Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins

Kyonoshin Maruyama1,2, Naoki Sato1,* and Niji Ohta2

1Department of Biochemistry and Molecular Biology, Faculty of Science, Saitama University, 255 Shimo-Ohkubo, Urawa, Saitama Prefecture 338-8570, Japan and 2School of Human Science, Waseda University, 2-579-15 Mikajima, Saitama Prefecture 359-1192, Japan

Received January 4, 1999; Revised and Accepted March 8, 1999 DDBJ/EMBL/GenBank accession nos + AB003331, AB003332 and AB003690

ABSTRACT

The rbp gene family of the cyanobacterium Anabaena variabilis strain M3 consists of eight members that encode small RNA-binding proteins containing a single RNA recognition motif (RRM). Similar genes are found in the genomes of Synechocystis sp. PCC6803, Helicobacter pylori and Treponema pallidum, but are absent from the other completely sequenced prokaryotic genomes. The expression of the rbp genes of Anabaena is induced by low temperature, with the exception of the rbpD gene. We found four stretches of conserved sequences in the 5′-untranslated region of the cyanobacterial rbp genes that are known to be induced by low temperature. The cold-regulated Rbp proteins contain a short C-terminal glycine-rich domain. In this respect, these proteins are similar to plant and mammalian glycine-rich RNA-binding proteins (GRPs), which also contain a single RRM domain with a C-terminal glycine-rich domain and are highly expressed at low temperature. Detailed phylogenetic analysis showed, however, that the cyanobacterial Rbp proteins and the eukaryotic GRPs do not belong to a single lineage, but that the glycine-rich domains are likely to have been added independently. The cold-regulation of both types of proteins is also likely to have evolved independently. Furthermore, the chloroplast RNA-binding proteins are not likely to have originated from the Rbp proteins of endosymbiont cyanobacterium, but are supposed to have diverged from the GRPs. These results suggest that the cyanobacterial Rbp proteins and the eukaryotic GRPs are similar in both structure and regulation, but that this apparent similarity has resulted from convergent evolution.

INTRODUCTION

We recently isolated five rbp genes encoding RRM (RNA recognition motif)-type RNA-binding proteins from the cyanobacterium Anabaena variabilis strain M3 (1–3). Similar genes have been reported in a closely related strain, Anabaena sp. PCC7120 (4), as well as in distantly related cyanobacteria, such as Chlorogloeopsis sp. PCC6912 (4), Synechococcus sp. PCC6301 (5), Synechococcus sp. PCC7942, Synechococcus sp. PCC7002 (6) and Synechocystis sp. PCC6803 (7,8). Cyanobacterial RNA-binding proteins contain a single RRM and are smaller than other reported RNA-binding proteins containing RRMs (see below). In cyanobacteria, there are two types of RRM-type RNA-binding proteins: those which contain a very short C-terminal glycine-rich domain and are regulated by low temperature and those which lack a glycine-rich domain and are constitutively expressed (1–3,5). An Anabaena disruptant of the cold-regulated rbpA1 gene starts the initial steps of differentiation to a heterocyst (nitrogen-fixing cell) during growth at low temperature in the presence of nitrate in the medium (9). This suggests an important role of the RbpA1 protein in maintaining a correct gene expression pattern or cell identity in cyanobacteria.

Curiously, until now, RRM-type RNA-binding proteins have been reported predominantly in various species of cyanobacteria (1–8). In most cases, there are multiple copies of genes that encode these RNA-binding proteins (rrp genes). Among other prokaryotes, only the completely sequenced genomes of Helicobacter pylori (10) and Treponema pallidum (11) contain a single copy of an rbp gene. No other bacteria, including the completely sequenced bacterial genomes of Haemophilus influenzae (12), Mycoplasma genitalium (13), Methanococcus jannaschii (14), Bacillus subtilis (15) and Escherichia coli K-12 (16), are known to contain rbp homologs. Several different kinds of RNA-binding proteins that contain RRMs are known in eukaryotes. These proteins contain one or more copies of the RRM (about 80 residues long) with various sizes and numbers of auxiliary domains, such as a glycine-rich domain, an arginine-rich domain or an acidic domain, among others. The three-dimensional structure of the RRM domain of the U1A protein (17) and the SXL protein (18) among others have been determined. The results indicated the presence of a four-stranded β-sheet, which provides an RNA-interacting surface, and two α-helices that sustain the β-sheet from the back side. A protruding loop (loop 3) is also supposed to be involved in the
interaction with RNA. In the two central β-strands, an octapeptide called the ‘RNP-1 motif’ and a hexapeptide called the ‘RNP-2 motif’ have been identified as the RNA-binding sites (19). Within these motifs, the hydrophobic sidechains are supposed to interact with the RNA bases by stacking.

Many of the eukaryotic RRM-type RNA-binding proteins are known to be involved in pre-mRNA splicing or alternative splicing in the nucleus or in the control of mRNA translation efficiency in the cytoplasm (19,20). In plants, two additional classes of RRM-type RNA-binding proteins are known. First, the proteins localized within the chloroplast are supposed to be involved in the 3′-end processing of chloroplast mRNA (21). Second, glycine-rich RNA-binding proteins are localized in the nucleolus and are supposed to be involved in the processing of pre-rRNA (20). These proteins are known to respond to various stresses such as low temperature (22,23), wounding (24) and heavy metals (25). Recently, homologous glycine-rich RNA-binding proteins (GRPs) were also isolated from human (26), mouse (27) and Xenopus (28) and the expression of all of these is induced at low temperatures. Based on these findings, we are interested in a possible relationship between the cyanobacterial Rbp proteins and the eukaryotic glycine-rich RNA-binding proteins. Both are characterized by the presence of a single RRM, a C-terminal glycine-rich domain and induction by low temperature.

In the present study, we characterized the hitherto uncloned three genes in the *rbp* gene family in *A.variabilis* strain M3 and analyzed their expression at low temperature. We found four stretches of conserved sequences in the 5′-untranslated region (5′-UTR) of all the cold-regulated cyanobacterial *rbp* genes. We also re-analyzed the molecular phylogeny of the prokaryotic, organellar and eukaryotic RRM domains. We provide evidence for a monophyletic origin of the RRM domains, but against the endosymbiotic origin of chloroplast RRMs. We also propose that the presence of the glycine-rich domain and cold-regulation should be considered to be a result of convergent evolution of cyanobacterial Rbp proteins and eukaryotic GRPs.

### MATERIALS AND METHODS

#### DNA techniques

*Anabaena variabilis* strain M3 was grown photoautotrophically as described previously (1). DNA was isolated from the *A.variabilis* cells as described previously (1). A genomic mini-library of *A.variabilis* in pBluescript SK+ was screened with digoxigenin (DIG)-labeled *rbpA1* gene (XbaI–HindIII fragment) at low stringency (20% formamide during the hybridization, final washing in 2× SSC at 50°C). Other details of DNA blot hybridization have been described previously (1,2). Detection of DIG-labeled DNA probe was performed using an alkaline phosphatase–anti-DIG antibody conjugate (Roche Diagnostics) and a luminescence substrate CDP-Star (Roche Diagnostics). Nucleotide sequence was determined on both strands by the chain-termination method with the Taq Dye Terminator Sequencing Kit (Applied Biosystems, Foster City, CA). RNA analysis was performed as described previously (2).

#### Computer-assisted sequence analysis

A similarity search of open reading frames against the GenBank database was performed at the Genome Net WWW Server (http://www.genome.ad.jp) through the Internet. The sources of amino acid sequences used in the phylogenetic analysis (Figs 4–6) are listed in Table 1. Alignment of the 5′-UTR of the *rbp* genes was constructed by the Clustal X program v.1.64 for the
Figure 1. Multiple alignment of all of the RRM-type RNA-binding proteins of *A. variabilis* strain M3. The amino acid sequences were aligned with Clustal X software v.1.64 and the resulting alignment was shaded and modified by drawing software. The residues that are conserved in all of the sequences are shown by white letters on a black background. The residues that are conserved in six and seven sequences are shown by black letters on a gray background. Other residues are shown in black letters on a white background. Allowed conservative amino acid groupings are (I,V,L,M), (F,Y,W), (Q,E), (D,N), (R,K) and (T,S). The residues of probable cAMP- and cGMP-dependent protein kinase phosphorylation sites are shown by white boxes.

Figure 2. Time course of accumulation of the *rbp* transcripts after a temperature shift from 38 to 22°C. Cells of *A. variabilis* strain M3, which had been grown at 38°C, were transferred to 22°C at time zero. Aliquots of 25 ml of culture were taken at 10 and 30 min and 1, 2, 3, 5 and 10 h and then RNA was extracted. The total RNA was glyoxylated, electrophoresed and then transferred to a nylon membrane (Biodyne A). The membrane was probed with a DIG- or fluorescein-labeled *rbp* probe (insets). The intensity of each band was quantified by densitometry and plotted as a function of time. In each case, the maximum level of the transcript was taken as 100. The two lines for *rbpA3* represent results with the two transcripts observed for this gene (arrowheads in the inset).

Power PC (29) with the following parameter sets: gap open penalty = 5.00, gap extension penalty = 0.05, protein weight matrix = identity matrix. The alignment was finally adjusted manually. Conserved motifs were detected by the Coresearch program (30) and then adjusted by visual inspection. An alignment of RRM* s* was created by the Clustal X program with the following parameter sets: gap open penalty = 5.00, gap extension penalty = 0.05. The alignment was finally adjusted manually. Phylogenetic trees were constructed from the protein sequence alignment by the neighbor-joining method (31) using the Phylib package (the programs Seqboot, Protdist, Neighbor and Consense). To estimate the confidence level of monophyletic groups, bootstrap analysis of 100 replicates was conducted (32). We also performed the following analyses: UPGMA analysis (the program Neighbor with the UPGMA option); parsimony analysis with the program Protpars; neighbor-joining analysis of the DNA sequence alignment (the program Dnadist in place of Protdist); quartet puzzling analysis with the Puzzle program (33); maximum likelihood analysis of the protein sequence alignment with the Protml program (34). These calculations were performed with a SUN Enterprise 6000 server (Solaris 2.5.1), a Silicon Graphics O2 workstation (IRIX 6.3) or an Apple Power Macintosh 7600, depending on the software employed and the calculation load. The TreeView program (35) for the Power PC was used to prepare graphical representation of the phylogenetic trees.

RESULTS AND DISCUSSION

Identification of *rbp* genes in *A. variabilis* strain M3

Previously, we estimated that the cyanobacterium *A. variabilis* strain M3 contains eight genes that encode RNA-binding proteins (*rbp* genes) and identified five of the *rbp* genes (*rbpA1, rbpA3, rbpB, rbpC* and *rbpD*). Now we have cloned all of the remaining members of the *rbp* gene family of *A. variabilis*. Based on the results of Southern blot hybridization, we estimated the size of the *EcoRV* fragments that contain the uncloned *rbp* genes to be 5.1, 2.1 and 1.6 kb, respectively. These DNA clones were isolated from
respective mini-libraries of size-fractionated, EcoRV-digested genomic DNA fragments. The resulting genes, rbpA2, rbpE and rbpF, encode polypeptides of 103, 99 and 105 amino acids, respectively (Fig. 1). These sizes of polypeptides roughly corresponded to the bands (104, 99 and 107 amino acid residues, respectively) that were previously detected in a high resolution immunoblot analysis using an antibody raised against the RbpA1 protein (2).

One of us reported in a previous publication (2) that the nucleotide sequence of the rbpA2 gene in A. variabilis strain M3 seemed identical to that of the rbpA gene of Anabaena sp. strain PCC 7120. However, we have now found a small number of differences between the two genes at nucleotide positions 55, 299 and 300 from the initiation codon. These differences in the nucleotide sequence are reflected in differences in the amino acid sequence at the corresponding sites.

Each of the eight RNA-binding proteins of A. variabilis strain M3 has a single RRM and is smaller than the reported eukaryotic RNA-binding proteins containing RRMs. Except for RbpD protein, the seven Rbp proteins (note that the protein name ‘Rbp’ is derived from the gene name ‘rbp’ and is not an abbreviation for ‘RNA-binding protein’, thus we can say ‘Rbp protein’) contain as an auxiliary domain a very short glycine-rich domain at their C-terminus. However, only RbpD contains a probable cAMP- and cGMP-dependent protein kinase phosphorylation site in the C-terminal region.

Cold-regulation of the newly cloned genes

We previously showed that the expression of the rbpA1, rbpA2, rbpB, rbpC and rbpA3 genes is regulated by low temperature (1–3). The newly cloned genes, rbpE and rbpF, are also regulated by low temperature (Fig. 2). In the same figure, the time courses of the levels of the previously cloned genes are also shown for comparison. The level of the rbpF transcript increased after a temperature shift from 38 to 22°C and attained its maximum at the first hour. The level of the rbpE transcript attained its maximum at the second hour. It should be noted that the induction of the rbpC, rbpA3, rbpE and rbpF genes was transient. The two lines for the rbpA3 transcripts in Figure 2 represent results with the two transcripts detected for this gene (3). The level of the rbpA1 and rbpA2 transcripts remained high at steady-state at low temperature.

Conserved sequences in the 5′-UTR of the rbp genes

In a previous study (9), the rbpA1 gene was not required for cold-regulation of itself. In a recent study (36), both transcriptional and post-transcriptional regulation was found to be involved in the regulation of the rbpA1 gene. The finding that three kinds of proteins bind to the 5′-UTR that is required for the cold-regulation of the rbpA1 gene suggests that these factors are trans-acting
regulators of transcription. Now we have compared the 5'-UTR of all known rbp genes of cyanobacteria (Fig. 3). The diagram includes a 160 bp region that lies upstream of the translational initiation codon. We found four conserved regions in the 5'-UTR of the rbp genes. The conserved region RBS (5'-TTCCG-GA-3') is located about six bases upstream of the translational initiation codon and this motif is highly conserved in all of the cold-regulated rbp genes. The conserved region Box II (5'-TCTCCGAA-3') is located about 40 bases upstream of the translational initiation codon and includes a 160 bp region that lies upstream of the translational initiation codon. We found four conserved regions in the 5'-UTR of the cold-regulated rbp genes. Some conserved sequences are also known in the 5'-UTR of the des genes, which were not found in the 5'-UTR of the rbp genes, namely the cold box (11 bases long) in E.coli (38) and the CSBox1 and CSBox2 (each 15 bases long) in B.subtilis (37). But, these sequences were not found in the 5'-UTR of the cold-regulated rbp genes in the cyanobacteria. Also, neither Box I, II nor III are found in the 5'-UTRs of the des genes. These results suggest that the conserved sequences (RBS, Box I and Box III) in the 5'-UTR of the rbp genes might represent novel types of putative cis-acting elements of cold-regulated genes.

**Comparative analysis of the RRM-type RNA-binding protein**

As described above, rbp genes are widely distributed among various species of cyanobacteria that do or do not fix dinitrogen (N2). In addition, rbp genes are also found in some species of prokaryotes. The prokaryotic Rbp proteins are small (about 100 amino acid residues) and contain only a single RRM domain. It is interesting to note that the copy number of rbp genes correlates with the genome size as far as cyanobacteria are concerned. For example, the 2.6 Mb genome of Synechococcus sp. PCC6301 includes two copies of the rbp genes (40). The 3.6 Mb genome of Synechocystis sp. PCC6803 includes three copies of the rbp genes (7). The genome sizes of A.variabilis M3 and Anabaena sp. PCC7118 (eight copies of the rbp genes) were shown to be similar to that of Anabaena sp. PCC7120 (41), which was estimated to be 6.4 Mb (42). Herdman et al. (41) suggested, based on the measurement of the genome size of 128 cyanobacterial strains, that the cyanobacterial genomes have evolved by

![Figure 5. Multiple alignment of 30 selected RRMs. The alignment was obtained by Clustal X software v.1.64, with final manual adjustments. The sequence names are listed in Table 1. The inference of the secondary structures is taken from Oubridge et al. (17), Lee et al. (18) and Birney et al. (19). The positions of conserved residues are colored by three grade shading with red, blue and green. The threshold values were 80, 65 and 50%, respectively. Allowed conservative amino acid residues are listed in Table 1. The inference of the secondary structures is taken from Oubridge et al. (17), Lee et al. (18) and Birney et al. (19).](image)
multiplication of the whole genome, with the unit genome size being \( \sim 1.2 \times 10^9 \) Da (1.8 Mb). Since the copy numbers of the cyanobacterial rbp genes in various cyanobacteria are consistent with this theory, the increase in the copy number of rbp genes is likely to have occurred by duplication of the whole genome and not by simple gene duplication.

To estimate a possible relationship between the prokaryotic Rbp proteins and the eukaryotic RNA-binding proteins of type RRM (Fig. 4), we constructed an alignment (Fig. 5) and a phylogenetic tree of 30 representative RRMs from various organisms (Fig. 6). In the multiple alignment, highly conserved residues are highlighted in color, so that we can easily recognize characteristic residues. As reported previously (3,19), aromatic amino acids within the motifs RNP-1 and RNP-2 are highly conserved. In addition, some residues are conserved in almost all RRMs analyzed. In the cyanobacterial RRMs, A, E, W and M in loop 5 are highly conserved (Fig. 1). In the prokaryotic and chloroplast RRMs, M is highly conserved at the last position of the RNP-1 motif.

The phylogenetic tree was constructed by the neighbor-joining method (Fig. 6). We also constructed trees by the UPGMA, the parsimony and the maximum likelihood methods and obtained essentially similar results (not shown). We used only representative RRMs in Figure 6 to avoid over-representation of certain groups, which would change the tree artificially. To show the phylogeny of individual groups, we constructed individual trees that include all known sequences of a group with an outgroup. All of the known cyanobacterial RRMs are highly conserved and form a single group (Fig. 7A). The cyanobacterial RRMs are most similar to other bacterial RRMs, forming a single group (Fig. 6). We can recognize several clearly identifiable clusters in the phylogenetic tree. This tree suggests that RRMs are classified according to groups of proteins of the same function and the position of the RRM, such as the small nuclear ribonucleoproteins, poly(A)-binding proteins, alternative splicing factors, heterogeneous nuclear ribonucleoproteins and chloroplast RNA-binding proteins. This classification is not related to the plant and animal kingdoms. Even the plant RRMs of other classes are quite divergent from the chloroplast and plant glycine-rich RNA-binding proteins. These results suggest that the eukaryotic RRM-type RNA-binding proteins are of ancient origin, probably dating from the ancestral eukaryote.

Cyanobacterial RNA-binding proteins and plant and mammalian GRPs share a lot of common features. Most of these RNA-binding proteins have a single RRM and a C-terminal glycine-rich domain (20,26,27). There are two types of cyanobacterial RNA-binding proteins: those that have a very short glycine-rich C-terminal domain (about 15 residues) and those that are devoid of a glycine-rich domain. The C-terminal domains of plant GRPs are longer (about 80 residues) than the corresponding cyanobacterial domains. The glycine-rich domains of mammalian proteins consist of two types: short ones of about 80 residues and longer ones of about 100 residues. All members with glycine-rich C-terminal domains contain Tyr residues along their entire length at regular spacing. A significant number of these RNA-binding proteins are synthesized in response to low temperature stress:
cyanobacterial RNA-binding proteins are highly abundant (~2% of total cellular protein) at low temperature in *A. variabilis* (2). The expression of the plant genes *Ccr1, Ccr2* and *blt801*, each encoding a GRP, is also induced by low temperature (22,23). In animals, the expression of the human genes *RBM3* and *CIRP* (25,27), as well as the *Xenopus* gene *XCRIP* (28), is induced by low temperature. We wanted to know if these cold-regulated RNA-binding proteins of prokaryotes and eukaryotes have a common origin.

**The origin of the RNA recognition motif**

Based on the phylogenetic tree, we can estimate the origin and evolution of the RRM domain. RRMs are present in both prokaryotic and eukaryotic proteins (arrow in Fig. 6). Most authors before this study showed a close relationship between the cyanobacterial RRMs and the chloroplast RNA-binding proteins before addition of a glycine-rich domain. Therefore, this is a case of convergent evolution. Although the function of the glycine-rich domain is still unknown, this domain might be functionally important for both cyanobacterial and eukaryotic RNA-binding proteins.

Another important point to note is that the chloroplast RNA-binding proteins are not direct descendents of cyanobacterial Rbp proteins (Fig. 6). Most authors before this study showed a close relationship between the cyanobacterial RRMs and the chloroplast RNA-binding proteins based on the similarity of the RRMs of both types of proteins. If this were the case, it would be anticipated that the chloroplast RNA-binding protein lineage would be a sister group of the cyanobacterial lineage as in the case of GroEL (fig. 11 in ref. 43). But this is not the case. The chloroplast RNA-binding proteins must have diverged from other eukaryotic RNA-binding proteins, before addition of a glycine-rich domain or duplication of RRM domains.

**Relationship with CSD and probable functions of the cyanobacterial Rbps**

It is appropriate to mention the relationship between the RRM proteins and the cold-shock domain (CSD) proteins. The
synthesis of many CSD proteins is induced by cold shock (44). The CSD proteins contain motifs similar to the RNP-1 and RNP-2 motifs of RRM, but these counterparts are present in the CSD in a different arrangement. In fact, there is no simple sequence similarity between CSD and RRM proteins. But CSD proteins also bind to single-stranded DNA or RNA. CSD proteins are present in prokaryotes such as E.coli and B.subtilis, as well as in eukaryotes. Curiously, no CSD proteins are reported in cyanobacteria or other prokaryotes that contain RRMs. Apparently, eukaryotes lack proteins that contain CSD or RRM motifs. An interesting hypothesis is that the CSD and RRM proteins play similar roles in the prokaryotic cell and that only one of them is necessary for the activity of the cell. Both types of proteins are known to be synthesized to high levels after a cold shock or during growth at low temperatures. A suggested role of the Csp proteins in bacteria is as “RNA chaperones”, i.e. to unfold RNA at low temperature. The cyanobacterial Rbp proteins are also likely to play a role as RNA chaperones (2,4,36), despite a lack of direct experimental evidence for this possibility.

ACKNOWLEDGEMENTS

We thank Dr Tetsuo Iino (School of Human Science, Waseda University) for helpful suggestions. This work was supported in part by Grants-in-Aid for Scientific Research to N.S. (no. 09874167) and to N.O. (no. 085404) from the Ministry of Education, Science, Sports and Culture, Japan.

REFERENCES