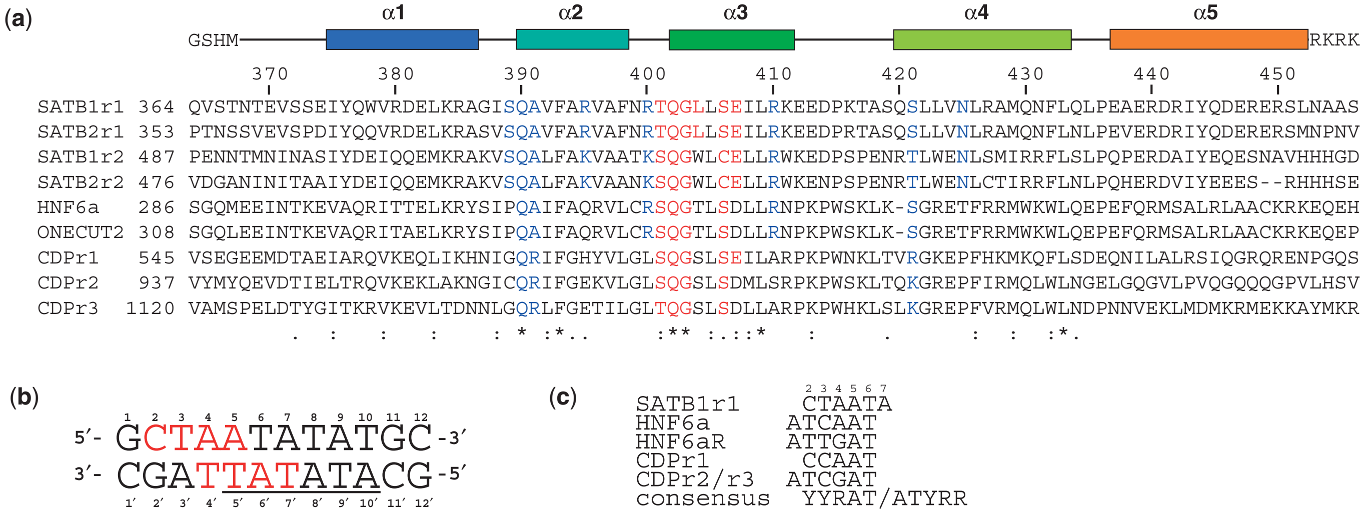

(a)
α1
α2
α3
α4
α5
GSHM
RKRK
370
380
390
400
410
420
430
440
450
SATB1r1 364 QVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAAS
SATB2r1 353 PTNSSVEVSPDIYQQVRDELKRASVSQAVFARVAFNRTQGLLSEILRKEEDPRTASQSLLVNLRAMQNFLNLPEVERDRIYQDERERSMNPNV
SATB1r2 487 PENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGD
SATB2r2 476 VDGANINITAAIYDEIQQEMKRAKVSQALFAKVAANKSQGWLCELLRWKENPSPENRTLWENLCTIRRFLNLPQHERDVIYEEES--RHHHSE
HNF6a 286 SGQMEEINTKEVAQRITTELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEH
ONECUT2 308 SGQLEEINTKEVAQRITAELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEP
CDPr1 545 VSEGEEMDTAEIARQVKEQLIKHNIGQRIFGHYVLGLSQGSLSEILARPKPWNKLTVRGKEPFHKMKQFLSDEQNILALRSIQGRQRENPGQS
CDPr2 937 VYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSLSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
CDPr3 1120 VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSLSDLLARPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
. : : : : * :*.. :** :.::* : . : : :*.
(b)
1 2 3 4 5 6 7 8 9 10 11 12
5′- GCTAATATATGC -3′
3′- CGATTATATACG -5′
1′ 2′ 3′ 4′ 5′ 6′ 7′ 8′ 9′ 10′ 11′ 12′
(c)
2 3 4 5 6 7
SATB1r1 CTAATA
HNF6a ATCAAT
HNF6aR ATTGAT
CDPr1 CCAAT
CDPr2/r3 ATCGAT
consensus YYRAT/ATYRR

Figure 1. (**a**) Amino acid sequence alignment of CUT domains of human proteins produced by Clustal X (44). Sequences of two SATB proteins (SATB1 and SATB2), each containing two repeated CUT domains (r1 and r2 from the N-terminus), two ONECUT group proteins [hepatocyte nuclear factor 6α (HNF6a) and ONECUT2] and CCAAT displacement protein (CDP), containing three repeated CUT domains (r1, r2 and r3), were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov). Entry codes are NP_002962 for SATB1, Q9UPW6 for SATB2, AAD00826 for HNF6α, CAB38253 for hsONECUT2 and AAB26579 for CDP. Residue numbers above are for SATB1-CUTr1, while those on the left are of the first residues of the individual sequences. Base-contacting residues of SATB1-CUTr1 are colored red, while those contacting sugar-phosphate backbone only are colored blue (Figure 5a). For the other CUT domains, residues that are conserved or expected to possess conserved interactions are colored similarly (classification is marginal in some cases; e.g. for residues in position 404, Thr and Ser are expected to show only a part of interactions similar to Leu residues; see text). Colored boxes above the sequence alignment indicate regions of the helices of SATB1-CUTr1, accompanied by a horizontal bar indicating a region used in crystallography, where residues shown in the both ends are from the expression vector (N-terminal Gly-Ser-His-Met) or introduced by mutagenesis (C-terminal Arg-Lys-Arg-Lys). Below the sequence, identical and similar residues are marked as produced by the program. (**b**) Double-strand dodecamer DNA used in crystallography. Bases contacted by the protein (Figure 5a) are shown in red. The horizontal bar shows the BUR nucleation sequence (4). (**c**) DNA sequences recognized by CUT domains shown from the 5′-end. Base numbers defined in (b) are shown above the SATB1-CUTr1 recognition sequence. For HNF6α, the two reversed sequences (HNF6a and HNF6aR) are also shown. In the consensus core sequence, Y and R represent pyrimidine and purine, respectively. Sequences were aligned so that the key recognition base pair, T6-A6′, is conserved and that commonly recognized regions are maximized.

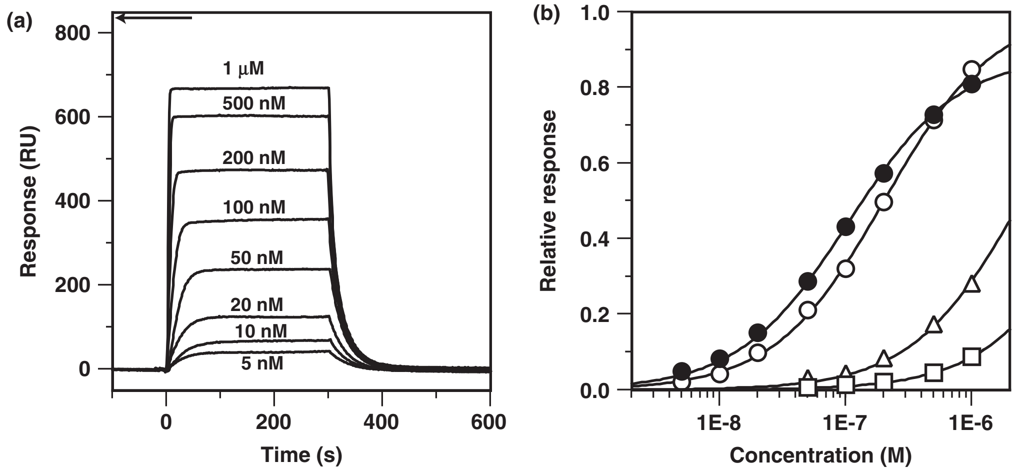


Figure 2. DNA binding of SATB1 fragments observed by surface plasmon resonance (SPR). (**a**) SPR difference sensorgrams (the responses in the control flow cell without immobilized DNA were subtracted) for binding of the SATB1-CUTr1 fragment used in the crystallography to a 16-mer double-stranded DNA, where the protein concentrations are 5 nM^{-1} μM, as indicated beside the sensorgram lines. The protein solutions were injected during a period of 0–300 s. The arrow indicates the maximum response value expected when the protein molecules bind to all the immobilized DNA molecules at 1:1 stoichiometry. (**b**) Equilibrium response values relative to the expected maximum values at a 1:1 stoichiometry, as functions of the protein concentration. Presented are the binding profiles of SATB1-CUTr1 (closed circle), wild-type MBD (Val353–Asn490) [open circle; (13)], MBD with Q402A mutation (open square) and MBD with G403A mutation (open triangle). Fitting curves to the simple 1:1 binding model are shown. Binding constants are $9.5(\pm1.2)\times10^{6}\,\mathrm{M}^{-1}$, $4.8(\pm0.6)\times10^{6}\,\mathrm{M}^{-1}$ (13), $9.7(\pm0.9)\times10^{4}\,\mathrm{M}^{-1}$ and $4.5(\pm0.6)\times10^{5}\,\mathrm{M}^{-1}$ for SATB1-CUTr1, wild-type MBD, Q402A MBD and G403A MBD, respectively, where error levels are estimated from four repeated experiments. Note that 50 mM NaCl was added to the running buffer for the SATB1-CUTr1 fragment, which contains four additional basic residues.

and a standard B-DNA structure. The statistics for the diffraction data, refinement results and structural properties are shown in Table 1.

Overall structure in the crystal

A triclinic crystal lattice contained a single structure of the complex of SATB1-CUTr1 and a dodecamer DNA possessing overall dimensions of ~40 Å × 40 Å × 40 Å. Crystal packing was promoted by DNA–DNA, protein–protein and protein–DNA interactions between the adjacent lattices (Supplementary Data and Figures S1–S3), which involves an artificially introduced Arg residue. Since the two structures determined from the two datasets are essentially the same (Supplementary Figure S4), we will describe on the structure from the laboratory data, which possesses better *R* values (Table 1), unless otherwise stated.

The protein moiety consists of five α-helices (α1: Ile375–Ala386, α2: Gln390–Phe398, α3: Gln402–Lys411, α4: Gln420–Leu433 and α5: Glu437–Leu452; a short helical region of Pro415–Thr417 appears only in the structure from the laboratory data, which is likely to be related to the crystal packing; Supplementary Data), as identified by the program Procheck (23) (Figures 1a and 3a). The structure of DNA is essentially in the standard B-form without significant deformation. The α3-helix, especially in its N-terminal half, deeply enters the major groove of DNA acting as the recognition helix, in a manner where the helical axes of the α-helix and DNA are perpendicular to each other. The backbone structure of the protein in the crystal is similar to that of the solution structure determined by NMR (the minimized average structure of the ensemble) (Figure 3b). The RMSD value for the Cα atoms in the Val371-Gln445 region is 0.99 Å after superimposition. However, the two structures significantly differ in Arg400–Gln402 in the N-terminal cap region of α-helix 3, with RMSD of 2.59 Å (Figure 3b). In this region, α-helix 3 starts from Gln402 in the crystal structure, even though Gly at position 403 generally acts as a helix breaker, although it starts from Leu404 in the solution structure. This is likely to be caused by the interaction with DNA, since the side-chains of the residues in this region directly contact DNA bases and phosphates, as described below. The difference in the C-terminal region is more noticeable, with RMSD of 6.12 Å for Asp446–Leu452, where the helix of the solution structure is truncated at Gln445 while that in the crystal structure elongates to Leu452. This is not likely to be the result of the protein–DNA interaction, since the previous NMR titration experiment did not show significant chemical-shift changes in this region (13). Alternatively, we suggest that the solution structure of the domain also contains the long helix at least in some conformations, although NMR structural constraints based on nuclear Overhauser enhancements and amide hydrogen exchange protection were not obtained because of the significant flexibility in this region. The Asp446–Leu452 region shows relatively large temperature factors, with average of 39.2 ± 8.9 $Å^2$ for backbone atoms, where the whole determined region (Glu370–Arg453) shows 22.4 ± 9.2 $Å^2$. In addition, the two crystal structures showed relatively large RMSD value of 1.29 Å in this region, which also indicates the high flexibility (Supplementary Figure S4).

The mode of the protein–DNA interaction in this crystal structure is consistent with our previous observation by the NMR titration analysis and SPR experiments indicating that α-helix 3 enters the major groove of DNA (13). A computational model of the complex based on these results shown in the same report is indeed similar to the present crystal structure.

Protein–DNA interface

Residues in α-helix 3 and its N-terminal cap (Thr401–Glu407) contact eight bases in six consecutive base pairs, through direct or water-mediated hydrogen bonds, and contacts between apolar atoms (apolar contacts) and other van der Waals interactions (Figures 4 and 5a) (see Methods section, for definition). All the contacts including those to the sugar-phosphate backbone cover the range of eight base pairs (Figure 5a). Details of the contacts are as follows.

The O^{γ} atom of Thr401 and the backbone amide of Gly403 form hydrogen bonds to a same water molecule that possesses hydrogen bonds to the N^6 and N^7 atoms of A4 (Figure 4a). At the same time the O^{γ} atom of Thr401 forms van der Waals contacts to C^{5M} and C^6 of T3 (data not shown). Also, the C^{α} and backbone O atoms of Gly403 form contacts to the C^{5M} and/or O^4 atoms of T5′ (Figure 4a). In addition, Gly403 is important in that it does not contain side-chain, since presumable C^{β} and H^{β} atoms would have steric hindrance with N^6H_2 of A4 and O^4 of T5′ (data not shown). Consistently, a mutation of Gly403 to Ala resulted in a reduction of DNA-binding activity by ~10-fold (Figure 2b).

The O^{ε} and N^{ε} atoms of Gln402 form hydrogen bonds to N^6 and N^7, respectively, of A6′ in a manner typical of the DNA base recognition by proteins (25,26) (Figure 4b). At the same time, the C^{γ}, C^{δ}, O^{ε}, and N^{ε} atoms form contacts to the C^5, C^{5M} and/or C^6 atom of T7′. In addition, O^{ε} of Gln402 and N^6 of A5 form a water-mediated hydrogen bond (Figure 5a). These interactions are very important, as confirmed by a mutation of Gln402 to Ala, which results in an impairment of DNA-binding activity by ~50-fold (Figure 2b).

The $C^{\delta 1}$ atom of Leu404 forms an apolar contact to the C^6 atom of the C2 base, with distance of 4.3 Å (data not shown). Although this contact is not likely to strongly exclude other bases, a pyrimidine base may be preferable, since C^5 may expand the area of apolar contacts, whilst the presumable position of C^8 of a purine base, when simply replaced, would be slightly more distant to Leu404 $C^{\delta 1}$, with a distance of 4.4 Å (data not shown). In addition, the C^{β}, $C^{\delta 1}$ and backbone N atoms of Leu404 form contacts to C^{5M} of T3.

The O^{γ} atom of Ser406 form van der Waals contacts to C^{5M} of T5′ and C^8 of A6. The C^{γ}, C^{δ} and O^{ε} atoms of Glu407 form contacts to the C^{5M} methyl groups of T4′ and/or T5′ (Figures 4c and 5a). It is noteworthy that two basic residues in α-helix 3, Arg410 and Lys411, form hydrogen bonding and/or electrostatic contacts with

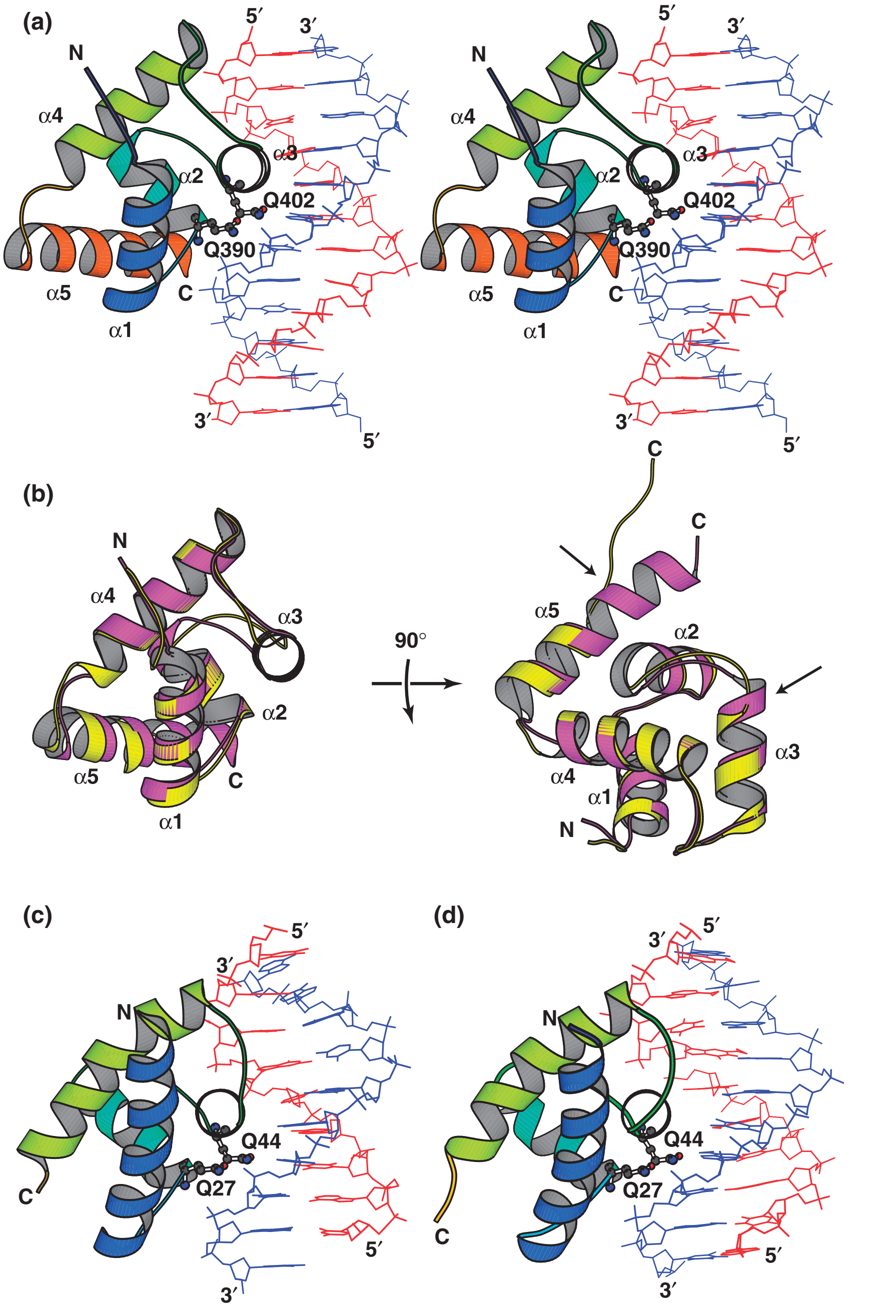


Figure 3. Structures of SATB1-CUTr1 and POU-specific domains. Shown are crystal structure of the SATB1-CUTr1/DNA complex in stereo view (**a**), comparison of the crystal (magenta) and solution (yellow) structures of SATB1-CUTr1 (**b**), and structures of POU-specific domain/DNA complex for OCT-1 (**c**) and Pit-1 (**d**). Helices of SATB1-CUTr1 in (a) and equivalent helices of POU-specific domain in (c) and (d) are colored similarly as in Figure 1a. The two Gln residues that are most contributing to defining the DNA-binding framework are shown in ball-and-stick in (a), (c) and (d). Residues numbers in (c) and (d) are those of the common POU domain, which are used in the PDB coordinates files. Arrows in the right panel of (b) show regions with large conformational differences. In (b), the backbone C and N atoms in the Val371–Gln445 region were superimposed by MOLMOL (24). The figures were produced by Molscript (45).

Glu407. Especially, Arg410 form two hydrogen bonds to Glu407 (Figure 4c). These interactions are likely to constrain the conformation of Glu407 side-chain so that the C^{γ} and C^{δ} atom are closer to the C^{5M} atoms of T4′ and T5′ than the O^{ε} atoms, which is essential for the apolar contacts. Importance of the Glu407–Arg410 interaction is evident from our previous observation that mutation of Arg410 to Asn results in a decrease of DNA-binding affinity by ~40-fold (13).

DNA sugar-phosphate backbones are contacted through direct or water-mediated hydrogen bonds, electrostatic attractions and apolar and van der Waals contacts (Figure 5a). Contacting residues are Ser389, Gln390 (Figure 4b), Ala391, Arg395, Arg400, Thr401,

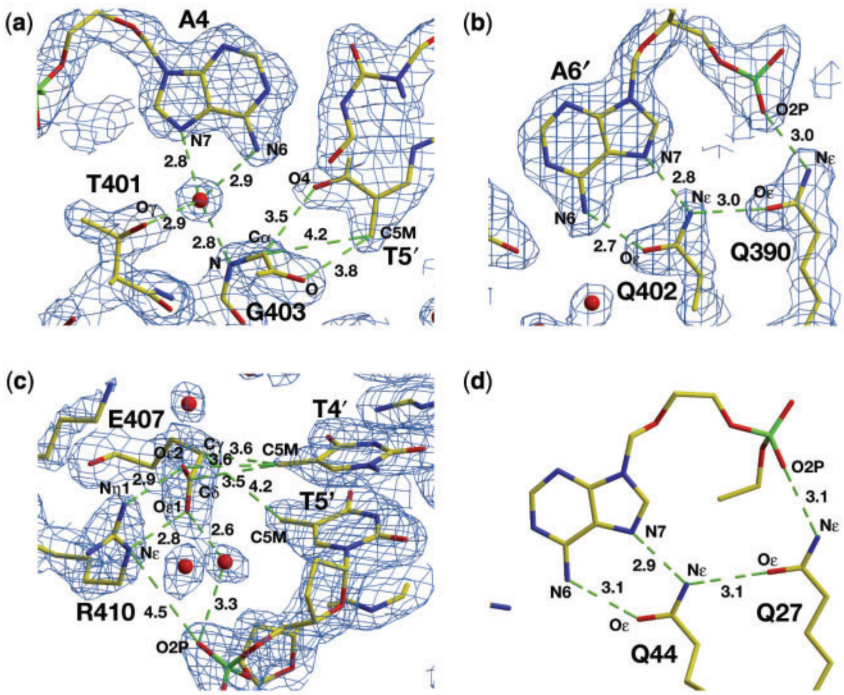


Figure 4. The interface of the base recognition in crystal structures. Interfaces around Thr401 and Gly403 (**a**), Gln402 (**b**) and Gln407 (**c**) of SATB1-CUTr1 and Gln44 of Pit-1 (**d**), which is equivalent to Gln402 of SATB1, are shown. Protein and DNA models are shown in sticks colored according to the atom type (carbon: yellow, nitrogen: blue, oxygen: red and phosphorus: green), while waters are shown in red spheres. For (a)–(c), the $2F_o$–F_c electron densities contoured at 1.5 σ are shown in blue. Atom pairs possessing hydrogen bonds, electrostatic attractions, apolar contacts and other van der Waals contacts are indicated by green broken lines with distances in angstrom. The contacts to the sugar-phosphate backbone within the panels are also highlighted. Figures were produced by XtalView (22) and Raster 3D (46).

Gln402, Leu404, Ser406, Glu407 (Figure 4c), Arg410 (Figure 4c), Ser421 and Asn425. In addition, O^{γ} of Ser421 is in a position allowing a hydrogen bond to the backbone phosphate of G1, if the DNA is not artificially truncated. Among these residues, Gln390 is of special note in that, in addition to two hydrogen bonds to two phosphate groups (Figure 5a), the O^{ε} atom forms a hydrogen bond to the N^{ε} atom of Gln402 that is the only residue forming direct hydrogen bonds to the base (Figure 4b). This hydrogen-bonding network is likely to contribute primarily to fixing the relative position of protein and DNA, namely, defining the DNA-binding framework.

The contacts to the DNA backbone should be important in the protein–DNA binding, enhancing the affinity, although not directly contributing to the sequence-specific recognition. It should be noted that Thr401, Gln402, Leu404, Ser406 and Glu407 contact to both the DNA bases and backbones, simultaneously. It was shown for Ser406 that a mutation to Ala reduces the DNA-binding activity by more than 10-fold (13).

It should be emphasized that mutations of the base-contacting residues conserved among CUT domains (Figure 1a), i.e. Gln402 and Gly403, on the MBD fragment (Val353–Asn490), as well as those of Ser406 and Arg410, significantly reduced the DNA-binding activity [Figure 2b; (13)]. Therefore, the binding mode described above is unlikely to be induced by the basic residues introduced at the C-terminus of the present SATB1-CUTr1 fragment.

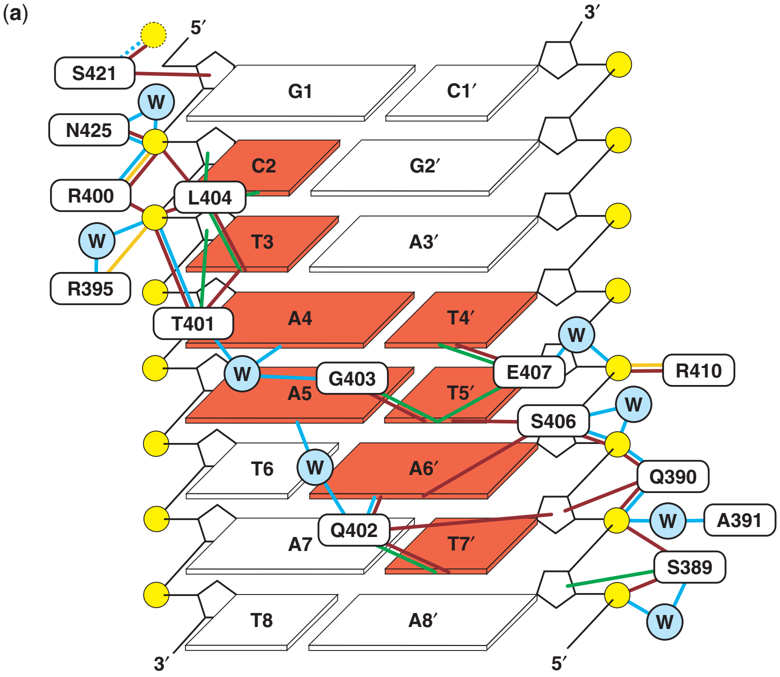
SATB1 recognition sequence in the MAR DNA

The base contacts described above are likely to define the sequence specificity of SATB1-CUTr1 to the MAR DNA, resulting in a recognition sequence of CTAATA/TATTAG, or (Y)TAATA/TATTA(R) considering the relatively weak contact and the presumable preference to pyrimidine (Y) rather than purine (R) at the C2 position. The recognition sequence only partially includes the BUR nucleation sequence (4), ATATAT (Figure 1b), to which SATB1 is initially expected to bind (6). The mutation that impaired the binding of SATB1 changed the CTAATA sequence to CTACTG (6), which should interfere the contacts by Gln402, Gly403, Ser406 and Glu407, at least partly, according to the present contacting profile (Figure 5a). Consistently, a missing nucleotide experiment in the same report showed that SATB1 contacts the DNA in the region centered at a sequence CTAATA, but not at ATATAT. We have previously introduced methylation at the N^6 atom of adenine base at position of A3′ or A6′ and observed that DNA binding was significantly interfered only when A6′ was methylated (13), which is also consistent with the present observation that the O^{ε} atom of Gln402 forms a hydrogen bond with the N^6 atom of A6′, while the A3′ base is not contacted by the protein (Figures 4b and 5a)

DNA recognition by other CUT domains

This study is the first report regarding the geometry of sequence-specific DNA recognition by a CUT domain.

(a)



(b)

SATB1r1	401	TQGLLSE
OCT1	43	SQTTISR
PIT1	43	SQTTICR
HNF1a	140	NQSHLSQ

(c)

	2 3 4 5 6 7 7′ 6′ 5′ 4′ 3′ 2′
SATB1r1	CTAATA/TATTAG
OCT1	TGCATA/TATGCA
PIT1	TGTATA/TATACA
HNF1a(1)	TTAATA/TATTAA
HNF1a(2)	TGAATT/AATTCA

(d)

	C/T(2)	T/G(3)	A/C/T(4)	A (5)	T (6)	A/T(7)
	G/A(2′)	A/C(3′)	T/G/A(4′)	T (5′)	A (6′)	T/A(7′)
T/S/N (401)		H[h^{a}],V[s,o,h]	H[s^{b}],V[p]			
Q (402)				H[s^{b}]		
					H[s,o,p,h], V[s,p,h]	A[s,p,h^{a}], V[s,o,p,h^{a}]
G/T/S (403)		V[h^{a}]	H[s^{b},o,p^{a},h], V[o,p,h^{a}]	H[h^{a}],V[o,p^{a}]		
			H[h^{a}]	H[o,p^{a},h^{a}], A[s,o,p,h],V[s,o,p,h]		
L/T/H (404)	A[s,h]	A[s,h^{a}],V[s,h^{a}]				
S/C (406)						
				A[p],V[s,h]	A[p],V[s]	
E/R/Q (407)	V[p]	H[o,p,h^{a}],A[h^{a}], V[o,p^{a},h^{a}]				
		V[o]	H[o^{c},h^{a}],A[s], V[s,h]	A[s,h]		

Figure 5. (**a**) Summary of contacts of SATB1-CUTr1 and DNA duplex (bases in squares are numbered as in Figure 1b), drawn looking from the major groove side of DNA. Circles in cyan with ‘W’ and those in yellow represent water molecules and DNA phosphates (a phosphorus and four oxygens), respectively. Bases contacted by protein were highlighted in red. Lines in cyan, yellow, green and brown represent hydrogen bonds, electrostatic attraction, apolar contacts and other van der Waals contacts respectively, as defined in the Methods section. Broken yellow circle and cyan line represent presumable phosphate and hydrogen bond, respectively, if the 5′-terminal G1 nucleotide would possess a phosphate group in an expected position for a standard B-DNA, where a van der Waals contact between Ser421 to the O$^{5'}$ atom of G1 is also shown. Hydrogen bonds mediated by two waters or more were excluded. (**b**) Amino acid sequences of the regions of SATB1-CUTr1 and POU-specific domains involved in base recognition, which are aligned so that Gln402 of SATB1 is conserved. Residue numbers of the first residues in the alignment are indicated, where numbers for OCT-1 and Pit-1 are those of the common POU domain, used in the PDB coordinates files. Base-contacting residues are highlighted in red. (**c**) DNA sequences recognized by SATB1-CUTr1 and POU-specific domains in the crystal structures, aligned so that the A bases hydrogen bonding with the conserved Gln residues are in the same position. Sequences of the both strands are shown from the 5′-end. Base numbers as in (a) are shown above. Bases contacted by the proteins are colored red. For HNF1α, the two different sequences of the DNA regions contacted by the dimeric protein in the crystal are shown. (**d**) Comparison of the base-contacting profiles of SATB1-CUTr1 and POU-specific domains. Contacting residues and bases are ordered vertically and horizontally, respectively, where the residue and base numbers for SATB1-CUTr1/DNA complex are shown in parentheses. For each residue, upper and lower rows are assigned for positions for the bases of the two strands. Intermolecular hydrogen bonds, apolar contacts and other van der Waals contacts are indicated by ‘H’, ‘A’ and ‘V’, respectively, in the definitions described in the Methods section, except that apolar C-S contacts within 4.5 Å are included. Accompanied ‘s’, ‘o’, ‘p’ and ‘h’ in parentheses indicate that the contacts are observed for SATB1, OCT1, Pit-1 and HNF1α, respectively. Red color shows that contacts are observed for SATB1 as well as at least one of the POU-specific domains, while blue and green colors show that contacts are observed solely for SATB1, and solely for at least one of the POU-specific domains, respectively. Annotations stand for, *a*: contacts are observed only in one of the two equivalent contacting sites in the crystal of Pit-1 or HNF1α; *b*: water-mediated hydrogen bonds; *c*: marginal hydrogen bonds with donor–accepter distances of 3.44 Å described in the original literature (29).

There are three subgroups of CUT-domain proteins, i.e. (i) SATB proteins including SATB1 and SATB2, possessing two CUT-domain repeats, (ii) ONECUT group including HNF6α and ONECUT proteins, possessing a single CUT domain and (iii) human CCAAT displacement protein (CDP) and its murine autholog, Cut homeobox (Cux), possessing three repeats (Figure 1a). Sequence specificity and/or DNA-binding activities of these CUT domains are significantly different among one another. Namely, HNF6α recognizes the ATCAAT sequence (27), and the N-terminal CUT domain, repeat 1, of CDP/Cux recognizes the CCAAT sequence, while the other two CUT domains, repeat 2 and repeat 3, recognize ATCGAT (28), which are different from the present SATB1-binding sequence (Figure 1c). In addition, Cut repeat 2 of SATB1 protein is dispensable for binding to the MAR DNA (12). However, many DNA-contacting residues observed in this study are conserved among these CUT domains (Figure 1a), indicating that the DNA-binding mode is essentially shared. Especially, the important base-contacting residues, Gln402 and Gly403, are perfectly conserved, which form hydrogen bonds to A6′ base, and apolar and van der Waals contacts to T5′, respectively. Also, Thr or Ser is conserved at position 401, which is, together with Gly at position 403, capable of forming the water-mediated hydrogen bonds to the N^6 and N^7 atoms of the A4 base. With regard to these hydrogen bonds, the A4 base can be replaced by a G base, since O^6 and N^7 of the G base would act as two hydrogen-bond acceptors. Therefore, a sequence of the three base pairs R4A5T6/A6′T5′Y4′ is likely to be recognized commonly by the CUT domains. Considering this, the recognition sequences were aligned as in Figure 1c, resulting in the consensus core recognition sequence of YYRAT/ATYRR.

The recognition sequence of HNF6α, ATCAAT/ATTGAT, fits the YYRAT consensus in the both directions (Figure 1c). It should be noted that HNF6α possesses Thr and Asp in the positions equivalent to Leu404 and Glu407, respectively, of SATB1-CUTr1 (Figure 1a). When Leu404 is simply replaced by Thr in the present SATB1-CUTr1/DNA structure, it is still possible to form apolar contacts to the C^{5M} atom in the T3 base, but not to atoms of the C2 base. In contrast, when Glu407 is replaced by Asp, apolar contacts to the C^{5M} atoms of the T4′ and T5′ bases are likely to be disrupted. These observation suggest that HNF6α recognizes the T3/A3′ base pair, but not the A4/T4′ base pair, and, therefore, that the ATTGAT sequence ('HNF6aR' in Figure 1c), but not the ATCAAT, is the direction equivalent to the YYRAT consensus core sequence.

The three CUT domains of CDP possess Ser at the position equivalent to Leu404 of SATB1-CUTr1 (Figure 1a). When Leu404 is simply replaced by Ser in the SATB1-CUTr1/DNA structure, an apolar contact between the Ser C^{β} atom and the T3 C^{5M} atom would still be possible depending on the conformation (data not shown), although it should be much weaker than those from the C^{β} and $C^{\delta 1}$ atoms of the Leu residue and may allow a C base at the T3 position (Figure 1c). It should be noted that CUT repeat 1 possesses Glu at the position equivalent to Glu407, although it is Asp for CUT repeats 2 or 3. For CUT repeat 1, the interactions between Glu and the T4′ and T5′ bases should be conserved, although for CUT repeats 2 and 3, those between Asp and the T bases are unlikely. Therefore, repeat 1, but not repeat 2 or 3, is likely to require an A base at position 4.

SATB1-CUTr2 is not likely to contribute to the MAR-DNA binding (12), which may be explained as follows. Among the several DNA-contacting residues that differ between CUTr1 and CUTr2, those likely to largely influence the binding are Trp at position 404 and Cys at position 406 (Figure 1a). The former possesses a bulky side-chain compared to Leu and, when simply replaced, shows a steric hindrance with DNA sugar-phosphate backbone or other amino acids, such as Arg400 and Asn425, probably disrupting contacts including hydrogen bonds to the phosphate. Also, the side-chain of Cys at position 406 is unable to form a hydrogen bond to a DNA phosphate.

Comparison with POU-specific domains

We previously predicted that the framework of DNA binding by the five-helix SATB1-CUTr1 structure is similar to that by the basically four-helix POU-specific domains of POU-homologous proteins, in that the third helix acts as the recognition helix and deeply enters the major groove of DNA, judging from the structural similarity and the NMR titration and SPR experiments (13), which was confirmed by the present crystal structure. In the crystal structures of the complex of DNA and POU-homologous proteins, Oct-1, Pit-1 and HNF1α (PDB entries 1OCT, 1AU7 and 1IC8, respectively), the third helix of the common four helices of the POU-specific domains acts as the recognition helix, achieving the sequence-specific DNA recognition (29–31) (Figure 3c and d). We have re-analyzed contacting profiles in these crystal structures, by the same criteria adopted in this study.

Although the amino acid sequences in the recognition helices of POU-specific domains are not very similar to those of CUT domains (Figure 5b), a Gln residue that is strictly conserved among the POU-specific domains forms a pair of hydrogen bonds to adenine N^6 and N^7 atoms, essentially in the same manner as Gln402 of SATB1 (Figure 4b and d). In addition, another Gln residue is also conserved, which forms hydrogen bonds to the side-chain of the above Gln residue and at the same time to phosphate groups from both the side-chain and backbone N atoms, in the similar manner as Gln390 of SATB1 (Figures 3a, c and d, and 4b and d). Since this conserved hydrogen-bonding network is most important in defining the DNA-binding framework, as described above, SATB1-CUTr1 and POU-specific domains are likely to share the DNA-binding framework.

Therefore, in order to compare the DNA-contacting profiles, amino acid sequences of the recognition helices are aligned so that the Gln residue equivalent to Gln402 of SATB1 (hereafter we use term 'residue 402', etc.) is conserved among the proteins (Figure 5b). Also, sequences of the DNA regions contacted in the crystal structures are aligned in reference to the T6-A6′ base pair that is contacted by Gln402 (Figure 5c). Based on the alignments, the contacting profiles i.e. locations and modes, of SATB1-CUTr1 and POU-specific domains are summarized in Figure 5d. In addition to the conserved hydrogen bonds to the A base at position equivalent to A6′ of SATB1-CUTr1 recognition sequence (hereafter we use term 'base 6′', etc.), there are several points where the contacting profiles of the SATB1-CUTr1 and POU-specific domains are similar to each other, as indicated in red color in Figure 5d. For example, hydrogen bonds between residue 403 and base 4 are observed for SATB1 and all the POU-specific domains. It should be noted, however, that the amino acids of residue 403 (Gly, Thr or Ser) and bases of base 4 (A or C or T) are different and therefore, manners of hydrogen bonding are different. Namely, for OCT1, $O^{\gamma1}$ of Thr accepts a hydrogen bond from N^4 of C base, while for Pit-1, $O^{\gamma1}$ of Thr donates a hydrogen bond to O^4 of T base, and for HNF1α, O^{γ} of Ser accepts a hydrogen bond from N6 of A base (data not shown), which are also different from SATB1-CUTr1, where backbone N of Gly donates a hydrogen bond to a water that behaves as both hydrogen donor and acceptor to the A4 base (Figure 4a).

There are four contacts that are observed only for SATB1-CUTr1, while 16 are observed only for either of the three POU-specific domains (Figure 5d). Among the latter 16, only four are observed for all the POU-specific domains, i.e. hydrogen bonds between residue 403 and base 5′, hydrogen bonds between residue 407 and base 3, van der Waals contacts between residue 403 and base 4, and van der Waals contacts between residue 407 and base 3. Even for these four, manner of contacts are diverse, similarly to as described above. Thus, the difference between the base-contacting profiles of SATB1-CUTr1 and POU-specific domains are at a similar level to that among the POU-specific domains. In fact, common and different contacts between SATB1-CUTr1 and HNF1α are both 14 in Figure 5d, which is better than the case between OCT-1 and HNF1α, showing 11 common and 15 different contacts, and the case between Pit-1 and HNF1α with 11 common and 18 different contacts.

Contacts to DNA backbone are also similar among these domains at least partially. Namely, contacts equivalent to the hydrogen bonds from Gln390 (Gln27 of OCT-1 and of Pit-1, and Gln130 of HNF1α; mentioned above; Figure 4d), Thr401 (Ser43 of OCT-1 and Pit-1), Ser406 (Ser48 of OCT-1 and Ser145 of HNF1α), Ser421 (Ser56 of OCT-1 and Pit-1) and Asn425 (Asn59 of OCT-1 and Pit-1) of SATB1-CUTr1 to phosphate groups (Figure 5a) are observed also for either of the POU-specific domains. In addition, apolar contacts similar to that from Ser389 of SATB1 are observed for Thr26 of OCT-1 and Pit-1. It should be noted that the contacted phosphates and sugars are equivalent to those of the nucleotides contacted by SATB1-CUTr1 on the basis of the alignment in Figure 5c.

It is concluded, therefore, that the SATB1-CUTr1 shares its DNA-binding framework with POU-specific domains, where a significant number of contacts to DNA bases and backbones are conserved. Within the framework, contacts of bases by similar or different types of amino acids located at similar positions yield a partly similar recognition sequences. In the alignment in Figure 5c, a preference to the ATA/TAT sequence at positions 5–7/7′–5′ is clear. It should be noted that one of the sequences contacted by HNF1α (TTAATA) matches the (Y)TAATA recognition sequence of SATB1 when the relatively weak contact and the presumable preference to pyrimidine (Y) rather than purine (R) at the C2 position was considered, as described above.

Evolutionary implications

As described above, the CUT domains and POU-specific domains are two very similar subtypes of helix–turn–helix DNA-binding domains, which are highly likely to be evolutionarily related to each other. The conserved hydrogen-bonding network involving two Gln and an A base (Figure 3a,c and d; Figure 4b and d) is also observed for some other proteins with helix–turn–helix DNA-binding motif, such as 434 repressor (32), 434 Cro (33) and λ repressor (34), which all are from phages, but not for Trp repressor (35), CAP (36), lac repressor (37,38) or purine repressor (39), which are of bacterial origin. Therefore, this hydrogen-bonding network can be a strong clue to classifying the subfamilies of the helix–turn–helix DNA-binding domains, and CUT and POU-specific domains are likely to be related to phage repressors, rather than bacterial ones.

The proteins containing CUT or POU-specific domains are also likely to be related to each other. They commonly possess homeodomain regions more C-terminal to CUT or POU-specific domain, except for minor entries possessing only CUT or POU-specific domain, which appear in the NCBI Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Both the groups of the proteins are identified from *bilateria* group of animals that includes nematoda, insects and vertebrates, but not from more primitive animals, yeast or plants, while homeodomain proteins exist more widely in eukaryotes. Together with the structural relationship with phage repressors, we propose a hypothesis of a lateral transfer in which a phage repressor-like DNA-binding domain was incorporated into a homeodomain protein of a nematoda-like animal that is fed on bacteria occasionally transfected with phages. After the transfer, the proteins became divided into two groups along with the evolution of animals. During the evolution, these proteins are likely to keep functioning as transcriptional repressors or activators, although the target systems became quite diverse, e.g. nervous systems including pituitary gland, and blood and immune systems (40,41).

Regulation of the DNA-binding activity of the full-length SATB1

The most characteristic point in the interaction between SATB1–CUTr1 and a MAR–DNA is that only a single pair of direct hydrogen bonds are used to recognize bases (Figures 4b and 5a), which is atypical of DNA recognition from the major groove side mostly driven by direct hydrogen bonds (25,26). Although water-mediated hydrogen bonds, apolar contacts and other van der Waals contacts compensate them to achieve the sequence specificity, the affinity is not expected to be very strong. In fact, this CUT domain itself did not show significant DNA-binding activity when isolated from the other regions of MBD (13), and we attached four basic residues at C-terminus for the present crystallographic study. To gain stable DNA binding by the full-length protein, attachment of homeodomain at more C-terminal region (42) and dimerization driven by the N-terminal PDZ domain (43) are likely to be necessary. When the PDZ domain is removed by proteolysis by caspase 6, the DNA-binding activity of the protein as well as the activity in the transcription regulation is significantly reduced (43). Recently it was reported that acetylation by histone acetyltransferase at the N-terminal PDZ-domain region impairs the DNA-binding activity, implying a possibility of loss of dimerization (11). Therefore, affinity in the DNA-binding domain itself should not be too strong, in order to allow post-translational regulation of the DNA-binding activity of the protein in cells, and thereby effective transcriptional regulation.

Accession number

The co-ordinates of the structures determined from the laboratory and synchrotron data sets as well as the relevant structural factors have been deposited to the Protein Data Bank under accession IDs 2O49 and 2O4A, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank H. Yamaguchi for his skillful technical assistance in producing crystals, K. Chiba-Kamoshida, S. Kidokoro and Y. Shikamoto for helpful advice and valuable discussions. They also thank K. Suzuki, H. Shinshi, M. Ikemoto and S. Kurachi for the use of experimental facilities, and K. Kurachi for continuous support. This work was supported in part by the Ministry of Education, Science, Sports and Culture, Grand-in-Aid for Scientific Research (C), 17570103, 2005. Funding to pay the Open Access publication charges for this article was provided by AIST.

Conflict of interest statement. None declared.

REFERENCES

1. Gasser,S.M. and Laemmli,U.K. (1987) A glimpse at chromosomal order. *Trends Genet.*, **3**, 16–22.
2. Anderson,J.N. (1986) Detection, sequence patterns and function of unusual DNA structures. *Nucleic Acids Res.*, **14**, 8513–8533.
3. von Kries,J.P., Phi-Van,L., Diekmann,S. and Strätling,W.H. (1990) A non-curved chicken lysozyme 5′ matrix attachment site is 3′ followed by a strongly curved DNA sequence. *Nucleic Acids Res.*, **18**, 3881–3885.
4. Kohwi-Shigematsu,T. and Kohwi,Y. (1990) Torsional stress stabilizes extended base unpairing in suppressor sites flanking immunoglobulin heavy chain enhancer. *Biochemistry*, **29**, 9551–9560.
5. Bode,J., Kohwi,Y., Dickinson,L., Joh,T., Klehr,D., Mielke,C. and Kohwi-Shgematsu,T. (1992) Biological significance of unwinding capability of nuclear matrix-associating DNAs. *Science*, **255**, 195–197.
6. Dickinson,L.A., Joh,T., Kohwi,Y. and Kohwi-Shigematsu,T. (1992) A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. *Cell*, **70**, 631–645.
7. Yasui,D., Miyano,M., Cai,S., Varga-Weisz,P. and Kowhi-Shigematsu,T. (2002) SATB1 targets chromatin remodeling to regulate genes over long distances. *Nature*, **419**, 641–645.
8. Alvarez,J.D., Yasui,D.H., Niida,H., Joh,T., Loh,D.Y. and Kohwi-Shigematsu,T. (2000) The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. *Genes Dev.*, **14**, 521–535.
9. Wen,J., Huang,S., Rogers,H., Dickinson,L.A., Kohwi-Shigematsu,T. and Noguchi,C.T. (2005) SATB1 family protein expressed during early erythroid differentiation modifies globin gene expression. *Blood*, **105**, 3330–3339.
10. Cai,S., Han,H.J. and Kohwi-Shigematsu,T. (2003) Tissue-specific nuclear architecture and gene expression regulated by SATB1. *Nat. Genet.*, **34**, 42–51.
11. Kumar,P.P., Purbey,P.K., Sinha,C.K., Notani,D., Limaye,A., Jayani,R.S. and Galande,S. (2006) Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. *Mol. Cell*, **22**, 231–243.
12. Nakagomi,K., Kohwi,Y., Dickinson,L.A. and Kohwi-Shigematsu,T. (1994) A novel DNA-binding motif in the nuclear matrix attachment DNA-binding protein SATB1. *Mol. Cell. Biol.*, **14**, 1852–1860.
13. Yamaguchi,H., Tateno,M. and Yamasaki,K. (2006) Solution structure and DNA-binding mode of the matrix attachment region-binding domain of the transcriptional factor SATB1 that regulates the T-cell maturation. *J. Biol. Chem.*, **281**, 5319–5327.
14. Sheng,W., Yan,H., Rausa,F.M.III, Costa,R.H. and Liao,X. (2004) Structure of the hepatocyte nuclear factor 6α and its interaction with DNA. *J. Biol. Chem.*, **279**, 33928–33936.
15. Dekker,N., Cox,M., Boelens,R., Verrijzer,C.P., van der Vliet,P. and Kaptein,R. (1993) Solution structure of the POU-specific DNA-binding domain of Oct-1. *Nature*, **362**, 852–855.
16. Assa-Munt,N., Mortishire-Smith,R.J., Aurora,R., Herr,W. and Wright,P.E. (1993) The solution structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage λ repressor DNA-binding domain. *Cell*, **73**, 193–205.
17. Evans,P.R. (1993) Data reduction. In Sawyer,L., Issacs,N.W. and Bailey,S. (ed.), *Proceeding of the CCP4 Study Weekend: Data Collection and Processing*. Daresbury Laboratory, Warrington, pp. 114–122.
18. French,G.S. and Wilson,K.S. (1978) On the treatment of negative intensity observations. *Acta Crystallogr. A*, **34**, 517–525.
19. Collaborative Computational Project. Number 4. (1994) The CCP4 Suite: programs for protein crystallography. *Acta Crystallogr. D*, **50**, 760–763.
20. Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.*, **276**, 307–326.
21. Brünger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuzewski,J., Nilges,M. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D*, **54**, 905–921.
22. McRee,D.E. (1999) XtalView/Xfit – a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.*, **125**, 156–165.
23. Laskowski,R.A., MacArthur,W.M., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.*, **26**, 283–291.
24. Koradi,R., Billeter,M. and Wüthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.
25. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.

26. Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
27. Lannoy,V.L., Brüglin,T.R., Rousseau,G.G. and Lemaigre,F.P. (1998) Isoforms of hepatocyte nuclear factor-6 differ in DNA-binding properties, contain a bifunctional homeodomain, and define the new ONECUT class of homeodomain proteins. *J. Biol. Chem.*, **273**, 13552–13562.
28. Truscott,M., Raynal,L., Wang,Y., Bérubé,G., Leduy,L. and Nepveu,A. (2004) The N-terminal region of the CCAAT displacement protein (CDP)/Cux transcription factor functions as an autoinhibitory domain that modulates DNA binding. *J. Biol. Chem.*, **279**, 49787–49794.
29. Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,CO. (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell*, **77**, 21–32.
30. Jacobson,E.M., Li,P., Leon-del-Rio,A., Rosenfeld,M.G. and Aggarwal,A.K. (1997) Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. *Genes Dev.*, **11**, 198–212.
31. Chi,Y.I., Frantz,J.D., Oh,B.C., Hansen,L., Dhe-Paganon,S. and Shoelson,S.E. (2002) Diabetes mutations delineate an atypical POU domain in HNF-1α. *Mol. Cell*, **10**, 1129–1137.
32. Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science*, **242**, 899–907.
33. Mondragon,A. and Harrison,S.C. (1991) The phage 434 Cro/O_{R1} complex at 2.5 Å resolution. *J. Mol. Biol.*, **219**, 321–334.
34. Beamer,L.J. and Pabo,C.O. (1992) Refined 1.8 Å crystal structure of the λ repressor-operator complex. *J. Mol. Biol.*, **227**, 177–196.
35. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P. (1988) Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
36. Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90°. *Science*, **253**, 1001–1007.
37. Chuprina,V.P., Rullmann,J.A., Lamerichs,R.M., van Boom,J.H., Boelens,R. and Kaptein,R. (1993) Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J. Mol. Biol.*, **234**, 446–462.
38. Bell,C.E. and Lewis,M. (2000) A closer view of the conformation of the Lac repressor bound to operator. *Nat. Struct. Biol.*, **7**, 184–187.
39. Schumacher,M.A., Choi,K.Y., Zalkin,H. and Brennan,R.G. (1994) Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science*, **266**, 763–770.
40. Ruvkun,G. and Finney,M. (1991) Regulation of transcription and cell identity by POU domain proteins. *Cell*, **64**, 475–478.
41. Nepveu,A. (2001) Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development. *Gene*, **270**, 1–15.
42. Dickinson,L.A, Dickinson,C.D. and Kohwi-Shigematsu,T. (1997) An atypical homeodomain in SATB1 promotes specific recognition of the key structural element in a matrix attachment region. *J. Biol. Chem.*, **272**, 11463–11470.
43. Galande,S., Dickinson,L.A., Mian,I.S., Sikorska,M. and Kowhi-Shigematsu,T. (2001) SATB1 cleavage by caspase 6 disrupts PDZ domain-mediated dimerization, causing detachment from chromatin early in T-cell apoptosis. *Mol. Cell. Biol.*, **21**, 5591–5604.
44. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgin,D.G. (1997) The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
45. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
46. Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.