IDENTIFICATION OF COGNATE RNA FOR THE RNA-BINDING PROTEIN, RbpA, FROM *Synechococcus* sp. PCC 7942

TIMOTHY G. STROZEN

Identification of cognate RNA for the RNA-binding protein, RbpA, from *Synechococcus sp.* PCC 7942

by

Timothy G. Strozen

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science in Biochemistry

Department of Biochemistry

Memorial University of Newfoundland

January, 2005
St. John's, Newfoundland

Abstract

The role RRM-type RNA-binding proteins (Rbps) perform in the cyanobacterial cell is unknown. However, evidence suggests that RRM-type RNA-binding proteins that contain a glycine-rich C-terminal auxiliary domain such as RNA-binding protein A from *Synechococcus sp.* PCC 7942 are involved in the cyanobacterial cold-shock response. In this investigation, the genomic SELEX technique was used to gain insight into the function of cold-inducible RbpA in the cyanobacterial cell by determining the nucleic acid binding specificity of the protein and by identifying the genes potentially regulated by RbpA.

The genomic SELEX technique involved use of a dsDNA library that contained 38-88 bp portions of *Synechococcus* 7942 genomic DNA sequence flanked by fixed DNA sequences. Representation of the entire *Synechococcus* 7942 genome in the genomic SELEX library was verified by nested-PCR analysis of a 43 base portion of the *rbpB* (RNA-binding protein B) gene. The library contained, in a staggered arrangement, molecules whose genomic portion terminated at 8 of the possible 13 nucleotides that correspond to bases 322 to 334 of the *rbpB* gene. These results indicated that if the genome was as equally represented in the library as that of the 13 base region of *rbpB*, then the genomic SELEX library would contain 3.38×10^6 different molecules. This number of library molecules corresponds to one molecule per 1.6 bases in the *Synechococcus* 7942 genome and thereby provided evidence that the library was a sufficient representation of the *Synechococcus* 7942 genome and could be used in the genomic SELEX procedure.

The genomic SELEX procedure involved multiple rounds of the same basic steps. The library was transcribed into RNA, RNA was mixed with N-terminal histidine-tagged RbpA (H_6RbpA) in a protein-RNA binding reaction. Isolation of H_6RbpA-RNA complexes was accomplished by Ni^{+2}-NTA metal chelate chromatography and the RNA molecules retained by this purification method were reverse-transcribed and PCR amplified to generate the dsDNA library used in the next round of selection.

To identify the nucleic acid binding specificity of RbpA, representative clones from the genomic SELEX library were sequenced after rounds 10 and 14 of selection. Following round 10 of selection, the amplified dsDNA was cloned into pUC19 and sequenced. Due to a lack of obvious sequence homology of the molecules selected after round 10 of SELEX, an additional 4 rounds of selection were performed under conditions of increased stringency. These conditions included an increase in salt concentration from 75mM to 150mM and a decrease in the RNA/protein ratio from 200:1 to 1:1. These conditions greatly enhanced selection, as evidenced by a substantial increase in nucleic acid sequence homology of representative members of the round 14 SELEX library. RNA selected in both rounds were of three types, RNA molecules poor in G/U residues (less than 50% G/U), rich in G/U (greater than 50% G/U) and G/U very rich sequences (greater than 90% G/U). Existence of G/U poor sequences was suprising because like many RRM-type RNA-binding proteins, RbpA has been shown to have a strong binding preference for GTP and UTP RNA homopolymers. Sequence homology analysis of the most homologous group of RNA molecules, the round 14 G/U rich RNA, identified a highly conserved putative RbpA consensus binding sequence 5'UGAAUGGGAGGUG 3'. The six 3' terminal ribonucleotides could be the most important nucleotides of the putative

consensus sequence due to the similarity of the 5' GAGGUG 3' sequence with a sequence 5' GUGGUG 3' present in many G/U very rich RNA sequences.

Comparative genomic sequence analysis of all sequences retained in rounds 10 and 14 of genomic SELEX allowed me to determine the genes whose RNA products are potentially regulated by RbpA via a binding interaction. Comparative BLAST analysis of the SELEX DNA sequences with that of the *Synechococcus* 7942 genomic sequence identified many genes. These include cold-shock inducible genes such *dsg*, *putP* and *sodB* that encode fatty acid desaturase, proline permease and superoxide dismutase respectively. Other interesting genes identified by comparative analysis include two transcription regulator proteins that contain conserved helix-turn-helix motifs and *ntrB* that encodes a nitrate reductase. Interestingly, some sequences identified by genomic SELEX corresponded to a sequence on the non-coding strand within the open reading frame of a gene. This result suggests that if RbpA is involved in regulating expression of these genes, it would involve the expression of a *cis*-encoded RNA molecule. This mechanism of gene regulation has not been studied extensively in prokaryotes and has not yet been characterized in cyanobacteria, however it has been identified in the eukaryotes. These results suggest that RbpA is involved in regulation of many genes that in some cases could involve a previously uncharacterized mechanism of gene regulation in cyanobacteria.

Acknowledgements

I would like to thank my supervisor Dr. Martin Mulligan for his guidance and patience in my work. I would also like to thank Donna Jackman, Janice Mayne, Shobhitha Ratnam and Toby Frost for their help in the technical aspects of my research.

Outside of the lab, I would like to gratefully acknowledge my girlfriend Rachael who not only has given me an infinite amount of support during this long process but had even moved across the country with me so that I could pursue graduate studies. In addition I thank my family and friends who have given me words of encouragement through this entire process.

Table of Contents

Section **Page**

Chapter 2: Materials and Methods

Chapter 3. Construction and Validation of a *Synechococcus* 7942 genomic SELEX library

List of Tables

List of Figures

List of Abbreviations

ARM	arginine-rich motif
bp	base-pair
BSA	bovine serum albumin
cDNA	complementary deoxyribonucleic acid
COGS	clusters of orthologous groups
CSD	cold-shock domain
DEPC	diethyl pyrocarbonate
DIG	digoxygenin
dNTP	deoxynucleotide triphosphate
ds	double-stranded
dsDNA	double-stranded deoxyribonucleic acid
EDTA	ethylenediamine tetraacidic acid
eIF	eukaryotic initiation factor
GTP	guanosine triphosphate
hnRNP	heterogeneous nuclear ribonucleoprotein
H_6RbpA	N-terminal 6x histidine labeled RNA-binding protein A
HTH	helix-turn-helix
IPTG	isopropylthiogalactoside
JGI	Joint genome institute
NBT	Nitroblue Tetrazolium salt
Ni^{+2}-NTA	Nickel Nitrilo-Tetracidic acid

NMR	nuclear magnetic resonance
ORF	open-reading frame
PABP	polyadenylate binding protein
PCR	polymerase chain reaction
PMSF	phenyl methyl sulfonyl flouride
RBS	ribosome-binding site
RNP	ribonucleoprotein
RRM	RNA recognition motif
Rbp	RNA-binding protein
rRNA	ribosomal ribonucleic acid
RT-PCR	reverse-transcription polymerase chain reaction
SELEX	systematic evolution of ligands by exponential enrichment
SDS-PAGE	sodium dodecylsulfate polyacrylamide gel electrophoresis
snRNP	small nuclear ribonucleoprotein
SOD	superoxide dismutase
SRBB	SELEX RNA-binding buffer
ssDNA	single-stranded deoxyribonucleic acid
ssRNA	single-stranded ribonucleic acid
UB	upstream box
UTP	uridine triphosphate
UTR	untranslated region
X-gal	bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

Introduction

1.1 Cyanobacteria

The cyanobacteria, once known as the "blue-green algae", are an important group of bacteria in terms of the role they played in the evolution of eukaryotic organisms as well as the role they played in creating the environment we know today. As bacteria capable of performing oxygenic photosynthesis, cyanobacteria played a role in leading to the creation of the oxygen-containing environment during the Archaean and Proterozoic eras (Schopf, 1983), an important step necessary for the evolution of life on this planet.

Cyanobacteria are integral to the evolutionary process not only by being involved in creating our oxygen-containing environment but also by playing a major role in the evolution of some eukaryotes. Through a process described by the endosymbiont theory (Gray, 1989), the evolution of green plants, red algae and glaucophytes was derived from a relationship involving a cyanobacterium living inside a eukaryotic host (Moriera *et al.*, 2000). Evidence of this relationship has been generated by biochemical analysis and more recently by comparison of whole genomic sequences of cyanobacteria and plants (Sato, 2002; Moriera *et al.*, 2000). For example, the conservation of extrinsic proteins that comprise the oxygen evolving centre of photosystem II of the photosynthetic apparatus is well conserved from cyanobacteria (*Synechocystis sp.* PCC 6803) to higher plants (*Arabidopsis thaliana*) according to genomic comparative analysis (Las Rivas *et al.*, 2004).

The structural conservation of cellular components such as proteins that exists between cyanobacteria and other prokaryotes and to a lesser extent with eukaryotes, is

one feature that makes the cyanobacteria interesting organisms of study. They can provide insight into the most basic cellular processes that, upon elucidation, can generate a greater understanding of more complex analogous systems of other highly evolved organisms. For example, the existence of RNP-type RNA-binding proteins is conserved from cyanobacteria to humans. Investigation of the role these proteins play in the cyanobacterial cell can not only provide information regarding the function performed by these proteins in cyanobacteria but can also provide a clue to the function of these proteins in higher organisms.

Many cyanobacterial species are capable of nitrogen fixation. In nitrogen starved conditions, some filamentous cyanobacterial species can convert atmospheric nitrogen into ammonia within a specialized anaerobic cell known as a heterocyst by the activity of the enzyme nitrogenase. Ammonia generated by the heterocyst provides an important source of nitrogen not only for itself but also for plants, protists and fungi which can have a symbiotic relationship with cyanobacteria.

1.2 *Synechococcus sp.* PCC 7942

Synechococcus sp. strain PCC7942 (hereafter referred to as *Synechococcus* 7942) is a unicellular obligate photoautotroph (Herdman *et al.*, 2001) that is nonheterocyst-forming and does not reduce nitrogen. Light is captured by means of the photosynthetic pigment molecules chlorophyll A and phycocyanin; the combination of these pigments generates the blue-green colour of the bacterium.

The first cyanobacterium demonstrated to be transformable was *Synechococcus* 7942 (Shestakov *et al.*, 1970) and has thus been the subject of extensive molecular genetic study. The genome of *Synechococcus* 7942 has been sequenced by the Joint

Genome Institute and at the time of analysis was available to the scientific community as a draft sequence (http://www.jgi.doe.gov). The genomic sequence has since been upgraded to that of a complete finished sequence.

1.3 Cyanobacterial cold-shock response

Both prokaryotic and eukaryotic cells have developed a diverse set of mechanisms designed to maintain cell viability in response to a wide range of environmental conditions. Cold temperatures present a variety of problems to the cell including decreased fluidity of the cell membrane, lowered processivity of enzymes due to an alteration of secondary and tertiary structures that affect the function of the protein and increased secondary and tertiary structures of DNA and RNA. RNA in particular is very susceptible to temperature change; a decrease in temperature introduces additional structures that can affect all aspects of RNA metabolism. Therefore in an effort to maintain viability at low temperature, cells generate a cold-shock response. In cyanobacteria, the cold-shock response is comprised of an induced expression of certain genes including those that encode fatty acid desaturases, protein chaperones and RNA-binding proteins. Together, the activity of these proteins is enough to prevent cell death and maintain cellular viability at lower temperature.

Cold temperature stress causes the membrane of a cell to become more rigid, which adversely affects the ability of the membrane to perform its normal function by preventing molecules from passing through the membrane. In cyanobacteria, fluidity of the membrane is re-established by desaturation of saturated fatty acids in the membrane by the activity of fatty acid desaturase enzymes, a process characterized in both *Anabaena variabilis* (Sato and Murata, 1980) and *Synechocystis sp.* PCC 6803 (Wada and Murata,

1990). Desaturation introduces a double bond in the carbon backbone of membrane fatty acids which in turn introduces a kink in the fatty acid structure resulting in a less ordered arrangment of phospholipid molecules. Since membrane fluidity is inversely related to the packing of phospholipid molecules in the membrane, introduction of a kink in the fatty acid structure increases fluidity. The importance of the fatty acid desaturases in the cold-shock response is evident in the upregulation of these genes in conditions of colder temperature. For instance, Northern blot analysis revealed the cold-induced upregulation of desaturase genes such as: lipid desaturaseA (*desA*) (Los *et al.*, 1993) and *desD* in *Synechocystis* 6803 as well as *desA* and *desB* in *Synechococcus sp.* strain PCC 7002 (Sakamoto and Bryant, 1997).

It is believed that rigidification of the membrane itself is the primary sensor that induces expression of genes upon conditions of cold-shock (Murata and Los, 1997). In *Synechocystis* 6803, a signal transduction pathway is initiated by a plasma membrane associated histidine kinase Hik33 (Suzuki *et al.*, 2000) through detection of membrane rigidity. However, DNA microarray analysis of a *Synechocystis* 6903 Hik33 mutant failed to provide evidence that the *desA*, *desB* and *desD* genes are under control of Hik33. Expression of all three genes was largely unaffected by the absence of Hik33 (Suzuki *et al.*, 2001). In the case of *desB* activation, experiments involving a luciferase fusion under control of the *desB* promoter (Suzuki *et al.*, 2000) demonstrated that another histidine kinase Hik19 and a response regulator protein Rer1, were required for the induction of *desB* under conditions of cold-shock. This result led to the description of a low temperature sensor responsible for *desB* induction. In the model, membrane-spanning Hik33 becomes autophosphorylated upon membrane rigidification generated under

conditions of cold-shock, the phosphate group is subsequently transferred to Hik19 and eventually to Rer1 which in turn activates transcription of *desB* (Suzuki *et al.*, 2000).

A recent development in the elucidation of the cold-shock response is the role of protein chaperones. Protein chaperones are large multisubunit proteins that are involved in the proper folding of nascent polypeptides or misfolded protein aggregates initially characterized as performing an important role in the heat-shock response. However, by western blot analysis, Hossain and Nakamoto (2002) showed that upon shifting a culture of *Synechococcus* 7942 to conditions of cold-shock (from 30°C to 16°C), the levels of chaperone proteins HtpG and GroEL increased. The requirement for HtpG in the cold-shock response was further supported by the finding that *Synechococcus* 7942 showed a reduced ability to recover from cold-shock conditions in the absence of functional HtpG.

A third set of genes that are induced upon conditions of cold-shock are those that encode two types of RNA-binding proteins. Cold-shock conditions introduce a propensity to generate additional secondary and tertiary structures in RNA molecules. These additional structures could have deleterious effects such as premature transcription termination by introducing transcription pause or termination signals, alteration of RNA half-life or prevention of mRNA translation that in turn could compromise cellular viability. The two cold-inducible gene types are those whose protein products contain an RNA Recognition Motif (RRM) domain and those that contain a DEAD (consecutive amino acids Aspartate, Glutamate, Alanine and Aspartate) box domain. The function of RRM-domain-containing RNA-binding proteins is not known, however due to the structural similarity and limited amino acid similarity of these proteins to that of the cold shock domain (CSD) of cold-shock proteins (Csps) in *Escherichia coli*, they could

function as RNA chaperones or transcription antiterminators. Members of the DEAD box group of proteins include CrhB and CrhC in *Anabaena* 7120 (Chamot *et al.*, 1999), which function as RNA helicases. These proteins unwind RNA secondary structures by use of energy derived from hydrolysis of ATP and are considered to be part of the cold-adapted ribosomal complex that re-establishes translation of mRNA at low temperatures.

1.4 RNA-binding proteins

The metabolism of RNA in prokaryotic and eukaryotic cells is accomplished by the activity of a multitude of RNA-binding proteins. Amino acid sequence comparative analysis of RNA-binding proteins characterized in animals, plants, fungi and bacteria has revealed the conservation of specific residues in groups of RNA-binding proteins. The most characterized group of RNA-binding proteins are the Ribonucleoproteins (RNPs) that contain the highly conserved RRM; this group is described in detail in the next section (1.4.1).

In addition to the RRM, other conserved RNA-binding domains include the Arginine-rich motif (ARM), RGG box, KH motif, cold-shock domain (CSD) and others. The ARM is a short sequence ranging from 10 to 20 amino acids in length that contains numerous arginine residues. With the exception of the arginine residues, the ARM does not have any additional sequence homology (Lazinski and Grzadzielska, 1989). The RGG box is 20 to 25 amino acids in length and is characterized by repeated arginine-glycine-glycine sequences within close proximity to each other (Kiledjian and Dreyfuss, 1992). The KH motif is composed of the consensus sequence VIGxxGxxI and can exist in multiple copies within an individual RNA-binding protein (Gibson *et al.*, 1993).

1.4.1 RNP-type RNA-binding proteins

RRM containing RNA-binding proteins (RNPs) are conserved throughout all kingdoms of life, thereby revealing the importance of these proteins in all cells. A multitude of cellular functions related to RNA metabolism from transcription to translation have been attributed to RNP activity. Examples include the activity of heterogeneous nuclear ribonucleoprotein (hnRNP) U1A and small nuclear ribonucleoprotein (snRNP) A2 Simian virus 40 (SV40) in splicing (Lutz and Alwine, 1994), human polyadenylate binding protein (PABP) in recognition of the mRNA polyA tail (Bernstein et al., 1989), sex-lethal (Sxl) protein involved in splice-site selection of the transformer gene *tra* involved in sex determination in *Drosophila melanogaster* (Baker, 1989), eukaryotic initiation factor eIF-4B involved in translation initiation of mRNA in rabbit reticulocytes (Benne and Hershey, 1978). In plants, chloroplast RNPs are involved in 3' mRNA processing. In *Escherichia coli*, RNPs/Csps are proposed to function as transcription antiterminators (Bae *et al.*, 2000) or RNA chaperones (Jiang *et al.*, 1997) both functions attributed to the activity of CspA.

The characteristic feature of the RNP family of RNA-binding proteins is the RRM (Swanson *et al.*, 1987). The RRM is approximately 90 to 100 amino acids in size and there can be one or more copies within a given RNA-binding protein. Conservation of amino acid residues in the RRM is focused within two regions termed the ribonucleoprotein 1 (RNP-1) and ribonucleoprotein 2 (RNP-2) sequences of 8 and 6 amino acids in length respectively (Dreyfuss *et al.*, 1988). In addition to the RRM, many RNA-binding proteins contain auxiliary domains rich in one or two amino acids. Examples include the glycine-rich domain of RbpA from *Synechococcus* 7942 (Mulligan

and Belbin, 1995), the glutamine-rich domain of the human cytolytic granule protein TIA-1 (Tian *et al.*, 1991), the proline rich domain of the murine protein mSAP49 (Ruiz-Lozano, 1997), the arginine-serine domain of the mammalian splicing factor U2AF (Zamore and Green, 1992), and the glycine-arginine rich domain of nucleolin (Ghisolfi *et al.*, 1992).

The RRM has a conserved $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ secondary structure, with the most conserved amino acid sequences RNP1 and RNP2 located on β-strands β_3 and β_1 respectively. Three-dimensional structural analysis of three eukaryotic RRM-domain-containing proteins, human U1A (Oubridge *et al.*, 1994) and *Drosophila melanogaster* proteins Sxl (Lee *et al.*, 1994) and Y14 (Shi and Xu, 2003), revealed a similar tertiary structure. The conserved $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ secondary structure of the RRM when folded creates a four-stranded antiparallel β-sheet flanked by two α helices. The interior β-strands create a scaffold that position the RNP-1 and RNP-2 conserved sequences in close proximity to one another (Nagai *et al.*, 1990) (Figure 1.1). NMR studies involving the small nuclear ribonucleoprotein U1A (U1A) and its 21 nucleotide cognate U1 snRNA revealed that RNA-protein interactions are mediated by stacking interactions between the 10 base loop of the hairpin RNA and highly conserved aromatic amino acids in RNP-1 and RNP-2 (Oubridge *et al.*, 1994). These conserved regions are required for the binding interaction between RNA and protein, however it has been shown that the nucleic acid specificity of the protein lies outside the conserved domains. In particular, RNA recognition is initiated by an interaction between variable portions of the RRM, namely the β_2-β_3 and β_1-α_1 loops (loops 3 and 1 respectively), and seven nucleotides located

Figure 1.1. Theoretical tertiary structure of residues 2 to 76 of the 107 amino acid RbpA protein of *Synechococcus* 7942. Structure was generated by use of the molecular viewer Deepview (Guex and Peitsch, 1997) accessed at the Swiss Institute of Bioinformatics website (www.expasy.org/spdbv/). The location of the RNP1 and RNP2 conserved sequences is shown.


RNP-2
RNP-1

within a well-structured helical region of the U1 snRNA. Subsequently, the RNA undergoes a conformational change such that a region of single-stranded RNA is localized against the β-sheet, whereupon stacking interactions between RNA and the conserved aromatic residues in RNP-1 and RNP-2 are initiated. When the stacking interaction is taking place, the C-terminal domain of the protein is reorganized, and this conformational change increases the interaction between the β-sheet region and the RNA by further exposing the highly conserved hydrophobic β-sheet (Allain *et al.*, 1996).

1.4.2 Cyanobacterial RNPs

RNA-binding protein encoding genes (*rbp*) are not well conserved throughout prokaryotes. With the exception of one gene detected in each of the completed genomes of *Helicobacter pylori* (Tomb *et al.*, 1997) and *Treponema pallidum* (Fraser *et al.*, 1998), *rbp* genes have not been found in prokaryotes other than cyanobacteria. Within cyanobacteria however, a large number of *rbps* have been characterized including eight *rbp* genes (*rbpA1*, *rbpA2*, *rbpA3*, *rbpB-F*) in *Anabaena variabilis* strain M3 (Sato, 1994; Sato, 1995; Sato and Maruyama, 1997; Maruyama *et al.*, 1999), four genes (*rbpA-D*) in *Anabaena sp. PCC* 7120 (Mulligan *et al.*, 1994), one gene (*rbpA*) in *Chlorogleopsis sp.* PCC 6912 (Mulligan *et al.*, 1994), two genes (*12rnp1*, *12rnp2*) in *Synechococcus sp.* PCC 6301 (Sugita and Sugiura, 1994) and two genes (*rbpA*, *rbpB*) from *Synechococcus* 7942 (Mulligan and Belbin, 1995).

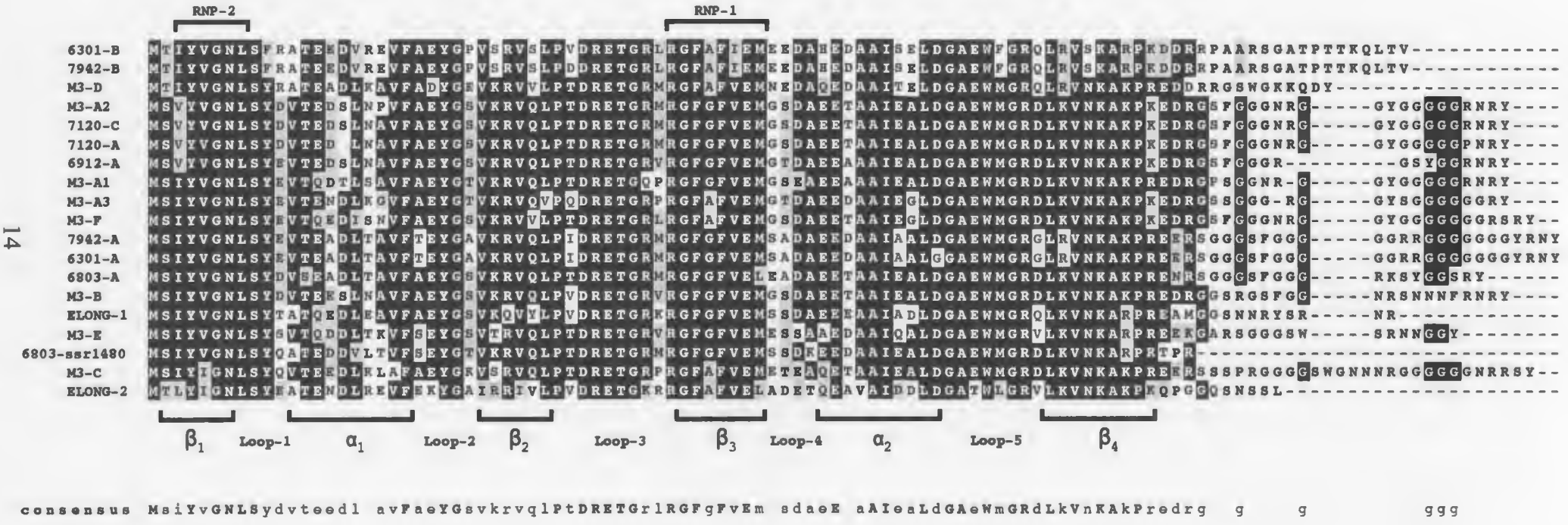
The cyanobacterial RNA-binding proteins are the smallest RRM-domain containing RNA-binding proteins characterized to date. Unlike RNA-binding proteins of eukaryotes which can contain multiple copies of the RRM module, all cyanobacterial

RNA-binding proteins contain one RRM. In cyanobacteria, there are two types of RNA-binding proteins termed type I and type II (Maruyama and Sato, 2001) which differ in numerous respects most notably being the number of amino acid residues in the two types.

1.4.2.1 Type I cyanobacterial RNA-binding proteins

The type I cyanobacterial RNA-binding proteins range in size from 83 to 110 amino acids in length, they all contain one N-terminal RRM with or without a short C-terminal auxiliary domain that, in many cases, is glycine-rich. A comparison of 17 Type I cyanobacterial RNA-binding proteins (Figure 1.2) illlustrates the high degree of conservation of amino-acid residues throughout the RRM. In particular, conservation of amino acid residues in loops 1 and 3, regions implicated in RNA sequence selection of the RRM (Allain *et al.*, 1996), is high. With exception of position 3 of loop 1, and positions 2 and 9 of loop 3, the loops are highly conserved suggesting that they could not be the regions responsible for nucleic acid sequence selectivity unless they recognize the same RNA sequence. From the alignment, α_1, loop 4 and the C-terminal domain could be the most important regions that confer RNA sequence specificity due to the degree of sequence heterogeneity at these positions. As evident in the alignment, the auxiliary domain is not highly conserved however most proteins contain a glycine-rich C-terminal auxiliary domain, some have an auxiliary domain rich in both asparagine and glycine (Av-RbpB, SyEl-tsl2255, Av-RbpE, Av-RbpD) or have no auxilliary domain at all (*Synechocystis* 6803 ssrl1480) (Figure 1.2).

Figure 1.2. Amino acid sequence alignment of 17 cyanobacterial RNP-type I RNA-binding proteins. Sequences were aligned using the Clustal W 1.8 multiple alignment program accessed at the Baylor College of Medicine website and converted into a presentable form by use of the Boxshade 3.21 program (K. Hoffman and M. Baron, unpublished) accessed at the European Molecular Biologists Swiss node website (www.ch.embnet.org). Residues present in 70% or greater at each position in the alignment are shown in reverse type. Conservation of similar amino acids at a given position are shaded. Amino acid sequences are as follows: RbpA (7942-A) (Belbin, 1999), RbpB (7942-B) from *Synechococcus* 7942, 12RNP1 (6301-A) and 12RNP-2 (6301-B) from *Synechococcus* 6301, RbpA1 (M3-A1), RbpA2 (M3-A2), RbpA3 (M3-A3), RbpB (M3-B), RbpC (M3-C), RbpD (M3-D), RbpE (M3-E) and RbpF (M3-F) from *Anabaena variabilis* M3, RbpA (6912-A) from *Chlorogleopsis* 6912, RbpA (6803-A) and gene product ssr1480 from *Synechocystis* 6803, and proteins encoded by genes tsl2255 (Elong-1) and tsr1443 (Elong-2) from *Thermosynechococcus elongatus BP-1.* An amino acid consensus sequence is given below the alignment, residues conserved in all 17 protein sequences are capitalized and in bold-type, residues not completely conserved yet present in greater than 50% of sequences are given in normal type and positions that do not have a particular amino acid residue in greater than 50% of sequences are blank.

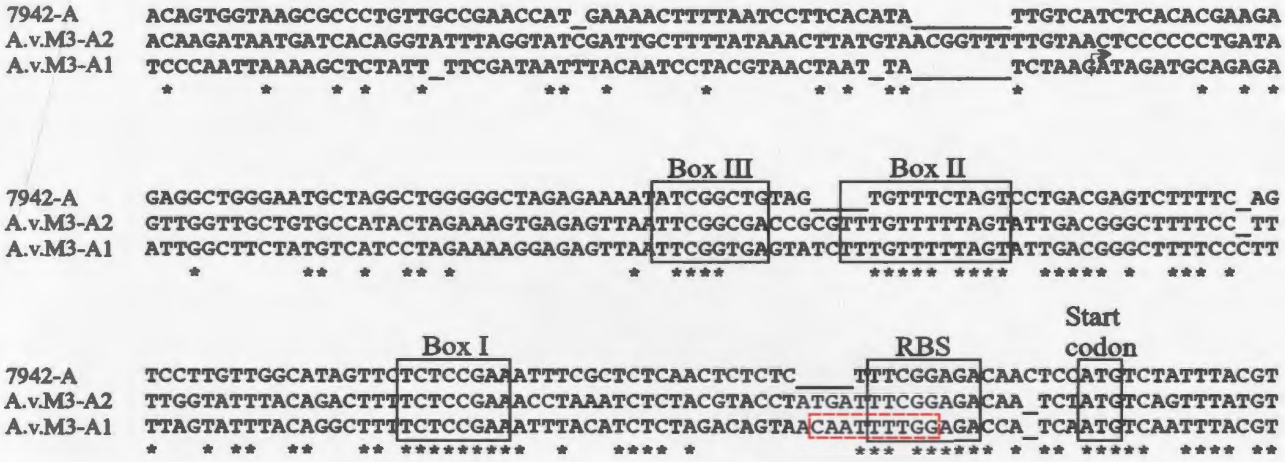

RNP-2
RNP-1
6301-B
7942-B
M3-D
M3-A2
7120-C
7120-A
6912-A
M3-A1
M3-A3
M3-F
7942-A
6301-A
6803-A
M3-B
ELONG-1
M3-E
6803-ssr1480
M3-C
ELONG-2
β1
Loop-1
α1
Loop-2
β2
Loop-3
β3
Loop-4
α2
Loop-5
β4
consensus MsiYvGNLSydvteedl avFaeYGsvkrvqlPtDRETGrlRGFgFvEm sdaeE aAIeaLdGAeWmGRdLkVnKAkPredrg g g ggg

The function of type I RNA-binding proteins in the cyanobacterial cell is not known, however, RNA-binding proteins that contain a glycine-rich C-terminal domain could be involved in the cold-shock response. It has been shown that the synthesis of some type I RNA-binding proteins increased upon introduction of a cold temperature stimulus. Examples include: RbpA in *Synechococcus* 7942 (Mulligan and Belbin, unpublished), 12RNP1 in *Synechococcus* 6301 (Sugita and Sugiura, 1994) RbpA1, RbpA2, RbpB, RbpC (Sato, 1995) and RbpE and RbpF (Maruyama *et al.*, 1999) in *Anabaena variabilis* M3, RbpA1, RbpA2, RbpB, RbpC, RbpE, RbpF in *Anabaena* 7120 (Ehira *et al.*, 2003) and Rbp1 in *Thermosynechococcus elongatus* (Ehira *et al.*, 2003).

1.4.2.2 Mechanisms of cold-shock induction of type I cyanobacterial RNA-binding proteins

Two mechanisms of cold-induced regulation of *rbp* genes have been described which may involve a set of conserved sequences termed box elements located in the 5' untranslated regions of the genes (Maruyama *et al.*, 1999) (Figure 1.3). It is not known exactly how these *cis*-acting regulatory regions are involved in regulation, however it is interesting that the Box I element and the RBS as well as Box III and Box I exhibit complementary base pairing. The box elements could be involved in transcriptional regulation by a set of *trans*-acting DNA-binding proteins. Three proteins isolated from *Anabaena variabilis* M3 grown at 38°C were able to bind to a DNA fragment encompassing nucleotides +1 to +151 in relation to the start point of the *rbpA1* transcript as evidenced by gel mobility shift analysis (Sato and Nakamura, 1998). Formation of DNA-protein complexes was not observed when protein fractions from cells grown at 22°C were used. In addition, deletion of nine nucleotides upstream of the *Anabaena*

Figure 1.3. Alignment of the 5' untranslated region of 3 cyanobacterial *rbp* genes. The nucleic acid sequences of *rbpA* (7942-A) from *Synechococcus* 7942, and *rbpA1* (A.v.M3-A1) and *rbpA2* (A.v.M3-A2) from *Anabaena variabilis* M3 were aligned using the Clustal W 1.8 multiple alignment program accessed at the Baylor College of Medicine website and converted into a presentable form by use of the Boxshade 3.21 program (K. Hoffman and M. Baron, unpublished) accessed at the European Molecular Biologists Swiss node website (www.ch.embnet.org). Complete nucleotide conservation is indicated by an asterisk under the alignment. The cold-shock inducible cyanobacterial *rbp* gene conserved box elements I, II and III (Maruyama *et al.*, 1999) are outlined by a box. The experimentally determined transcriptional start site of the *Anabaena variabilis* M3 *rbpA1* gene (Sato and Nakamura, 1998) is indicated by an arrow. In addition, the nine base sequence implicated as necessary for cold-shock induction of the *Anabaena variabilis* M3 *rbpA1* gene (Sato and Nakamura, 1998) is identified by a dotted box.


7942-A ACAGTGGTAAGCGCCCTGTTGCCGAACCAT_GAAAACTTTTAATCCTTCACATA_____TTGTCATCTCACACGAAGA
A.v.M3-A2 ACAAGATAATGATCACAGGTATTTAGGTATCGATTGCTTTTATAAACTTATGTAACGGTTTTTGTAACTCCCCCCTGATA
A.v.M3-A1 TCCCAATTAAAAGCTCTATT_TTCGATAATTTACAATCCTACGTAACTAAT_TA_____TCTAAGATAGATGCAGAGA
Box III
Box II
7942-A GAGGCTGGGAATGCTAGGCTGGGGGCTAGAGAAAATATCGGCTGTAG____TGTTTCTAGTCCTGACGAGTCTTTTC_AG
A.v.M3-A2 GTTGGTTGCTGTGCCATACTAGAAAGTGAGAGTTAATTCGGCGACCGCGTTTGTTTTTAGTATTGACGGGCTTTTCC_TT
A.v.M3-A1 ATTGGCTTCTATGTCATCCTAGAAAAGGAGAGTTAATTCGGTGAGTATCTTTGTTTTTAGTATTGACGGGCTTTTCCCTT
Start
codon
Box I
RBS
7942-A TCCTTGTTGGCATAGTTCTCTCCGAAATTTCGCTCTCAACTCTCTC____TTTCGGAGACAACTCCATGTCTATTTACGT
A.v.M3-A2 TTGGTATTTACAGACTTTTCTCCGAAACCTAAATCTCTACGTACCTATGATTTCGGAGACAA_TCTATGTCAGTTTATGT
A.v.M3-A1 TTAGTATTTACAGGCTTTTCTCCGAAATTTACATCTCTAGACAGTAACAATTTTGGAGACCA_TCAATGTCAATTTACGT

variabilis M3 *rbpA1* gene within the RBS prevented the binding of these DNA-binding proteins and resulted in constitutive transcription as detected by *rbpA1-lacZ* fusion analysis (Sato and Nakamura, 1998) (Figure 1.3). Therefore, a set of DNA-binding proteins could be involved in preventing the transcription of *rbp* genes in conditions other than cold-shock.

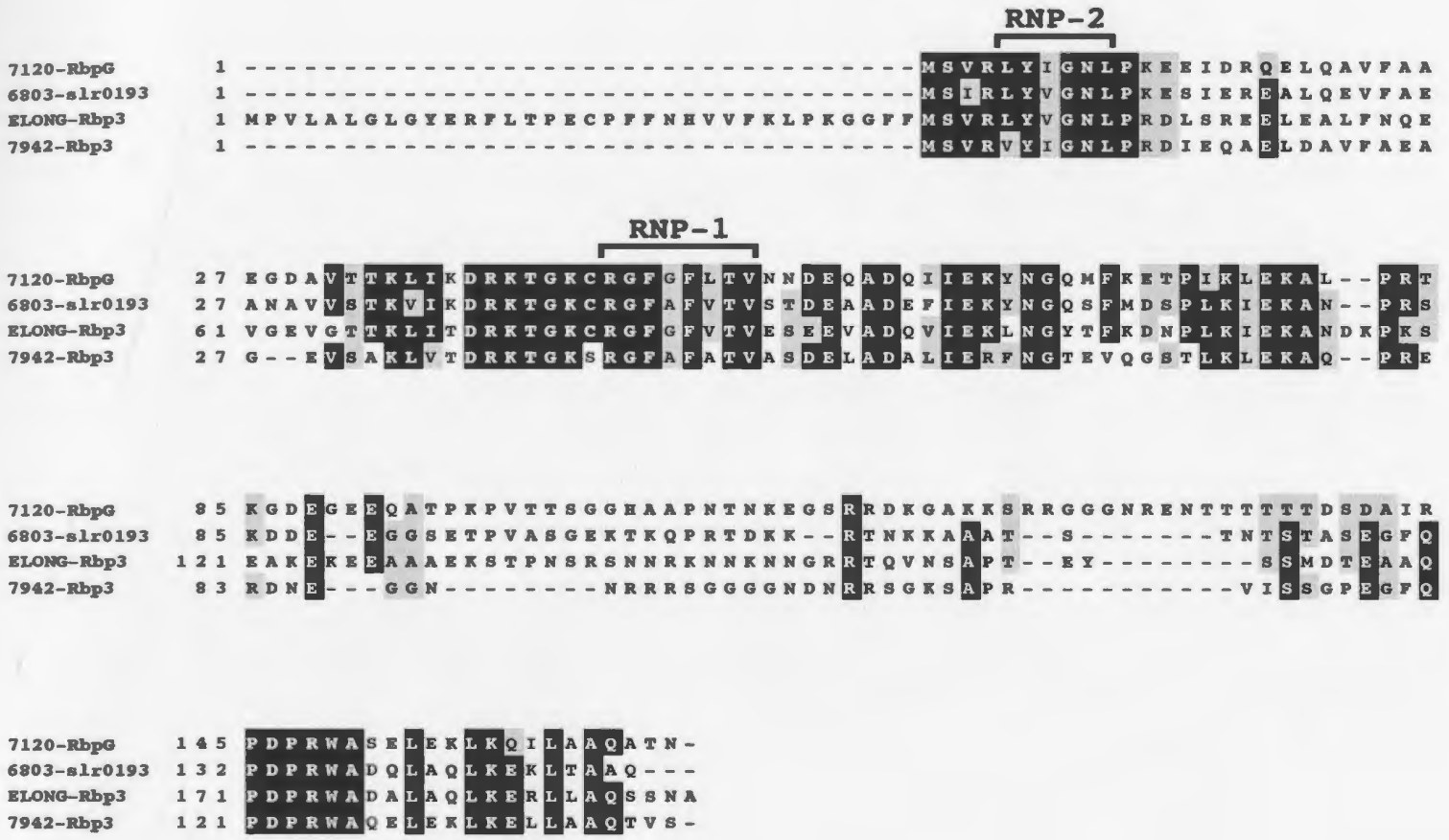
The other proposed mechanism of cold-induced regulation of *rbp* gene expression involving the box elements of the 5'UTR involves mRNA stability. The half-life of the *rbpA1* transcript in *Anabaena variabilis* M3 increases substantially (greater than four fold) upon shifting the cell culture to conditions of cold-shock in the presence of rifampicin (Sato and Nakamura, 1998). The mechanism of RNA half-life regulation is not known; however, it could involve putative RNA hairpins generated between box elements and the RBS in the 5'UTR. As discussed below in section 1.6, the conserved 5'UTR of cyanobacterial *rbp* genes could be responsible for increased stability at low temperature and thereby function as the main regulatory mechanism of RbpA expression.

It has been shown that the expression of type I RNA-binding proteins that do not contain a glycine-rich domain is not cold-induced. Examples include RbpB in *Synechococcus* 7942 (Mulligan and Belbin, unpublished), 12RNP2 in *Synechococcus* 6301 (Sugita and Sugiura, 1994) and RbpD in *Anabaena variabilis* M3 (Sato, 1995). Interesingly, the 5'UTR of genes *rbpB* in *Synechococcus* 6301 and *rbpD* in *Anabaena variabilis* M3 do not contain the box elements implicated in cold-regulated expression of the cold-shock genes (Maruyama *et al.*, 1999).

1.4.2.3 Type II cyanobacterial-RNA-binding proteins

The second type of cyanobacterial RNA-binding proteins are the type II proteins (Maruyama and Sato, 2001). These proteins are significantly larger than type I RNA-binding proteins. Open reading frame (ORF) detection analysis of genomic sequence data from *Anabaena* 7120, *Synechocystis* 6803 and *Thermosynechococcus elongatus BP-1* detected single type II RNA-binding proteins encoded in the genomes of each organism, with sizes of 166, 151 and 193 amino-acids respectively. A protein BLAST search of the draft sequence of *Synechococcus* 7942 (described in section 2.20) using the slr0193 protein of *Synechococcus* 6803 as the query sequence identified an additional homologous protein encoded by gene 1105 (now designated ser2048, refer to table A1). Amino-acid sequence alignment (Figure 1.4) of the four proteins reveals that the RNP-1 and RNP-2 modules are highly conserved, however, a number of differences exist with the type I RNA-binding proteins (Figure 1.2). In the RNP-2 module, the first residue which is isoleucine in type I RNA-binding proteins is replaced by leucine or valine in type II RNA-binding proteins. In the RNP-1 module, glutamate and methionine residues at postions 7 and 8 in type I RNA-binding proteins are replaced by tyrosine and valine in type II proteins. The most notable difference between type I and II RNA-binding proteins is the absence of a glycine or glycine/asparagine C-terminal auxiliary domain in type II RNA-binding proteins. However, type II RNA-binding proteins do exhibit amino acid conservation near the C-terminus with the existence of an absolutely conserved six amino acid sequence PDPRWA, and complete conservation of three leucine residues, one lysine residue and one alanine residue. The purpose of this conserved C-terminal region is unknown; an attempt to identify similar motifs of known function by comparative amino

Figure 1.4. Amino acid sequence alignment of 4 cyanobacterial RNP-type II RNA-binding protein sequences. The proteins are as follows: that encoded by gene 1105 from *Synechococcus* 7942 (7942-p.1105), Rbp3 from *Thermosynechococcus elongatus BP-1,* RbpG from *Anabaena* 7120, and the protein encoded by gene slr0193 from *Synechocystis* 6803. Alignments were generated as described in figure 1.2. Identical residues are shown in reverse type, residues that are not completely conserved yet similar are shaded.


RNP-2
RNP-1
7120-RbpG
6803-slr0193
ELONG-Rbp3
7942-Rbp3

acid analysis (Interpro database accessed at www.ebi.ac.uk) failed to detect homology to any characterized amino acid motif.

The function of the type II RNA-binding proteins in the cyanobacterial cell has not been determined. However, Northern blot and immunoblot analysis of *rbpG* and *rbp3* gene expression in *Anabaena* 7120 and *Thermosynechococcus elongatus* grown at a variety of temperatures revealed that the genes are constitutively expressed and the proteins are constitutively present (Ehira *et al.*, 2003). Efforts to construct a *Themosynechococcus elongatus rbpG* mutant were unsuccessful (Ehira *et al.*, 2003). These results suggested that type II RNA-binding proteins are absolutely required for cell viability, and are likely involved in a basic aspect of RNA metabolism as opposed to being involved exclusively in a stress response.

1.4.3 *Synechococcus* 7942 RNP-type RNA-binding proteins

To date three RNP-type RNA-binding proteins have been found in *Synechococcus* 7942: two type I RNA-binding proteins and one type II RNA-binding protein. The type I RNA-binding proteins RbpA and RbpB have been identified previously (Mulligan and Belbin, 1995) and I identified the third RNA-binding protein, a type II RNA-binding protein, by BLAST comparative analysis of the *Synechococcus* 7942 genome. RbpA expression is cold-inducible. Both the transcript and amount of this protein in the cell increase upon conditions of cold shock and is discussed in detail in the next section. In contrast, RbpB expression is not cold-inducible, expression remains constant regardless of growth temperature. The function of type II RNA-binding proteins, like those in *Synechococcus* 7942, is not known.

1.4.3.1 *Synechococcus* 7942 RNA-binding protein A

RNA-binding protein A (RbpA) from *Synechococcus* 7942 is a 107 amino-acid protein with an 83 amino-acid RRM domain at its N-terminus and a 24 amino-acid auxilliary domain rich in glycine residues at its C-terminus. The RRM of RbpA is very highly conserved in comparison to other cyanobacterial type I RNA-binding proteins (Figure 1.2). The RNP-1 and RNP-2 sequences are completely identical to the type I consensus sequence. The cyanobacterial protein that is most homologous in sequence to RbpA is 12RNP1 from *Synechococcus* 6301 which is nearly identical. The only difference occurs at postion 63 where RbpA has an aspartic acid residue and 12RNP1 has a glycine residue.

The function of RbpA, like the other cyanobacterial RNA-binding proteins is not known. RbpA has been shown to bind preferentially to RNA sequences solely composed of guanine or uracil nucleotides (Belbin, 1999 and figure 2.2), a characteristic feature of RRM-domain-containing proteins. The nearly identical 12RNP-1 protein from *Synechococcus* 6301 has also been shown to bind preferentially to poly(U) and poly(G) homopolymers (Sugita and Sugiura, 1994). RbpA could be involved in some way with regulating the photosynthetic apparatus as evidenced by a decreased growth rate and altered photosynthetic pigment composition of a *Synechococcus* 7942 strain in which the *rbpA* gene had been interrupted via a streptomycin/spectinomycin cassette (Mulligan and Belbin, 1995). In *Synechococcus* 6301, *12rnp1* mRNA decreased gradually over 12 hours after transfer of the culture from light to dark (Sugita and Sugiura, 1994), thereby providing evidence that expression is light dependent.

RbpA could also be involved in the cold-shock response since transcription of *rbpA* is induced upon cold temperature (15°C) incubation (Belbin, 1999). 12RNP1 is also induced upon conditions of cold-shock: mRNA levels doubled in response to shifting the *Synechococcus* 6301 culture from 30°C to 20°C (Sugita and Sugiura, 1994). In *Anabaena variabilis sp.* M3, the *rbpA1* and *rbpA2* genes are induced upon shift of the cell culture from 38°C to 22°C (Maruyama *et al.*, 1999). Another indication of cold-induced regulation of RbpA is found in the sequence of the *rbpA* 5'UTR which contains regions of significant homology to the conserved box elements and RBS (Maruyama *et al.*, 1999) characteristic of type I cyanobacterial cold-induced genes. As shown in figure 1.3, the 5'UTR of *rbpA* contains a completely conserved ribosome binding site and Box I element. In addition, *rbpA* also contains a Box II element that is identical to that present in the 5' UTR of *Anabaena variabilis sp.* M3 genes *rbpA1* and *rbpA2* at nine of twelve nucleotide positions and a Box III element that is not as highly conserved, having homology at four of eight nucleotide positions. As with type I cyanobacterial RNA-binding proteins, expression of the *rbpA* gene is cold-regulated via an unknown mechanism likely associated with the 5' UTR of the gene, a process that could also involve the stability of the transcript (refer to section 1.6).

1.5 Convergent Evolution of the RRM and CSD

Interestingly, as mentioned in section 1.4.2, RNP-type RNA-binding proteins are not well conserved in prokaryotes other than cyanobacteria. Therefore the question arises, what protein performs the analogous function of the RNA-binding proteins in other prokaryotes such as *E. coli*? The most likely candidates thus far are the cold-shock proteins (Csps). These have significant structural similarities with the RRM but only

limited amino acid similarity. Csps are able to bind to both ssDNA and RNA via an amino acid motif known as the cold-shock domain (CSD) that contains conserved regions similar to the RNP-1 (Landsman, 1992) and RNP-2 (Schroder *et al.*, 1995) modules of the RRM. With the exception of the limited similarity of the RNP-1 and RNP-2 modules, the amino acid sequences are not homologous. However, conservation of protein tertiary structure has been observed. NMR analysis of the CSD containing CspA protein of *E. coli* revealed a $\beta_1\beta_2\beta_3\beta_4\beta_5$ secondary structure that when folded positioned the RNP-1 and RNP-2 conserved sequences on β-strands β_2 and β_3 respectively (Schindelin *et al.*, 1994; Newkirk *et al.*, 1994) just as is the case with the RRM. This result suggests that the two types of proteins function as RNA-binding proteins via a mechanism in which conserved aromatic amino acid residues located within a β-sheet scaffold interact with RNA (Graumann and Marahiel, 1996).

If RNA-binding proteins perform the same function in cyanobacteria as Csps in *E. coli*, investigation of the cold-shock mechanism of *E. coli* and the function of Csps in the cold-shock response may provide an understanding of a putative function of RNA-binding proteins in cyanobacteria.

1.6 The cold-shock response in *Escherichia coli*

In *E. coli*, the cold-shock proteins (CSPs) are involved in generating the cold-shock adaptation response. When the environmental temperature decreases, the cell exhibits a sharp decline in cell metabolism known as the acclimation phase. During this phase, initiation of translation in the cell is minimal (Jones and Inouye, 1996). To re-establish translation, the synthesis of two classes (I and II) of CSPs is induced and non-CSP protein synthesis is halted. The difference between class I and II CSPs is their

relative induction levels within the acclimation phase. The class I CSPs are virtually non-existent at 37°C but are induced at least 10 fold upon cold-shock. Examples include cold-shock proteins A, B, G (CspA, CspB, CspG), an ATP-dependent RNA helicase (CsdA), and ribosome binding factor A (RbfA). Class II CSPs are induced less than 10 fold, examples include RecA, DNA gyrase (GyrA) and Initiation factor-2 (IF-2) (reviewed by Thieringer *et al.*, 1998). Together the class I and II cold-shock proteins perform a variety of functions including the re-establishment of translation of non-CSP encoding mRNA by formation of a cold-adapted ribosomal complex requiring CSPs, RbfA, IF2, and CsdA. Another function of CSPs is the destabilization of RNA secondary structures introduced in conditions of low temperature, a process involving the Csd RNA helicases and Csp RNA chaperone family of proteins.

As a member of the cold-adapted ribosomal complex, CsdA functions as an RNA helicase. The RNA helicases are a group of RNA-binding proteins that unwind dsRNA into ssRNA through use of energy derived from ATP hydrolysis. This group of proteins is of importance in the cold-shock response because conditions of lower temperature (approximately 20°C) promote the formation of dsRNA, creation of which can prevent translation of the mRNA by blocking the ribosome binding site. There are three groups of RNA helicases known as the DEAD (Linder *et al.* 1989), DEAH and DEXH families, members of which are found in *E. coli*. These familial classifications are based on the sequence and arrangement of eight amino acid motifs within the primary structure of the protein (Gorbalenya *et al.*, 1993). During the cold-shock response, the amount of DEAD box RNA helicase CsdA increases dramatically (greater than 10 fold) and is proposed to

be involved in maintaining translation of many mRNAs including those that encode the *E.coli* Csp proteins (Jones *et al.*, 1996).

Destabilization of RNA secondary structures is also accomplished by the activity of Csp proteins. These proteins maintain normal metabolism of RNA by acting as RNA chaperones, proteins that destabilize RNA secondary structures (Jiang *et al.*, 1997). The function of Csps has not been completely elucidated however CspA, CspC and CspE have been shown to decrease transcription termination of the cold-shock inducible *E. coli metY-rpsO* operon (Bae *et al.*, 2000) *in vivo*. Also, CspE was shown to increase the stability of mRNA by binding to the 3' poly (A) tail of mRNAs thereby preventing degradation by nucleases polynucleotide phosphorylase and RNase E (Feng *et al.*, 2001). The functions of the Csp proteins likely overlap due to the finding that cellular viability under cold-shock conditions was compromised only by inactivation of four of nine *E. coli csp* genes (Xia *et al.*, 2001), therefore the nucleic acid binding specificity of the Csps could be limited. SELEX experiments (refer to section 1.8) did not identify a consensus sequence for CspA but did detect a degree of RNA binding specificity for CspB, CspC and CspE of UUUUU, AGGGAGGGA and AAAUUU sequences respectively. The latter proteins selected RNA sequences with binding affinity in the 10^{-6} to 10^{-5} M range (Phadtare and Inouye, 1999).

1.7 Function of the 5'UTR of *csp* and *rbp* genes

The similarity of the 5' UTR of *csp* and *rbp* genes suggest that these genes could be regulated in a similar manner. Csp protein concentration in the cell during the cold-shock response is increased by up regulation of the *csp* transcription, increasing the stability of *csp* transcripts at low temperature, and by improving the efficiency of

translation (Gualerzi *et al.*, 2003). Analysis of the 5' UTR of *E. coli cspA* mRNA revealed a long 159 base untranslated region (Tanabe *et al.*, 1992) which is common to several *csp*-family cold-shock inducible genes including *cspB* (Etchegaray *et al.*, 1996), *cspG* (Nakashima *et al.*, 1996) and *cspI* (Wang *et al.*, 1999) as well as *rbp* genes.

The most conserved region within the 5'UTR of cold inducible *csp* genes is a 13 base sequence known as the upstream box (UB) that in *cspA* is located 11 bases upstream of the Shine-Dalgarno sequence. This sequence is implicated as a *cis*-acting element that regulates translation efficiency of the mRNA by providing an additional 16S rRNA-binding site upstream of the Shine-Dalgarno sequence for the cold-adapted ribosome (Yamaka *et al.*, 1999). Comparison of the 5'UTR of *cspA* and *cspB* to that of *rbpA* revealed that the highly conserved upstream box of the *csp* genes overlapped that of the highly conserved Box I element of *rbpA*. As shown in figure 1.5, the Box I element of the *rbpA* 5'UTR is nearly identical (8 of 9 nucleotides) to the 5' portion of the *cspA* UB. Therefore, this degree of homology implies that like *csp* genes, regulation of *rbp* cold-shock expression could exist at the level of translation efficiency mediated by the Box I element of the 5'UTR.

The *csp* 5'UTR regulates, in addition to translation efficiency, expression at the level of mRNA stability. Upon deletion of the first 80 bases of the *cspA* transcript, the half-life of the transcript increased significantly at both high and low temperature (Goldenberg *et al.*, 1996). Also, deletion of the entire 5'UTR of the *cspA* transcript resulted in constitutive expression (Fang *et al.*, 1997). Stability of mRNA mediated by the 5'UTR is also a regulatory mechanism of *rbp* expression. In a paper by Sato and Nakamura, (1998) the stability of *A. variabilis* M3 *rbpA1* mRNA was dependent upon

Figure 1.5. Alignment of DNA sequences (coding strand) from the 5' untranslated region of three prokaryotic cold-shock regulated genes. **(A)** The nucleotide sequence alignment of *rbpA* (7942-A) from *Synechococcus* 7942 and *cspA* (K12-cspA) and *cspB* (K12-cspB) from *Escherichia coli* K12.. The location of the ATG start codon in the coding strand of each sequence is indicated. **(B)** Important features of the sequences contained in the alignment such as the location of the Box I element implicated in cold-shock regulation of cyanobacterial *rbp* expression and the upstream box (UB) implicated in cold-shock induction of *Escherichia coli csp* genes are shown. In addition, the location of the Shine-Dalgarno (SD) sequence and start codon in the coding strand of each gene is indicated. Identical residues are shown in reverse type.

A.


Start codon
7942-rbpA TGGCATAGTTCTCTCCGAAATTTCGCTCTCAACTCTCTCTTTCGGAGACAACTCCATG
K12-cspA -GCCATA----TCGCCGAAAGG-CACACTTAATTA----TTAAAGGTAATACACTATG
K12-cspB -ACCATT--TAT-GCCGAAAGG-----CTCAAG------TTAAGG--AATGTAGAATG
consensus g CAT T CCGAAA CT AA TT gG A a ATG

B.


Box I
SD
Start codon
7942-rbpA TGGCATAGTTCTCTCCGAAATTTCGCTCTCAACTCTCTCTTTCGGAGACAACTCCATG
UB
SD
K12-cspA -GCCATA----TCGCCGAAAGG-CACACTTAATTA----TTAAAGGTAATACACTATG
UB
SD
K12-cspB -ACCATT--TAT-GCCGAAAGG-----CTCAAG------TTAAGGA--ATGTAGAATG

temperature. The half-life of the transcript was only 4.6 minutes at 38°C and increased to 18.5 minutes at 22°C. In addition, deletion of a stretch of nine nucleotides located from 10 to 19 bases upstream of the initiation codon partially within a putative ribosome binding site (Figure 1.3) (Maruyama *et al.*, 1999) resulted in constitutive expression of the transcript in the cell independent of temperature. This evidence suggests that expression of both *csps* and *rbps* is negatively regulated by the 5'UTR of the transcript in conditions other than cold-shock.

1.8 Identification of the RNA-binding specificity of RbpA by genomic SELEX

Although similarities with the Csp family of proteins provides an intriguing look into a potential function of RbpA, the function of RbpA can be further characterized by investigation of the nucleic acid specificity of the protein. As RNA-binding proteins, these proteins bind to stretches of nucleic acid sequence in either a specific or a non-specific manner. In an attempt to identify the RNA-binding specificity of RbpA protein I employed a variant of the SELEX technique known as genomic SELEX (Singer *et al.*, 1997).

1.9 Genomic SELEX

SELEX (systematic evolution of ligands by exponential enrichment) is a method whereby the nucleic acid binding specificity of a protein can be determined. This technique, independently and simultaneously developed by Tuerk and Gold (1990) and Ellington and Szostak (1990), uses partially degenerate DNA libraries generated by synthetic combinatorial chemistry to identify nucleic acid aptamers recognized by nucleic acid binding proteins. The technique involves multiple rounds of the same basic steps. First, the protein of interest is incubated with a library of nucleic acid molecules. Those

that are bound by the protein are captured and retained, those not bound by the protein are removed. Nucleic acids retained are amplified and become the pool of molecules used in the next round of selection. In this way, following multiple rounds of selection, only those molecules bound by the protein remain. Use of this technique has expanded rapidly, since its advent in the early 1990's. The use of SELEX has grown to the degree that construction of two aptamer databases, SELEX_DB (Ponomarenko *et al.*, 2000) and Aptamer Database (Lee *et al.*, 2004), have been developed.

The SELEX technique has been used to identify nucleic acid aptamers recognized by a multitude of proteins including bacteriophage T4 DNA polymerase (Tuerk and Gold, 1990), and HIV type 1 Rev protein (Jensen *et al.*, 1995). In addition to identifying nucleic acid-protein interactions, SELEX has been extended to identify aptamers recognized by nucleotide cofactors (Burke and Gold, 1997) and to isolate RNA with enzymatic activity (ribozymes) (Bartel and Szostak, 1993).

In regular SELEX, huge libraries (10^{15} unique sequences) of synthetically synthesized nucleic acids are required to generate the few sequences potentially recognized by a protein. The length of the nucleic acid sequence that can be varied in order to identify the binding specificity of the protein is restricted due to the enormous number of possible combinations that exist once the length of sequence is more than minimal.

A variation of the SELEX technique known as genomic SELEX, developed by Singer (1997), utilizes a library of DNA molecules that contain genomic nucleic acid sequences from the organism of interest from which the protein in question was isolated. This technique has been employed in the identification of aptamers recognized by

proteins such as the bacteriophage MS2 coat protein (Shtatland *et al.*, 2000) and human TAP(NXF1) (Zolotukhin *et al.*, 2001) as well as splicing substrates of the *Drosophila* pre-mRNA splicing factor B52 (Kim *et al.*, 2003). The use of genomic sequences provides an advantage over the use of the synthetically constructed libraries in terms of the size of the library and the information that can be generated from the analysis.

The genomic SELEX technique utilizes a small library (10^6 unique sequences) that contains all sequences that are potentially recognized by the protein because the library is composed of DNA sequences from the genome of the organism itself. In addition, the genomic SELEX technique allows one to select the length of the variable portion of the sequence that is used to identify nucleic acid binding specificity. Since the library is composed of genomic sequences, the location in the genome of any sequence selected by the protein is identifiable. This information in turn, can provide an insight into the function of a nucleic acid binding protein by identifying the gene or transcript that is bound by the protein. The one disadvantage of the genomic SELEX technique however is construction of the library. Since genomic nucleic acid sequences are used as a template, the library cannot be constructed synthetically, instead a lengthy protocol that involves the copying of genomic DNA sequences is required to construct a genomic SELEX library, a process described in chapter 3.

1.10 Goals

The initial goal of this study was to create a *Synechococcus* 7942 genomic SELEX library that sufficiently represented the entire genome. The library was created by a series of steps (Singer *et al.*, 1997) involving the copying of genomic DNA into small dsDNA fragments. The validity of the library was verified by determining the representation of a

portion of the *rbpB* gene in the library. The second goal was to use the library to determine the RNA-binding specificity of RbpA. The genomic SELEX process consisted of multiple rounds of selection under conditions of increasing stringency to select RNA molecules bound by RbpA. Numerous controls were utilized in an effort to provide evidence that selection of RNA was mediated solely by RbpA. The final goal of this study was to analyze the RNA sequences selected by genomic SELEX and to identify the genes that encode or are adjacent to these RNA sequences.

Materials and Methods

2.1 Primers, plasmids and strains

Oligonucleotide primers used in construction and analysis of the *Synechococcus* 7942 genomic SELEX library were purchased from Cortec Laboratories Inc. (Kingston, Ontario) and are listed in Table 2.1 (refer to page 41). Cloning procedures utilized the plasmid pUC19 (Yannisch-Perron *et al.*, 1985) and the *E. coli* strain JM109 (Yannisch-Perron *et al.*, 1985). A prokaryotic expression vector construct containing the *rbpA* gene (courtesy of Tom Belbin) in *E. coli* BL21 (DE3) pLacI was used for *in vivo* production of RbpA.

2.2 Purification of DNA

DNA generated by PCR amplification or by restriction enzyme digestion was purified by one of two methods: phenol/chloroform extraction followed by ethanol precipitation, and column chromatography via the Qiaquick DNA purification column (Qiagen).

A sample of DNA (10uL to 500uL) was extracted once with an equal volume of phenol and once with an equal volume of chloroform, using a centrifugation step of 14,000 rpm for 1 minute to separate phases. Following chloroform extraction, ½ volume 7M ammonium acetate (pH 7.5) and 3 volumes 100% ethanol were added to the aqueous phase. DNA was precipitated by incubation at -70^{o}C for 1 hour and pelleted by centrifugation at 14,000 rpm for 15 minutes. The pellet was washed with 80% ethanol, dried for approximately 10 minutes at room temperature, and resuspended in TE buffer (10mM Tris pH 7.5; 1mM EDTA).

For Qiaquick column chromatographic purification, 5 volumes buffer (composition proprietory to manufacturer) were added to the DNA sample. The solution was applied to the column and passed through by centrifugation at 13,000 rpm for 1

minute. The column was washed with 750uL buffer PE (composition proprietory to manufacturer) followed by centrifigation at 13,000 rpm for 1 minute. Once the filtrate was removed, the column was centrifuged again at 14,000 rpm for 1 minute to remove residual liquid on the column. The DNA was then eluted by addition of buffer EB (10mM Tris-HCl, pH 8.0) or distilled water followed by centrifugation at 14,000 rpm for 1 minute into a clean 1.5mL microfuge tube. DNA recovered was quantified by spectrophotometry utilizing wavelengths of 340nm and 260nm. The results obtained were then used in the following equation to determine the concentration of DNA.

Concentration (ng/uL) = (abs 260 nm -abs 340 nm) x 50 ng/uL x dilution factor

2.3 Isolation of plasmid DNA

Plasmid DNA was isolated from *E. coli* liquid cultures according to the Merlin mini-prep protocol described by Iyer (1994). Overnight cultures of *E. coli* grown in LB broth containing an appropriate antibiotic were pelleted by centrifugation and resuspended in 200uL of Cell Resuspension Solution (50mM Tris-HCl, pH 7.5; 10mM EDTA; 100ug/mL RNase A). Cellular contents were released upon addition of 200uL Cell Lysis Solution (0.2M NaOH; 1% SDS) and the resultant mixture was neutralized by the addition of 200uL Neutralization Solution (1.25M potassium acetate; 1.24M acetic acid) in the process precipitating protein and genomic DNA. The precipitated cellular components were pelleted by centrifugation at 14,000 rpm for 5 minutes and the supernatant was transferred to a clean tube containing 1mL of DNA-binding resin (Celite resin slurry in 7M guanidine hydrochloride). Following mixing by inversion, a 3cc syringe was used to force the slurry through a mini-column (Promega, Madison WI) which retained non-soluble matter and plasmid DNA. The column was then washed with 2mL Wash Solution (20mM Tris-HCl, pH 7.5; 200mM NaCl; 5mM EDTA; 50% Ethanol). Plasmid DNA was eluted by application of warm distilled water or TE,

incubation at room temperature for 1 minute, and centrifugation at 14,000 rpm for 1 minute.

2.4 Isolation of Genomic DNA from *Synechococcus* 7942

A 500mL culture of *Synechococcus* 7942 was grown in BG-11 liquid medium (Rippka *et al.*, 1979) at room temperature with constant illumination for approximately 2 weeks. Genomic DNA was isolated according to the protocol described by Golden (1997). Briefly, cells were harvested by centrifugation (5000 rpm, 5 min) and resuspended in 2.5mL buffer TE_{100} (100mM Tris-HCl, pH 8.0; 100mM EDTA). To the resuspended cells, 125uL 20% SDS, 50uL 1M DTT, 2.5mL phenol-chloroform-isoamyl alcohol (25:24:1) and 2.5mL autoclaved glass beads (150 - 212 microns in size) were added. Cells were ruptured by 3 rounds of vortexing at top speed for 3 minutes followed by incubation on ice for 30 seconds. Non-soluble material in the mixture was pelleted by centrifugation (10000 rpm, 10 min) and discarded. The supernatant was transferred to a clean 25mL Corex tube and was extracted with an equal volume once with phenol, once with phenol-chloroform-isoamyl alcohol (25:24:1) and once with chloroform-isoamyl alcohol (24:1). The phases were separated during each extraction by centrifugation (10000 rpm, 10 min). Nucleic acids were precipitated from solution by addition of one half volume 7.5M ammonium acetate (pH 7.5) and two and one half volumes 100% ethanol. The mixture was incubated on ice for 30 minutes followed by centrifugation (10000 rpm, 10 minutes) to pellet nucleic acids.

The nucleic acid pellet was resuspended in 450uL TE and transferred to a 1.5mL microfuge tube. The solution was purified further by successive extractions with equal volumes of phenol, phenol-chloroform (1:1), and chloroform. RNA was removed from the aqueous solution by addition of one half volume 7.5M ammonium acetate (pH 7.5), incubation on ice for 30 minutes and centrifugation (10000 rpm, 10 min.). DNA

remaining in the supernatant was precipitated by addition of 2 volumes of cold 100% ethanol. The precipitated DNA was spooled out of solution with a drawn pasteur pipette, dried in air, and resuspended in 100uL TE buffer. The concentration of DNA was determined spectrophotometrically at 260 nm, purity was assessed by comparison of absorbance at 260 nm to that at 340 nm.

2.5 Purification of RNA

Following transcription reactions the DNA used as the template was degraded by the addition of RQ1 DNase (Promega) to a final concentration of 0.06 - 0.08 units/uL in RQ1 DNase buffer (40mM Tris-HCl, pH 8.0; 10mM $MgSO_4$; 1mM $CaCl_2$). The resultant solution was incubated at 37°C for 1 hour.

Following DNase digestion, RNA was purified using an RNeasy purification kit (Qiagen). For every 100uL of DNase digestion mixture, 350uL of buffer RLT (composition proprietory to manufacturer) containing 0.01% 2-mercaptoethanol and 250uL 100% ethanol were added and mixed by inversion. The sample was then applied to an RNeasy minicolumn and centrifuged at 14,000 rpm for 15 seconds. Following centrifugation, the filtrate was emptied, 500uL buffer RPE (composition proprietory to manufacturer) was added to the column and it was centrifuged at 14,000 rpm for 2 minutes. Again the filtrate was emptied and the column was centrifuged at 14,000 rpm for 1 minute to remove any residual solution. The RNA was then eluted by addition of 20-50uL Diethyl pyrocarbonate (DEPC)-treated distilled water followed by centrifugation at 14,000 rpm for 1 minute into a clean 1.5mL microfuge tube.

Purified RNA was quantified by spectrophotometry utilizing wavelengths of 340 nm and 260nm. The results obtained were then used in the following equation to determine the concentration of RNA.

Concentration (ng/uL) = (abs 260 nm -abs 340 nm) x 40 ng/uL x dilution factor

2.6 Restriction endonuclease digestion and cloning

DNA molecules containing a *BamHI* restriction endonuclease site were digested by addition of template to a solution containing a final concentration of 1 unit/uL *BamHI* restriction endonuclease (New England Biolabs) in 1X NEB buffer *BamHI* (10mM Tris-HCl, pH 7.9 at 25°C; 150mM NaCl; 1mM DTT; 10mM $MgCl_2$) with 100ug/mL Bovine Serum Albumin (BSA). The solution was then incubated at 37°C for 1 hour. Similarly, DNA fragments containing a *SacI* site were digested by addition of template to a solution containing a final concentration of 1 unit/uL *SacI* restriction endonuclease (New England Biolabs) in 1X NEB 1 buffer (10mM Bis-Tris-Propane-HCl, 10mM $MgCl_2$, 1mM DTT, pH 7.0 at 25°C) with 100ug/mL BSA. The solution was incubated at 37°C for 1 hour.

Plasmids and DNA fragments were purified prior to cloning as described in sections 2.3 and 2.2 respectively. When required following restriction endonuclease digestion, plasmid DNA was dephosphorylated by the addition of purified DNA to a solution containing 0.1 unit/uL Shrimp Alkaline Phosphatase (SAP) (Amersham) in 1X SAP buffer. The solution was then incubated at 37°C for 1 hour followed by a 15 minute incubation at 65°C to inactivate the enzyme.

DNA fragments were cloned into plasmid vectors by incubation of purified, restriction endonuclease digested fragment and vector in a solution containing a final concentration of 0.15 unit/uL T4 DNA ligase (Promega) and 1X ligation buffer (30mM Tris-HCl, pH 7.8; 10mM $MgCl_2$; 10mM DTT; 1mM ATP) for 3 hours at room temperature.

Ligation product was transformed into competent *E. coli* JM109 (Promega) by a heat shock procedure (Huff *et al.*, 1990). A portion of the ligation product was mixed with 50uL competent *E. coli* JM109 cells (>10^8 cfu/ug) and incubated on ice for 30 minutes. The next step was a heat shock incubation of the mixture at 42°C for 45

seconds. Cell recovery was promoted by the addition of 1mL LB broth followed by incubation at 37°C for 1 hour with shaking at 250 rpm. Transformed cells were grown overnight on LB/agar plates containing 100ug/mL Ticarcillin (Sigma) with 800ug 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside (X-gal) spread on the plate to enable detection of recombinant plasmid containing colonies by blue/white selection.

2.7 Agarose and urea/acrylamide gel electrophoresis, electrophoretic transfer of nucleic acids to membrane

Samples of dsDNA and RNA were analyzed by agarose gel electrophoresis. Nucleic acids were separated in agarose gels that contained from 0.8% to 2% agarose in 0.5X TBE buffer (1X TBE is 100mM Tris-Cl, pH 8.0; 100mM boric acid; 2.5mM EDTA) with 50ug/mL ethidium bromide at 5-10 volts/cm. Results were recorded using an Alpha Innotech Corporation ChemiImager 4400 Low light imaging gel documentation system. Nucleic acid size determination and densitometric analysis was performed with Optiquant Image Analysis Software (Packard).

Denaturing urea/acrylamide gel electrophoresis was used to analyze ssDNA samples. Gels were prepared that contained 8M urea, 4% acrylamide and 0.2% bisacrylamide in 1X TBE buffer. Once the gel solution was made, it was filtered through 1 mm Whatman filter paper and placed on ice. To induce polymerization, 10uL of TEMED and 20uL of 10% ammonium persulfate were added to the gel solution. The solution was then poured into the mini-Protean III gel casting system (Bio-Rad), allowed to set, then placed into the Mini-Protean III electrophoresis system. Prior to loading of samples, the gel was electrophoresed in 1X TBE buffer at 20-40 volts until the anode buffer chamber reached a temperature of approximately 45°C. The wells were then flushed out using a pasteur pipette, samples were loaded and electrophoresed at 400 volts (required approximately 8 minutes for tracking dye to reach the bottom of the gel).

Samples were prepared by addition of an equal volume of 2X urea loading buffer (4M urea; 10mM EDTA; 2.5mM Tris-Cl, pH 7.5; 0.25% (w/v) bromophenol blue) followed by incubation at 90°C for 5 minutes. Nucleic acids were detected by incubation of the gel in a 150mL 1X TBE buffer solution containing the SYBR Green II (Molecular Probes) dye (1:10,000 dilution) for approximately 30 minutes at room temperature in the absence of light. Visualization of the fluorescent dye was achieved by use of an activation wavelength of 302 nm, light emitted from the dye was filtered through the SYBR green II filter (blocks light of wavelength 500nm or lower) and captured by the ChemiImager 4400 gel documentation system.

Nucleic acids separated either by agarose or by urea/acrylamide gel electrophoresis were transferred to positively charged nylon membrane (Hybond) by electrophoretic transfer utilizing the mini-Protean III transfer apparatus (Bio-rad). Transfer was executed in 0.5X TBE buffer at 40 volts for approximately 5 hours at 4°C. Following transfer the membrane was dried at 80°C for 30 minutes. The efficiency of transfer was determined by SYBR Green II re-staining of nucleic acids that remained in the gel.

2.8 Detection of Digoxygenin (DIG)-labeled Nucleic acids

Following transfer of DIG-labeled nucleic acids (refer to section 2.13) to nylon membrane, the membrane was equilibrated by a brief wash in Buffer I (100mM Tris-HCl, pH 7.5; 150mM NaCl). Following equilibration, non-specific antibody-membrane interactions were prevented by a 30 minute incubation in Buffer 2 [Buffer 1 containing 0.5% (w/v) blocking reagent (Roche)]. The membrane was then briefly rinsed in Buffer 1. Detection of the DIG label was initiated by incubation in 10mL of a 150mU/ml solution of polyclonal sheep anti-DIG Fab-fragments conjugated to alkaline phosphatase (AP). The membrane was then rinsed twice for 15 minutes in Buffer 1 and once for 2 minutes in

Buffer 3 (100mM Tris-HCl, pH 9.5; 100mM NaCl; 50mM $MgCl_2$). Detection of alkaline phosphatase was accomplished by incubation of the membrane in colour solution [40mL Buffer 3; 0.4mM Nitroblue Tetrazolium salt (NBT); 0.4mM 5-bromo-4-chloro-3-indolyl phosphate (X-phos)] until coloured precipitate was evident. The colour reaction was stopped by incubation in 100mL of Buffer 4 (10mM Tris-HCl, pH 8.0; 1mM EDTA).

2.9 *In vivo* synthesis of RbpA

N-terminal histidine-tagged (six histidine residues) RbpA protein (H_6RbpA) was generated *in vivo* by induction of a prokaryotic expression vector construct containing the *rbpA* gene (courtesy of Dr. Tom Belbin) in *E. coli* BL21 (DE3) pLacI. An overnight culture (approximately 5mL) was added to a larger volume (200mL to 1L) of liquid LB broth containing 100ug/mL Ticarcillin and incubated at 37°C with shaking (250 rpm). When cell growth reached mid-log phase (A_{595} of 0.6-0.8) H_6RbpA protein production was induced with addition of isopropylthiogalactoside (IPTG) to a final concentration of 2mM. Following a 30 minute induction period, the cell culture was harvested by centrifugation at 6000 rpm for 10 minutes and stored at -70°C until the next step in the purification protocol was conducted.

Total protein extracts of cellular contents were generated by use of a non-denaturing protocol (Petty, 1987). Pellets were resusupended in 10mL of Resuspension Buffer [50mM Na_2HPO_4, pH 8.0; 500mM NaCl; 10% glycerol; 1mM phenyl methyl sulfonyl flouride (PMSF) (Sigma)], Triton X-100 was added to a final concentration of 0.1% and cells were sonicated in ice water for 10 intervals of 30 seconds followed by 30 seconds of recovery (to dissipate heat generated by sonication). To prevent protease activity, a 150X inhibitor cocktail was added composed of 1mg/mL pepstatin A (Sigma) and 1mg/mL leupeptin (Sigma). The resultant solution was incubated on ice for 15 minutes, followed by centrifugation at 15,000 rpm for 15 minutes to remove insoluble

cellular material. The supernatant (total protein extract) was removed and kept at 4°C until H_6RbpA was purified by column chromatography utilizing a Ni^{+2}-NTA spin column.

2.10 Purification of RbpA with Ni^{+2}-NTA spin columns

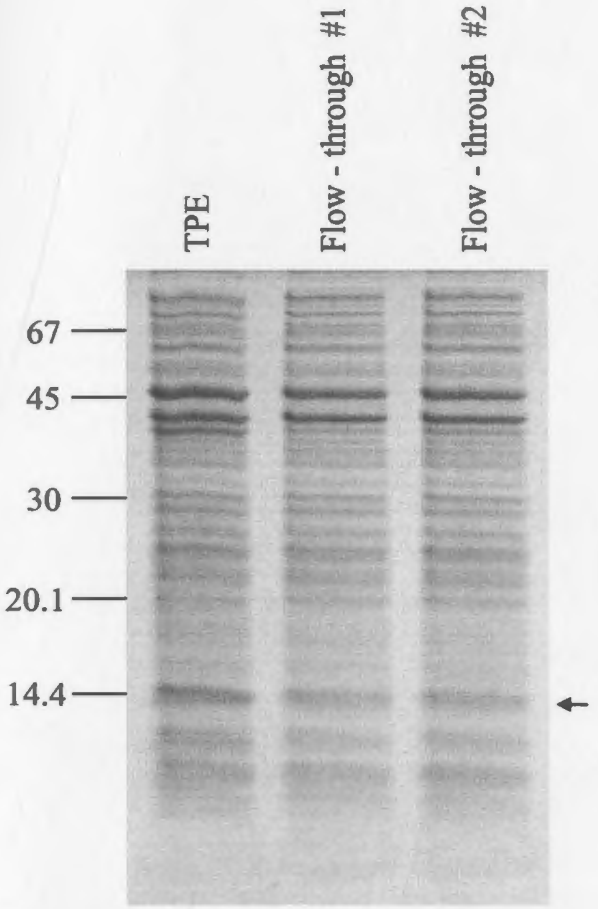
H_6RbpA protein was purified from other components of the total protein extract by use of a pre-made Ni^{+2}-NTA spin column (Qiagen). The entire procedure was performed at 4°C. The column was equilibrated by the addition of 600uL Resuspension Buffer (section 2.18) and spun at 2,000 rpm for 2 minutes. A volume of total protein extract (600uL) was applied to the column which was centrifuged again for 3 minutes at 1,000 rpm. The filtrate was applied to the column and centrifuged again at 1,000 rpm for 3 minutes. These two steps were repeated until 5mL of total protein extract had passed through the column twice. The spin column was then washed by addition of 600uL of Wash Buffer (50mM Na_2PO_4, pH 7.5; 300mM NaCl; 40mM Imidazole) and spun 2,000 rpm for 2 minutes, this process was repeated an additional 14 times to maximize the removal of contaminating proteins from the column. Proteins bound to the resin were eluted upon application of 100uL Elution Buffer (50mM Na_2PO_4, pH 7.5; 300mM NaCl; 250mM Imidazole) and centrifugation at 2,000 rpm for 2 minutes. The elution step was repeated 9 additional times. The effectiveness of this procedure was analysed by Tricine SDS-PAGE (section 2.17) (Figure 2.1).

2.11 Dialysis and Quantification of RbpA

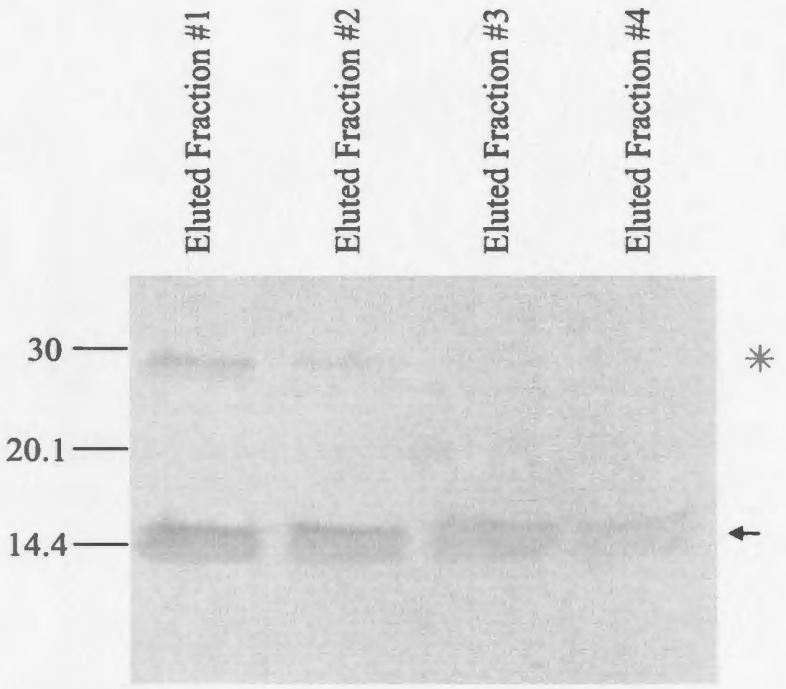
Eluted fractions containing RbpA protein detected by discontinuous Tricine SDS-PAGE were pooled together (approximately 2.5mL to 5mL) and placed in 10cm of snakeskin dialysis tubing (molecular weight cut-off of 3 kDa, Pierce). The tubing was placed in 500mL of DEPC-treated 1X SELEX RNA-binding buffer [SRBB (50mM Na_2HPO_4, pH 7.5; 50mM NaCl; 2.5mM $MgCl_2$; 0.5% Triton X-100; 40mM Imidazole)] and gently stirred overnight at 4°C. Following dialysis, the amount of RbpA was

Figure 2.1. Evaluation of the H_6RbpA purification protocol. **(A)** Tricine SDS-PAGE separation analysis of the total protein extract (TPE) following a 30 minute induction before and after the preparation had passed through a Ni^{+2}-NTA affinity chromatography spin column, H_6RbpA protein is indicated by an arrow. **(B)** Analysis of eluted fractions 1 to 4 generated upon elution of H_6RbpA from the Ni^{+2}-NTA affinity column with Elution buffer. The location of the H_6RbpA protein and a higher molecular weight contaminant are indicated by an arrow and an asterisk respectively. Gels were stained with Coomassie G-250 blue strain. The location of applicable low molecular weight standards (Pharmacia) are indicated at left.

A.


TPE
Flow - through #1
Flow - through #2
67
45
30
20.1
14.4

B.


Eluted Fraction #1
Eluted Fraction #2
Eluted Fraction #3
Eluted Fraction #4
30
20.1
14.4

quantified by comparison of the band intensities of samples of RbpA and lysozyme produced by either SYPRO orange or Coomassie blue staining. A standard curve based on lysozyme was used to quantify the concentration of RbpA (Figure 2.2A).

2.12 RNA homopolymer analysis of RbpA

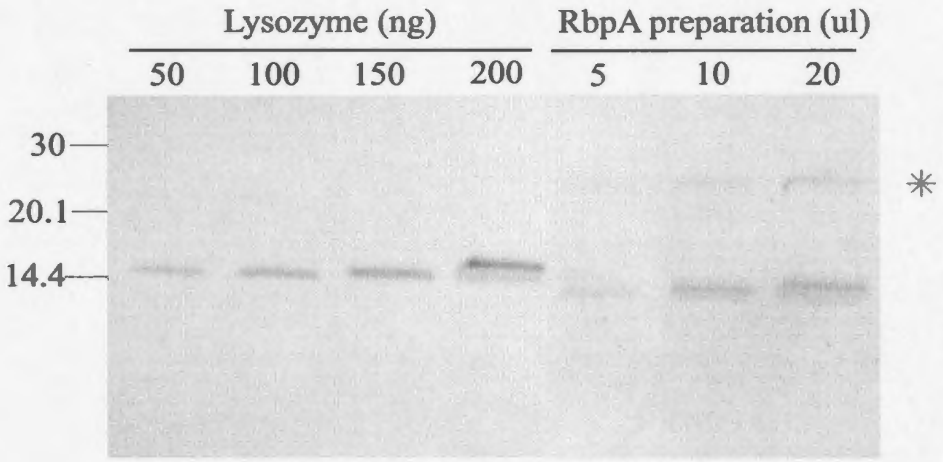
Binding analysis of RNA homopolymers was performed as described by Sugita and Sugiura (1994) and Belbin (1999). RNA homopolymer resins used in binding analysis were as follows: polyadenylic acid [poly(A) RNA] attached to cross-linked 4% beaded agarose (0.5 to 1.0mg of RNA per mL resin), polycytidylic acid [poly(C) RNA] attached to cross-linked beaded agarose (0.25 to 1.0mg RNA per mL resin), polyguanylic acid [poly(G) RNA] complexed to polyacrylhydrazido-agarose (1.5 to 4.5mg RNA per mL resin) and polyuridylic acid [poly (U) RNA] complexed to polyacrylhydrazido-agarose (0.1 to 1.0mg RNA per mL resin) (Sigma). RbpA protein (1ug) was incubated with 25ug of respective RNA homopolymers in 1mL of 1X SRBB and mixed by rotation at 4^{o}C for 10 minutes. The mixture was added to a mini-column (Promega) used to retain the polymer and protein bound to it. Non-bound protein was removed by washing with 1mL 1X SRBB containing 1mg/mL heparin, 1mL SRBB and 1mL dH_20. Removal of bound protein was achieved upon addition of 50uL Tricine sample buffer (100mM Tris-HCl, pH 6.8; 20% glycerol; 1% SDS; 0.02% Coomassie Blue G-250; 62.5 mM DTT) and centrifugation at 14,000 rpm for 20 seconds. Protein bound to the resin was detected by discontinuous Tricine SDS-PAGE as shown in Figure 2.2B.

2.13 SELEX: First and second-strand extension reactions

First-strand extension reactions consisted of 28ug of *Synechococcus* 7942 genomic DNA mixed with 54uM Aran primer (Table 2.1), 1X USB Klenow buffer (50mM Tris-HCl, pH 7.5; 10mM $MgCl_2$; 1mM DTT; 50ug/mL Bovine Serum Albumin), 300uM each dNTP and 135 Units of the Klenow fragment of *E. coli* DNA polymerase I

Figure 2.2. Quantification and RNA homopolymer binding analysis of H_6RbpA. **(A)** Tricine SDS-PAGE analysis of known amounts of Lysozyme (50ng, 100ng, 150ng and 200ng) used as a standard to determine the unknown concentration of H_6RbpA expressed and purified as described in sections 2.13 to 2.16. The location of a higher molecular weight protein contaminant in the H_6RbpA preparation is indicated by an asterisk. **(B)** RNA homopolymer binding activity of H_6RbpA was assessed by Tricine SDS-PAGE analysis of H_6RbpA eluted from columns containing RNA homopolymers as indicated. The location of low molecular weight standards (Pharmacia) are indicated at left.

A.


Lysozyme (ng)
RbpA preparation (ul)
50
100
150
200
5
10
20
30
20.1
14.4
*

B.

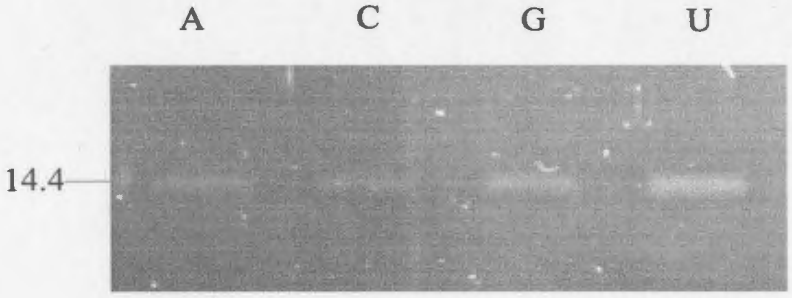

A
C
G
U
14.4

Table 2.1. Primers used in Construction and Analysis of the *Synechococcus* 7942 genomic SELEX library

Name	Length (bases)	Sequence	Tm (°C)*	Reference
Aran	25	5' AGGGAGGACGATGCGGNNNNNNNNN 3'	58.3	Singer *et al.*, 1997
Bran	25	5' TCCCGCTCGTCGTCTGNNNNNNNNN 3'	58.3	Singer *et al.*, 1997
T7A	37	5' GAAATTAATACGACTCACTATAGGGAGGACGATGCGG 3'	66.1	Singer *et al.*, 1997
T3B	41	5' CTAGCAATTAACCCTCACTAAAGGGACCCGCTCGTCGTCTG 3'	67.7	This work
B	16	5' TCCCGCTCGTCGTCTG 3'	44.8	Singer *et al.*, 1997
rbpB 1	25	5' CGCAGTTGGCGACCAAACCACTCAG 3'	58.3	This work
rbpB 2	30	5' AAACCACTCAGTGCCATTGAGCTCGCTGAT 3'	62.3	This work
T7A(BamHI)	41	5' AGACGGATCCAATACGACTCACCACAGGGAGGACGATGCGG 3'	67.7	Singer *et al.*, 1997
B(BamHI)	20	5' ATCTGGATCCTCGTCGTCTG 3'	52.3	Singer *et al.*, 1997
Elongated Forward	25	5' CGCCAGGGTTTTCCCAGTCACGACG 3'	58.3	This work
Elongated Reverse	28	5' GCGGATAACAATTTCACACAGGAAACAG 3'	60.9	This work
Forward Sequencing	17	5' ACGTTGTAAAACGACGG 3'		Sanger *et al.*, 1977

* Melting temperature calculated by Cortec Laboratories Inc. according to the formula :

$$Tm = 81.5^{\circ}C - 16.6(\log_{10}[Na^{+}]) + 0.41(\%\ G+C) - 0.63(\%\ \text{formamide}) - (600/L)$$

where L is the length of the oligonucleotide in base pairs

Table 2.2 Components of Polymerase Chain Reactions

	Template	Amount Template (ng)	First primer name	First primer conc (uM)	Second primer name	Second primer conc (uM)	Reaction volume (uL)	Annealing temperature (C)o
Construction of the genomic SELEX library	Second-strand Extension product	- *	T7A	4	B	4	20	61
	cDNA	-	T7A	5	T3B	5	50	61
Validation of the genomic SELEX library	Nested PCR step 1	180	T7A or T3B	5 5	rbpB1	5	20	51
	Nested PCR step 2	200	T7A(BamHI)	5	rbpB2	5	20	56
	cDNA	-	T7A	1.25	T3B	1.25	160	61
	detection of *rhpB* gene containing library molecules	200	T7A	10	rbpB1	10	20	51
Cloning and Sequencing	Cloning	-	T7A(BamHI)	5	B(BamHI)	5	20	46
	Analysis of plasmid inserts	~100	Reverse	5	Forward	5	20	52

*Concentration of template DNA was unknown

(*exo*+) (New England Biolabs) in a 150uL reaction volume. The genomic DNA and primer were denatured for 3 minutes at 93°C then incubated on ice for 5 minutes during which the remaining components of the reaction were added. The mixture was then incubated for 25 minutes at 25°C and 5 minutes at 50°C. Inactivation of the Klenow enzyme was achieved by the addition of 1.5uL 0.5M EDTA and incubation at 75°C for 10 minutes.

Second strand extension reactions were carried out in the same way as that of the first extension reactions except that the Bran primer (Table 2.1) was used at a concentration of 11.9uM.

Aran and Bran extension reactions incorporating Digoxygenin-labelled nucleotides were carried out in the same way as the first-strand extension reaction with the exception that the nucleotide concentrations were as follows : 70uM dATP, 70uM dCTP, 70uM dGTP, 45uM dTTP and 25uM DIG-dUTP. Incorporation of the digoxygenin label was detected following application or transfer of labeled ssDNA to a nylon membrane via a dot blot or electrophoretic transfer from an 8M Urea/4% acrylamide denaturation gel respectively (section 2.7).

2.14 SELEX: Removal of excess Aran primer by filtration

Removal of excess Aran primer (8.2kDa) was achieved by filtration with Microcon-10 filters (10kDa molecular weight cut-off). The entire first-strand extension reaction (150uL) was brought up to 500uL with TE buffer and added to the retentate end of the filtration device. The filter was then wrapped in paper towel to ensure a tight fit when placed in a rotor sleeve and in turn placed in a Sorvall SS-34 type rotor within a Sorvall RC-58 Superspeed Centrifuge and spun at 4100 rpm for 15 minutes. Approximately 300ul of the initial volume was removed as filtrate and discarded. The retentate was recovered by inversion of the retentate vial into a clean tube, followed by

centrifugation for 2 minutes at 4100 rpm. The entire process was repeated a second time with a clean Microcon 10 filtration device.

2.15 SELEX: Purification of second-strand extension product by electroelution

One-third the volume (100uL) of second-strand extension product was added to an equal volume of USB stop solution (95% formamide; 20mM EDTA; 0.05% bromophenol blue; 0.05% xylene cyanol) and electrophoresed using two 8M urea/4% polyacrylamide denaturing gels (20uL per well, 10 wells per gel). Three segments of the gel were excised that corresponded to regions 3-4 cm, 4-5 cm and 5-6 cm from the top of the gel. These segments were placed individually into snakeskin dialysis tubing (Pierce) (3000 Dalton molecular weight cut-off) containing 1mL 0.5X TBE and electroeluted for 3 hours at 100V. Following electroelution, the ssDNA separated in each gel fraction was purified by ethanol precipitation (section 2.2) and resuspended in 50uL TE buffer.

2.16 Amplification of DNA by Polymerase Chain Reaction

Polymerase Chain Reactions contained 1X Singer PCR buffer (10mM Tris-HCl, pH 9.0; 50mM KCl; 0.1% Triton X-100; 3mM $MgCl_2$) (Singer *et al.*, 1997), 50uM dNTPs and 0.1 - 0.2 units/uL Taq polymerase with an overlay of mineral oil. The remaining components such as the identity and amount of template DNA, identity and concentration of primers, volume of reaction and annealing temperature are given in Table 2.2. All reactions were conducted by the "hot start" method in a PTC-200 Peltier Thermal Cycler (MJ Research), whereby nucleotides and Taq polymerase were added following a 5 minute incubation period at 95°C to denature template and primers. Once all components were added, template DNA was amplified using 30 cycles with the following parameters: denaturation at 93°C for 30 seconds, annealing of primer to template at the temperature specified in Table 2.2 for 10 seconds, and extension of primer at 72°C for 45 seconds. Amplified products were purified by phenol/chloroform and

ethanol precipitated or by Qiaquick spin column chromatography (Qiagen) (section 2.2).

2.17 SELEX: Transcription

Two transcription reactions will be discussed in this section, the first pertains to transcription of the PCR amplified second-strand extension product involved in construction of the *Synechococcus* 7942 genomic SELEX library, the second pertains to transcription of PCR amplified cDNA used to create RNA in each round of SELEX.

Second-strand extension product was transcribed by a reaction that contained 2ug of the PCR product of the second-strand extension reaction amplified with primers T7A and T3B in 1X Gibco transcription buffer [40mM Tris-HCl, pH 8.0; 10mM $MgCl_2$; 20mM DTT; 25mM NaCl; 2mM Spermidine-$(HCl)_3$], 6.25mM NTPs and 20 units of T7 RNA polymerase in a final volume of 80uL. The reaction was incubated at 37°C for 2 hours, then 4°C overnight.

To generate RNA for a round of SELEX, transcription reactions were performed in duplicate; each contained 1.5ug of PCR amplified cDNA from the preceding round of SELEX in 1X NEB RNA polymerase buffer (40mM Tris-HCl, pH 7.9; 6mM $MgCl_2$; 2mM spermidine; 10mM DTT), 0.5mM each NTP, and 250 units of T7 RNA polymerase in a 500uL volume. The reaction was subsequently placed in either a thermocycler or water bath at 37°C for 14 to 18 hours. The resultant RNA was purified and quantified by RNeasy column chromatography and spectrophotometry respectively as described in section 2.5.

2.18 SELEX: Reverse-transcription

RNA was reverse-transcribed by the addition of RNA template to 1uM T3B primer (Table 2.1), 0.5mM each dNTP, 5units/uL Omniscript reverse transcriptase (Qiagen) in 1X Buffer RT (composition proprietory to manufacturer) and incubation at 37°C for 1 hour. To inactivate the reverse transcriptase, the reaction was incubated at

93°C for 5 minutes.

2.19 SELEX: Isolation of RbpA-RNA complexes

RNA was denatured by incubation at 95°C for 1 minute, then kept on ice for at least 10 minutes before use. The RNA-protein binding reaction utilized in rounds 1 to 10 of SELEX consisted of 1nmol (40ug, 10,000nM final concentration) denatured RNA and 0.05nmol (540ng, 500nM final concentration) H_6RbpA in a reaction volume of 100uL containing 1X SELEX RNA-binding Buffer (SRBB) (50mM Na_2HPO_4, pH 7.5; 50mM NaCl; 2.5mM $MgCl_2$; 0.5% Triton X-100; 40mM Imidazole). In rounds 11 to 14, components of the RNA-protein binding reaction were altered and consisted of 0.025nmol (1ug, 500nM final concentration) denatured RNA and 0.025nmol (27ng, 500mM final concentration) H_6RbpA in a reaction volume of 50uL containing 1X SELEX RNA-binding Buffer with High Salt (SRBBHS) (50mM Na_2HPO_4, pH 7.5; 150mM NaCl; 2.5mM $MgCl_2$; 0.5% Triton X-100). In addition to the RNA-protein binding reaction, a control reaction was performed for each round that contained all components with exception of the protein itself. Once all components were added, the reaction was incubated at room temperature for 1 hour with rotation (tube was inverted approximately 20/min).

After the 1 hour incubation period, 2uL of resuspended Ni^{+2}-NTA resin (5 to10mg/ml binding capacity of a 20 kDa protein, Qiagen) was added to both the control and RNA-protein reactions and incubated further at room temperature with rotation for 20 minutes. Protein-RNA complexes bound to the resin were deposited at the bottom of the tube by centrifugation at 1,000 rpm for 2 minutes and the resulting supernatant containing free RNA, protein or protein-RNA complexes not bound to the resin were discarded. The remaining solution containing the resin (approximately 50uL to 100uL) was washed by addition of 200uL 1X SRBB (rounds 1 to 10) or 200uL of SRBBHS (rounds 10 to 14)

followed by a 2 minute incubation at room temperature with rotation. The mixture was then centrifuged at 1000 rpm for 2 minutes, followed by removal of the supernatant. This process was repeated 4 additional times to maximize the removal of uncomplexed RNA. RNA-protein complexes were then eluted from the resin by addition of 100uL SELEX Elution Buffer (SEB) (50mM Na_2HPO_4, pH 7.5; 50mM NaCl; 2.5mM $MgCl_2$; 0.5% Triton X-100; 250mM Imidazole), incubation at room temperature for 5 minutes with rotation, centrifugation at 1000 rpm for 2 minutes and removal of the supernatant which was retained.

The RNA isolated in the presence and absence (control) of protein was purified using an RNeasy column and quantified by spectrophotometry as described in section 2.5.

2.20 Computational data analysis

The theoretical tertiary structure of RbpA was generated by use of the molecular viewer program Deepview (Guex and Peitsch, 1997) developed at the Swiss Institute of Bioinformatics (www.expasy.org/spdbv/). The structure was created by first identifying two proteins of known tertiary structure that have significant sequence homology with RbpA. These proteins; 1HD1A and 1HD0A (human heterologous nuclear ribonucleoproteins) were used as a scaffold onto which the sequence of RbpA was threaded to create the theoretical structure given in figure 1.1. Note that, although RbpA is 107 amino acids in length, the three-dimensional structure only encompasses residues 2-76 because the remaining amino acids were not recognized as possessing homology to any domain of known tertiary structure.

Comparison of RNA sequences identified by the SELEX procedure to those of other prokaryotes was accomplished by use of the BLAST and Gapped-BLAST programs (Altschul *et al.*, 1997) accessed at the European Biotechnological Institute (EBI) website www.ebi.ac.uk.

The alignment of RNA sequences was accomplished by use of the multiple sequence alignment program Clustal W (Jeanmougin *et al.*, 1998) accessed at the Baylor College of Medicine website http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html. Alignments were made visually suitable by use of the boxshade 3.21 program (K. Hoffman and M. Baron, unpublished) accessed at the European Molecular Biologists Swiss node website (www.ch.embnet.org).

Computational analysis to find homologous of both nucleic acid and protein sequences was performed using the FASTA program (Pearson, 1990). Protein sequences were also analyzed in terms of the presence of conserved amino acid residues and motifs and the likelihood that they belong to a group of conserved proteins by InterproScan (Zdobnov and Apweiler, 2001). Both programs were accessed at the EBI website www.ebi.ac.uk. Protein sequences within ORFs encoded in the *Synechococcus* 7942 genome were analyzed by the Joint Genome Institute using the program Clusters of Orthologous Groups (COGS) to determine the identity of *Synechococcus* 7942 proteins based on similarity to other known groups of proteins, this analysis is referred to numerous times in section 5.6.

A graphical version of the alignment used to define the RbpA cognate RNA consensus sequence was generated by the Weblogo program (Crooks *et al.*, 2004) accessed at the following website (http:://weblogo.berkeley.edu/).

RNA secondary structural determination and analysis was accomplished by use of the program mfold version 2.3 (Zuker *et al.*, 1999; Mathews *et al.*, 1999) as developed by Michael Zuker at the Rensselaer Institute and accessed at the website www.bioinfo.rpi.edu/~zukerm/.

2.21 Tricine SDS-PAGE analysis

Proteins were analysed by discontinuous Tricine SDS-PAGE according to the

protocol of Schagger and von Jagow (1987). An equal volume of protein was added to 2X Tricine Sample Buffer (200mM Tris-HCl, pH 6.8; 40% glycerol; 2% SDS; 0.04% Coomassie Blue G-250; 125mM DTT) and boiled 5 minutes. Prior to loading, the gel composed of a 3.8% acrylamide stacking gel (750mM Tris-HCl, pH 8.5; 0.075% SDS; 3.8% acrylamide; 0.12% bis-acrylamide) and a 16% acrylamide resolving gel (750mM Tris-HCl, pH 8.5; 0.075% SDS; 16% acrylamide; 0.5% bis-acrylamide; 10.5% Glycerol) was pre-run at 60 volts for 30 to 60 minutes in 1X Anode buffer (200mM Tris-HCl, pH 8.9) and 1X Cathode buffer (100mM Tris-HCl, pH 8.25; 100mM Tricine). The wells of the gel were flushed out following the pre-run, samples were loaded and gel was electrophoresed at 30 volts through the stacking gel and 60 volts through the resolving gel till the bromophenol blue dye front reached the bottom of the gel.

Following electrophoresis, gels were stained in 200mL 1X SYPRO Orange dye (Molecular Probes, 1X final concentration in 200mL 7.5% acetic acid diluted from a 10,000X concentrate) for 10 to 40 minutes, destained 30 minutes in 100mL 7.5% acetic acid and detected through a SYPRO orange filter using an excitation wavelength of 305 nm. Gels were photographed by the ChemiImager 4400 gel documentation system then incubated in a Coomassie blue solution overnight at room temperature, destained in 10% methanol, photographed under normal light and dried between sheets of cellulose for preservation.

Construction and validation of a *Synechococcus* 7942 genomic SELEX library

3.1 Introduction

This chapter describes the theory and methodology used to create and validate the genomic SELEX library composed of *Synechococcus* 7942 genomic DNA sequences, which was used to identify RNA sequences recognized by the *Synechococcus* 7942 RNA-binding protein, RbpA (RNA-binding protein A). Sections 3.2 and 3.3 discuss the theory and the results, respectively, involved in the creation of the *Synechococcus* 7942 genomic SELEX library. Sections 3.4 and 3.5 deal with the theory and the results, repectively, involved in validation of the *Synechococcus* 7942 genomic SELEX library to ensure that the library is a sufficient representation of the *Synechococcus* 7942 genome.

3.2 Construction of a *Synechococcus* 7942 genomic SELEX library: Theoretical Outline

A genomic SELEX library is comprised of a pool of dsDNA molecules individually composed of both fixed and non-fixed (genomic) nucleic acid sequences. The non-fixed portion is located in the middle of each molecule and refers to the multitude of different DNA sequences copied from the genome of the organism of interest. It is this non-fixed or genomic sequence portion that contains the nucleic acid sequence ultimately recognized by the nucleic acid binding protein. The fixed portions are located at both ends of each library molecule. The identity of these fixed nucleic acid sequences enables one to transcribe, reverse-transcribe and PCR amplify the library, necessary steps in the SELEX procedure. To create a genomic SELEX library, genomic DNA isolated from the organism of interest is first copied and the product generated is then manipulated by transcription, reverse-transcription and PCR amplification in an effort to isolate molecules with specific fixed sequences at either end. Once generated, the

genomic SELEX library must then be validated to ensure that it sufficiently represents the genome of the organism of interest. This section discribes the protocol (outlined in Figure 3.1) developed by Singer *et al.*, (1997) that I used to create a *Synechococcus* 7942 genomic SELEX library.

3.2.1 Step 1: First-strand extension

The first step in construction of the *Synechococcus* 7942 genomic SELEX library is copying the *Synechococcus* 7942 genome. This is accomplished by extension of the partially degenerate primer Aran (Table 2.1) by the Klenow fragment of *E. coli* DNA polymerase following hybridization of the primer to template (genomic) DNA, a procedure termed "first-strand" extension. The Aran primer consists of a 16 nucleotide fixed "**A**" sequence at the 5' end and a nine nucleotide single-standed DNA random sequence located at the 3' end of the molecule. During first-strand extension, Aran molecules hybridize to every nine-nucleotide sequence within the genome via the degenerate portion of the primer. Following hybridization, the 3' end of the primer is extended by Klenow enzyme in the reaction mixture (Figure 3.2, reaction 2). In addition to extension of hybridized Aran primer during the first-strand extension reaction, another extension product is generated by intra-strand hybridization of template strands to themselves (Figure 3.2, reaction 1). However, only Aran extended products will ultimately be used to prepare the genomic library.

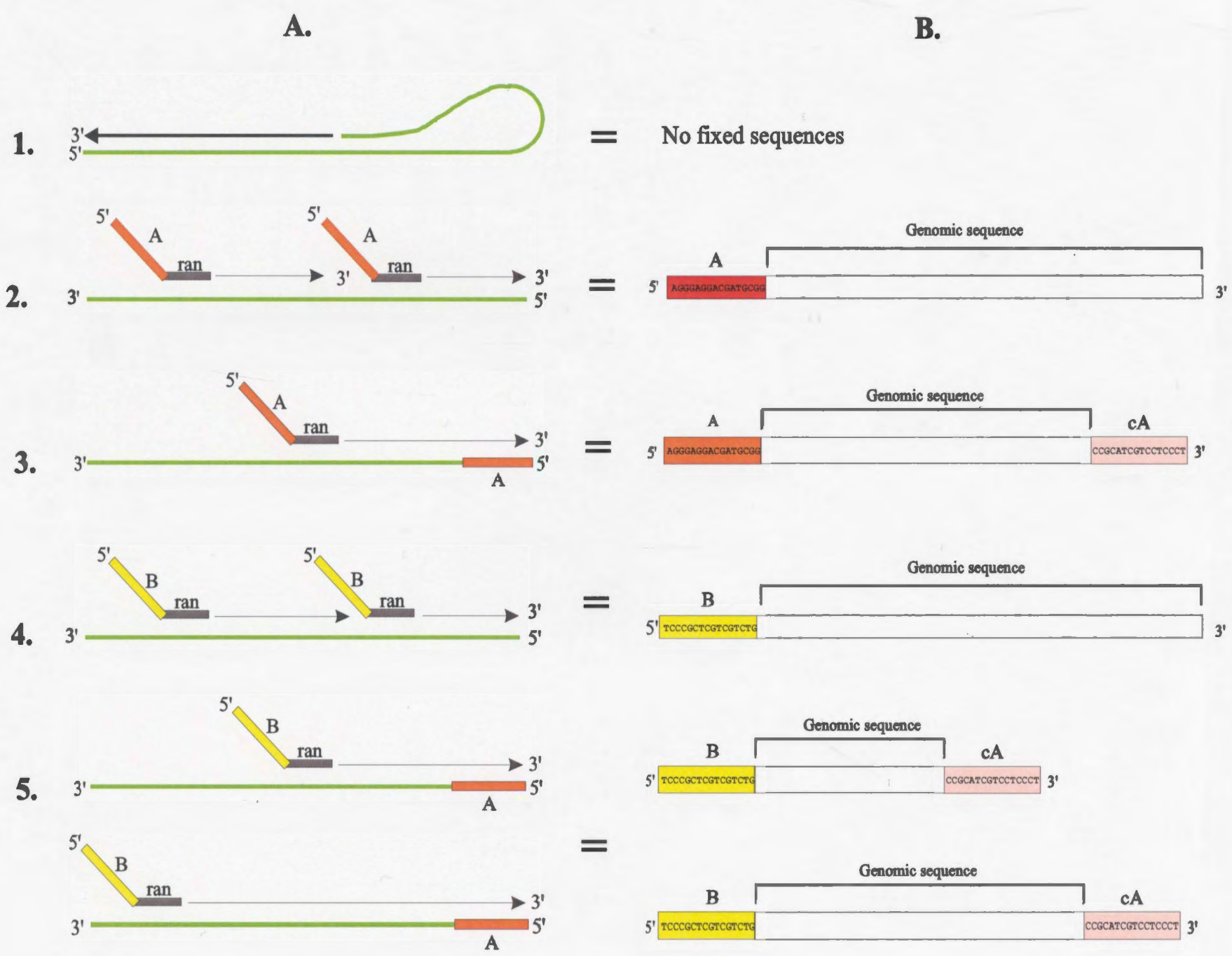
3.2.2 Step 2: Purification of first-strand extension product

Excess Aran primer molecules that are not extended during first-strand extension are removed by filtration (Singer *et al.*, 1997). This procedure is required because in the next step (second-strand extension) a different primer, Bran, is used instead of Aran. The presence of excess Aran primer molecules would interfere with Bran extension by acting as a competitor for extension by Klenow enzyme.

Figure 3.1. Flow chart describing the steps used to create a *Synechococcus* 7942 genomic SELEX library. The section that explains each corresponding step of the procedure is indicated at left.

Section	Procedure
3.2.1	First-strand extension of Aran primer using *Synechococcus* 7942 genomic DNA as a template
	↓
3.2.2	Removal of excess Aran primer by filtration
	↓
3.2.3	Second-strand extension of Bran primer using the first-strand extension product as a template
	↓
3.2.4	Selection of size range for library by urea-PAGE and electroelution of subfragments
	↓
3.2.5	PCR amplification of subfragments with T7A and B primers
	↓
3.2.6	Purification of dsDNA by column chromatography
	↓
3.2.7	Transcription with T7 RNA polymerase
	↓
3.2.8	Purification of RNA by column chromatography
	↓
3.2.9	Reverse-transcription with the T3B primer
	↓
3.2.10	PCR amplification with T7A and T3B primers

Figure 3.2. Schematic diagram of **(A)** the mechanism and **(B)** the ssDNA products generated in the first and second strand extension reactions. All possible reactions and their corresponding products are shown. **(1)** Intra-strand genomic reaction generating ssDNA lacking any fixed primer sequences. **(2)** Primer extension of genomic DNA (green) generating ssDNA product with the A fixed sequence (red) at the 5' end followed by a length of genomic sequence (as indicated). **(3)** Aran primer extension of synthesized DNA containing the A sequence at the 5' end generating ssDNA product flanked by the A sequence at the 5' end and complementary A (cA) sequence (pink) at the 3' end. **(4)** Primer extension of genomic DNA generating ssDNA product with the B fixed sequence (yellow) at the 5' end followed by a length of genomic sequence. **(5)** Bran primer extension products of synthesized template containing the A sequence at the 5' end, the ssDNA products have genomic sequence flanked by the B sequence at the 5' end and the cA sequence at the 3' end. Note that reactions 1 and 2 occur during first-strand extension synthesis, and reactions 1, 3, 4, and 5 occur during second-strand extension synthesis.


A.
B.
1.
3'
5'
=
No fixed sequences
2.
5'
A
ran
3'
=
Genomic sequence
AGGGAGGACGATGCGG
3.
cA
CCGCATCGTCCTCCCT
4.
B
TCCCGCTCGTCGTCTG
5.

3.2.3 Step 3: Second-strand extension

The second-strand extension reaction is performed using the partially degenerate primer Bran (Table 2.1) that, like Aran, consists of a fixed sequence at the 5' end of the molecule ("**B**" sequence) and a random sequence within the nine nucleotide positions on the 3' end of the molecule. The purpose of the second extension reaction is to introduce a second fixed sequence on the opposite 5' end of the Aran extended products generated by first-strand synthesis, thereby creating molecules with *Synechococcus* 7942 genomic DNA sequences flanked by two different fixed sequences. Second-strand extension reactions generate five possible products as shown in figure 3.2. The first results from intra-strand hybridization as before (number 1). The second and third types of products result from the hybridization of Aran primers not removed by filtration to either genomic DNA (number 2) or to Aran primer extended products of the first-strand extension reaction (number 3). The fourth type of product is the result of Bran primer hybridization to genomic DNA (number 4). The fifth type corresponds to Bran primer extended ssDNA generated upon hybridization to Aran extended products of first-strand extension (number 5). The size of these products is dependent on the distance between the location of the sequence complementary to the random portion of the Bran primer and the 5' end of the molecule. This extension event generates ssDNA with the B fixed sequence at the 5' end and the complementary A (cA) sequence at the 3' end and constitutes the desired product of second-strand extension reactions.

3.2.4 Step 4: Purification of second-strand extension product

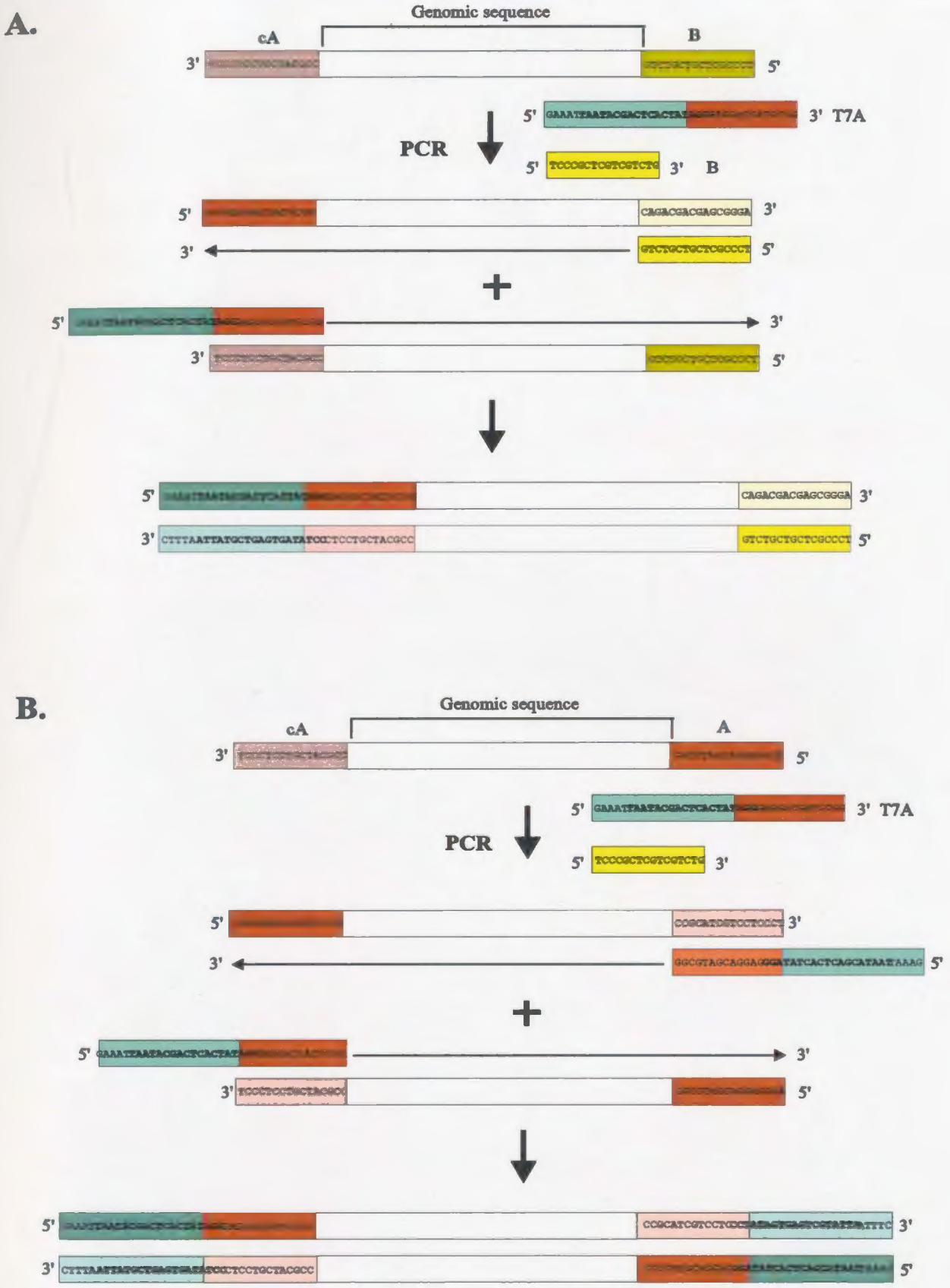
Once the *Synechococcus* 7942 genome is theoretically copied as a result of the first and second strand extension reactions, ssDNA extension products of a specific size must be separated from the genomic DNA template. Purification and size fractionation is achieved using denaturing urea polyacrylamide gel electrophoresis (PAGE), and

subsequent electroelution into a suitable buffer. In this way the molecular size of the genomic SELEX library and thus the length of *Synechococcus* 7942 genomic sequence located in the middle of each library molecule is selected.

3.2.5 Step 5: PCR amplification of second-strand extension product

After purification via electroelution, the ssDNA second-strand extension products are amplified by PCR utilizing primers T7A and B (Table 2.1) (Figure 3.3). The T7A primer contains the "A" sequence of the partially degenerate primer Aran at its 3' end, and a T7 RNA polymerase promoter at the 5' end. The B primer is composed of the "B" sequence of the other partially degenerate primer Bran. Amplification by PCR is necessary for three reasons. First, PCR amplification converts ssDNA to dsDNA and generates sufficient product that analysis by agarose gel electrophoresis is possible. Purification of the various sized second strand extension products cannot be verified directly due to the minute amount of ssDNA generated by the Klenow enzyme extension reaction. Second, PCR amplification selects against two unwanted ssDNA byproducts of the second strand extension reaction. Of the five ssDNA molecules produced by second-strand extension reactions (Figure 3.2), only those that contain the "complementary A" ("cA") or "B" sequences at both ends of the molecule are amplified (Figure 3.2, numbers 3 and 5) That is, it is unlikely that PCR can amplify molecules that contain only the "A" or "B" sequence at one end of the molecule (Figure 3.2, numbers 2 and 4). Third, introduction of a T7 promoter by the T7A primer allows an RNA copy of the library to be generated via a transcription reaction with T7 RNA polymerase. The purpose of creating the *Synechococcus* 7942 genomic SELEX library is to identify the nucleic acid binding sequences of an RNA-binding protein (RbpA), therefore, transcription of the library is necessary.

Figure 3.3. Schematic diagram of the two second-strand extension products that can be amplified by PCR with the T7A and B fixed primers. **(A)** Second-strand extension products composed of genomic DNA (as indicated) surrounded by the complementary A (cA) (pink) and B (yellow) fixed sequences are amplified by primers T7A and B. The T7A primer consists of two parts. One part (green) contains most of the T7 promoter (bold), the other part is the A sequence (red) which contains 3 bases on the 3' end of the T7 promoter and hybridizes to the A primer used in the Klenow enzyme extension reactions. **(B)** Second-strand extension product containing cA and A fixed sequences are recognized by the T7A primer, and will be amplified to generate dsDNA products with the T7A sequence at both ends.


A.
Genomic sequence
cA
B
3'
5'
5'
3' T7A
PCR
5'
3' B
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
B.
Genomic sequence
cA
A
3'
5'
5'
3' T7A
PCR
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'

3.2.6 Step 6: Purification of dsDNA PCR amplified second-strand extension product

Column chromatography using a Qiaquick DNA purification column (Qiagen) was employed in purification of dsDNA following amplification by PCR. This step is necessary to remove excess dNTPs, enzyme and buffer components from the solution because these components may potentially interfere with the transcription reaction used in the next step of library construction.

3.2.7 Step 7: Transcription of PCR amplified second-strand extension product

The next step in library construction is a conversion of DNA to RNA in an effort to select molecules that contain the fixed sequences T7A and B. This conversion is accomplished by T7 RNA polymerase transcription initiated at the T7 promoter located within the T7A primer sequence (Figure 3.4, reaction A). This step in itself does not allow the selection of T7A/B molecules because both types of molecules generated by PCR amplification in step 5 can serve as templates for transcription due to the existence of at least one T7 promoter in each type of molecule (Figure 3.4, reaction B). However, conversion of RNA back to DNA (cDNA) by reverse-transcription in the next step does.

3.2.8 Step 8: Purification of RNA transcript

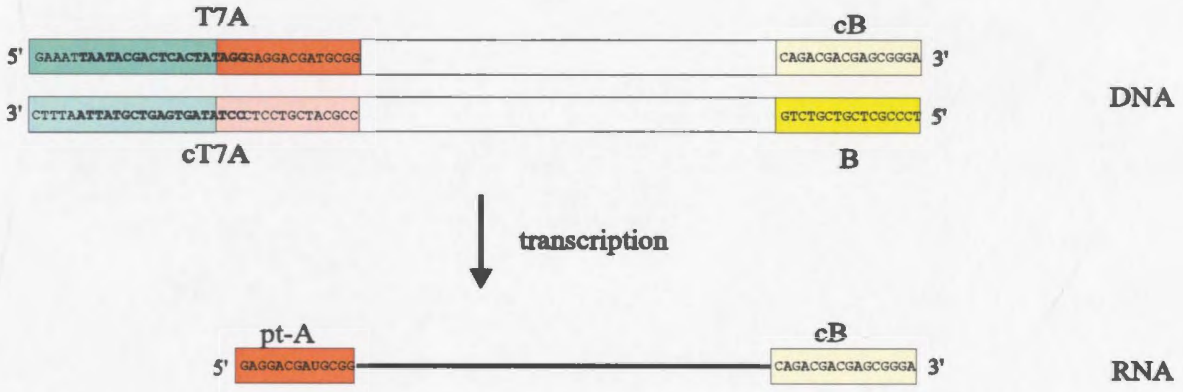
RNA molecules generated by the transcription reaction are purified from the DNA template in the reaction mixture first by degrading the DNA in solution via a DNase digestion reaction, then by the use of a column chromatographic purification strategy designed to retain only RNA molecules (RNeasy purification system, Qiagen).

3.2.9 Step 9: Reverse-transcription of purified RNA

In this reaction, (Figure 3.5) RNA is converted into cDNA by use of reverse transcriptase and the T3B primer (Table 2.1). This primer contains the "B" sequence at the 3' end of the molecule and a 26 nucleotide extension at the 5' end containing a T3 promoter. The incorporation of the T3 promoter enables each library molecule to be

Figure 3.4. Transcription of PCR amplified second-strand extension products. **(A)** The dsDNA synthesized molecules that contain sequences T7A and B at each 5' end respectively are transcribed by T7 RNA polymerase from the T7 promoter (bold) located within the T7A primer sequence. The RNA transcript is composed of the partial A sequence (pt-A, the sequence downstream of the promoter), genomic sequence (black line) and the complementary B (cB) sequence. **(B)** The dsDNA synthesized molecules that contain genomic sequence flanked by T7A fixed primers at both ends give rise to RNA molecules that contain genomic sequence flanked by partial-A (pt-A) and cT7A sequences. Colour coding is the same as that in figure 3.3.

A.


T7A
cB
5'
GAAATTAATACGACTCACTATAGGGAGGACGATGCGG
CAGACGACGAGCGGGA
3'
DNA
3'
CTTTAATTATGCTGAGTGATATCCCTCCTGCTACGCC
GTCTGCTGCTCGCCCT
5'
cT7A
B
transcription
pt-A
cB
5'
GAGGACGAUGCGG
CAGACGACGAGCGGGA
3'
RNA

B.

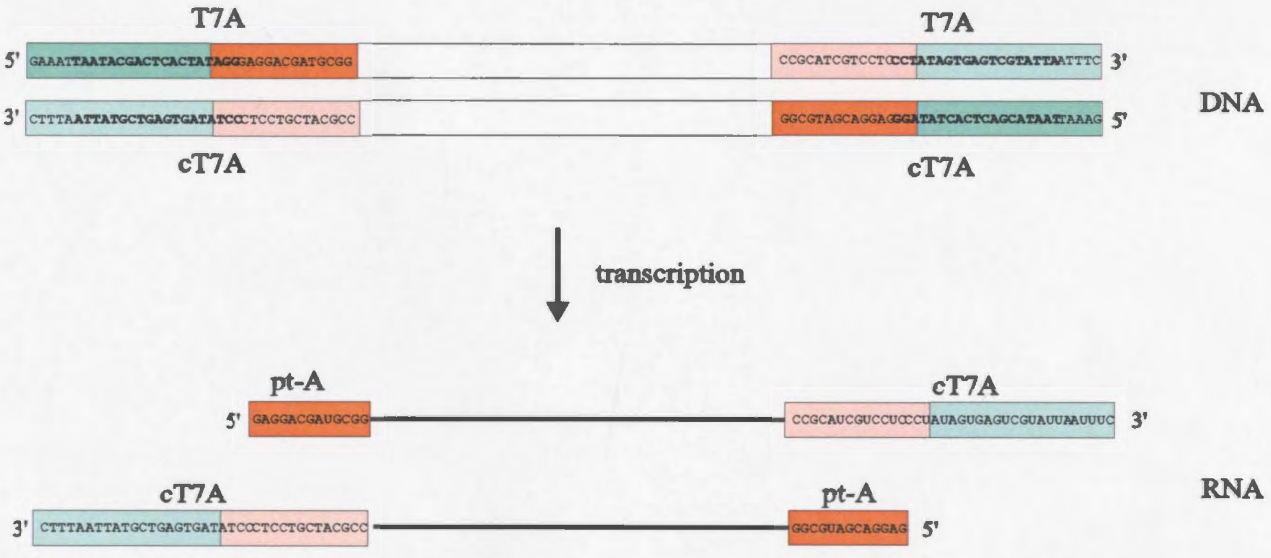
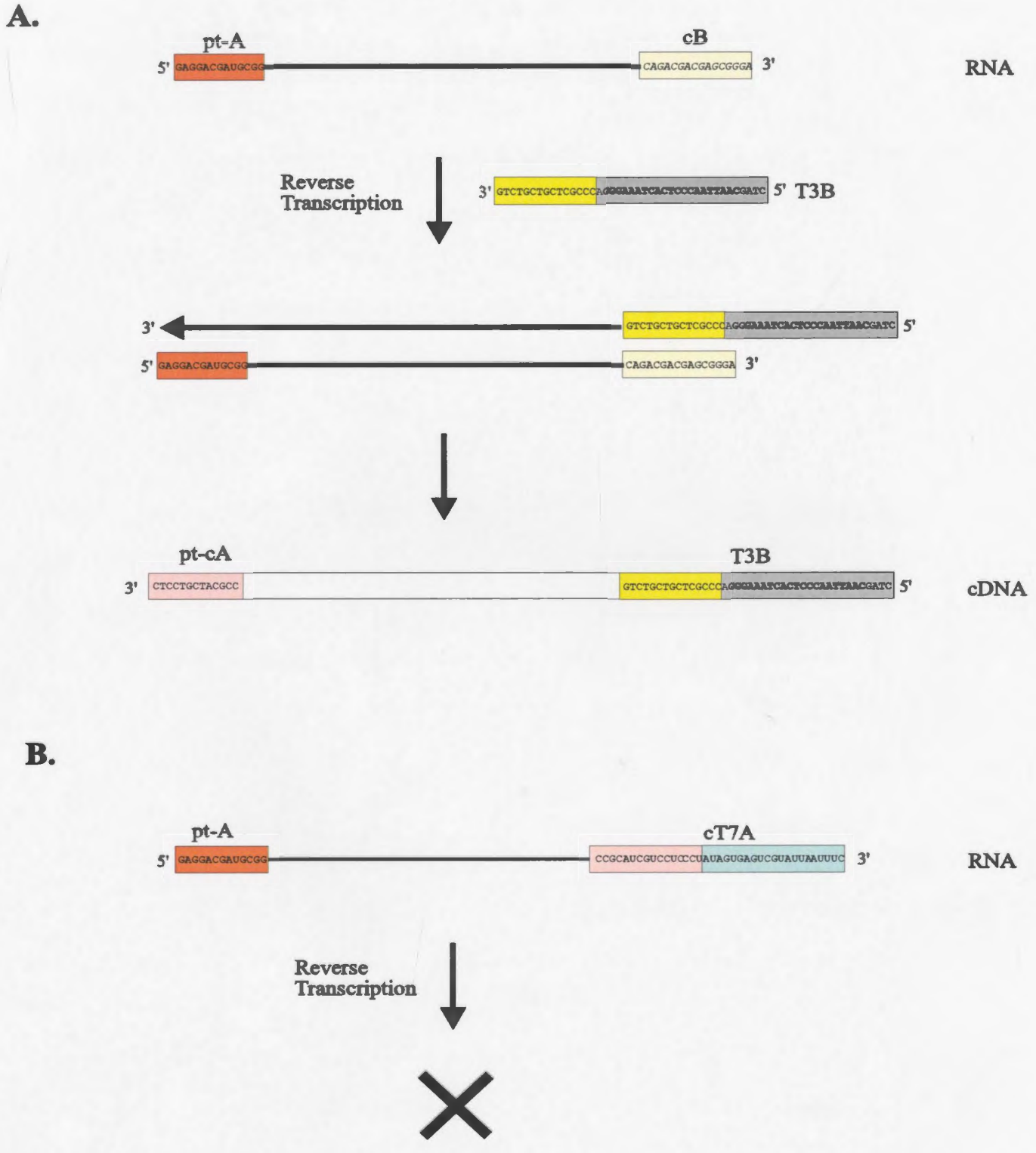

T7A
T7A
5'
GAAATTAATACGACTCACTATAGGGAGGACGATGCGG
CCGCATCGTCCTCCCTATAGTGAGTCGTATTAATTTC
3'
DNA
3'
CTTTAATTATGCTGAGTGATATCCCTCCTGCTACGCC
GGCGTAGCAGGAGGGATATCACTCAGCATAATTAAAG
5'
cT7A
cT7A
transcription
pt-A
cT7A
5'
GAGGACGAUGCGG
CCGCAUCGUCCUCCCUAUAGUGAGUCGUAUUAAUUUC
3'
cT7A
pt-A
RNA
3'
CTTTAATTATGCTGAGTGATATCCCTCCTGCTACGCC
GGCGUAGCAGGAG
5'

Figure 3.5. Schematic diagram of the reverse-transcription reaction used to select molecules containing the complementary B sequence. **(A)** RNA molecules with partial A (pt-A) and complementary B (cB) sequences are reverse-transcribed with the T3B primer, which is composed of the B sequence at the 3' end (yellow) of the molecule and a 26 nucleotide extension containing the T3 promoter (bold) at the 5' end (grey). The product is single-stranded complementary DNA (cDNA) containing the T3B sequence at the 5' end and the partial complementary A (pt-cA) sequence at the 3' end. **(B)** RNA molecules with pt-A and cT7A sequences are not reverse-transcribed because the T3B primer does not hybridize to the 3'end of the molecule. Colour coding is the same as that in figure 3.3.


A.
pt-A
cB
5' GAGGACGAUGCGG
CAGACGACGAGCGGGA 3'
RNA
Reverse Transcription
3' GTCTGCTGCTCGCCCAGGGAAATCACTCCCAATTAACGATC 5' T3B
3'
GTCTGCTGCTCGCCCAGGGAAATCACTCCCAATTAACGATC 5'
5' GAGGACGAUGCGG
CAGACGACGAGCGGGA 3'
pt-cA
T3B
3' CTCCTGCTACGCC
GTCTGCTGCTCGCCCAGGGAAATCACTCCCAATTAACGATC 5'
cDNA
B.
pt-A
cT7A
5' GAGGACGAUGCGG
CCGCAUCGUCCUCCCUAUAGUGAGUCGUAUUAAUUUC 3'
RNA
Reverse Transcription

transcribed twice, thereby increasing the number of RNA sequences generated by transcription of the library by a factor of two. Reverse-transcription with the T3B primer enables a selective conversion of RNA molecules that exclusively contain the complementary B (cB) sequence at the 3' end to cDNA. Therefore the only product of the reverse-transcription reaction should be single-stranded cDNA molecules with the T3B primer sequence at the 5' end and a portion of the cA sequence at the 3' end.

3.2.10 Step 10: The second and final PCR amplification step

The final step involved in creation of the *Synechococcus* 7942 genomic SELEX library is another PCR amplification. In this step, single-stranded cDNA generated by reverse transcription in step 9 is converted to double-stranded DNA by the use of the T7A and T3B primers (Figure 3.6). The final product is a library of dsDNA molecules that contain *Synechococcus* 7942 genomic DNA sequences flanked by sequences T7A and T3B.

3.3 Construction of a *Synechococcus* 7942 genomic SELEX library: Outcome

3.3.1 Step 1: First-strand extension

First-strand extension reactions were carried out as described in sections 2.9 and 3.2.1. To analyze the efficiency of the reaction, newly-synthesized products were labeled by incorporation of digoxygenin-dUTP (DIG-dUTP) (section 2.9). This method of labeling was used because the minute amount of product generated by first-strand extension would not be detected by direct staining methods following separation by gel electrophoresis. Colorimetric detection of DIG-dUTP labeled first-strand extension product following separation by urea-PAGE electrophoresis and transfer to nylon membrane revealed that both intra-genomic and extended products were synthesized when either Aran or Bran primers were used (Figure 3.7A). The relative amounts of Aran extension product (E) compared to that of the intra-genomic product (I) was estimated by

Figure 3.6. PCR amplification used to convert single-stranded cDNA to dsDNA in the final step of library construction. The cDNA preparation is composed of molecules containing the partial complementary A sequence (pt-cA) at the 3' end and the T3B sequence at the 5' end. A PCR amplification reaction utilizing primers T7A and T3B generates the final product, the *Synechococcus* 7942 genomic SELEX library, composed of genomic sequence flanked by fixed sequences T7A and T3B. Colour coding is the same as that in figure 3.3.

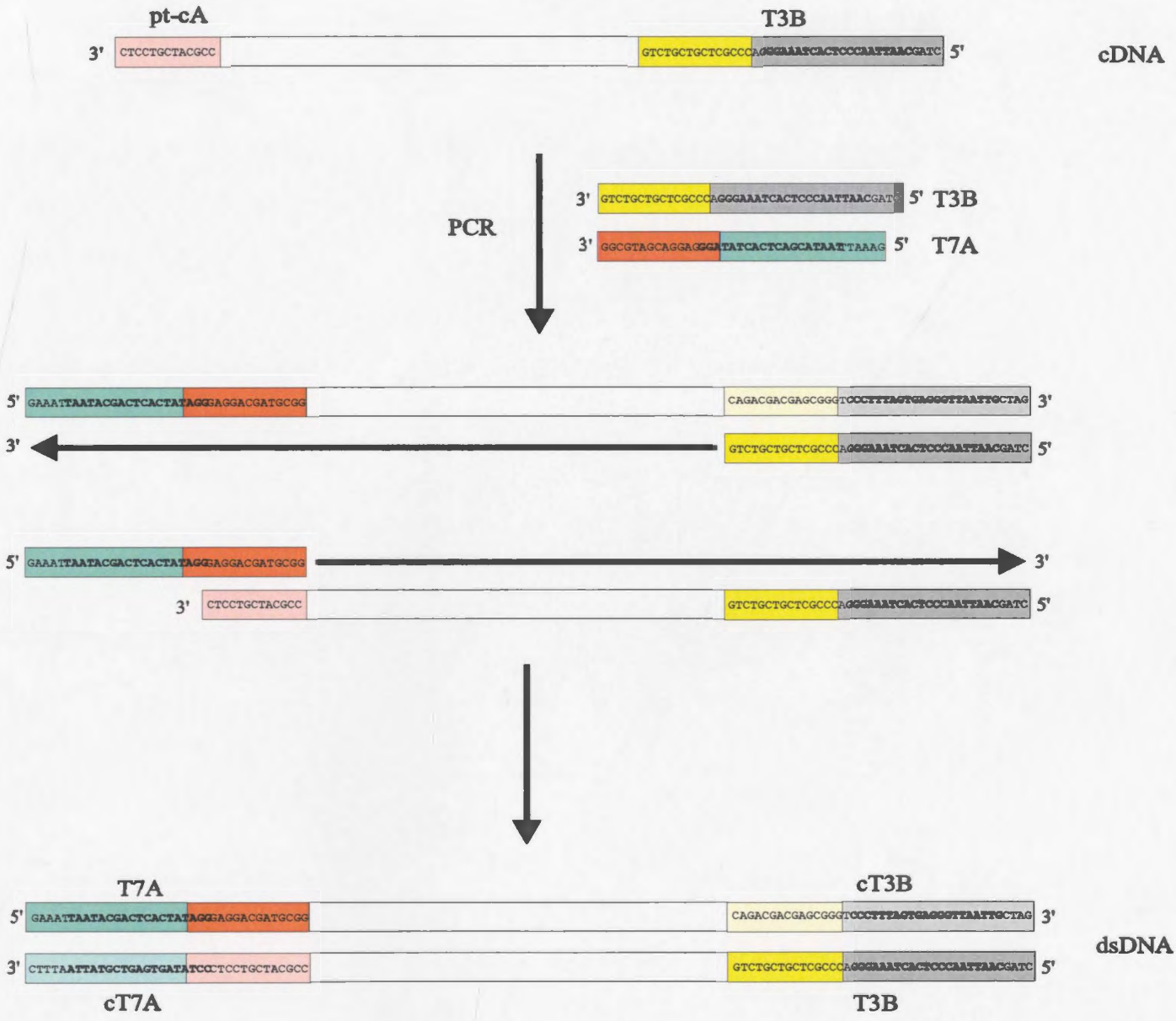
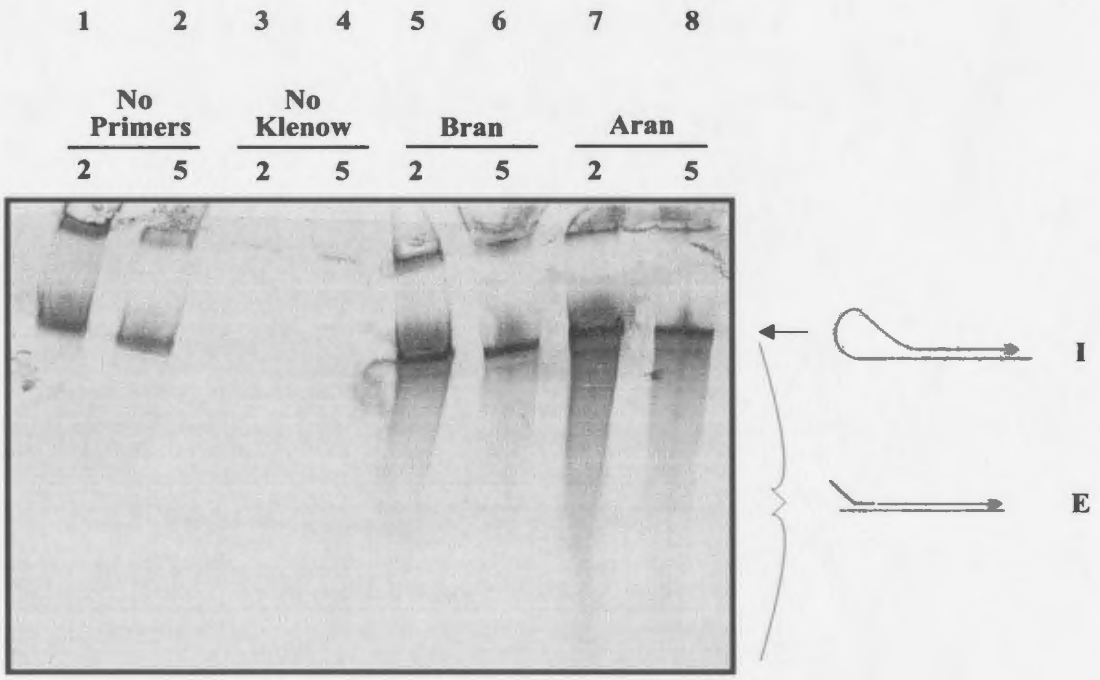
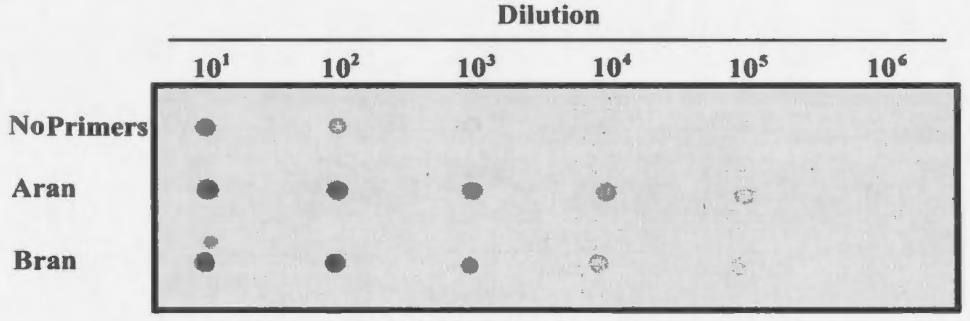

pt-cA
T3B
3'
CTCCTGCTACGCC
GTCTGCTGCTCGCCCA
5'
cDNA
PCR
3'
GTCTGCTGCTCGCCCAGGGAAATCACTCCCAATTAACGATC
5'
T3B
3'
GGCGTAGCAGGAGGGATATCACTCAGCATAATTAAAG
5'
T7A
5'
GAAATTAATACGACTCACTATAGGGAGGACGATGCGG
CAGACGACGAGCGGGT
3'
3'
GTCTGCTGCTCGCCCA
5'
5'
GAAATTAATACGACTCACTATAGGGAGGACGATGCGG
3'
3'
CTCCTGCTACGCC
GTCTGCTGCTCGCCCA
5'
T7A
cT3B
5'
GAAATTAATACGACTCACTATAGGGAGGACGATGCGG
CAGACGACGAGCGGGT
3'
3'
CTTTAATTATGCTGAGTGATATCCTCCTGCTACGCC
GTCTGCTGCTCGCCCA
5'
dsDNA
cT7A
T3B

Figure 3.7. Analysis of the first-strand extension reaction synthesis by DIG-dUTP labelling. **(A)** Aliquots (2uL and 5uL) of the first-strand extension reaction product generated in the absence of primer (lanes 1 and 2), in the absence of Klenow DNA polymerase (lanes 3 and 4), with Bran primer only (lanes 5 and 6) and Aran primer only (lanes 7 and 8) were separated in an 8M urea/4% polyacrylamide gel. Following separation, DNA was transferred to nylon membrane by electrophoretic transfer and the DIG label was detected colorimetrically. The location of intra-genomic (I) and primer extended (E) products are indicated on the right. **(B)** Serially diluted aliquots (as indicated) of DIG-dUTP labeled extension reactions were dot blotted onto a nylon membrane and the DIG label was detected colorimetrically. Densitometric quantification (by Optiquant analysis software) of the intensity of each colorimetric signal (dot) was used to compare the amount of extension product generated in the presence of a primer (Aran or Bran) with that in the absence of a primer.

A.


1 2 3 4 5 6 7 8
No Primers
No Klenow
Bran
Aran
2 5 2 5 2 5 2 5
I
E

B.


Dilution
10¹ 10² 10³ 10⁴ 10⁵ 10⁶
NoPrimers
Aran
Bran

dot blot analysis (Figure 3.7B). Densitometric analysis revealed that approximately 20 fold and 14 fold more product was generated in the presence of Aran and Bran respectively, than in the absence of either primer. These results indicated that the first-strand extension reaction was successful in creating a significant amount of primer extended product. The process was therefore repeated in the absence of DIG-dUTP to synthesize ssDNA for the next step of library construction.

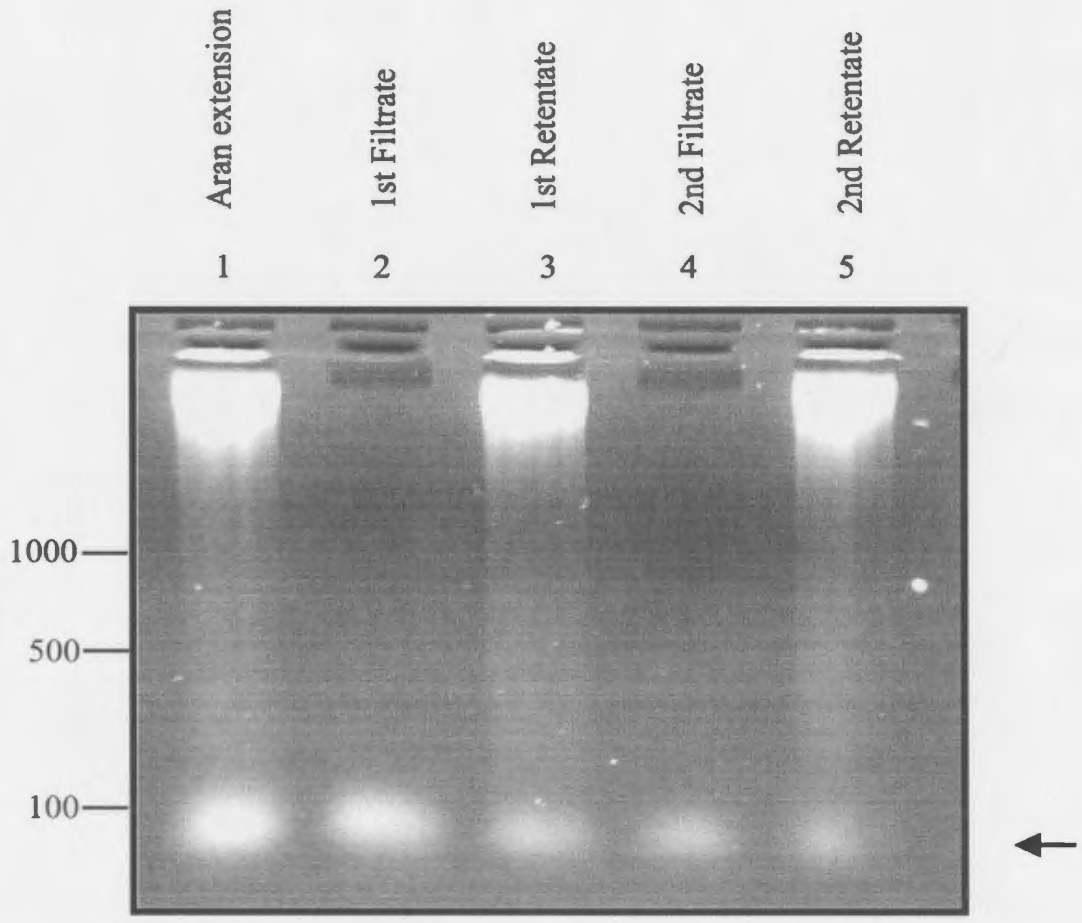
3.3.2 Step 2: Purification of first-strand extension product

Following first-strand extension, molecules less than 10kDa including non-extended Aran primer molecules (8.2kDa) were removed by filtration. Filtration was performed twice to maximize the removal of free primer. To assess the performance of two rounds of filtration, samples of retentate (the first-strand reaction) and filtrate (molecules less than 10kDa) were examined by agarose gel electrophoresis (Figure 3.8). Filtration of molecules less than 10kDa in size was qualitatively evaluated by a reduction in band intensity corresponding to the primer (designated by arrow) in samples from the first-strand reaction and retentate following each filtration. In addition to the reduction in band intensity in the retentate, a band corresponding to primer was evident in the filtrate as was expected. These results indicated that the vast majority of non-extended primer was removed from the first-strand product and thereby should not interfere considerably in the next step of library construction.

3.3.3 Step 3: Second-strand extension

Second-strand extension utilized the Bran oligonucleotide as the primer and Microcon-10 filtered first-strand extension product as the template. The efficiency of this reaction was not analyzed but the ability of Bran to be extended by Klenow DNA polymerase was similar to that of Aran as stated in section 3.3.1 and shown in Figure 3.7A.

Figure 3.8. The effectiveness of Microcon-10 filtration in removing excess Aran primer from the first-strand synthesis reaction. Samples corresponding to 1.3% total volume of Aran extension product (lane 1), 2.7% total volume of filtrate 1 (lane 2) and filtrate 2 (lane 4), and 1% of retentate 1 (lane 3) and retentate 2 (lane 5) were separated by electrophoresis in a 1.5% agarose gel. The band corresponding to non-extended Aran primer and extended products less than 10kDa in size is indicated (arrow). The location of 100bp dsDNA ladder molecules used as size markers are indicated.


Aran extension
1st Filtrate
1st Retentate
2nd Filtrate
2nd Retentate
1
2
3
4
5
1000
500
100

3.3.4 Steps 4 and 5: Purification and PCR amplification of second-strand extension product

Second-strand extension products were purified by electroelution of gel slices following separation by urea-PAGE. Three portions of the urea-PAGE gel containing second-strand extension product corresponding to the region between 3-4 cm, 4-5 cm and 5-6 cm from the top of the gel (Figure 3.9A) were excised, electroeluted, and DNA was purified and PCR amplified in a reaction containing primers T7A and B (Figure 3.3A). Agarose gel electrophoretic analysis of the PCR amplified product revealed that a range of ssDNA sizes were recovered from each gel section (Figure 3.9B). That is, PCR amplification of ssDNA from gel slices 3-4 cm, 4-5 cm and 5-6 cm generated dsDNA in the size range of 141-195 bp, 110-154 bp and 91-141 bp respectively. The PCR amplified second-strand extension product from the 5-6 cm gel slice was chosen for construction of the library due to its size range. The genomic sequence portion of a library of 91-141bp dsDNA fragments ranges in size between 38 bp and 88 bp. RNA expressed from this library is short enough that any RNA consensus sequence recognized by RbpA would be present, and long enough to form the necessary RNA secondary structures involved in an interaction with RbpA.

3.3.5 Step 6: Purification of dsDNA PCR amplified product

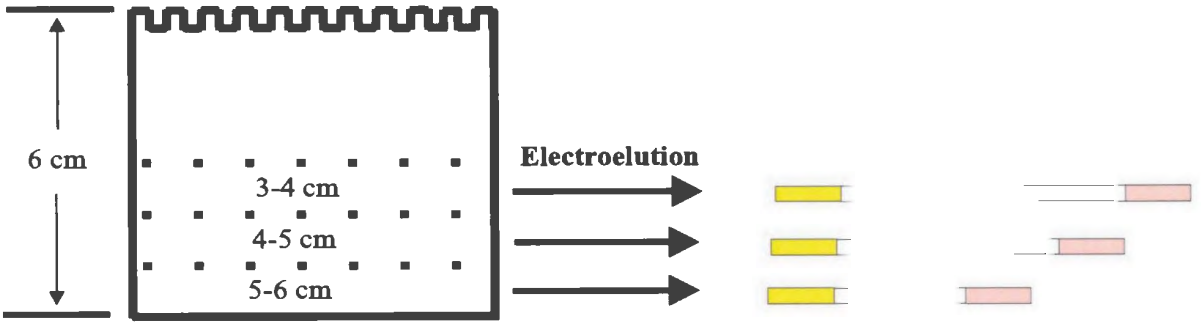
The PCR amplified second-strand product was purified prior to the next step in library construction by phenol/chloroform extraction and ethanol precipitation as described in Materials and Methods. Purification was required because it was necessary to remove excess dNTPs, enzyme and buffer components to prevent complications in the remaining steps of library construction.

3.3.6 Step 7: Transcription of PCR amplified second-strand extension product

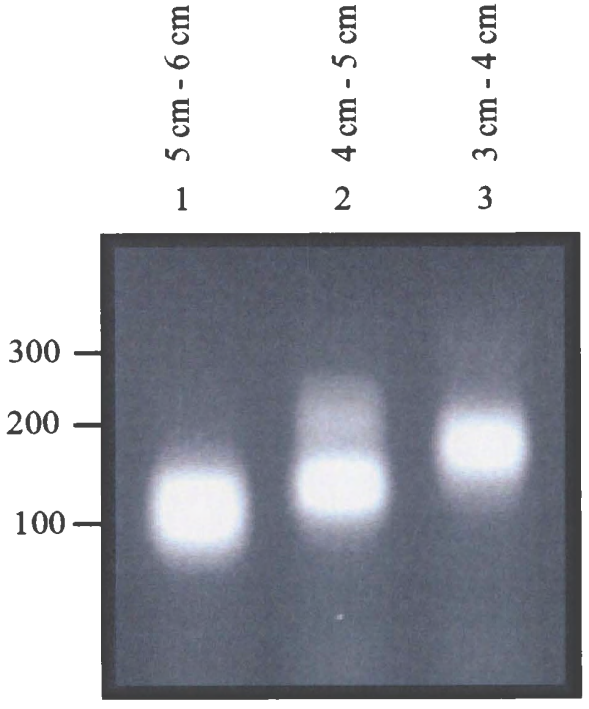
The PCR amplified second-strand extension product served as a template for

Figure 3.9. Purification of second-strand extension reaction product. **(A)** Schematic diagram of the strategy. The second-strand extension reaction product was separated in an 8M urea/7% acrylamide gel of length 6 cm as indicated. Following electrophoresis, segments of the gel corresponding to regions 3-4 cm, 4-5 cm and 5-6 cm from the top of the gel were excised (dotted line), electroeluted and the DNA reaction product was purified. Each gel segment should correspond to a particular size of DNA fragment as shown. **(B)** Analysis of the product recovered from gel segments 5-6 cm (lane 1), 4-5 cm (lane 2) and 3-4 cm (lane 3) following PCR amplification with T7A and B primers and separation on a 2% agarose gel. The location of 100 bp DNA ladder molecules used as a marker are indicated at left.

A.


6 cm
3-4 cm
4-5 cm
5-6 cm
Electroelution

B.


5 cm - 6 cm
4 cm - 5 cm
3 cm - 4 cm
1
2
3
300
200
100

transcription by T7 RNA polymerase. The transcription reaction was successful as evidenced by a comparison of band intensities of the transcription product and 47 ng of template (the amount loaded in the transcription lane) upon agarose gel electrophoresis (Figure 3.10A). That is, the amount of nucleic acids (DNA) in the template only lane is considerably less than the amount of nucleic acids (DNA and RNA) in the transcription reaction sample lane, thereby providing evidence that the transcription reaction had worked.

3.3.7 Step 8: Purification of RNA transcript

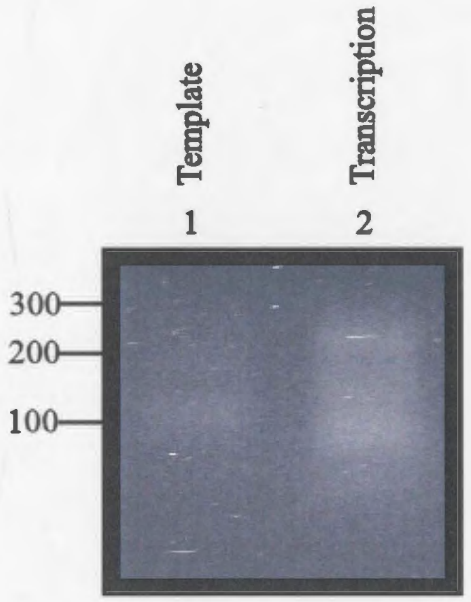
RNA was purified with the RNeasy purification system (Qiagen) as described in sections 2.5 and 3.2.8. In this protocol, DNA template was removed from the transcription reaction solution by RQ1 DNase (Promega) degradation followed by RNeasy column chromatography whereby RNA is retained and DNA is not. The effectiveness of the purification in terms of removing the DNA template from the RNA preparation could not be assessed directly after this step because I could not conveniently determine the relative amount of RNA versus DNA in the purified preparation. Therefore, as described in section 3.3.9, I had to wait until the final step of library construction to assess the efficiency of this important purification step.

3.3.8 Step 9 : Reverse-transcription of purified RNA

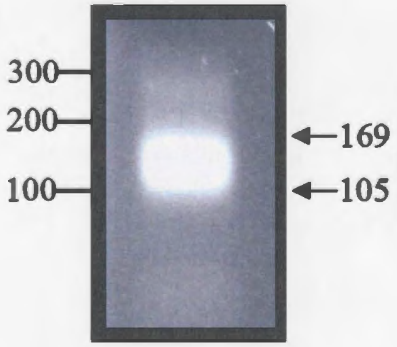
Purified RNA was converted into cDNA by reverse-transcription. The primer used in the reaction was T3B. RNA molecules that contained the cB sequence at the 3' end of the molecule would exclusively serve as a template for reverse-transcription. In addition to selection of RNA molecules containing the cB sequence, the T3B primer incorporated a T3 promoter into the end of the molecule opposite to that of the T7 promoter. This alteration was made with the intention of increasing the diversity of RNA molecules generated by transcription of the genomic SELEX library. In this way, both T7 and T3

Figure 3.10. Theoretical and experimental size range of the transcribable *Synechococcus* 7942 genomic SELEX library. **(A)** Verification that transcription of the PCR amplified second-strand extension product was successful. The same amount of DNA template (47ng) as in the transcription sample (lane 2) was loaded in a template only control (lane 1) and electrophoresed in a 2% agarose gel. **(B)** Analysis by 2% agarose gel electrophoresis of the dsDNA *Synechococcus* 7942 genomic SELEX library following the final PCR amplification step. The estimated size range of library molecules is shown on the right. A table displaying how the theoretical size range of library molecules was generated is also included. **(C)** A graphical representation of the genomic SELEX library. The length of each portion of a dsDNA library molecule is given in nucleotides. Colour coding is the same as that in figure 3.3. The location of 100bp DNA markers is given on the left.

A.



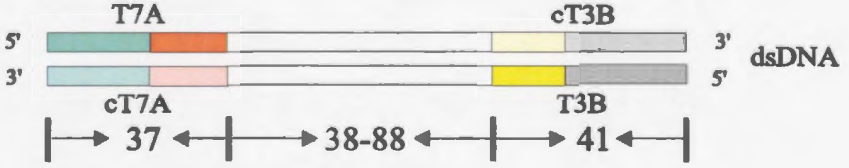
B.



Theoretical size range of library molecules

	Smallest	Largest
Electroeluted fraction 5-6cm T7A/B ssDNA (figure 3.3.3)	91	141
T3B primer sequence (without the B sequence)	25	25
Total	116	166

C.



RNA polymerase could be used to transcribe the library thereby increasing the diversity of possible RNA sequences generated from the library by a factor of two.

3.3.9 Step 10: The second and final PCR amplification step

The final step involved in creation of the *Synechococcus* 7942 genomic SELEX library was PCR amplification. In this step, single-stranded cDNA generated by reverse transcription was converted to double-stranded DNA by PCR amplification utilizing primers T7A and T3B. Agarose gel electrophoretic analysis following PCR amplification revealed that the library had an experimental size range of 105bp to 169bp (figure 3.10B). This range was comparable with that of the theoretical expected size range of 116 to 166 bp. The final product was a dsDNA library composed of molecules that contain *Synechococcus* 7942 genomic DNA sequences ranging from 38 to 88 nucleotides in length flanked by fixed primer sequences T7A and T3B (Figure 3.10C).

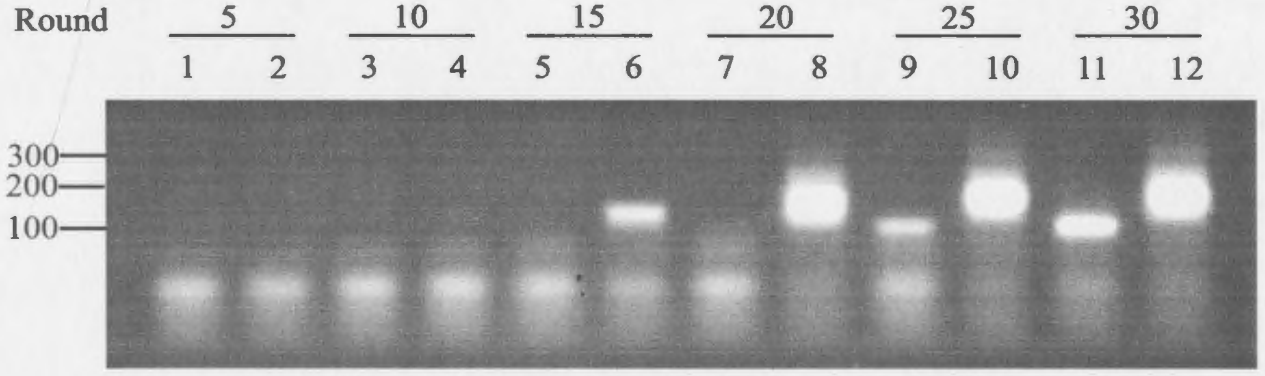
At this point in library construction, I was able to use PCR to qualitatively examine the efficiency of the RNA purification step. As described in section 3.3.7, following transcription the RNA product was subjected to DNase digestion for the purpose of removing the DNA template. Degradation of the template was important because I wanted all nucleic acid material used in creation of the library to "pass through" an RNA intermediate for the purpose of selecting those molecules that contain both T7A and T3B primer sequences. Therefore I qualitatively assessed the relative amounts of RNA and DNA in the purified RNA preparation by comparing the amount of product generated by RT-PCR versus PCR using the purified RNA preparation as template. The amount of RNA present was detected by RT-PCR whereas the amount of contaminating DNA was detected by PCR amplification alone without a reverse transcription step. The amount of RNA and DNA was comparable by this method because one RNA molecule will generate one cDNA molecule by reverse transcription.

Once amplified, the relative amount of RNA and DNA was determined by agarose gel electrophoresis of the "real-time" amplification products (Figure 3.11). After three DNase digestions and subsequent RNA purifications, trace amounts of DNA still remained in solution as evidenced by a band (lanes 9 and 11). Therefore I had to conclude that removal of DNA template was not 100% effective. However, the amount of contaminating DNA was negligible compared to the amount of RNA. A strong signal (band) was produced following only 15 rounds of PCR amplification of the RT product (lane 6). In contrast an additional 10 to 15 rounds (25-30 rounds in total) of PCR amplification without the RT step was required to generate a band of similar intensity, thereby indicating that 2^{10} or approximately 1000 fold less nucleic acids were present in the PCR sample generated without an RT step compared to the RT-PCR sample. Therefore I concluded that the vast majority of nucleic acid molecules in the purified RNA preparation was RNA, and that most genomic SELEX library molecules are of the T7A/T3B type because these molecules had passed through the RT-PCR step employed in library construction.

3.4. Validation of the *Synechococcus* 7942 genomic SELEX library by assessing the representation of the *rbpB* gene : Theory

Prior to use of a genomic SELEX library to identify the sequence specificity of a nucleic acid binding protein, it must be ensured that the library is sufficiently representative of the genomic sequences from the organism used to create the library. If the library does not contain a good representation of the genome, any results would be suspect due to the lack of all possible genomic sequences. Validation of a genomic SELEX library is accomplished by assessing the representation of a portion of the genome in the library (Singer *et al.*, 1997). If a particular sequence is sufficiently represented, then it can be concluded that other sequences from the entire genome should

Figure 3.11. A qualitative assessment of the efficiency of RNA purification by "real-time" PCR and RT-PCR. The amount of contaminating DNA template (detected by PCR amplification) and RNA transcript (detected by RT-PCR amplification) in an RNA preparation was compared by 2% agarose gel electrophoresis. Samples were taken from RT-PCR (odd numbered lanes) and PCR reactions (even numbered lanes) after a specified number of cycles as indicated (5, 10, 15, 20, 25, 30). The size of the PCR amplified band in the RT-PCR product is 25 bp larger than that of the PCR product due to the identity of the template molecules in each reaction. The RNA that served as the template for reverse-transcription and PCR amplification contains fixed sequences A and T3B whereas the DNA contaminant that served as the template for PCR contains fixed sequences A and B. The T3B sequence is 25 nucleotides longer than the B sequence. The location of 100 bp DNA molecules used as size markers are indicated at left.


Round
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
11
12
300
200
100

be as equally well represented.

To determine if a particular sequence is located within the library, those molecules containing the sequence of interest must first be isolated from the rest of the library. A simple method of isolating dsDNA molecules based on sequence is by nested PCR amplification. This method selectively amplifies and thus effectively isolates DNA molecules by the use two (or more) overlapping primers in successive polymerase chain reactions. Once isolated, the nucleic acid sequence of molecules containing the particular genomic sequence can be determined. To validate the *Synechococcus* 7942 genomic SELEX library, a portion of the genome corresponding to the *rbpB* gene was analyzed for its representation in the library.

3.4.1 The nested-PCR strategy

The region adjacent to nucleotide 335 of the *rbpB* gene was chosen as the sequence of interest (Figure 3.12) due to the presence of a *SacI* restriction endonuclease site that could be used in cloning. Selective amplification was achieved by the use of nested primers rbpB1 and rbpB2 (Table 2.1) which hybridize to nucleotides 354-378 and 335-364 of the *rbpB* gene respectively. The primer rbpB2 overlaps rbpB1 by 11 nucleotides and recognizes 19 nucleotides upstream of the rbpB1 primer. In this way, molecules containing at least nucleotides 335 to 378 of the *rbpB* gene are amplified twice (60 rounds of amplification in total). The chance of amplification of non-*rbpB* gene sequences is greatly decreased by employing two successive PCR reactions specific to the *rbpB* gene. In addition, nested-PCR creates a vast excess of these molecules in comparison to that of the rest of the library and therefore essentially isolates these molecules from the library.

Nested PCR amplification was performed in two steps. In the first step, the primers T7A and rbpB1 (Figure 3.13A) were used. The rbpB1 primer should hybridize to

Figure 3.12. Region of the *rbpB* gene used to validate the *Synechococcus* 7942 genomic SELEX library. Nucleotides 380 to 290 (3' to 5' orientation) of the 553 bp nucleotide sequence containing the *rbpB* gene (Genbank file accession number L25435, Dolganov and Grossman, 1993) (ser0515) are shown. Nested PCR primers rbpB1 (blue) and rbpB2 (blue and orange) hybridize to regions of nucleotide sequence from 378 to 354 and 364 to 335, respectively. Note that the region of primer overlap is indicated by identical coloration. The *SacI* site is boxed. Numbers indicate the location of important nucleotides referred to above.


CGCAGTTGGCGACCAAACCACTCAG
rbpB2 AAACCACTCAGTGCCATTGAGCTCGCTGAT
3' GCGCGTCGACCGCTGGTTTGGTGAGTCGCGGTAGCTCGAGCGACTATCGTCGTAGAAGCACTCGTAGAAG AAGGTAAAGTTACTTTCGCT 5'
SacI
378
364
354
335
322
290

members of the SELEX library that contained nucleotides 354 to 378 of the *rbpB* gene; and, the T7A primer should hybridize to the fixed sequence present at the other end of the molecule. PCR amplification should generate dsDNA molecules with the rbpB1 and T7A primer sequences located at opposite ends of the molecule. These PCR products were used as the template in a second PCR amplification reaction in which the primers T7A(BamHI) and rbpB2 were used (Figure 3.13B). The T7A(BamHI) primer should hybridize to the T7A sequence and incorporate a *BamHI* site to be used in cloning; the rbpB2 primer should hybridize to molecules that contain nucleotides 335 to 364 of the *rbpB* gene of which contains a *SacI* restriction site also used in cloning. In this way, the second *rbpB* gene specific primer (rbpB2) overlaps a portion of the first (bases 364 to 354) and specifically recognizes 19 bases upstream of the first (bases 353 to 335), thereby selecting and amplifying the sequence twice. The product of the second amplification reaction should be dsDNA molecules with the T7A(BamHI) sequence at one end and the rbpB2 sequence at the opposing end that contain *BamHI* and *SacI* restriction endonuclease sites respectively, to be employed in cloning.

3.4.2 Cloning nested PCR products and identification of potentially recombinant clones

The products of nested PCR amplification were cloned into pUC19 using the *BamHI* and *SacI* restriction sites as shown in figure 3.14. Cloning of fragments into pUC19 and subsequent transformation of ligation product into *E. coli* JM109 allowed detection of recombinant plasmid containing clones by blue/white selection (section 2.6). however, blue/white selection alone is not sufficient to identify recombinant clones due to the false positive results that can be generated. Therefore, since dideoxynucleotide sequencing is both expensive and time consuming, I devised a quick and easy method to identify potentially "positive" result generating clones, that is, plasmids that have an

Figure 3.13. Schematic diagram of the nested PCR amplification strategy used to verify representation of a portion of the *rbpB* gene in the SELEX library. **(A)** In the first PCR step, primers rbpB1 (blue) and T7A (green with A sequence in red) were used to amplify library molecules containing *Synechococcus* 7942 genomic sequence complementary to the rbpB1 nucleotide sequence (nucleotides 378 to 354, refer to Figure 3.12) and the T7A primer sequence. The T3B fixed sequence which should be present in all library molecules is shown in grey and yellow (B sequence). **(B)** In the second amplification step primers T7A(BamHI) and rbpB2 were used. The 5' end of rbpB2 (blue) hybridizes to the 3' complementary sequence of rbpB1 (light blue) and the 3' end of rbpB2 (orange) hybridizes to the sequence upstream of that recognized by rbpB1. The other end of the molecule is PCR amplified via the T7A(BamHI) primer (sequence different to that of T7A is shown in purple) which introduces a *BamHI* restriction endonuclease site (as indicated) in the amplified products. The naturally occuring *SacI* restriction endonuclease site located in the *rbpB* is indicated. Complementary nucleotide sequences are shown in a lighter shade of the same colour.

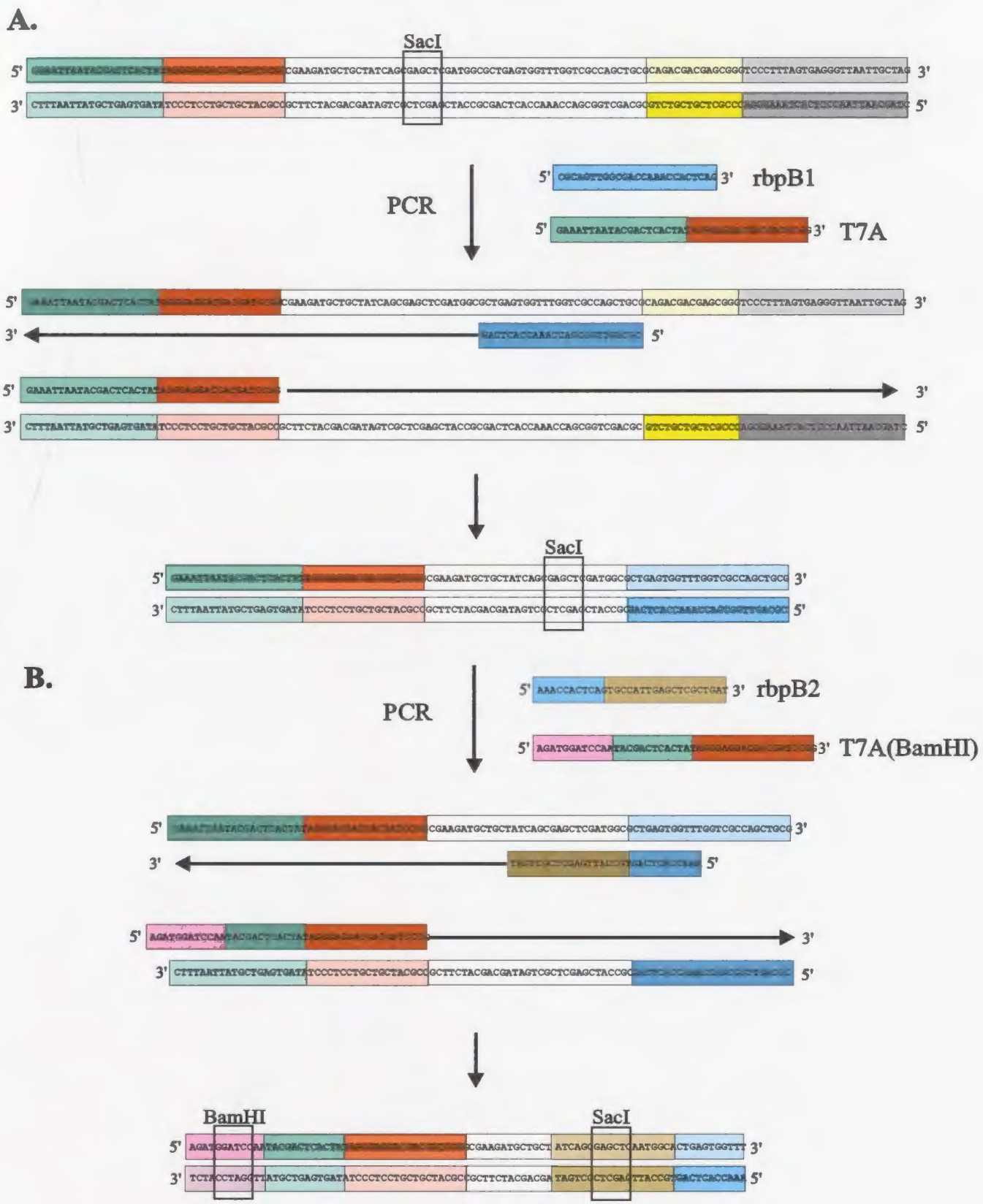
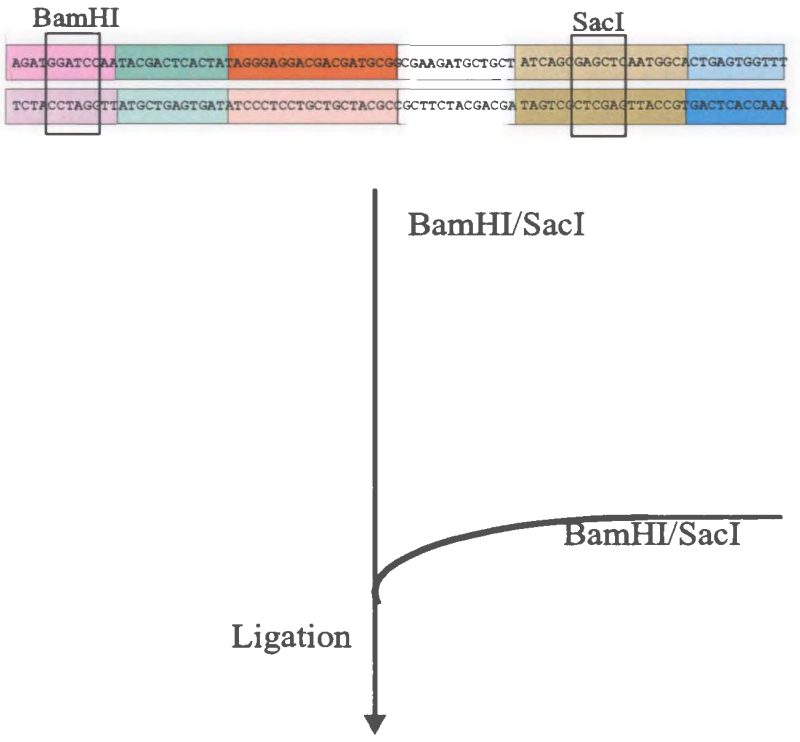

A.
SacI
PCR
rbpB1
T7A
SacI
B.
PCR
rbpB2
T7A(BamHI)
BamHI
SacI

Figure 3.14.. Cloning of nested PCR amplification products. A representative PCR amplification product containing a portion of the *rbpB* gene is cloned into pUC19 by use of the *BamHI* restriction site located at the 5' end of the T7A(BamHI) primer and the *SacI* site located in the rbpB2 primer. Colour coding is the same as in figure 3.13.


BamHI
SacI
BamHI/SacI
BamHI/SacI
Ligation
SacI
BamHI
EcoRI
SalI
HindIII
lacZ
pUC19
Amp
Ori
SacI
BamHI
EcoRI
SalI
HindIII
lacZ
pUC19
Amp
Ori

insert and plasmid preparations that contain a single recombinant plasmid.

The first criterion for "positive" clones is obvious because one does not want to sequence non-recombinant plasmids. The cloning process can generate non-recombinant plasmids or plasmids that contain an incorrect insert, and I did not want to sequence numerous plasmids that would not provide any data. The second criterion of a "positive" clone is recombinant purity because when colonies are chosen to inoculate a liquid culture of cells for plasmid isolation, a single colony may contain two strains with different recombinant plasmids. This would pose a problem during sequence analysis because the existence of more than one recombinant insert could lead to confusing results.

Two methods can be employed to identify potentially "positive" clones: restriction enzyme analysis and PCR analysis. Plasmids can be digested with the restriction enzymes *HindIII* and *EcoRI* because restriction sites recognized by these enzymes flank the insert within the multiple cloning site of pUC19 (Figure 3.14). To analyze plasmids by PCR (Figure 3.15), two standard sequencing primers that flank the multiple cloning site could be used. However, for the purpose of using PCR, both primers were modified so that the melting temperature for each primer was approximately 60°C (Elongated Forward and Elongated Reverse primers, Table 2.1). A relatively high melting temperature is important in PCR amplification to minimize the likelihood of non-specific primer-template interactions which lead to the amplification of undesirable sequences.

An ideal *Synechococcus* 7942 genomic SELEX library would contain molecules whose genomic sequence portion is in a staggered arrangement such that there would exist one molecule per base pair of the genome. Therefore, upon nested-PCR analysis of the *rbpB* gene, molecules would be identified that contain each position downstream of nucleotide 335 of the *rbpB* gene in a staggered arrangement (Figure 3.16).

Figure 3.15. Diagram of the PCR amplification strategy used to identify potential positive clones prior to sequencing. The multiple cloning site of plasmid pUC19 containing a representative nested *rbpB* PCR insert is shown. The insert was cloned via the *BamHI* site located in the T7A(BamHI) primer sequence and the *SacI* site located in the rbpB2 primer sequence. The cloning site is flanked and thus can be PCR amplified (as denoted by arrows) using the Elongated Forward (25mer) (green boxed outline) and Elongated Reverse (28mer) (red boxed outline) primers. Colour coding is the same as in figure 3.13.

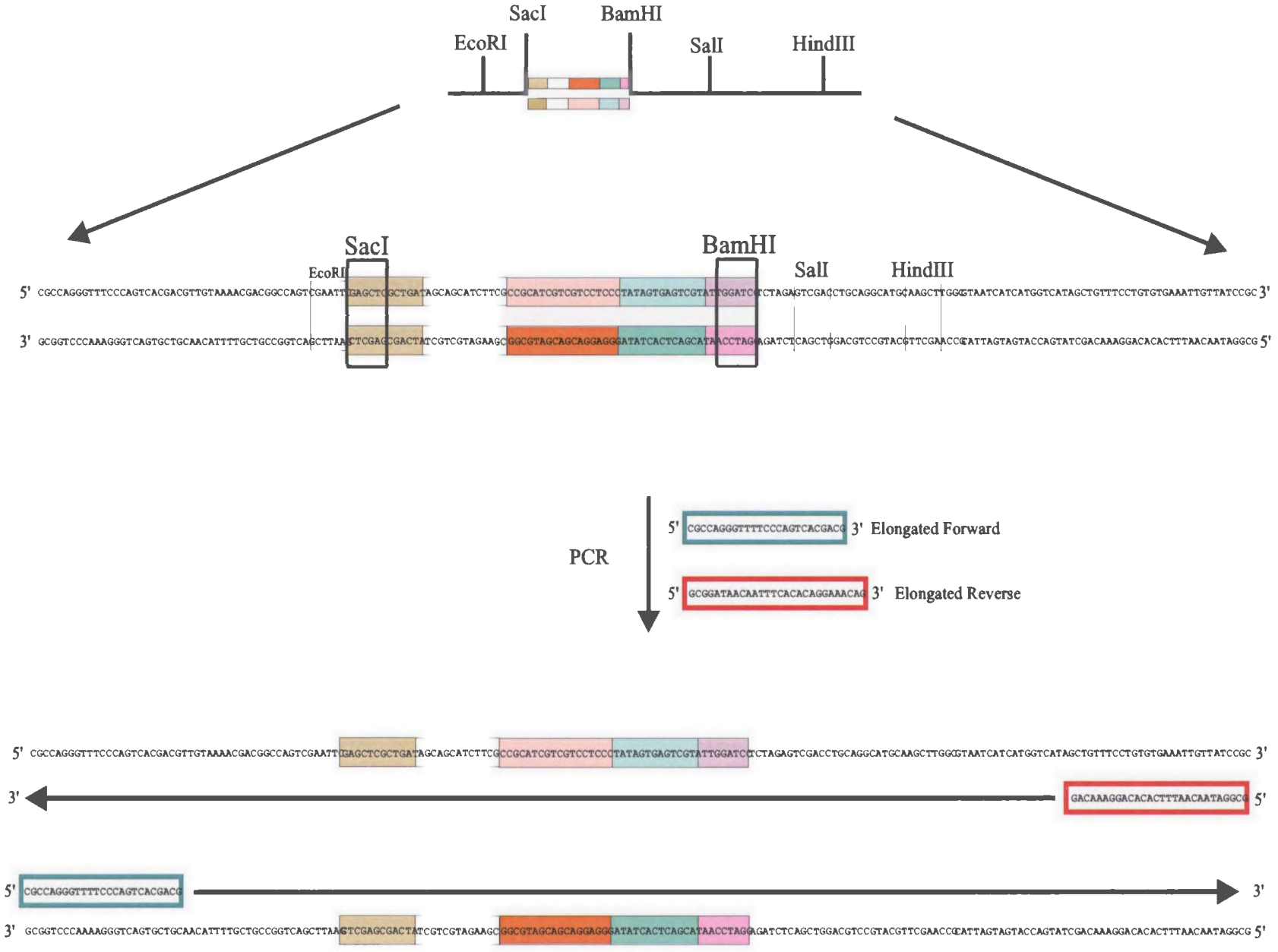
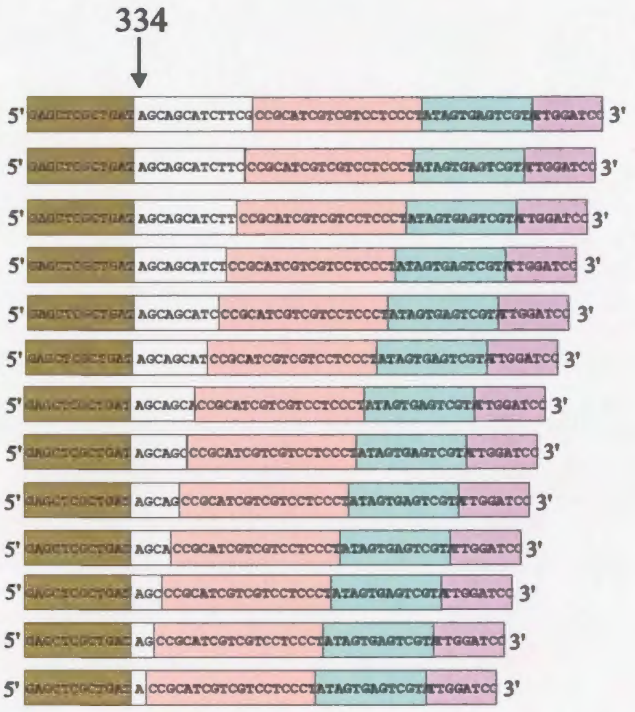

SacI
BamHI
EcoRI
SalI
HindIII
5' CGCCAGGGTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTCGAATTCGAGCTCGCTGATAGCAGCATCTTCGCCGCATCGTCGTCCTCCCTATAGTGAGTCGTATTGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGGGTAATCATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGC 3'
3' GCGGTCCCAAAGGGTCAGTGCTGCAACATTTTGCTGCCGGTCAGCTTAAGTCGAGCGACTATCGTCGTAGAAGCGGCGTAGCAGCAGGAGGGATATCACTCAGCATAACCTAGGAGATCTCAGCTGGACGTCCGTACGTTCGAACCGTATTAGTAGTACCAGTATCGACAAAGGACACACTTTAACAATAGGCG 5'
PCR
5' CGCCAGGGTTTTCCCAGTCACGACG 3' Elongated Forward
5' GCGGATAACAATTTCACACAGGAAACAG 3' Elongated Reverse
GACAAAGGACACACTTTAACAATAGGCG

Figure 3.16. Schematic diagram of the results generated by nested-PCR analysis of an ideal *Synechococcus* 7942 genomic SELEX library. There would exist molecules that contain, in a staggered arrangement, the nucleotide sequence upstream and including nucleotide 334 of the *rbpB* gene (figure 3.12). Colour coding is the same as that in figure 3.13.


334
5'
3'

3.5 Validation of the *Synechococcus* 7942 genomic SELEX library by assessing the representation of the *rbpB* gene: Outcome

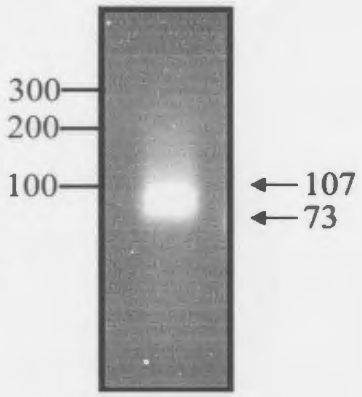
3.5.1 Nested-PCR amplification and identification of potentially recombinant clones

Nested PCR amplification of the *Synechococcus* 7942 genomic SELEX library generated dsDNA in the range of 73bp to 107bp in size (Figure 3.17A). This PCR product size range was consistent with the expected size range of 74bp to 117bp. Theoretically, the smallest PCR product would be generated from a library molecule that contains the complementary rbpB2 sequence directly adjacent to T7A and the largest PCR product would be generated from a library molecule with the largest genomic segment (88bp) that contained the complementary rbpB2 sequence located adjacent to T3B.

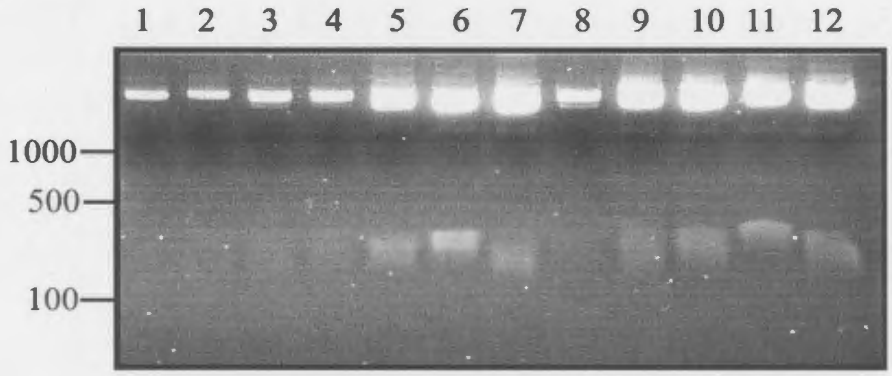
Nested PCR products were cloned into plasmid pUC19 and positive recombinant clones were identified using the two procedures as described in section 3.4.2. The restriction enzyme digestion method did not prove to be useful in detecting the existence and confirming the purity of an insert in a plasmid preparation due to the extensive smearing observed after gel electrophoresis (Figure 3.17B). The smearing of bands and the fact that a large amount of plasmid DNA was required to generate a questionable result prompted me to develop the PCR based method. Agarose gel electrophoretic analysis of the PCR product generated by use of the Elongated Forward and Elongated Reverse primers revealed that the PCR-based method was accurate in assessing the existence and purity of an insert in a plasmid preparation (Figure 3.17C). Non-recombinant plasmids were identified as a PCR product of approximately 140 bp in size. By this method, amplified plasmids shown in lanes 6, 7, 10 and 11 were identified as non-recombinant due to the existence of one band approximately 140bp in size and were thus

Figure 3.17. Agarose gel electrophoretic analysis used to identify true recombinant clones that contain nested-PCR product inserts. **(A)** Analysis of T7A(BamHI)/rbpB2 nested-PCR product. The experimental size range of PCR product as determined by use of Optiquant image analysis software is shown at right (arrows). **(B)** Identification of pUC19 plasmid molecules that contain inserts by agarose gel electrophoretic analysis following restriction endonuclease digestion with *EcoRI* and *HindIII*. Plasmid samples analyzed were isolated from colonies that were blue (lane 2), blue/white (lane 3) and white (lanes 3 to 12) in colour with use of non-recombinant pUC19 as a standard (lane 1). **(C)** PCR amplification with Elongated Forward and Elongated Reverse sequencing primers of plasmids isolated from colonies blue/white (lane 2) and white (lanes 3 to 15) in colour. The PCR product generated by non-recombinant pUC19 plasmid was used as a standard (lane 1). The location of 100bp DNA sized markers are indicated on the left side of each figure.

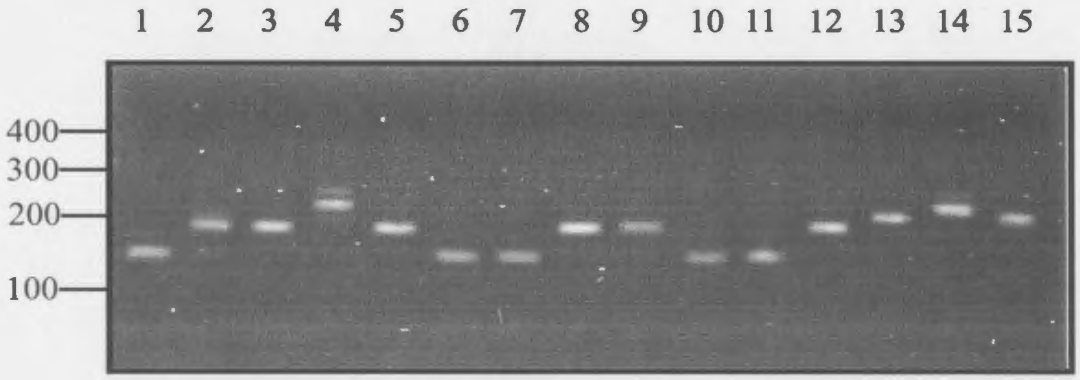
A.


300
200
100
107
73

B.


1 2 3 4 5 6 7 8 9 10 11 12
1000
500
100

C.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
400
300
200
100

excluded from dideoxynucleotide sequencing. In addition, the amplified plasmid preparation shown in lane 4 contained more than one recombinant plasmid as indicated by the existence of more than one band following PCR amplification. Therefore, this preparation was also excluded from dideoxynucleotide sequencing. In total, 48 plasmids were analyzed by the PCR-based method, 25 contained a single recombinant type, 12 contained more than one recombinant type and 11 were non-recombinant.

3.5.2 Sequence analysis of cloned nested-PCR product

Of the 25 plasmid preparations that were identified with a single insert, 20 were sequenced. Of these 20 sequences, 15 contained a portion of the *rbpB* gene sequence as listed in figure 3.18. The end-points of these sequences lie within nucleotides 334 to 322 of the *rbpB* gene as shown in figure 3.19. In total, 8 of 13 positions (61.5%) were represented in a staggered arrangement within the *Synechococcus* 7942 genomic SELEX library (Figure 3.19A). Assuming that both strands of the entire genome of *Synechococcus* 7942 (5.5×10^6 bases) are as equally represented in the library as nucleotides 334 to 322 of the *rbpB* gene, the genomic SELEX library would be composed of 3.38×10^6 different molecules. This corresponds to one library molecule per 1.6 bases of the *Synechococcus* 7942 genome (Figure 3.19B).

The genomic representation in the *Synechococcus* 7942 genomic SELEX library was similar to that of other genomic SELEX libraries as listed in figure 3.19C. The percentage of analyzed nucleotide sequence represented in the library was comparable in each case. A notable difference is the length of sequence analyzed, since I analyzed the shortest sequence length of 13 nucleotides. The reason why a larger region of variability was not observed was unclear. The largest size of genomic sequence encoded by a library molecule was 88 bp (Figure 3.10B). Of this, the rbpB1 and rbpB2 primers hybridize to 44 bases (taking into account the 11 nucleotide overlap of nested primers) in molecules

Figure 3.18. Results of sequence analysis of 15 recombinant clones containing an *rbpB* nested PCR amplified insert. The sequence of the insert is shown starting from the *SacI* restriction site in the rbpB2 primer sequence followed by various lengths of genomic sequence (white) corresponding to nucleotides 322 to 334 of the *rbpB* gene and ending in the complementary T7A(BamHI) primer sequence. The colour coding is the same as in figure 3.13.

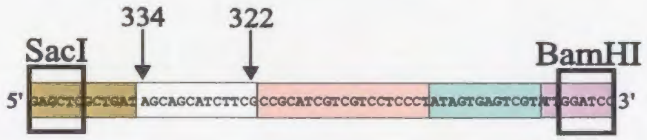
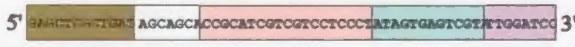
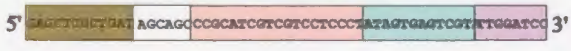
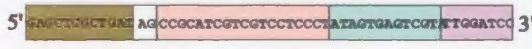
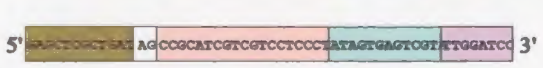
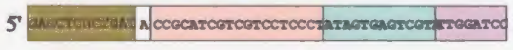
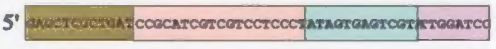
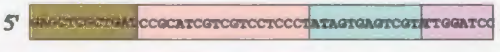
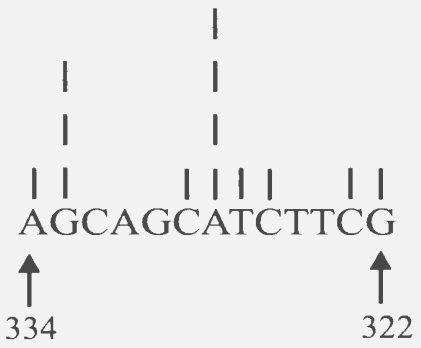
Clone number	Sequence
17	
23	
9	
39	
4	
8	
43	
16	
45	
10	
48	
41	
12	
22	
15	

Figure 3.19. Representation of the *rbpB* gene in the genomic SELEX library and a comparison of the results to that of other libraries. **(A)** Representation of nucleotides 334 to 322 of the *rbpB* gene in the *Synechococcus* 7942 genomic SELEX library. Each vertical bar represents an individual molecule whose genomic sequence portion ends in the specified nucleotide. **(B)** Calculations based on the information shown in part A. In total, 8 of 13 positions (61.5%) were represented in a staggered arrangement within the genomic SELEX library. Assuming that both strands of the entire genome of *Synechococcus* 7942 ($2 \times 3.5 \times 10^6$ bases) are as equally represented in the library as nucleotides 334 to 322 of the *rbpB* gene, the genomic SELEX library would be composed of 3.38×10^6 different molecules. This corresponds to one library molecule per 1.6 bases of the *Synechococcus* 7942 genome. **(C)** The results of nested PCR analysis of other genomic SELEX libraries.

A.



B.

- 8 of 13 nucleotides represented

- Calculations

⇒ 8/13 X 100% = 61.5%

⇒ $2 \times 2.75 \times 10^6$ X 0.615 = 3.38×10^6

⇒ 0.615 molecules per base = 1 molecule per 1.6 bases

C.

Organism	**gene analyzed**	**length of sequence analyzed**	**number of positions represented**	**% of positions represented**	**Reference**
Synechoccocus	*rbpB*	13	8	62	This work
E. coli	*metB*	33	23	70	Singer *et al.*, 1997
Yeast	*NDC1*	29	19	66	Singer *et al.*, 1997
Human	*ada*	34	18	53	Singer *et al.*, 1997
	U1A	55	28	51	

containing nucleotides 378 to 335 of the *rbpB* gene. Therefore, theoretically the longest possible nucleotide sequence upstream of nucleotide 334 was 44 bp (88 minus 44) and not 13 bp as observed.

My analysis indicated that my genomic SELEX library is a good representation of the *Synechococcus* 7942 genome, that is, all genomic sequences and thus potential RNA transcripts are encoded in the library. Therefore, the library is suitable to identify RNA aptamers recognized by RNA-binding protein A (RbpA) via the SELEX method as will be described in Chapter 4.

Isolation of H_6RbpA RNA aptamers by genomic SELEX

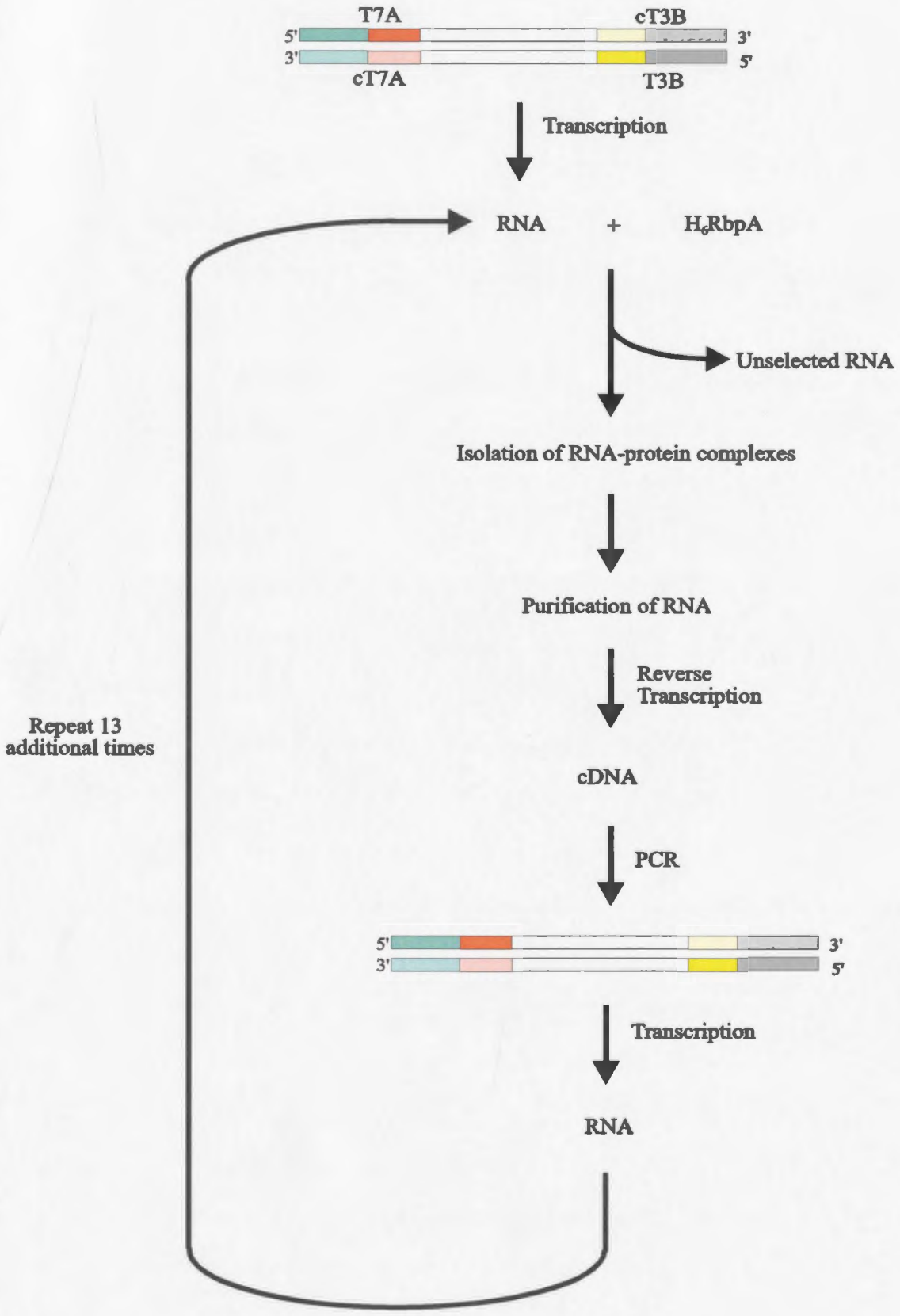
4.1 Introduction

The SELEX technique is used to select nucleic acid molecules that possess a desired characteristic such as an ability to be bound by a nucleic acid binding protein. The procedure involves multiple rounds of the same basic steps. First a library of nucleic acid molecules are incubated with the protein of interest, those molecules that are bound by the protein are captured and retained, those not bound by the protein are removed. Nucleic acid molecules retained are amplified and become the pool used in the next round of selection. In principle, with each round of SELEX, true nucleic acid aptamers constitute a greater proportion of the pool of molecules used at the start of each round. In this way, an enrichment of nucleic acid sequence(s) that bind to the protein is generated with each round.

4.2 The SELEX technique: strategy

Once I had established that the *Synechococcus* 7942 genomic SELEX library that I had constructed was sufficiently representative of the entire genome, I was able to use the library to determine the RNA binding specificity of H_6RbpA. The SELEX procedure outlined in Figure 4.1 was initiated by transcription of the genomic SELEX library, and the resultant RNA was mixed with H_6RbpA in an RNA-protein binding reaction. In the next step, RNA-protein complexes were isolated by Ni^{+2}-NTA metal chelate chromatography via the histidine tag located on the N-terminus of H_6RbpA. The RNA was purified from the protein, reverse-transcribed and PCR amplified to generate a new pool of dsDNA molecules, this pool was subsequently transcribed to generate RNA for the next round of selection. In addition to the H_6RbpA binding reaction, a control reaction consisting of RNA without H_6RbpA was conducted as a negative control. This procedure was repeated 13 times to promote selection and thus identification of RbpA cognate

Figure 4.1. Outline of the steps involved in the SELEX procedure. Colour coding is the same as in figure 3.13.


T7A
cT3B
5'
3'
3'
5'
cT7A
T3B
Transcription
RNA + HcRbpA
Unselected RNA
Isolation of RNA-protein complexes
Purification of RNA
Reverse Transcription
cDNA
PCR
5'
3'
3'
5'
Transcription
RNA
Repeat 13 additional times

RNA.

4.2.1 Step 1: transcription of the library

RNA was generated for each round of SELEX by T7 RNA polymerase transcription of the genomic SELEX library (round 1), or the PCR amplified cDNA generated in each round (refer to section 2.17).

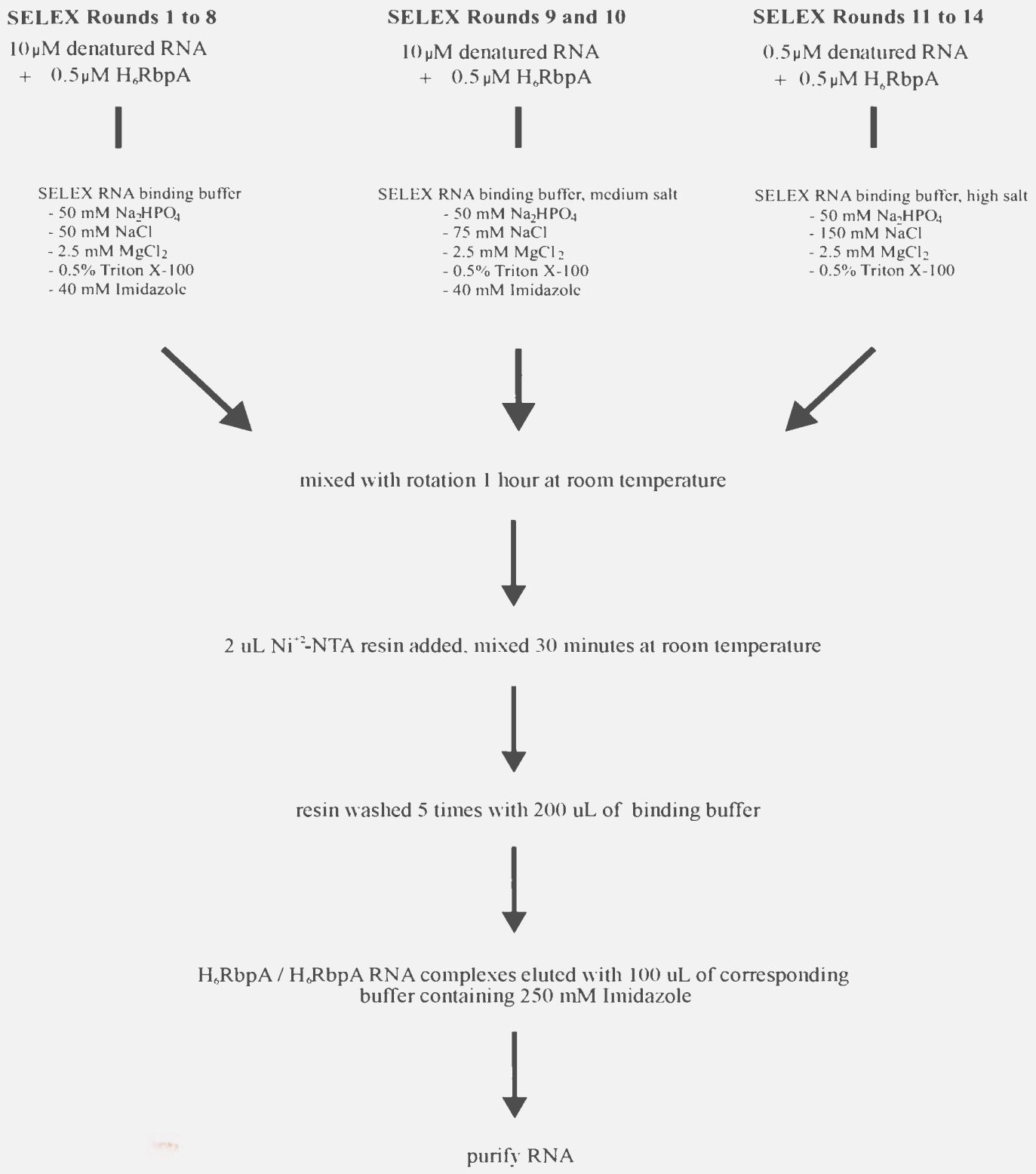
4.2.2 Step 2: The RNA-protein binding reaction

Selection of H_6RbpA binding RNA molecules was carried out by incubation of RNA with H_6RbpA in a protein-RNA binding reaction (section 2.19). To select the RNA molecules with greatest affinity to the protein, the stringency of the reactions was altered in later rounds (Figure 4.2) by varying the concentration of some components of the binding reaction, namely the molar ratio of RNA to protein and the concentration of salt in the binding buffer.

In the initial eight rounds of SELEX, reaction conditions of low stringency were used to promote complex formation between protein and RNA. These conditions used a high RNA to protein ratio (20:1) and low salt concentration (50mM) in the reaction buffer. A 20:1 molar ratio of RNA to protein was employed because in the initial rounds of SELEX, the number of each library molecule was low (sequence diversity is high). In addition, each molecule may not constitute an equal proportion of the library, therefore a large amount of RNA was added to ensure that the protein has an opportunity to bind to any RNA sequence that it recognizes with some affinity.

In the intermediate and late rounds of selection, the stringency of the binding reaction was increased substantially. In the intermediate rounds (rounds 9 and 10), a 25mM increase in salt concentration within the binding buffer was used, and in the late rounds of selection, stringency was raised by both an increase in salt concentration (to

Figure 4.2. Methodology and RNA-H_6RbpA binding reaction conditions used in the various rounds of the genomic SELEX process. Note that components of the three SELEX RNA binding buffers (SRBB) are identical with exception of the concentration of NaCl and the absence of 40mM Imidazole in the buffer used in rounds 11 to 14. The ratio of RNA to H_6RbpA was also varied: a ratio of 20:1 was used in rounds 1 to 10 and a ratio of 1:1 was used in rounds 11 to 14.


SELEX Rounds 1 to 8
10 µM denatured RNA
+ 0.5 µM H6RbpA
SELEX Rounds 9 and 10
10 µM denatured RNA
+ 0.5 µM H6RbpA
SELEX Rounds 11 to 14
0.5 µM denatured RNA
+ 0.5 µM H6RbpA
SELEX RNA binding buffer
- 50 mM Na2HPO4
- 50 mM NaCl
- 2.5 mM MgCl2
- 0.5% Triton X-100
- 40 mM Imidazole
SELEX RNA binding buffer, medium salt
- 50 mM Na2HPO4
- 75 mM NaCl
- 2.5 mM MgCl2
- 0.5% Triton X-100
- 40 mM Imidazole
SELEX RNA binding buffer, high salt
- 50 mM Na2HPO4
- 150 mM NaCl
- 2.5 mM MgCl2
- 0.5% Triton X-100
mixed with rotation 1 hour at room temperature
2 uL Ni+2-NTA resin added, mixed 30 minutes at room temperature
resin washed 5 times with 200 uL of binding buffer
H6RbpA / H6RbpA RNA complexes eluted with 100 uL of corresponding buffer containing 250 mM Imidazole
purify RNA

150mM) as well using equimolar amounts of RNA.

The remaining components of the RNA binding buffer (Figure 4.2) are based on the buffer used in RNA homopolymer binding analysis of RbpA (Section 2.18) initially described by Sugita and Sugiura (1994). Under these reaction conditions RbpA exhibits RNA homopolymer binding activity, however some alterations were required for the SELEX application. First, a sodium phosphate buffer was used instead of a Tris buffer because the secondary or tertiary amines found in buffers such as Tris will reduce nickel ions on Ni-NTA, thereby preventing efficient binding of His-tagged protein (Ausubel *et al.*, 1996). Second, in rounds 1 to 10, Imidazole (40mM) was added to the buffer in an attempt to prevent non-specific interactions between negatively charged RNA and Nickel cations in the resin as described in the next section (Qiagen, personal communication).

4.2.3 Step 3: Isolation of RNA-H_6RbpA protein complexes

RNA-protein complexes formed in the binding reaction must be purified from free RNA and protein so only those RNA molecules bound to H_6RbpA are retained. The method generally used in SELEX experiments uses nitrocellulose filter binding (Tuerk and Gold, 1990) to retain RNA-protein complexes on the filter while uncomplexed RNA and protein molecules pass through. In my experiments, the histidine tag present at the N-terminus of RbpA was used to purify RNA-protein complexes by use of Ni^{+2}-NTA affinity chromatography (refer to section 2.19).

One problem with this purification strategy is the background retention of RNA, that is, selection of RNA molecules independent of protein by non-specific retention of RNA during purification. Using Ni^{+2}-NTA to select RNA-protein complexes, background retention of RNA is possible via an interaction between negatively charged RNA and positively charged Nickel ions on the resin. To minimize the potential interaction between RNA and the resin, a few precautions were used. First, the smallest amount of resin that

could be detected visually during purification (enough to bind 0.6 nmol of a 20kDa protein, Qiagen) was used. Second, the highest concentration of Imidazole (40mM) that would not elute H_6RbpA from the resin (refer to section 2.10) was used because Imidazole may be helpful in preventing an interaction between RNA and the Ni^{+2}-NTA resin (Qiagen, personal communication). Imidazole was excluded from the binding buffer in rounds 11 to 14 because the RNA molecules that exist at the higher rounds of SELEX should be bound tightly by the protein (high affinity) and therefore non-specific, weak interactions with the resin should not be a factor. In addition, in rounds 10 to 14 the stringency of the reaction was high, thereby discouraging the formation of weak interactions between RNA and resin.

4.2.4 The remaining steps in a round of SELEX

Following purification of RNA-protein complexes, a round of SELEX (figure 4.1) was completed by purification of RNA selected in the binding reaction by use of an RNeasy column (Qiagen) (section 2.5). The purified RNA was then reverse-transcribed (section 3.2.9) and the resultant cDNA product was PCR amplified using primers T7A and T3B (section 3.2.10). DNA was then purified by use of the QIAquick DNA purification system (Qiagen, section 2.4) and transcribed by T7 RNA polymerase to generate RNA for the next round of SELEX.

4.2.5 Methods used to determine if selection is taking place

Prior to sequence analysis of nucleic acids isolated by SELEX, it would be advantageous to determine if the SELEX technique is working. In this way, one could find out if selection was occuring and make adjustments as necessary to ensure selection. Two methods that I used to evaluate nucleic acid sequence selection are discussed below.

The amount of RNA isolated in the presence of H_6RbpA versus that in the absence of H_6RbpA can be used to indicate if selection is taking place and if RNA is being bound

by protein or the resin. One would expect that with each round of SELEX, the amount of RNA isolated in the presence of H_6RbpA would increase due to the existence of a greater proportion of RNA sequences that bind to the protein in the pool of RNA molecules until the number of bound RNA molecules equals that of H_6RbpA. If RNA was isolated independent of protein via a specific interaction with the Ni^{+2}-NTA resin, the amount of RNA isolated in the presence and absence of protein would increase with each round of SELEX due to the selection of RNA molecules that bind to the resin. To determine the amount of RNA isolated with and without protein, RNA concentration was estimated quantitatively by spectrophotometry (section 2.5) and qualitatively by gel electrophoresis of "real-time" RT-PCR products. Real-time RT-PCR served as a more sensitive method of assessing the amount of RNA isolated from RNA-protein complexes. Comparison of band intensity during the course of the reaction can assess the relative amount of RNA in the two samples, as described in section 3.3.9.

A second method that can be used to indicate selection of RNA sequences in the SELEX process is to observe the loss of a particular nucleic acid sequence from the pool of nucleic acids present at the start of each round of SELEX. If a particular nucleic acid sequence is not bound by the protein, the absence of an amplified PCR product for that sequence would be evidence of loss of the sequence. The same strategy for analyzing the genomic representation of the SELEX library can be applied to the problem of analyzing selection during the SELEX process. Therefore, amplification of library molecules containing nucleotides 354 to 378 of the *rbpB* gene was used for this purpose. Agarose gel electrophoretic analysis of the PCR product (utilizing primers T7A and rbpB1, section 3.4) produced using the pool of DNA molecules generated at the end of each round of SELEX should indicate if this sequence has been removed and thus if selection is taking place assuming, that is, that H_6RbpA does not bind to the sequence. We did not know if

nucleotides 354 to 378 of the *rbpB* gene are actually recognized by RbpA. This sequence was selected based solely on the fact that we had the primers to amplify this sequence. If H_6RbpA actually binds to this sequence or selection is not taking place at all, a PCR amplified product will be observed.

4.2.6 Cloning of selected nucleic acid molecules into pUC19 and identification of recombinant clones

Members of the genomic SELEX library selected following multiple rounds of SELEX were sequenced to determine the binding specificity of H_6RbpA. To accomplish this, the pool of molecules generated following the final round of SELEX were PCR amplified utilizing the T7A(BamHI) and B(BamHI) primers which hybridize to primer sequences T7A and T3B respectively (Figure 4.3) and incorporate a *BamHI* restriction site at both ends of a DNA molecule. *BamHI* restriction digestion of the library was used to clone the fragments into pUC19 (Figure 4.4). Prior to nucleic acid sequence determination, plasmid preparations that contained a single recombinant plasmid were identified by PCR amplification using the Elongated Forward and Elongated Reverse primers as described in section 3.4.2. In this case however, the sole determinant of a desired clone for sequencing is the presence of an insert, regardless of its size.

4.3 Isolation of RbpA RNA aptamers by genomic SELEX : outcome

4.3.1 Did selection take place?

The amount of RNA isolated in each round of SELEX in the presence and absence of protein was determined in an effort to obtain evidence of RNA sequence binding mediated by H_6RbpA. Quantitative analysis of the amount of RNA was determined spectrophotometrically and the results are given in Table 4.1. The amount and percentage of RNA retained in the absence of protein (background) remained relatively constant through rounds 1 to 10 as would be expected if the Ni^{+2}-NTA resin has no effect on RNA

Figure 4.3. Introduction of *BamHI* restriction endonuclease sites after multiple rounds of genomic SELEX. The final library of molecules selected after multiple rounds of genomic SELEX are PCR amplified using the primers T7A(BamHI) and B(BamHI) (light yellow, sequence different than primer T3B is shown in dark yellow). This introduces *BamHI* restriction sites at each end of the molecule so that they can be cloned. Other colour coded sequences are as shown in figure 3.13.

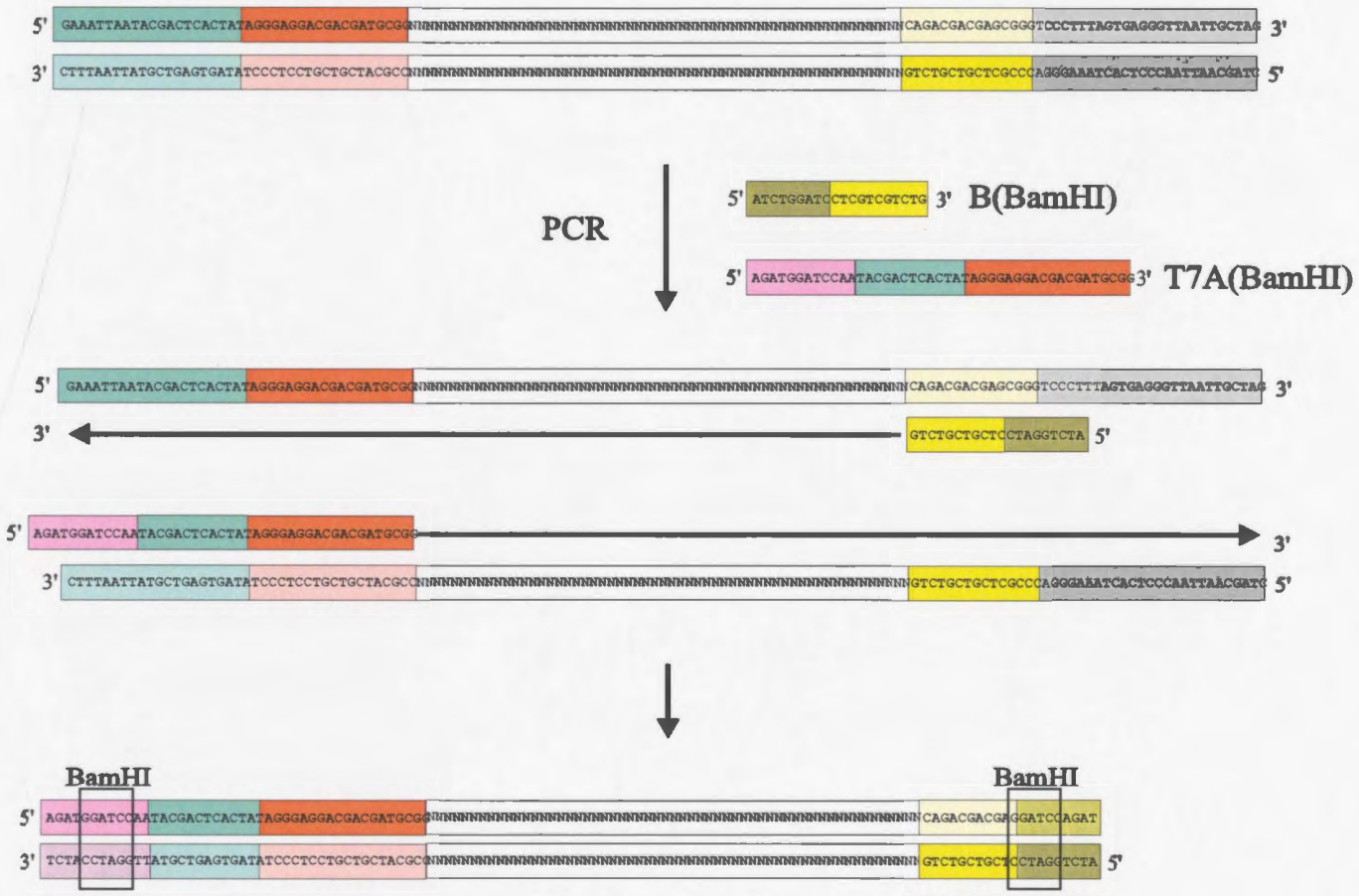

PCR
B(BamHI)
T7A(BamHI)
BamHI
BamHI

Figure 4.4. Cloning of molecules isolated following rounds 10 and 14 of genomic SELEX. The PCR amplification products containing different portions of the *Synechococcus* 7942 genome were cloned into pUC19 by use of two *BamHI* restriction sites introduced at both ends of the fragment. Since the fragment is not directionally cloned, it can be inserted into the plasmid in two orientations as shown. Colour coding is the same as in figure 3.13.

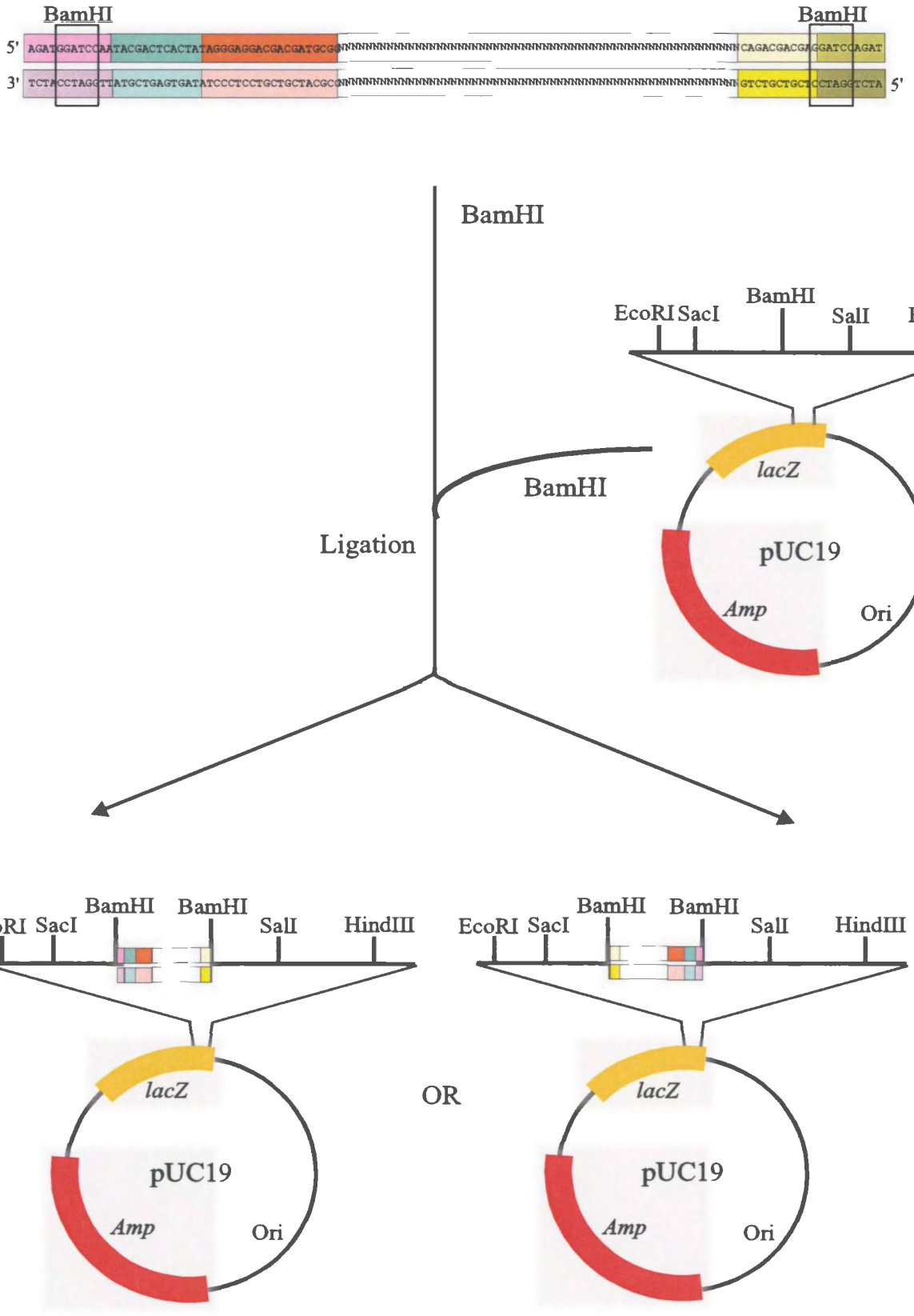

BamHI
BamHI
5' AGATGGATCCAATACGACTCACTATAGGGAGGACGACGATGCGN
CAGACGACGAGGATCCAGAT
3' TCTACCTAGGTTATGCTGAGTGATATCCCTCCTGCTGCTACGCN
GTCTGCTGCTCCTAGGTCTA 5'
BamHI
EcoRI SacI
BamHI
SalI
HindIII
lacZ
BamHI
Ligation
pUC19
Amp
Ori
EcoRI SacI
BamHI
BamHI
SalI
HindIII
lacZ
pUC19
Amp
Ori
OR
EcoRI SacI
BamHI
BamHI
SalI
HindIII
lacZ
pUC19
Amp
Ori

Table 4.1 Amount of RNA and H_6RbpA in the binding reaction of each round of genomic SELEX and the amount of RNA retained in the presence and absence of protein as determined by spectrophotometric quantification.

Round	RbpA (nM)	RNA		H_6RbpA absent		H_6RbpA present	
		ug	uM	Amount retained (ng) *	Percentage retained (%)	Amount retained (ng) *	Percentage retained (%)
1	500	40	10	n/a	n/a	320	0.797
2	500	40	10	520	1.29	520	1.29
3	500	40	10	400	1.00	600	1.49
4	500	40	10	500	1.24	500	1.24
5	500	40	10	600	1.49	750	1.87
6	500	40	10	600	1.49	550	1.37
7	500	40	10	350	0.871	550	1.37
8	500	40	10	450	1.12	450	1.12
9	500	40	10	550	1.37	550	1.37
10	500	40	10	550	1.37	500	1.25
11	500	1	0.5	260	26	260	26
12	500	1	0.5	240	24	260	26
13	500	1	0.5	220	22	240	24
14	500	1	0.5	220	22	240	24

* Concentration (ng/uL) = (absorbance 260 nm - absorbance 340 nm) x 40 ng/uL x dilution factor

sequence selection.

Qualitative analysis of the retention of RNA in each round of the SELEX procedure by "real-time" RT-PCR provided an indication of protein-mediated selection. In 10 of the 12 rounds in which qualitative analysis was performed, a greater amount of PCR-amplified DNA was generated from cDNA template reverse-transcribed from RNA isolated in the presence of H_6RbpA (Figure 4.5A) than that isolated in the absence of protein. For instance, RT-PCR amplification of RNA selected in the tenth round of SELEX revealed a stronger signal (band) from RNA template isolated in the presence of protein than that isolated in the absence of protein following 10 rounds of PCR amplification (Figure 4.5B). In RNA samples isolated in round 8 (Figure 4.5C) and round 4 of SELEX, the opposite was observed as RT-PCR revealed a greater signal generated from RNA isolated in the absence of protein than in the presence of protein. Therefore, qualitative RT-PCR analysis was not definitive in determining whether RNA retained in the selection process was due to an interaction with H_6RbpA, however, since 10 of 12 rounds indicated protein-mediated binding of H_6RbpA, it was more likely that the protein was involved in RNA sequence selection.

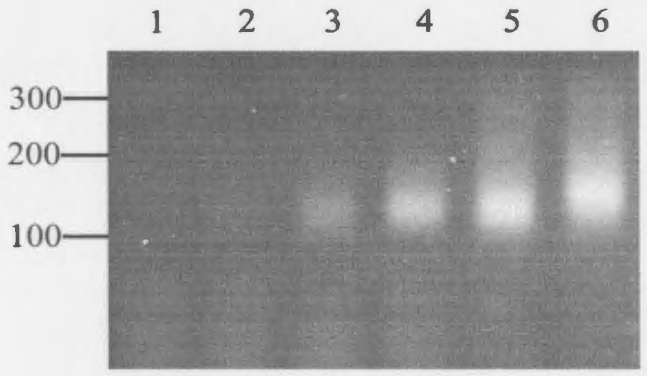
Loss of molecules containing nucleotides 354 to 378 of the *rbpB* gene from the pool of double-stranded cDNA molecules generated in rounds 1 to 9 of SELEX provided better evidence that nucleic acid sequence selection was taking place. As shown in figure 4.6, a high intensity band was observed in the lane corresponding to the double-stranded cDNA product of PCR amplification utilizing the T7A and rbpB1 primers generated in round one of SELEX. However, the amount of PCR amplified product decreased when cDNA templates generated from the later rounds of SELEX were employed. Therefore, this result suggested a gradual selection against molecules containing nucleotides 354 to 378 of the *rbpB* gene throughout the SELEX process and thus provided a positive

Figure 4.5. Results of the qualitative "real-time" RT-PCR method used to detect H_6RbpA-mediated retention of RNA during the SELEX process. (A) Table of results generated from RT-PCR analysis of the amount of RNA retained in the presence or absence of protein in the binding reaction in each round of genomic SELEX as shown for rounds 10 and 8 below. The greater amount of DNA present in one of the two samples as detected upon agarose gel electrophoresis following PCR amplification is denoted by a plus (+) sign. (B) Agarose gel electrophoresis of "real-time" RT-PCR products generated from the RNA isolated in round 10 of genomic SELEX in the absence (lanes 1, 3 and 5) and presence (lanes 2, 4 and 6) of protein. Samples taken following 20 cycles (lanes 1 and 2), 25 cycles (lanes 3 and 4) and 30 cycles (lanes 5 and 6) of RT-PCR were separated by agarose gel electrophoresis and compared in terms of band intensity. Since samples were taken at points during the RT-PCR amplification reaction, the analysis is referred to as "real-time". (C) Agarose gel electrophoresis of "real-time" RT-PCR products generated from RNA isolated in round 8 of SELEX in the absence (lanes 1, 3 and 5) and presence (lanes 2, 4 and 6) of protein. Samples of the RT-PCR were taken following 20 cycles (lanes 1 and 2), 25 cycles (lanes 3 and 4) and 30 cycles (lanes 5 and 6). The location of 100bp markers are indicated at left.

A.

Round	Background (absence of protein)	Reaction (presence of protein)
1		+
2		+
3		+
4	+	
5		+
6		+
7		+
8	+	
9		+
10		+
11		+
12	n/a	n/a
13	n/a	n/a
14		+

B.



C.

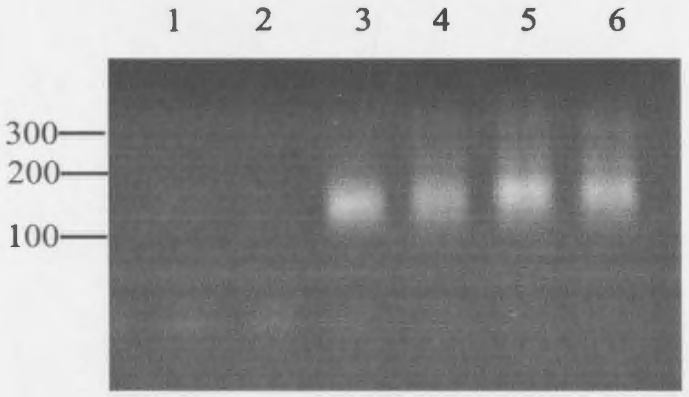
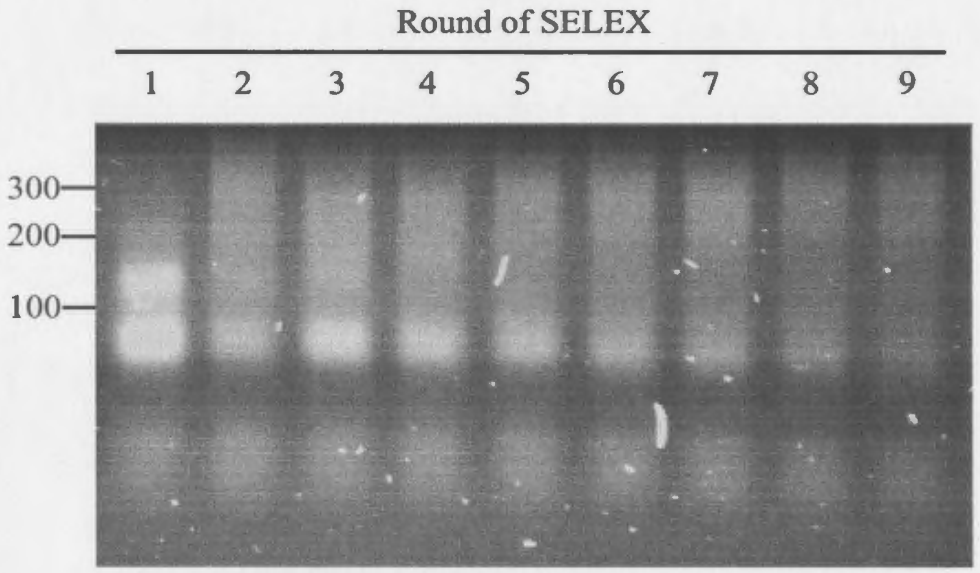


Figure 4.6. Loss of *rbpB* sequences from the library during genomic SELEX with H_6RbpA. Agarose gel electrophoresis of dsDNA generated by PCR amplification of the pool of library molecules used in genomic SELEX rounds 1 to 9 with primers T7A and rbpB1. The resultant PCR products were visualized by agarose gel electrophoresis. Location of 100bp markers are indicated at left.


Round of SELEX
1
2
3
4
5
6
7
8
9
300
200
100

indication of nucleic acid sequence selection. However, it does not provide evidence of the mode of selection, that is, whether the protein was involved in selection or not.

4.3.2 Cloning of round 10 and round 14 double-stranded cDNA into pUC19 and identification of potentially positive result generating clones

The pools of DNA generated following rounds 10 and 14 of genomic SELEX were cloned and sequenced as described in section 4.2.6. Agarose gel electrophoretic analysis of the PCR product generated by amplification of plasmid DNA with primers T7A(BamHI) and B(BamHI) was successful in identifying recombinant and non-recombinant plasmids as well as heterogeneous plasmid preparations. Non-recombinant plasmids were identified by the presence of one band 142bp in size (Figure 4.7, control plasmid lane 1). Recombinant plasmids that contained one insert were identified by the presence of one band of size greater than 142bp (Figure 4.7, clones 11, 12, 14, 15 to 18). The presence of more than one band identified a recombinant plasmid preparation that contained more than one insert sequence (Figure 4.7, clones 13, 19, and 20), these clones were therefore excluded from sequence analysis.

4.3.3 Nucleic acid sequences selected by genomic SELEX

Following round 10 of SELEX, the library was cloned and sequenced. Of the 36 sequenced clones, 22 provided reliable sequence information (*i.e.* few unidentified bases). Individual sequences are given in table 4.2. Comparative sequence analysis (section 2.16) of these 22 sequences (refer to Chaper 5) did not provide an indication of a consensus sequence recognized by RbpA, therefore in an effort to maximize the selection process, an additional four rounds of SELEX were performed in reaction conditions of higher stringency (section 4.2.2). Following round 14, 34 cloned library molecules were sequenced of which 27 provided reliable sequence information (Table 4.3) The additional four rounds of selection proved to be beneficial as evidenced by the dramatic increase in

the similarity of those sequences obtained (refer to section 5.1).

Figure 4.7. Identification of recombinant clones by agarose gel electrophoresis of the PCR amplified product. Recombinant pUC19 molecules that contained an insert were amplified with primers T7A(BamHI) and B(BamHI) and analyzed by agarose gel electrophoresis of the product. PCR amplification of non-recombinant plasmid, used as a comparative control (C), generated a 142bp fragment. Clones 11, 12, 14, 15, 16, 17, and 18 are recombinant and contain one library molecule type as indicated by a single band whereas clones 13, 19 and 20 are recombinant but contain more than one insert as evidenced by the presence of multiple bands. Location of 100bp markers are indicated at left.

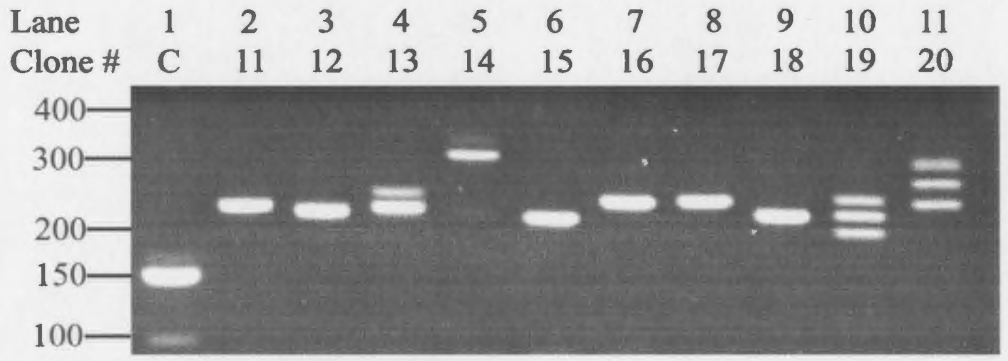

Lane 1 2 3 4 5 6 7 8 9 10 11
Clone # C 11 12 13 14 15 16 17 18 19 20
400
300
200
150
100

Table 4.2 Genomic SELEX round 10 RNA sequences

	Clone Number	Genomic sequence length (bases)	% G/U	Sequence*
G/U rich‡	1	57	61	5' UGAAAUUAAUACGACUCACUAUAGGGAGGACGAUGCGGAUGGUGUGAUGUGUGGGUG 3'
	4	32	84	5' UGCUGGGGGACGAGCUGGGUGUGGUGUUGUGG 3'
	6	26	85	5' UGGCUUGGGCAGUGGUGUGGAGUUUG 3'
	7	21	67	5' AUGAUGGGGUUCAGCCAUGGU 3'
	12	27	88	5' UCUGUGGGGACGUUGUUGUGGGUGGUG 3'
	13	33	64	5' UGUGACAAGAUGAAAGCGAUUGGGAGGUGGUGC 3'
	14	24	83	5' GUAGGUGGUGGGAAUCGGUUUGGG 3'
	26	37	70	5' UGUGUUGGUUUGCUUACUGAGCUUGUCGCCAUUCGGC 3'
	29	23	74	5' UGAGGGGUAGUUGGUUAGGGCAA 3'
	30	29	76	5' GUGGUUGCAGGACUGAUGGUGUUGCGGGC 3'
	41	27	67	5' AUAGGGUUAGCGAUGAUGCCUUGGUGC 3'
	45	25	60	5' CUCUGGUCUGCAGCCCUAGUGAUGG 3'
G/U very rich ‡	5	45	98	5' UGUGUGGUGUGUGGUGUGUGGUGUGUGGUGUGUGGUGUGGGUGUA 3'
G/U poor‡	8	23	30	5' ACUGAGCUGCCGUAACCAAACAC 3'
	10	25	36	5' UGGCCGACUAAACCAUCACCACUGG 3'
	11	51	41	5' UAAUGCAAGCCAGCUCCAGCGGGUUCGCAGCAGUCGAGCGCGAGGCAGACC 3'
	19	44	46	5' UNUGNNAGCGANCGCCAGAGGAGAUAGCGCACCGUCAGCAGNGC 3'
	23	31	23	5' ACUGAGCUGCCGUAACCAAACACCACCACAC 3'
	25	16	25	5' GGAUCCACCACCAGCC 3'
	42	44	33	5' UAGGUCAGCACUAUCACUAGGGAGAGAACACCCANUCCCAGNCA 3'
	43	27	31	5' ACCGAGCUGCCGUAACNCGAGCACCAC 3'
	49	24	29	5' CCUCAAGCUUGGAACCACAGCCAC 3'

* sequence is the *Synechococcus* 7942 genomic sequence portion of each libary molecule, primer sequences are not given

‡ sequences are grouped according to composition of GTP and UTP residues. G/U very rich sequences contain greater than 90% G/U, G/U rich sequences contain between 50% and 90% G/U and G/U poor sequences contain less than 50% G/U residues

Table 4.3 Genomic SELEX round 14 RNA sequences

	Clone Number	Genomic sequence length (bases)	% G/U	Sequence [+]
G/U rich [‡]	1	24	75	5′ UCUAGUCGGGGUGAAUGGGAGGUG 3′
	3	41	78	5′ UGUAUGGGANGUGGACGUAGGUUGUGGUUUCGUUGGACUGG 3′
	6	25	84	5′ UGGUGGUGAUGAUAGAGGUGGUGGU 3′
	7	40	70	5′ UGACUGGCUUGGCGGUGAGUGGAGUGAAUGGGAAGUGUAC 3′
	11	32	69	5′ ACCUGGUCAGAGUGGAGGUGAUGGUGGCGGUA 3′
	16	41	80	5′ UGUAUGGGAGGUGGACGUAGGUUGUGGUUCCGUUGGGCUGG 3′
	18	24	79	5′ UUUAGUCGGGGUGAAUGGGAGGUG 3′
	36	15	73	5′ UGUAUGGGAGGUGAA 3′
	37	24	79	5′ UUUAGUCGGGGUGAAUGGGAGGUG 3′
	39	26	69	5′ GACUGGAGUGAAGUGGGAGGUGUAAU 3′
	47	24	79	5′ UUUAGUCGGGUGAAUGGGAGGUG 3′
	52	24	79	5′ UUUNGUCGGGGUGAAUGGGUGGNG 3′
	53	33	79	5′ CAGUGGUGGUGGAGAUAGUAGUGGUGUGAGUGU 3′
	54	24	79	5′ UUUAGUCGGGGUGAAUGGGAGGUU 3′
	56	41	78	5′ AAGCCGUUUGAGUGUGAGUGAUGGAGUGUUUGUGUGGGUG 3′
	59	24	75	5′ UUUAGUCGGGGUGAAUGGGAGGUU 3′
	64	34	68	5′ GUGCGAUCGCGAGGCUGCGUGAAUGGGAGGUGAU 3′
	67	43	70	5′ AUCUGCACCUAGAGCUGAUGCGUNUGGGUGGAUGGGGGGUGUU 3′
G/U very rich [‡]	35	36	92	5′ GACUGCGGUGUGUGGUGUGUGGUGUGUGGUGUGUGG 3′
	43	24	92	5′ NAGUGUGGUGUGUGGUGUGUGGUG 3′
G/U poor [‡]	5	25	52	5′ AACUCCGUCCCAGACGUGUUGGUGN 3′
	8	21	24	5′ AGCACCACAGCACCAAUGCAU 3′
	15	18	28	5′ GGAGCGGACACACCACCA 3′
	34	30	23	5′ AAGCCCACGACAGNNGACAAUAGCCCCAAU 3′
	45	24	25	5′ AGCCGGGCGACACCAGACACACAC 3′
	55	30	27	5′ ACCUGACGCUGCCGUAACCCAACACCACAU 3′
	63	29	34	5′ CGGCACCAACCAGAUCCGUACUCCGCAU 3′

+ sequence is the *Synechococcus* 7942 genomic sequence portion of each libary molecule, primer sequences are not given

‡ sequences are grouped according to composition of GTP and UTP residues. G U very rich sequences contain greater than 90% G U, G U rich sequences contain between 50% and 90% G U and G U poor sequences contain less than 50% G U residues

Results and Discussion

5.1 Analysis of nucleic acid sequences isolated by the genomic SELEX procedure

The RNA sequence binding specificity of RbpA was examined by analyzing the RNA sequences bound by RbpA (tables 4.2 and 4.3) in terms of base composition and the location of these sequences in the *Synechococcus* 7942 genome. First, the percentage of guanine and uracil nucleotides in each sequence was determined to see if RbpA preferentially bound to RNA sequences predominantly comprised of these bases (Section 5.2). Second, RNA sequences from rounds 10 and 14 were aligned independently to determine if a conserved RNA binding sequence was evident (Section 5.3). Third, the location in the *Synechococcus* 7942 genome of each sequence was determined to identify genes potentially regulated by RbpA (Section 5.6). Together, this information might provide an indication of the function of RbpA in the cyanobacterial cell.

5.2 Guanine and uracil nucleotide composition

The abundance of guanine and uracil nucleotides in an RNA sequence is important because like many RNP-type RNA-binding proteins, RbpA has been found to exhibit preferential binding to guanine and uracil nucleotide homopolymers (Belbin, 1999). Therefore, to determine if selective homopolymer binding specificity of RbpA was evident in the RNA sequences isolated by the SELEX procedure, the G/U content of each sequence was determined. Analysis of the RNA revealed the existence of three classes of RNA sequences in each set of data (rounds 10 and 14). They are: sequences that are largely not composed of G/U residues (G/U poor) (less than 50% G/U), sequences rich in G/U residues (between 50% and 90% G/U) and sequences very rich in G/U residues (greater than 90%) and (tables 4.2 and 4.3).

Comparison of the number of RNA sequences in each G/U class indicated that conditions of increased stringency likely generated a selection preference for RNA

molecules that are G/U rich. The percentage of G/U rich sequences that comprise the total number of RNA sequences isolated in round 14 (78%) is greater than that of round 10 (59%). In addition, two very-rich G/U sequences were isolated in round 14 whereas one was isolated in round 10 of selection. These results suggested that RbpA has a greater binding specificity for RNA sequences that contain a high percentage of guanine and uracil nucleotides. This finding however, does not rule out the possibility that RNA molecules that contain a low percentage of guanine and uracil residues are recognized by RbpA and are biologically relevant, it only corroborates previous data (Mulligan and Belbin, unpublished) regarding the binding specificity of RbpA, that the protein has a preference for RNA predominantly composed of guanine and uracil residues.

5.3 Sequence alignment of RNA from rounds 10 and 14

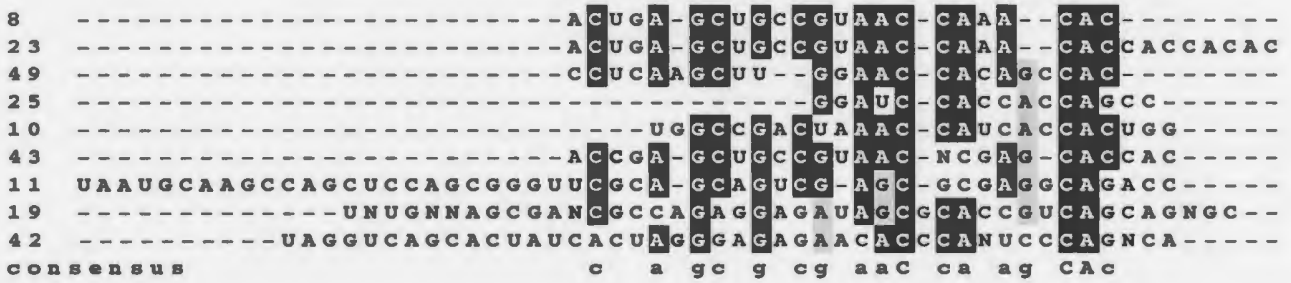
The next question regarding the sequence of RNA molecules obtained from rounds 10 and 14 of SELEX was the extent of sequence homology. If a protein has ribonucleic acid binding specificity, RNA sequences retained in the SELEX process will become more alike with each round of selection. If homology is high, alignment of RNA sequences could identify a conserved RNA binding site of the protein. Since the percentage of G/U residues in the RNA sequence is important in the analysis of the cognate RNA for RbpA, the RNA sequences were aligned according to G/U classification (G/U poor, rich and very rich) and round of selection (10 or 14).

5.3.1 G/U poor RNA

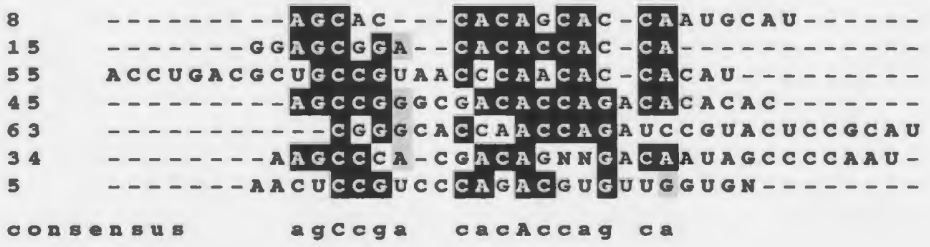
A qualitative comparison of the alignments of G/U poor RNA from round 10 and round 14 (Figure 5.1, A and B, respectively) revealed that round 14 RNA sequences had a higher degree of RNA sequence conservation. This is evident by the presence of fewer conserved bases in the consensus sequence of the round 10 G/U poor RNA than in the round 14 G/U poor RNA. Since round 14 used reaction conditions of higher stringency, a

Figure 5.1 Alignment of G/U poor RNA sequences selected in **(A)** round 10 and **(B)** round 14 of genomic SELEX. **(C)** Comparative alignment of two RNA molecules bound by Ni^{+2}-NTA resin alone (N1 and N9) (Hofmann *et al.*, 1997) with G/U poor RNA molecules selected in rounds 10 and 14 of genomic SELEX. Sequence alignments were performed by use of the CLUSTAL W1.8 (Jeanmougin *et al.*, 1998) multiple alignment program at the Baylor College of Medicine Search Launcher website (http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html). Shading was introduced into the alignment by use of the Boxshade 3.21 program located at the website (http://www.ch.embnet.org/software/BOX_form.html). Conservation of residues that occur in at least 70% of the number of sequences is indicated by reverse type face, conservation of the same class of RNA bases (purines or pyrimidines) are indicated by light shading.

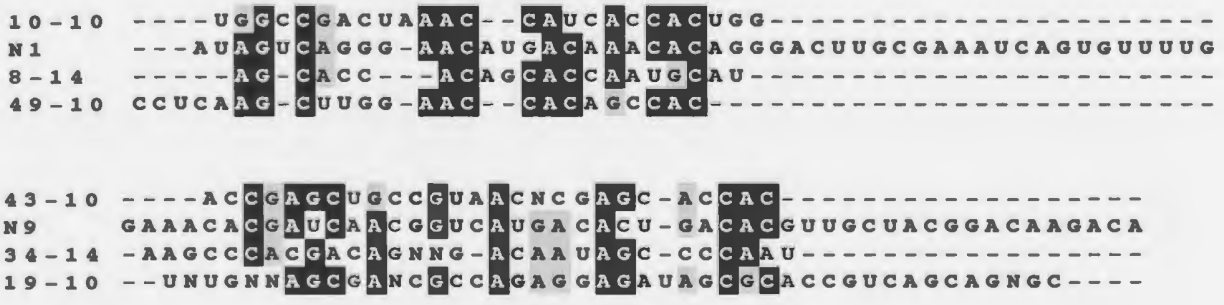
A.


8 ACUGA-GCUGCCGUAAC-CAAA--CAC
23 ACUGA-GCUGCCGUAAC-CAAA--CACCACCACAC
49 CCUCAAGCUU--GGAAC-CACAGCCAC
25 GGAUC-CACCACCAGCC
10 UGGCCGACUAAAC-CAUCACCACUGG
43 ACCGA-GCUGCCGUAAC-NCGAG-CACCAC
11 UAAUGCAAGCCAGCUCCAGCGGGUUCGCA-GCAGUCG-AGC-GCGAGGCAGACC
19 UNUGNNAGCGANCGCCAGAGGAGAUAGCGCACCGUCAGCAGNGC
42 UAGGUCAGCACUAUCACUAGGGAGAGAACACCCANUCCCAGNCA
consensus c a gc g cg aaC ca ag CAc

B.


8 AGCAC---CACAGCAC-CAAUGCAU
15 GGAGCGGA--CACACCAC-CA
55 ACCUGACGCUGCCGUAACCCAACAC-CACAU
45 AGCCGGGCGACACCAGACACACAC
63 CGGGCACCAACCAGAUCCGUACUCCGCAU
34 AAGCCCA-CGACAGNNGACAAUAGCCCCAAU
5 AACUCCGUCCCAGACGUGUUGGUGN
consensus agCcga cacAccag ca

C.


10-10 UGGCCGACUAAAC--CAUCACCACUGG
N1 AUAGUCAGGG-AACAUGACAAACACAGGGACUUGCGAAAUCAGUGUUUUG
8-14 AG-CACC---ACAGCACCAAUGCAU
49-10 CCUCAAG-CUUGG-AAC--CACAGCCAC
43-10 ACCGAGCUGCCGUAACNCGAGC-ACCAC
N9 GAAACACGAUCAACGGUCAUGACACU-GACACGUUGCUACGGACAAGACA
34-14 AAGCCCACGACAGNNG-ACAAUAGC-CCCAAU
19-10 UNUGNNAGCGANCGCCAGAGGAGAUAGCGCACCGUCAGCAGNGC

correlation exists between stringency and sequence homology. This result suggests that the G/U poor sequences are not retained in a non-specific manner but are likely specifically bound by RbpA. If RNA was bound in a completely non-specific manner, an increase in sequence similarity would not have been observed.

The binding preference of RbpA for guanine and uracil nucleotide homopolymers would suggest that these G/U poor sequences were retained in a non-RbpA mediated manner. One possible explanation is that selection of G/U poor RNA occurred due to an interaction with the positively-charged Ni^{+2}-NTA resin used to purify RNA-H_6RbpA complexes in the genomic SELEX procedure. Ni^{+2}-NTA resin has been shown to exhibit RNA binding specificity in a study utilizing the regular SELEX protocol whereby RNA sequences were randomized at 50 positions (Hofmann *et al.*, 1997). The resin exhibited a selective binding for RNA motifs composed of a purine rich internal loop and a mismatch G-A base pair with a dissociation constant as low as 0.8 uM. Comparative sequence alignment analysis of the two most selected RNA sequences (N1 and N9) from the Hofmann study and the G/U poor RNA sequences selected during rounds 10 and 14 identified partial homology (Figure 5.1C). This result suggests that G/U poor RNA molecules were selected by the resin but it still does not rule out the involvement of RbpA. Interestingly, the sequence of many G/U poor RNA molecules are slight variations of the sequence ACACCAC, a sequence complementary to the repeated UGUGGUG sequence found in G/U very rich sequences (refer to section 5.3.3). Therefore, G/U poor RNA could have been retained via an interaction with G/U very rich RNA bound by RbpA.

5.3.2 G/U rich RNA

Comparative alignment of G/U rich sequences from round 10 and round 14 identified greater homology of RNA selected in conditions of higher stringency (Figure

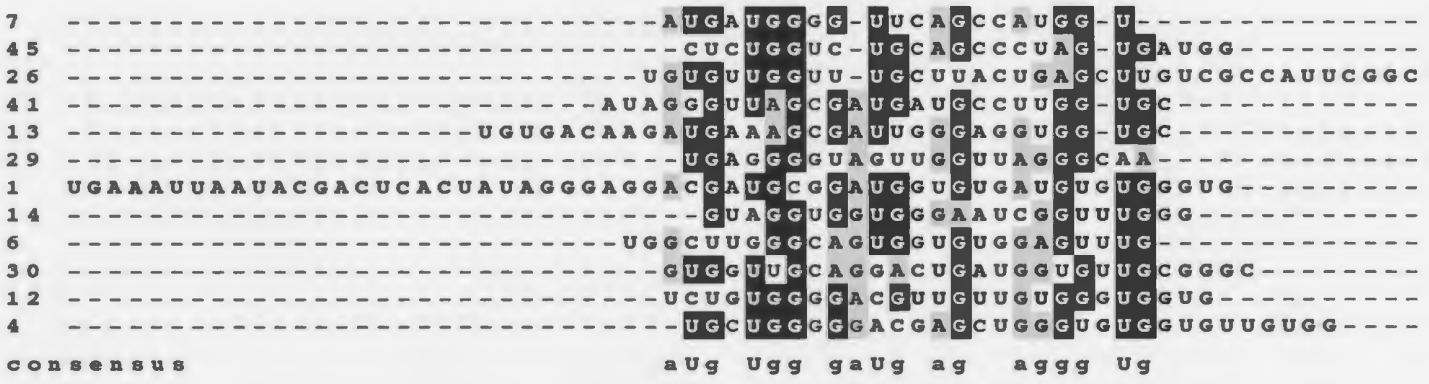
5.2). The lack of sequence conservation evident in the alignment of round 10 RNA molecules contrasts with a strikingly high amount of conservation in the round 14 RNA sequences. The round 10 G/U rich RNA sequences do not contain any sequence that was selected more than once, and there is no position in the alignment that is completely conserved in all 12 sequences (Figure 5.2A). The eighteen round 14 G/U rich RNA were highly homologous due to selection of a nearly identical sequence seven times (clones 1, 18, 37, 47, 52, 54, 59) (Figure 5.2B). In fact, the most selected sequence was identical to the consensus sequence generated from the alignment with the exception of one base difference at position 4 (as indicated) in which the most selected sequence contains a cytosine nucleotide but the other sequences in this group are aligned with a gap, a guanine nucleotide or an undefined residue (N). A high degree of sequence homology and the multiple selection of the same RNA sequence indicate that this group of sequences are likely bound with greater affinity by the protein compared with the G/U very rich and G/U poor classes of RNA and is consistent with previous evidence indicating a preference of RbpA for G/U homopolymers. Therefore, the sequence 5'AGUNGGGGUGAAUGGGAGGUG3' can be postulated as the most likely consensus sequence of RbpA; it is further characterized in section 5.5.

5.3.3 G/U very rich sequences

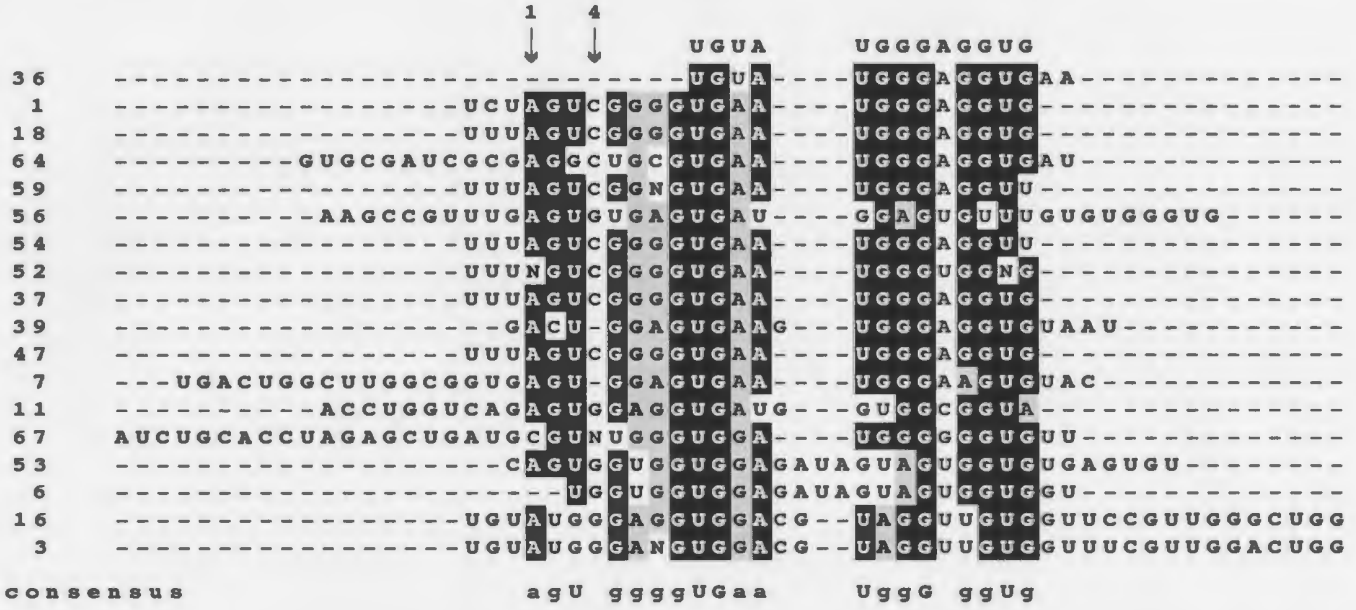
Alignment of the RNA sequences very rich in G/U residues (Figure 5.2C) identified the conservation of a repeated UGUGGUG sequence. In fact, all three G/U very rich RNA sequences are composed almost entirely of this repeat. Some G/U rich and G/U poor RNA sequences identified from both rounds 10 and 14 of SELEX also contain UGUGGUG or a closely related sequence such as GUGGUG or UGUGGGUG (Figure 5.3). Therefore, the consistent degree of conservation of this sequence in all three groups of RNA sequences from both rounds of genomic SELEX indicate that the GUGGUG

Figure 5.2 Alignment of G/U rich RNA sequences selected in **(A)** round 10 and **(B)** round 14 of genomic SELEX. The revised consensus sequence outlined in figure 5.4 is given above the alignment. Designation of residues 1 and 4 are shown. **(C)** Alignment of G/U very rich RNA sequences from round 10 (5-10) and round 14 (35-14, 43-14) of genomic SELEX. Alignments were performed as described in figure 5.1. Consensus sequences generated by the alignment program are given at the bottom of each alignment.

A.


7 AUGAUGGGG-UUCAGCCAUGG-U
45 CUCUGGUC-UGCAGCCCUAG-UGAUGG
26 UGUGUUGGUU-UGCUUACUGAGCUUGUCGCCAUUCGGC
41 AUAGGGUUAGCCGAUGAUGCCUUGG-UGC
13 UGUGACAAGAUGAAAGCCGAUUGGGAGGUGG-UGC
29 UGAGGGGUAGUUGGUUAGGGCAA
1 UGAAAUUAAUACGACUCACUAUAGGGAGGACGAUGCGGAUGGUGUGAUGUGUGGGUG
14 GUAGGUGGUGGGAAUCGGUUUGGG
6 UGGCUUGGGCAGUGGUGUGGAGUUUG
30 GUGGUUGCAGGACUGAUGGUGUUGCGGGC
12 UCUGUGGGGGACGUUGUUGUGGGUGGUG
4 UGCUGGGGGGACGAGCUGGGUGUGGUGUUUGUGG
consensus aUg Ugg gaUg ag agggg Ug

B.


1 4
UGUA UGGGAGGUG
36 UGUA UGGGAGGUGAA
1 UCUAGUCGGGGUGAA UGGGAGGUG
18 UUUAGUCGGGGUGAA UGGGAGGUG
64 GUGCGAUCGCGAGGCUGCGUGAA UGGGAGGUGAU
59 UUUAGUCGGNGUGAA UGGGAGGUU
56 AAGCCGUUUGAGUGUGAGUGAU GGAGUGUUUGUGUGGGUG
54 UUUAGUCGGGGUGAA UGGGAGGUU
52 UUUNGUCGGGGUGAA UGGGUGGNG
37 UUUAGUCGGGGUGAA UGGGAGGUG
39 GACU-GGAGUGAAG UGGGAGGUGUAAU
47 UUUAGUCGGGGUGAA UGGGAGGUG
7 UGACUGGCUUGGCGGUGAGU-GGAGUGAA UGGGAAGUGUAC
11 ACCUGGUCAGAGUGGAGGUGAUG GUGGCGGUA
67 AUCUGCACCUAGAGCUGAUGCGUNUGGGUGGA UGGGGGGUGUU
53 CAGUGGUGGUGGAGAUAGUAGUGGUGUGAGUGU
6 UGGUGGUGGAGAUAGUAGUGGUGGU
16 UGUAUGGGAGGUGGACG--UAGGUUGUGGUUCCGUUGGGCUGG
3 UGUAUGGGANGUGGACG--UAGGUUGUGGUUUCGUUGGACUGG
consensus agU ggggUGaa UggG ggUg

C.


35-14 GACUGCGGUGUGUGGUGUGUGGUGUGUGGUGUGUGG
43-14 NAGUGUGGUGUGUGGUGUGUGGUG
5-10 UGUGUGGUGUGUGGUGUGUGGUGUGUGGUGUGGGUGUA
consensus GUGUGGUGUGUGGUGUGUGGUG

Figure 5.3 RNA sequences identified in rounds 10 and 14 of genomic SELEX that contain a UGUGGUG or closely related sequence (boxed). The round of selection in which the RNA sequence was identified, the clone number and the percentage of G/U residues in the RNA sequence is given.

UGUGGUG repeat

Round	Clone	%G/U	Sequence (5'-3')
10	1	61	UGAAAUUAAUACGACUCACUAUAGGGAGGACGAUGCGGAUGGUGUGAUGUGUGGGUG
10	4	84	UGCUGGGGGACGAGCUGGGUGUGGUGUUGUGG
10	5	98	UGUGUGGUGUGUGGUGUGUGGUGUGUGGUGUGUGGUGUGGGUGUA
10	6	85	UGGCUUGGGCAGUGGUGUGGAGUUUG
10	12	88	UCUGUGGGGACGUUGUUGUGGGUGGUG
10	13	64	UGUGACAAGAUGAAAGCGAUUGGGAGGUGGUGC
10	14	83	GUAGGUGGUGGGAAUCGGUUUGGG
14	6	84	UGGUGGUGAUGAUAGAGGUGGUGGU
14	35	92	GACUGCGGUGUGUGGUGUGUGGUGUGUGGUGUGG
14	56	78	AAGCCGUUUGAGUGUGAGUGAUGGAGUGUUUGUGUGGGUG
14	53	79	CAGUGGUGGUGGAGAUAGUAGUGGUGUGAGUGU
14	43	92	NAGUGUGGUGUGUGGUGUGUGGUG

motif could be a binding site for RbpA.

5.4 Characterization of an RbpA consensus binding sequence

Identification of a putative RbpA consensus binding site was difficult due to retention of three apparently different groups of RNA sequences (G/U poor, G/U rich and G/U very rich). If there is one, the G/U rich group of RNA sequences from round 14 most likely contain the RbpA consensus sequence for three reasons: these sequences were selected in conditions of higher stringency; they represent the majority of sequences identified in round 14 (18 of 27, Table 4.3); and, because alignment identifies this group as containing the greatest degree of sequence homology.

Careful examination of the round 14 G/U rich RNA sequence alignments allows the putative consensus sequence identified by the alignment to be further defined. Clone 36 was the shortest sequence (15 nucleotides) recovered during round 14 of selection. Nearly the entire sequence of clone 36 is highly conserved in all of the other G/U rich group of sequences selected in this round as shown in figure 5.4. However clone 36 does not contain the initial UUUAGUNGGGG portion (where N is GTP, CTP or a gap) of the conserved sequence shown in figure 5.2B, therefore this initial region may not be important in RbpA binding or recognition. As a result, a revised consensus sequence UGAAUGGGAGGUG (Figure 5.4A) of 13 nucleotides in length is considered to be the best candidate for an RbpA consensus binding sequence. The first 2 nucleotide positions are completely conserved in all 18 sequences with bases U and G respectively; positions 8 and 12 are also completely conserved with G and U residues respectively. The remaining nine positions are degenerate (Figure 5.4B). However, with the exception of position 9, whose most prominent residue is an A present in only 53% of sequences, the remaining positions in this consensus sequence are dominated by one particular base present in 68% to 95% of all sequences (Figure 5.4C). A graphic representation of the relative abundance

Figure 5.4 Characterization of the RbpA consensus binding sequence. **(A)** Comparisons used to generate a revised RbpA consensus binding sequence. Note that the G/U rich consensus sequence is largely a tandem repeat of 10 (I) and 11 (II) nucleotides as shown. **(B)** The degeneracy of conservation (of RNA bases) at each position in the 13 base consensus sequence. **(C)** Degeneracy of nucleic acid conservation is given as the percentage in which each base occurs in each position of the 13 base revised consensus sequence. **(D)** Graphic representation of ribonucleic acid conservation in the RbpA consensus binding sequence created by use of the Weblogo version 2.6 program (Schneider and Stephens, 1990) accessed at the University of California at Berkeley website (http://weblogo.berkeley.edu/).

A.

Round 14 G/U rich consensus sequence	I II UUUAGUNGGGGUGAAUGGGAGGUG
Clone 36	UGAAUGGGAGGUGAA
Revised consensus sequence	UGAAUGGGAGGUG

B.

Degeneracy of revised consensus sequence

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
	U	G	A	A	U	G	G	G	A	G	G	U	G
			G	U	G	A	A		U	U	U		U
			U	G		U	U		C	A			A
									G				

C.

Position

1. U 100%
2. G 100%
3. A 68%, G 21%, U 11%
4. A 79%, U 16%, G 5%
5. U 79%, G 21%
6. G 74%, A 16%, U 10%
7. G 84%, A 10%, U 5%
8. G 100%
9. A 53%, U 37%, C 5%, G 5%
10. G 84%, U 10%, A 5%
11. G 95%, U 5%
12. U 100%
13. G 79%, U 16%, A 5%

D.

of a particular base within each position of the revised consensus sequence is given in figure 5.4D.

Interestingly, the six bases, GAGGUG, on the 3' end of the putative RbpA binding site are nearly identical to the GUGGUG repeated motif shown in figure 5.3. Five of the six residues of the GAGGUG portion of the RbpA binding site (positions 8 to 13) are highly conserved. The exception is position 9 which is the most degenerate of all positions in the 13 base consensus sequence (Figure 5.4D) and corresponds to the only difference between GAGGUG and GUGGUG. The most prominent residue at position 9 is an A, present in only 53% of sequences. In addition, the second most selected residue at position 9 is a U residue, present in 37% of round 14 G/U rich SELEX sequences. In this way, the round 14 G/U rich sequences and the group of sequences from rounds 10 and and 14 that contain the GUGGUG motif are very similar, thereby defining the six and 14 that contain the GUGGUG motif are very similar, thereby defining the six nucleotides on the 3' end of the revised consensus sequence as potentially being very important in sequence recognition by RbpA. A six base RNA recognition sequence for RbpA is consistent with the length of another RRM-type RNA binding protein U1A, which recognizes a seven base sequence: AUUGCAC (Oubridge *et al.*, 1994).

The mechanism in which H_6RbpA binds to the putative consensus sequence is not known. Potentially H_6RbpA could recognize a conserved binding sequence, a conserved structure, or a sequence-independed set of nucleotides. The alignment of round 14 G/U rich RNA suggests that RbpA recognizes a consensus sequence because of the sequence conservation in this group of RNA molecules. If it was a structure that was conserved, a greater degree of sequence variability in this group of molecules would likely have been the case. In addition, an attempt to identify an RNA consensus structure of the RNA molecules selected in both rounds 10 and 14 by use of an RNAfold program was

unsuccessful (refer to section 2.20) These results suggest that a conserved sequence, not a structure, is recognized by H_6RbpA.

5.5 Comparison of the putative RbpA consensus sequence with the consensus binding sequence of three Csps

RbpA may perform a function in cyanobacteria analogous to that of the Csp proteins in *Escherichia coli* described in section 1.5. Therefore, once the putative RbpA consensus sequence of RbpA was identified, I wanted to compare the binding specificity of RbpA with that of the Csps. The RNA binding specificity of CspB, CspC and CspE was identified by SELEX as UUUUU, AGGGAGGGA and AAAUUU residues respectively (Phadtare and Inouye, 1999). None of the 49 RbpA binding sequences identified by genomic SELEX here contain an entire Csp consensus sequence, however a portion of the putative RbpA consensus sequence UGAAUGGGAGGUG is homologous to a portion of the CspB consensus sequence. Both the RbpA and CspC consensus sequences contain the six nucleotide sequence GGGAGG. This result identifies a similarity in the binding specificity of RbpA and CspB and thereby provides further evidence that these proteins could perform the same function in *Synechococcus* 7942 and *Escherichia coli* respectively. RbpA is a cold-shock protein and could perform the same function as CspB of the CspA-family of cold-shock proteins in *E. coli* by acting as an RNA chaperone or a transcription antiterminator.

5.6 Genomic Analysis of SELEX isolated sequences

Since the RNA sequences identified in the genomic SELEX protocol are derived from genomic sequences, one can identify the location in the genome in which the RNA sequence is encoded and determine which genes are potentially regulated by RbpA. Therefore, identification of the genes that contain an RbpA target sequence may provide information regarding a putative function of RbpA.

Genomic similarity analysis detected a high degree of conservation between most SELEX sequences isolated from rounds 10 and 14 of SELEX and the *Synechococcus* 7942 draft genome sequence (Tables 5.1 to 5.6). All alignments with the *Synechococcus* 7942 genome sequence with the exception of round 14 clones 7 and 39 and the G/U very rich sequences are 75% or greater. Theoretically, the degree of homology between the genomic SELEX RNA and the *Synechococcus* 7942 genome sequences should be 100%, however the fact that there is a lack of complete homology is not surprising for a number of reasons. Firstly, amplification of dsDNA generated in each round of genomic SELEX was performed using Taq polymerase, a polymerase that compared with other thermostable DNA polymerases, has low fidelity. However, although the likelihood of a mistake incorporated into the SELEX sequences is low, given that the error rate of Taq polymerase is 8.0×10^{-6} errors/bp/duplication (Cline *et al.*, 1996) which translates into one mistake in 4166 bp in 30 cycles of amplification, mistakes could make an RNA molecule better able to be recognized by RbpA, thereby greatly increasing the chance that the molecule is retained during the many rounds of selection. Secondly, the use of primers completely randomized at 9 nucleotides at the 3' end of the primer for first and second strand extension reactions can generate mispairing between primer and template. The lack of complete complementarity can therefore lead to mistakes in the genomic portion of library sequences. In a paper by Singer *et al.*, 1997, the introduction of mistakes in genomic SELEX library sequences is well documented, with the greatest likelihood of a mistake generated by the random nucleotide adjacent to the first fixed nucleotide of the primer. Indeed, of the 49 sequences identified by genomic SELEX, 23 sequences did not align with the *Synechococcus* 7942 genome at the first position. Lastly, molecules in a PCR can undergo recombination events, generating new sequences in the process (Jensen and Straus, 1993).

Table 5.1 Comparative sequence analysis of round 10 G/U very rich RNA and the *Synechococcus* 7942 genome

Clone Number(s)	Alignment [a]	Homology (%)	Location in Synechococcus [b] 7942 genome	Gene Name/ Function	Location in Gene
5	UG-UGUGGUGUGUGGUGUGUGGUGUGUGGUG--UGUGGUGUGGGUGUA TGATTTGGGGTGTGGTGTGTGGTGCG-GCTGCTTCTGGCGGCTATGAA	75.6	contig 135, gene 1922	Chlorophyll synthase 33kDa subunit	In ORF, coding strand

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

Table 5.2 Comparative sequence analysis of round 10 G/U rich RNA and the *Synechococcus* 7942 genome

Clone Number(s)	Alignment[a]	Homology (%)	Location in *Synechococcus*[b] 7942 genome	Gene Name/ Function	Location in Gene
1	UGAAAUUAAUACGACUCACUAUAGGGAGGACGAUGCGGAUGGUGUGAUGUGUGGGUG GCAGATGCGATCGCCTGTCTCCAGCAAACGGGGTGCGGATGGTGTGATGTGTGGGCG	63.2	contig 136, gene 2374	probable tRNA dihydrouridine synthetase (Dus)	In ORF, coding strand
4	UGCUGGGGGACGAGCUGGGUGUGGUGUUGUGG CGCTGGAGGATGAGCTGGGTGTGGTGTTGCTA	81.2	contig 125 gene 169	lrrA, transcriptional regulator	In ORF, coding strand
6	UGGCUUGGGCAGUGGUGUGGAGUUUG TGG-TCGGGCAGTGGTGTGGAGGAAA	76.9	contig 128, gene 467	GDP-D mannose dehydratase	In ORF, coding strand
7	AUGAUGGGGUUCAGCCAUGGU GTGAGGGGGTTCAGCCATGGT	90.4	contig 127 gene 343	Leu/Phe tRNA protein transferase, Aat	overlaps 3' end of ORF , non-coding strand
12	UCUGUGGGGACGUUGUUGUGGGUGGUG ACAGTAGGGACGTTGTTGTGGGTGGTG	88.8	contig 136, gene 2533	phosphoglycolate phosphatase	starts at stop codon, coding strand
13	UGUGACAAGAUGAAAGCGAUUGGGAGGUGGUGC CGAGACAAGATGAAAGCGATTGGGAGGTGGTGC	93.9	contig 122, gene 13	uncharacterized conserved bacterial protein	In ORF, coding strand
14	GUAGGUGGUGGGAAUCGGUUUGGG GTAGGTGGTGGGAATCGGTTGGGG	95.8	contig 133 gene 1155	unknown protein	In ORF, non-coding strand
26	UGUGUUGGUUUGCUUACUGAGCUUGUCGCCAUUCGGC CGCCTTGGTTTGCTTACTGAGCTTGTCGCCATTCGGC	91.9	contig 126, gene 284	class I aminoacyl-tRNA synthetase (E and Q amino acids)	In ORF, non-coding strand
29	UGAGGGGUAGUUGGUUAGGGCAA CCTGGGGTAGTTGGTTAGGGCAA	-	contig 136, gene 2748		between ORFs
30	GUGGUUGCAGGACUGAUGGUGUUGCGGGC GTGGCTGCAGGACTGATGGTGCTGCGGGC	93.1	contig 135, gene 1766	uncharacterized flavoprotein, FpaA	In ORF, coding strand
41	AUAGGGUUAGCGAUGAUGCCUUGGUGC TTAGGGTTAGCGATGATGCCTTGGTGC	96.3	contig 135, gene 2061	Aspartate-semialdehyde dehydrogenase, Asd	In ORF, coding strand
45	CUCUGGUCUGCAGCCCUAGUGAUGG GGTTTGGTTGCAGCCCTAGTGATGG	82.6	contig 136, gene 2061	conserved hypothetical protein	In ORF, coding strand

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

Table 5.3 Comparative sequence analysis of round 10 G/U poor RNA and the *Synechococcus* 7942 genome

<table>
<tr><th>Clone Number(s)</th><th>Alignment[a]</th><th>Homology (%)</th><th>Location in Synechococcus[b] 7942 genome</th><th>Gene Name/ Function</th><th>Location in Gene</th></tr>
<tr><td>8, 23, 43</td><td>ACUGAGCUGCCGUAACCAAACACCACCACAC
||||||||||||||||||||||||||||
ACTGAGCTGCCGTAACCAAACACCACCATTG</td><td>90.3</td><td>contig 135,
gene 2106</td><td>putP, sodium/proline symporter</td><td>In ORF,
non-coding strand</td></tr>
<tr><td>10</td><td>UGGCCGACUAAACCAUCACCACUGG
||||||||||||||||||||||||
UGGCCGACUAAACCAUCACCACUGG</td><td>100</td><td>contig 136,
gene 2771</td><td>uncharacterized conserved protein</td><td>In ORF,
non-coding strand</td></tr>
<tr><td>11</td><td>UAAUGCAAGCCAGCUCCAGCGGGUUCGCAGCAGUCGAGCGCGAGGCAGACC
|||| |||||||||||||||||||||||||||||||||||||||||||||
UAAUCCAAGCCAGCUCCAGCGGGUUCGCAGCAGUCGAGCGCGAGGCAGACC</td><td>98</td><td>contig 136,
gene 2816</td><td>two-component histidine kinase</td><td>In ORF,
non-coding strand</td></tr>
<tr><td>19</td><td>UNUGNNAGCGANCGCCAGAGGAGAUAGCGCACCGUCAGCAGNGC
| | ||||| ||||||||||||||||||||||||||||| ||
TGAGAGAGCGAACGCCAGAGGAGATAGCGCACCGTCAGCAGAGC</td><td>86.4</td><td>contig 136,
gene 2809</td><td>UDP-glucose-beta-D-glucan-glucosyltransferase</td><td>In ORF,
non-coding strand</td></tr>
<tr><td>25</td><td>GGAUCCACCACCAGCC
||||||||||||||
CTATCCACCACCAGCC</td><td>87.5</td><td>contig 133,
gene 1184</td><td>conserved hypothetical protein</td><td>upstream of start codon,
non-coding strand</td></tr>
<tr><td>42</td><td>UAGGUCAGCACUAUCACUAGGGAGAGAACACCCANUCCCAGNCA
||||||||||||||||||||||||||||||||| ||||| ||
AAGGTCAGCACTATCACTAGGGAGAGAACACCCAGCCCCAGCCA</td><td>90.7</td><td>contig 133,
gene 1258</td><td>signal transduction protein, contains PAS, PAC, GGDEF and EAL domains</td><td>In ORF,
non-coding strand</td></tr>
<tr><td>49</td><td>CCUCAAGCUUGGAACCACAGCCAC
||||||||||||||||||||
GCTCAAGCTTGGAACCACAGCTTT</td><td>83.3</td><td>contig 132,
gene 996</td><td>sodB, Iron superoxide dismutase</td><td>In ORF,
coding strand</td></tr>
</table>

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

Table 5.4 Comparative sequence analysis of round 14 G/U very rich RNA with the *Synechococcus* 7942 genome

Clone Number(s)	Alignment [a]	Homology (%)	Location in *Synechococcus* [b] 7942 genome	Gene Name/ Function	Location in Gene
35, 43	GACUGCGGUGUGUGGUGUGUGGUGUGUGGUGUGUGG TGATTTGGGGTGTGGTGTGTGGTGCG-GCTGCTTCT	65.7	contig 135, gene 1922	Chlorophyll synthase 33kDa subunit	In ORF, coding strand

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

Table 5.5 Comparative sequence analysis of round 14 G/U rich RNA with the *Synechococcus* 7942 genome

Clone Number(s)	Alignment[a]	Homology (%)	Location in *Synechococcus*[b] 7942 genome	Gene Name/ Function	Location in Gene
1, 18, 36, 37, 47, 52, 54, 59	UUUAGUCGGGGUGAAUGGGAGGUG ||||||||||||||||||| | TTTAGTCGGGGTGAATGGGGCTGG	83.3	contig 126 gene 274	Homology to ABC transporter, Uup protein of E.coli	In ORF, coding strand
1, 18, 36, 37, [c] 47, 52, 54, 59	UUUAGUCGGGGUGAA**UGGGAGGUG** |||| || | ||| ||||||||| TTTA-TCCGTGTGTTTGGGAGGTG	83.3	contig 135, gene 2154	phosphorylase kinase	In ORF, coding strand
3, 16	UGU-AUGGGAGGUGGACGUAGGUUGUGGUUCCGUUGGGCUGG || | |||||||| |||||||||||||| || |||| ||| GGTAAAAGGAGGTGGGCGTAGGTTGTGGTTTCGGTGGGATGG	85.4	contig 124 gene 79	ntrB, nitrate reductase	In ORF, non-coding strand
5	UUGAGGCAGGGUCUGCUGUUGGUGN ||||| ||||||||||||||| TTGAG-CAGGGTCTGCTGTTGGTG	100	contig 126 gene 240	ArgD, ornithine/acetylornithine aminotransferase	In ORF , coding strand
6, 53	CAGUGGUGGUGGAGAUAGUAGUGGUGUGAGUGU |||||||||||||||||||||||| ||||| GCGTGGTGGTGGAGATAGTAGTGGTGGGAGTGG	87.9	contig 135, gene 2163	Protein with Winged-helix DNA-binding domain and Glycine-rich domain	In ORF, coding strand
7	UGACUGGCUUGGCGGUGAGUGGAGUGAAUGGGAAGUGUAC | ||| ||||||||||||||| GATCCCAACCAAGCCTGACCGGAGTGAATGGGAAGTGGCA	52.5	contig 126, gene 269	ABC transporter	In ORF, coding strand
11	ACCUGGUCAGAGUGGAGGUGAUGGUGGCGGUA | ||||||||||||||||||||||||||| TCTTGGTCAGAGTGGAGGTGATGGTGGCGGTG	90.6	contig 129 gene 578	dsg gene, fatty acid desaturase	In ORF, non-coding strand
39	GACUGGAGUGAAGUGGGAGGUGUAAU | | ||||||||||||||| ATGGGAAATGAAGTGGGAGGTGTTGC	65.3	contig 133 gene 1215	Beta 1,4 xylanase	In ORF, coding strand
56	AAGCCGUUUGAGUGUGAGUGAUGGAGUGUUUGUGUGGGUG |||||||||||||||||||||||||||||||| |||| || TAGCCGTTTGAGTGTGAGTGATGGAGTGTTTTTGTGATTG	90	contig 122, gene 18	unknown function	In ORF, non-coding strand
64	GUGCGAUCGCGAGGCUGCGUGAA-UGGGAGGUGAU |||||||||||||||| ||||| |||||| || GTGCGATCGCGAGGCTGAGTGAAATGGGAG-TGGA	88.2	contig 135, gene 1928	Type II secretory pathway ATPase PulE/Tfp pilus assembly pathway	In ORF, coding strand
67	AUCUGCACCUAGAGCUGAUGCGUNUGGGUGGAUGGGGGGUGUU | |||||| ||||||||| || |||||| |||| ||||| AATTGCACCAAGAGCTGATGGGTCTGGGTGAATGGTGGGTGGG	79.1	contig 124, gene 141	Hypothetical protein, unknown function	In ORF, coding strand

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

[c] Second alignment of this group of RNA sequences, generated by using the C-terminal nine base sequence (in bold) as the query sequence in BLAST analysis of the *Synechococcus* 7942 genome.

Table 5.6 Comparative sequence analysis of round 14 G/U poor RNA with the *Synechococcus* 7942 genome

Clone Number(s)	Alignment[a]	Homology (%)	Location in *Synechococcus*[b] 7942 genome	Gene Name/ Function	Location in Gene
8	AGCACCACAGCACCAAUGCAU ||||||||||||||||| | AGCACCACAGCACCAATCCTC	81	contig 136, gene 2663	gamma-glutamyltransferase	downstream of stop codon, coding strand
15	GGAGCGGACACACCACCA | ||||||||||| || GATGCGGACACACCCCCT	77.8	contig 123, gene 56	Molybdopterin-guanine dinucleotide biosynthesis protein	In ORF, coding strand
34	'AAGCCCACGACAGNNGACAAUAGCCCCAAU ||||||||||||| |||||||||||| AAGCCCACGACAGCCAGCAATAGCCCCAAA	83.3	contig 136, gene 2520	uncharacterized, low complexity protein	In ORF , non-coding strand
45	AGCCGGGCGACACCAGACACACAC |||||||||||||||||||| | AGCCGGGCGACACCAGACACCGCC	87.5	contig 135, between genes 1700 and 1701		
55	ACCUGACGCUGCCGUAACCCAACACCACAU |||| |||||||||||| |||||||| GACTGA-GCTGCCGTAACCAAACACCACCA	82.7	contig 135, gene 2106	putP, sodium/proline symporter	In ORF, non-coding strand
63	CGGGCACCAACCAGAUCCGUACUCCGCAU |||||||||||||||||||||||||||| CGGGCACCAACCAGATCCGTACTCCGCAC	96.7	contig 129, between genes 615 and 616		

[a] Sequences identified by genomic SELEX are given as RNA and are shown in the top line of each alignment, sequences from the *Synechococcus* 7942 genome are given as DNA are shown in the bottom line.

[b] Designations according to the *Synechococcus* 7942 genomic sequence given by the Joint Genome Institute, consult table A1 for permanent gene designations.

The location in the *Synechococcus* 7942 genome of each RNA sequence is given in tables 5.1 to 5.6. Of the 49 sequences obtained, 44 were identified as being located within the open reading frame (ORF) of a gene, a result that was expected given that 89.5% of the Synechococcus 7942 genome encodes annotated genes. Genomic analysis of some sequences identified homology with a gene whose protein product is hypothetical, unknown or isn't well characterized in the literature; these results are not discussed in further detail. Also, some RNA sequences had poor homology with the genomic sequence or were not located within an open reading frame (ORF); these are also excluded from further discussion. The RNA sequences discussed in further detail (section 5.6.1) are those that were selected multiple times such as the group of seven G/U rich round 14 sequences as well as those that have nearly complete homology to a gene that is well characterized.

The majority of RNA sequences identified by genomic SELEX (44 of 49) are within the open reading frame of a gene. Of these 44 sequences, 28 are encoded on the coding strand and 16 are encoded on the non-coding strand. Therefore, in the majority of cases (28 of 44), RbpA could function by binding directly to the mRNA transcript of a gene within the open-reading frame. By doing so, RbpA would not likely function in altering translation initiation, as that would require the RbpA binding site to be located near the ribosome binding site. Likewise, RbpA would likely not function as a transcription terminator or antiterminator since sequences that regulate expression of the transcript in this way are generally located within the 5' untranslated regions of the transcript, (*e.g.* include the *bgl* operon of *E. coli*, and the *trp* and *pyr* operons of *Bacillus subtilis* (reviewed by Yanofsky, 2000)). Therefore the most obvious function of RbpA when bound to the mRNA transcript within an ORF would be to destabilize RNA secondary structures generated in conditions of cold-shock or to regulate the stability

(half-life) of the mRNA.

The remaining sequences (16 of 44) were located in the non-coding strand of an ORF. In these cases, if RbpA is involved in regulation of gene expression via the mRNA transcript, it could involve a mechanism mediated by a *cis*-encoded antisense RNA. Regulatory antisense RNAs are encoded either in *trans* or in *cis*. Those that are encoded in *trans* are located in a region of genome that does not encode the gene being regulated. Those that are encoded in *cis* are located within the regulated gene on the non-coding strand. Antisense RNA regulation in prokaryotes was first characterized by Tomizawa and Itoh (1981) but the amount of knowledge concerning chromosomally encoded antisense RNA regulation in prokaryotes is very limited, with only nine examples characterized thus far (reviewed by Wagner *et al.*, 2002). Interestingly, a role for two antisense RNAs has been characterized in *E. coli* stress responses. These antisense RNAs, MicF and DrsA, regulate the expression of genes by acting as translational inhibitors or activators (Lease *et al.*, 1998) and at the level of mRNA stability (Lease and Belfort, 2000). MicF regulates the expression of outer membrane porins such as OmpF in response to changes in temperature and osmolarity (Andersen *et al.*, 1989). DrsA acts as both a translational activator of the *rpoS* gene (encodes stationary phase sigma factor S) in response to cold temperature (Sledjeski, *et al.*, 1996) and as a repressor of the *hns* gene (encodes a histone-like repressor protein) (Lease *et al.*, 1998; Majadalani *et al.*, 1998). Therefore, the RNA sequences retained by RbpA in the genomic SELEX that do not correspond to known transcribed RNA but are located on the non-coding strand within a transcript, could be part of a *cis*-encoded antisense RNA mechanism of gene regulation in cyanobacteria.

Interestingly, a correlation exists between the percentage of G/U residues in the RbpA recognition sequence and the location of the sequence in terms of being encoded in the coding or non-coding strand of a gene. Of the 28 RNA sequences located in the

coding strand of a gene, the vast majority of these sequences (26) are G/U very rich or G/U rich sequences (93%). This result suggests that an RbpA-mediated mechanism of post-transcriptional regulation that involves binding of RbpA to the coding strand within the ORF of an mRNA would involve an RbpA recognition sequence predominantly composed of G and U residues. In addition, of the 16 RNA sequences encoded in the non-coding strand of a gene, the majority of these sequences (9 or 56%) are G/U poor sequences. While not as dramatic as that for the coding strand outlined above, it identifies the tendency of G/U poor sequences bound by RbpA to be located on the non-coding strand of an ORF. This result suggests that an RbpA-mediated *cis*-encoded antisense RNA mechanism of regulation would involve binding of RbpA to a G/U poor sequence.

5.6.1 Comparative analysis of round 10 sequences and the *Synechococcus* 7942 genome

With exception of the clone 49 sequence, which is located within the *sodB* gene (Table 5.3) encoding a superoxide dismutase protein that could be involved in maintaining the photosynthetic apparatus under conditions of cold-shock (*c. f.* section 5.6.1.6), the genes identified by homology to SELEX round 10 sequences have no obvious connection to the cold-shock response or with a possible function of RbpA. Therefore, I selected genes for further analysis and discussion based on (a) good homology between the gene and the corresponding round 10 RNA sequence (greater than 80% homology) and (b) if the function for the encoded protein is related to nucleic acid function or gene regulation.

5.6.1.1 Clone 4

The clone 4 sequence has an 81.2% homology with a sequence within the ORF of the coding strand of gene 169 which encodes the protein LrrA (Table 5.2). The function of this protein is not known however it is likely a transcriptional regulator. Interpro analysis of the 294 amino acid protein revealed that it contains an N-terminal helix-turn-helix domain (HTH) of the LysR family (Anandan *et al.*, 1996), and a LysR substrate binding domain at the C-terminal end. The LysR family of proteins are involved in the activation of a diverse set of operons and genes (Schell, 1993) including *lysR* (autoregulation) and *lysA* involved in lysine biosynthesis (Stragier, 1983). In *E. coli*, 27 putative LysR-type proteins have been detected in the genome (Karp *et al.*, 2004). FASTA analysis of the *E. coli* proteome with LrrA detected greatest homology to the uncharacterized LysR-type transcriptional regulator Yfer. Interestingly, efforts to create an *lrrA* inactivated mutant strain of *Synechococcus* 7942 were unsuccessful, suggesting that the protein could be involved in an integral cellular function (Anandan *et al.*, 1996).

5.6.1.2 Clone 7

Clone 7 has greatest homology (90.4%) to a region in the non-coding strand of gene 343 of the *Synechococcus* 7942 genome (Table 5.2). This gene encodes the Leu/Phe tRNA protein transferase protein Aat, a protein that transfers a leucine or phenylalanine from their respective aminoacyl-tRNA to the N-terminus of an acceptor protein (Kaji *et al.*, 1965). Acceptor proteins are those destined for the N-end rule pathway of protein degradation characterized in *E. coli* (Varshavsky, 1992) and are recognized by Aat via an N-terminal Arg or Lys residue. The existence of an N-end pathway of protein degradation in cyanobacteria has not been characterized, however evidence for the pathway is

supported by complementation of an N-end defect in *E. coli* by the *Synechocystis* 6803 Aat protein (Ichetovokin *et al.*, 1997).

5.6.1.3 Clone 11

Clone 11 has 98% homology to the non-coding strand within the ORF of gene 2816 (Table 5.3). The JGI classification of the protein is a two-component histidine kinase that by COGS comparative analysis, is most closely related to the BaeS family of histidine kinases. However, when I analyzed protein 2816 by Interpro it revealed that the protein contains ATPase and histidine kinase domains like BaeS, but that it also contains a GAF domain. The GAF domain is present in phytochrome proteins as well as cGMP-specific phosphodiesterases and NifA, a transcriptional regulator of *nif* operons involved in nitrogen fixation. Regulation of protein 2816 expression by RbpA would have to be mediated by an antisense RNA mechanism, since RbpA recognizes a sequence on the non-coding strand of the gene.

5.6.1.4 Clone 26

Clone 26 has 91.9% homology to the non-coding strand of gene 284 of the *Synechococcus* 7942 genome (Table 5.2). Gene 284 encodes a class I glutamyl-tRNA synthetase. This protein catalyzes the attachment of the amino acids glutamic acid and glutamine to their respective tRNA molecule by means of the Rossman fold tertiary structure, a conserved structure characteristic of all type I aminoacyl-tRNA synthetases (Steitz *et al.*, 1993).

5.6.1.5 Clone 42

Clone 42 has 90.7% homology to the non-coding strand within the ORF of gene 1258 (Table 5.3) which encodes a protein that contains the following four domains

normally associated with signal transduction proteins: PAS, PAC, GGDEF and EAL. The only published data regarding a protein with all four domains relates to the FimX protein of *Pseudomonas aeruginosa.* Inactivation of *fimX* in *Pseudomonas aeruginosa* identified a potential involvement in regulation of twitching motility involving type IV pili (Huang *et al.*, 2003). The involvement of type IV pili in movement of cyanobacterial species has been characterized in *Synechocystis* 6803, whereby the pili are involved in gliding motility of the cyanobacterium (reviewed by McBride, 2001).

5.6.1.6 Clone 49

The sequence identified (83.3% homology) by clone 49 is found in the coding strand of gene 996 (Table 5.3). BLAST analysis of the protein identified it as SodB, an iron superoxide dismutase (SOD) protein. *Synechococcus* 7942 contains two SODs, an iron SOD located in the cytoplasm and a thylakoid-associated manganese SOD (Laudenbach *et al.*, 1989). Together the SOD proteins function as antioxidants by catalysing the conversion of harmful superoxide radicals (O_2^-) into hydrogen peroxide (H_2O_2) and molecular oxygen (O_2). In conditions of light irradiation and cold shock, oxidative stress in photosynthetic organisms including cyanobacteria and plants occurs. This effect is created because cold temperature decreases enzymatic activity of all proteins including those of the Calvin cycle, whereas the ability to harvest light energy is not compromised leading to the formation of reactive O_2 molecules. Oxidative stress in turn damages the cell photosynthetic apparatus such as photosystem I (Thomas *et al.*, 1998). Antioxidants such as SodB are involved in maintaining cellular viability in conditions of cold shock and oxidative stress as indicated by the reduced viability of

Synechococcus 7942 in moderate cold shock conditions (17°C-10°C) in the presence of light (Thomas *et al.*, 1999).

Recognition of a sequence within the coding strand of the *sodB* gene ORF suggests that RbpA could be involved in regulating expression of SodB in cold-shock conditions. The sequence recognized by RbpA is located near the middle of the 690 base ORF, comprising nucleotides 303 to 327. Therefore, RbpA could potentially be involved in increasing the stability of the *sodB* mRNA as a means of increasing the amount of SodB in the cell. The reason for an RbpA-mediated increase in the concentration of SodB in conditions of cold-shock is due to the involvement of SodB in the oxidative stress response and because induction of Fe-SOD has been observed in the chloroplast of tobacco plants in conditions of cold-shock under normal light (Tsang *et al.*, 1991).

5.6.2 Comparative analysis of round 14 sequences and the *Synechococcus* 7942 genome

To further characterize the genes potentially regulated by RbpA in the cyanobacterial cell, round 14 SELEX selected RNA sequences that exhibited nearly complete homology (greater than 80%) with a portion of a gene in the *Synechococcus* 7942 genome that encodes a protein that is well characterized in the literature were further investigated. In addition, genes associated with the cold-shock response that were identified as containing a SELEX RNA sequence were also analyzed due to the involvement of RbpA in the cold-shock response.

5.6.2.1 Clones 1, 18, 36, 37, 47, 52, 54, 59

The sequence of clones 1, 18, 36, 37, 47, 52, 54 and 59 are almost entirely identical and therefore are considered together. Since this group contains the entire

revised consensus sequence and comprised the greatest proportion of the final pool of nucleic acid molecules sequenced from round 14 of genomic SELEX, determination of the location of this sequence group in the *Synechococcus* 7942 genome was important.

Comparative analysis of the *Synechococcus* 7942 genome with all clones revealed an identical result, that these sequences have 83.3% homology with two regions of the genome (Table 5.5). The first region is located at nucleotides 45304 to 45327 of contig 126 within the coding region of gene 274. The problem with this alignment is the absence of homology within one of the most conserved regions of the sequences, namely AGGU near the 3' end (*c. f.* Figure 5.2B). Since the 3' end of this group of sequences is highly conserved among the round 14 G/U rich RNA selected sequences, and is therefore an important sequence, I tried to detect a region of the *Synechococcus* 7942 genome that is homologous to the 3' end of this group of sequences by using the sequence UGGGAGGUG as the query sequence (BLAST search with a word size of 9 was used). The result was 83.3% homology with nucleotides 99425 to 99440 of contig135 within the coding region of gene 2154, which encodes a phosphorylase kinase protein. The lack of complete homology between the genome and this group of sequences was disappointing. However, the reason for the lack of complete homology could be due to mistakes in the SELEX sequences themselves, as explained in section 5.6.

Gene 274 (ser2217) encodes a 569 amino acid protein with 75% identity and 86% similarity to an uncharacterized ATP-ase ABC (ATP-binding cassette) transporter of *Nostoc sp.* PCC 7120 (Anabaena 7120). Interpro analysis (Mulder *et al.*, 2003) of the protein encoded by gene 274 detected the presence of two NTP-binding motifs termed P-loops, one ATP-binding P-loop at the N-terminus (amino acids 29-50) and a GTP-binding

loop near the middle of the protein (amino acids 352-372). Both P-loops are part of individual AAA (ATPases Associated with diverse cellular Activities) ATPase domains (Karata *et al.*, 1999) characteristic of ABC transporters.

Gene 2154 encodes a phosphorylase kinase protein that, according to Interpro analysis, is the alpha subunit of phosphorylase kinase. This purpose of this enzyme is to phosphorylate glycogen phosphorylase at its Serine 14 residue, a process that activates the protein. Active glycogen phosphorylase in turn functions in the breakdown of glycogen to yield glucose-1-phosphate. Cyanobacteria like many prokaryotes and eukaryotes, store glucose in the form of glycogen, to be used accessed in energy depleting conditions (Preiss, 1982). The amount of knowledge concerning cyanobacterial glycogen is very limited therefore it is difficult to make any inferences about a potential function of glycogen breakdown during the cold-shock response.

5.6.2.2 Clone 11

Comparative analysis with the clone 11 sequence revealed nearly complete homology (90.6%) to a region of the non-coding strand within the *dsg* gene (Table 5.5). *Synechococcus* 7942 and 6301 are desaturase group I strains of cyanobacteria (Murata and Wada, 1992), a group that does not have the ability to introduce a second double bond into monounsaturated fatty acids. The desaturase Dsg is a non-specific delta 9 desaturase that can generate a *cis*-double bond at the delta 9 position of both 16 and 18 carbon saturated fatty acids (Ishizaki-Nishizawa, 1996).

Desaturases have been shown to be extremely important in maintaining cellular viability in conditions of cold shock. As mentioned in section 1.3, cold temperature stress causes the membrane of a cell to become more rigid, adversely affecting the ability of the

membrane to perform normal functions of export and import of molecules. In cyanobacteria, fluidity of the membrane is re-established by desaturation of saturated fatty acids in the membrane, a process characterized in both *Anabaena variabilis* (Sato and Murata, 1980) and *Synechocystis sp.* PCC 6803 (Wada and Murata, 1990). The existence of one desaturase and thus the ability to generate only monounsaturated fatty acids in *Synechococcus* 7942 and 6301 has been implicated as the reason why these cyanobacteria are somewhat cold-sensitive (Wada *et al.*, 1990). In fact, increased desaturation of membrane lipids of *Synechococcus* 7942 by introduction of the *desA* gene that encodes a delta 12 desaturase from *Synechocystis* 6803 enhanced the ability of the strain to survive at cold temperature (22°C) (Wada *et al.*, 1994).

Evidence suggests that the mechanism of induced desaturase expression in cold-shock conditions is increased stability of desaturase transcripts. In a study by Sakamoto and Bryant (1997), upon shift of *Synechococcus* 7002 cells from 38°C to 22°C, desaturase transcripts that encode DesA and DesB increased in stability by 3.5 fold and 15 fold respectively. Interestingly, the level of the *desC* transcript was shown to be not dependent upon temperature. However this result does not disprove the possible role of the *Synechococcus* 7942 DesC homologue, Dsg, in cold shock. Although *Synechococcus* 7942 has only one desaturase and is somewhat temperature sensitive, the importance of the Dsg protein to act as a desaturase in cold-stress conditions is evidenced by a study in which expression of the *Synechococcus* 7942 Dsg protein in tobacco plants decreased the level of saturated fatty acids and increased resistance to cold (Ishizaki-Nishizawa, 1996). If RbpA regulates expression of *dsg*, it could do so via a *cis*-encoded antisense RNA that is complementary to nucleotides 288-320 of the ORF located in the *dsg* mRNA. Since

RbpA could presumably increase the expression of Dsg in cold-shock conditions, the binding of RbpA to the antisense RNA could either prevent binding of the inhibitory antisense RNA to the *dsg* mRNA or activate the antisense RNA if binding of *dsg* and the antisense RNA increases expression of Dsg.

5.6.2.3 Clones 3 and 16

Clones 3 and 16 are 41 bases in length and differ in sequence at two nucleotide positions, therefore the two sequences are considered together. The most homologous sequence in the *Synechococcus* 7942 genome (85.4%) was to a region on the non-coding strand within the ORF of the *narB* gene which encodes nitrate reductase (NarB) (Table 5.5). Homology to the non-coding strand of nitrate reductase indicates that any RbpA-mediated mechanism of nitrate reductase regulation would involve a *cis*-encoded antisense RNA (refer to section 5.6).

NarB is involved in the processing of nitrate to nitrite, a two electron reduction reaction involved in the assimilation of nitrate (Rubio *et al.*, 1996). A role for RNA-binding proteins in the regulation of nitrogen metabolism has been identified by Mori *et al.* (2003). In the nitrogen fixing heterocystous cyanobacterium *Anabaena* 7120, the level of the *rbpA*, *rbpB*, *rbpC* and *rbpD* transcripts upon cold-shock were dependent upon nitrogen status suggesting that a relationship exists between nitrogen status and stress responses. Unfortunately the study did not measure the effect of *rbp* gene expression in the presence of nitrate. A role for RNA-binding proteins in repression of heterocyst formation in conditions of low temperature was outlined by Sato and Wada, (1996). In wild-type heterocyst forming cyanobacteria, both nitrate and ammonia repress initiation of heterocyst formation at all temperatures. However, upon inactivation of the *rbpA1* gene, *Anabaena variabilis* M3 was unable to repress heterocyst formation at low temperature when nitrate was present, while the more potent inhibitor ammonia was still

able to repress differentiation. This finding suggests that RbpA1 is involved in the mechanism of nitrate induced repression of heterocyst formation at low temperature.

In non-heterocyst forming cyanobacteria such as *Synechococcus* 7942 and all *Synechococcus* species other than PCC 7335 (a *Synechococcus* species that fixes nitrogen) (Bergman *et al.*, 1997), RbpA would not be involved in repression of heterocyst formation, therefore the question arises: what function would regulation of *narB* levels in the cell by RbpA in conditions of cold-shock serve? In a study by Sakamoto and Bryant (1999), a decrease in the growth rate of *Synechococcus* 6301 in conditions of low temperature (15°C) is attributed to a decrease in the rate of nitrate consumption. The authors believe this could be due to a decreased ability of the cell to import nitrate because of the compromised function of nitrate permease (NtrA) in conditions of cold temperature due to decreased membrane fluidity. Therefore, although nitrate permease is not fully functional, RbpA could be involved in maintaining cellular levels of nitrate reductase to process the limited amount of nitrate imported during conditions of cold shock. In fact, when nitrate uptake is completely shut down at a temperature of 15°C, active nitrate reductase was still present (Sakamoto and Bryant, 1999).

5.6.2.4 Clones 6 and 53

RNA sequences from clone 6 and 53 are considered together because clone 53 contains the entire clone 6 sequence. Comparative sequence analysis of the *Synechococcus* 7942 genome detected significant homology to the coding strand of gene 2163 (Table 5.5). The gene is 1298 nucleotides in length and encodes a 432 amino acid protein characterized as a protein with unknown function by JGI. Interpro analysis of the protein revealed the existence of a winged helix-turn-helix DNA binding domain located within N-terminal 73 amino acids, a glycine-rich sequence encoded in amino acids 122 to 149 and two hypothetical conserved domains of unknown function encoded by amino

acids 151-330 and 333-410. The winged-helix DNA binding domain is found in a large superfamily of proteins that contain a myriad of protein domains in tandem with a winged-helix domain. These proteins function as transcription factors, involved in the transcription regulation of many genes in both prokaryotes and eukaryotes (Gajiwala and Burley, 2000).

Efforts to further characterize the potential function of this protein by identification of homologous proteins by FASTA analysis were unsuccessful since all proteins with significant homology were uncharacterized, hypothetical proteins. Therefore, a putative function of this protein in cyanobacteria cannot be proposed. However, involvement of winged-helix containing proteins in the cold-shock response has been suggested in Archaea by the presence of a gene encoding a helix-turn-helix protein in cold-adapted bacteria (Saunders *et al.*, 2003).

5.6.2.5 Clone 55 of round 14, clones 8, 23 and 43 of round 10

The only sequence that was identified in the *Synechococcus* 7942 genome in both rounds 10 and 14 of genomic SELEX was a sequence located on the non-coding strand within gene 2106 encompassing nucleotides 1713 to 1739 of the 1786 base ORF. This G/U poor sequence was identified by clones 8, 23 and 43 of round 10 (Table 5.3) and clone 55 of round 14 (Table 5.6). COGS homology analysis performed by JGI identified the product of gene 2106 as *putP*, a proline permease that acts as a sodium: proline symporter. PutP acts as the main uptake mechanism of proline (Wood and Zadworny, 1979) in *Salmonella* and *E. coli*, whereby proline taken in by the cell is used not only as a nitrogen and carbon source, and also as a cryoprotectant in conditions of cold temperature (Rudolph and Crowe, 1985). In response to varying conditions of cold, heat and salt concentration, some bacteria have been shown to accumulate a variety of low molecular weight compounds, including proline, that act as chemical chaperones (Tatzelt *et al.*,

1996). These chemical chaperones likely function by increasing the stability of proteins and preventing denaturation and subsequent aggregation. For instance, addition of proline was shown to be involved in heat shock adaptation by activating heat-shock protein chaperones GroEL, DnaK and ClpB in *E. coli* (Diamant *et al.*, 2001). Therefore, in theory, RbpA could upregulate the expression of PutP (via binding to a regulatory antisense RNA) in conditions of cold-shock, to increase the amount of the chemical chaperone proline as part of the diverse cold-response mechanism of maintaining cellular viability.

5.6.2.6 Clones 35 and 43 from round 14 and clone 5 from round 10

The sequences of clones 35 and 43 from round 14 like that of clone 5 from the round 10 RNA are very rich in G/U residues and contain numerous consecutive copies of the sequence UGUGGUG. Clone 5 contains 5 consecutive UGUGGUG repeats and both clones 35 and 43 contain 3 consecutive repeats. The only portion of the *Synechococcus* 7942 genome that has some similarity (75.6%) with this repeat sequence is the coding strand of gene 1922 that encodes the 33kDa subunit of chlorophyll synthase (Tables 5.1, 5.4). The reason for the lack of complete homology in the *Synechococcus* 7942 genome is unclear, the absence of homology could be due to the mistakes in the *Synechococcus* 7942 genomic draft sequence or highly repetitive sequences could have been excluded from the genomic sequence provided by JGI.

Conclusion

The genomic SELEX procedure was an effective method to determine both the ribonucleic acid binding specificity of RbpA and the identity of genes potentially regulated by it. The genomic SELEX library was shown to be sufficiently representative of the *Synechococcus* 7942 genome. Determination of the representation of a 43 bp portion of the *rbpB* gene in the genomic SELEX library by nested-PCR revealed that on average, the library contains in a staggered arrangement, one library molecule per 1.6 bases of the genome or a total of 3.38×10^6 library molecules. This result showed that all sequences potentially bound by RbpA should be present within the RNA binding reaction of the initial round of selection.

It was surprising to find that three classes of RNA sequences (G/U poor, G/U rich and G/U very rich) were isolated in both rounds 10 and 14 of genomic SELEX given that RbpA and other glycine-rich RRM-type RNA-binding proteins have shown a binding preference for poly U and poly G homopolymers. However, since RNA sequences rich or very rich in G/U residues comprise the majority of sequences identified following rounds 10 and 14 of selection and exhibit a greater degree of sequence homology than those poor in G/U residues, RbpA could have a higher binding affinity for RNA sequences predominantly composed of G and U residues.

A putative RbpA consensus binding sequence was characterized by analysis of the RNA sequence alignment of round 14 G/U rich and G/U very rich RNA sequences. Identification of a nearly identical 13 base RNA sequence UGAAUGGGAGGUG sequence eight times in the round 14 G/U rich RNA pool that is highly conserved among

all round 14 G/U rich sequences suggests that this sequence likely contains the RNA-binding sequence of RbpA. Since RNA-binding proteins usually bind to a sequence 5-10 ribonucleotides in length, the actual binding sequence may be shorter than 13 bases in length. Interestingly, the six nucleotide sequence at the 3' end of UGAAUGGGAGGUG is almost identical to a six base sequence, GUGGUG, present in many RNA sequences selected in both rounds 10 and 14 and are present in the three G/U very rich RNA identified in rounds 10 and 14 of genomic SELEX. This suggests that GUGGUG could be the target of RbpA binding. In order to verify this, experiments to detect specific RNA-RbpA complexes by non-denaturing PAGE gel shift analysis must be conducted. In this way, one should be able to investigate the relative binding affinity of RbpA not only to the putative revised consensus sequence but also to sequences in each of the three classes of RNA sequences isolated by genomic SELEX (G/U rich, G/U very rich, G/U poor). In addition, this should definitively prove that RNA selected in the genomic SELEX process was due to RbpA binding and not to an interaction with the Ni^{+2}-NTA resin.

Induction of RbpA expression in conditions of cold-shock suggests that the protein is involved in the cyanobacterial cold-shock response. Alteration of prokaryotic gene expression in cold-shock conditions suggests that the cold-inducible RbpA could be involved in the mechanism of altered gene expression at the RNA level. I was able to identify some intriguing candidate genes potentially regulated by RbpA in the *Synechococcus* 7942 cell. Nearly all round 10 and round 14 RNA sequences were found within an open reading frame (90%). Most sequences were encoded on the coding strand of the ORF (64%) and interestingly some sequences were located on the non-coding strand 36%). Nearly all encoded on the coding-strand were G/U rich or very rich

sequences whereas most sequences located in the non-coding strand were G/U poor. Therefore, RbpA could regulate gene expression by binding to G/U rich sequences within the ORF of an mRNA in an attempt to either increase the half-life of the mRNA or maintain efficient translation by preventing the formation of additional secondary structures that can exist in cold temperature conditions. RbpA could also regulate gene expression by binding to G/U poor sequences within regulatory cis-encoded antisense RNAs. One could probe for the presence of *cis*-encoded antisense RNA in *Synechococcus* 7942 cells when subjected to a variety of conditions including cold-shock to demonstrate that these RNAs are in fact transcribed and potentially function in regulating gene expression by a process involving RbpA.

The most interesting genes identified by comparative genomic analysis with the genomic SELEX sequences were those known to be important for maintaining cellular viability in cold-shock conditions because of the putative function of RbpA as a cold-shock regulator of gene expression. For instance, the *dsg* gene encoding fatty acid desaturase is involved in maintaining cellular viability in cold shock conditions by maintaining the fluidity of the cell membrane by desaturation of membrane fatty acids. The *putP* gene encodes a proline permease protein involved in the accumulation of proline in conditions of cold temperature. Proline is used as a cryoprotectant by some prokaryotes and has been characterized as a chemical chaperone, required to prevent protein denaturation in conditions of heat stress. The *sodB* gene encodes the iron superoxide dismutase protein involved in preventing the accumulation of harmful superoxide radicals in conditions of cold temperature. Therefore, the fact that RbpA recognizes a portion of the RNA sequence encoded within these genes provides evidence

that RbpA may play a role in regulating the expression of genes involved in the cold-shock response.

In addition to cold-regulated genes, other interesting genes potentially regulated by RbpA include genes that encode transcriptional factors such as LrrA and the protein encoded by gene 2163. RbpA could also be involved in nitrate assimilation by regulation of nitrate reductase, in maintaining the cyanobacterial circadian rhythm by regulation of a protein similar to CikA that acts as an environmental sensor, and in the breakdown of glycogen by regulation of glycogen phosphorylase kinase.

In conclusion, although the function of RbpA in the *Synechococcus* 7942 cell has not been elucidated, the RNA sequence binding specificity of the protein is now better characterized and genes potentially regulated by RbpA have been identified for further analysis.

Suggestions for future work

By use of the genomic SELEX protocol, I isolated a large set of RNA molecules potentially recognized by RbpA and I identified a putative RbpA consensus binding sequence. To verify binding of RbpA to these RNA molecules, quantitative assays are required because evidence that protein-mediated selection of RNA sequences occurred during the SELEX procedure was primarily qualitative. Quantitative assays such as an electrophoretic mobility shift assay (EMSA) or a filter binding assay would enable one to verify and thus validate the protein-RNA interaction and also to obtain binding constants for the interaction. In addition, with the use of quantitative assays one could investigate the relative binding affinity of G/U poor, rich and very rich RNA classes to RbpA and determine if G/U poor RNA (that contain the sequence CACCAC) was retained due to a direct interaction with RbpA or via an indirect mechanism such as hybridization to G/U-rich sequences (that contain the sequence GUGGUG).

The best candidate sequences for initial binding studies should be from round 14 of genomic SELEX and be those with the most extensive similarity to the *Synechococcus* 7942 genome sequence such as clone 63 (in the G/U-poor class of selected RNA sequences), or have been selected the greatest number of times such as clone 18 (G/U-rich RNA class) and clone 35 (G/U-very rich RNA class) because retention of multiple copies of a sequence may indicate that these sequences are bound with greatest affinity by the protein. In addition, since the SELEX-selected sequences contain mismatches in comparison to the *Synechococcus* 7942 genome sequence, the genomic sequences themselves should also be used in quantitative binding studies to determine if RbpA

binding to the sequences retained by genomic SELEX is biologically relevant. If RbpA does not bind to the genomic sequence, then RbpA would have retained the SELEX-selected sequence after it was altered by mutation and thus the retention of the SELEX-selected sequence would not be biologically relevant. I would expect that RbpA would have greater binding affinity in each case for the SELEX-selected sequence than for its genomic equivalent, therefore providing evidence that the relatively large number of differences between the SELEX-selected sequences and the *Synechococcus* 7942 genomic sequence) may have occurred due to mutation and subsequent selection of those RNA sequences.

The existence of *cis*-encoded antisense RNA molecules retained in the genomic SELEX process indicated that RbpA could be involved in a mechanism of gene regulation that is poorly understood in prokaryotes. To determine if RbpA actually binds to these sequences, a quantitative assay would be required. The synthesis of these *cis*-encoded antisense RNAs in *Synechoccocus* 7942 would need to be investigated by Northern blot analysis to determine if these molecules are produced in the cyanobacterial cell. Similarly, one could investigate the expression of genes potentially regulated by the RbpA-antisense RNA mechanism by Northern blot analysis in a variety of conditions including cold-shock and in strains that contain wild-type or mutant *rbpA* alleles.

Table A1: *Synechoccus* 7942 genome gene numbers.

Old gene[a] designation	new gene[b] designation
13	-[c]
18	-
51	sef0001
56	ser0006
79	sef0035
141	ser0107
169	ser0142
240	sef0226
269	sef0263
274	-
343	ser0349
467	ser0485
578	ser0609
615	ser0648
616	ser0649
996	ser1109
1105	sef1230
1155	ser1286
1184	-
1215	sef1357
1258	-
1700	-
1701	sef1925
1766	sef2003
1922	sef2200
1928	sef2207
2061	ser2359
2106	ser2417
2154	ser2470
2163	ser2479
2374	ser2719
2520	sef2641
2533	ser1339
2663	-
2748	-
2771	-
2809	-
2816	-

[a] arbitrary gene designation numbers used by the Joint Genome Institute (JGI) during create of a draft sequence of the *Syn.* 7942 genome

[b] Permanent gene names generated upon completion of the Syn. 7942 genome, gene names were accessed at http://ramsites.net/~biolingua/

[c] classified as no longer considered a gene

References

Allain, F.H., Gubser, C.C., Howe, P.W., Nagai, K., Neuhaus, D., Varani, G. (1996) Specificity of ribonucleoprotein interaction determined by RNA folding during complex formation. *Nature*, **380**, 646-650

Anandan, S., Nalty, M,S., Cogdell, D.E., Golden, S.S. (1996) Identification of two classes of transcriptional regulator genes in the cyanobacterium *Synechococcus sp.* strain PCC 7942. *Arch. Microbiol.*, **166**, 58-63

Andersen, J., Forst, S.A., Zhao, K., Inouye, M., Delihas, N. (1989) The function of *micF* RNA. *micF* RNA is a major factor in the thermal regulation of OmpF protein in *Escherichia coli. J. Biol. Chem.*, **264**, 17961-17970

Bae, W., Xia, B., Inouye, M., Severinov, K. (2000) *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proc. Natl. Acad. Sci. USA*, **97**, 7784-7789

B. S. Baker (1989) Sex in flies: the splice of life. *Nature (London),* **340**, 521-524

Bartel, D. and Szostak, J. (1993) Isolation of New Ribozymes from a Large Pool of Random Sequences. *Science*, **261**, 1411-1418

Belbin, T. (1999) Characterization of a gene encoding an RNA-binding protein (*rbpA*) in the cyanobacterium *Synechococcus sp.* PCC7942. Ph.D. thesis, Memorial University of Newfoundland.

Benne, R., and Hershey, J. W. B. (1978) The mechanism of action of protein synthesis initiation factors from rabbit reticulocytes, *J. Biol. Chem.*, **253**, 3078-3087.

Bergman, B., Gallon, J.R., Rai, A.N., Stal, L.J. (1997) N_2 Fixation by non-heterocystous cyanobacteria *FEMS Microbiol. Rev.*, **19**, 139-185

Bernstein, P.L., Peltz, S.W. Ross, J. (1989) The poly(A)-poly(A)-binding protein complex is a major determinant of mRNA stability in vitro. *Mol. Cell Biol.* , **9**, 659–670

Brandi, A., Pietroni, C., Gualerzi, O., Pon, C. (1996) Post-transcriptional regulation of CspA expression in *Escherichia coli. Mol. Microbiol.*, **19**, 231-240

Burd, C.G. and Dreyfuss, G. (1994) Conserved structures and diversity of functions of RNA-binding proteins. *Science*, **265**, 615-621

Burke, D. and Gold, L. (1997) RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX. *Nucleic Acids Res.*, **25**, 2020-2024

Carpousis, A.J. (2002) The *Escherichia coli* RNA degradosome: structure, function and relationship to other ribonucleolytic multienzyme complexes. *Biochem. Soc. Trans.*, **30**, 150-155

Chamot D., Magee, W.C., Yu, E., Owttrim, G.W. (1999) A Cold Shock-Induced Cyanobacterial RNA Helicase. *J.Bacteriol.*, **181**, 1728-1732

Cline, J., Braman, J.C., Hogrefe, H.H. (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acid. Res.*, **24(18)**, 3546-3551

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004) Weblogo: A sequence logo generator. *Genome Res.* **14**, 1188-1190

Cruz, J., Kressler, D., Linder, P. (1999) Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. *Trends Biochem. Sci.*, **24**, 192-198

Diamant, S., Eliahu, N., Rosenthal, D., Goloubinoff, P. (2001) Chemical chaperones regulate molecular chaperones in vitro and in cells under combined salt and heat stresses. *J. Biol. Chem.*, **276**, 39586-39591

Ellington, A. and Szostak, J. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818-822

Ehira, S., Hamano, T., Hayashida, T., Kojima, K., Nakamoto, H., Hiyama, T., Ohmori, M., Shivaji, S., Sato, N. (2003) Conserved temperature-dependent expression of RNA-binding proteins in cyanobacteria with different temperature optima. *FEMS Microbiol. Lett.*, **225**, 137-142

Erickson, H.P. (1997) FtsZ, a tubulin homologue in prokaryote cell division. *Trends Cell Biol.*, 7, 362-367

Etchegaray, J., Jones, P., Inouye, M. (1996) Differential thermoregulation of two highly homologous cold-shock genes, *cspA* and *cspB*, of *Escherichia coli*. *Genes Cells*, **1**, 171-178

Feng, Y., Huang, H., Liao, J., Cohen, S. (2001) *Esherichia coli* poly(A)-binding proteins that interact with components of degradosomes or impede RNA decay mediated by polynucleotide phosphorylase and RNase E. *J. Biol. Chem.*, **276**, 31651-31656

Franks, F. (1995) Protein destabilization at low-temperatures. *Adv. Prot. Chem.*, **46,** 105-139

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, **281**, 375-388

Fraunholz, M.J., Moerschel, E., Maier, U.G. (1998) The chloroplast division protein FtsZ is encoded by a nucleomorph gene in cryptomonads. *Mol. Gen. Genetics*, **260**, 207-211

Gajiwala, K.S. and Burley, S.K. (2000) Winged helix proteins. *Curr. Opin. Struct. Biol.*, **10,** 110–116

Ghisolfi L, Kharrat A, Joseph G, Amalric F, Erard M. (1992) Concerted activities of the RNA recognition and the glycine-rich C-terminal domains of nucleolin are required for efficient complex formation with pre-ribosomal RNA. *Eur. J. Biochem.*, **209**, 541-548

Gibson, T.J., Thompson, J.D., Heringa, J. (1993) The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. *FEBS Lett.*, **324**, 361-6.

Goldenburg, D., Azar, I., Oppenheim, A. (1996) Differential mRNA stability of the cspA gene in the cold-shock response of *Escherichia coli*. *Mol. Microbiol.*, **19**, 241-248

Gomis-Ruth, F. (2003) Structural aspects of the metzincin clan of metalloendopeptidases. *Mol. Biotechnol.*, **24**, 157-202

Gorbalenya, A.E. and Koonin, E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419-429

Graumann, P. and Marahiel, M. (1996) A case of convergent evolution of nucleic acid binding modules. *Bioessays*, **18**, 309-315

Gray, M.W. (1989) The evolutionary origins of organelles. *Trends Genet.*, **5**, 294-299

Gualerzi, C., Giuliodori, A., Pon, C. (2003) Transcriptional and Post-transcriptional Control of Cold-shock Genes. *J. Mol. Biol.*, **331**, 527-539

Guex, H. and Peitsch, M.C. (1997) Swiss-model and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis*, **18**, 2714-2723

Herdman, M., Castenholz, R., Iteman, I., Waterbury, J., Rippka, R. (2001) in Subsection I: Bergey's Manual of Systematic Bacteriology, eds. D.R. Boone, R.W. Castenholz, G.M. Garrity, pp. 721., Springer-Verlag, New-York

Hofmann, H., Limmer, S., Hornung, V., Sprinzl, M. (1997) Ni^{+2}-binding RNA motifs with an asymmetric purine-rich internal loop and a G-A base pair. *RNA*, **3**, 1289-1300

Hossain, M. and Nakamoto, H. (2002) HtpG plays a role in cold acclimation in cyanobacteria. *Curr. Microbiol.*, **44**, 291-296

Ikeuchi, M. (1996) Complete genome sequence of a cyanobacterium *Synechocystis sp.* PCC 6803, the oxygenic photosynthetic prokaryote. *Tanpakushitsu Kakusan Koso.*, **41**, 2579-83

Ichetovkin, I.E., Abramochkin, G., Shrader, T.E. (1997) Substrate recognition by the Leucyl/Phenylalanyl-tRNA-protein transferase *J. Biol. Chem.*, **272**, 33009-33014

Ishizaki-Nishizawa O, Fujii T, Azuma M, Sekiguchi K, Murata N, Ohtani T, Toguri T. (1996) Low-temperature resistance of higher plants is significantly enhanced by a nonspecific cyanobacterial desaturase. *Nat Biotechnol.*, **14**, 1003-6

Jacobs, J.S., Anderson, J., Parker, R. (1998) The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the Ski2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. *EMBO J.*, **17**, 1497-1506

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. and Gibson, T. J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, **23**, 403-5

Jensen, K., Atkinson, B., Willis, M., Koch, T., Gold, L. (1995) Using *in vitro* selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands. *Proc. Natl. Acad. Sci. USA.*, **92,** 12220-12224

Jensen, M.A. and Straus, N. (1993) Effect of PCR conditions on the formation of heteroduplex and single-stranded DNA in the amplification of bacterial ribosomal DNA spacer regions. *PCR Methods Applic.* **3**:186-194

Jiang, W., Jones, P., Inouye, M. (1993) Chloramphenicol induces the transcription of the major cold shock protein of *Escherichia coli*, CspA. *J. Bacteriol.*, **175**, 5824-5828

Jiang, W., Hou, Y., Inouye, M. (1997) CspA, the major cold-shock protein of *Escherichia coli,* is an RNA chaperone. *J. Biol. Chem.*, **272**, 196-202

Jones, P.G., Mitta, M., Kim, Y., Jiang, W., Inouye, M. (1996) Cold shock induces a major ribosomal-associated protein that unwinds double-stranded RNA in *Escherichia coli. Proc. Natl. Acad. Sci. USA*, **93**, 76-80

Kaji, A., Kaji, H., Novelli, D. (1965) Soluble Amino Acid-incorporating System. I. Preparation of the system and nature of the reaction. *J. Biol. Chem.*, **240**, 1185-1197

Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara, *et al.* (2001) Complete

genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena sp.* strain PCC 7120. *DNA Res.*, **8**, 205-13, 227-53

Karp, P.D., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I.T., Saier, M.H. (2004) The *E.coli* Ecocyc Database: No Longer Just a Metabolic Pathway Database. *ASM News*, **70**, 25-30

Kiledjian, M. and Dreyfuss, G. (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *EMBO J.*, **11**, 2655-64

Kim, S., Shi, H., Lee, D-K., Lis, J.T. (2003) Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nucleic Acids Res.*, **31**, 1955-1961

Landsman, D. (1992) RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. *Nucleic Acids Res.*, **20**, 2861-2864

Laudenbach, D.E., Trick, C.G., Straus, N.A. (1989) Cloning and characterization of an *Anacystis nidulans* R2 superoxide dismutase gene. *Mol. Gen. Genet.*, **216**, 455-461

Las Rivas, J., Balsera, M., Barber, J. (2004) Evolution of oxygenic photosynthesis: genome-wide analysis of the OEC extrinsic proteins. *Trends Plant Sci.*, **9**, 18-25

Lazinski, D., Grzadzielska, A., Das, A. (1989) Sequence-specific recognition of RNA hairpins by bacteriophage antiterminators requires a conserved arginine-rich motif. *Cell*, **59**, 207-18

Lease, R.A. and Belfort, M. (2000) Riboregulation by DsrA RNA: trans-actions for global economy. *Mol. Microbiol.*, **38**, 667-672

Lease, R.A., Cusick, M.E., Belfort, M. (1998) Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl. Acad. Sci. USA*, **95**, 12456-12461

Lee, F.J., Hesselberth, J.R., Meyers, L.A., Ellington, A.D. (2004) Aptamer Database. *Nucleic Acids Res.*, **32**, D95-D100

Lee, A.L., Kannar, R., Rio, D.C., Wemmer, D.E. (1994) Resonance assignments and solution structure of the second RNA-binding domain of sex-lethal determined by multidimensional heteronuclear magnetic resonance. *Biochemistry*, **33**, 13775-13786

Los, D.A., Horvath, I., Vigh, L., Murata, N. (1993) The temperature-dependent expression of the desaturase gene *desA* in *Synechocystis* PCC 6803. *FEBS Lett.*, **378**, 57-60

Lutz, C. and Alwine, J. (1994) Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. *Genes Dev.*, **8**, 576-86.

Majdalani, N., Cunning, C., Sledjeski, D., Elliot, T., Gottesman, S. (1998) DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, indepedent of its action as an antisilencer of transcription. *Proc. Natl. Acad. Sci. USA*, **95**, 12462-12467

Maruyama, K., Sato, N., Ohta, N. (1999) Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins. *Nucleic Acids Res.*, **27**, 2029-2036

Maruyama,K. and Sato,N (2001) Characterization of Type 2 RNA-binding proteins in cyanobacteria. Submitted to the EMBL/GenBank/DDBJ databases

McBride, M.J. (2001) Bacterial gliding motility: multiple mechanisms for cell movement over surfaces. *Annu. Rev. Microbiol.*, **55**, 49-75

McFadden, G.I. (1999) Endosymbiosis and evolution of the plant cell. *Curr. Opin. Plant Biol.*, **2**, 513-519

Moreira, D., Le Guyader, H., Philippe H. (2000) The origin of red algae and the evolution of chloroplasts. *Nature*, **405**, 69-72

Mori, S., Castoreno, A., Mulligan, M.E., Lammers P.J. (2003) Nitrogen status modulates the expression of RNA-binding proteins in cyanobacteria. *FEMS Microbiol. Lett.*, **227**, 203-210

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U. et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315-318

Mulligan, M.E. and Belbin, T.J. (1995) Characterization of RNA-binding protein genes in cyanobacteria. *Nucleic Acids Symp. Ser.*, **33**, 140-142

Mulligan, M.E., Jackman, D.M., Murphy, S.T. (1994) Heterocyst-forming Filamentous Cyanobacteria Encode Proteins that Resemble Eukaryotic RNA-binding Proteins of the RNP family *J. Mol. Biol.*, **235**, 1162-1170

Murata, N., Los, D.A. (1997) Membrane Fluidity and Temperature Perception. *Plant Physiol.*, **115**, 875-879

Murata, N., Wada, H., Gombos, Z. (1992) Modes of fatty-acid desaturation in cyanobacteria. *Plant Cell Physiol.*, **33**, 933-941

Mutsuda, M., Michel, K.P., Zhang, X., Montgomery, B.L., Golden, S.S. (2003) Biochemical properties of CikA, an unusual phytochrome-like histidine protein kinase the resets the circadian clock in *Synechococcus elongatus* PCC 7942. *J. Biol. Chem.*, **278**, 19102-19110

Nagai, K., Oubridge, C., Jessen, T.H., Li, J., Evans, P.R. (1990) Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature*, **348,** 515-520

Nakamura Y, Kaneko T, Sato S, Ikeuchi M, Katoh H, Sasamoto S, Watanabe A, Iriguchi M, Kawashima K, Kimura T, Kishida Y, Kiyokawa, *et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res.*, **9**,123-30

Nakashima, K., Kanamaru, K., Mizuno, T., Horikoshi, K. (1996) A novel member of the *cspA* family of genes that is induced by cold shock in *Escherichia coli*. *J. Bacteriol.*, **178**, 2994-2997

Neuhaus, K.S., Rapposch, K., Francis, P., Scherer, S. (2000) Restart of exponential growth of cold-shocked *Yersinia enterocolitica* occurs after downregulation of *cspA1/A2* mRNA. *J. Bacteriol.*, **182**, 3285-3288

Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S.D., Inouye, M., Montelione, G.T. (1994) Solution NMR structure of the major cold shock protein (CspA) from *Escherichia coli*: identification of a binding epitope for DNA. *Proc. Natl. Acad. Sci. USA*, **91**, 5114-5118

Nierhaus, K. (1991) The assembly of prokaryotic ribosomes. *Biochimie*, **73**, 739-755

Nobles, D.R., Romanovicz, D.K., Brown, R.M. (2001) Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase? *Plant Physiol.*, **127**, 529-542

Osteryoung, K.W., Stokes, K.D., Rutherford, S.M., Percival, A.L., Lee, W.Y. (1998) Chloroplast division in higher plants requires members of two functionally divergent gene families with homology to bacterial *ftsZ*. *Plant Cell*, **10**, 1991-2004

Osteryoung, K.W. and Pyke, K.A. (1998) Plastid division: evidence for a prokaryotically derived mechanism. *Curr. Opin. Plant Biol.*, **1**, 475-479

Oubridge, C., Ito, N., Evans, P.R., Teo, C.H. Nagai, K. (1994) Crystal structure at 1.92Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, **372**, 432-438

Paquin, B., Kathe, S.D., Nierzwicki-Bauer, S.A. and Shub, D.A. (1997) Origin and evolution of group I introns in cyanobacterial tRNA genes. *J. Bacteriol.*, **179**, 6798-6806

Park, H.I. and Ming, L.J. (2002) Mechanistic studies of the astacin-like Serratia metalloendopeptidase serralysin: highly active (>2000%) Co(II) and Cu(II) derivatives for further corroboration of a "metallotriad" mechanism. *J. Biol. Inorg. Chem.*, **7**, 600-610

Pearson, W.R. and Lipman, D.J. (1988) Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444- 2448

Pearson, W.R. (1990) Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Meth. Enzymol.*, **183**, 63 - 98.

Phadtare, S., Inouye, M. (1999) Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of *Esherichia coli*. *Mol. Microbiol.*, **33**, 1004-1014

Ponomarenko, J.V., Orlova, G.V., Ponomarenko, M.P., Lavryushev, S.V., Frolov, A.S., Zybova, S.V., Kolchanov, N.A. (2000) SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. *Nucleic Acids Res.*, **28**, 205-208

Porankiewicz, J. and Clarke, A. K. (1997) Induction of the heat shock protein ClpB affects cold acclimation in the cyanobacterium *Synechococcus sp.* PCC 7942. *J. Bacteriol.*, **179**, 5111-5117

Porankiewicz, J., Schelin, J. & Clarke, A. K. (1998) The ATP-dependent Clp protease is essential for acclimation to UV-B and low temperature in the cyanobacterium *Synechococcus*. *Mol. Microbiol.*, **29,** 275–283.

Preiss, J. (1982) Bacterial glycogen synthesis and its regulation. *Annu. Rev. Microbiol.*, **38**, 419-458.

Reinhold-Hurek, B. and Shub, D.A. (1992) Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature*, **357**, 173-176

Rippka, R., Deruelles, J., Waterbury, J.B., Herdman, M., Stanier, R.Y. (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J. Gen. Microbiol.*, **111**, 1-61

Rubio, L.M., Herrero, A., Flores, E. (1996) A cyanobacterial *narB* gene encodes a ferrodoxin-dependent nitrate reductase. *Plant Mol. Biol.*, **30**, 845-850

Rudolph, A.S., Crowe, J.H. (1985) Membrane stabilization during freezing: the role of natural cryoprotectants, trehalose and proline. *Cryobiology*, **22**, 367-377

Ruiz-Lozano P., Doevendans, P., Brown, A., Gruber, P.J., Chien, K.R. (1997) Developmental expression of the murine spliceosome-associated protein mSAP49. *Dev. Dyn.*, **208**, 482-490

Sakamoto, T. and Bryant, D.A. (1999) Nitrate transport and not photoinhibition limits growth of the freshwater cyanobacterium *Synechococcus* species PCC 6301 at low temperature. *Plant Physiol.*, **119**, 785-794

Sakamoto, T., Bryant, D.A. (1997) Temperature regulated mRNA accumulation and stabilization for fatty acid desaturase genes in cyanobacterium *Synechococcus sp.* strain PCC 7002. *Mol. Microbiol.*, **23**, 1281-1292

Sato, N. (1994) A cold regulated cyanobacterial gene cluster encodes RNA-binding protein and ribosomal protein S21. *Plant Mol. Biol.*, **24**, 819-823

Sato, N. (1995) A family of cold-regulated RNA-binding protein genes in the cyanobacterium Anabaena variabilis M3. *Nucleic Acids Res.*, **23**, 2161-2167

Sato, N. (2002) Comparative analysis of the genomes of cyanobacteria and plants. *Genome Informatics*, **13**, 173-182

Sato, N. and Maruyama, M. (1997) Differential regulation by low temperature of the gene for an RNA-binding protein, *rbp A3*, in the Cyanobacterium *Anabaena variabilis* strain M3. *Plant Cell Physiol.* **38**, 81-86

Sato, N., Murata, N. (1982) Temperature shift-induced responses in lipids in the blue-green alga *Anabaena variabilis*: the central role of diacylmonogalactosylglycerol in thermo-adaptation. *Biochim. Biophys. Acta*, **619**, 353-366

Sato, N. and Nakamura, A. (1998) Involvement of the 5'-untranslated region in cold-regulated expression of the *rbpA1* gene in the cyanobacterium *Anabaena variabilis* M3. *Nucleic Acids Res.*, **26**, 2192-2199

Sato, N. and Wada, A. (1996) Disruption analysis of the gene for a cold-regulated RNA-binding protein, *rbpA1*, in Anabaena: cold-induced initiation of the heterocyst differentiation pathway. *Plant Cell Physiol.*, **37(8)**, 1150-1160

Saunders NF, Thomas T, Curmi PM, Mattick JS, Kuczek E, Slade R, Davis J, Franzmann PD, Boone D, Rusterholtz K, Feldman R, Gates C, Bench S, Sowers K, Kadner K, Aerts A, Dehal P, Detter C, Glavina T, Lucas S, Richardson P, Larimer F, Hauser L, Land M, Cavicchioli R (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.*, **13**,1580-8

Scanlan, D. J. and Mann, N.H. (1996) Sequence Analysis of a Genomic Region (Accession No. U70379) from the Cyanobacterium *Synechococcus sp.* PCC7942 Encoding a NarL Homologue and a Two-Component System Pair. *Plant Physiol.*, **112**, 1398

Schagger, H. and von Jagow, G. (1987) Tricine-Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis for the Separation of Proteins in the Range from 1 to 100 kDa. *Anal. Biochem.*, **166**, 368-379

Schell, M.A. (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu. Rev. Microbiol.*, **47**, 597-626

Schindelin, H., Jiang, W., Inouye, M., Heinemann, U. (1994) Crystal structure of CspA, the major cold shock protein of *Esherichia coli. Proc. Natl. Acad. Sci. USA*, **91**, 5119-5123

Schopf, J. W.(ed.) 1983. Earth's Earliest Biosphere. *Princeton Univ. Press, NJ, USA*.

Schmitz O, Katayama M, Williams SB, Kondo T, Golden SS (2000) CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. *Science,* **289,** 765–768

Schneider, T.D., Stephens, R.M. (1990) Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, **18**:6097-6100

Schröder, K., Graumann, P., Schnuchel, A., Holak, T. A. & Marahiel, M. A. (1995) Mutational analysis of the putative nucleic acid-binding surface of thecold-shock domain, CspB, revealed an essential role of aromatic and basic residues in binding of single-stranded DNA containing the Y-box motif. *Mol. Microbiol,*, **16,** 699–708.

Shestakov, S. and Khyen, N. (1970) Evidence for genetic transformation in blue-green alga *Anacystis nidulans* R2. *Molec. Gen. Genet.*, **107**, 372-375

Shi, H., and Xu, R-M. (2003) Crystal structure of the *Drosophila* Mago nashi-Y14 complex. *Genes Dev.*, **17**, 971-976

Shtatland, T., Gill, S.C., Javornik, B.E., Johansson, H.E., Singer, B.S., Uhlenbeck, O.C., Zichi, D.A., Gold, L. (2000) Interactions of *Escherichia coli* RNA with bacteriophage MS2 coat protein: genomic SELEX. *Nucleic Acids Res.*, **28**, E93

Singer, B., Shtatland, T., Brown, D., Gold, L. (1997) Libraries for genomic SELEX. *Nucleic Acids Res.*, **25**, 781-786

Sledjeski, D.D., Gupta, A., Gottesman, S. (1996) The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in *Escherichia coli*. *EMBO J.*, **15**, 3993-4000

Steitz, T.A., Rould, M., Perona, J.J. (1993) Structural basis for Transfer RNA Aminoacylation by *Escherichia coli* Glutaminyl-tRNA Synthetase. *Biochemistry*, **32**, 8758-8771

Stragier, P. and Patte, J.C. (1983) Regulation of diaminopimelate decarboxylase synthesis in *Escherichia coli* III. Nucleotide sequence and regulation of the *lysR* gene. *J. Mol. Biol.*, **168**, 333-350

Strepp, R., Scholz, S., Kruse, S., Speth, V., Reski, R. (1998) Plant nuclear gene knockout reveals a role in plastid division for the homolog of the bacterial cell division protein FtsZ, an ancestral tubulin. *Proc. Natl. Acad. Sci. USA*, **95**, 4368-4373

Sugita, C., Mutsuda, M., Sugiura, M., Sugita, M. (1999) Targeted deletion of genes for eukaryotic RNA-binding proteins, Rbp1 and Rbp2, in the cyanobacterium *Synechococcus* sp. strain PCC7942: Rbp1 is indispensable for cell growth at low temperatures. *FEMS Microbiol. Lett.*, **176**, 155-161

Sugita, M., Sugishita, H., Fujishiro, T., Tsuboi, M., Sugita, C., Endo, T., Sugiura, M. (1997) Organization of a large cluster encoding ribosomal proteins in the cyanobacterium *Synechococcus* sp. strain PCC 6301: comparison of gene clusters among cyanobacteria, eubacteria and chloroplast genomes. *Gene*, **195**, 73-79

Sugita, M. and Sugiura, M. (1994) The existence of eukaryotic ribonucleoprotein consensus sequence-type RNA-binding proteins in a prokaryote, *Synechococcus* 6301. *Nucleic Acids Res.*, **22**, 25-31

Suzuki, I., Kanesaki, Y., Mikami, K., Kanehisa, M., Murata, N. (2001) Cold-regulated genes under control of the cold sensor Hik33 in *Synechocystis*. *Mol. Microbiol.*, **40**, 235-244

Suzuki, I., Los, D.A., Kanesaki, Y., Mikami, K., Murata, N. (2000) The pathway for perception and transduction of low-temperature signals in *Synechocystis*. *EMBO J.*, **19**, 1327-1334

Tanabe, H., Goldstein, J., Yang, M., Inouye, M. (1992) Identification of the Promoter Region of the *Escherichia coli* Major Cold Shock Gene, *cspA*. *J. Bacteriol.*, **174**, 3867-3873

Tatzelt, J., Prusiner, S.B., Welch, W.J. (1996) Chemical chaperones interfere with the formation of scrapie prion protein. *EMBO J.*, **15**, 6363-6373

Thieringer, H.A., Jones, P.G., Inouye, M. (1998) Cold shock and adaptation. *Bioessays*, **20**, 49-57

Thomas, D.J., Thomas, J.B., Prier, S.D., Nasso, N.E., Herbert, S.K. (1999) Iron superoxide dismutase protects against chilling damage in the cyanobacterium *Synechococcus* species PCC 7942. *Plant Physiol.*, **120**, 275-282

Thomas, D.J., Avenson, T.J., Thomas, J.B., Herbert, S.K. (1998) A cyanobacterium lacking iron superoxide dismutase is sensitized to oxidative stress induced with methyl viologen but not sensitized to oxidative stress induced with norflurazon. *Plant Physiol.*, **116**, 1593-1602

Tian, Q., Streuli, M., Saito, H., Schlossman, S.F., Anderson, P. (1991) A polyadenylate binding protein localized to the granules of cytolytic lymphocytes induces DNA fragmentation in target cells. *Cell*, **67**, 629-639

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG., and 37 others. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539-547

Tomizawa, J. and Itoh, T. (1981) Plasmid ColE1 incompatibility determined by interaction of RNA I with primer transcript. *Proc. Natl. Acad. Sci. USA*, **78**, 6096-6100

Tsang, E.W.T., Bowler, C., Hérouart, D., Van Camp, W., Villarroel, R., Genetello, C., Van Montagu, M. and Inzé, D. (1991) Differential regulation of superoxide dismutases in plants exposed to environmental stress. *Plant Cell,* **3**, 783–92

Tuerk, C. and Gold, L. (1990) Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science*, **249**, 505-510

Varshavsky, A. (1995) The N-end rule. *Cold Spring Harb. Symp. Quant. Biol.*, **60**, 461-478

Wada, H., Murata, N. (1990) Temperature-induced changes in the fatty acid composition of the cyanobacterium, *Synechocystis* PCC 6803. *Plant Physiol.*, **92**, 1062-1069

Wada, H., Gombos, Z., Murata, N. (1990) Enhancement of chilling tolerance of a cyanobacterium by genetic manipulation of fatty acid desaturation. *Nature*, **347**, 200-203

Wada, H., Gombos, Z., Murata, N. (1994) Contribution of membrane lipids to the ability of the photosynthetic machinery to tolerate temperature stress. *Proc. Natl. Acad. Sci. USA*, **91**, 4273-4277

Wagner, E.G., Altuvia, S., Romby, P. (2002) Antisense RNAs in Bacteria and Their Genetic Elements. In *Advances in Genetics*, vol. 46, pp. 361-398, Elsevier Science (USA)

Wang, N., Yamanaka, K., Inouye, M. (1999) CspI, the ninth member of the CspA family of *Escherichia coli*, is induced upon cold shock. *J. Bacteriol.*, **181**, 1603-1609

Wood, J.M., Zadworny, D. (1979) Characterization of an inducible porter required for L-proline catabolism by *Escherichia coli* K12. *Can. J. Biochem.*, **57**, 1191-1199

Xia, B., Ke, H., Inouye, M. Aquirement of cold sensitivity by quadruple deletion of the *cspA* family and its suppression by PNPase S1 domain in *Escherichia coli. Mol. Microbiol.*, **40**, 179-188

Yamanaka, K., Mitta, M., Inouye, M. (1999) Mutation analysis of the 5' untranslated region of the cold shock *cspA* mRNA of *Escherichia coli. J. Bacteriol.*, **181**, 6284-6291

Yanofsky, C. (2000) Transcription attenuation: Once viewed as a Novel Regulatory Strategy. *J. Bacteriol.*, **182(1)**, 1-8

Zamore, P.D. and Green, M.R. (1992) Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, **355**, 609-614

Zdobnov, E.M. and Apweiler, R. (2001) InterproScan - an integration platform for the signature-recognition methods in Interpro. *Bioinformatics*, **17,** 847-848

Zolotukhin, A.S., Michalowski, D., Smulevitch, S., Felber, B.K. (2001) Retroviral constitutive transport element evolved from cellular TAP(NXF1)-binding sequences. *J. Virol.*, **75**, 5567-5575