Molecular evolution of DNA-(cytosine-N4) methyltransferases: evidence for their polyphyletic origin

Janusz M. Bujnicki* and Monika Radlinska

Molecular Biology Research Program, Henry Ford Health System, One Ford Place Suite 5D, Detroit, MI 48202, USA

Received June 22, 1999; Revised and Accepted September 21, 1999

ABSTRACT

DNA N4-cytosine methyltransferases (N4mC MTases) are a family of *S*-adenosyl-L-methionine (AdoMet)-dependent MTases. Members of this family were previously found to share nine conserved sequence motifs, but the evolutionary basis of these similarities has never been studied in detail. We performed phylogenetic analysis of 37 known and potential new family members from the multiple sequence alignment using distance matrix, parsimony and maximum likelihood approaches to infer the evolutionary relationship among the N4mC MTases and classify them into groups of orthologs. All the treeing algorithms employed as well as results of exhaustive sequence database searching support a scenario, in which the majority of N4mC MTases, except for M.*Bal*I and M.*Bam*HI, arose by divergence from a common ancestor. Interestingly, MTases M.*Bal*I and M.*Bam*HI apparently originated from N6-adenine MTases and represent the most recent addendum to the N4mC MTase family. In addition to the previously reported nine sequence motifs, two more conserved sequence patches were detected. Phylogenetic analysis also provided the evidence for massive horizontal transfer of MTase genes, presumably with the whole restriction-modification systems, between Bacteria and Archaea.

INTRODUCTION

DNA methylation is catalyzed by DNA MTases, transferring the methyl group from the AdoMet molecule to certain N and C atoms of nucleotides. Modification of genomic DNA of most organisms plays a role in a variety of biological processes, including regulation of gene expression, DNA replication, mismatch repair and defense of the host against foreign DNA (reviewed in 1,2). DNA methylation leads to the formation of three kinds of products: N6-methyladenine (N6mA), N4-methylcytosine (N4mC) and 5-methylcytosine (5mC). Because of the chemical character of the reaction catalyzed by N4mC and N6mA DNA MTases (methylation of exocyclic -NH_2 group), they are both grouped as one class, N-MTases (3). The methylation of 5mC is widespread in all branches of the tree of life. N6-adenine methylation, common to all bacteria, has been also reported in the ciliated protozoa (4). To our knowledge, however, N4mC has been found only in Prokaryota and Archaea. Moreover, contrary to the diversity of the biological function of DNA modification, N4-methylation seems to be primarily a component of restriction–modification systems (R–M) with the exception of M.*Ngo*MXV MTase (5; former name M.*Ngo*MV), for which no corresponding endonucleolytic activity has been found. Nowadays the number of known MTase sequences is difficult to estimate precisely as genome and other sequence data continue to pour into databases at a fast rate, but despite the growing number of putative N4mC MTases, this group remains minor compared to N6mA and 5mC MTases (6).

All DNA MTases share a common building plan, with a pattern of highly conserved amino acid sequence blocks. A set of ten motifs arranged in a constant linear order is found among most 5mC MTases along with a variable region, which confers sequence specificity (7). Although N-MTases seem to be a much less homogenous class than 5mC MTases, Malone *et al.* (8) were able to identify nine segments of similarity in the sequence alignment of 45 N-MTases (36 N6mA and only nine N4mC) corresponding to motifs I–VIII and X in 5mC MTases. Based on relative position of two most conserved of these motifs (I and IV) and the variable region N-MTases were classified as α, β and γ (9). Group α is arranged in the order, motif I–variable region–motif IV; group β, motif IV–variable region–motif I; and group γ, motif I–motif IV–variable region (9). The N6mA MTases were found in all these classes, while the majority of N4mC MTases aggregated into the β group with only one representative in the α group, and none in the γ group (8). Only recently a N4mC MTase was described with an order of motifs similar to that of γ-MTases, however lacking the typical variable region at the C-terminus (5,10).

Crystal structures have been determined for a number of AdoMet-dependent MTases, including two 5mC, M.*Hha*I (11) and M.*Hae*III (12); two N6mA, M.*Taq*I (13) and M.*Dpn*M (14); and one N4mC DNA MTase, M.*Pvu*II (15). All of these enzymes share a remarkably similar catalytic domain structure, resembling an α/β Rossmann-fold with conserved binding patterns for the cofactor AdoMet and modified base corresponding mainly to conserved motifs I and IV (16). In all cases the substrate to be methylated is bound or expected to bind in a pocket adjacent to the AdoMet binding site, which is formed by different amino acids in different MTases. The binding

*To whom correspondence should be addressed. Tel: +1 313 874 61 28; Fax: +1 313 876 23 80; Email: iamb@ibbrain.ibb.waw.pl
Permanent address:
Monika Radlinska, Institute of Microbiology, University of Warsaw, Nowy Swiat 67, 00-046 Warsaw, Poland

mode of N-MTases for their DNA target (different in all examined enzymes) has been suggested from the relative orientation of either additional target recognition domains (TRD) or assemblies of flexible loops, and accumulation of positive electrostatic charge in certain regions of protein surface (14–16). The site of the flipped-out nucleotide binding has been also postulated which has suggested a possible reaction mechanism, different from that of 5mC MTases (15,16).

It has been proposed that N6mA and N4mC MTases, which closely resemble one another, derive from a common ancestor (17). Recently, Jeltsch *et al.* (18) demonstrated that the catalytic activities of these two families overlap to some degree. However, a phylogenetic analysis of MTases utilizing superposition of tertiary structures and resulting rmsd values along with a structure-guided sequence alignment, which included representatives of N4mC and N6mA families, argues against their close common origin (19). N4mC and N6mA MTases are found on distinct branches of a tree, suggesting very ancient divergence of both subfamilies of N-MTases and opening possibilities for subsequent functional convergence.

In this paper we investigate the phylogenetic history of the N4mC MTase family and ask whether discrepancies between their function and degree of sequence similarity arose by divergence, or are evidence for convergent evolution. We compare enzymes from different structural classes (α, β and γ-like) and propose a non-trivial scheme describing their divergence from a common ancestor.

MATERIALS AND METHODS

Amino acid sequences of all previously characterized members of the N4mC family were taken from publicly available databases through the REBASE catalog (6) and the PSI-BLAST program (20) was used for iterative multiple database searches with all of them as queries. The databases used in this search were the non-redundant (NR) database and both the complete and unfinished genomes obtained through the BLAST interface (http://www.ncbi.nlm.nih.gov/BLAST/) at the NCBI. The assignment of putative protein sequences as members of N4mC MTase family was based on high homology to known N4mC sequences according to the BLAST default cutoff values. All sequences were subsequently aligned using the CLUSTALX program (21). After the refinement of poorly aligned regions or subsets of sequences, manual adjustments were introduced based on the PSI-BLAST pairwise comparison and secondary structure prediction [carried out using consense JPRED approach (22), data not shown]. All sequences that appeared truncated, defective or only marginally similar to N4mC MTases were excluded from further analysis.

The phylogenetic trees were inferred from the sequence alignments using distance, parsimony and maximum likelihood algorithms implemented in programs available in the PHYLIP package (23 and references therein). In a distance matrix method, evolutionary distances (representing an estimate of the number of amino acid substitutions per site) were computed for all protein pairs, and a phylogenetic tree was reconstructed by using an algorithm of Fitch and Margoliash (24). According to the principle of maximum likelihood, for a possibly large set of trees a search for the maximum likelihood value was carried out for the patterns of amino acid differences among the sequences considering each site separately, and the tree with the largest value was chosen as the preferred one. Using a maximum parsimony method a tree was generated, which required the possibly smallest number of evolutionary changes to explain the differences observed among the sequences under study (methodology comprehensively reviewed in 25,26).

Since in all methods employed, each alignment position is assumed to include residues sharing common ancestry, regions of ambiguous alignment and most extensive gaps were excluded from the phylogenetic analysis. The distances proportional to the number of amino acid replacements per sequence position separating each pair of sequences were estimated using the JTT model (27) and the phylogenetic tree that best fits the sequence-to-sequence distances was generated with the KITSCH program. The trees that best fit the parsimony and maximum likelihood criteria were generated with the PROTPARS and PROTML programs respectively. Multiple runs were conducted using up to 20 different input orders, with global rearrangements and the subreplicates options used wherever possible to find an optimal (or nearly optimal) tree. The length of branches in each consensus tree computed using the majority-rule method CONSENSE was calculated with the FITCH program. The consistency of each tree was evaluated by the bootstrap resampling of the original sequence data using the SEQBOOT program. In this technique all alignment positions were randomly sampled with replacement from the original sequence set (28). The process was repeated 100 times, and a set of randomized alignments was used for reconstruction of new phylogenetic trees. The clusters with high proportion of occurrence among all the trees were considered to be statistically significant (26).

RESULTS

Taking advantage of all sequences deposited in databases and the 18 completed (four archaeal and 14 bacterial) and 34 (including three eukaryotic) unfinished genome sequences we have identified 37 proteins and putative proteins with extensive amino acid sequence similarity to the N4mC MTases. Nine homologs of N4mC MTases from Archaea and 28 from Bacteria have been identified, their absence from eukaryotic sequences has been also confirmed (Table 1). Many sequences of new family members have been obtained by genome sequencing projects that do not provide any information about biological function or biochemical activity of putative proteins and even if such information exist it is often incomplete and sometimes incorrect (29). It is worth emphasizing that homologs of known MTases were found only in two of 14 completely sequenced bacterial genomes, but in two of four archaeal genomes.

Multiple sequence alignment

All retrieved sequences were aligned using computer programs and criteria described in Materials and Methods. The resulting multiple sequence alignment (Fig. 1) was analyzed from the point of conservation of sequence patterns specific for N4mC and their closest relatives. Pairwise comparison of most N4mC MTase sequences indicated a moderate degree of sequence similarity restricted mainly to nine motifs composed of groups of conserved residues (8). However, only two residues are invariant, found not surprisingly in the two most conserved

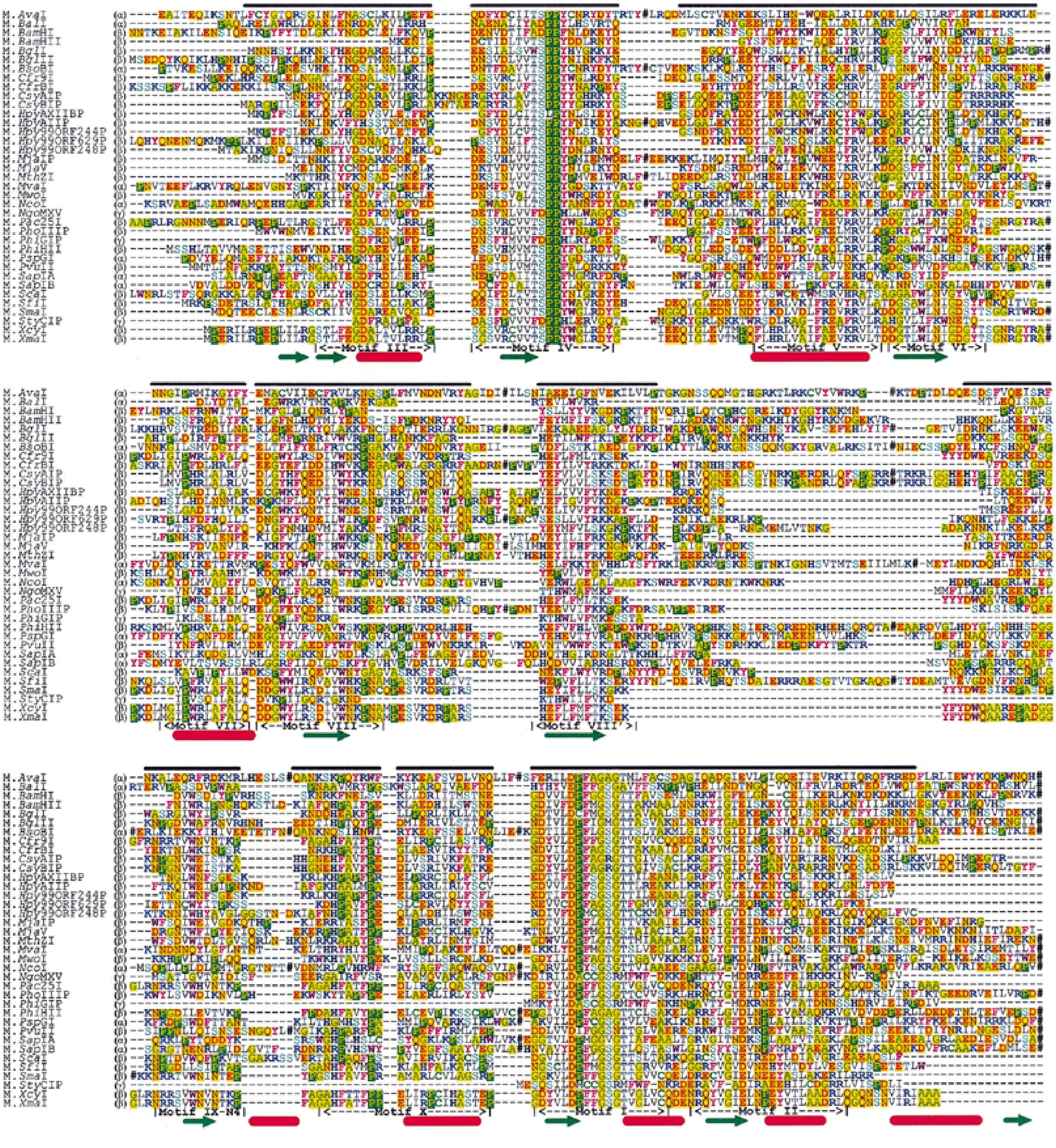

Figure 1. Multiple sequence alignment of 37 members of the N4mC MTase family classified as ‘α’, ‘β’ or ‘γ-like’ (Materials and Methods). The order is as in Table 1. # indicates the site of deletion in the loop regions or the topological breakpoint introduced into the alignment. The secondary structure of M.*Pvu*II (15) is shown at the bottom. Conserved motifs are outlined. Sequence blocks used for phylogenetic calculations are delineated using black bars above the alignment.

motifs: second proline in the core of the motif IV, ‘SPPY’ hallmark of the N4mC MTase active site (30) and the middle glycine in ‘FxGxG’ motif I—more generally conserved in all AdoMet-dependent MTases (31). This is due to the relatively large number of protein sequences used in the alignment and the inclusion of atypical (i.e. other than ‘SPPY’-bearing) and hypothetical proteins in initial calculations of the consensus sequence. The difficulty in obtaining unambiguous alignment of several regions, including for example the segment of M.*Pvu*II, for which structure could not be solved, suggests either the presence of structural or functional features unique to each protein (such as specific sequence recognition determinants)

Table 1. The 37 known and potential N4mC MTases analyzed in this study

Name	Target sequence	Host (strain)	Growth conditions	Accession #	Citation
M.*Ava*I	CYCGRG	*Anabaena variabilis* ATCC 27892	26°C	X98339	(42)
M.*Bal*I[*1]	TGGCCA	*Brevibacterium albidum*	30°C	D82028	(43)
M.*Bam*HI	GGAT*CC	*Bacillus amyloliquefaciens* H	37°C	X55285	(44)
M.*Bam*HII	GGAT*CC	*Bacillus amyloliquefaciens* H prophage H2	37°C	X53032	(45)
M.*Bgl*I	GCCN$_5$GGC	*Bacillus globigii*	30°C	AF050216	(46)
M.*Bgl*II	AGAT*CT	*Bacillus globigii* plasmid pTsp45s	30°C	U49842	(47)
M.*Bso*BI	CYCGRG	*Bacillus stearothermophilus* JN2091	55°C	X98287	(42)
M.*Cfr*9I	C*CCGGG	*Citrobacter freundii* RFL9	37°C	X17022	(30)
M.*Cfr*BI	CCWWGG	*Citrobacter freundii* 4111 plasmid pZE8	37°C	X57945	(48)
M.*Csy*AIP	ND	*Cenarchaeum symbiosum* strain A	10°C	AF083071	(49)
M.*Csy*BIP	ND	*Cenarchaeum symbiosum* strain B	10°C	AF083072	(49)
M.*Hpy*AXIIBP[*2]	ND	*Helicobacter pylori* 26695	37°C	AE000545	(50)
M.*Hpy*AIIP	ND	*Helicobacter pylori* 26695	37°C	AE000637	(50)
M.*Hpy*99ORF244P	ND	*Helicobacter pylori* J99	37°C	AE001462	(51)
M.*Hpy*99ORF629P	(CCWWGG)	*Helicobacter pylori* J99	37°C	AE001495	(51)
M.*Hpy*99ORF248P	(GGATCC)	*Helicobacter pylori* J99	37°C	AE001439	(51)
M.*Mja*IP	*CTAG	*Methanococcus jannaschii*	85°C, 250 atm	U67541	(52)
M.*Mja*V	GTA*C	*Methanococcus jannaschii*	85°C, 250 atm	U67590	(52)
M.*Mth*ZI	*CTAG	*Methanobacterium thermoformicum* Z-250	55°C	X67212	(53)
M.*Mva*I	C*CWGG	*Micrococcus varians* RFL19	26°C	X16985	(54)
M.*Mwo*I	GCN$_7$GC	*Methanobacterium wolfei*	60°C	AF051376	(55)
M.*Nco*I	CCATGG	*Nocardia corallina*	26°C	AF068761	(56)
M.*Ngo*MXV	GC*CHR	*Neisseria gonorrhoeae* MS11	37°C	AJ004687	(5)
M.*Pac*25I	CCCGGG	*Pseudomonas alcaligenes* NCIB 9867	32°C	U88088	(57)
M.*Pho*IIIP	(*CTAG)	*Pyrococcus horikoshii* OT3	98°C	AP000002	(58)
M.*Phi*GIP	(GC*CHR)	*Lactobacillus* phage phi g1e	37°C	X98106	(59)
M.*Phi*HII	ND	*Halobacterium salinarum* phage phi-H	37°C, 3M salt	X80164	(60)
M.*Psp*GI	CCWGG	*Pyrococcus* sp. GI-H	95°C	AF067805	(61)
M.*Pvu*II	CAG*CTG	*Proteus vulgaris*	37°C	X13778	(62)
M.*Sap*IA	GCTCTTC	*Saccharopolyspora* sp.	30°C	AF045021	(63)
M.*Sap*IB	GCTCTTC	*Saccharopolyspora* sp.	30°C	AF045021	(63)
M.*Sca*I	AGTACT	*Streptomyces caespitosus*	26°C	AF044681	(63)
M.*Sfi*I	GGCCN$_5$GGCC	*Streptomyces fimbriatus*	26°C	AF039750	(64)
M.*Sma*I	C*CCGGG	*Serratia marcescens* Sb	26°C	X16458	(65)
M.*Sty*CIP	(GC*CHR)	*Salmonella typhi* CT18	37°C	not assigned	[*3]
M.*Xcy*I	C*CCGGG	*Xanthomonas cyanopsidis* 13D5	26°C	M98768	(66)
M.*Xma*I	C*CCGGG	*Xanthomonas malvacearum*	26°C	AF051091	(67)

The data are presented according to the REBASE catalog (6) or taken from the corresponding references. Putative target DNA sequences inferred from the phylogenetic relationships (Fig. 1) are shown in parentheses. ND, not determined; P, indicates putative proteins; *C, methylated cytosine, [*1], originally predicted as 5mC MTase; [*2], originally predicted as N6mA MTase; [*3], the sequence data produced by the Pathogen Sequencing Group at the Sanger Centre can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/st/

or some degree of structural plasticity and lack of amino acid sequence constraints in these regions (15; our unpublished data).

Our results comparing N4mC MTases presented here in the form of the multiple sequence alignment and phylogenetic trees are more complete than previous studies, as they are based on all 37 sequences available to date. In addition, recent crystallographic results for M.*Dpn*M (14) and M.*Pvu*II (15) showed that several sequence motifs and local supersecondary structure predictions assigned by Malone *et al.* (8) as common features of all DNA N-MTases were in fact inconsistent between analyzed subfamilies. Therefore in our analysis, we attempted to rationalize the classification of conserved motifs of N4mC MTases based on similarities to the motifs of other classes of DNA MTases in respect to the common super-secondary structural and functional elements.

The assignment of conserved motifs I–VIII and X in our final alignment differs slightly from the widely cited results of Malone *et al.* (8), especially in respect to the position of weakly conserved motif III, but is essentially identical to the structure-based alignment presented by Gong *et al.* (15) (Fig. 1). Many residues are conserved throughout the sequence, most of them forming common structural features: both Rossmann-fold-like core (16,32) and several conserved loops with catalytic or ligand-binding functions, as inferred from M.*Pvu*II structure (15). We suggest a modification in nomenclature regarding motif VIII, building an antiparallel β-hairpin localized at the 'edge' of the common core of AdoMet-dependent MTases (33). In all structurally characterized DNA and RNA MTases this region forms a part of a target nucleotide binding pocket (16); however, the length of the loop between antiparallel β-strands may dramatically vary even between proteins belonging to the same class (Fig. 1). For that reason different locations of motif VIII were proposed, in either one of the β-strands (14,15) or the intervening loop (8,16). This discrepancy is clearly caused by the inability to bridge two conserved patches, for convenience we suggest referring to the C-terminal part of motif VIII as to the submotif VIII', so that parts VIII and VIII' would correspond to either of β-strands, respectively (Fig. 1).

We have also localized a previously overlooked, weakly conserved sequence patch present in most of N4mC MTase sequences. This patch (N/Q/D-V/I-W-N/E/D-I/V) can be found after motif VIII, between the variable region and motif X. In M.*Pvu*II MTase this region precedes helix F, postulated as a DNA-binding element similar to 5mC MTases and not present in N6mA MTases (15,34). However, there is no significant sequence similarity between motif IX in 5mC MTases and the newly described region in N4mC MTases, and to avoid confusion we labeled it as motif IX-N4. In M.*Pvu*II this region forms a short β-strand, while in M.*Hha*I it forms a loop and an α-helix, moreover, the presented alignment suggests that the helix F of M.*Pvu*II might be not conserved among many of N4mC MTases (Fig. 1; J.M.Bujnicki, unpublished structure predictions). Due to low sequence conservation in this region (e.g. M.*Pvu*II is lacking central Trp residue) structure prediction is ambiguous and would certainly benefit from further experimental investigation.

Phylogenetic trees

In phylogenetic inference there are two computational steps: estimation of the topology (branching pattern of a tree) and estimation of branch lengths for that topology (35). While the statistical estimation of branch lengths is relatively simple for a known topology, the number of possible topologies for a sizeable number of sequences is enormously large (for 37 sequences the number of bifurcating unrooted trees is in the order of 10^{49} given by $N = (2n - 5)!/[2^{n-3}(n - 3)!]$) (26). We therefore had to resort to a heuristic search to estimate a good tree. In the absence of a priori knowledge, the ultimate criterion for determining phylogenetic reliability rests on tests of congruence among results of different algorithms, which enable detection and minimization of systematic errors caused by the partially false assumptions of the implemented methods. Because the issue of phylogenetic reconstruction is controversial, with some disagreement coming even from personal preference or philosophy of researches in the field, we decided to use methods, which rely on substantially different assumptions about the molecular evolutionary process and have different limitations.

Phylogenetic trees of the N4mC MTases were inferred from the alignment using distance, maximum likelihood and parsimony methods (Materials and Methods). The distance method is based on a probabilistic model of amino acid transitions, which does not take explicit account of the genetic code or differences in preferred directions of substitutions of residues from different secondary structures. Its performance depends on the linear relationship with the number of substitutions and the standard error of the estimate of the distance measure. The maximum parsimony procedure is the only one that can easily take care of insertions and deletions, which may carry important phylogenetic information, but when the rate of multiple substitutions per site in the alignment is relatively high, it can be expected to converge onto the wrong tree. Under assumption that all amino acid residues diverge at the same expected rate the maximum likelihood estimation yields quite robust trees, but is computationally most expensive and to reduce the number of calculations of the maximum likelihood values for all alternative trees heuristics leading to relatively greatest simplifications are necessary. All phylogenetic algorithms that we used assume correct alignment of positional homologs. For this reason, areas of questionable alignment including regions with gaps in >50% of sequences have been omitted from consideration prior to the process of tree inference (Fig. 1). Such regions, where the sequences appear randomized with respect to evolutionary history are evolving at rates too high for effective phylogenetic analysis (26). Therefore restriction of our analysis to regions that are likely to have the highest signal-to-noise ratio seems justified.

Due to the possibility of processes such as domain swapping and recombination with genes coding for MTases other than N4mC-specific (which would generate a hybrid with mosaic similarity to N4mC and other MTase subfamilies), we inferred and compared the evolutionary trees based solely on regions forming the catalytic (motifs III–VIII) and cofactor-binding (motifs X, I and II) subdomains. For each method, the topologies of both trees were nearly identical and the separate alignment of sequences from classes α and β also gave similar distribution of branches in corresponding subtrees (data not shown). This congruence strongly suggests that both subdomains coevolved and that the recombination events leading to the permutation of the catalytic and AdoMet binding regions in the N4mC MTase family did not involve 'domain stealing' (36) from any other family of MTases. This justifies the approach of artificial unification of the order of conserved motifs in sequences from different classes to base the phylogenetic inference on one alignment (Fig. 1).

The subtopologies of most branches of the evolutionary trees obtained by maximum parsimony, maximum likelihood and distance criteria are nearly identical (Fig. 2). These topologies are reliable by the criterion of bootstrap and even the removal of putative MTases does not significantly alter the relationship between other lineages (data not shown). This suggests that the markedly different assumptions used by the three algorithms were in agreement with the nature of evolutionary processes governing the divergence of N4mC MTases and small differences most likely come from unequal efficiency of the algorithms in the exploration of the huge space of possible results (see above).

A clear correlation exists between the distribution of the proteins into clusters and the nature of the recognized sequence. All of MTases recognizing the same target form individual branches with the subtopology unchanged between trees and strongly supported by bootstrap values, suggesting that they recently diverged from a common ancestor.

Homologs of M.*Sma*I and M.*Ngo*MXV form coherent clades with bootstrap values close to 100 in all trees and with low estimated branch lengths (Fig. 1). The group of α-MTases bearing the 'SPPY' version of motif IV (M.*Mva*I homologs) and three MTases from the thermophilic Archaea (M.*Mth*ZI, M.*Pho*IIIP and M.*Mja*I) also form separate clades in all trees, but with branches rather longer with respect to the common stem. MTases from *Helicobacter pylori*, M.*Hpy*AXIIBP and M.*Hpy*99ORF244P (certainly a pair of orthologs) and M.*Hpy*AIIP usually group together, but the subtopology is not congruent between trees. Both 'DPPY' MTases, namely M.*Bam*HI and M.*Bal*I, despite their different motif permutations (typical for α and β classes, respectively) are usually found together, branched out at the central part of the tree. The sequence database searching using sequences of these proteins as queries resulted in almost exclusively N6mA MTases and putative proteins assigned to this family based on sequence similarity (Table 2). These results taken together led us to the conclusion that both M.*Bam*HI and M.*Bal*I MTases diverged relatively recently from N6mA MTases.

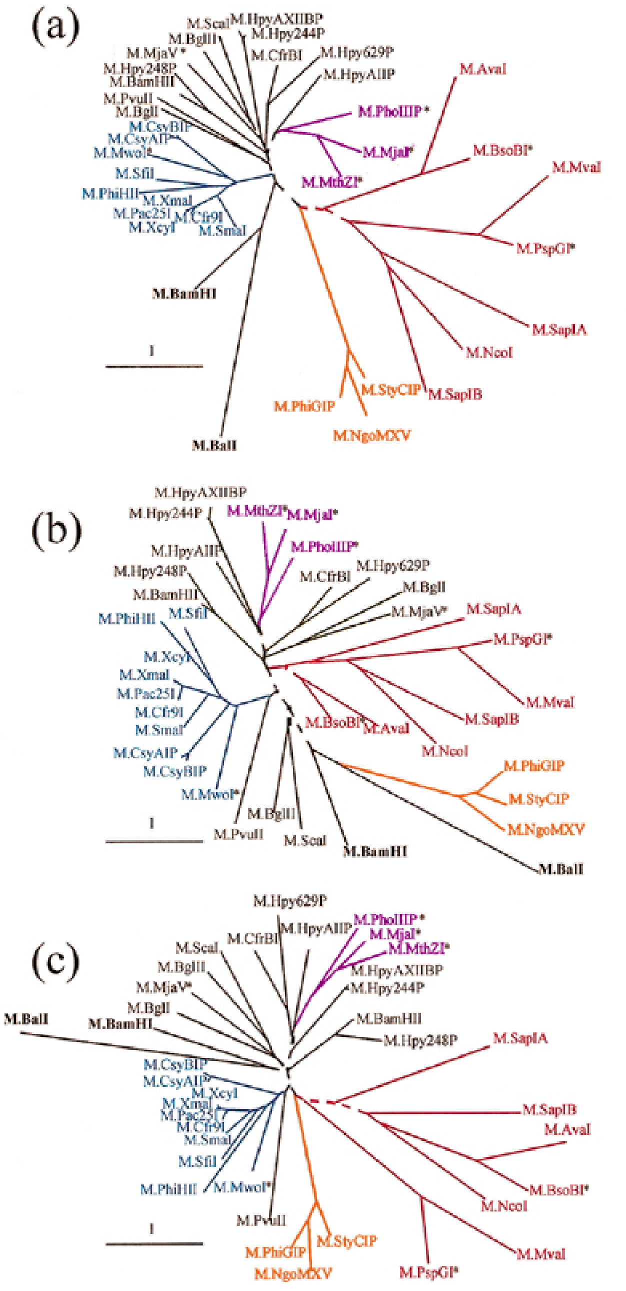

Figure 2. Dendrograms representing the relationship between N4mC MTases inferred using (**a**) distance method, (**b**) parsimony and (**c**) maximum likelihood approaches. Conserved subfamilies are shown in color, thermophilic enzymes are indicated by asterisks. For clarity of the presentation M.*Hpy*99ORF244P, M.*Hpy*99ORF629P and M.*Hpy*99ORF248P have been labeled as M.*Hpy*244P, M.*Hpy*629P and M.*Hpy*248P, respectively. Branches with bootstrap values below 50% are shown as broken lines. The bars at the bottom of each phylogeny are scaled to an amino acid replacement distance of 1 (corrected for multiple substitutions).

The evolutionary relationships among subfamilies are less strongly resolved than those within the subfamilies, bootstrap values for the nodes that define the deep branching pattern are low, indicating that changes in the sampling of alignment position used to generate the trees affect the inferred relationships among subfamilies, especially in the parsimony-based tree. This can be explained as indication of their simultaneous differentiation from the common ancestor. The mutual position of M.*Ngo*MXV, M.*Mva*I and M.*Sma*I clades is ambiguous, similarly the position of several single MTases, e.g. M.*Pvu*II. More accurate determination of their relationship should be possible after identification of further members of each subfamily; however, we believe that most of the overall topology of the relationships among subfamilies will not change significantly from that presented here. In our opinion, it would be most reasonable to root the main branches based on more unequivocal data, e.g. comparison of atomic coordinates, if they were available for more N4mC MTases than only M.*Pvu*II.

Table 2. Proteins homologous to M.*Bal*I and M.*Bam*HI MTases

A

Protein name	Organism	Acc. #	Score (bits)	E-value
Hypothetical protein Upf31.0	*Enterobacter aerogenes*	U67194	62	4.00E-09
Hypothetical N6mA MTase	*Pyrococcus horikoshii*	AP000004	54	1.00E-06
N6mA MTase M.*Lla*DCHIA	*Lactococcus lactis*	U16027	47	1.00E-04
Hypothetical protein sll0729	*Synechocystis sp.* PCC6803	D90917	46	2.00E-04
N6mA MTase M.*Mbo*IA	*Moraxella bovis*	D13968	46	4.00E-04
Hypothetical MTase (truncated)	*Rhodobacter capsulatus*	AF010496	45	6.00E-04
N6mA MTase Dam	*Serratia marcescens*	X78412	45	6.00E-04
N6mA MTase M.*Pgi*I	*Porphyromonas gingivalis*	M63469	43	0.002
N6mA MTase M.*Dpn*II	*Streptococcus pneumoniae*	M11226	42	0.005
N6mA MTase Dam	*Escherichia coli*	U18997	41	0.009

B

Protein name	Organism	Acc. #	Score (bits)	E-value
N6mA MTase M.*Hpa*I	*Haemophilus parainfluenzae*	D10668	131	7.00E-30
N6mA MTase M.*Mbo*II	*Moraxella bovis*	X56977	115	8.00E-25
Hypothetical N6mA MTase yhdJ	*Escherichia coli*	U00096	110	2.00E-23
Hypothetical N6mA MTase HP1367	*Helicobacter pylori*	AE000637	99	4.00E-20
N6mA MTase M.*Hin*fI	*Haemophilus influenzae* Rf	M22862	98	1.00E-19
N4mC MTase M.*Bgl*II	*Bacillus subtilis*	U49842	96	4.00E-19
N6mA MTase M.*Xba*I	*Xanthomonas campestris*	AF051092	90	2.00E-17
N6mA MTase	*Brucella abortus*	AF011895	89	6.00E-17
N6mA MTase	*Sinorhizobium meliloti*	AF011894	87	3.00E-16
N6mA MTase M.*Lla*DCHIB	*Lactococcus lactis*	U16027	86	5.00E-16

(**A**) Results of protein sequence database screening with PSI-BLAST using M.*Bal*I as a query. (**B**) Results of protein sequence database screening with PSI-BLAST using M.*Bam*HI as a query.

DISCUSSION

Protein families are often categorized as the result of the possession of conserved motifs. The N4mC MTases share nine weakly conserved regions with N6mC MTases and to some degree with other AdoMet-dependent MTases. The amino acid sequences of N4mC MTases exhibit great divergence, including permutation of structural and functional modules within a common three-dimensional fold, a feature characteristic for all DNA MTases (15). Therefore the issue of evolutionary history and phylogenetic origin of these enzymes is not straightforward.

Traditionally, N4mC and N6mA DNA MTases have been considered to be very similar (3,8,9,30). However, the determination of three crystal structures (two for N6mA and one for N4mC) did not fully clarify their relationships, showing incompatibility of target DNA recognition determinants between classes and presenting features common also to N4mC and 5mC MTases and absent from N6mA MTases (11,13–15). The inference of evolutionary relationship among different MTases based on similarity of the three-dimensional fold of their catalytic domains directly supports the scenario, in which the bulk of N4mC and N6mA MTases diverged prior to the specialization of N6mA MTases into DNA and RNA-specific subfamilies (19). At least in certain cases equivalence at the level of target specificity and local sequence similarity between members of different classes could be explained by subsequent convergence.

The simplest assumption would be that all genes of the known N4mC MTases evolved from one or several recombining common precursor genes, similarly to 5mC MTases (37). Considering the limited number of proteins in the family and their fairly unique role, namely protection of bacterial DNA against digestion by the restriction endonuclease (ENase) from its ‘own’ R–M system, but also non-cognate ENases (38), one might expect a high level of sequence conservation. However, in striking contrast to the 5mC MTases, the N4mC MTase family encompasses extensive diversity. The phylogenetic trees inferred with different methods suggest that the N4mC family underwent a radical restructuring, leading to inversion of the linear order of two main subdomains and establishing two major highly diverged branches: α and β. In addition to that, all data support the relationship of M.*Bam*HI and M.*Bal*I not with other N4mC, but rather with N6mA MTases, suggesting a polyphyletic origin of the N4mC MTase subfamily (Table 2, Fig. 3). Recently it has been speculated that M.*Ngo*MXV (and presumably its homologs) might be related to the common ancestor of both N6mA and N4mC MTases, as it shows comparable degree of similarity to representatives of both N-MTase subfamilies (10). Modeling of M.*Ngo*MXV, which exhibits relaxed sequence specificity, indicated its single-domain structure and lack of extended loops (10), which are properties usually assigned to the ancestors of modern enzyme families (39). The ‘ancient’ character of M.*Ngo*MXV would be consistent with the hypothesis that the most highly specific MTases evolved later in the history of this family by acquiring additional target-recognizing determinants (40).

In Figure 3 we propose a general model of polyphyletic evolution of the N4mC MTase family, in which after separation of two main lineages a few widely diverged enzymes narrowed or switched their preference for a methylated base to N4mC-specificity.

Recently, Jeltsch *et al.* (18) demonstrated that certain N6mA MTases are able to methylate mismatched cytosines in artificial substrates. The authors argued that this result supports the hypothesis of independent origin of α and β subfamilies of N4mC MTases from α and β N6mA MTases, respectively, considered by Malone *et al.* (8). Jeltsch *et al.* (18) suggested that the permutation events must have been so rare that simultaneous use of the same ‘topological switchpoint’ in two families should be considered improbable. However, they did not support this conjecture by any of the established methods of phylogenetic inference, and their biochemical data might equally support our model of late convergence and specificity switching between the N4mC and N6mA MTase families. Whereas the significantly higher degree of overall sequence similarity between α and β N4mC MTases than between N4mC and N6mA MTases within α or β groups (our unpublished data) clearly argues for independence of permutation events in the N4mC and N6mA lineages. Even if N4mC and N6mA MTase families were in fact more closely related to one another than to other MTases—a hypothesis not supported by either structure- or sequence-based trees (19)—we suggest that the ancestral N4mC MTases would rather evolve from the relatively most similar β-N6mA lineage. In other words, the α-topology would independently appear among N6mA and N4mC MTases. We believe that structure solution of any β-N6mA MTase and/or α-N4mC MTase and including it in a recalculated carbon-α distance-based tree might help to resolve that controversy.

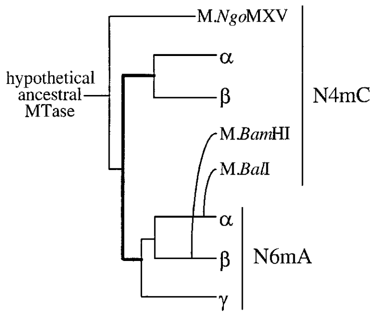

Figure 3. Proposed schematic phylogeny of N4mC MTases. The branch lengths are arbitrary and indicate only relative time of divergence of different lineages. The present data do not exclude alternative rooting, e.g. with M.*Ngo*MXV group radiated soon after the major N6mA/N4mC bifurcation.

The distribution of ‘modern specificities’ among 5mC or N6mA MTases could result from shuffling of ‘mobile’ TRD units between independently evolving catalytic domains (37,40). However, the analysis of the structure and the docking model of M.*Pvu*II MTase (15) shows that this and related N4mC MTases do not maintain potential DNA-recognizing determinants in one distinct domain (neither in amino acid sequence nor three-dimensional structure). Instead, they seem to be embedded in several loops protruding from between conserved segments of the structural scaffold common to all AdoMet-dependent MTases. Therefore, contrary to 5mC and probably also N6mA MTases, which presumably gained target specificity primarily through fusion with distinct domains, modern specificities of most of N4mC MTases bearing ‘DPPH’ and ‘SPPY’ versions of motif IV may have arisen by extension of flexible loops accommodating substrate nucleic acids in a V-shaped cleft (10,15).

Our results indicate that particular specificities evolved only once in the evolutionary history of N4mC MTases. Proteins with similar target recognition properties usually display significant sequence homology and form coherent branches of the evolutionary tree, suggesting that they derive from a common ancestor. The sole exception is a pair of M.*Bam*HI and M.*Bam*HII MTases, extremely diverged at the amino acid sequence level, but recognizing identical target DNA sequence (Fig. 1, Table 1). Docking the substrate DNA onto the three-dimensional models of these two MTases suggests that the possible determinants of sequence specificity are located within dissimilar secondary structural elements, further supporting the case of functional convergence or at least ‘domain shuffling’ (our unpublished data). If sequence specificity is conserved to some degree within subfamilies, then the specificity of some of the uncharacterized proteins in the N4mC family can be predicted by comparison with other members of the same subfamily (Table 1).

The presence of N4mC MTases both in Bacteria and Archaea indicates that these enzymes had their origins in the common ancestor of these kingdoms or that one of them acquired the N4mC MTase gene(s) from another by horizontal transfer of genetic material. It also suggests that there is something specific that prevented or at least did not support the diversification of this protein family in the higher organisms. It is hypothesized that the last common ancestor of all cellular organisms was a hyperthermophilic prokaryote (41). However, even if the ancient N4mC MTase was present in the thermophilic cenancestor, many of thermophilic enzymes are more related to their mesophilic homologs, than to each other (e.g. M.*Psp*GI to M.*Mva*I and M.*Mja*V to M.*Bgl*I, see Fig. 2), indicating that hyperthermophilicity or hyperthermostability of N4mC MTases evolved relatively late and independently from various mesophilic lineages. Also the N4mC MTases from psychrophilic (M.*Csy*AIP and M.*Csy*BIP) or halophilic (M.*Phi*HII) Archaea seem to originate from a mesophile (the bulk of MTases in the 'blue' clade in Fig. 2), suggesting multiple events of horizontal transfer of genetic material from already diverged Bacteria.

We are aware that the accuracy of our analysis depends on the assumption that sequences and functional annotations of putative proteins are correct. However, we hope that it will stimulate and help to advance the experimental verification of presented premises. As additional complete genome sequences become available, especially from eukaryotic and archaeal genome projects, and relation of the phylogeny of the N4mC MTase family to the organismal phylogeny becomes less obscure it shall be possible to answer the question of whether Eukaryota lost the N4mC activity during evolution or it evolved in Bacteria and/or Archaea after the establishment of the main branches of the tree of life. We believe that combining genomics, molecular phylogeny and comparative biochemistry of DNA MTases will help to solve the problem of the last universal common ancestor, whilst highlighting the pitfalls of horizontal transfer and molecular convergence between paralogs, which when unnoticed may obscure the organismal phylogeny inferred from sequence data.

ACKNOWLEDGEMENTS

We wish to thank Dr David T. F. Dryden for critical reading of the manuscript, provision of unpublished materials and useful discussions and two anonymous referees for helpful suggestions. We also thank all genome-sequencing groups, especially the Pathogen Sequencing Group at the Sanger Centre, for making their preliminary data publicly available.

REFERENCES

1. Adams,R.L. (1990) *Biochem. J.*, **265**, 309–320.
2. Noyer-Weidner,M. and Trautner,T.A. (1993) *EXS*, **64**, 39–108.
3. Timinskas,A., Butkus,V. and Janulaitis,A. (1995) *Gene*, **157**, 3–11.
4. Hattman,S., Kenny,C., Berger,L. and Pratt,K. (1978) *J. Bacteriol.*, **135**, 1156–1157.
5. Radlinska,M. and Piekarowicz,A. (1998) *Biol. Chem.*, **379**, 1391–1395.
6. Roberts,R.J. and Macelis,D. (1999) *Nucleic Acids Res.*, **27**, 312–313.
7. Kumar,S., Cheng,X., Klimasauskas,S., Mi,S., Posfai,J., Roberts,R.J. and Wilson,G.G. (1994) *Nucleic Acids Res.*, **22**, 1–10.
8. Malone,T., Blumenthal,R.M. and Cheng,X. (1995) *J. Mol. Biol.*, **253**, 618–632.
9. Wilson,G.G. (1992) *Methods Enzymol.*, **216**, 259–279.
10. Radlinska,M., Bujnicki,J.M. and Piekarowicz,A. (1999) *Proteins*, in press.
11. Cheng,X., Kumar,S., Posfai,J., Pflugrath,J.W. and Roberts,R.J. (1993) *Cell*, **74**, 299–307.
12. Reinisch,K.M., Chen,L., Verdine,G.L. and Lipscomb,W.N. (1995) *Cell*, **82**, 143–153.
13. Labahn,J., Granzin,J., Schluckebier,G., Robinson,D.P., Jack,W.E., Schildkraut,I. and Saenger,W. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 10957–10961.
14. Tran,P.H., Korszun,Z.R., Cerritelli,S., Springhorn,S.S. and Lacks,S.A. (1998) *Structure*, **6**, 1563–1575.
15. Gong,W., O'Gara,M., Blumenthal,R.M. and Cheng,X. (1997) *Nucleic Acids Res.*, **25**, 2702–2715.
16. Schluckebier,G., Labahn,J., Granzin,J. and Saenger,W. (1998) *Biol. Chem.*, **379**, 389–400.
17. Wilson,G.G. and Murray,N.E. (1991) *Annu. Rev. Genet.*, **25**, 585–627.
18. Jeltsch,A., Christ,F., Fatemi,M. and Roth,M. (1999) *J. Biol. Chem.*, **274**, 19538–19544.
19. Bujnicki,J.M. (1999) *In Silico Biol.*, **1**, http://www.bioinfo.de/isb/1999–01/0016/ .
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
21. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
22. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) *Bioinformatics*, **14**, 892–893.
23. Felsenstein,J. (1988) *Annu. Rev. Genet.*, **22**, 521–565.
24. Fitch,W.M. and Margoliash,E. (1967) *Science*, **155**, 279–284.
25. Hillis,D.M., Allard,M.W. and Miyamoto,M.M. (1993) *Methods Enzymol.*, **224**, 456–487.
26. Li,W.-H. (1996) *Molecular Evolution.* Sinauer Associates, Sunderland, MA.
27. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Comput. Appl. Biosci.*, **8**, 275–282.
28. Efron,B. and Tibshirani,R. (1991) *Science*, **253**, 390–395.
29. Brenner,S.E. (1999) *Trends Genet.*, **15**, 132–133.
30. Klimasauskas,S., Timinskas,A., Menkevicius,S., Butkiene,D., Butkus,V. and Janulaitis,A. (1989) *Nucleic Acids Res.*, **17**, 9823–9832.
31. Kagan,R.M. and Clarke,S. (1994) *Arch. Biochem. Biophys.*, **310**, 417–427.
32. Rossmann,M.G., Moras,D. and Olsen,K.W. (1974) *Nature*, **250**, 194–199.
33. Schluckebier,G., O'Gara,M., Saenger,W. and Cheng,X. (1995) *J. Mol. Biol.*, **247**, 16–20.
34. Klimasauskas,S., Kumar,S., Roberts,R.J. and Cheng,X. (1994) *Cell*, **76**, 357–369.
35. Nei,M. (1996) *Annu. Rev. Genet.*, **30**, 371–403.
36. Doolittle,R.F. (1995) *Annu. Rev. Biochem.*, **64**, 287–314.
37. Bujnicki,J. and Radlinska,M. (1999) *Acta Microbiol. Pol.*, **48**, 19–33.
38. Piekarowicz,A. and Radlinska,M. (1998) *Acta Microbiol. Pol.*, **47**, 405–407.
39. Murzin,A.G. (1998) *Curr. Opin. Struct. Biol.*, **8**, 380–387.
40. Lauster,R. (1989) *J. Mol. Biol.*, **206**, 313–321.
41. Forterre,P. (1995) *C.R. Acad. Sci. III*, **318**, 415–422.
42. Ruan,H., Lunnen,K.D., Scott,M.E., Moran,L.S., Slatko,B.E., Pelletier,J.J., Hess,E.J., Benner,J., Wilson,G.G. and Xu,S.Y. (1996) *Mol. Gen. Genet.*, **252**, 695–699.
43. Ueno,H., Kato,I. and Ishino,Y. (1996) *Nucleic Acids Res.*, **24**, 2268–2270.
44. Brooks,J.E., Benner,J.S., Heiter,D.F., Silber,K.R., Sznyter,L.A., Jager-Quinton,T., Moran,L.S., Slatko,B.E., Wilson,G.G. and Nwankwo,D.O. (1989) *Nucleic Acids Res.*, **17**, 979–997.
45. Vanek,P.G., Connaughton,J.F., Kaloss,W.D. and Chirikjian,J.G. (1990) *Nucleic Acids Res.*, **18**, 6145.
46. Duncan,C.H., Wilson,G.A. and Young,F.E. (1978) *J. Bacteriol.*, **134**, 338–344.
47. Anton,B.P., Heiter,D.F., Benner,J.S., Hess,E.J., Greenough,L., Moran,L.S., Slatko,B.E. and Brooks,J.E. (1997) *Gene*, **187**, 19–27.
48. Zakharova,M.V., Kravetz,A.N., Beletzkaja,I.V., Repyk,A.V. and Solonin,A.S. (1993) *Gene*, **129**, 77–81.
49. Schleper,C., DeLong,E.F., Preston,C.M., Feldman,R.A., Wu,K.Y. and Swanson,R.V. (1998) *J. Bacteriol.*, **180**, 5003–5009.

50. Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al*. (1997) *Nature*, **388**, 539–547.
51. Alm,R.A., Ling,L.S., Moir,D.T., King,B.L., Brown,E.D., Doig,P.C., Smith,D.R., Noonan,B., Guild,B.C., deJonge,B.L. *et al*. (1999) *Nature*, **397**, 176–180.
52. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., Fitzgerald,L.M., Clayton,R.A., Gocayne,J.D. *et al.* (1996) *Science*, **273**, 1058–1073.
53. Nolling,J. and de Vos,W.M. (1992) *Nucleic Acids Res.*, **20**, 5047–5052.
54. Butkus,V., Klimasauskas,S., Kersulyte,D., Vaitkevicius,D., Lebionka,A. and Janulaitis,A. (1985) *Nucleic Acids Res.*, **13**, 5727–5746.
55. Lunnen,K.D., Morgan,R.D., Timan,C.J., Krzycki,J.A., Reeve,J.N. and Wilson,G.G. (1989) *Gene*, **77**, 11–19.
56. Van Cott,E.M. and Wilson,G.G. (1988) *Gene*, **74**, 55–59.
57. Yeo,C.C., Tham,J.M., Kwong,S.M. and Poh,C.L. (1998) *Plasmid*, **40**, 203–213.
58. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A. *et al.* (1998) *DNA Res.*, **5**, 55–76.
59. Yamada,K. and Taketo,A. (1997) *Gene*, **187**, 45–53.
60. Stolt,P., Grampp,B. and Zillig,W. (1994) *Biol. Chem.*, **375**, 747–757.
61. Morgan,R.D., Xiao,J.P. and Xu,S.Y. (1998) *Appl. Environ. Microbiol.*, **64**, 3669–3673.
62. Blumenthal,R.M., Gregory,S.A. and Cooperider,J.S. (1985) *J. Bacteriol.*, **164**, 501–509.
63. Xu,S.Y., Xiao,J.P., Ettwiller,L., Holden,M., Aliotta,J., Poh,C.L., Dalton,M., Robinson,D.P., Petronzio,T.R., Moran,L., Ganatra,M., Ware,J., Slatko,B.E. and Benner,J. (1998) *Mol. Gen. Genet.*, **260**, 226–231.
64. Van Cott,E.M., Moran,L.S., Slatko,B.E. and Wilson,G.G. (1997) GenBank accession no. AF0039750
65. Heidmann,S., Seifert,W., Kessler,C. and Domdey,H. (1989) *Nucleic Acids Res.*, **17**, 9783–9796.
66. Withers,B.E., Ambroso,L.A. and Dunbar,J.C. (1992) *Nucleic Acids Res.*, **20**, 6267–6273.
67. Butkus,V., Petrauskiene,L., Maneliene,Z., Klimasauskas,S., Laucys,V. and Janulaitis,A. (1987) *Nucleic Acids Res.*, **15**, 7091–7102.