CALL FOR PAPERS | Comparative Genomics

Genome-wide analysis of restriction-modification system in unicellular and filamentous cyanobacteria

Fangqing Zhao,1 Xiaowen Zhang,1 Chengwei Liang,1 Jinyu Wu,2 Qiyu Bao,2 and Song Qin1

1Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China; and
2Institute of Biomedical Informatics, Wenzhou Medical College, Wenzhou, China

Submitted 17 October 2005; accepted in final form 13 December 2005

Zhao, Fangqing, Xiaowen Zhang, Chengwei Liang, Jinyu Wu, Qiyu Bao, and Song Qin. Genome-wide analysis of restriction-modification system in unicellular and filamentous cyanobacteria. Physiol Genomics 24: 181–190, 2006. First published December 20, 2005; doi:10.1152/physiolgenomics.00255.2005.—Cyanobacteria are an ancient group of gram-negative bacteria with strong genome size variation ranging from 1.6 to 9.1 Mb. Here, we first retrieved all the putative restriction-modification (RM) genes in the draft genome of Spirulina and then performed a range of comparative and bioinformatic analyses on RM genes from unicellular and filamentous cyanobacterial genomes. We have identified 6 gene clusters containing putative Type I RMs and 11 putative Type II RMs or the solitary methyltransferases (MTases). RT-PCR analysis reveals that 6 of 18 MTases are not expressed in Spirulina, whereas one hsdM gene, with a mutated cognate hsdS, was detected to be expressed. Our results indicate that the number of RM genes in filamentous cyanobacteria is significantly higher than in unicellular species, and this expansion of RM systems in filamentous cyanobacteria may be related to their wide range of ecological tolerance. Furthermore, a coevolutionary pattern is found between hsdM and hsdR, with a large number of site pairs positively or negatively correlated, indicating the functional importance of these pairing interactions between their tertiary structures. No evidence for positive selection is found for the majority of RMs, e.g., hsdM, hsdS, and hsdR, and Type II restriction endonuclease gene families, while a group of MTases exhibit a remarkable signature of adaptive evolution. Sites and genes identified here to have been under positive selection would provide targets for further research on their structural and functional evaluations.

coevolution; comparative genomics; horizontal gene transfer; molecular evolution; Spirulina platensis

CYANOBACTERIA ARE AMONG the oldest life on Earth with the capacity of oxygenic photosynthesis similar to the process found in higher plants. Cyanobacteria may be unicellular or filamentous, and they can be found in many different environments ranging from marine to freshwater habitats and also in soil, on rocks, and on plants. Spirulina, a nonheterocystous, filamentous cyanobacterium, is proposed as an important mariculture crop for both food and food supplement and as a potential chemical factory for a variety of useful products if effective genetic engineering can be accomplished. Unlike other unicellular (e.g., Synechococcus and Synechocystis) and filamentous cyanobacteria (e.g., Anabaena), genetic transformation of Spirulina has had limited success to date (18, 52). The predominant barrier appears to be the presence of restriction-modification (RM) systems in Spirulina platensis, which may decrease or prevent the uptake of exogenous DNA (53). RM systems comprise a restriction endonuclease (REase) activity that specifically cleaves DNA and a corresponding methyltransferase (MTase) activity that specifically methylates the host DNA, thereby protecting it from cleavage (4, 55). RM systems can be subdivided into four groups (I, II, III, and IV) based on the subunit composition, cofactor requirements, sequence specificity, and cleavage position (40, 55). Moreover, those MTases not associated with REases are termed solitary MTases. RM systems in the host genome have been viewed as a defense mechanism against DNA invasion or as the parasitic selfish genetic elements (22, 42). The selfish gene hypothesis is now well supported by much evidence from genome analysis and experimentation (23). The traditional method used to screen for REases is to culture individual strains and test their extracts for the ability to produce specific fragments from small DNA molecules (55). A large number of both filamentous and unicellular cyanobacteria have been shown to contain one or more Type II site-specific REases (29). Lyra et al. (29) employed principal component analysis to demonstrate the relationship between certain REases and their hosts and found that all 28 cyanobacterial strains studied contain Type II REases, with the most common cyanobacterial REase types being AvalI, Aval, and AsuII. Matveev et al. (31) used experimental and bioinformatics approaches to give a full insight on the MTases of the cyanobacterium Anabaena sp. PCC7120, which has four Type II MTases associated with REases (Aval–IV) and five solitary MTases. Comparative analysis of cyanobacterial MTases revealed that the solitary ones appear to be of ancient origin in cyanobacteria, while the Type II MTases appear to have arrived by horizontal gene transfer (31). Now 15 cyanobacterial genomes have been fully sequenced, 3 in the draft format, and >20 are in the process of being sequenced (http://img.jgi.doe.gov/pub/main.cgi?page=restrictedMicrobes&domain=Bacteria; http://www.ncbi.nlm.nih.gov/). These genome-sequencing projects undoubtedy bring a great convenience to the search for novel RM genes using bioinformatic tools. We have generated >7-Mbp genomic sequences with 8× coverage of redundancy from the shotgun genome-sequencing project of S. platensis (Spirulina Genome Sequencing Consortium, unpublished data). Genome-wide identification of RM systems in Spirulina will help to
develop an efficient gene transfer system in this organism. Here, we reported all the putative RM superfamily from the draft genome of *S. platensis* and presented a comparative genomic analysis of RM genes in the unicellular and filamentous cyanobacteria.

**MATERIALS AND METHODS**

Computational search for novel RM genes. The cyanobacteria species examined included *Synechocystis*, *Synechococcus*, *Prochlorococcus*, *Anabaena*, *Nostoc*, *Spirulina*, and *Trichodesmium*. An initial set of RM genes was obtained from CyanoBase (http://www.kazusa.or.jp/cyano/cyano.html) and Integrated Microbial Genomes (IMG) v.1.1 (http://img.jgi.doe.gov/v1.1/main.cgi). This data set, including well-characterized and putative cyanobacterial RMs, was used to construct a query protein set. Each protein in this query data set was used to search the potential novel sequences in all cyanobacteria species with whole genome sequences available, by using the basic local alignment search tool (BLAST) and TBLASTN programs, with *e*-value < 1E-10; the searches were iterated until convergence. Hits produced by different query sequences were combined, and duplicate records, which were identified on the basis of accession, were removed. Because RM genes tend to form clusters (55), all cyanobacterial genomic clusters containing RM sequences were also retrieved and translated into six open reading frames (ORFs) and examined manually for the presence of the RM motifs to discover potential sequences with distant homology. Then profile recognition programs (45, 46), including PFAM (http://pfam.wustl.edu/) and simple modular architecture research tool (SMART; http://smart.embl-heidelberg.de/), were used to determine the presence of RM motifs, and multiple sequence alignments were used to finally classify a protein as RM homolog. Overall, six motifs (HSDR, DEXDC, Methylase-M, Methylase-S, N6_Mtase, and MethyltransfID12) were used in this study. All the putative RM genes identified in *Spirulina* genome have been submitted to the GenBank. Accession numbers are as follows: DQ239732-DQ239748.

**RT-PCR analysis of RM genes in *Spirulina***. Total RNA was extracted from fresh *Spirulina* using RNeasy Mini kit (Qiagen) by following the manufacturer’s protocol. Residual DNA in RNA preparations was eliminated by digestion with RNase-free DNase (Promega). The absence of DNA was checked by PCR. The purified total RNA was reverse transcribed into cDNA using AMV RT (Takara). The following primers used for RT-PCR were designed based on the coding region of each gene: I-hsdM: forward primer CGATGC-GACGGAGTITTAG, reverse primer GCTTTTGTTCCGACTT; II-hsdM: forward primer TACTCGGCCAGCTGTTC, reverse primer CCCTGCCAACCTTTTAT; III-hsdM: forward primer GCGTCT-CGTTTTCGTAG, reverse primer TGGCTTTGGCCATATTGC; IV-hsdM: forward primer GCCATATCGCTTACC, reverse primer GTGAGGTCTCTCGGTC; IV-mtase: forward primer GACCCCTTTGCCGAGTGC, reverse primer TGGTCTGCAATCATGC; V-mtase: forward primer ATCAAAATCCACAGAAGC; V-hsdR: forward primer GCCTCAATACGTTCG; V-hsdM: forward primer ATCCAAATCCACAGAAGC; reverse primer ATGGTG-GAGCCAAAAGAGC; VI-hsdR: forward primer TATGCGGCTCT-GGTTAT, reverse primer GCGGAGTCCTGGGTAT; MTase1: forward primer GACGTAGCAACACTT; reverse primer GCAAATAGAAAAGCCTTGG; MTase2: forward primer TTAACC-CACGAAATCATCC; reverse primer CGAGACGTGTTAAACCTCC; MTase3: forward primer GGAGAAAAACTCAACACAC; reverse primer GCCTAAAAAGGGGGGAC; MTase4: forward primer TTGGCCTTATCATCCG; reverse primer TGGCCCTGATAGTTTTC; reverse primer CCCGATCAATAC; MTase5: forward primer GCGACGAGTGTTAAACCTCC; reverse primer GCCTAAAAAGGGGGGAC; MTase6: forward primer TGGCCCTGATAGTTTTC; reverse primer CCCGATCAATAC; MTase7: forward primer AAATTATGGCCTCAACAGC, reverse primer CTTGCCACATATGCC; MTase8: forward primer CCAAAACGAATACAGCAC, reverse primer CCAGCGGACGAGAAAAAC; MTase9: forward primer TTATCGGGCGG-CAATTACC, reverse primer TGGCTCATTTTGCAATGCT; MTase10: forward primer TTGGCAGATCTCCCTTACA, reverse primer AATTA-ATTACATGTTGCG; MTase11: forward primer TAGAAATTGC-CTCCAGCC, reverse primer AGACATCCTTGGACTCC.

Multiple sequence alignment and phylogenetic analysis. Multiple sequence alignments were carried out using the Clustal W program (48) and then adjusted manually. Graphical representations of the multiple amino acid sequence alignments, the sequence logos, were presented using WebLogo (7). To maximize the number of sites available for analysis, partial sequences and certain sequences with large deletions were excluded from the analysis. The neighbor-joining (NJ) method and maximum parsimony method in MEGA3 (28) were used to construct the phylogenetic tree, and the reliability of each branch was tested by 1,000 bootstrap replications.

*Molecular evolutionary analysis*. Detecting structural interactions and statistical covariance among separate amino acid sites is of great significance for understanding protein structure and evolution (2, 38). Such analyses are based on the assumption that functionally significant coordinated residues in proteins result from physicochemical interactions (e.g., hydrophobic, electrostatic). These interactions are dependent on the physicochemical properties (e.g., volume, charge, and hydrophobicity) of the residues (50). Here, we estimated the pairwise correlation of residue substitutions among several motifs in hsdM and hsdR. This approach, based on the estimation of the linear correlation coefficients between the amino acid physicochemical property values, was fully described by Afonnikov and KolchanoV (1). Different characteristics of amino acids (volume, hydrophobicity, and polarity) were chosen for coevolution analysis that reflect physical and chemical interactions between residues. Different sequence weighting methods (14, 54) were also used in these tests to avoid sequence bias.

One of the most stringent methods to detect positive selection is a comparison of the rate of nonsynonymous substitutions (dS) to the rate of synonymous substitution (dS). The ratio between these rates (ω = dS/dD) is a reliable measure of the selective pressure acting on a protein-coding gene, with ω = 1 indicating neutral evolution, ω < 1 purifying selection, and ω > 1 positive selection (57). Likelihood ratio tests (LRTs) were performed using Codeml with the site-specific models (M1/M2, M7/M8) of the phylogenetic analysis by maximum likelihood (PAML) package (59). When parameter estimates under a model allowing positive selection suggested the presence of a number of sites with ω > 1, Bayes’ theorem was used to calculate their posterior probabilities (56).

**RESULTS**

The RM superfamily in *Spirulina*. To identify all members of the RM superfamily, we performed TBLASTN searches of the *Spirulina* in-finishing genome database with RM genes for the bacterial and archaeobacterial homologs. Similar searches were run with the Pfam domains (HSDR, DEXDC, Methylase-M, N6_Mtase, and MethyltransfID12) of the RM proteins to avoid the exclusion of highly diverged sequences with just limited conserved motifs. PFAM and SMART domain analyses with the derived sequences employed as queries were then carried out to eliminate false positives. After the above analyses, a total of 30 putative RM genes were obtained from the *Spirulina* genome database (Fig. 1 and Supplemental Table S1; available at the *Physiological Genomics* web site).1

1 The Supplemental Material for this article (Supplemental Table S1) is available online at http://physiolgenomics.physiology.org/cgi/content/full/00255.2005/DC1.
Type I RM systems comprise three closely linked subunits, which can distinguish them obviously from other type of RM genes without further experimental research. In this study, the symbol for Type I RM system was hsd, and thus the genes were hsdR, hsdM, and hsdS, whereas other types of RM genes or the solitary MTases were uniformly termed as REase or MTase.

Six clusters of the *Spirulina* genome, ranging from 4.7 to 11.7 kb, were found containing the putative Type I RM systems (Fig. 1). In *cluster I*, hsdM [688 amino acids (aa)] and hsdS (392 aa) should be cotranscribed from the same promoter, with a four-nucleotide overlap. The hsdR gene was downstream of the hsdM/S gene partitioned by two ORFs (putative ATP-binding protein and orf1). In *cluster II*, the hsdS subunit was interrupted by a stop codon (DQ239733, 2642-8E-37), AgeI (ACCGGT, P value = 8E-37), and HindVP (GRCGYC, P value = 2E-50), respectively. These MTases can be annotated into three groups: m4C (IV-MTase, V-MTase), m5C (MTase1–9), and m6A (MTase10, -11) based on querying the Restriction Enzyme database (REBASE) (41).

**RT-PCR analysis of RM genes in Spirulina.** Because gene expression patterns can provide important clues for gene function, we employed RT-PCR experiments to characterize gene expression profiles of the *Spirulina* MTases. Previous studies revealed that many of the RM systems in *Anabaena* sp. 7120 or *Helicobacter pylori* J99 were partially or completely inactive (25, 31). Similar results were also found in the RM systems of *S. platensis*. As shown in Fig. 2, 6 of 18 MTase genes were not expressed in *Spirulina*. Notably, the hsdM gene in the II-hsd system was expressed, whereas its cognate hsdS gene was defective, with a stop codon in the coding frame resulting in two split TRD domains. Most of the HsdS subunits carry two transD12) and the adjacent putative REases. Because the sequences of type II MTases possess a moderate level of sequence conservation, most of them should be identified with the computational searches. In contrast, REases only share sequence similarity when they have similar recognition specificities (43). Therefore, only four MTases could be identified to possess cognate REases in the *Spirulina* genome (Fig. 1), and whether other closely linked ORFs function as the endonucleases can be clarified by an experimental approach. Four cognate REases (REase1, -2, -6, and -7) share significant sequence similarities with known AvaIVAP (GCTNAGC, P value = 8E-37), AgeI (ACCGGT, P value = 2E-81), BsuBI (CTGCAG, P value = 2E-50), and HindVP (GRCGYC, P value = 1E-82), respectively. These MTases can be annotated into three groups: m4C (IV-MTase, V-MTase), m5C (MTase1–9), and m6A (MTase10, -11) based on querying the Restriction Enzyme database (REBASE) (41).

![Fig. 1. Organization of the gene clusters potentially encoding restriction-modification (RM) system of *Spirulina platensis*. Arrows represent the direction of translation and the relative sizes of open reading frames (ORFs) deduced from analysis of the nucleotide sequence. Gene names are given at top of corresponding region. Open arrows indicate the flanking ORFs; other types of arrows indicate different subfamilies of RM system. “×” in the arrow indicates the nonsense codon. Gene clusters encoding Type I RM genes are marked as clusters I–VI; the putative Type II RM genes or solitary methyltransferases (MTases) are marked nos. 1–11. REase, restriction endonuclease.](http://physiolgenomics.physiology.org/)

![Fig. 2. RT-PCR analysis of the expression of RM genes in *S. platensis*. See MATERIALS AND METHODS for details.](http://physiolgenomics.physiology.org/)
Anabaena sp. 7120, and BLAST score two species if they could be aligned over, at minimum, a tentatively assumed to identify homologous counterparts in locations. (Fig. 3). Because the onto the unicellular and filamentous cyanobacterial genomes complete genomic structure of the RM genes, we mapped them RMs are listed in Supplemental Table S1. To elucidate the REBASE were used as queries for database. All the putative Protein sequences of RM genes previously described in http://img.jgi.doe.gov/v1.1/main.cgi?page=ncbi.nlm.nih.gov/BLAST Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST) and Joint Genome Institute (JGI; http://img.jgi.doe.gov/v1.1/main.cgi?page=geneSearchForm)). Protein sequences of RM genes previously described in REBASE were used as queries for database. All the putative RM s are listed in Supplemental Table S1. To elucidate the complete genomic structure of the RM genes, we mapped them onto the unicellular and filamentous cyanobacterial genomes (Fig. 3). Because the S. platensis genomic sequences were still in the draft format, all the putative RM s were then mapped onto the chromosome evenly, and did not represent their actual locations.

Sequences giving the better reciprocal BLAST hits were tentatively assumed to identify homologous counterparts in two species if they could be aligned over, at minimum, a BLAST score > 90 and an e-value < 1E-10. By these criteria, all the RM domains derived from the genome of S. platensis, Anabaena sp. 7120, and Anabaena variabilis ATCC 29413 were examined to have quite different sequence similarities. Some MTases (MTase1–3) in A. variabilis ATCC 29413 had more homologs with high similarities in the other two filamentous cyanobacteria, while other MTases had less or no counterparts in Spirulina and Anabaena sp. 7120. Moreover, no obvious similarities (usually < 20%) were found among the specificity genes (HsdS), which encode polypeptide sequences responsible for the recognition of the target sequence. A striking example of evolutionary conservation is provided by the genes encoding Type I RMs in Anabaena sp. 7120 (hsd2 and hsd3) and A. variabilis ATCC 29413 (hsd2). The amino sequences of these gene products (HsdR and HsdM) are nearly identical in Anabaena sp. 7120; 96% identity was preserved in HsdR/M between Anabaena sp. 7120 and A. variabilis ATCC 29413. Interestingly, all three Type I RMs in A. variabilis ATCC 29413 shared some similarities with the HsdR/M in Anabaena sp. 7120, whereas the Type I RMs in Spirulina exhibited much less sequence similarity with those in the Anabaena species.

Genomic distribution and sequence analysis of RM genes. Computer-based nucleotide and protein searches were performed using different BLAST search programs at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST) and Joint Genome Institute (JGI; http://img.jgi.doe.gov/v1.1/main.cgi?page=geneSearchForm)). Protein sequences of RM genes previously described in REBASE were used as queries for database. All the putative RM s are listed in Supplemental Table S1. To elucidate the complete genomic structure of the RM genes, we mapped them onto the unicellular and filamentous cyanobacterial genomes (Fig. 3). Because the S. platensis genomic sequences were still in the draft format, all the putative RM s were then mapped onto the chromosome evenly, and did not represent their actual locations.

Sequences giving the better reciprocal BLAST hits were tentatively assumed to identify homologous counterparts in two species if they could be aligned over, at minimum, a BLAST score > 90 and an e-value < 1E-10. By these criteria, all the RM domains derived from the genome of S. platensis, Anabaena sp. 7120, and Anabaena variabilis ATCC 29413 were examined to have quite different sequence similarities. Some MTases (MTase1–3) in A. variabilis ATCC 29413 had more homologs with high similarities in the other two filamentous cyanobacteria, while other MTases had less or no counterparts in Spirulina and Anabaena sp. 7120. Moreover, no obvious similarities (usually < 20%) were found among the specificity genes (HsdS), which encode polypeptide sequences responsible for the recognition of the target sequence. A striking example of evolutionary conservation is provided by the genes encoding Type I RMs in Anabaena sp. 7120 (hsd2 and hsd3) and A. variabilis ATCC 29413 (hsd2). The amino sequences of these gene products (HsdR and HsdM) are nearly identical in Anabaena sp. 7120; 96% identity was preserved in HsdR/M between Anabaena sp. 7120 and A. variabilis ATCC 29413. Interestingly, all three Type I RMs in A. variabilis ATCC 29413 shared some similarities with the HsdR/M in Anabaena sp. 7120, whereas the Type I RMs in Spirulina exhibited much less sequence similarity with those in the Anabaena species.

Phylogenetic analysis. RMs identified from cyanobacteria and other prokaryotic organisms, such as proteobacteria, archaea, chlorobi, firmicutes, spirochaetes, and actinobacteria, were used to construct phylogenies to discern the evolutionary history of RM system. The 16s rDNA phylogeny clearly showed separation of prokaryotes into several monophyletic clades, and all these branches had significant statistical support (see Fig. 5A). The hsdM and hsdR phylogenies illustrated in Fig. 4, however, exhibited quite different topologies, where different sources of hsd genes scattered throughout the tree. Although several monophyletic groups were inferred with strong bootstrap support, these groups did not correspond to their 16s rDNA phylogeny. Furthermore, in the filamentous species of cyanobacteria, the hsd genes were usually present in several copies, and, based on the tree topology and their genomic organization, these copies might have quite different evolutionary histories. For example, hsd2 and hsd3 of

Fig. 3. Genomic organization of RM genes in unicellular (D. Synechocystis sp. 6803; E. Synechococcus elongatus) and filamentous cyanobacteria (A: S. platensis; B: Anabaena variabilis ATCC 29413; C: Anabaena sp. 7120). Vertical bars show the locations of RM genes, with different colors representing different Pfam domains: red, Methylose_S; light green, HSDR_N; purple, DEXDc; blue, Methylose_M; black, N6_Mtase; orange, MethyltransfD12; dark green, REase (not the Pfam nomenclature). Long horizontal line indicates the chromosome, whereas short horizontal lines denote the extranuclear plasmids. A vertical bar above a horizontal line indicates the transcriptional direction opposite to that below a horizontal line. Positions and orders of RM genes in horizontal line indicates the transcriptional direction opposite to that below a horizontal line. Positions and orders of RM genes in

Physiol Genomics • VOL. 24 • www.physiolgenomics.org
Anabaena sp. 7120 were adjacent to each other along the chromosome and were nearly identical at the amino acids level. These two copies most likely have evolved through recent duplication, whereas evolution of other RMs (hsd1, hsd4, and hsd5) is more complex and probably involves horizontal gene transfers.

Unlike the wide, dispersed distribution of hsd genes through the phylogeny, most of the cyanobacterial MTases in Fig. 5B formed a separate clade (77% bootstrap value), with the exception of six sequences (A7-MTase1, A7-MTase3, Se-MTase1, Se-MTase2, S6-MTase2, and S6-MTase3). The archaebacterial sequences were located elsewhere on the tree. Most of the γ-proteobacterial sequences formed a separate clade with 99% bootstrap support. The exception was the sequences from Moraxella bovis (PG-BAA03071) and Chromohalobacter salexigens (PG-ZP_00473324), which grouped with other prokaryotes. Moreover, two viral MTases were clustered with four β-proteobacterial sequences (100% bootstrap value), indicating horizontal transfers of MTases between the bacterial host and the phage. These results evidently suggested multiple lateral transfers of RM genes both within and between the major clades of the tree (Figs. 4 and 5).

We also performed phylogenetic analysis separately for each Pfam domain of HsdR and HsdM (HSDR_N, DEXDc, Methylase_M, and N6_Mtase), and phylogenies obtained from these different domains were quite similar to the phylogenetic trees illustrated in Fig. 4 (data not shown). By contrast, phylogenetic analyses using different domains of HsdS gave different phylogenetic topologies with no strong bootstrap support (data not shown). HsdS usually contains two Pfam domain (Methylase_S) repeats, and little similarity exists outside of these motifs. The diversification of hsdS genes to achieve maximal variation of specificity should be driven by strong selective pressures (55).

**Evolutionary analysis of RM genes.** HsdM and HsdR, as with interacting domains, must coevolve to form an integral and active complex with HsdS subunit. Phylogenetic trees (Fig. 4) based on the NJ method showed, extensively, similarity in clustering of cognate hsdM and hsdR genes, also indicating the coevolution of two subunits of Type I RMs. Previous studies revealed several conserved motifs in HsdM, among which motif I (D/E/SXFXGXG) can bind the AdoMet and motif IV (N/DPPF/Y/W) can be involved in substrate binding or in the catalytic activity (34). As for HsdR, several conserved motifs were also found in the DEAD-box domain, which formed a nucleoside triphosphate (NTP)-binding pocket and a portion of a nucleic acid-binding site (13, 34).
We estimated the pairwise correlation of five domains containing the above functional motifs. Different sequence weighting methods (14, 54) used to avoid sequence bias gave similar results, and only the results derived from the Vingron and Argos (54) approach were shown. We first considered such amino acid properties as volume (size) for coevolution analysis. From Fig. 6, we can see that many site pairs are highly correlated in a positive or negative way at the 99.9% level. In the $hsdM$-motif I, seven sites (sites 7–11, 13, 15) were correlated with the sites in the other $hsdR$-motifs, and these sites mostly resided in or were surrounding the featured residues. Similar results were found in the $hsdM$-motif IV. Sites close in the primary sequence are likely to be close in the tertiary structure and are therefore likely to have coevolved. In this study, however, the observed numbers of positively or negatively coevolving site pairs within the distant sites were significantly higher than those close pairs, indicating that the amino acids size interactions, especially those distant site pairs, should play an important role in the tertiary structure formation. Another amino acid characteristic, hydrophilicity, was also used to perform the coevolution analyses. From a structural and functional perspective, correlated amino acid replacements may have occurred among sites to maintain the same hydrophilicity properties, size constraints, electrostatic charges, etc. Many pairs of residues showed high correlations, and nearly one-third of the coevolving site pairs (70 of 212) were identical to those computed for the residue size effect analysis (Fig. 6). The hydrophilicity effect, however, was different with the size correlations in the $hsdM$-motif I, where much fewer coevolving site pairs were found in this domain. Nearly all the sites in the $hsdR$-motifs I, Ia, II, and III exhibited strong positive or negative correlations, revealing the great significance of these sites to form a hydrophobic hydrolysis pocket.

To analyze the possibility of positive selection acting on specific codons in the RM family, likelihood ratio test (LRT) was used to test for variation in the $d_{s}/d_{a}$ ratio among sites in sets of closely related RM genes. However, most of them were highly divergent, and it was problematic to align such cases with confidence. For these reasons, we performed most of our analyses with local clades of 10–15 sequences, which could keep genetic divergence reasonably small and keep alignments robust for LRT analysis. In the $hsdM$, $hsdS$, $hsdR$, and Type II REase gene families, all site-specific models provided little or no evidence for positive selection (data not shown).

In contrast, with the MTase gene family, we obtained results indicating positive selection, and it was estimated that a total of 10.2% sites were likely to be positively selected ($\omega = 31.1, P < 0.001$). Twelve highly homologous MTases (Sp-Mtase1, Sp-Mtase4, Av-Mtase2, A7-Mtase5, A7-Mtase10, A7-Mtase8, S6-Mtase1, Np-40208720, AAA50432, YP_199246, BAC81824, ZP_00517809), which could keep the alignment robust for LRT analysis, were used to construct a phylogenetic tree. Codon usage bias is known to affect the estimation of synonymous and nonsynonymous substitution rates (58). Thus two different codon usage bias models were used in all of our analyses: F3$\times$4 and F61. The results under these two models were quite similar, and those under F3$\times$4 were presented only. To examine the relationship of sites under positive selection to the protein structure, we mapped these sites onto the MTase tertiary structure (Fig. 7). First, the sequence alignment of MTase protein sequences (m5C family) previously used for PAML analysis was submitted to the protein-modeling server: SWISS-MODEL [http://swissmodel.expasy.org/SWISS-MODEL.html](http://swissmodel.expasy.org/SWISS-MODEL.html) with PDB-1dctA (39) as the modeling template. The derived structure using Av-Mtase2 as the target sequence is shown in Fig. 7A. Several positively selected residues (482S, 484M, 494G, 507S, 508S) were located at the MTase COOH-terminal extension of ~50 aa, where no coordinate residues were found in the template (1dctA), and thus they could not be mapped onto the tertiary structure. Although the other predicted positively selected sites appeared to be relatively uni-
formly distributed along the protein sequence, a clear pattern was revealed when the sites were examined in relation to the tertiary structure: they were all exposed amino acids, and were clustered on one face of the dimer toward the nucleic acids (Fig. 7B). Two sites were located in the DNA-binding pocket: one was site 376A at the scaffold region of the small domain; another site, 197T, was located at one of the DNA backbone-binding motifs adjacent to the active site (39). Thus these sites were available to participate in DNA-protein interactions. Moreover, other sites were located at the helix D (206Q, 213E, 217N) or the coil linking H3 and H4 (126S, 131I).

DISCUSSION

To date, more than 100 cyanobacterial strains are known to produce REases (41). Generally, only one or two REases were found in each strain of cyanobacteria through conventional methods to test the crude cell extracts for their restriction activity (29, 37). The main reason is the difficulty of detecting Type I RM or solitary DNA MTases by biochemical or genetic assay (34, 41). Genome-wide screening of RM systems based on the genome-sequencing project, however, leads to a virtual explosion in the number of putative RM genes, which provides us a new and comprehensive insight into the cyanobacterial RM systems.

In this study, we have identified and analyzed a large number of putative RM genes from all the finished or ongoing cyanobacterial genomes. From Fig. 2 and Supplemental Table S1, we can see that more RM genes are found in filamentous cyanobacteria (Anabaena, Spirulina, Nostoc) than in unicellular species (Synechocystis, Synechococcus, Prochlorococcus). We also found that the bacterial or archaeobacterial genomes in REBASE with a tremendously large number of RM genes are usually pathogenic bacteria (e.g., Campylobacter jejuni, H. pylori, Neisseria gonorrhoeae) or extremophiles (e.g., Chloroflexus aurantiacus, Methanocaldococcus jannaschii). Despite a wide variation in RM contents among closely related species, these pathogenic bacteria or extremophiles have a significantly higher number of RMs than the other species (t-test; P < 0.001). It is considered that such disproportionate expansion of certain gene families allows these organisms to adapt to a wider range of environmental conditions, or exploit a greater variety of nutrient sources (26). When we ruled out the chromosome size effect and calculated the relative number of defense-related genes out of the putative cluster of orthologous groups (COGs) in cyanobacteria, it is still significantly greater than that in unicellular species, and similar results were also found in the cyanobacterial signal transduction systems (P < 0.001; data not shown). Unicellular Prochlorococcus
living in the oligotrophic open ocean possesses no, or a limited number of, RM systems, as well as other genes involved in environmental sensing and regulatory systems. Dufresne et al. (9) assumed that the major driving force behind genome reduction within the Prochlorococcus radiation has been a selective process favoring the adaptation of this organism to its environment. Most filamentous cyanobacteria exhibit a wide range of ecological tolerance and are found in freshwater, marine, and terrestrial habitats. A good example is Nostoc punctiforme, a filamentous nitrogen-fixing cyanobacterium that displays an extraordinarily wide range of vegetative cell developmental alternatives, physiological properties, and ecological niches, including broad symbiotic competences with plants and fungi (32). The identification of a new family of putative PD-(D/E)XXK nuclease in various cyanobacteria also indicated that the expansion of such a nuclease family should be correlated with their host's ecological plasticity (11).

Nobusato et al. (36) analyzed all of the RM genes identified from the complete genome sequences of two H. pylori strains and found that these genes were scattered among diverse groups of bacteria in the phylogeny. Horizontal gene transfer may have a considerable influence on the widespread distribution of RM systems among bacteria (23). Similar results were found in the phylogenies of cyanobacterial RMs. However, most of the Type II or the solitary MTases in cyanobacterial genomes seemed to form a separate clade, in contrast to the wide, dispersed distribution of Type I RMs in the phylogeny. Matveev et al. (31) considered the solitary MTases to be the ancient origin within cyanobacteria. Here, the MTases that fell out of the cyanobacterial clade were all solitary ones (Fig. 5), which shared higher similarities with their homologs in proteobacteria or firmicutes, whereas the solitary MTases in A. variabilis ATCC 29413 and S. platensis were all tightly grouped with the Type II MTases, indicating the horizontal gene transfer of the solitary MTases between cyanobacteria and other bacteria. In Spirulina, most of the MTases were identified to be orphan genes with the absence of their cognate REases in the closely linked loci. One explanation may be a failure in detection of the REase gene because of its high divergence or pseudogenization. The other concern was that RM genes were sequentially transferred into the host genome, with the modification gene being transferred first. These solitary genes may acquire a unique function compared with other MTases in RM systems (23). Most of the Type I RM systems in cyanobacteria, however, seemed to be transferred en bloc. Three subunits of the Type I RM system, hsdM, hsdS, and hsdR, are usually closely linked. As shown in Figs. 4 and 6, the similar phylogenies of HsdM and HsdR and several highly coevolved motifs between two subunits suggest that en bloc lateral transfer of such linked genes and cooperative evolution should be more common in the Type I RM systems. After the host acquires the novel RM genes, sequence permutation or gene duplication seems to be the major mechanism in the diversification of RM systems (5, 16).

In Type I RM system, HsdR, HsdM, and HsdS form a complex (R2M2S1) to perform both endonuclease and MTase activities (4, 34, 55). Compared with the extensive literatures describing the conserved motifs in HsdM and HsdR and elucidating their functions, covariation analyses of both subunits are rare. Previous studies identified seven conserved motifs in the DEAD-Box of HsdR, and all these motifs are involved in the ATP binding/hydrolysis reaction (13). Here, we applied the CRASP method (1) to estimate correlation coefficients of six conserved domains in hsdM and hsdR proteins, and found a significant (at the 99.9% level) pairwise correlation among these domains. In the hsdR subunit, motifs I and II are directly involved in catalyzing the NTP hydrolysis reaction; motif Ia interacts with the sugar-phosphate backbone, whereas motif III is involved in hydrogen bond and stacking interactions with the DNA bases (13). Korolev et al. (27) found that residues in motifs Ia and III are in close proximity to a loop containing the aspartate and glutamate residues in motif II, suggesting a potential role for these regions in the transduction of energy from the ATPase site to the DNA molecule. Site-directed mutations in these motifs usually result in greatly decreased ATP hydrolysis and eliminated helicase activity (8, 13). Here, we identified that a large number of site pairs in these motifs are positively or negatively correlated, indicating these residues should function in forming a flexible hydrophobic hydrolysis pocket. It appears likely that subtle changes in ATP or DNA binding characteristics might arise through minor conformational changes associated with the structural interac-
tions among these motifs. However, further experiments are required to examine such amino acid changes for their functional implications.

RM genes seem to be frequently exchanged between species (24) and to evolve quickly after invading the new host (17). To elucidate the selective pressure under the diversification of RM genes, the likelihood ratio test was employed to identify whether those common ancestral genes were under positive selection after they were transferred into the new hosts. Unlike the “arms race” effect occurring on other defense-related genes, no signature of adaptive evolution was found in the RM systems except some groups of MTases. Although we cannot rule out that some groups of genes in these families are under positive selection, it is the case in fact that many genes, especially for Type I RM systems, are pseudogenes in various states of decomposition (25, 31). Pairwise dS/dN analysis of Type I RM genes from closely related species also showed strong signatures of relaxed selective constraints rather than positive selection (data not shown). These findings strongly suggest that the evolution of RM genes does not follow the same pattern as most of the defense-related genes.

An exception is the MTase gene family, which has a total of 10.2% sites with elevated rates of nonsynonymous substitution. The majority of these residues could be precisely located on the MTase tertiary structure, which provides a structural interpretation of adaptive evolution. Two positively selected residues are located in the DNA-binding pocket, whereas the other five residues are located between motifs III and IV (12S, 13I) or VI and VII (206Q, 213E, 217N) when they are mapped onto the secondary structure of M.HaeII (3, 39). Extensive studies revealed that the variable region between motifs VIII and IX determines specificity (6, 20, 33, 39). However, Timar et al. (49) found that two amino acid replacements between motifs VI and VII resulted in relaxed recognition specificity of SinI DNA-MTase, suggesting that some specificity-determining elements should be located in the large domain. LRT analysis also suggested that these regions were subjected to strong adaptive evolution, and some residue replacements in these motifs might have provided structural variations and thereby generated novel characteristics for the recognition of new target sites. It still remains unclear how positive selection really operated on such specificity-determining domains. Further studies focusing on the intraspecific homologous genes would be useful to address the role of selective pressure in such families.

GRANTS
This work was supported by grants from the Key Innovative Project (KZCX3-SW-215) of the Chinese Academy of Sciences.

REFERENCES