Sequence-specific Methyltransferase-Induced Labeling of DNA (SMILing DNA)

Goran Pljevaljčić,[b] Falk Schmidt,[a] and Elmar Weinhold*[a]

A new concept for sequence-specific labeling of DNA by using chemically modified cofactors for DNA methyltransferases is presented. Replacement of the amino acid side chain of the natural cofactor S-adenosyl-L-methionine with an aziridine group leads to a cofactor suitable for DNA methyltransferase-catalyzed sequence-specific coupling with DNA. Sequence-specifically fluorescently labeled plasmid DNA was obtained by using the DNA methyltransferase from Thermus aquaticus *(M.*Taq*I) as catalyst and attaching a fluorophore to the aziridine cofactor. First results suggest that all classes of DNA methyltransferases with different recognition sequences can be used. In addition, this novel method for DNA labeling should be applicable to a wide variety of reporter groups.*

1. Introduction

The completion of the human genome sequencing project[1] emphasizes the need for more structural and functional studies. A major task in investigating bionanostructures is to make them visible. In this respect, intrinsic physical properties of biopolymers or reporter molecules specifically conjugated to them have been utilized for visualization, and monitoring binding and dynamics. The attached reporter molecule must be chosen such that it does not interfere with the structure of a biopolymer or hinder its interaction with other molecules. Bearing in mind the size of biopolymers and the recurrence of a small number of monomeric units, sequence-specific labeling is a challenging task. Here we present a new concept for sequence-specific labeling of DNA which should also be applicable for sequence-specific labeling of RNA and proteins.

2. Labeling of DNA

In principle, DNA can be labeled either by chemical or enzymatic methods. Phosphoramidite solid-phase DNA synthesis is the standard chemical method to synthesize labeled short oligodeoxynucleotides by incorporation of modified phosphoramidites.[2] The modified unit can either contain the reporter group or a linker with a functional group for post-synthetic modification. Chemical modification of nucleic acids can also be achieved by bisulfite-catalyzed transamination of the N4-amino group of cytosine residues, bromination of the C5-position of pyrimidines and the C8-position of purines, followed by reaction with derivatized amines or photochemical coupling of azides like photobiotin, to name just a few methods.[3] In addition, platinum-based reagents have been used to label nucleic acids.[4] Enzymatic labeling methods mainly rely on the ability of DNA polymerases or terminal transferases to incorporate modified deoxynucleoside triphosphates.[3] Random labeling can be performed with Klenow polymerase (random primed labeling), *Escherichia coli* DNA polymerase I (nick translation) or *Taq* DNA polymerase (polymerase chain reaction, PCR), and end-labeling can be achieved with terminal deoxynucleotidyl transferase. In general, these chemical and enzymatic methods are not suitable for sequence-specific internal labeling of long DNA molecules.

3. Sequence-Specific Recognition of DNA by Small Molecules

A requirement for sequence-specific DNA labeling is the aptitude of molecular structures for sequence-specific DNA recognition. Several types of small molecules have been developed for this purpose.

DNA intercalators generally bind with low sequence-specificity to DNA by incorporating the planar aromatic systems between two base pairs. Great efforts have been made to find intercalators with increased sequence-specificities, and a peptide-linked *bis*-intercalator which sequence-specifically recognizes a six-base-pair sequence was recently presented.[5]

Short triple-helix-forming oligodeoxynucleotides (TFO) can be used for sequence-specific targeting of double-stranded DNA. They preferentially bind to homopurine·homopyrimidine sequences by forming triple helices.[6] The two strands of the targeted double helix form usual Watson–Crick base pairs, and

[a] *F. Schmidt, Prof. Dr. E. Weinhold*
Institut für Organische Chemie der RWTH Aachen
Professor-Pirlet-Strasse 1, 52056 Aachen (Germany)
Fax: (+49) 241-80-92528
E-mail: elmar.weinhold@oc.rwth-aachen.de

[b] *Dr. G. Pljevaljčić*
The Scripps Research Institute, Department of Molecular Biology, MB19
10550 North Torrey Pines Road, La Jolla, CA 92037 (USA)

 DOI: 10.1002/cbic.200300739

the third strand binds via Hoogsteen hydrogen bonding in the major groove. Although different TFO binding motifs have been observed, the approach still remains limited to purine-rich sequences. Very stable triple helices can be obtained with peptide nucleic acid (PNA). Interestingly, PNA does not preferentially lead to classical triplexes with double-stranded DNA, but forms triplex invasion complexes in which two PNA molecules bind to DNA.[7]

In addition to the DNA major groove, the minor groove can be targeted. Several natural products like Netropsin and Distamycin bind in the minor groove. Based on structural DNA binding studies of Distamycin (a polyamide containing *N*-methylpyrrole repeating units), hairpin polyamides have been developed. They are composed of pyrrole, imidazole and 3-hydroxypyrrole amino acids, which give them the ability to distinguish between all four possible base-pair combinations in the minor groove and to recognize DNA sequence-specifically.[8]

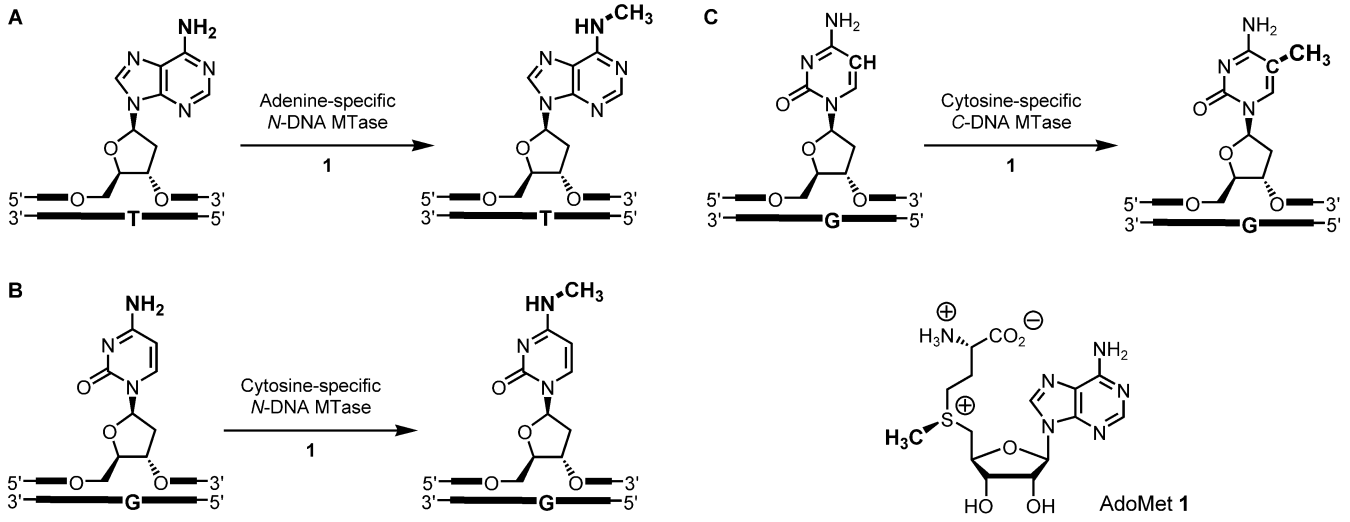
Although enormous progress has been made in the field of sequence-specific DNA recognition by small molecules, bioorganic chemists are just starting to attach reporter groups to these molecular structures and use them for sequence-specific DNA labeling. Recently, sequence-specific but not covalent labeling of DNA was achieved with a hairpin polyamide[9] and a TFO-based approach.[10] In addition, sequence-specific covalent biotinylation was reported with biotin – TFO – psoralen conjugates by photochemical coupling, reaching a maximum yield of 80% after lengthy sequence optimization.[11] Given the difficulties and limitations associated with these synthetic systems for sequence-specific DNA labeling, it is interesting to have a look at biological systems.

4. Sequence-Specific DNA Labeling with a Methyl Group

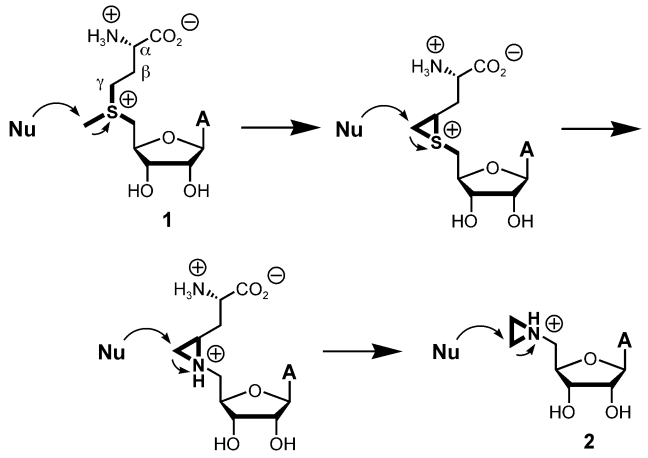
Nature has evolved many different proteins that interact specifically with DNA. They can exert their biological function by binding to appropriate sequences or by covalent modification of DNA. In this respect, DNA methyltransferases (MTases) are a very interesting class of enzymes because they combine both the ability of sequence-specific DNA recognition and sequence-specific covalent-bond formation. They catalyze the nucleophilc attack of either adenine or cytosine residues within specific double-stranded DNA sequences onto the activated methyl group of the cofactor *S*-adenosyl-L-methionine (AdoMet, **1**), and can be regarded as enzymes that label DNA with a methyl group. DNA MTases are categorized into two classes according to the atom that is modified. *N*-DNA MTases catalyze the methylation of the exocyclic amino group of either adenine or cytosine, and *C*-DNA MTases catalyze the methylation of the 5-position of cytosine (Scheme 1).[12]

5. A Novel Cofactor for DNA MTases

Since the de novo design and synthesis of molecular entities capable of sequence-specific DNA recognition and modification is still a very challenging task, we considered utilizing DNA MTases for sequence-specific DNA labeling. Evidently, the methyl group is not an attractive reporter group if one wants to avoid the safety, stability and disposal problems associated with radioactive isotopes. Thus, engineering novel cofactors for DNA MTases that change the function of the enzymes and lead to transfer of larger chemical entities onto DNA could be very interesting for sequence-specific DNA labeling. In principal, replacement of the methylsulfonium moiety in AdoMet **1** with a thiiranium ring, by linking the activated methyl group with Cγ of the amino acid side chain, could lead to nucleophilic ring opening and hence coupling of the whole cofactor to the target sequence (Scheme 2). However, thiiranium compounds are unstable in nucleophilic solvents.[13] Therefore, we decided to introduce the more stable aziridine ring, which can be activated for nucleophilic attack by protonation of the ring nitrogen atom. Since cofactor binding studies with different DNA MTases



Scheme 1. *DNA methyltransferases (MTases) naturally catalyze the methyl group transfer from the cofactor S-adenosyl-L-methionine (AdoMet, **1**) onto nucleobases within double-stranded DNA. Adenine-specific N-DNA MTases form N6-methyladenine (A), cytosine-specific N-DNA MTases form N4-methylcytosine (B) and cytosine-specific C-DNA MTases form C5-methylcytosine (C) within their DNA recognition sequences (indicated by thick lines).*



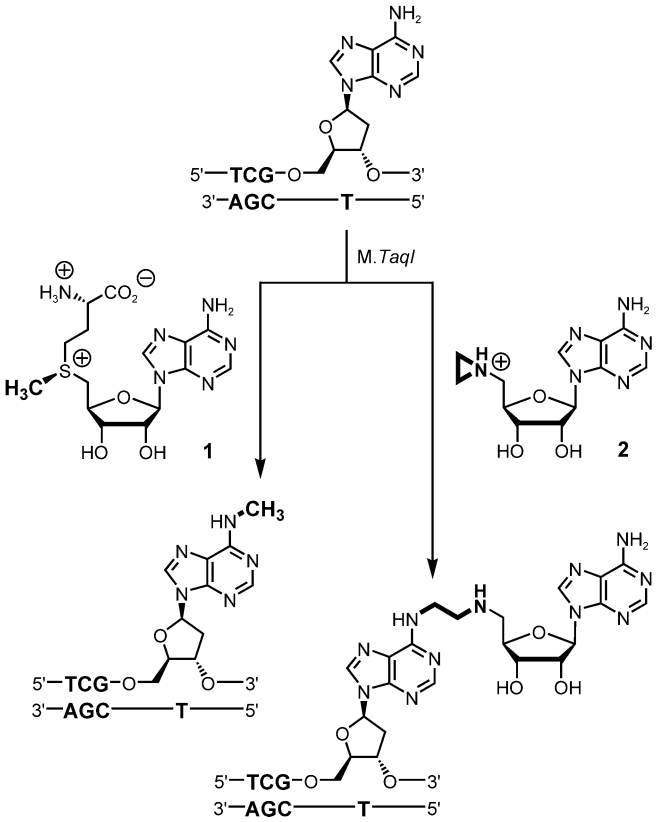
Scheme 2. *Design of the new cofactor* N*-adenosylaziridine (**2**) starting from the natural cofactor* S*-adenosyl-*L*-methionine (**1**) for MTase-catalyzed coupling with DNA (A = adenine).*

revealed that the amino acid side chain is not essential for cofactor binding,[14] we synthesized the new nucleoside *N*-adenosylaziridine (**2**), which lacks the amino acid part, by nucleophilic substitution of 5′-deoxy-5′-tosyladenosine with aziridine. Using the DNA MTase from *Thermus aquaticus* (M.*Taq*I) as a paradigm, we established a DNA MTase-catalyzed transfer of more than just a single carbon unit (methyl group), namely the cofactor itself, onto DNA.[15] The new cofactor **2** is quantitatively coupled with the exocylic amino group of adenine within the 5′-TCGA-3′ recognition sequence of M.*Taq*I (Scheme 3).

Most importantly, this cofactor design does not introduce additional steric bulk compared with the natural cofactor and does not take into account the individual mechanisms of *N*- and *C*-DNA MTases for activating their target bases as nucleophiles. Thus, it is not surprising that *N*-adenosylaziridine (**2**) functions as a general cofactor for all classes of DNA MTases, as demonstrated by the sequence-specific coupling catalyzed not only by N6-adenine but also by N4-cytosine and C5-cytosine DNA MTases.[16]

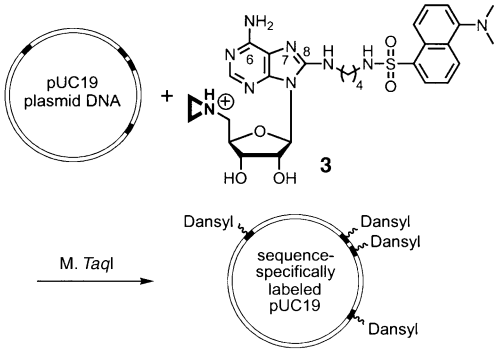
6. Sequence-Specific *M*ethyltransferase-*I*nduced *L*abel*ing* of DNA (SMILing DNA)

In principle, selective 1,2-diol cleavage of the attached adenosine residue with periodate followed by reaction of the resulting dialdehyde with a primary amine containing a reporter group and subsequent reduction,[17] could lead to sequence-specific labeling of DNA. Although such a multistep procedure appears feasible, a one-step labeling reaction would be much more convenient. We, therefore, investigated the possibility of using the new aziridine cofactor in combination with DNA MTases as a delivery system for reporter groups. For this purpose, a suitable attachment position at the aziridine cofactor must be found. Inspection of the three-dimensional structure of M.*Taq*I in complex with a cofactor analogue and DNA[18] revealed that the 8-position of the adenine ring points towards the solvent. Therefore, attaching a reporter group via a flexible linker at this position should not block cofactor binding (Figure 1). Recently,



Scheme 3. *Reactions catalyzed by the adenine-specific DNA MTase M.*Taq*I. Nucleophilc attack of the exocyclic adenine amino group on the activated methyl group of the natural cofactor* S*-adenosyl-*L*-methionine (**1**) leads to sequence-specific methyl-group transfer (left) and nucleophilc ring opening of the new cofactor* N*-adenosylaziridine (**2**); this results in sequence-specific coupling of the whole cofactor with DNA (right).*

we demonstrated that such an aziridine cofactor **3** containing a dansyl fluorophore at the 8-position is quantitatively and sequence-specifically coupled by M.*Taq*I with small duplex oligodeoxynucleotides as well as plasmid DNA (Scheme 4).[19] In addition, the aziridine cofactor **4**, which contains a biotin group instead of the dansyl group, functions as a cofactor for M.*Taq*I, as demonstrated by almost complete protection of



Scheme 4. *Sequence-specific fluorescence labeling of pUC19 plasmid DNA (2686 base pairs). The aziridine cofactor **3** containing a dansyl fluorophore attached via a flexible linker at the 8-position of the adenine ring, is quantitatively and sequence-specifically coupled with all four 5′-TCGA-3′ recognition sequences (black bars) by M.*Taq*I.*

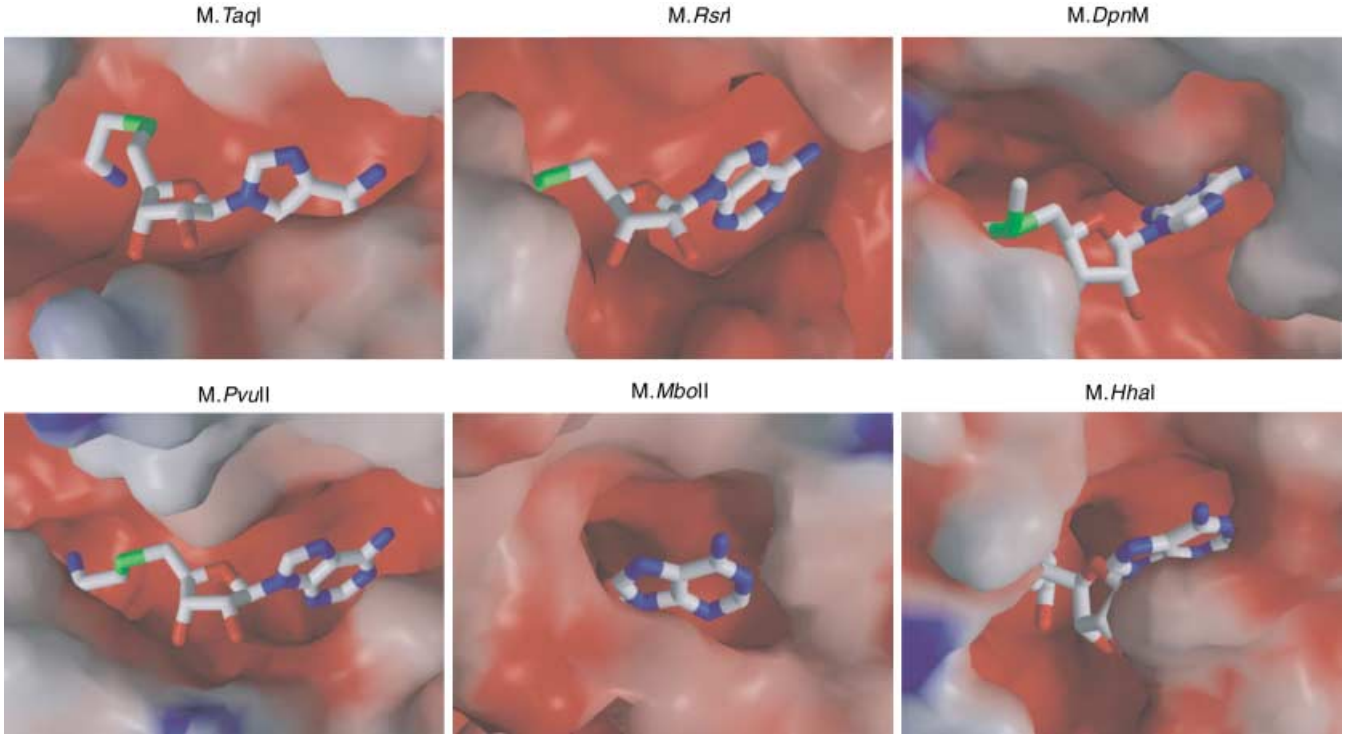


Figure 1. *Cofactor binding pockets of DNA MTases. The enzymes are shown as an electrostatic surface potential structure (red = negative potential, blue positive potential and grey is neutral). The cofactors or cofactor analogues are presented as stick models. The following Protein Data Bank entry numbers of cocrystal structures were used to prepare the pictures: 1G38 (M.*TaqI*), 1EG2 (M.*RsrI*), 2DPM (M.*DpnM*), 1BOO (M.*PvuII*), 1G60 (M.*MboII*), 1MHT (M.*HhaI*).*

linearized plasmid DNA against fragmentation by the corresponding restriction endonuclease, R.*Taq*I (Figure 2). Furthermore, preliminary experiments in our laboratory indicate that an aziridine cofactor with the (negatively charged) cyanine5 fluorophore attached to the 8-postion can be used in a M.*Taq*I-catalyzed coupling reaction for the sequence-specific labeling of DNA. Thus, it is expected that many more reporter and affinity groups can be sequence-specifically delivered to DNA.

Most interestingly, the known crystal structures of six DNA MTases with different sequence-specificities in complex with AdoMet **1** or cofactor analogues (and DNA for M.*Taq*I and M.*Hha*I) reveal that the 8-position of the cofactor adenine ring is not only accessible in M.*Taq*I but also in M.*Rsr*I, M.*Dpn*M and M.*Pvu*II (Figure 1). Thus, aziridine labeling cofactors with reporter groups at the 8-position could also be substrates for these DNA MTases. However, in the M.*Mbo*II and M.*Hha*I structure the 8-position appears to be blocked by the cofactor binding pocket. For the latter two enzymes, attachment of the reporter group to the 7-position (7-deazaadenosine derivative) or to the 6-position of the aziridine cofactor should be more suitable.

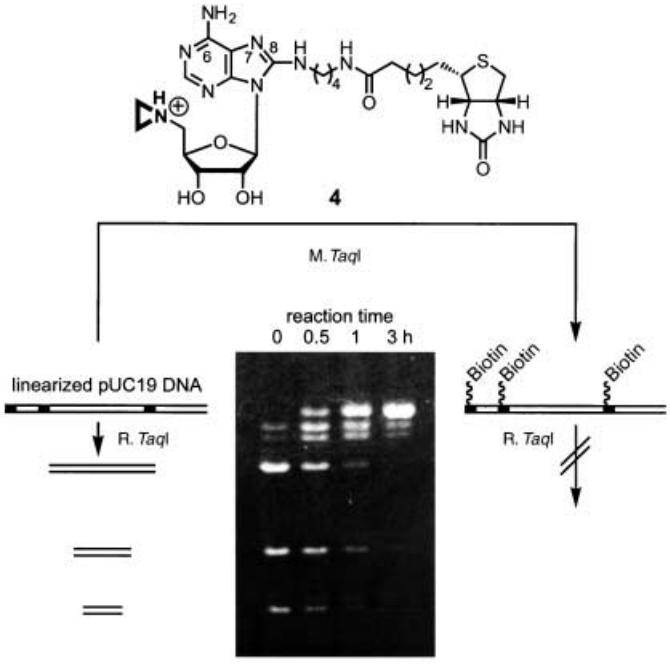


Figure 2. *M.*TaqI*-catalyzed labeling of linearized plasmid DNA with biotin. The progress of the labeling reaction with the aziridine cofactor **4** is analyzed in a DNA protection assay. At the beginning of the labeling reaction (reaction time 0 h) fragmentation of R.*EcoRI*-linearized pUC19 DNA with the restriction endonuclease R.*TaqI *leads to three major DNA fragments on an agarose gel (one of the four* TaqI *recognition sequences in pUC19 is destroyed by R.*EcoRI *linearization). With increasing reaction times (0.5 and 1 h) these bands disappear and bands corresponding to longer intermediates or the full-length linearized plasmid appear. After 3 h the DNA is almost completely protected against fragmentation by R.*TaqI*; this indicates that the three* TaqI *recognition sequences are blocked by covalent modification. No DNA protection against cleavage by R.*TaqI *is observed in the absence of either M.*TaqI *or the aziridine cofactor **4** (not shown).*

7. Perspectives

SMILing DNA has the potential to become a new general tool for sequence-specific DNA labeling and is characterized by high versatility and flexibility. It can be expected that not only a wide range of reporter groups and affinity labels but also larger chemical entities like peptides, oligosaccharides, oligonucleotides, gold clusters or proteins are suitable for sequence-specific attachment to DNA. In addition, REBASE,[20] a data base for restriction endonucleases and DNA MTases, currently lists about 900 DNA MTases with a multitude of recognition sequences ranging from two to eight base pairs. Thus, a great number of DNA sequences could be targeted. SMILing DNA can be regarded as a new labeling technology with potential applications in functional studies of DNA modifying enzymes, medical diagnosis, molecular biology, and nanobiotechnology.

The new labeling cofactors could also be useful for sequence-specific labeling of RNA and proteins by using AdoMet-dependent RNA and protein MTases. Table 1 lists the accessibility of the 6-, 7- and 8-positions of the cofactor adenine ring in several cocrystal structures. In most cases at least one position appears suitable for attaching a reporter group. Thus, *S*equence-specific *M*ethyltransferase-*I*nduced *L*abel*ing* (SMILing) could become a general labeling technique for sequence-specific incorporation of reporter molecules into DNA, RNA and protein biopolymers.

Table 1. *Accessibilities of the 6-, 7- and 8-positions of the cofactor adenine ring in known crystal structures of RNA and protein MTases in complex with the cofactor or analogues.*

MTase	Substrate	Cofactor[a]	Accessibility[b]			Pdb[c]
			6	7	8	
RNA MTases						
VP39 (Vaccinia virus)	mRNA	AdoHcy	+	+	(+)	1AV6
FtsJ/RrmJ (*E. coli*)	23S rRNA	AdoMet	(+)	+	−	1EIZ
NS5 (Dengue virus)	mRNA	AdoHcy	(+)	+	−	1L9K
ErmC′ (*B. subtilis*)	23S rRNA	AdoMet	+	+	−	1QAO
TrmD (*H. influenza*)	tRNA	AdoMet	−	−	+	1UAM
Protein MTases						
CheR (*S. typhimurium*)	chemotaxis receptor	AdoHcy	−	−	(+)	1BC5
PIMT (*T. maritima*)	proteins isoaspartyl	AdoHcy	−	−	+	1DL5
PRMT3 (*R. norvegicus*)	proteins arginine	AdoHcy	−	−	−	1F3L
PIMT (*P. furiosus*)	proteins isoaspartyl	Ado	−	−	−	1JG3
hPIMT	proteins	AdoHcy	(+)	−	(+)	1I1N
(*H. sapiens*)	isoaspartyl	AdoHcy	(+)	−	(+)	1KR5
LSMT (*P. sativum*)	proteins lysine	AdoMet	+	(+)	−	1P0Y
SET7/9	histone H3	AdoMet	(+)	−	−	1N6C
(*H. sapiens*)	lysine	AdoHcy	(+)	−	−	1O9S
HemK/PrmC (*T. maritima*)	proteins glutamic acid	AdoMet	(+)	+	−	1NV8
hDOT1L (*H. sapiens*)	histone H3 lysine	AdoMet	(+)	+	−	1NW3
PRMT1 (*R. norvegicus*)	proteins arginine	AdoHcy	+	+	+	1ORI

[a] AdoMet = S-adenosyl-L-methionine, AdoHcy = S-adenosyl-L-homocysteine, Ado = adenosine. [b] + = accessible, (+) = less accessible, − = blocked. [c] Pdb = Protein Data Bank entry number.

Acknowledgements

Research in our laboratory was supported by the Deutsche Forschungsgemeinschaft, the Max Planck Society and the QIAGEN GmbH.

Keywords: bioorganic chemistry · cofactors · DNA methylation · enzyme catalysis · transferases

[1] a) J. C. Venter et al., *Science* **2001**, *291*, 1304 – 1351; b) International Human Genome Sequencing Consortium, *Nature* **2001**, *409*, 860 – 921.
[2] M. H. Caruthers, A. D. Barone, S. L. Beaucage, D. R. Dodds, E. F. Fisher, L. J. McBride, M. Matteucci, Z. Stabinsky, J. Y. Tang, *Methods Enzymol.* **1987**, *154*, 287 – 313.
[3] C. Kessler in *Nonisotopic Probing, Blotting, and Sequencing* (Ed. L. J. Kricka), Academic Press, San Diego, **1995**, pp. 41 – 109.
[4] R. J. Heetebrij, E. G. Talman, M. A. v. Velzen, R. P. M. v. Gijlswijk, S. S. Snoeijers, M. Schalk, J. Wiegant, F. v. d. Rijke, R. M. Kerkhoven, A. K. Raap, H. J. Tanke, J. Reedijk, H.-J. Houthoff, *ChemBioChem* **2003**, *4*, 573 – 583.
[5] V. Guelev, J. Lee, J. Ward, S. Sorey, D. W. Hoffman, B. L. Iverson, *Chem. Biol.* **2001**, *8*, 415 – 425.
[6] a) H. E. Moser, P. B. Dervan, *Science* **1987**, *238*, 645 – 650; b) T. Le Doan, L. Perrouault, D. Praseuth, N. Habhoub, J.-L. Decout, N.-T. Thuong, J. Lhomme, C. Helene, *Nucleic Acids Res.* **1987**, *15*, 7749 – 7760.
[7] P. E. Nielsen, *Methods Enzymol.* **2001**, *340*, 329 – 340.
[8] S. White, J. W. Szewczyk, J. M. Turner, E. E. Baird, P. B. Dervan, *Nature* **1998**, *391*, 468 – 471.
[9] V. C. Rucker, S. Foister, C. Melander, P. B. Dervan, *J. Am. Chem. Soc.* **2003**, *125*, 1195 – 1202.
[10] T. Roulon, C. Helene, C. Escude, *Angew. Chem.* **2001**, *113*, 1571 – 1574; *Angew. Chem. Int. Ed.* **2001**, *40*, 1523 – 1526.
[11] C. Pfannschmidt, A. Schaper, G. Heim, T. M. Jovin, J. Langowski, *Nucleic Acids Res.* **1996**, *24*, 1702 – 1709.
[12] a) X. Cheng, *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 293B318; b) A. Jeltsch, *ChemBioChem* **2002**, *3*, 275 – 293.
[13] W. A. Smit, M. Z. Krimer, E. A. Vorob'eva, *Tetrahedron Lett.* **1975**, 2451 – 2454.
[14] a) M. Pignot, G. Pljevaljčić, E. Weinhold, *Eur. J. Org. Chem.* **2000**, 549 – 555; b) M. Pignot, Dissertation, Universität Dortmund (Germany), **1999**.
[15] M. Pignot, C. Siethoff, M. Linscheid, E. Weinhold, *Angew. Chem.* **1998**, *110*, 3050 – 3053; *Angew. Chem. Int. Ed.* ***1998***, *37*, 2888 – 2891.
[16] G. Pljevaljčić, Dissertation, Universität Dortmund (Germany), **2002**.
[17] R. E. Hileman, K. M. Parkhurst, N. K. Gupta, L. J. Parkhurst, *Bioconjugate Chem.* **1994**, *5*, 436 – 444.
[18] K. Goedecke, M. Pignot, R. S. Goody, A. J. Scheidig, E. Weinhold, *Nat. Struct. Biol.* **2001**, *8*, 121 – 125.
[19] G. Pljevaljčić, M. Pignot, E. Weinhold, *J. Am. Chem. Soc.* **2003**, *125*, 3486 – 3492.
[20] a) R. J. Roberts, T. Vincze, J. Posfai, D. Macelis, *Nucleic Acids. Res.* **2003**, *31*, 418 – 420; b) http://rebase.neb.com/rebase/rebase.html

Received: August 12, 2003 [C739]

