UNIVERSITY OF LATVIA
Faculty of Biology


Jānis Rūmnieks

Doctoral Thesis

# Studies of protein and genome structure in the single-stranded RNA bacteriophages

Promotion to the degree of Doctor of Biology
Molecular Biology

Rīga 2015

This work has been carried out in the Biomedical Research and Study Center during years 2008 – 2014.

Latvian Biomedical Research and Study Centre

This work has been supported by the Latvian Council of Science (Project 09.1294).

This work has also been supported by the European Social Fund within the project "Support for Doctoral Studies at University of Latvia".

EIROPAS SOCIĀLAIS FONDS

EIROPAS SAVIENĪBA

IEGULDĪJUMS TAVĀ NĀKOTNĒ

LATVIJAS UNIVERSITĀTE ANNO 1919

This Doctoral Thesis is a summary of publications.

Supervisor: Dr. biol. Kaspars Tārs

Reviewers: Dr. biol. Jānis Kloviņš, Biomedical Research and Study Center
Dr. chem. Kristaps Jaudzems, Latvian Institute of Organic Synthesis
David Peabody, Ph.D., University of New Mexico

University of Latvia, Promotion Council of Biology

Chairman: Dr. biol. Jānis Kloviņš

# SUMMARY

Bacteriophages of the *Leviviridae* family are the simplest known phages with single-stranded, positive-sense RNA genomes about 3500-4200 nucleotides long that encode only four proteins. They have been extensively used as models to study translational control mechanisms, virus evolution, structure, and assembly. However, despite their simplicity, several aspects of their life cycle are still poorly understood, particularly from a structural viewpoint, and the available genome sequences of the RNA phages leave some open questions about their evolution. Therefore the aim of this thesis was to gain new knowledge about the three-dimensional structure of RNA phage proteins and protein – RNA complexes and to better understand the evolution of RNA phage genomes and RNA secondary structures in them.

First, the crystal structure of the read-through domain of the A1 protein from bacteriophage Qβ was determined at 1.8 Å resolution. A1 is a minor coat protein species in Qβ capsids that is formed when ribosomes occasionally read-through the leaky stop codon of the coat protein gene. The structure revealed that the read-through domain has a mixed α/β architecture and a prominent polyproline type II helix at the N-terminal part. The overall fold of the domain was not similar to other known proteins. Protein – RNA interactions in the RNA phages were studied by determining the crystal structure of Qβ coat protein in complex with an RNA operator hairpin of the replicase gene which the coat protein binds to downregulate production of the replicase. The structure showed that the RNA binding mode of the Qβ coat protein shares several features with that of the widely studied phage MS2, but only the adenine base in the hairpin loop makes sequence-specific contacts with the protein. Unlike MS2 and other RNA phages, the Qβ coat protein uses a stacking interaction with a tyrosine side chain to accommodate a bulged adenine base in the hairpin stem. The structure also revealed that the extended loop between β strands E and F of Qβ coat protein makes contacts with the lower part of the RNA stem, explaining the greater length-dependence of the RNA helix for Qβ.

To study evolution of the RNA phages, genome sequences of the IncM conjugative plasmid-dependent phage M and *Caulobacter* phage φCb5 were determined and analyzed. The genomes had the canonical maturation-coat-replicase genome organization, but, surprisingly, in both cases the lysis genes completely overlapped with the replicase gene in a different reading frame. Analysis of conserved RNA secondary structures in the genomes provided more insight into the evolution of the RNA phages infecting different bacterial genera and the diversification of those using distinct conjugative pili for infection. Consequently, a phylogenetic tree is proposed in an attempt to reconstruct the evolutionary history of the *Leviviridae* family.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| Å | ångström ($10^{-10}$ m) |
| aa-tRNA | aminoacyl-transfer ribonucleic acid |
| ATP | adenosine triphosphate |
| ATPase | adenosine triphosphatase |
| cDNA | complementary deoxyribonucleic acid |
| dATP | deoxyadenosine triphosphate |
| dGTP | deoxyguanosine triphosphate |
| DNA | deoxyribonucleic acid |
| dsDNA | double-stranded deoxyribonucleic acid |
| dsRNA | double-stranded ribonucleic acid |
| EM | electron microscopy |
| GDP | guanosine diphosphate |
| GTP | guanosine triphosphate |
| MIR | multiple isomorphous replacement |
| MIRAS | multiple isomorphous replacement with anomalous scattering |
| MME | monomethyl ether |
| MR | molecular replacement |
| mRNA | messenger ribonucleic acid |
| NTP | nucleotide triphosphate |
| ORF | open reading frame |
| PCR | polymerase chain reaction |
| PDB | Protein Data Bank |
| PEG | polyethylene glycol |
| RdRp | RNA-dependent RNA polymerase |
| rmsd | root mean square deviation |
| RNA | ribonucleic acid |
| RNAse | ribonuclease |
| SD | Shine-Dalgarno |
| SDS | sodium dodecyl sulfate |
| SDS-PAGE | sodium dodecyl sulfate polyacrylamide gel electrophoresis |
| ssDNA | single-stranded deoxyribonucleic acid |
| ssRNA | single-stranded ribonucleic acid |
| T4S | type 4 secretion |
| Tris | tris(hydroxymethyl)aminomethane |
| tRNA | transfer ribonucleic acid |
| UTR | untranslated region |

# INTRODUCTION

Bacteriophages or phages are the most abundant biological entities on Earth. It has been estimated that there are more than $10^{30}$ phage particles on this planet, and the vast majority of those are DNA-containing viruses, dominated by large tailed phages with double-stranded DNA genomes. Among the rest, there are some that contain single-stranded RNA (ssRNA) as their genome, and these are phages of the *Leviviridae* family, the smallest of the known bacteriophages with simple spherical capsids and just four genes. Despite their seeming simplicity, studies on ssRNA phage proteins and RNA have turned out to be a remarkably rich source of information about translational control mechanisms, protein – RNA interactions, assembly of virus particles, RNA secondary structure and virus evolution. Nonetheless, several aspects of ssRNA phage biology are still poorly understood.

The structure and function of biological macromolecules are inseparably linked, and in order to truly understand the way the phage proteins work and the biological mechanisms they accomplish, one must determine their three-dimensional structure. The structure of the genomic RNA is also of crucial importance, and three out the four phage proteins are RNA binding proteins that recognize specific secondary and tertiary RNA structures at some point during the viral life cycle. Therefore studies on the three-dimensional structure of phage proteins alone often cannot give complete answers about how they function and have to be complemented with structural studies of the protein in complex with the RNA it interacts with. Such studies can provide valuable information about the co-evolution of protein and RNA structure, but even more insight can often be gained when the structures are considered in context of the evolution of whole genomes and secondary structure elements in them.

The aim of my thesis was to gain new knowledge about the three-dimensional structure of ssRNA phage proteins and protein – RNA complexes and to better understand the evolution of ssRNA phage genomes and RNA secondary structures in them. The specific tasks for achieving this were

- to determine and analyze the three-dimensional structure of the minor coat protein A1 from bacteriophage Qβ;
- to determine the three-dimensional structure of bacteriophage Qβ coat protein in complex with an RNA operator hairpin of the replicase gene and compare it to the coat protein – RNA complex structures from other ssRNA phages;
- to determine and analyze the complete genome sequence of bacteriophage M and perform analysis of the genome sequence of bacteriophage φCb5.

This thesis is based on four original research papers that will be referred to by their Roman numerals:

I. Rumnieks J, Tars K. 2011. Crystal structure of the read-through domain from bacteriophage Qβ A1 protein. *Protein Sci.* **20**, 1707-1712.

II. Rumnieks J, Tars K. 2014. Crystal structure of the bacteriophage Qβ coat protein in complex with the RNA operator of the replicase gene. *J Mol Biol.* **426**, 1039-1049.

III. Kazaks A., Voronkova T., Rumnieks J., Dishlers A., Tars K. 2011. Genome structure of Caulobacter phage phiCb5. *J Virol.* **85**, 4628-4631.

IV. Rumnieks J, Tars K. 2012. Diversity of pili-specific bacteriophages: genome sequence of IncM plasmid-dependent RNA phage M. *BMC Microbiol.* **12**, 277.

Reprints of the papers are included in the Appendix.

# 1. RNA BACTERIOPHAGES: A LITERATURE REVIEW

The scientific history of the small RNA bacteriophages began in the late 1950s when Tim Loeb at the Rockefeller institute decided to look if there were phages that could infect only specific "mating types" of *Escherichia coli* (Zinder, 1975). Conjugation, a way of transferring genetic material between bacterial cells, was well known at the time, and three "mating types", F-, F+ and Hfr, had been identified (Cavalli et al., 1953). These differed in their ability to serve as recipients or donors during conjugation, the resulting frequency of recombination, the ability to transfer the "fertility" or "F" factor to the recipient cells as well as some other physiological properties. Loeb was indeed able to isolate some phages from New York sewage that were able to infect only the donor or "male" type (F+ and Hfr) of bacteria (Loeb, 1960). The second phage stock, f2, turned out to be a small, spherical virus containing RNA as the genetic material and was the first known RNA-containing bacteriophage at the time (Loeb and Zinder, 1961).

In the following years, several other RNA phages like MS2 (Davis et al., 1961), R17 (Paranchych and Graham, 1962) and fr (Marvin and Hoffmann-Berling, 1963) were isolated that were closely similar to f2, but soon a serologically distinct phage Qβ was also discovered (Overby et al., 1966; Watanabe, 1964). The "male"-specificity of the phages turned out to be determined by the F factor-encoded pili which the phages used as the cellular receptors for adsorbing to bacteria (Figure 1) (Crawford and Gesteland, 1964). Besides the F plasmid, there are a lot of other conjugative plasmids in nature which often encode pili that are very different from those of the F plasmid, and RNA phages specific for such non-F pili like PRR1 (Olsen and Shipley, 1973), t (Bradley et al., 1981), C-1 (Sirgel et al., 1981), Iα (Coetzee et al., 1982), M (Coetzee et al., 1983), D (Coetzee et al., 1985a), pilHα (Coetzee et al., 1985b) and Hgal1 (Nuttall et al., 1987) were later isolated. It was also discovered that not all RNA phages are plasmid-specific, and there are some that infect Gram-negative bacteria by attaching to their genome-encoded pili, like *Pseudomonas* phages 7s (Feary et al., 1964) and PP7 (Bradley, 1966), several phages infecting *Caulobacter* (Schmidt and Stanier, 1965) and the *Acinetobacter*-infecting phage AP205 (Klovins et al., 2002).

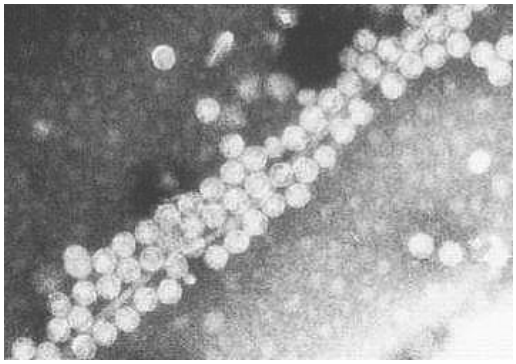

**Figure 1.** MS2 phages attached to an *E.coli* F pilus. Image taken from ICTVdb Picture gallery, © Hans Ackermann.

## 1.1. Genome organization

The small RNA phages have linear, single-stranded, positive-sense genomes approximately 3400 – 4300 nucleotides in length. All of the known ssRNA phages are evolutionary related and are classified into the *Leviviridae* family [from Latin *levis* meaning "light" (not heavy)]. Many of the phages are further divided into two genera, Levivirus and Allolevivirus. Despite considerable sequence variation, all of the known ssRNA phages have a remarkably similar core genome organization with three common genes – maturation, coat and replicase – following each other in the 5' to 3' direction (Figure 2). In addition, phages of the Levivirus genus encode a small lysis protein that

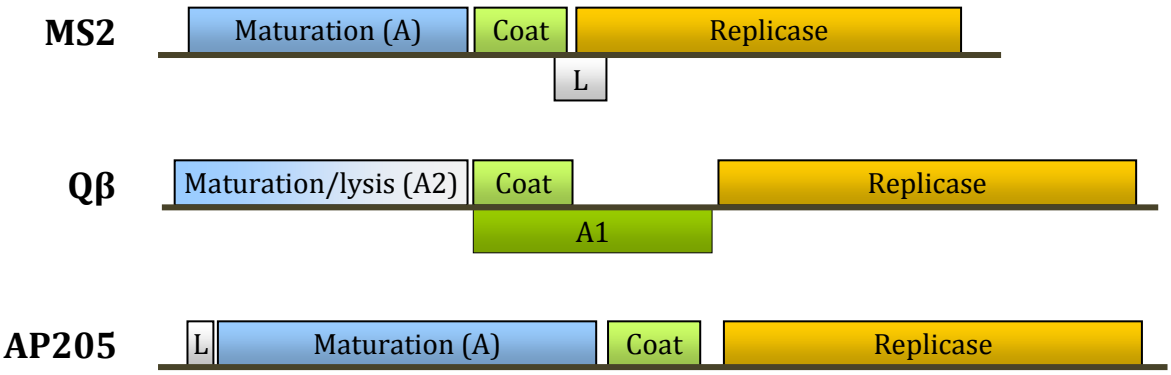


**Figure 2**. Genome organization of the *Leviviridae* phages. Genomes from representative phages MS2 (levivirus), Qβ (allolevivirus) and AP205 (unclassified) are shown. Genes are represented by rectangles; L designates the gene encoding the lysis protein

overlaps with the coat and replicase ORFs in a different reading frame. The Levivirus genus is rather diverse with representatives infecting different conjugative pili-harboring Enterobacteria and also includes the *Pseudomonas* phage PP7, but all of the currently known alloleviviruses are a rather closely related group of F pili-specific *Escherichia coli* phages. The hallmark feature of Allolevivirus phages is that they encode a minor coat protein A1, a C-terminally extended version of the coat protein that is produced by ribosomal read-through of a leaky termination codon of the coat gene (Weiner and Weber, 1971). The other distinct feature of alloleviviruses is that they do not have a separate lysis gene; instead, cell lysis is mediated by a bi-functional maturation protein (Karnik and Billeter, 1983; Winter and Gold, 1983). The more distantly related *Acinetobacter* phage AP205 encodes a lysis protein with similar properties to those of leviviruses, but the lysis gene is located at the 5' end of the genome preceding the maturation gene (Klovins et al., 2002). Therefore, AP205 is not usually recognized as a levivirus but rather as an unclassified *Leviviridae* phage.

## 1.2. Life cycle of the ssRNA phages

The *Leviviridae* virion consists of a single genomic RNA molecule packaged in a small, roughly spherical protein shell that is made up of 180 coat protein molecules and a single copy of the maturation protein. The infection cycle begins when the maturation protein in the virion binds to the shaft of a bacterial pilus. Subsequently, the genomic RNA is released from the capsid as a maturation protein – RNA complex which then enters the bacterial cell through a poorly understood mechanism. In the cytoplasm, the genome serves directly as an mRNA molecule and directs the production of replicase, an RNA-dependent RNA polymerase, as an early product. The replicase synthesizes complementary "minus" strands of the genomic RNA which are then used as templates for producing more "plus" strands. As the number of "plus" strands and, consequently, the amount of coat protein in the cell rapidly grows, coat proteins assemble to form a capsid around a "plus" RNA strand bound to the maturation protein. Lastly, the bacterial cell is lysed and the newly assembled virions are released in the environment.

### 1.2.1. Adsorption, genome ejection and penetration

As the first step of the infectious cycle, all of the known ssRNA phages use some kind of pili to attach to bacterial cells, but the particular pili that they utilize can be very different for distantly related phages. However, almost everything that is known about virion adsorption, genome ejection and its transport into the bacterial cell comes from studies of the closely related phages MS2, f2 and R17, which infect *Escherichia coli* harboring the F plasmid-encoded "F pili". Still, these stages in the phage life cycle are very poorly understood.

Early studies on ssRNA phages quickly revealed that, besides the coat protein and RNA, the virions also contain a minor protein species encoded by the first or "A" cistron of the genome. Phage mutants lacking the "A protein" in their capsids were noninfectious and, in contrast to normal virions, contained smaller amount of RNA (Engelhardt and Zinder, 1964). Since the A protein appeared to be necessary for the production of correctly assembled, "mature" particles, it was also called the "maturation" protein. Further studies showed that the A protein-deficient particles are unable to adsorb to F-pili (Lodish et al., 1965) which suggested that this is the phage attachment protein. The coat protein was convincingly shown to have no role in adsorption, as an *in vitro* reconstituted complex of only RNA and the A protein turned out to be infectious (Shiba and Miyake, 1975).

F pili are fibrillar extracellular structures usually present at up to five copies per cell for the laboratory strains of *E. coli*. The pili measure about 8 nm in diameter, have an approximately 2 nm wide central lumen and can reach several micrometres in length, considerably exceeding the length of the bacterium (Lawley et al., 2004). They are

assembled from subunits of F pilin, a small 7.2 kDa protein with a predicted unstructured N-terminal part and two hydrophobic α helices (Silverman, 1997). Mutational studies have revealed that a region in the N-terminal part and the very C-terminus of the pilin monomers are involved in ssRNA phage binding (Frost and Paranchych, 1988). It is not known which parts of the A protein participate in the interaction, and no high-resolution structures of either the A protein or the F pilus have been determined. After adsorption, the genome is ejected from the capsid as an RNA – A protein complex, leaving empty capsids in the medium. At this point, the genome becomes sensitive to RNAse, although it is protected while inside the virion. The A protein is bound to the RNA at two locations; one in the A protein coding sequence and the other in the 3' untranslated region (Shiba and Suzuki, 1981). During the ejection reaction, the A protein is cleaved in two fragments (Krahn et al., 1972), both of which are transported into the bacterial cell along with the RNA. From this it seems reasonable to assume that there are two RNA binding domains in the A protein and that the cleavage occurs between them, but currently there is no experimental evidence to support this. It is also not clear what is the exact trigger for the cleavage.

The genome ejection only happens if the virions are adsorbed to cell-attached F pili at 37° C (Danziger and Paranchych, 1970). The virions equally well adsorb to cell-bound pili also at 4° C as well as to cell-free F pili, but in these cases, the binding is reversible and the virions remain infectious after desorption (Valentine and Strand, 1965). Therefore, the interaction between the pilus and the A protein alone is apparently insufficient to trigger the genome ejection. The normal function of F pili in conjugation is to sample the surrounding medium for recipient cells, bind to them and then retract which brings the two cells together and enables to form a stable bridge for DNA transfer. The ability to retract is a characteristic feature of F pili and obviously requires that the pili are attached to cells and that the cells are metabolically active. Since the requirements for phage RNA ejection and pilus retraction are similar, there is a reason to believe that the RNA ejection might be linked in some way to the ability of the pili to retract. There is no evidence, however, that binding to a recipient cell or adsorption of a phage particle to the side of a pilus would actively initiate its retraction. On the contrary, there is experimental support that F pili undergo cycles of retraction and extension without an apparent trigger. This was demonstrated in a study where fluorescently labeled phage R17 was incubated together with piliated bacteria and the dynamics of F pili were monitored in real time using confocal fluorescence microscopy (Clarke et al., 2008). The pili could clearly extend despite the phages being attached, effectively ruling out the possibility that phage adsorption causes their retraction. Therefore the most likely scenario appears to be that a virion binds to the side of a pilus and remains there until by an accidental pilus retraction it gets transported to the cell

surface where the cleavage of the A protein is triggered and the RNA ejection takes place.

Virtually nothing is known about the further events in which the ejected phage RNA crosses the bacterial cell envelope, but some of the F plasmid-encoded proteins that make up the conjugation apparatus are likely involved in the process. The F plasmid encodes more than twenty proteins which together create a complex machinery for replicating plasmid DNA and transferring it to recipient cells (Frost et al., 1994). Although the F plasmid has been widely studied and has become a paradigm for the conjugative plasmids in many areas, structural biology is not one of those and little is known about the three-dimensional structure of the F plasmid-encoded proteins and their organization into macromolecular assemblies. In this respect, other plasmids, namely pKM101 and R388, have been much more widely studied. The conjugation machineries that all of these plasmids encode belong to what is known as the type IV secretion (T4S) system, and although there is essentially no sequence identity between proteins encoded by the F, pKM101 and R388 plasmids, the general architecture of all T4S systems is believed to be similar.

The T4S systems are large, several megadalton-scale protein assemblies that cross the entire bacterial cell envelope and contain a pore as a central structural element through which the DNA is transported during conjugation. For the pKM101 plasmid, the core complex of the T4S system is formed by 14 copies of proteins TraF, TraO and TraN at a 1:1:1 stoichiometric ratio which assemble into a double-membrane spanning channel with a prominent chamber in the middle (Fronzes et al., 2009) (Figure 3A). In case of the F plasmid, the respective homologous proteins TraB, TraK and TraV have also been shown to interact and form an envelope-spanning structure using genetic and

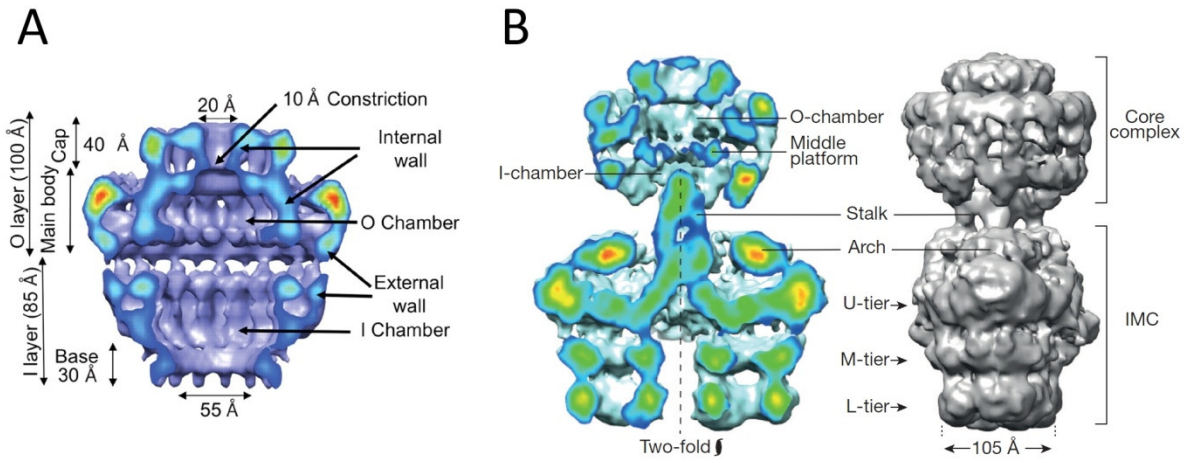


**Figure 3.** Structure of the conjugative pore. A, cryoEM reconstruction of the core complex. A cut-away side view is shown. Image taken from (Fronzes et al., 2009). B, negative-stain EM reconstruction of the complete T4S system. Left, a cut-away front view; right, side view. Image taken from (Low et al., 2014).

biochemical methods (Harris et al., 2001), thus presumably the overall three-dimensional structure of the core complex of the F plasmid is similar. In addition to the core complex, structure of the complete T4S system from plasmid R388 revealed an inner-membrane complex with intriguing features such as two barrel-like structures with several tiers on the cytosplasmic side, connected to an arch and a stalk which inserts into the chamber of the core complex (Low et al., 2014) (Figure 3B). In this case, similarities with the F plasmid-encoded proteins are less clear. From the current T4S system structures it remains unknown how the pilus is incorporated into the system, but other studies have shown that pilin monomers first accumulate in the inner membrane (Paiva et al., 1992) and then polymerize in a helical arrangement (Marvin and Folkhard, 1986) to form a pilus, likely through the pore. When the pilus retracts, the pilin subunits depolymerize in a similar manner, and if an A protein – RNA complex is bound to a pilin molecule it might get dragged to the periplasm along with it.

There is evidence that other proteins that make up the conjugative pore are involved in phage RNA penetration as well. A number of mutants in the pore-forming genes have been isolated and while they confer resistance to ssRNA phage infection to the cell, they are also DNA transfer deficient, which suggests that the formation of the conjugative apparatus is likely disrupted and that no pili are produced as a consequence (Willetts and Achtman, 1972). An exception is the traG gene which encodes a transmembrane protein that localizes in the inner membrane complex of the conjugative pore and has large periplasmic domains (Frost et al., 1994). The TraG protein appears to be bifunctional with the N-terminal part of the protein involved in F pilus assembly and the C-terminal domain in mating pair stabilization (Firth and Skurray, 1992). Mutations throughout the protein are detrimental to DNA transfer, but only those in the N-terminal part also abolish ssRNA phage infectivity (Willetts and Achtman, 1972). An F plasmid with an uncharacterized mutation in the traG gene has also been isolated that produces normal pili and is able to transfer DNA to recipient cells, but provides resistance to phage Qβ infection and reduced sensitivity to phage R17 infection (Frost and Paranchych, 1988).

Another notable exception is the traD gene, mutants of which are DNA transfer-deficient and resistant to phage f2 infection but, interestingly, not to Qβ. The traD mutants produce functional F pili and allow normal adsorption and RNA ejection for the f2 virions, but the infection is aborted during the genome penetration step (Achtman et al., 1971). The TraD protein is not a structural component of the envelope-spanning conjugative pore but is associated with the inner membrane by two N-terminal transmembrane helices with the rest of the protein residing on the cytoplasmic side (Lee et al., 1999). The cytoplasmic domain contains a nucleotide binding site and serves to transport the replicated DNA strand of the F plasmid to the recipient cell (Lanka and

Wilkins, 1995). The three-dimensional structure of the cytoplasmic domain of TrwB, a TraD homolog from the R388 plasmid, revealed that six TrwB monomers assemble in a hexameric ring structure with a channel in the middle that is remarkably similar to the $F_1$-ATPase (Gomis-Ruth et al., 2001). This implies that the TraD protein functions as an ATP-driven pump that translocates the outgoing DNA strand to the periplasm during conjugation. It is tempting to speculate that the A protein – RNA complex might use this channel to enter the cytoplasm, but there is no experimental evidence for this.

Allolevivirus, exemplified by phage Qβ, in addition to the coat and maturation proteins contain a few copies of the minor coat protein species A1 in their capsids (Horiuchi et al., 1971). A1 is a C-terminally prolonged version of the coat protein that is formed when ribosomes occasionally read-through the leaky stop codon of the coat protein (Weiner and Weber, 1971). The A1 protein is required to produce infectious virus particles (Hofstetter et al., 1974), but its specific role in the infection process has remained unknown.

Although leviviruses and alloleviviruses both use F pili as the cellular receptors, the notable differences in the infection process suggest that there is likely no conserved mechanism of how the RNA genome enters the bacterial cell, and clearly much more research needs to be done to a gain a mechanistic understanding for the ssRNA phage attachment, genome ejection and penetration phases of their life cycle.

### 1.2.2. RNA replication

The first intracellular step that needs to happen after the genome has crossed the cell envelope is the replication of phage RNA. The bacterial cells, however, do not have an enzyme that would be capable of synthesizing complementary RNA strands from an RNA template. Therefore, RNA viruses must encode their own enzyme, an RNA-dependent RNA polymerase (RdRp) or "replicase" for making copies of their genome, and the ssRNA phages are no exception. ssRNA phage replicase activity was for the first time detected in extracts from cells infected with bacteriophages f2 and MS2 (August et al., 1963; Haruna et al., 1963), but further purification and biochemical analysis of these enzymes was hampered by their marked instability and rapid inactivation of the preparations. A few years later, an enzyme from bacteriophage Qβ was isolated (Haruna and Spiegelman, 1965a) which turned out to be much more stable and easier to work with, and it quickly became the prototype for studying the ssRNA phage replicases. Since then, virtually everything that is known about RNA replication in *Leviviridae* phages comes from studies of this enzyme, the Qβ replicase.

All of the known ssRNA phages encode an approximately 60-65 kDa polypeptide which provides the enzymatic RNA-dependent RNA polymerase activity; however, this protein alone is not sufficient to replicate phage RNA. The complete RNA replication

complex or holoenzyme contains three other proteins: ribosomal protein S1 (Wahba et al., 1974) and elongation factors EF-Tu and EF-Ts (Blumenthal et al., 1972), which the phage-encoded protein, usually designated as the β subunit of the complex, recruits from the host cell. Ribosomal protein S1 functions as a translational initiation factor and is required for translation of most mRNAs (Sørensen et al., 1998). EF-Tu is a GTP-containing elongation factor that binds an aminoacyl-tRNA molecule and delivers it to the active site of an elongating ribosome, after which the EF-Tu-bound GTP is hydrolyzed to GDP. Elongation factor EF-Ts then recycles EF-Tu by stripping the GDP from it and allowing the regenerated EF-Tu to bind another molecule of GTP and aa-tRNA and enter another elongation cycle (Krab and Parmeggiani, 1998). The S1 protein differs somewhat from the other subunits in that it is required only for the recognition of the Qβ genomic "plus" strand as a template, while a replication complex consisting of the β subunit, EF-Ts and EF-Tu, often referred to as the "core replicase", is sufficient to synthesize "plus" strands from "minus" strands (Kamen et al., 1972).

The Qβ replicase has an impressive processivity, and the amplification process has been compared in a way to a "room-temperature PCR" (Ugarov and Chetverin, 2008), for as long as the enzyme is in a molar excess over the template, the number of product strands grows exponentially (Haruna and Spiegelman, 1965b). In order for an RNA molecule to be amplified, both "plus" and "minus" strands have to serve as good templates for the replicase, and not any RNA molecule fulfills these criteria. The Qβ replicase initiates RNA synthesis *de novo*, i.e., it requires no primer for initiation, but the template must contain a sequence CCA at the very 3' terminus, and, correspondingly, begin with GG at the 5' terminus in order to be amplified (Chetverin and Spirin, 1995). The 3'-terminal adenosine is added during chain termination and does not serve as a template nucleotide for the complementary strand; instead, the RNA synthesis begins at the penultimate nucleotide, C (Weber and Weissmann, 1970). The nucleotide sequence at the ends of the RNA alone is, however, insufficient to make an RNA molecule a good template for the replicase. Some RNAs are much better templates than others, and several *in vitro* selected molecules have been characterized that can be replicated to estimated $10^{10}$ copies in 10 minutes (Chetverina and Chetverin, 1993). Reasons for such efficiency are not exactly clear, but the secondary and tertiary structure of the folded RNA molecule likely plays a major role in this.

Structure of the natural template, the phage genome, certainly plays an essential role in replication. The genome contains large amounts of adjacent self-complementary sequences that fold back to each other to form hairpin structures, often with unpaired nucleotides in them. In total, about 75% of the nucleotides in the genome are involved in base pairing (Skripkin et al., 1990) which demonstrates that the ssRNA phages are not that single-stranded after all. The Qβ replicase cannot initiate RNA synthesis with

double-stranded RNA (Weissmann et al., 1967), and the high degree of secondary structure in the genome effectively prevents the "plus" and "minus" genomic strands from forming a double-stranded RNA duplex during replication. The RNA hairpins in the genome further engage in higher order structures by long-distance base pairing between them which results in a complex three-dimensional shape, therefore the genome presumably is a largely globular structure somewhat similar to a ribosome, although the phage genome is undoubtedly much more dynamic as the higher-order structures need to be temporarily disrupted for replication or translation. The extensive secondary and tertiary structures serve many roles throughout the life cycle of the phage, and recognition of the RNA template for replication is a prime example for this.

Studies on Qβ RNA replication have shown that besides the 5'-GG and CCA-3' sequences, the Qβ genomic "plus" strand is recognized in a remarkably complex manner that involves two internal regions in the genome and some long-distance interactions. First of the internal regions, called the S site, is an approximately 100 nucleotide long U-rich stretch preceding the initiation codon of the coat protein with a rather poorly defined secondary structure (Meyer et al., 1981). The other one, the M site, is a branched stem-loop structure that comprises some 100 nucleotides within the replicase coding region (Schuppli et al., 1998). Both sites are bound simultaneously by the S1 protein, and the S1 protein alone, without the involvement of the other subunits of the replicase complex, is sufficient for binding Qβ RNA (Miranda et al., 1997). For *in vitro* replication, the S site is dispensable but the M site cannot be removed without a dramatic loss of template activity (Schuppli et al., 1998). Two other crucial elements for Qβ "plus" strand recognition are a long-distance interaction that bridges thousand nucleotides within the replicase coding sequence (Klovins et al., 1998) and an RNA pseudoknot structure which connects the loop of the 3'-terminal hairpin with an unpaired region adjacent to the M site (Klovins and Van Duin, 1999). These interactions apparently bring the 3' domain spatially close to the M site and impose a specific three-dimensional structure to the RNA such that the S1 protein-mediated binding of the replicase holoenzyme to the M site positions the 3' terminus in the active site of the enzyme and allows the initiation of RNA synthesis to proceed.

The ssRNA phage replicases are not only very efficient in recognizing and replicating their own genomes, but also in discriminating against other RNAs. Somewhat predictably, Qβ and MS2 replicases do not replicate heterologous RNAs like bacterial ribosomal RNA or the genomes of some plant RNA viruses. More surprisingly, the Qβ replicase is able to replicate the Qβ genomic RNA but not that of phage MS2 and, similarly, the MS2 replicase copies MS2 RNA but ignores the Qβ one (Haruna and Spiegelman, 1965a). Both Qβ and MS2 genomes contain the S and M sites and have the long-range interactions, and the remarkable template selectivity of Qβ and MS2

replicases still remains to be explained. The 3' domains in Qβ and MS2 RNAs have a remarkably different structure (Klovins et al., 2002) that probably confers some of the template specificity, but other differences in RNA tertiary structure likely play a role as well. Determination of the three-dimensional organization of the phage genome at high resolution would be necessary to ultimately resolve this problem, but such studies are extremely challenging due to the presumed high flexibility and heterogeneity of the several thousand nucleotide-long RNA molecules.

### 1.2.3. Control of gene expression

The ssRNA phages have just four genes which makes them the simplest phages and some of the simplest viruses in general. Therefore it might appear that not much of a control of gene expression is going on in these phages like there are, for example, early, middle and late genes in the tailed dsDNA phages. However, the levels and time course of ssRNA phage protein synthesis during infection (Figure 4) clearly show that some regulatory mechanisms must be present. As it turns out, synthesis of all of the phage proteins is regulated in some way and this is achieved by a surprising variety of mechanisms, including initiation codon availability to ribosomes, RNA secondary structure and folding kinetics, translational coupling of one gene to another, ribosomal read-through and specific protein-RNA interactions.

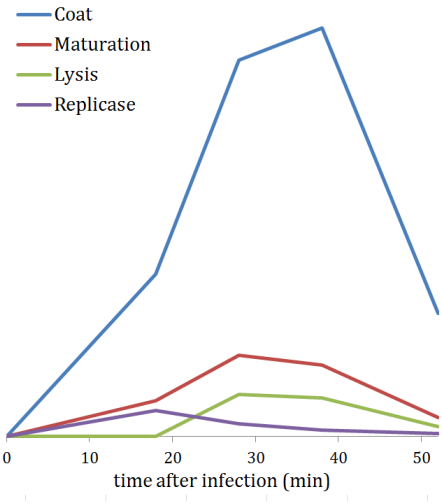


**Figure 4.** Levels (arbitrary units) and time-course of phage protein synthesis during infection of bacteriophage f2. Image adapted from (Beremand and Blumenthal, 1979).

As with the RNA replication, the three-dimensional structure of the genomic RNA is very important for the control of gene expression. In the folded state, only the initiation codon of the coat gene is accessible to ribosomes, while those of the other genes are buried inside the structure and unavailable (Van Duin and Tsareva, 2006). Therefore when a phage genome first arrives in the bacterial cell, the coat protein ORF is the first one that is translated. The translating ribosome on its way disrupts RNA hairpins and breaks long-range base-pairing within the RNA, and as a result the initiation codon of the replicase gene becomes available to ribosomes as well. This in turn allows the production of the β subunit and assembly of the replicase complex for RNA replication. As discussed in the previous section, the S1 protein in the replicase holoenzyme is required to recognize the genomic RNA by binding to the S and M sites. However, the S1 protein is also required for the translation of the coat protein gene as it does not have a

classical Shine-Dalgarno sequence, and the ribosome-bound S1 protein recognizes the same region, the S site, as the one in the replicase complex. This creates a situation where the S1 protein molecules from ribosome and replicase both compete for the same binding site on RNA, and since the coat initiation site is the only one normally available to ribosomes, masking of the S site by the replicase complex effectively prevents the genome from being translated. The repression of translation is particularly important early in the infection as the phage is interested to quickly produce many copies of the genome and not translate a few existing ones. Such mechanism also helps to avoid a situation when a ribosome and replicase would collide on the same RNA strand and halt both translation and RNA replication. Competition for the same template also regulates the amount of the genomic "plus" and "minus" strands that are synthesized. When the Qβ genome is used as a template *in vitro*, the number of the synthesized "plus" and "minus" strands is equal (Kamen, 1975), but in the infected cell, the "plus" strands are in an about tenfold excess over the "minus" strands (Chetverin and Spirin, 1995). In case of the "plus" strands, the replicase must constantly compete with ribosomes that translate the RNA and, later in infection, A protein and coat protein molecules that bind to the genome to encapsidate it. The "minus" strands, on the other hand, are always available for copying, and the increased rate of replication initiation on the "minus" strands results in an excess of the "plus" strands (Kamen, 1975).

Besides its structural role in forming the capsid, the coat protein also acts as a translational repressor that controls the synthesis of the replicase. The replicase is a characteristic early gene product that is required at the beginning of the infection, but later on, when a substantial amount of phage RNA has already been synthesized, there is no need for more replicase to be produced. As the amount of genomic "plus" strands in the cell increases, the quantity of the synthesized coat protein also rapidly grows which in turn causes the translation of the replicase gene to shut down. The control element or the "operator" of the replicase gene is an RNA hairpin at the beginning of the ORF that comprises the initiation codon of the gene (Gralla et al., 1974; Weber, 1976) which the coat protein binds to, thereby masking the initiation codon from ribosomes and preventing translation of the gene. Operators of the studied ssRNA phages have seven to eight base pair-long stems with an unpaired adenosine on the 5' side of the stem and three to six nucleotide long loops. The operator structures and the specific requirements for interaction between the coat protein and RNA can be rather different and have been characterized biochemically for several phages (Lim and Peabody, 2002; Romaniuk et al., 1987; Witherell and Uhlenbeck, 1989) (Figure 5). The specific binding of the coat protein to phage RNA might also mark the genome for encapsidation, as discussed in the next section.

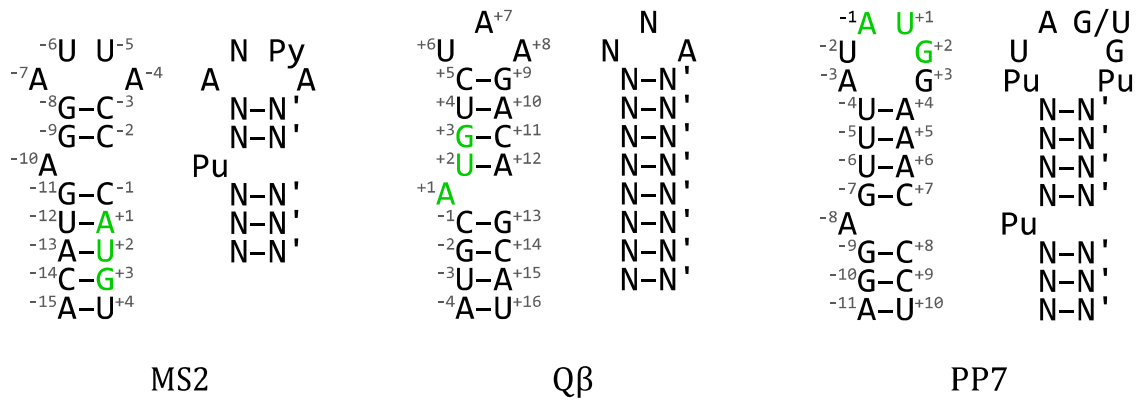


**Figure 5.** Secondary structure of ssRNA phage operators. For each phage, the wild-type operator is shown on the left and the minimal sequence requirements for binding to the coat protein on the right (Py, pyrimidine (C or U), Pu, purine (A or G), N, any nucleotide, N', a nucleotide complementary to N). The initiation codons of the replicase gene are shown in green, the numbering of nucleotides is relative to the first nucleotide of the replicase ORF (+1).

In leviviruses, the lysis gene overlaps with the coat and replicase genes in a different reading frame (Beremand and Blumenthal, 1979). In MS2 genome the initiation codon of the lysis gene is located some 50 nucleotides upstream the termination codon of the coat protein and does not have an SD sequence. Like with the replicase, ribosomes are unable to initiate at the lysis start codon in the folded RNA, and production of the lysis protein requires that the coat protein gene is translated first. To translate the lysis gene, the phage has made use of a property of the ribosomes that after the synthesis of a protein chain has been terminated, the ribosome does not immediately dissociate from the mRNA but for a short while randomly drifts along it in both directions (Adhin and Van Duin, 1990). This way, after the termination codon of the coat protein gene has been reached, there is some chance that the ribosome will slide back and reach the lysis initiation codon before it dissociates from the RNA. In effect, the ribosome reinitiates translation at the lysis start codon some 5% of the time (Van Duin and Tsareva, 2006), and this ensures that the lysis protein accumulates slowly and the cells lyse late in the infection cycle when there has been enough time to complete particle assembly.

In contrast to leviviruses, the allolevivirus encode an additional protein A1 which is a C-terminally prolonged variant of the coat protein (Weiner and Weber, 1971). The A1 protein is essential for infection (Hofstetter et al., 1974) and assembles into capsids along with the normal coat protein, but only a few copies per virion are required. This is achieved by having a "leaky" UGA termination codon at the end of coat protein gene instead of the "strong" UAA codon (Hirsh and Gold, 1971). In about 3% of the cases, the ribosomes do not terminate at the UGA codon but instead insert a tryptophan residue and continue translation until the end of the A1 ORF that actually ends with two tandem termination codons UAG and UAA. The proportion of the A1 protein incorporated into

capsids corresponds to the coat protein – A1 ratio produced in the cell, and thus by employing the ribosomal ability to read-through certain termination codons the phage is able to achieve an optimal amount of the A1 protein in the virion.

Expression of the A protein gene in leviviruses is controlled by RNA folding kinetics of the 5'-untranslated region (UTR) of the genome. In the thermodynamically stable equilibrium state, the 5'-UTR consists of a 5'-terminal hairpin and a four-way junction of three hairpins brought together by a long-range interaction (Groeneveld et al., 1995). The long-range interaction at the base of the structure contains the SD sequence of the A protein gene base-paired to a complementary stretch some 90 nucleotides upstream, and the stable structure prevents ribosomes from initiating A protein synthesis in the folded RNA molecule. However, as the 5' end of a newly synthesized genomic "plus" strand emerges from the replicase enzyme, the equilibrium structure of the 5'-UTR is not formed immediately. Instead, after some 45 nucleotides have been synthesized, a temporary hairpin structure forms that includes the upstream complementary sequence of the long-distance interaction and prevents formation of the four-way junction (Figure 6) (Van Meerten et al., 2001). This gives ribosomes an opportunity to bind to the SD sequence before the 5'-UTR rearranges to form the more stable equilibrium structure. Consequently, initiation of the A protein synthesis occurs only once or a few times per RNA strand. This strategy allows the phage to produce the right amount of the A protein as only a single copy is needed per virion. In alloleviviruses,

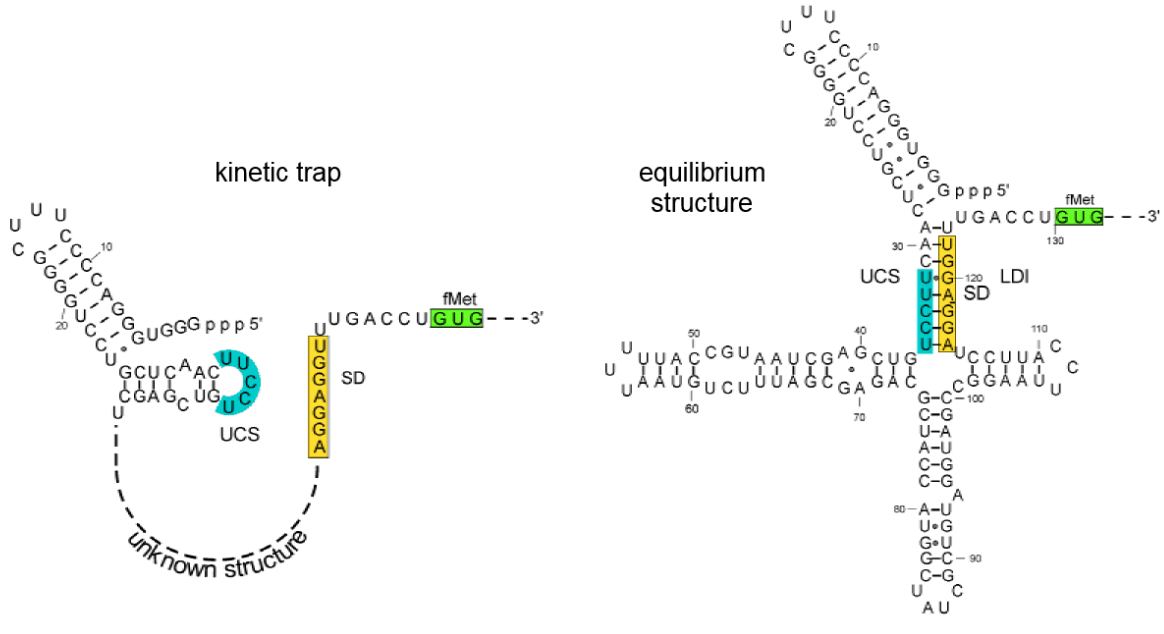


**Figure 6.** Regulation of A protein synthesis in bacteriophage MS2. During the synthesis of new genomic "plus" strands, a temporary hairpin forms that comprises an upstream complementary sequence (UCS, cyan) of the Shine-Dalgarno (SD) sequence (yellow) of the A protein gene, creating a "kinetic trap" for RNA folding that allows ribosomes to initiate translation of the A protein from its initiation codon (green). In a while, RNA rearranges to form the more stable equilibrium structure that prevents the translation of the gene. Figure adapted from (Van Meerten et al., 2001).

RNA folding is also believed to control the maturation or A2 protein synthesis, but in this case the SD sequence is base-paired with a complementary region some 400 nucleotides downstream (Beekwilder et al., 1996). This presumably gives ribosomes more time to initiate synthesis of the A2 protein and leads to a higher quantity of it compared to the A protein in leviviruses. In allolevivirus the A2 protein also mediates cell lysis, which might explain why the phage needs more of it.

### 1.2.4. Assembly of virions

Given the size of phage genomes, the confined spherical space inside capsids certainly puts some restraints on the RNA molecules that can be packaged inside them, and the theme of how important the RNA structure is in the life cycle of ssRNA phages continues to the assembly stage of the new virions as well. To fit inside the particle, the RNA must apparently adopt a globular and roughly spherical shape, and the A protein appears to play an important role in organizing the genome. The A protein binds to two sites in the genome, one in the A protein-coding sequence some 400 nucleotides from the 5' terminus and the other in the 3'-UTR (Shiba and Suzuki, 1981). Mutants lacking the A protein produce particles where some 30% of the genome dangles outside of the capsid and becomes sensitive to RNAse (Argetsinger and Gussin, 1966; Lodish et al., 1965). Thus it appears that in the absence of the A protein the RNA adopts some shape that is unable to fit optimally inside the capsid while binding of the A protein to the genome might bring the two ends of the RNA together and confine the RNA in a packaging-competent state.

When the virions are assembled in the infected cell, cellular RNAs are not packaged into particles. The specificity is quite pronounced, since even when the same cell is co-infected with Qβ and MS2 phages, only authentic progeny virions are formed and no mixed particles are produced (Ling et al., 1970). However, when coat protein genes are cloned and expressed from a plasmid, they encapsidate various cellular RNAs and form virus-like particles morphologically identical to phage virions (Pickett and Peabody, 1993), therefore specific RNA is not required to initiate capsid assembly *per se*. Virus-like particles can be readily assembled also *in vitro* from mixtures of coat protein and a variety of heterologous RNAs such as ribosomal RNA, genomes of some plant RNA viruses and poly(U) (Hohn, 1969). Due to their specific interaction, coat protein bound to the replicase operator hairpin is often assumed to be the nucleation point for assembling the capsid around the genome *in vivo*; however, this might not be the only factor for specific RNA encapsidation. *In vitro*, short RNA oligonucleotides corresponding to the replicase operator indeed induce capsid formation at a slightly lower protein concentration, but the advantageous effect of the operator almost disappears when the length of the RNA increases (Beckett et al., 1988). The role of the

operator is further questioned by the fact that some MS2 pseudorevertants have been isolated that have a defective translational operator that cannot bind the coat protein (Peabody, 1997) or lack the operator entirely (Licis et al., 2000) but which are nevertheless capable to form normal virions.

Although there is no strong experimental evidence that the maturation protein specifically interacts with the coat protein, such possibility seems rather obvious since both are structural components of the capsid, and thus the A protein might also play a role in the assembly of phage virions. There is evidence that in the infected cells, association of the A protein, coat protein and the genome is a strictly sequential process where binding of the A protein to the RNA is an early event preceding the assembly of coat protein molecules around the RNA (Kaerner, 1970). Therefore when coat protein molecules are forming a protein shell around the A protein-RNA complex, the A protein must be accommodated at the surface of the capsid at some point, and specific protein-protein interactions seem to be the most obvious way of doing this.

Some evidence that the A protein contributes to the specificity of RNA packaging comes from *in vitro* reassembly studies with phages Qβ and MS2. When MS2 and Qβ total capsid protein preparations were reassembled together with either of the phage RNAs, some specificity for each phage protein to its cognate RNA was observed (Ling et al., 1969). However, it was known that during prolonged storage, the maturation protein gets inactivated in the protein preparations (Hung and Overby, 1969), and when such "aged" protein preparations were used for the reassemblies, the species specificity was lost. In conclusion, there probably is not a single determinant for the specific encapsidation of phage RNA, but a combination of factors like an optimal three-dimensional shape of the folded RNA molecule, binding of the coat protein to replicase operator and interactions involving the A protein might all contribute to the successful assembly of infectious virions.

### 1.2.5. Lysis

When the new virions have been assembled, they face the last challenge in their life cycle in the cell – to get out of it. Bacterial cells have a rigid envelope with a thick cell wall made of peptidoglycan or murein, a three-dimensional mesh-like structure of long sugar chains cross-linked by short oligopeptides (Vollmer et al., 2008a), and all lytic phages have to find a way to break the peptidoglycan barrier to escape the cell. dsDNA phages encode multicomponent lysis systems that consist of a holin that makes holes in the inner membrane, an endolysin that enzymatically degrades the murein layer and a spanin that disrupts the outer membrane (Young, 2013). Other phages like the small ssDNA phage phiX174 encode a single lysis protein that inhibits a protein in the peptidoglycan biosynthesis pathway (Bernhardt et al., 2000). This way, when the

bacterial cell grows and starts to divide, there is no new peptidoglycan being synthesized which destroys the integrity of the cell wall and leads to lysis.

ssRNA phages of the levivirus and allolevivirus genera use remarkably distinct mechanisms for lysing the cells. The levivirus-type lysis proteins are some 35-75 residue long polypeptides that vary greatly in sequence and their only unifying feature appears to be a hydrophobic transmembrane helix within the protein. Although MS2 and the closely related phages have a rather long unstructured region N-terminal to the helix with an increased prevalence of positively charged amino acids, the first 40 amino acids of the MS2 lysis protein can be removed without affecting lysis activity (Berkhout et al., 1985), therefore the transmembrane helix alone appears to be sufficient for cell lysis. The mechanism by which the lysis is achieved is not entirely clear, but it has been shown that *in vitro* a synthetic peptide corresponding to the 25 C-terminal amino acids of the MS2 lysis protein is able to dissipate the proton motive force in *Escherichia coli* membrane vesicles by generating hydrophilic pores in them (Goessens et al., 1988). In bacteria, disruption of the proton motive force across the membrane is associated with the activation of autolysins, enzymes that normally catalyze a highly regulated process of breaking up the peptidoglycan in small pieces to allow the cells to grow and divide (Vollmer et al., 2008b). With their small lysis protein, the leviviruses appear to have evolved a very simple, yet effective way to activate the autolysins in an unregulated way that leads to uncontrolled degradation of the cell wall and resulting cell lysis.

After the lysis gene of leviviruses had been identified, it turned out that the alloleviviruses do not encode theirs in a similar way, but, somewhat unexpectedly, cloning and expression of the A2 gene resulted in cell lysis (Karnik and Billeter, 1983; Winter and Gold, 1983). No smaller fragment of the A2 coding sequence was able to achieve this, which led to a conclusion that there are no overlapping lysis genes in the A2 ORF and that the entire A2 protein, in addition to its role as a maturation protein, mediates the lysis. It was later discovered that the A2 protein inhibits MurA, an enzyme that catalyzes the first step in peptidoglycan biosynthesis (Bernhardt et al., 2001). Further studies revealed that A2 binds to a “closed” conformation of MurA with its bound substrate, uridine diphosphate-N-acetylglucosamine (UDP-NAG) and that MurA mutants that are resistant to A2 have the mutated residues on the surface of the protein in proximity to the catalytic loop (Reed et al., 2012). This suggests that the A2 protein inactivates MurA by blocking its active site and achieves cell lysis using a similar strategy as the phiX174 phage by inhibiting the murein biosynthesis pathway.

## 1.3. Structural studies of ssRNA phages

### 1.3.1. Capsids

From a structural point of view, ssRNA phage coat proteins and capsids are very well characterized. High-resolution capsid structures of phages MS2 (Golmohammadi et al., 1993; Valegård et al., 1990), fr (Liljas et al., 1994), Qβ (Golmohammadi et al., 1996), GA (Tars et al., 1997), PP7 (Tars et al., 2000), PRR1 (Persson et al., 2008) and φCb5 (Plevka et al., 2009) have been determined, and although the coat protein sequences of these phages are often very different, the three-dimensional structure of coat proteins and capsids is remarkably similar. The *Leviviridae* coat proteins adopt a fold not observed in any other virus family with an N-terminal β-hairpin, five-stranded antiparallel β-sheet and two C-terminal α-helices (Figure 7). Two coat protein molecules form a very stable dimer with a single continuous ten-stranded β sheet on one side of the dimer and the N-terminal β-hairpins and α-helices interlocked with each other on the opposite side. The assembled capsid is roughly spherical with a diameter of about 280-300 Å and consists of 90 coat protein dimers. The capsid has icosahedral symmetry where the subunits follow a quasi-equivalent arrangement (Caspar and Klug, 1962). The quasi-equivalence principle presents a way where an icosahedral structure can be built using certain multiples of 60 subunits (60*T*, where *T* is called the triangulation number), but the subunits, although chemically identical, must adopt slightly different conformations to form the particle. The subunit arrangement in ssRNA phage capsids is described by triangulation number *T*=3, and the protein monomers adopt three different conformations, denoted A, B and C. In some phages like MS2 and Qβ, there are prominent differences in conformations of the loops connecting β strands F and G (the FG loops) in the different conformers, while in other phages like PRR1, PP7 and φCb5, the loops have a nearly identical structure. There are two types of dimers in the capsid, one where the monomers are in conformations A and B (called an AB dimer, Figure 8A) and the other where both monomers are in a C conformation (a CC dimer, Figure 8B). In the capsid, the AB dimers form pentamers around fivefold symmetry axes which are interconnected with CC dimers around twofold axes (Figure 8C). The assembled capsids are generally very rigid, but in some phages like Qβ and PP7, coat proteins form inter-subunit disulfide

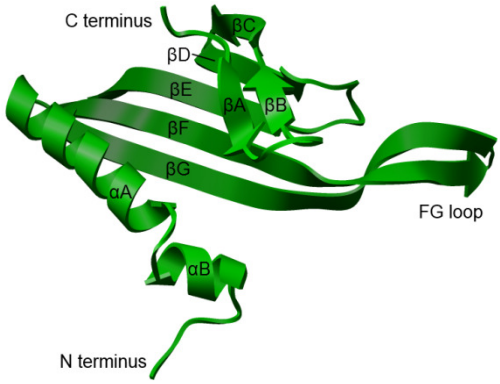


**Figure 7.** Three-dimensional structure of a levivirus coat protein monomer.

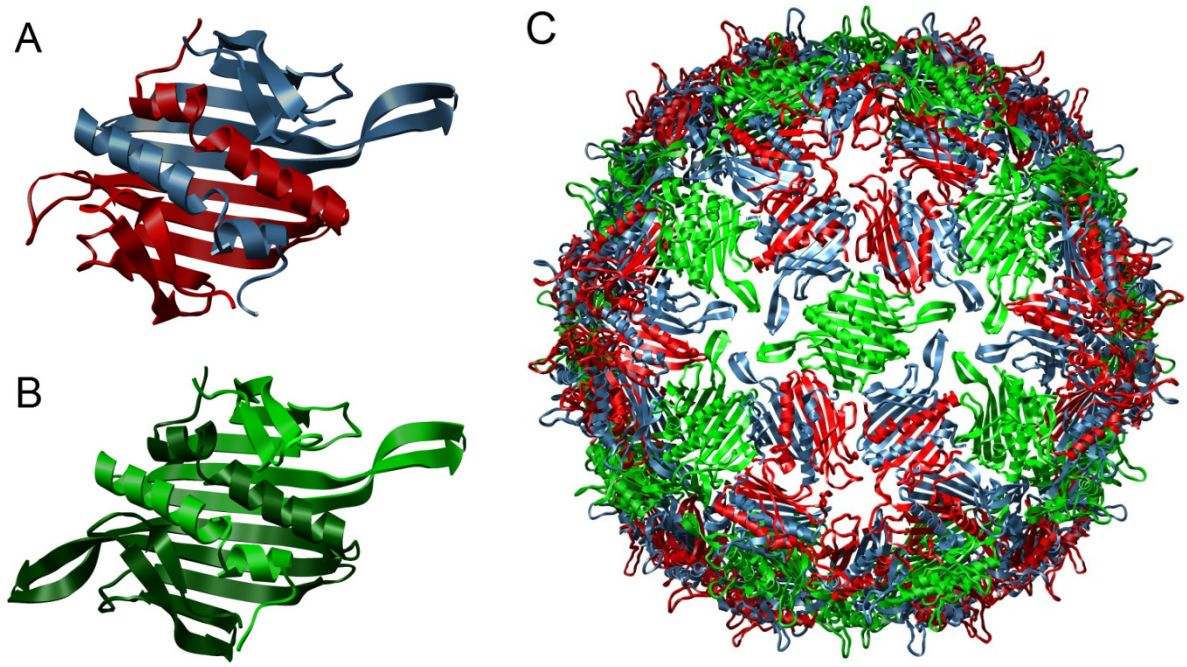


**Figure 8.** Structure of the levivirus capsid. 60 coat protein dimers in an AB conformation (panel A) and 30 in a CC conformation (panel B) assemble in an icosahedral *T*=3 capsid (panel C) with a quasi-equivalent symmetry. The back of the capsid is partly removed for clarity. Subunits in the A conformation are shown in blue, B in red and C in green.

bonds that make them even more robust, and some other phages like PRR1 and φCb5 utilize divalent ions to stabilize their capsids.

Despite the extensive structural studies, molecular details of the capsid assembly pathway are still unclear. Without RNA, the coat protein exists in a dimeric form, but once RNA is present, particles are formed very rapidly, and no assembly intermediates have been isolated (Stockley et al., 1994). The RNA appears to be necessary only during the assembly stage, but does not have a structural role after the particles have been completed as it can be removed from inside of the capsids without impairing their stability (Hooker et al., 2004). Some studies suggest that in solution isolated coat protein dimers have a CC-like conformation while those bound to an RNA hairpin adopt an AB-like conformation (Stockley et al., 2007). Therefore the RNA appears to act as a switch for attaining the different quasi-equivalent conformations necessary for capsid formation. In addition, long RNA molecules can simultaneously bind many coat protein dimers and bring them together which would further facilitate the formation of capsids.

### 1.3.2. CryoEM studies of phage RNA and A protein

Several cryoEM studies on ssRNA phages have been carried out to visualize the RNA genome inside the capsid. The first such study clearly showed density below the known RNA binding site of coat protein dimers, but at a lower contour level additional features emerged that showed a continuous RNA network lining the inside of the particle with triangles of density around the threefold axes and pentagons around the fivefold axes

(Koning et al., 2003). A follow-up study confirmed similar RNA arrangement in different phages, although in the distantly related phage AP205 the RNA appeared to interact more loosely with the capsid (Van den Worm et al., 2006). Later a higher-resolution structure from a different group showed that the genome inside the capsid is organized predominantly in two concentric shells, one immediately below the capsid surface and another some 55 Å from the center of the particle, both connected along the fivefold axes (Toropova et al., 2008).

Recently, a cryoEM study of phage MS2 bound to an F pilus was undertaken in an effort to shed more light on the interaction (Dent et al., 2013). In contrast to the prior structures, this study did not use icosahedral averaging of the electron density. While averaging considerably improves the signal-to-noise ratio of icosahedrally symmetric features, it also averages out features that do not obey such symmetry. The phage genome is intrinsically asymmetric and the A protein is present only in a single copy per virion, therefore to gain structural information about them, no averaging can be used. The structure revealed a continuous stretch of density from the pilus extending towards the center of the capsid, but the low resolution (about 40 Å) of the structure did not allow to identify the boundaries between the pilus, the A protein and the genome. Still, it let the authors to suggest with some confidence that the maturation protein replaces a coat protein dimer in the particle. Such interpretation implies that separation of the A protein from the virion upon RNA ejection leaves a hole in the capsid that serves as an exit route for the genome. Although such model seems attractive, some additional evidence would be necessary to confirm this idea.

### 1.3.3. Coat protein – RNA interaction

The interaction between the coat protein of bacteriophage MS2 and RNA has been extensively studied and has become one of the best structurally characterized protein-RNA interactions. To date, more than 25 structures of complexes between different variants of MS2 coat protein and RNA hairpins have been determined, and, for a large part, these studies were made possible because of the ability to obtain the protein-RNA complexes in pre-crystallized MS2 capsids. The MS2 capsids have prominent pores around the threefold and fivefold symmetry axes, and these allowed diffusion of small RNA hairpins into the particles. Once MS2 capsid crystals had been prepared using well-established crystallization conditions, it was then easy to investigate many different RNAs by simply adding an RNA solution to the crystals. Another factor crucial for determining the many high-resolution structures was that the RNA hairpins bound asymmetrically to coat protein dimers in the AB conformation. The coat protein dimers and the RNA binding site have a twofold rotational symmetry and therefore the RNA can bind to the protein in two possible orientations. The RNA hairpins actually bind in both

orientations to CC dimers, resulting in two overlapping structures in the electron density map, but due to steric restraints imposed by bending of the FG loops near fivefold axes, only in a single orientation to AB dimers.

The RNA binding surface of the MS2 coat protein is located on the ten-stranded antiparallel β sheet that in the assembled particles faces the interior of the capsids. The structure of the MS2 coat protein – operator complex revealed that the operator adopts a crescent-like shape and three out of the four nucleotides that were known to be important for binding, A-10, U-5 and A-4, directly interact with the protein (Figure 9A, B) (Valegård et al., 1994). A key element for the protein – RNA interaction is binding of two adenine bases, the unpaired A-10 in the stem and the A-4 in the hairpin loop, to two symmetrical adenine-recognizing pockets in the coat protein dimer. The interaction is stabilized by aromatic stacking that continues from the helical stem to the A-7 and U-5 bases in the hairpin loop and a conserved tyrosine residue. In addition, the U-5 base makes contact with an asparagine side chain in the coat protein. The sugar-phosphate backbone of the operator also makes extensive sequence-unspecific contacts with the protein in the stretch between the A-10 and A-4 adenines.

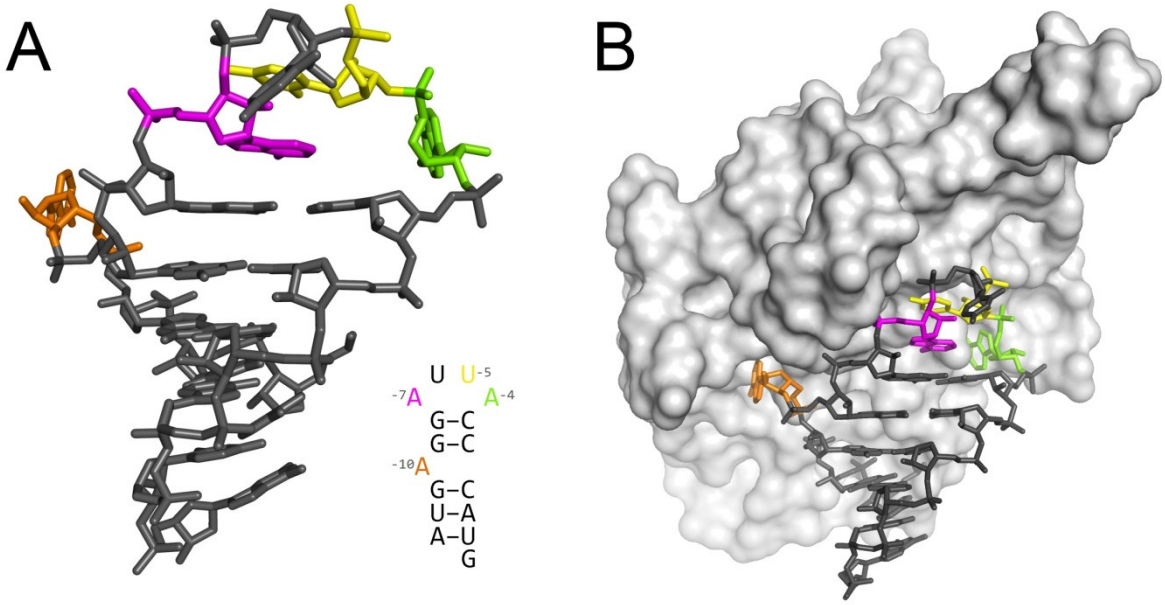


**Figure 9.** Three-dimensional structure of the coat protein – operator complex from bacteriophage MS2. A, structure of the operator. Nucleotides important for binding are showed in color. B, Structure of a coat protein dimer in an AB conformation bound to the operator hairpin. The nucleotides are colored as in A.

Using variants of the operator, it was soon determined that a small RNA stem-loop consisting of only eight nucleotides from the upper part of the hairpin is still able to bind to the protein but a loopless seven base-pair long stem with the bulged adenosine is not (Grahn et al., 1999). Another study examined the effect of amino acid substitutions in the adenine-binding pockets to operator binding and found that despite decreased affinity *in vivo*, the mutant coat proteins still bind the RNA the same way as

the wild-type (Van den Worm et al., 1998). The effects of base substitutions at the -5, -7 and -10 positions have also been extensively investigated. Substitution of the wild-type uracil base at the -5 position with a cytosine results in an operator variant with an about 50 times higher affinity for the protein (Lowary and Uhlenbeck, 1987). Structure of the MS2 coat protein – C-5 operator complex revealed that the cytosine base forms an intramolecular hydrogen bond that stabilizes the RNA in its bound conformation as well as an additional water-mediated contact with the protein (Valegård et al., 1997). At the -5 position, several different bases could be tolerated (Grahn et al., 2001) with one notable exception where substitution of the wild-type uracil with an unnatural base pyridin-4-one caused dramatic conformational rearrangements in the loop resulting in the modified -5 base facing away from the protein and the U-6 making the stacking interaction with the tyrosine side chain instead (Grahn et al., 2000). At the -10 position, substitution of the bulged adenine base with a guanine or cytosine still retained RNA binding and the -7 adenine could be replaced with a cytosine without apparent alterations of the structure (Helgstrand et al., 2002). Hence, although many of the operator variants displayed severely reduced affinity for the coat protein in biochemical RNA binding assays, the structural studies were often unable to find apparent alterations in the binding interactions except for some minor adjustments in protein side chain and RNA backbone orientation. This led to a conclusion that the conditions where highly concentrated RNA solutions are soaked into capsid crystals for a prolonged period of time permit binding of low-affinity RNAs to the protein while in solution where the coat protein is in small concentration and in vast excess over RNA, each affinity-reducing alteration in the RNA has a much more pronounced effect. Still, these studies clearly demonstrated that the loop is more important for binding to the protein than the bulged adenine, and that the continuous aromatic stacking from the RNA stem to the tyrosine chain is much more important for the stability of the complex than the identity of the bases making up the stack.

Another intensive area of research with the MS2 coat protein were structural studies on how the protein can bind several operator-like RNA hairpins called aptamers that were isolated using *in vitro* selection techniques (Hirao et al., 1999). In one study, structure of an aptamer F5 containing a non-Watson-Crick G-A base pair was determined that showed that the bases adopt a head-to-head orientation that does not disrupt the helical stem (Rowsell et al., 1998). Substitution of the unpaired A-10 in the stem of the F5 aptamer with an unnatural base 2'-deoxy-2-aminopurine led to an RNA hairpin (F5/2AP10) with a 65-fold higher affinity to the coat protein than the wild-type operator (Parrott et al., 2000). Structure of the MS2 coat protein – F5/2AP10 complex revealed that an extra hydrogen bond is formed between the modified -10 base and the protein that explained the increased affinity compared to the parental F5 aptamer

(Horn et al., 2004). In another study, an aptamer called F6 was investigated that had a three-nucleotide loop and a bulged adenosine three nucleotides prior to the loop (Convery et al., 1998). Although the secondary structure of the aptamer appeared to be rather different from the wild-type operator, the distance between the adenine at the last position in the loop and the unpaired one in the stem was the same in both hairpins, and the structure of the F6 – MS2 coat protein complex revealed that interactions between the protein and RNA are similar to those in the wild-type operator. Finally, an MS2 coat protein mutant able to bind the operator of phage Qβ with high affinity was investigated (Horn et al., 2006). Like the F6 aptamer, the Qβ operator has a three-nucleotide loop, and the RNA binding in both cases is quite similar, although the unpaired adenosine in the Qβ operator is further away from the loop and thus unable to fit in the respective adenine-binding pocket of the MS2 coat protein.

Compared to MS2, the coat protein – RNA interactions in other ssRNA phages have received much less attention from structural biologists. Recently, the structure of bacteriophage PRR1 coat protein was solved in complex with its cognate replicase operator (Persson et al., 2013). Like the MS2 hairpin, the PRR1 operator has an unpaired adenine two base pairs prior to the loop and another one at the last position of the loop; however, the loop is five nucleotides long instead of four. The adenine-binding pockets of the PRR1 coat protein are very similar to those of MS2, and interactions involving A-11 and A-4 of the PRR1 operator are equivalent to those of A-10 and A-4 of MS2, respectively. In contrast to MS2, the RNA in PRR1 capsids binds in two orientations to both the AB and CC dimers, resulting in overlapping electron density that complicated its analysis. The aromatic stacking extending from the helical stem to a tyrosine residue in the protein was clearly present, although it was hard to interpret the exact conformation of the hairpin loops.

The three-dimensional structure of the PP7 coat protein – operator complex revealed a very distinct RNA recognition mode (Chao et al., 2008). The operator of phage PP7 is quite different from those of MS2 and PRR1 with a six-nucleotide loop and a bulged adenosine four base pairs from the loop (Figure 5), and biochemical studies have shown that many of the loop nucleotides have to be sequence-specific for optimal binding (Lim and Peabody, 2002). Like in MS2 and PRR1, adenines in the bulge and the loop bind to symmetric adenine-recognizing pockets, but the pockets are very distinct and are located at a completely different place than in MS2 and PRR1. In the PP7 operator, in total four nucleotides in the bulge and the loop are involved in sequence-specific contacts with the protein, and three of the six bases in the hairpin loop form an aromatic stack in the RNA stem that further makes a van der Waals interaction to a valine residue in the protein.

### 1.3.4. Replicase

For a long time, almost everything that was known about the Qβ replicase came from biochemical studies of the enzyme *in vitro*. The situation changed significantly in 2010 when the core Qβ replicase was crystallized and its high-resolution structure was determined (Kidmose et al., 2010; Takeshita and Tomita, 2010). More structures soon followed that captured the enzyme in the initiation, elongation and termination phases of RNA synthesis (Takeshita and Tomita, 2012; Takeshita et al., 2012). Very recently, a structure of the Qβ holoenzyme containing a truncated version of the S1 protein was also determined (Takeshita et al., 2014). Together, these structures provide important mechanistic insights about how the replicase works at a molecular level.

The overall architecture of the catalytic β subunit of Qβ replicase resembles other RdRps with the canonical right-handed palm, thumb and finger domains (Ng et al., 2008). The active center contains a conserved YGDD amino acid motif where the aspartates coordinate two magnesium ions that catalyze the polymerization reaction. The overall structure of the core replicase resembles a boat where the catalytic center on the palm domain is facing towards the inside of the boat (Takeshita and Tomita, 2010). The β subunit makes extensive interactions with EF-Tu and EF-Ts which in the replicase complex are assembled in the same way as in the natural EF-Tu:EF-Ts binary complex when GDP is displaced from EF-Tu (Kawashima et al., 1996). The β subunit contains entrance channels for template RNA and NTPs, while the template exit channel is formed by both the β subunit and EF-Tu. From the current data, it appears that there are no other functions for EF-Tu and EF-Ts in RNA synthesis, and their main role seems to be the stabilization of the β subunit in its active conformation (Tomita, 2014).

To study the mechanism of the *de novo* initiation of Qβ RNA synthesis, the core replicase was co-crystallized with an RNA oligonucleotide ending with CCA-3' and a GTP analog, dGTP (Takeshita and Tomita, 2012). In addition to some hydrogen bonds between the protein and RNA and the expected coordinated divalent ions, the structure revealed extensive stacking interactions between the two 3'-terminal cytosine bases of the template, their complementary dGTPs and the 3'-terminal adenosine of the template RNA. The structure showed that the 3'-terminal adenosine is involved in contiguous stacking interactions that are required for the formation of a stable initiation complex. There are interesting similarities with a replicase from a dsRNA phage phi6 which also initiates RNA synthesis *de novo*, but instead employs a tyrosine side chain for maintaining similar stacking interactions with the template RNA at an equivalent position to the 3'-nontemplated adenosine in the Qβ replicase (Butcher et al., 2001). The Qβ and phi6 initiation complexes are closely similar when their three-dimensional structures are superimposed which offers some insight into the evolution and common ancestry of the ssRNA and dsRNA phages.

Like all nucleic acid polymerases, the Qβ replicase must necessarily form an RNA duplex during replication by base-pairing the template with the product. However, the template and the product are released single-stranded after the replication, and the mechanism of how the replicase achieves this had remained mysterious for a long time. To address this problem, a whole range of structures were solved with the replicase together with a template RNA oligonucleotide and 7, 8, 9, 10 and 14 nucleotide long products (Takeshita and Tomita, 2012). The structures revealed that after reaching eight base pairs in length, the RNA duplex collides with the C-terminal part of the β subunit which acts like a wedge to destabilize the helix and guide the template strand through its exit channel while the product strand is free to leave the complex from the open side. As a result, the template and product strands are effectively released each at the opposite side of the complex which provides enough separation and time for local secondary structures to form and prevent the two strands from reforming an RNA duplex. The structural studies have also revealed how the non-templated 3' adenosine is added to the product strand during termination (Takeshita et al., 2012). When the elongating replicase reaches the end of the template RNA strand, an active site-proximal region of the β subunit undergoes a slight conformational change that results in a pocket being formed by the β subunit from one side and the 5'-terminal guanine residue of the template RNA from the other side. The pocket serves to accommodate the adenine base of an ATP molecule which fits much better in the pocket than the bases of other NTPs. The RNA polymerization is then completed by catalyzing the addition of the adenosine to the product strand after which it is released from the enzyme.

Finally, a structure of the Qβ replicase holoenzyme with a shortened version of the S1 protein containing the first three of its six OB-fold domains (Takeshita et al., 2014) showed that the S1 protein binds to the β subunit using the two N-terminal domains while the third domain, which is involved in Qβ RNA binding, is free to rotate near the surface of the β subunit. Although the structure does not provide immediate answers about how the Qβ genomic "plus" strand might be recognized by the replicase complex, the current pace of the structural studies leads to believe that this moment is not too far off.

# 2. METHODS FOR STUDYING RNA PHAGES

To study the structure of ssRNA phage proteins and genomes, I have used a wide variety of different techniques, ranging from microbiology and recombinant DNA technology to bioinformatic sequence analysis and x-ray crystallography. In this chapter, I briefly outline these methods and how they relate to my work, while, for the most part, some of the more specific details like primer sequences and compositions of buffers can be found in the Materials and Methods sections of the respective papers.

## 2.1. Preparation of recombinant proteins for structural studies

The first x-ray structure of a protein was determined in the late 1950s (Kendrew et al., 1958), and for the next three decades some three hundred more were solved. In the late 1980s, however, the number of structures started to grow rapidly, and by 2015 more than 100,000 structures have been deposited in the Protein Data Bank. The explosion in the number of solved structures was in part because synchrotrons, facilities for generating high-intensity x-ray beams designed specifically for protein crystallography, became generally available, but just as important was the advent of recombinant DNA technology that allowed production of recombinant proteins for the structural studies. If in the early days, crystallographers had to work with proteins that could be easily obtained in large quantities from natural sources like hemoglobin from red blood cells, lysozyme from eggs or ribonuclease from bovine pancreas. Techniques like polymerase chain reaction (PCR) and cloning of genes in bacterial expression vectors allowed to produce large amounts of essentially any protein of interest regardless their source organism and abundance in the original cells and opened whole new horizons for structural biology.

In a structural biology project, the first task usually is to secure a steady supply of the target protein that often requires cloning of the protein-coding gene and establishing a high-level expression system for it. Traditionally, the cloning procedure involves amplification of the gene encoding the protein of interest with primers containing restriction enzyme recognition sites at their 5’ termini, after which the PCR product and the target cloning vector are both digested with the respective enzymes, ligated together and the construct is introduced into the target organism for protein production. The cloning vectors are usually medium- to high-copy bacterial plasmids containing a strong promoter upstream the cloning site for transcribing the insert. One of the most popular is the pET system that contains a very strong promoter from bacteriophage T7 (Rosenberg et al., 1987). For protein production, the plasmid construct is introduced into an *E.coli* strain expressing the T7 polymerase, and after

induction of the promoter usually very high yields, often up to 50% of the total cellular protein, can be achieved.

For structural studies, highly pure and homogeneous proteins are required since any contaminants might unspecifically bind to the target protein and interfere with crystallization. The recombinant DNA technology also greatly helps in protein purification by allowing a straightforward method for introducing affinity tags in proteins. This is usually done either by PCR with primers that append the coding sequence of the tag to the ORF or by in-frame cloning in expression vectors that already have the tags. The simplest and most widely used is the 6xHis tag, a stretch of six histidine residues in a row added to either the N- or C-terminus of the protein. The first step of the purification protocol then involves fractionation of the bacterial cell lysate on a column with immobilized $Ni^{2+}$ ions that strongly bind the 6xHis tag, and after elution from the column the protein is already at least 90% pure. Subsequent ion exchange chromatography and gel filtration is usually sufficient to remove the remaining contaminants and provide highly purified crystallization-grade protein.

Production and purification of the Qβ A1 extension followed a standard protocol and was rather simple. The sequence encoding the read-through domain of Qβ A1 protein was PCR-amplified from a plasmid containing the full-length A1 gene using a primer that introduced an initiation codon and a 6xHis tag at the N-terminal part of the A1 extension. The PCR fragment was cloned in a protein expression vector under control of the araBAD promoter which allows arabinose-induced control of protein production. The read-through domain could be readily purified on a $Ni^{2+}$ column and almost all remaining contaminants removed using ion-exchange chromatography on a Q Sepharose column. Gel filtration on a Superdex 200 column was used as the last "polishing" step after which the protein was concentrated using centrifugal spin filters and was ready for crystallization.

Construction and purification of the assembly-deficient Qβ coat protein mutant required a little bit more effort, but the procedure was still fairly straightforward. The starting point was the coding sequence of a Qβ coat protein without cysteines that had been cloned in some plasmid and was kindly provided by Dr. Indulis Cielēns. The background and rationale for construction of the particular assembly-deficient mutant is described in more detail in section 3.2.1. One of the amino acid substitutions, Asn129Arg, was very close to the C-terminus of the protein, and could be introduced by PCR with a reverse primer containing the mutated sequence. The fragment was cloned into a vector that contains the strong T7 promoter, and the construct was used as a template for introducing the second mutation, Pro42Arg. To achieve this, a different type of PCR was used with two perfectly complementary oligonucleotides as primers that contained the desired mutation in the middle of the sequence. In the PCR annealing

step, the mutant oligonucleotides hybridize to both template strands and in the elongation step direct synthesis of DNA around the entire circular template until the polymerase bumps into the 5' end of the primer. After the denaturation step, both complementary product strands can anneal in a double-stranded "fragment" with the length of the entire plasmid and single-stranded ends containing the self-complementary primer sequences. This allows the fragment to circularize and form a mutant "plasmid" with single-strand nicks in both strands. After PCR, the reaction mix is digested with methylated DNA-specific restriction enzyme DpnI which cuts the template plasmid that has been methylated in bacterial cells but not the newly synthesized PCR product that has no methyl groups. The mix is then directly transformed in bacteria that repair the nicks and replicate the newly created plasmid as usual. The method is fast and efficient, and the vast majority of clones usually contain the mutated plasmid.

For some reason, addition of a 6xHis tag negatively affected the solubility of the Qβ coat protein, therefore it had to be purified without the use of affinity chromatography. Still, it was possible to obtain highly purified preparations of the protein, mainly due of the fact that the mutant protein had an isoelectric point of 9.75, distinctly higher than most *E.coli* proteins. For this reason, ion-exchange chromatography was the method of choice, and a single-step purification of the lysate on an SP Sepharose column already yielded a remarkably pure preparation. Further purification on a high-resolution MonoS column removed the remaining contaminants and a final fractionation on a Superdex 200 gel filtration column was used for desalting and buffer exchange, after which the protein was concentrated, mixed with RNA and subjected to crystallization.

## 2.2. Protein x-ray crystallography

X-ray crystallography is the most powerful method for determining high-resolution structures of biological macromolecules, and the vast majority of the published structures have been determined using this method. In this work, I have used x-ray crystallography to solve the structures of the A1 protein and the coat protein – operator complex from bacteriophage Qβ.

When a wave of electromagnetic (EM) radiation meets an object in its path with a size comparable to its wavelength, the wave gets *scattered*, i.e., it changes its direction. In molecules, atoms are a few ångströms (Å, $10^{-10}$ m) apart, and if the wavelength of an incident EM wave falls within the x-ray range (0.5 – 2 Å), it gets scattered by the electrons in the molecule. The combined scattering from all of the electrons in a molecule results in a complex interference pattern that is dependent on the relative position of its constituent atoms and hence contains information about the three-dimensional structure of the molecule. However, scattering from a single molecule is so weak that currently there exist no instruments capable of measuring it. Therefore the

information about the position of the atoms in a molecule has to be extracted in some other, indirect way by combining signal from many identical molecules. If the molecules are in the same orientation and arranged in a repeating, symmetrical way, in certain directions the scattered x-rays combine and amplify each other, resulting in measurable *diffracted* waves. The directions and intensities of the diffracted x-rays contain information about the distribution of *electron density* in the molecules that can be used to infer the positions of their constituent atoms.

An ordered, periodical three-dimensional arrangement of atoms, molecules or ions is called a *crystal*, and because crystals diffract x-rays, they can be used to determine the spatial organization of their constituent entities. A crystal is described by its *unit cell*, the smallest element or a "box" from which an arbitrarily large crystal can be constructed by translation operations only, i.e., by stacking the cells next to each other leaving no gaps in-between. There can be one or several molecules in the unit cell, and these can be related by further symmetry elements like rotation or screw axes. The particular symmetry within the unit cell is described by a *space group*, and chiral molecules like proteins can crystallize in 65 different space groups.

### 2.2.1. Crystallization

While substances like salts and small organic molecules easily form crystals, the big, flexible and thermo-labile biological macromolecules usually do not, and getting them to arrange in a symmetrical pattern that would diffract x-rays is often a non-trivial task. The process of protein crystallization usually involves preparation of a concentrated protein solution of about 10 mg/ml, to which a low concentration of precipitant is added and the mixture is allowed to slowly concentrate. In an aqueous solution, water molecules interact with polar amino acid residues on the surface of a protein and form a hydration shell around it. A precipitant, such as a salt or a polymer like polyethylene glycol (PEG), competes with the protein molecules for water, and at a high enough concentration, there are not enough water molecules left to completely hydrate the protein. Consequently, the protein molecules start to make contacts to each other, and while this usually results in an amorphous precipitate, there is a chance that instead a few molecules form some kind of a symmetrical arrangement which then acts as a nucleation center for recruiting more protein molecules in the same arrangement and the growth of a crystal.

Every protein is unique with its complex shape and distribution of amino acids on its surface, and the conditions at which it would crystallize in general cannot be predicted in advance and have to be determined experimentally. Yet, some combinations of precipitants, salts, buffers and other additives have proven to be better in promoting crystal growth than others, and these are sold commercially as

crystallization screens, usually in 96-condition formats. As an initial trial, usually several hundred different conditions are screened, and if crystals or crystal-resembling objects are found in any of those, optimization of conditions usually follows by varying the concentrations of the constituent components, pH, various additives, etc., often resulting in another couple of hundred related conditions. The crystallization trials are greatly facilitated by pipetting robots that can handle small volumes down to 0.1 µl and greatly save time, effort and material and increase reproducibility. Crystallization is usually carried out using the so-called vapor diffusion technique in specifically designed plates containing individual wells with a large bottom compartment and a small upper compartment. The precipitant-containing crystallization solution is placed in the bottom compartment and a small drop of a concentrated protein solution in the upper, to which then an equal volume of the bottom solution is added and the well is sealed. Since the crystallization drop is now more dilute than the bottom solution, the water gradually evaporates from the drop until both solutions are equilibrated, and at some point during the slow concentration, crystals hopefully grow.

The first small plate- and needle-like crystals of the A1 read-through domain were obtained using the JCSG+ screen from Molecular Dimensions at several conditions that had pH 8.0 – 9.0, no salt and various PEGs as precipitants. Initial optimization of the crystallization conditions suggested that 40% PEG 300 at pH 8.5 results in the biggest crystals with maximum dimensions of 0.1 x 0.3 mm (Figure 10A), and one of these was used to collect a 1.8 Å dataset. Later a hexagonal crystal form (Figure 10B) was also discovered when any buffer was omitted from the crystallization drop (40% PEG 300 in water). The crystals reached a maximum size of 0.2 mm and diffracted to 2.9 Å.

Crystals of the Qβ coat protein – operator complex were also obtained using the JCSG+ screen at a condition containing 0.2 M zinc acetate, 10% PEG 3000 and 0.1 M sodium acetate, pH 4.5. The initial crystals were small, hexagonal bipyramid-shaped (Figure 10C) and did not exceed 0.1 mm in size. During optimization, some slightly

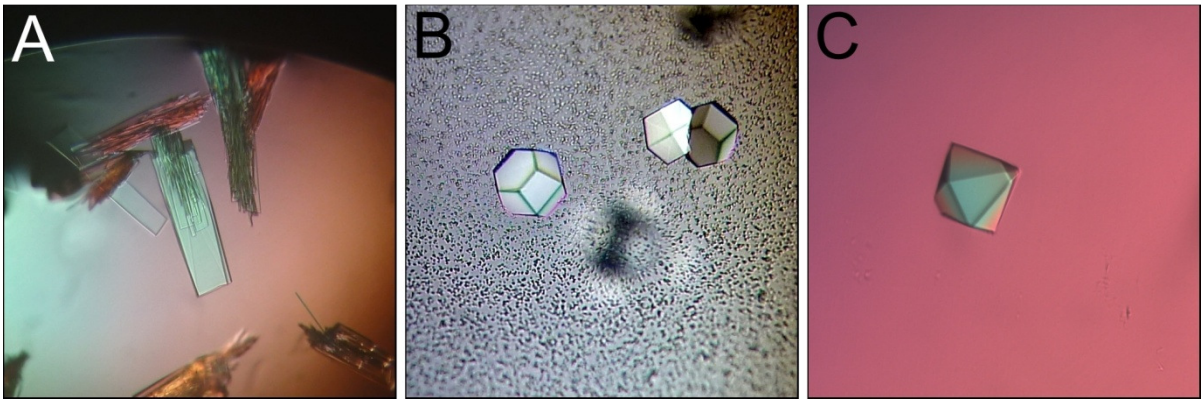


**Figure 10.** Crystals used in this study. A, monoclinic and B, hexagonal crystal forms of the read-through domain of the Qβ A1 protein. C, crystal of the Qβ coat protein – operator complex. The crystals are not shown to scale.

bigger crystals with maximum dimensions of about 0.15 mm grew when using 9% PEG 3000, and from one of those a 2.4 Å dataset was collected that was used for structure determination.

### 2.2.2. Data collection and processing

When crystals have been obtained, they are cryoprotected, usually by an addition of glycerol, flash-frozen in liquid nitrogen and x-ray diffraction data are collected while keeping the crystal at a cryogenic temperature. This is necessary because x-rays generate free radicals that react with protein molecules in the crystal and damage it, resulting in loss of diffraction, while at cryogenic temperatures the diffusion of the free radicals is greatly reduced and much more data can be collected. Nowadays the x-ray sources used for data collection are usually synchrotrons, dedicated facilities that can provide tightly focused x-ray beams with a very high intensity. The diffraction data are collected by rotating the crystal in the x-ray beam and capturing a series of images of the diffraction spots or *reflections* on a detector. Each reflection is then indexed, i.e., assigned a set of indices h, k and l that relate to a particular set of planes in the unit cell, and its intensity estimated by integrating the pixel values that represent the diffraction spot. The data then undergo scaling in which the integrated intensities from different images are normalized, followed by merging where intensities of reflections that are partially registered on adjacent images are combined and reflections with the same indices that are collected more than once due to crystal symmetry are assigned average values, and everything is combined in a single dataset.

### 2.2.3. Phase determination

The distribution of electron density in the crystal and the diffraction pattern are related to each other by a mathematical operation called a Fourier transform. Knowing either of those, the other can be calculated, and the desired electron density can be computed knowing the indices, amplitude, and phase of each reflection. The indices of the reflections are determined from their position in the diffraction pattern, the amplitudes can be easily calculated from intensities of the spots, but the phase information is lost and cannot be obtained from the diffraction images. Unfortunately, the electron density cannot be calculated without knowing the phases and they must be obtained by other, indirect means.

A classic method of solving the "phase problem" is multiple isomorphous replacement (MIR) that involves preparation of derivative protein crystals that contain several heavy atoms in them but are otherwise the same, or isomorphous to the "native" crystals. If the derivative and native crystals are isomorphous, their diffraction patterns will look the same, but since the scattering power of an atom is approximately proportional to the square of its atomic number, just a few atoms of an electron-rich

element like mercury or gold in the unit cell will cause measurable differences in intensities of the diffracted x-rays. The key to solving the phase problem is the fact that while it remains true that the electron density cannot be calculated without phases, for very simple structures consisting of just a few atoms, it is possible to deduce the position of the atoms in the unit cell by other means, using only amplitudes. The amplitudes of the heavy atom substructure can be easily calculated by subtracting the amplitudes of the respective reflections from the native protein crystal from those of the heavy atom derivative. When the positions of the heavy atoms have been deduced, from those a simulated diffraction pattern can be calculated containing amplitudes *and* phases, and the heavy atom phases then act as a stepping stone that in turn allow the deduction of protein phases. However, a single heavy atom derivative does not provide unambiguous values for the phases, and some three or four different derivatives are required to estimate the protein phases with a sufficient precision to generate an interpretable electron density map. If an x-ray source with an adjustable wavelength is available, as is the case in many synchrotron beamlines, a phenomenon called anomalous scattering can be utilized to obtain more precise phases. Anomalous scattering is observed when the energy of the beam is close to an x-ray absorption edge of the heavy atom in the crystal, and result in some additional changes in the intensities of the diffracted beams. While the theory behind this is somewhat more complicated than in MIR, the differences in intensities can be utilized to gain additional phase information and provide better phase estimates. Using this method, multiple isomorphous replacement with anomalous scattering (MIRAS), fewer derivatives are required, and even one good derivative may sometimes be sufficient to solve the structure.

In practice, preparation of the heavy atom derivative crystals is a laborious task that can set back the structure determination project for quite some time. This usually requires growing large amounts of crystals in optimized conditions and soaking them in heavy atom compound-containing solutions. In a largely trial-and-error manner, many different conditions have to be tried to find the right heavy atom compounds, concentrations and soaking times to derivatize the crystal but not to destroy it, and even if the crystal appears to be fine, there is a chance that binding of the heavy atoms slightly alters the packing of molecules and causes non-isomorphism, drop in resolution and other undesirable effects. The only reliable way to determine if the heavy atoms have bound to the protein at all and if the crystal is suitable for phasing is to collect diffraction data and process them. Many of the heavy atom compounds are also very toxic, requiring strict safety precautions while handling them. For these reasons, MIR and MIRAS are rarely used nowadays and other methods that employ genetically exchanging methionine residues in proteins to selenomethionines and collecting

anomalous signal from the selenium atoms have largely taken over. However, in some cases like the Qβ A1 protein where the number of methionines was too low, MIR and MIRAS can still be a method of choice for solving the structure.

The structure of the Qβ A1 protein described in this thesis was determined using MIRAS with two heavy atom derivatives, mercury and iodine. Derivatives with some gold and platinum compounds were also prepared, but these turned out to have no heavy atoms bound to the protein. Initial attempts using mercury compounds indicated that the mercury atoms indeed react with the A1 protein, but the crystals diffracted poorly and were too non-isomorphous to be useful. The fact that the A1 crystals were very thin and fragile further complicated the situation, but ongoing optimization efforts succeeded in slightly thicker and more robust crystals when a combination of 20% PEG 300 and 10% PEG 2000 MME was used instead of the original 40% PEG 300. After numerous trials, a suitable derivative was finally obtained by soaking a crystal in 20 mM mercury(II) nitrate for 30 minutes, followed by backsoaking in the original crystallization solution for 10 seconds to reduce background from unbound mercury atoms. The crystal diffracted to 3 Å and withstood collection of three datasets at different wavelengths. For iodine derivatization, a crystallization solution containing 0.1M $I_2$ in 0.1M KI was prepared, however, the iodine prominently precipitated in the PEG-containing solution. The undissolved iodine was removed by centrifugation, and the resulting solution with an unknown concentration of iodine was used for soaking the crystals overnight. One of the derivative crystals diffracted to 2.92 Å and allowed the collection of two datasets at different wavelengths. Data from the two derivatives permitted successful determination of the positions of the mercury and iodine atoms, after which some additional heavy atom refinement and density modification techniques in dedicated programs were used to obtain the initial phases. When the resulting map revealed protein-like features like recognizable strands and a piece of density that looked like an α helix, it was clear that the phase problem has been solved and the three-dimensional structure of the A1 protein could be determined.

In case a structure has already been determined for a similar protein (usually judged by sequence identity), the trouble of preparing heavy atom-derivative crystals can be avoided altogether and another method, molecular replacement (MR), can be used to obtain the initial phases. To do this, it is first necessary to prepare a "virtual crystal" in computer where molecules of the known structure are placed in the same unit cell, space group and in the same orientation as in the unknown structure. The orientation of molecules in the unknown structure is, however, *unknown*, therefore an exhaustive computer search has to be performed by translating and rotating the molecule of the known structure in the artificial unit cell, calculating simulated diffraction patterns and trying to find orientations where the calculated intensities best

match the experimentally measured ones. Since the simulated diffraction pattern from the virtual crystal contains both amplitudes and phases, the calculated phases can then be combined with the experimentally measured intensities of the unknown structure and an initial electron density map of the unknown structure can be calculated.

In this work, I have used the MR method to obtain phases for the Qβ coat protein – operator complex. In this case, part of the structure, the coat protein dimer, was already known from the previously determined Qβ capsid structure while that of the RNA operator was not. A MR run with the Qβ coat protein dimer as the search model successfully located the protein in the unit cell, and when the resulting electron density map was examined, a region of density clearly resembling RNA could be seen under the RNA-binding surface of the dimer.

### 2.2.4. Model building, refinement and validation

After the initial phases, obtained one way or another, have led to an interpretable electron density map, the structure is essentially considered "solved", but the structure determination process is yet far from finished. The next step involves placing atoms in the electron density map to build a model of the molecules in the unit cell. Usually the model does not have to be built completely from scratch; there are programs that attempt to auto-trace the protein chain in the map and in case of MR, the search model serves as the starting point. When a reasonable model is built using the available map, it undergoes a refinement process in which the atomic coordinates are statistically adjusted so that the calculated diffraction pattern from the updated model best fits the experimental data. The refinement employs setting several restraints such as bond lengths and angles to avoid generating a model that might appear to fit the data well but has physically implausible features. After refinement, a new electron density map is calculated using the improved phases, and another cycle of model rebuilding and refinement follows. After numerous cycles, at some point the crystallographer feels that no further improvement of the model is possible, and the model proceeds to structure validation. Validation involves checking the model for features like unusual side chain conformations, angles between backbone atoms, close contacts between molecules etc. that might signal for some errors in the model which have to be corrected in further rounds of model rebuilding and refinement. After the model has, hopefully, passed the validation checks, it is finally deposited in the Protein Data Bank (PDB), a global public repository of the determined three-dimensional structures of macromolecules where it gets assigned a four-character identification code, a "PDB ID". Models of the read-through domain of the Qβ A1 protein can be accessed from the Protein Data Bank using PDB IDs 3RLK (the monoclinic crystal form) and 3RLC (the hexagonal form) and the Qβ coat protein – operator complex using PDB ID 4L8H.

## 2.3. Propagation and purification of bacteriophages

Usually, propagation of *E.coli* phages is not an exquisitely complicated task – to a culture of bacteria at an early- or mid- log phase add phage at a multiplicity of some ten virions per cell, after an hour or two spin down the remains of what was once bacteria, and what is left is a high-titer phage lysate for your needs. Propagation of bacteriophage M, however, was not nearly that convenient, and required a different approach to provide an adequate amount of phage for genome sequencing purposes. All of the *Leviviridae* phages use pili as their cellular receptors, but the pili can be morphologically rather different. The F-pili that the "classic" ssRNA phages bind to are flexible and perform their functions well when the bacteria are in liquid medium. However, some other conjugative plasmids express what are called surface mating systems which have rigid pili and transfer DNA efficiently only when bacteria are growing on solid surfaces, apparently because the pili are fragile and break off easily due to shear forces in liquid media. Plasmids belonging to incompatibility group M (IncM) encode the surface mating type pili and transfer some three to four orders of magnitude better on plates than in liquid media (Bradley et al., 1980). The IncM plasmid-specific phage M thus could not be propagated in liquid media using the standard protocol. Instead, the bacteria were first grown in liquid media overnight without agitation to minimize breaking of the pili. To propagate the phage, a small volume of the host cell suspension and phage lysate were spotted on 1.5% LB agar plates, overlaid with molten 0.7% LB agar, mixed by gentle swirling and incubated overnight. The next morning, top agar layers containing a lawn of lysed bacteria were scraped off, transferred to centrifuge tubes and centrifuged to sediment the agar. The phage could then be concentrated by addition of PEG and NaCl to the supernatant.

One of the most powerful techniques for purification of phage virions is centrifugation in cesium chloride density gradient that separates macromolecules based on their buoyant density. While the density of proteins is around 1.3 g/cm$^3$ and that of DNA and RNA exceeds 1.6 g/cm$^3$, the virions of ssRNA phages contain both protein and RNA and have an intermediate density. Since the protein and RNA are in very precise proportions in the particles, the virions have a very sharply defined density of about 1.4 g/cm$^3$, e.g., MS2 virions have a density of 1.38 ± 0.01 g/cm$^3$ (Kuzmanovic et al., 2003). Accordingly, centrifugation of the concentrated M lysate in a CsCl density gradient resulted in a sharp band that was collected and dialyzed against a large volume of buffer to remove excess salt. In essentially a single-step purification, the preparation was pure from almost all bacterial proteins and contaminating cellular RNA and was suitable for RNA extraction and genome sequencing.

Bacteriophage φCb5 was propagated by Dr. Andris Dišlers and purified by Dr. Andris Kazāks as described previously (Plevka et al., 2009).

## 2.4. Sequencing and analysis of phage genomes

The genome of phage φCb5 was sequenced by Dr. Andris Kazāks. To extract RNA from M virions, the phage preparation was treated with a mixture of phenol and sodium dodecyl sulfate (SDS). Phenol and SDS denatures the capsid proteins and after the mixture is centrifuged, the proteins are dissolved in the bottom organic phase while the RNA is left in the upper aqueous layer from which it can then be precipitated using ethanol and used for sequencing. The genomic RNA was reverse-transcribed using a primer with a specific sequence at the 5' end followed by six random nucleotides at the 3' terminus. The resulting cDNA strands were dATP-tailed and their complementary strands synthesized in PCR using an oligo(T) primer and a primer corresponding to the sequence-specific 5' part of the one used in the reverse transcription. The PCR products were separated in an agarose gel and a slice corresponding to 1000 – 3000 base pair DNA fragments was cut out, the DNA extracted and cloned in plasmid vectors for sequencing. The insert-containing clones were sequenced using the classic dye-terminator Sanger sequencing method. Since the initial cloning procedure already involved 3'-tailing of cDNAs, it was possible to determine the 5' end of the genome from these clones. To determine the sequence of the 3' end, phage RNA was tailed with Poly(A) polymerase and reverse-transcribed using an oligo(T) primer followed by PCR using oligo(T) and a sequence-specific primer close to the then-known 3'end of the genome, and the PCR fragment was cloned and sequenced.

After the complete genome sequence had been obtained, phylogenetic and secondary structure analyses were carried out. The first and most important step in analyzing the evolutionary relationships among different phages is alignment of their genomic RNA and protein sequences. This is done using various computer algorithms that usually involve first calculating similarities for all possible sequence pairs and then iteratively aligning them starting from the most similar ones (Higgins and Sharp, 1988). From the multiple sequence alignment, a tree can be constructed that represents the relatedness of the sequences and infers their evolutionary history. The simplest and fastest methods for doing this involve the distance-matrix approach that in essence graphically represents arithmetically calculated "genetic distances" between the aligned sequences. Some more advanced techniques include approaches such as maximum likelihood and maximum parsimony that employ statistical methods to evaluate probabilities of different evolutionary events and aim to arrive at a tree that requires the smallest amount of such events. These methods can also take into account different mutation rates in various branches of the tree that the distance-matrix methods generally cannot.

## 2.5. RNA secondary structure analysis

The secondary structure of a single-stranded RNA molecule is defined by double-stranded regions in it, and determining its secondary structure in essence means to elucidate which nucleotides base pair to each other and which remain single-stranded. One obvious way of doing this is to determine it experimentally, and this can be done by using nucleases that cut single-stranded RNA like the S1 nuclease, RNAse A and RNAse T1. Analysis of the digestion products thus can identify double- and single-stranded regions in the genome like hairpin loops and bulged nucleotides. Such techniques were extensively used to study the secondary structure of MS2 RNA, the first-ever sequenced genome (Fiers et al., 1976). Nowadays, with ever-increasing computing power and more sophisticated algorithms, *in silico* prediction of the secondary structure has become the primary method for analyzing RNA structure. The prediction is based on thermodynamic principles by trying to find a structure with the lowest free energy. The free energy of the entire RNA molecule is calculated as a sum of free energies of all the base pairs and unpaired nucleotides, which are in turn based on experimentally measured values. The secondary structure predictions are usually done on dedicated servers like RNAfold (Hofacker, 2003) that provide web interfaces for inputting the sequence and secondary structure plots as an output.

Just like in proteins where the amino acid sequences can diverge while the fold remains the same, also in RNA the secondary structures may be conserved despite nucleotide changes. If RNA sequences from several related species are available, a comparative analysis can give more confidence that the predicted structures are indeed present in real life. For example, two distant nucleotide changes like G→A and C→U might not look very significant, but if the G and C in one RNA molecule form a base pair in a, say, long-distance interaction, the changes to A and U, respectively, in another RNA molecule preserves the base pairing and suggests that such interaction indeed exists. The ssRNA phages with their high mutation rates are prime examples where the structure of RNA is preserved over nucleotide sequence as was nicely illustrated for allolevivirus Qβ, M11 and NL95 (Beekwilder et al., 1996, 1995).

Although a combination of secondary structure prediction and sequence co-variation analysis often permits to draw rather confident conclusions about the RNA structure, such an approach inevitably has its limits. For example, the secondary structure prediction programs usually cannot predict RNA structures like pseudoknots that might be important in RNA function like was the case with a long-range pseudoknot in the Qβ genome (Klovins and Van Duin, 1999), therefore critical analysis and experimental verification of the predicted structures should never be neglected.

# 3. RESULTS

## 3.1. Structure of the Qβ A1 protein

While levivirus capsids consist of 180 copies of the coat protein and a single copy of the maturation or "A" protein, those of allolevivirus besides the coat protein contain two minor proteins, A1 and A2. The A2 protein is homologous to the A protein of leviviruses, but the presence of the A1 protein is a unique feature of allolevivirus. Soon after the finding that Qβ virions contain an extra protein (Garwes et al., 1969) it was discovered that phage mutants that are unable to synthesize coat protein also do not produce the A1 protein (Horiuchi et al., 1971), and that the first eight N-terminal residues of the coat and A1 proteins are identical (Weiner and Weber, 1971). This strongly suggested that the A1 protein is an extended variant of the coat protein that is generated when ribosomes read-through the termination codon of the coat gene. Indeed, when Qβ phage was propagated in a UGA stop codon-suppressor strain, the molar amount of the A1 protein in the capsid increased from 2% to 7% (Weiner and Weber, 1971), and subsequent amino acid sequence analysis definitively showed that the A1 protein contains the C-terminal sequence of the coat protein followed by a single tryptophan residue, after which the sequence of the read-through domain continues (Weiner and Weber, 1973). The N-terminal coat protein domain of the A1 protein forms a heterodimer with a coat protein molecule which then gets incorporated into the capsid along with "normal" coat protein homodimers (Takamatsu and Iso, 1982), while the read-through domain appears to be located on the outside of the capsid. When Qβ virions are subjected to native polyacrylamide gel electrophoresis, they form a wide diffuse band while the levivirus R17 migrates as a sharp narrow band (Strauss and Kaesberg, 1970). It has been shown that the mobility of Qβ virions in the gel and in sucrose density gradients depends on how many copies of the A1 protein are present in the capsid (Radloff and Kaesberg, 1973). The same study also showed that limited proteolysis of the Qβ virions results in a narrower and sharper band in the gel than native Qβ, presumably due to partial degradation of the A1 proteins. All of these findings are best explained if the read-through parts of the A1 protein are considered to be located on the exterior of the virions that would effectively increase their size and hydrodynamic drag and manifest in the observed behavior in gel electrophoresis and sucrose gradient sedimentation.

About the only thing that is known about the function of the A1 protein is that it is somehow necessary for the formation of infectious virions – when Qβ particles were reassembled *in vitro* by mixing purified phage components, addition of the A1 protein to the mixture was essential for producing infectious particles (Hofstetter et al., 1974). After this finding, for more than thirty years essentially nothing new has appeared

about the A1 protein, and its precise function has remained enigmatic. The amino acid sequence of the Qβ A1 protein does not offer many clues as it is not similar to any other protein except the A1 proteins of related allolevivirus. However, it is not uncommon that proteins with unrecognizable sequence similarity have similar folds and possibly similar functions; therefore information about the three-dimensional structure of the A1 protein has the potential to shed some light on its function. Thus, I picked Qβ as the prototype allolevivirus and set out to determine the structure of the A1 protein.

### 3.1.1. Structure determination and quality of the models

Qβ phage virions containing the A1 protein have been crystallized and their three-dimensional structure determined (Golmohammadi et al., 1996), but the read-through extensions were not visible in the crystal structure, apparently due to their low abundance and presumed random orientation in the capsid. When the A1 protein is produced from a recombinant plasmid, it alone is insoluble and cannot assemble into particles without the assistance of the coat protein (Vasiljeva et al., 1998). When the phage is propagated in a UGA stop codon-suppressor strain or when the coat and A1 proteins are co-expressed from a plasmid, the amount of the A1 protein that can be incorporated into the particles seems to be limited to about 15% (Vasiljeva et al., 1998; Weber and Konigsberg, 1975), too little to be useful for structure determination. Since all the evidence pointed towards that the read-through part of the A1 protein is a separate, soluble domain outside of the virions, it appeared reasonable to express the read-through domain separately.

When the amino acid sequence of the A1 protein is examined, six residues into the read-through part there is a sequence GGGSGS that looks like a flexible linker that could possibly separate the coat and read-through domains. Therefore a construct of a 6xHis-tagged variant of the read-through domain starting from the last serine of the putative linker (residues 144-328 of the full-length A1 protein) was prepared, and the resulting protein turned out to be highly soluble, could be readily purified and suitable to proceed with crystallization. The protein was crystallized in two crystal forms, monoclinic and hexagonal, which diffracted to 1.8 and 2.9 Å resolution, respectively. The structure of the monoclinic form was solved by multiple isomorphous replacement with anomalous scattering (MIRAS) using two derivatives, and that of the hexagonal form by molecular replacement using the model from the monoclinic crystal form. Except for the 6xHis-tag and the first two residues of the crystallized domain, the polypeptide chain for the monoclinic form could be traced unambiguously, without breaks, from residue 146 (the numbering of residues is as of full-length A1 protein) to the end of the chain. In the hexagonal form, another seven N-terminal residues could not be located in the electron density, and the chain was traced starting from residue 153. The domain adopts an

almost identical conformation in the two crystal forms, with a root mean square deviation (rmsd) of 0.76 Å for the main-chain atoms.

### 3.1.2. Overall structure

The overall fold of the read-through domain (Figure 11A) is not similar to any other published structure in the Protein Data Bank, according to the DALI server (Holm and Rosenström, 2010). Except for the N-terminal region, the domain has a compact, roughly globular shape with a mixed α/β architecture. The core of the domain is built of β-sheets: strands β2, β3, β6, β7 and β8 form a heavily deformed, five-stranded β barrel on one side of the protein, whereas β1 and β4 and β5, β9 and β10 form two anti-parallel sheets on the other side. There are three α helices and two $3_{10}$ helices in the protein, which are all short and are located predominantly on the surface. A remarkably long loop (23 residues) connects the first $3_{10}$ helix and strand β5, but it is well ordered and kept in place by extensive hydrogen bonding involving main-chain and side-chain atoms. Eight out of the first 15 residues that are visible in the electron density map are prolines. These residues form a polyproline type II helix that stretches for about 45 Å before turning 90 degrees toward the rest of the protein (Figure 11B). The polyproline

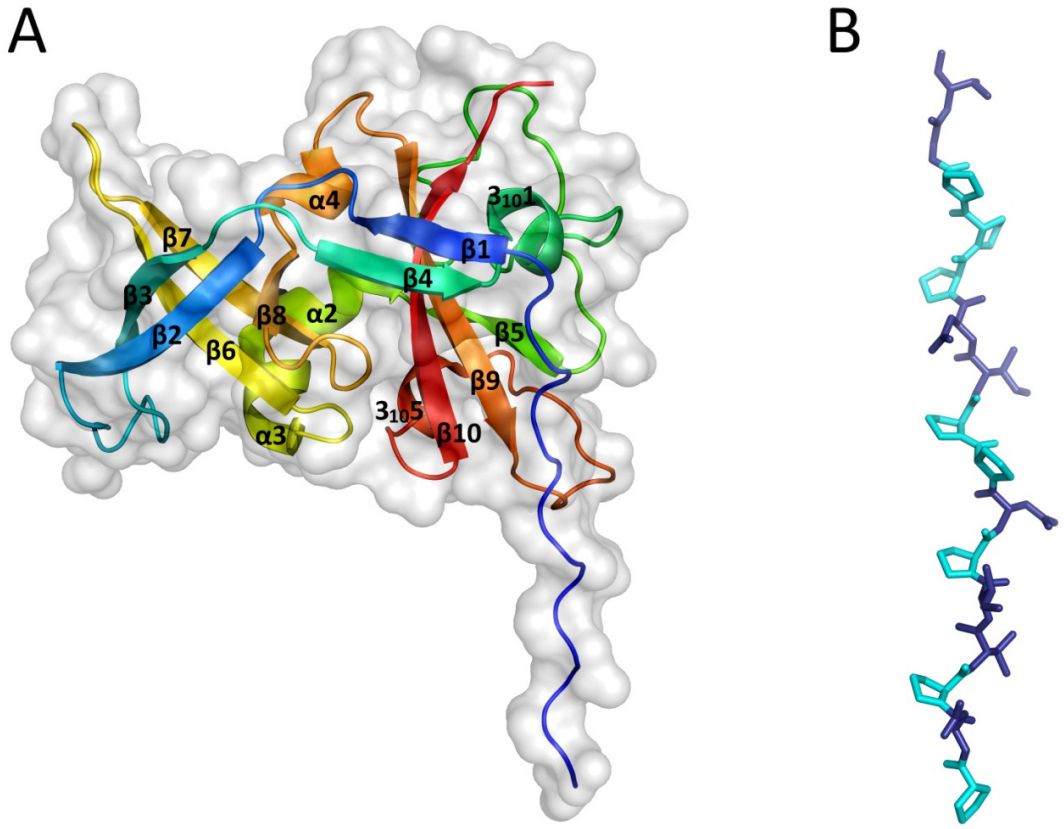


**Figure 11**. Structure of the read-through domain. A, overall structure of the domain. The protein is represented as a cartoon model rainbow-colored blue (N-terminus) to red (C-terminus) and overlaid with a surface representation of the domain (light grey). B, a detailed view of the polyproline helix. In the first 16 residues of the model, prolines are represented in cyan and other residues in deep blue.

helix is held in position by two crystal contacts with the globular part of neighboring molecules in the monoclinic crystal form but not in the hexagonal form. Consequently, the distant part of the helix is not visible in the hexagonal form, which suggests that it is flexible in solution.

### 3.1.3. Conserved regions

Based on phylogenetic and serological criteria, alloleviviruses cluster into two groups denoted III and IV (Bollback and Huelsenbeck, 2001; Furuse, 1987). Up to date, there are 15 allolevivirus genome sequences available, of these eight are from group III and seven from group IV. When all of the sequences are aligned, coat proteins are the most conserved (approximately 64% sequence identity), followed by the replicase (approximately 44% identity) and maturation proteins (approximately 29% identity). When sequences of all of the known A1 extensions are aligned, the total identity is only 26%, making them the most divergent part of all phage proteins. However, in a sequence alignment of A1 extensions from representative phages from group III (Qβ and MX1) and group IV (FI and SP) several conserved regions emerge (Figure 12A). First, in the N-terminal part (residues 146-159), approximately 50% of the residues are prolines in all alloleviviruses, suggesting that the polyproline helix is present in all allolevivirus A1 proteins and is probably important for their function. A short stretch of amino acids immediately following the helix is also conserved. The most prominent

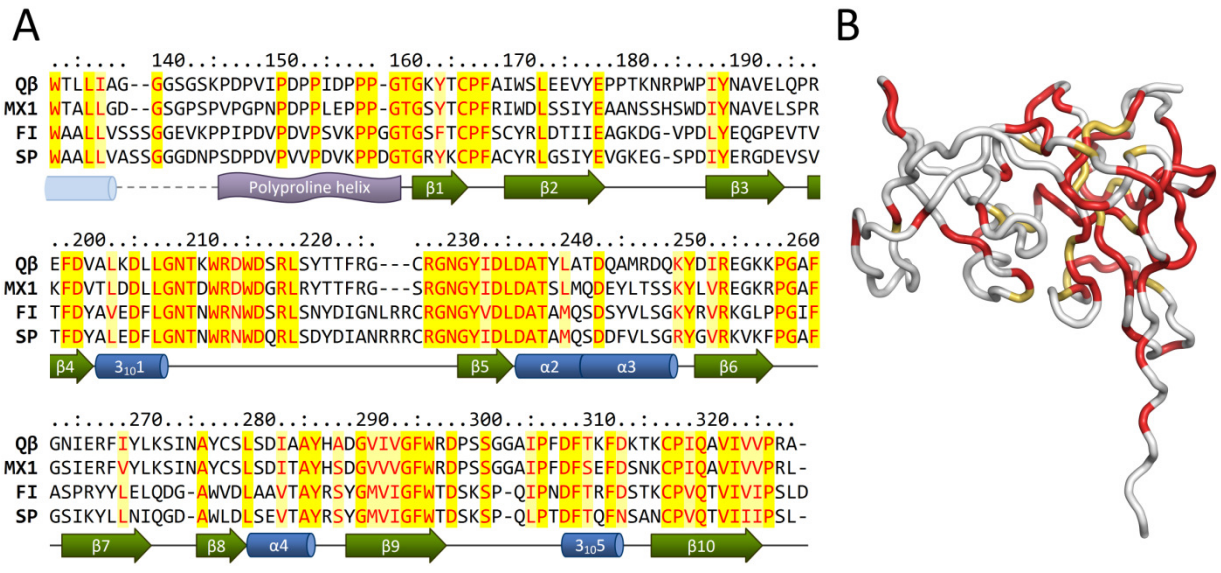


**Figure 12.** Conserved regions of the read-through domains. A, sequence alignment of the read-through domains from different alloleviviruses. Conserved residues are colored red; of these, identical residues are shaded yellow and non-identical light yellow. Assigned secondary structure elements are presented below the alignment. A dashed line represents the portion for which no experimental data are available; the α helix from secondary structure prediction is drawn as a pale blue cylinder. B, mapping of the conserved regions on the three-dimensional structure of the read-through domain. Identical and non-identical but conserved residues as of Fig. 12A are colored red and yellow-orange, respectively.

conserved regions are located at residues 207-219 and 228-238, which form part of the long loop between helix $3_{10}1$ and β5 and extend to strand β5 and the beginning of helix α2. The C-terminal region of the domain is also relatively conserved. Interestingly, the majority of the conserved residues cluster on one side of the protein closer to the polyproline helix (Figure 12B), suggesting that this part of the domain is the most critical for performing its function.

### 3.1.4. Possible function of the A1 protein

Just like the amino acid sequence, also the three-dimensional structure of the A1 extension is not similar to other known proteins, leaving no clues about its evolutionary origin, and despite the initial hopes, the actual function of the read-through domain has remained about as mysterious as before. Although there is hardly any doubt that the read-through domains are located on the outside of the particle, the current structure does not show how the prolonged A1 proteins are accommodated in the capsid as there is no structural information about residues 133-145 which separate the coat and read-through domains. Secondary structure prediction by JPred (Cole et al., 2008) suggests that this region is unstructured except for the coat protein-proximal six residues, which together with the last three residues of the coat protein may form a short α helix. In the Qβ capsid structure, the C-termini of coat proteins are located close to each other, but are not directly exposed on the surface of the capsid. With some minor structural re-arrangements they could however easily reach the outside of the particle, but if there is indeed an α helix in the A1 protein at the very end of the coat protein domain, the rearrangements required would be somewhat larger because of the increased diameter of the helix. Although the proportion of full-length A1 protein in the capsids never seems to exceed 15%, there have been some experiments that show that when the read-through extensions are shortened to some twenty amino acids, the capsids can contain about 50% of them, the same proportion as the coat and shortened-A1 proteins are produced in the cell (Vasiljeva et al., 1998). Thus the amount of the A1 protein that can be accommodated in the capsids appears to be limited by steric reasons likely imposed by the size of the domain.

An interesting feature of the A1 protein undoubtedly is the long polyproline type II helix at the N-terminal part of the read-through domain. Although polyproline type II helices are not uncommon in proteins, the vast majority of them are shorter than six residues (Berisio et al., 2006) and long helices are rare. The 15-residue-long polyproline helix in A1 is quite remarkable in this aspect, since, according to a statistical survey of polyproline helices in protein structures in 2006 (Berisio et al., 2006), the longest such helix observed in a crystal structure was that of the benzoylformate decarboxylase from *Pseudomonas putida* (Hasson et al., 1998) (PDB ID 1BFD), which is 14 residues long and

contains three prolines. The helix connects two subdomains of the enzyme but otherwise does not seem to have a specific function. Polyproline helices and proline-rich regions in general are relatively abundant in proteins and have different functions (Williamson, 1994), but they frequently serve as ligands for various protein-protein interaction domains, resulting in formation of protein complexes that are often involved in signaling and regulatory pathways in eukaryotic cells (Kay et al., 2000). In other proteins, proline-rich regions have a structural role and act as relatively rigid spacers to keep protein domains apart. For example, a 68 residue long proline-rich segment of the bacterial protein TonB has been shown to adopt a polyproline II-like conformation that spans the periplasm (Köhler et al., 2010).

The linker between the coat and read-through domains would stretch for estimated 35 Å, and is then followed by the 45 Å long polyproline helix, which is apparently also somewhat flexible. An obvious explanation for such a long linker would be that the read-through domain in the virion is positioned far away from the viral quasi-3-fold symmetry axis (relating the three quasi-equivalent subunits A, B and C) where the C-termini of coat proteins are located. Because both A1 and A2 proteins are required for infectivity, it seems possible that the two proteins might interact with each other and that the long linker would allow the read-through domain to reach the A2 protein, wherever in the capsid it is located. Experiments to test the association of the read-through domain with the maturation protein are underway in our laboratory. Thus, although the structure of the read-through domain did not provide immediate answers about its function, it gives a good starting point for further studies that could eventually lead to the understanding of the molecular mechanism by which the small RNA phages infect the bacterial host.

## 3.2. Structure of the Qβ coat protein – operator complex

The specific binding of the coat protein to an RNA hairpin at the beginning of the replicase gene to repress its translation is a mechanism conserved throughout much of the *Leviviridae* family. The details of the interaction have been extensively studied in phage MS2 both biochemically (Carey et al., 1983a; Romaniuk et al., 1987; Uhlenbeck et al., 1983) and structurally (Grahn et al., 2001; Helgstrand et al., 2002; Valegård et al., 1994, 1997; van den Worm et al., 1998), making it one of the best characterized protein-RNA interactions to date. The recognition mechanism involves the binding of two adenine bases, an unpaired one in the RNA operator stem and another in the hairpin loop, to symmetrical adenine-recognizing pockets in the protein dimer. The complex is further stabilized by aromatic stacking that extends from the helical RNA stem via two bases in the hairpin loop to a conserved tyrosine side chain in the coat protein. The coat protein – RNA interaction has also been characterized in *Pseudomonas* phage PP7 (Chao et al., 2008; Lim and Peabody, 2002; Lim et al., 2001). The operator of phage PP7 is remarkably different from MS2 and uses a distinct RNA binding mode. Nonetheless, the PP7 coat protein also uses symmetrical pockets to bind two adenine bases in the bulge and the loop, despite the fact that the pockets are very different from those found in MS2.

Bacteriophage Qβ is distantly related to MS2 with their coat proteins only about 20% identical. Both coat proteins preferentially bind their cognate translational operators, which are also rather different (see Figure 5). For strong binding to the MS2 coat protein, the operator helix needs to be at least five base pairs long and contain an unpaired purine nucleotide two base pairs prior to a four-nucleotide loop with adenosines as the first and last nucleotides and a pyrimidine nucleotide at the penultimate position (Romaniuk et al., 1987). For high-affinity binding to the Qβ coat protein, the operator requires a three-nucleotide loop and an eight-base pair stem with a bulged nucleotide four base pairs from the loop (Witherell and Uhlenbeck, 1989). The only critical nucleotide in the loop is an adenosine at the last position, whereas the unpaired adenosine in the stem can be mutated or removed altogether with a rather minor decrease in affinity (Lim et al., 1996). Despite the differences, several facts suggest that the RNA binding modes of the MS2 and Qβ coat proteins are related. Although the overall sequence identity is low, the three-dimensional structure of the two proteins is very similar, and many of the residues that are involved in RNA binding in MS2 are conserved in Qβ (Golmohammadi et al., 1996). Furthermore, MS2 and Qβ coat protein mutants that are able to tightly bind the operator of the other phage have been isolated (Lim et al., 1996; Spingola and Peabody, 1997) while analogous experiments have been unsuccessful with PP7 (Lim and Peabody, 2002).

The MS2 and PP7 RNA binding modes are so different that besides realizing that they are probably evolutionarily related, it is hard to tell something more about how the protein-RNA interactions co-evolved. In contrast, the MS2 and Qβ coat proteins are in an interesting same-but-different position with similar RNA binding modes but different specificities, and understanding the molecular details that allows the Qβ coat protein to achieve the specific binding has much greater potential to reveal how the co-evolution of protein and RNA structure took place. To address this issue, I set out to determine the three-dimensional structure of the Qβ coat protein in complex with its replicase operator hairpin.

### 3.2.1. Design of the assembly-deficient Qβ coat protein

Previous work with MS2 that led to numerous protein-RNA complex structures relied on the ability to soak small RNA hairpins into pre-crystallized capsids via pores that are present at their three-fold and five-fold symmetry axes. Initially, the same approach was applied to Qβ, however, when diffraction data from the RNA-soaked crystals were collected and an electron density map calculated, there were sadly no signs of RNA inside the capsid. This failure was first attributed to the fact that in contrast to MS2, the FG loops from neighboring Qβ coat protein dimers are linked to each other with disulfide bonds that result in covalent rings around the pores and presumably restrict RNA diffusion into the particles. To address this issue, we obtained a plasmid encoding a Qβ coat protein that had the cysteines in the FG loop mutated to glycines, produced capsids from the modified coat proteins and crystallized and repeated the RNA soaking experiments with those. Unfortunately, still no bound RNA was detected in the electron density maps, suggesting that perhaps the crystallization conditions (0.4 M NaCl at pH 7.5) are suboptimal for RNA binding and that the approach of soaking capsid crystals with RNA would not be successful with Qβ.

The structure of the PP7 coat protein in complex with its operator was determined not by the RNA-soaking method but via a different approach that utilized coat protein dimers that were incapable of assembling into capsids. This was done by truncating the FG loops of the protein which are important for capsid assembly, mixing the assembly-defective dimers with RNA and crystallizing the protein – RNA complexes. However, our initial attempts to truncate the FG loop of Qβ coat protein in a similar manner resulted in a largely insoluble protein that indicated that this method would not work with Qβ. Consequently, yet another approach was devised to introduce some amino acid point mutations into the coat protein that would prevent it from assembling into particles. Examination of the Qβ capsid structure suggested Asn129 as a good candidate for mutagenesis as its side chain forms two hydrogen bonds with the main chain of the adjacent dimer; thus, introduction of a bulkier side chain at this position would both

destroy the bonding and cause a steric clash with the nearby chain (Figure 13). A similar situation was observed for Pro42 in the CD loop where substitution with a longer side chain would likely result in a collision with the neighboring dimer. Mutation of the two residues to arginines (Pro42Arg, Asn129Arg) in the cysteine-less mutant (Cys74Gly, Cys80Gly) indeed resulted in a protein that produced a highly soluble and homogenous dimeric species suitable for structural studies.

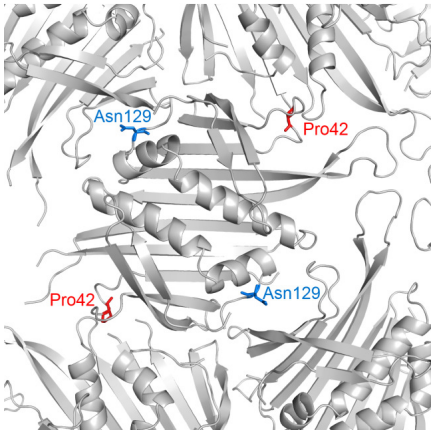


**Figure 13.** Locations of the mutated residues in Qβ coat protein.

### 3.2.2. Structure determination and quality of the model

The coat protein-RNA complex was prepared by mixing the purified assembly-deficient dimers and the RNA operator in a molar ratio of 1:1.2, and the mixture was immediately subjected to crystallization. Crystals were obtained that diffracted to 2.4 Å resolution, and the structure was solved by molecular replacement. The final model (Figure 14A) contains one Qβ coat protein dimer (chains A and B) and one RNA molecule (chain R). There are no crystal contacts close to the protein-RNA interface, suggesting that the model represents a biologically relevant structure. The unassembled dimer adopts a conformation highly similar to that found in the crystallized phage capsids (Golmohammadi et al., 1996), with an rmsd of Cα atoms of 0.8 Å. Notably, the EF

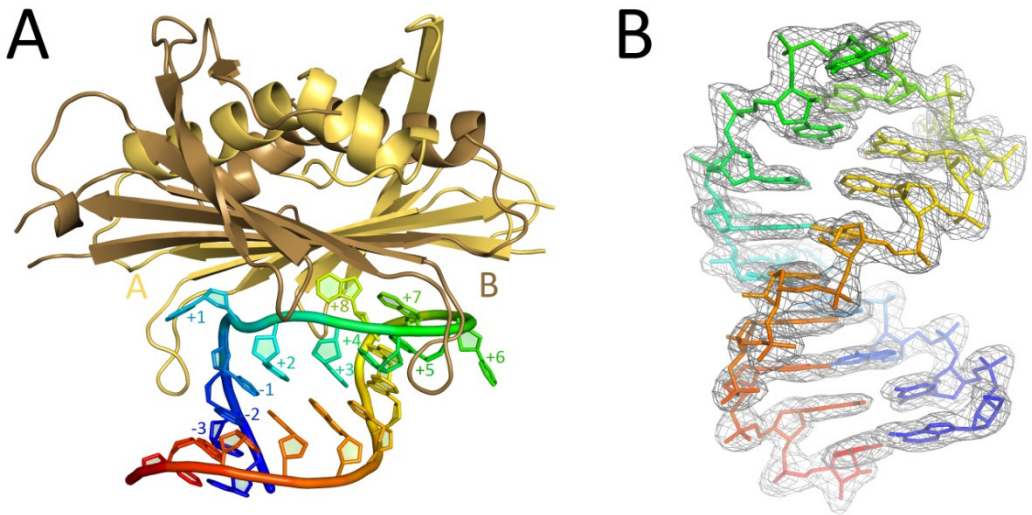


**Figure 14.** Three-dimensional structure of the Qβ coat protein-operator complex. A, overall structure of the complex. The coat protein dimer is represented in light orange (monomer A) and light brown (monomer B), and the RNA is rainbow-colored blue (5' end) to red (3' end). Nucleotide positions relative to the first nucleotide of the replicase initiation codon are indicated next to the bases. B, a close-up view of the RNA hairpin. The RNA is shown in a stick representation colored as in (A) as modeled into a 2Fo-Fc electron density contoured at 1.1 σ.

loops of the assembly-deficient dimer make contacts with RNA and can be reliably modeled, whereas they were only partly visible in the virus structure. In contrast, the FG loops (residues 74-84 of chain A and 75-83 of chain B) are disordered in the unassembled dimer and were not included in the final model. Electron density for the whole RNA molecule (20 nucleotides) was clearly visible (Figure 14B), and the complete hairpin was modeled without breaks.

### 3.2.3. Overall structure of the complex

The experimentally observed structure of the RNA hairpin is consistent with the predicted secondary structure and consists of an eight base pair stem, a three-nucleotide loop and an unpaired adenosine in the stem. The stem adopts a canonical A-form helical conformation with ribose puckers in the C3'-endo conformation except for loop nucleotides A+7 and A+8, which adopt more of a C2'-endo conformation. The majority of the contacts between the protein and RNA are sequence-independent interactions between the sugar-phosphate backbone of the RNA and the EF loop and β-strand F of both coat protein monomers. The adenine base of the A+8 nucleotide fits into an adenine-binding pocket formed by Val32, Thr49, Ser51, Gln65 and Lys67 of chain A in the coat protein dimer. The base of the A+7 nucleotide is stacked between C+5 in the stem and the aromatic side chain of Tyr89 of the A chain. In addition, the hydroxyl group of the tyrosine forms a stabilizing hydrogen bond with an oxygen atom in the phosphate backbone. The base of the last loop nucleotide, U+6, points away from the protein and does not make any contacts with it. The unpaired A+1 nucleotide bulges out from the stem and stacks with Tyr89 in chain B of the coat protein. There seem to be no additional stabilizing interactions involving the base, but the phosphate oxygen of A+1 forms an electrostatic interaction with the side chain of Lys63 in the A chain, and additional contacts with sugars and phosphates of C-1, G-2 and U-3 nucleotides in the lower part of the stem stabilize the hairpin in the observed orientation.

### 3.2.4. Comparison of RNA binding between Qβ and MS2

The top part of the Qβ hairpin that faces the protein (nucleotides +3 to +8) adopts a conformation remarkably similar to that of the MS2 operator (nucleotides -9 to -4, respectively) with an rmsd of 1.1 Å, which demonstrates that the two proteins indeed share a similar RNA binding mode. The number of hydrogen bonds and electrostatic interactions between the protein and RNA is about the same in Qβ and MS2, however in MS2 a higher proportion of the interactions involve contacts with the nucleotide bases rather than the sugar-phosphate backbone (Figure 15A). The adenine-binding pocket of the Qβ coat protein is almost identical to that of MS2, and all of the base-protein interactions within the pocket are the same in the two phages. However, the nearby interaction between LysA43 and the phosphate backbone in MS2 is not preserved as the

equivalent ArgA47 in Qβ is too far away from the RNA (4.4 Å) to make any significant contribution to the interaction. The similarities in RNA binding of the two proteins extend to the A+7 nucleotide, which in Qβ is stacked with TyrA89, while in MS2 an analogous interaction is found between U-5 and TyrA85, and a contact between the hydroxyl of the tyrosine and a phosphate of the RNA backbone is also conserved. Like U-

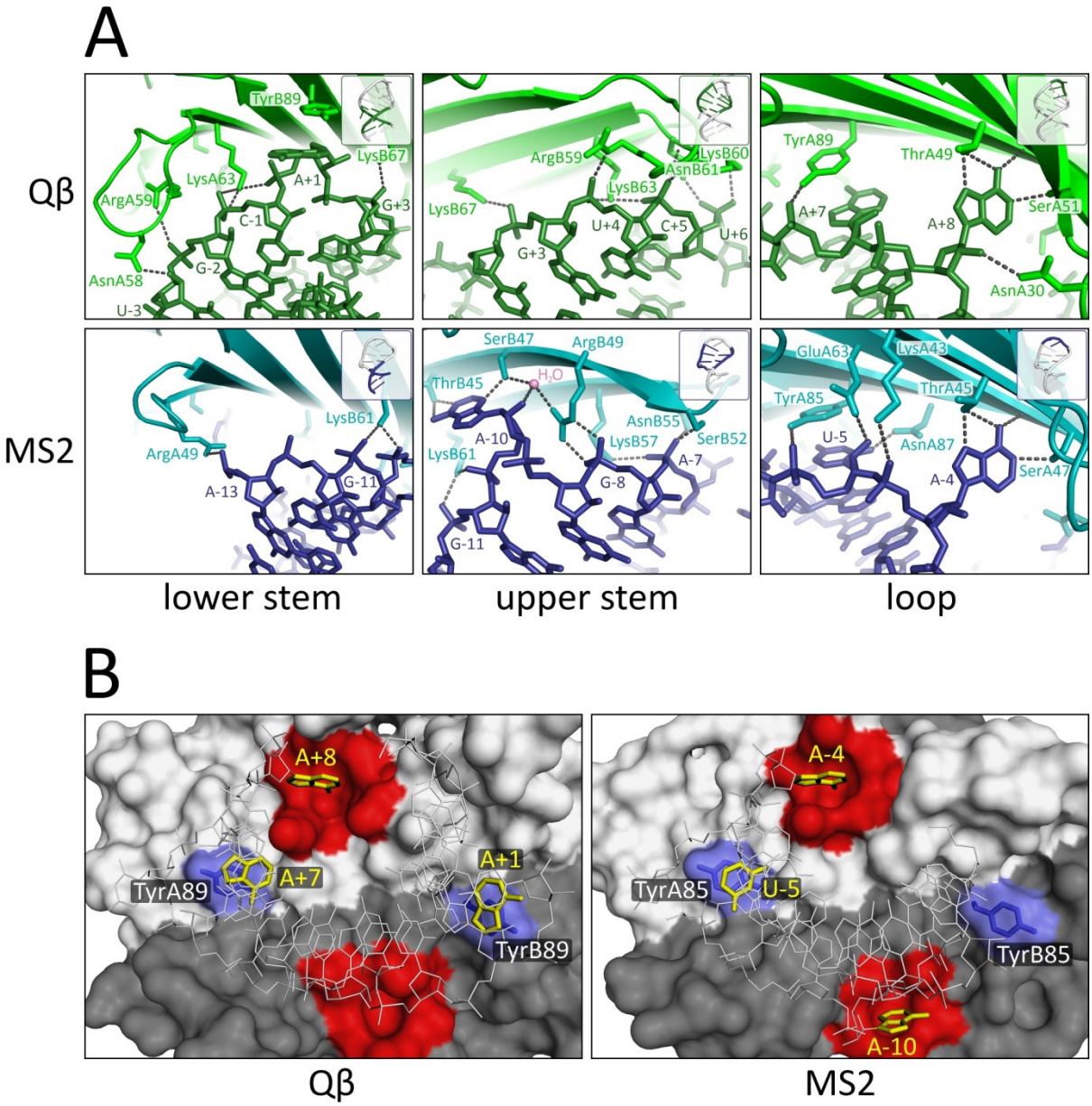


**Figure 15.** Differences in binding of the Qβ and MS2 coat proteins to their cognate operators. A, close-up views of the protein-RNA interactions in Qβ and MS2. Hydrogen bonds and electrostatic interactions in the lower and upper parts of the stem and the hairpin loops are indicated as grey dashed lines. The insets on top right highlight the approximate region of the operator hairpin that is visible in the particular close-up. B, comparison of protein-RNA interactions in Qβ and MS2 involving the loop and the bulged adenosine. The solvent-accessible surfaces of Qβ and MS2 coat protein dimers are shown in different shades of gray as for A and B monomers. The adenine-binding pockets are shown in red, while the tyrosine residues that stack with RNA bases are colored blue. The RNA is shown in light gray as a stick model except for the bases that occupy the adenine-binding pockets or stack with the tyrosine side chains, which are shown in yellow. In Qβ, only one of the symmetrical adenine-binding pockets is occupied and tyrosines from both monomers participate in base stacking. In contrast, both pockets are occupied by adenine bases in MS2, while only a single tyrosine is involved in base stacking.

6 of MS2, the U+6 in Qβ points away from the protein and does not make contacts with it. Finally, residues AsnB61 and LysB63, which make interactions with the sugar-phosphate backbone in Qβ, are conserved and provide the same function in MS2.

Away from the hairpin loop, the differences in protein-RNA interactions in the two phages become more pronounced. In the lower part of the hairpin, only a single electrostatic interaction exists between Arg49 of the A monomer and the -13 phosphate in MS2, but in Qβ the arginine residue is not conserved and interactions involving AsnA58, ArgA59 and LysA63 take place instead. The additional contacts are possible due to an extended EF loop that in Qβ is two residues longer than in MS2. However, the most profound difference between Qβ and other RNA phages involves the interaction with the bulged adenosine in the stem of the hairpin. In MS2, the bulged A-10 base fits into the same pocket as A-4 in the other monomer, albeit in a different orientation; however, in Qβ, the other adenine-binding pocket is empty, and the A+1 base is stacked with Tyr89 of the other monomer (Figure 15B). This configuration has not been observed in any other ssRNA phage coat protein-operator complex and thus represents a novel mechanism how an unpaired base in the stem can be accommodated.

### 3.2.5. RNA binding discrimination of Qβ coat protein

Since the conformation of the β sheet that makes up the RNA binding surface of the coat protein is very similar in MS2 and Qβ, superimposition of the two protein-RNA complexes using Cα atoms from strands D, E, F and G results in a very good alignment of the A+8/A-4 bases, the adenine-binding pockets and other conserved RNA-binding residues. A possible RNA discrimination mechanism for the Qβ coat protein can therefore be modeled by combining protein coordinates from the Qβ complex with RNA coordinates from the fitted MS2 complex.

In the modeled Qβ coat protein-MS2 operator complex, the A-10 and A-4 bases of the MS2 operator fit very well into the adenine-binding pockets of the Qβ coat protein, and many of the interactions with the RNA backbone in the upper stem seem to be preserved. There appear to be some differences regarding the interactions involving Arg49, which is found in the wild-type MS2 complex but is not conserved in Qβ. In the wild-type MS2 complex, Arg49 in the A monomer forms a salt bridge with the -13 phosphate, but this interaction is lost with the Qβ coat protein, which has a serine residue at the equivalent position. In the B monomer, Arg49 forms a salt bridge with the -8 phosphate and additionally coordinates a water molecule that forms a hydrogen bond with the O2’ atom of the A-10 ribose. In Qβ, the side chain of ArgB59 lies in approximately the same place as ArgB49 in MS2 and partly serves the same function by providing an electrostatic interaction with the phosphate of U+4. This interaction would likely be preserved in the complex with the MS2 operator, but the arginine side chain

would be too far away from the A-10 nucleotide to allow interactions similar to those observed in the cognate MS2 complex. Consequently, this might contribute to the weaker binding of the MS2 operator to the Qβ coat protein.

Another reason for the poor binding of the MS2 operator likely involves the -5 uracil base in the loop. The side chain of TyrA89 that stacks with A+7 in Qβ is tilted by approximately 20 degrees compared to TyrA85 in MS2. As a result, planes going through the U-5 base of the MS2 operator and the side chain of TyrA89 in Qβ coat protein would not be parallel which could lead to impaired binding of the RNA. In addition, the interaction between U-5 and AsnA87 that is present in the cognate MS2 complex is lost. The corresponding amino acid in Qβ is AspA89, and repulsion between the acidic side chain and the O2 carbonyl of the uracil base would prevent an analogous interaction with the Qβ coat protein. This is consistent with the observation that the interaction between an Asp91Asn Qβ coat protein mutant and the MS2 operator is 20 times stronger than with the wild-type Qβ coat protein (Lim et al., 1996). In contrast to aspartic acid, the asparagine side chain would permit formation of a hydrogen bond between the protein and RNA and result in the observed improvement in binding.

Finally, the three-nucleotide loop of the Qβ operator and the EF loops of Qβ coat protein appear to play a role in RNA binding discrimination as well. Biochemical studies have shown that addition of an extra nucleotide to the three-nucleotide loop severely reduces the binding, and that the Qβ coat protein requires a longer RNA stem than MS2 for a high-affinity interaction (Witherell and Uhlenbeck, 1989). The length-dependence is explained by the EF loops, which are longer in the Qβ coat protein than their MS2 counterparts and make contacts with the lower stem, which is likely necessary to compensate for the lack of some of the interactions in the upper part of the helix. However, from the model it appears that the size of the hairpin loop determines how effective the binding of the lower stem will be. When the stems of the MS2 and Qβ operators are superimposed, substantial differences in loop conformations are clearly evident due to the extra base pair at the top of the Qβ hairpin (Figure 16A). When the protein-RNA complexes of the two phages are superimposed, the smallest conformational differences are observed in the region comprising the loop and two preceding nucleotides and not in the stems (Figure 16B). Thus in the protein-bound state, the different-sized loops would cause the phosphate backbones of the Qβ and MS2 operators to follow different paths and impose different relative orientations of the RNA stems. As a direct consequence, binding of a hairpin with a three-nucleotide loop to the Qβ coat protein would position the lower part of the RNA stem in a more favorable orientation for making interactions with the EF loop than the binding of a four-nucleotide loop. The conformation with a three-nucleotide loop also restricts the ability to accommodate bulged nucleotides in the stem except those at a position four

nucleotides prior to the loop; in this case, an additional stacking interaction with the protein is formed that further stabilizes the complex. However, the unpaired adenosine is not critical for binding and results in only 1.5 to 5-fold reduction in affinity upon removal (Lim et al., 1996; Witherell and Uhlenbeck, 1989). The absence of the bulged adenosine would eliminate only a single stacking interaction since there are no additional contacts between the protein and the base, and would indeed result in a rather minor decrease in affinity. The lack of the unpaired base apparently does not impose significant conformational changes to the stem and still permits the EF loop to bind the lower part of the RNA hairpin, although the interactions are probably somewhat different than in the wild-type complex.

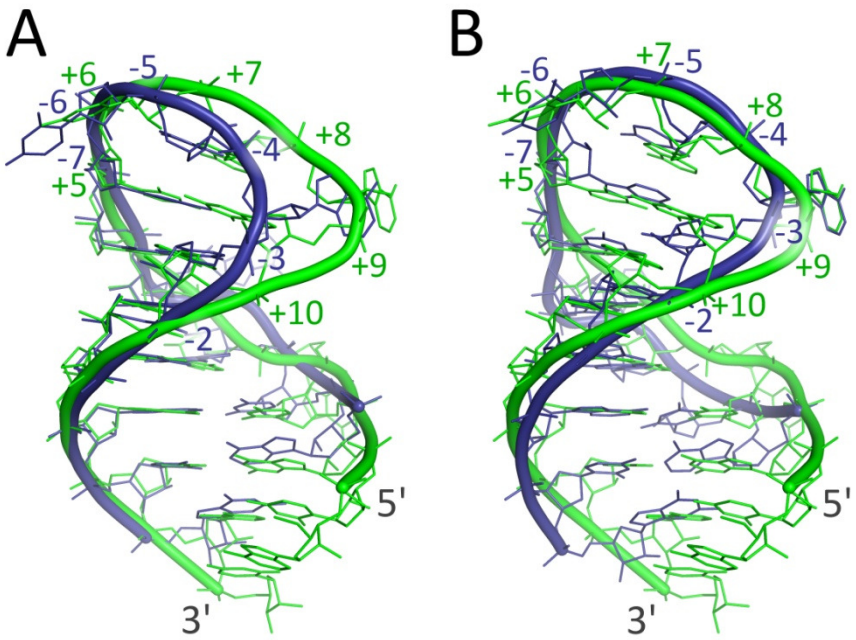


**Figure 16.** Conformational differences of the Qβ and MS2 operators. Superimposition of the helical stems (A) demonstrates the differences in hairpin loop conformations of the two operators. Superimposition of the RNA binding residues of the two cognate protein-RNA complexes (B) results in different relative orientations of the stems that in turn cause the phosphate backbones of the two RNAs to follow different paths. The Qβ (green) and MS2 (blue) operators are shown as stick models with the phosphate backbones represented by ribbon traces.

## 3.3. Genome structure of *Caulobacter* phage φCb5

From all of the isolated ssRNA bacteriophages, those infecting *Caulobacter* are perhaps the most distinct from the other ones. Bacteria of the *Caulobacter* genus are rather unusual in that they have a dimorphic life cycle with two cell types, stalked and swarmer cells. The stalked cells can attach to a surface by a polar stalk and asymmetrically divide to release a motile swarmer cell which later differentiates again to a stalked cell. The swarmer cells bear a flagellum and several pili, and it is those pili that the *Caulobacter* ssRNA phages use for attachment. Bacteriophage φCb5 was isolated in the 1960s along with several other ssRNA *Caulobacter* phages (Schmidt and Stanier, 1965). These phages formed three serologically distinct groups: group IV with phages φCb8r and φCb9, group V with phages φCb2, φCb4, φCb5, φCb12r and φCb15 and group VI with a single representative φCb23r. The three groups had distinct host ranges, with phages from group IV infecting *Caulobacter bacteroides*, group V *Caulobacter crescentus* and *Caulobacter vibrioides* and group VI *Caulobacter fusiformis*, respectively. Of the *Caulobacter* phages, only φCb5 has been propagated and purified in large scale and characterized in some detail (Bendis and Shapiro, 1970). Like the F pili-specific ssRNA phages, φCb5 capsids were small and spherical in appearance and contained the coat and maturation proteins. In stark contrast to the other known ssRNA phages, φCb5 virions were extremely salt-sensitive, and even milimolar amounts of magnesium salts were able to inactivate the particles. Interestingly, divalent cations could not be removed completely as addition of EDTA also completely inactivated phage preparations. In addition to that, RNA in φCb5 particles was sensitive to RNAse while that in the F pili-specific ssRNA phage capsids was not.

Some 35 years later, being interested in ssRNA phage biology, we obtained the largely “forgotten” phage φCb5 and began to study it in our laboratory. However, being primarily a structural biology lab, even before the sequencing of the φCb5 genome was completed, φCb5 capsids were already crystallized and their three-dimensional structure determined (Plevka et al., 2009). The structure revealed several interesting features and provided some explanation to the unusual properties of φCb5. The φCb5 coat protein is the shortest among the known *Leviviridae* phages with only 122 residues while those of the other phages range from 127 to 132 residues in length. Despite this, the fold of the protein is the same and also the capsids have approximately the same size as those of other ssRNA phages. The greatest savings in the φCb5 coat protein come from considerably shorter FG loops that in turn result in large, star-shaped apertures around the threefold symmetry axes of the capsid. These are sufficiently big to allow diffusion of RNAse molecules into the particles and explain the observed ribonuclease sensitivity. φCb5 capsids are stabilized by calcium ions and utilize salt bridges in

subunit-subunit interactions that explain their sensitivity to EDTA and instability in high salt. The structure also revealed that unlike other ssRNA phages, φCb5 capsids have nucleotide-binding pockets at the subunit interface collectively formed by three coat protein monomers and suggested that RNA might have a significant role in particle stability as well.

The distinction of φCb5 from other phages was also evident when the complete genome sequence of the phage was available. However, some time after our analysis of the φCb5 genome was published, a genome sequence of a "marine RNA phage MB" was deposited in the GenBank (accession code KF510034) that was distantly similar to φCb5 but not to other ssRNA phages. The sequence was identified during a metagenomic study to find new virus sequences in San Francisco wastewater. Sadly, no further information about the phage is available, thus the phage might be one from the other serological groups infecting different *Caulobacter* species, or infect some different bacterial genera altogether. Nevertheless, the sequence of phage MB provides some useful information when analyzing the genome of phage φCb5, thus in this chapter I provide an updated analysis of the φCb5 genome and compare it to the other known ssRNA phage genomes in light of the new data.

### 3.3.1. Overall structure of the genome and similarity to other phages

The genome of phage φCb5 is 3762 nucleotides long and organized similar to other ssRNA phages where after a short 5′ untranslated region (UTR), ORF1 encodes the maturation protein, ORF2 encodes the coat protein, and ORF3 encodes the replicase (Figure 17). With the single exception of phage MB, the nucleotide sequence of the φCb5

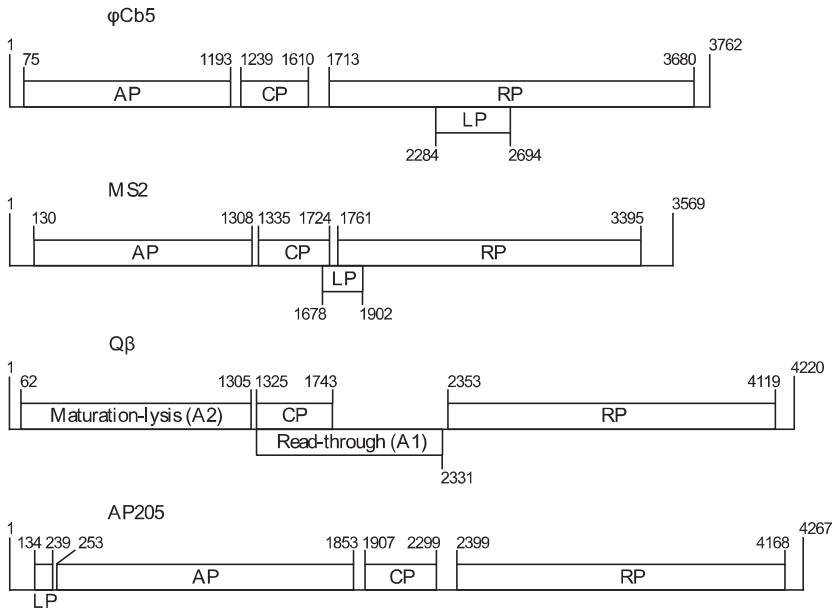


**Figure 17.** Genome organization of the ssRNA phages φCb5, MS2, Qβ, and AP205. Genes are drawn to their approximate scale.

genome and amino acid sequences of the individual proteins have very low homologies with their counterparts in other ssRNA phages. While the replicases of phages φCb5 and MB are about 39% identical, only the central part of the replicase (residues 295 to 537) can be aligned unambiguously also to the distantly related phages. Again with the exception of phage MB, the coat protein of φCb5 displays no detectable sequence similarity to other ssRNA phages, and also the maturation proteins, except for a short conserved stretch in the C-terminal part, are very different. The maturation and coat proteins have highly diverged also in phages φCb5 and MB with about 25% and 24% sequence identity, respectively.

### 3.3.2. Translation initiation site of the maturation protein and replicase

Initially, the most obvious initiation site for the maturation protein appeared to be the first AUG codon in the genome that also had a strong preceding Shine-Dalgarno sequence; however, mass spectrometry revealed the presence of a protein of a smaller mass than predicted from the sequence. To establish the actual translation start site of the maturation protein, the proteins of purified φCb5 virions were separated by SDS-PAGE, and the N-terminal sequence of the 40-kDa band was determined. The sequence was found to be ARIRN, corresponding to a nucleotide sequence 78 nucleotides from the 5′ end of the genome. The sequence is immediately preceded by a UUG codon, which can serve as an initiation codon in bacteria, and is probably the case for the maturation protein of φCb5. Although initially we felt that a possibility cannot be excluded that the upstream AUG codon is in fact used for translational initiation and that proteolytic cleavage occurs later, the sequence of phage MB indicates that this is likely not the case. Although weak, there is some amino acid sequence similarity at the very N-termini of the φCb5 and MB maturation proteins (Figure 18), and there are no upstream initiation codons in the MB genome, thus the UUG codon in φCb5 appears to be the actual initiation site of the maturation protein.

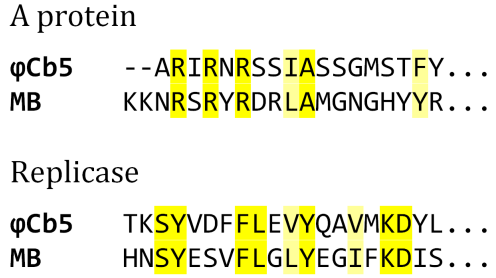


**Figure 18.** Alignment of the N-terminal sequences from bacteriophage φCb5 and MB A proteins and replicases.

Like the maturation protein, also the replicase ORF has several potential initiation codons, but the first AUG also has a significant SD sequence upstream. We therefore assume that the replicase is most likely translated starting from this codon, which is supported by the fact that there is reasonable sequence identity at the very N-termini of the φCb5 and MB replicases (Figure 18), and none of the alternative possible initiation codons in φCb5 are conserved in MB.

### 3.3.3. A non-canonical lysis gene

In φCb5, no obvious ORF at a position corresponding to the lysis protein of leviviruses or to AP205 could be detected. The levivirus and AP205 lysis proteins, albeit localized differently in the genome, both contain a transmembrane helix, and to see if a similar protein could be located somewhere in the φCb5 genome, the translated genome sequence was analyzed using the TMHMM 2.0 server (Krogh et al., 2001). In total three putative transmembrane helices were found that entirely overlapped with the replicase gene in a different reading frame. The first helix, encoded by nucleotides 2098 to 2226, lacked any suitable upstream initiation codon while the other two helices were found in a single potential ORF that had a strong SD sequence but an unusual start codon, UUG. Cloning and expression of the two ORFs (done by Dr. Andris Kazāks and described in more detail in paper III) revealed that the two-helix ORF is the lysis protein of the φCb5 phage, although it should be noted that the cloning experiments were conducted in *Escherichia coli*, and it is possible that in the natural host *Caulobacter crescentus* the ORF behaves differently. Interestingly, in the genome of phage MB there is also an ORF at a similar position within the replicase gene, almost with the same length and also in the +1 frame. The ORF-encoded protein contains a single transmembrane helix in the middle with almost 100% probability, and although it does not have a particular sequence identity with the φCb5 protein, this still adds some further support for the proposed location of the φCb5 lysis gene.

### 3.3.4. Secondary structure of the genome

Although many of the secondary structure elements in ssRNA phage genomes are highly variable, there are several that are conserved even in very distantly related phages. One of such is a characteristic stable stem-loop structure at the very 5′ end of the genome, believed to be necessary to ensure strand separation during replication (Beekwilder et al., 1996). A similar stem-loop is found near the 5′ end of the φCb5 genome (Figure 19A). Another characteristic feature is a hairpin around the initiation codon of the replicase gene that serves as a binding site for a coat protein dimer. There is indeed a hairpin around the putative initiation codon of the φCb5 replicase with the SD sequence comprising part of the loop and the 3’ part of the stem and the AUG codon nine nucleotides from the loop (Figure 19B). The structure does not resemble any of the known operator hairpins, and the stem-loop around the replicase initiation codon in phage MB also does not have any clear similarities to that of φCb5. As none of the conserved RNA-binding residues in other phages were identified in the φCb5 coat protein but instead RNA bases were observed between dimers in the capsid (Plevka et al., 2009), the RNA recognition mechanism of the φCb5 coat protein may be very different.

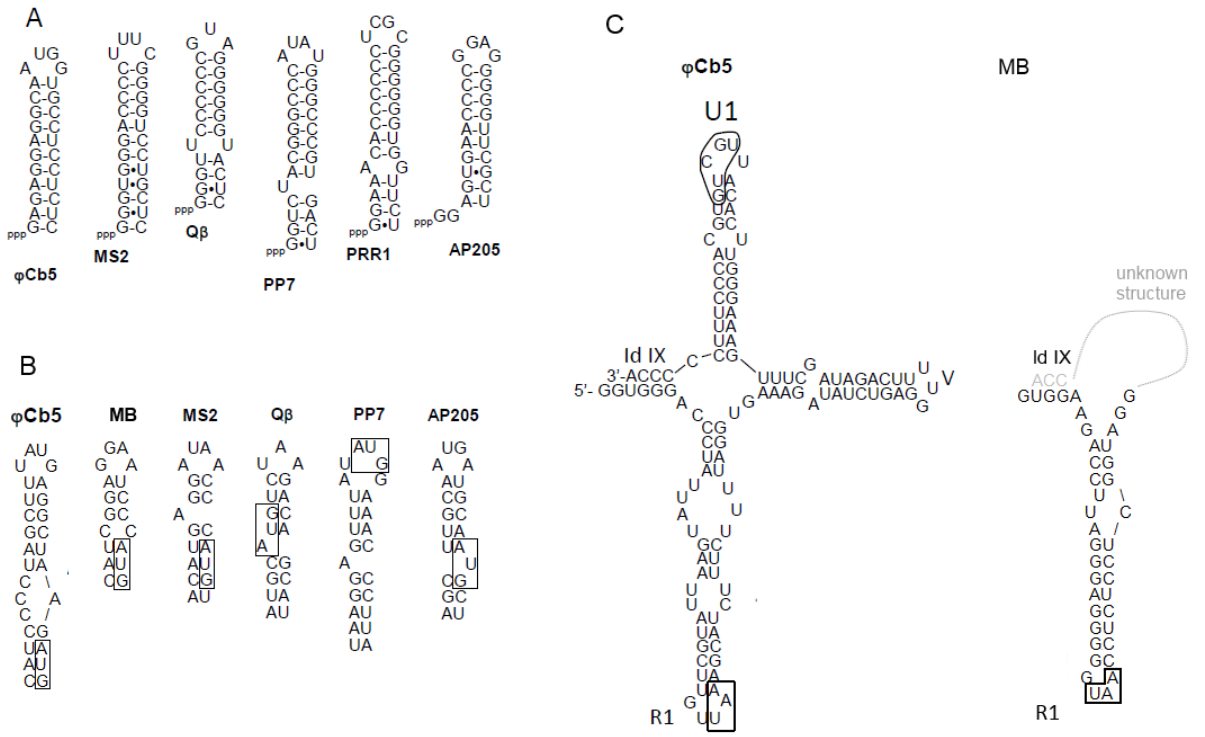


**Figure 19.** Comparison of RNA secondary structures in φCb5 genome to those in other ssRNA phages. A, hairpins at the 5' end of the genome. B, hairpins around the initiation codon of the replicase gene. C, the 3' domain of phage φCb5 compared to the R1 hairpin from phage MB.

The 3′ UTRs in ssRNA phage genomes fold into a separate domain composed of several stem-loops that are formed when the 3' terminus of the genome forms a long-distance interaction (ld IX) with a complementary nucleotide sequence upstream. φCb5 appears to have the simplest arrangement known so far with just three stem-loops (Figure 19C). In contrast to the other known phages where the 3' domain does not contain any protein-coding sequences, the 3' terminus of φCb5 RNA base-pairs with a sequence within the replicase ORF, and the 3' domain includes also the replicase termination codon-containing hairpin R1. In the deposited MB genome, the sequence of the 3' terminal part is missing, but from the available data it appears that the 3' domain also contains the R1 hairpin (Figure 19C). This lets to carefully suggest that the 3' domains might be similar in the φCb5/MB clade of ssRNA phages. In leviviruses and AP205, there is a conserved UGCUU sequence some 15 nucleotides from the 3′ end that in the case of phage Qβ has been shown to regulate replication via formation of a long-distance pseudoknot (Klovins and Van Duin, 1999). The last stem-loop of φCb5 RNA is somewhat similar to the U1 loops in other ssRNA phages, and contains a sequence UGCUG 16 nucleotides from the 3′ end. A sequence complementary to UGCUG is found in two positions in the replicase gene, but due to the insignificant sequence similarity of φCb5 and Qβ genomes, no conclusions can be made whether a long-distance interaction takes place in φCb5 as well.

## 3.4. Genome structure of RNA phage M

Historically, the ssRNA phages that infect *Escherichia coli* cells by adsorbing to the F plasmid-coded pili were the first isolates of the *Leviviridae* family (Davis et al., 1961; Loeb and Zinder, 1961), and to date these "male-specific" phages, with type species MS2 and Qβ, have been the most intensively studied and best characterized of this family. However, the F plasmid is just one of the many conjugative plasmids that are present in nature. These plasmids are often highly divergent from F and are most often grouped according to their mutual compatibility, or the ability to stably coexist in the same cell. In *Enterobacteriaceae*, the conjugative plasmids form more than twenty different incompatibility (Inc) groups which are denoted by capital Latin letters (Taylor et al., 2004). All of these plasmids encode conjugative pili, but the pilin subunits often share no recognizable sequence similarity.

Several ssRNA phages specific for conjugative pili other than that of plasmid F have been discovered. Phage PRR1 (Olsen and Thomas, 1973) which adsorbs specifically to IncP plasmid-encoded pili was the first such example, and later other phages specific for Inc group C (Sirgel et al., 1981), D (Coetzee et al., 1985a), H (Coetzee et al., 1985b; Nuttall et al., 1987), I (Coetzee et al., 1982), M (Coetzee et al., 1983) and T (Bradley et al., 1981) plasmids followed. Phage PRR1 has become somewhat of a prototype non-F plasmid-specific phage with its genome sequenced (Ruokoranta et al., 2006), capsids crystallized (Persson et al., 2008) and coat protein – operator structure determined (Persson et al., 2013). Phages C-1 (IncC-specific) and Hgal1 (IncH-specific) have also been sequenced (Kannoly et al., 2012), but no research has been done on the other plasmid-specific phages since their isolation.

The IncM plasmid-specific RNA phage M (Coetzee et al., 1983) was isolated from sewage in Pretoria, South Africa in the beginning of the 1980s. IncM plasmids have a broad host range, code for rigid pili and transfer efficiently only when bacteria are growing on solid media (Bradley et al., 1980). Likewise, phage M is able to propagate in different strains of *Escherichia*, *Salmonella*, *Klebsiella*, *Proteus* and *Serratia*, provided they contain an IncM plasmid. To see how phage M relates to the other known ssRNA phages and to obtain more insight into their evolution, I determined the genome sequence of phage M.

### 3.4.1. Overall structure of the genome and similarity to other phages

The genome of phage M is 3405 nucleotides long and follows the canonical *Leviviridae* genome organization with maturation, coat and replicase cistrons following each other in the 5'-3' direction (Figure 20). An unusual feature of the genome is that the lysis gene appears to be located in a different position than in other leviviruses, as discussed in the next section. It is also the smallest known *Leviviridae* genome to date,

about 60 nucleotides shorter than that of the F pili-specific phage GA (Inokuchi et al., 1986). The protein coding regions of phage M are of similar length to those of phage GA, with maturation and coat genes being a bit longer and replicase somewhat shorter; the greatest savings in M's genome come from the terminal untranslated regions, the 5' UTR being about 45 nucleotides and the 3' UTR about 20 nucleotides shorter.

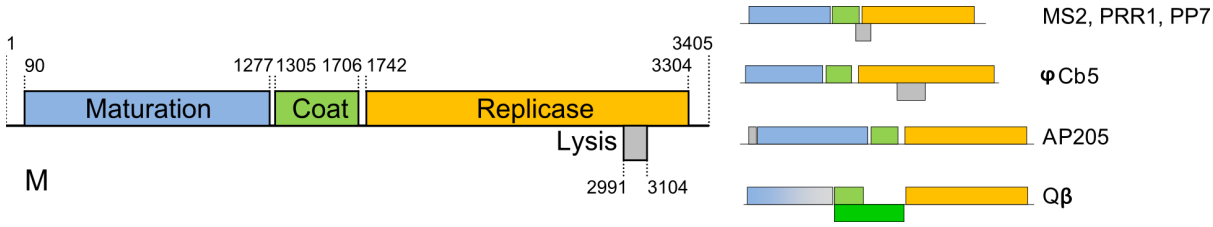


**Figure 20.** Genome organization of phage M. Start and end positions of phage genes are indicated. For comparison, the other known genome organizations of *Leviviridae* phages are represented on the right with genes color-coded as in the M genome. In phage Qβ, protein A1 (bright green) is an extended read-through variant of the coat protein and the lysis function is performed by the maturation protein.

The maturation protein of phage M is most similar to those of the other plasmid-specific RNA phages, but the sequence identity is only 24.5% to phage PRR1, around 22% to C-1, Hgal1, GA and MS2 and drops to 17% when compared to alloleviviruses SP and Qβ. The very low sequence identity of the maturation proteins is unsurprising due to the vast diversity of pili they have evolved to bind. The coat proteins are more conserved and here M groups clearly with phages PRR1, C-1 and Hgal1 with amino acid identities of 48-51%. The identity with F pili-specific phages is significantly lower and ranges from 27.1% for group II levivirus KU1 to 19% for group IV allolevivirus NL95. Notably, M coat protein shares 24.6% amino acids with that of the *Pseudomonas* phage PP7, which is the only plasmid-independent phage for which the sequences could be reasonably aligned. For replicase, the trend is similar as for the maturation protein: the replicase of phage M most resembles that of PRR1 with 41% amino acid identity, followed by other plasmid-dependent phages C-1, Hgal1, MS2 and GA (33-37% identity) and alloleviviruses (27-29% identity). Again, M replicase turns out to be more closely related to that of phage PP7 (25.5% identity) than to the other plasmid-independent phages AP205 and φCb5 (17.7 % identity).

### 3.4.2. Identification of the lysis gene

All members of the levivirus genus encode a short polypeptide that mediates cell lysis. In all of the known *Enterobacteria*-infecting leviviruses, the lysis gene overlaps with coat and replicase genes in a different reading frame and is translationally coupled with the coat gene (Van Duin and Tsareva, 2006). However, in the genome of phage M, no candidate ORFs at this location could be identified: in the +2 frame relative to the coat gene there are no termination codons until the start of replicase and in the +1

frame only a 17 amino acid long ORF that would encode a non-hydrophobic peptide is found.

Up to now, there have been two reported cases in the *Leviviridae* family where the lysis gene in is in a different location: *Acinetobacter* phage AP205 has a short lysis gene preceding the maturation gene (Klovins et al., 2002), while *Caulobacter* phage φCb5 codes for a longer, two-helix protein that completely overlaps with the replicase gene (Paper III). To test the possibility that phage M also has a non-canonical localization of the lysis gene, I utilized the fact that the pJET1.2 plasmid, where the cDNA copies of the genome were cloned for sequencing, contains a T7 promoter that can be used to transcribe the insert. Several clones with inserts in the correct orientation with respect to the T7 promoter were selected and transformed to a T7 polymerase-producing *E.coli* strain. When expression of the T7 polymerase was induced, a clone containing an approximately 1000 nucleotide long fragment spanning nucleotides 2098-3129 of the phage genome resulted in a clear cell lysis. Examination of this sequence located a likely candidate for the lysis gene between nucleotides 2991-3104 (Figure 21A). This was based on several criteria: (1) it was the only ORF in the fragment with a significant length (37 amino acids; the shortest known *Leviviridae* lysis protein is that of phage AP205 with 34 amino acids); (2) according to the TMHMM server (Krogh et al., 2001), the ORF-encoded protein was predicted to contain a transmembrane helix with over 95% probability; (3) although the ORF had an unusual initiation codon UUG, there was a rather strong Shine-Dalgarno (SD) sequence GAGG nine nucleotides upstream; (4) RNA secondary structure prediction using the RNAfold server (Hofacker, 2003) revealed that the initiation codon of the ORF is located on top of an AU-rich stem-loop that would presumably have sufficiently low thermodynamic stability to promote the initiation of

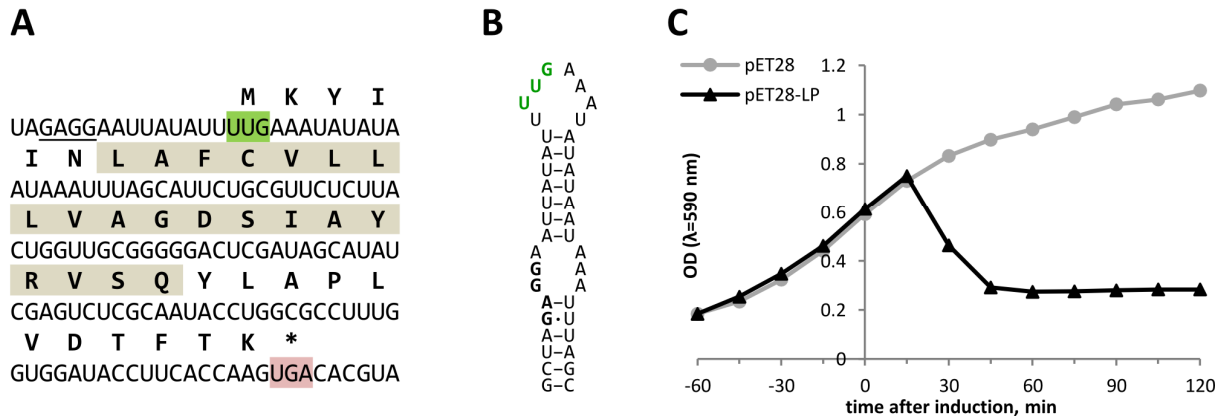


**Figure 21.** Lysis protein of phage M. A, the lysis gene. The Shine-Dalgarno sequence is underlined and initiation and termination codons are indicated by green and pink shading, respectively. The translated amino acid sequence is given above the RNA sequence and the putative transmembrane helix is shaded gray. B, an RNA hairpin around the initiation codon of the lysis gene. The initiation codon and the Shine-Dalgarno sequence are indicated. C, verification of the lysis gene. Growth of *E.coli* cells harboring either empty vector (pET28) or a plasmid with the cloned lysis gene (pET28-LP) before and after the induction of protein synthesis is shown.

translation (De Smit and Van Duin, 1990) (Figure 21B). To verify the lytic function of the gene, the ORF together with the original SD sequence and UUG initiation codon was cloned in an inducible protein expression vector. Induction resulted in almost complete cell lysis some 45 minutes after (Figure 21C), thus demonstrating that the approximately 150 nucleotide long stretch is sufficient to encode a functional lysis protein. The abovementioned evidence therefore lets to suggest with some confidence that this is the actual lysis gene of phage M.

### 3.4.3. Conserved RNA secondary structures

For ssRNA phages the secondary and tertiary structure of the genome is very important, and in many cases where nucleotide sequences from different phage genomes show no similarity, the secondary structures they fold into are nevertheless well preserved. One such example lies at the very 5' end of all of the sequenced ssRNA phage genomes, where there is a stable GC-rich hairpin that has been suggested to play an important role in phage RNA replication (Beekwilder et al., 1996). Phage M is no

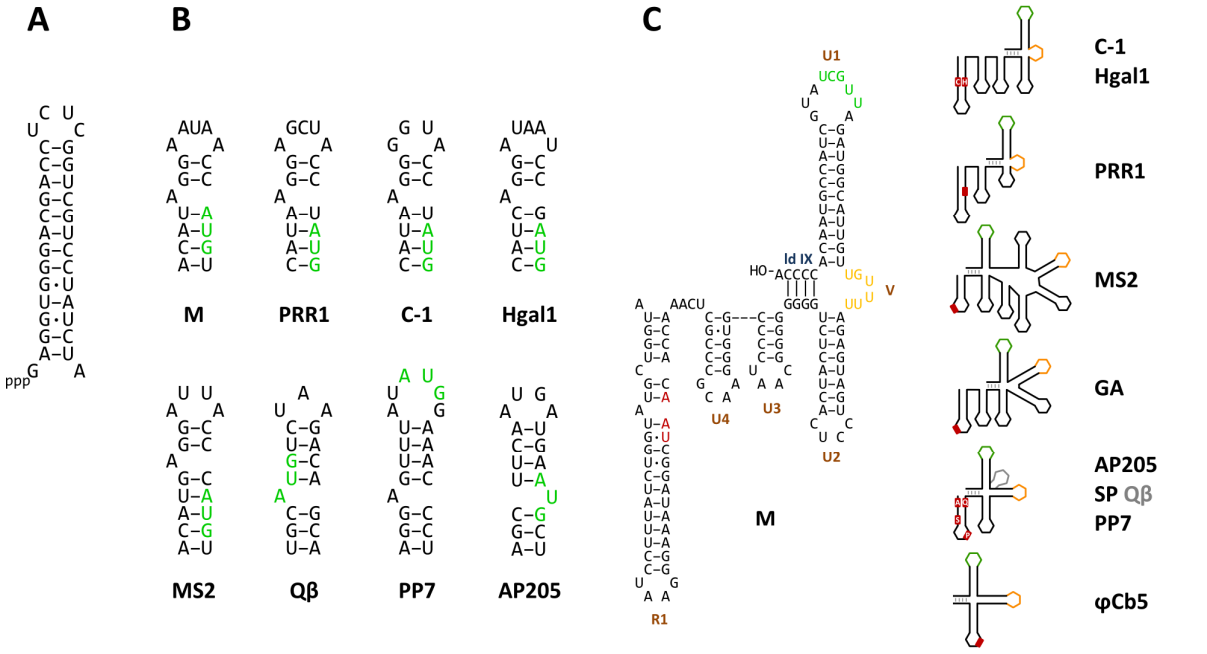


**Figure 22.** RNA secondary structures in M genome. A, a stable hairpin at the very 5' end of the genome important for phage RNA replication. B, the operator hairpin around the initiation codon of replicase. The analogous hairpins from other *Leviviridae* phages are shown for comparison. Start codons of the replicase gene are colored green. C, structure of the 3' untranslated region. The termination codon of replicase is colored dark red, the unpaired stretch corresponding to loop V or V2 in other phages in orange and the conserved nucleotide sequence in the loop of hairpin U1 that potentially forms a long-distance pseudoknot in green. On the right, schematic representations of the 3' UTRs from other phages based either on published data (Beekwilder et al., 1995; Klovins et al., 2002; Olsthoorn et al., 1995) or RNA secondary structure predictions are given for comparison. The 3' UTR of phage Qβ is closely similar to that of phage SP except for a short extra helix which is depicted in gray. The locations of replicase gene termination codons are represented as red boxes. RNA secondary structures were predicted by the RNAfold server (Hofacker, 2003).

exception (Figure 22A). Another important RNA structure is the translational operator of the replicase gene. When the operator hairpin of phage M is compared to those of other ssRNA phages, it is evident that it groups with the conjugative pili-dependent phages PRR1, C-1, Hgal1 and MS2 (Figure 22B). An adenine residue in the loop four nucleotides upstream of the replicase initiation codon and an unpaired purine residue in the stem which are critical for RNA-protein binding in phages MS2 (Carey et al., 1983b), GA (Gott et al., 1991) and PRR1 (Persson et al., 2013) are both preserved also in phage M, therefore the mechanism of interaction is probably similar.

It is also interesting to take a look at the 3' untranslated region of the phage genome. The configurations of 3' UTRs vary between different phages, but nevertheless some similarities exist. In all of the known *Leviviridae* phages, a long-distance interaction designated ld IX bridges the very 3' terminus with a complementary nucleotide stretch upstream, forming the 3' terminal domain (Beekwilder et al., 1995). The domain usually consists of at least three hairpins, denoted U1, U2 and V. In phage M, the 100-nucleotide-long 3' UTR is made up from four hairpins U4, U3, U2 and U1 (Figure 22C). In all ssRNA phages the 3'-terminal helix U1 has a remarkably conserved nucleotide sequence in the loop: UGCUU in phages as diverse as MS2, SP and AP205, UGCUG in φCb5 and CGCUC in PP7. In the case of Qβ, this loop forms a long-distance pseudoknot with a complementary sequence approximately 1200 nucleotides upstream that is essential for phage replication (Klovins and Van Duin, 1999). In phage M, the sequence of the U1 loop is AUUGCUAUG. It has not been experimentally verified that phages other than Qβ have the pseudoknot, but in M genome the sequence UUGCU in the loop could potentially basepair with a sequence AGCAA that is found in the replicase gene some 1215 nucleotides upstream. The other notable feature of the 3' domains, although less pronounced, is hairpin V (designated V2 in some phages) which in phages MS2, Qβ, SP and AP205 contains a large, adenine-rich loop. There is some evidence that in MS2 this might be one of the sites where the maturation protein binds to the RNA (Shiba and Suzuki, 1981). In phage φCb5, however, the candidate hairpin V lacks analogous features and in phages PRR1, C-1 and Hgal1 it does not seem to exist at all; instead, there is a stretch of unpaired nucleotides (UAUAAACA in PRR1, UAUA in Hgal1 and UUAAU in C-1) that connects hairpins U2 and U1 and might serve the same function as hairpin V in other phages. In phage M the situation is similar, but the loop sequence is UUUUGU and contains no adenine residues. When the overall structures of the 3' UTRs from different phages are compared (Figure 22C, right), it is evident that in the distantly related phages φCb5, AP205, PP7 and SP the 3' domain is remarkably simple with just three hairpins, while it is considerably expanded in the plasmid-specific leviviruses, culminating in seven hairpins in phage MS2. In this respect, phages M, C-1, Hgal1 and PRR1 form their own group where the 3' UTR adopts a characteristic fold of only two

hairpins between the ld IX, a stretch of unpaired nucleotides instead of hairpin V and one or two hairpins between the terminal replicase hairpin R1 and ld IX.

### 3.4.4. Phylogenetic relationship to other ssRNA phages

The high mutation rates and resulting sequence variability in RNA viruses makes reconstruction of their evolutionary history not a trivial task. Based on similarities between the maturation and replicase proteins, phage M seems more related to phage PRR1, while the coat protein sequences and structures of the 3' UTRs suggest that it might be closer to phages C-1 and Hgal1. To further address this question, I conducted a phylogenetic analysis of 15 representative *Leviviridae* phages using both the complete genome sequences and also the replicase protein sequences since the RNA-dependent RNA polymerases are the most conserved proteins of all positive-sense RNA viruses (Koonin and Dolja, 1993). Both trees (Figure 23) confirm that phage M is more closely related to the IncC, IncH and IncP than to the IncF plasmid-dependent phages but they show differences in the clustering of the non-F plasmid specific phages. Although phylogenetic analysis of the coat proteins (not shown) gives the same (M(C-1(Hgal1,PRR1))) clustering as the replicase, low bootstrap values for the IncC, IncH and IncP branches indicate that confidence in that particular branching order is not high and suggest that phages C-1, Hgal1 and PRR1 have radially diverged from a similar ancestral sequence. In both trees phage M represents a lineage that branched off early in the course of specialization on different plasmids after the separation of the IncF lineage had occurred but before the diversification on IncC, IncH and IncP plasmids took place.

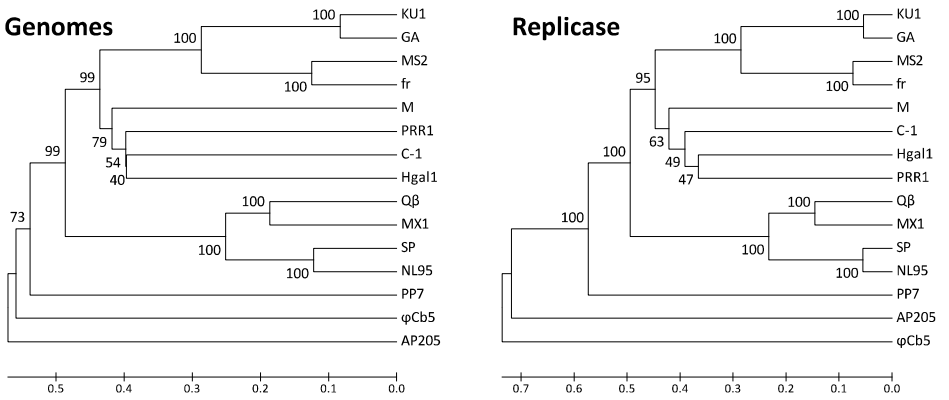


**Figure 23.** Phylogeny of RNA phages. The phylogenetic analysis was based on the complete genomic RNA sequences {left) and amino acid sequences of the replicase (right) which is the most conserved of all ssRNA phage proteins. Trees were constructed by the unweighted pair group method with arithmetic mean (UPGMA) and tested using the bootstrap method with 500 replicates. The bootstrap values are expressed as percentages next to the nodes.

# 4. DISCUSSION

In this thesis, I have explored the ssRNA phages in a variety of ways, both zooming-in to the very atomic details of how the phage proteins are built and interact, and stepping back to see the bigger picture of how whole genomes change and evolve. I think that together this has allowed me to better understand these viruses and provided material for interesting thoughts and speculations about their evolutionary history.

## 4.1. The A1 protein

The A1 protein is perhaps the most mysterious of all ssRNA phage proteins and that was one the main reasons why I chose to study it. When the three-dimensional structure of the read-through domain of the A1 protein was solved, it was of course intriguing to find a new protein fold that had never been observed before; at the same time, it also did not give immediate answers about its function but rather provided material for asking new questions, as sometimes happens in science. It has always been puzzling why the allolevivirus need two proteins for infectivity while the other phages suffice with just a single one. One possibility why the A1 protein might be necessary could be related to the fact that allolevivirus do not encode a separate lysis protein but the A2 protein mediates cell lysis (Karnik and Billeter, 1983; Winter and Gold, 1983). The lysis mechanism involves binding of A2 to the bacterial MurA protein that catalyzes a step in the murein biosynthesis pathway, and blocking its enzymatic activity. Since the maturation proteins of ssRNA phages also bind to bacterial pili, are specific RNA-recognizing proteins and mediate genome ejection and penetration, it is possible that yet another function is too much for the protein and in allolevivirus, the additional role of the A2 protein has led to the transfer of some of its other functions to the A1 protein. It is not known whether the A1 and A2 proteins form some kind of a complex in the capsid or act separately, but the long linker and flexible polyproline helix in the read-through domain might allow some of the A1 proteins in the capsid, which are presumably randomly distributed, to reach the A2 protein in the virion. Since in leviviruses, only the A protein – RNA complex enters the bacterial cell, leaving empty capsids outside, as structural components of the virion the A1 proteins might, for example, be involved in binding to the F pili. Another difference in leviviruses (at least the F pili-specific ones) and allolevivirus is that in the genome penetration step the leviviruses are dependent on the bacterial TraD protein while the allolevivirus are not (Achtman et al., 1971). Since at the same time allolevivirus have the A1 protein and leviviruses do not, it is possible that the A1 protein is somehow involved in this.

The origin of the A1 protein in allolevivirus is also enigmatic. Sequences encoding large protein domains with novel folds do not just suddenly appear, and the very fast

replication cycles and high mutation rates of the ssRNA phages tend to rapidly remove any disadvantageous sequences from their genomes. How, then, the A1 protein did appear? As already pointed out by Hofstetter (Hofstetter et al., 1974), there are two main possibilities, either that the A1 protein represents a surviving ancestral lineage and the levivirus-type phages have evolved a way to live without the A1 protein and deleted in from their genome; or vice versa, that the ancestral sequence was levivirus-like and the A1 protein appeared during genome expansion. For a number of reasons, the second possibility seems to be the more probable one. Up to now, only allolevivirus that are F pili-specific have been isolated while all of the sequenced other conjugative pili-specific phages and those infecting distinct bacterial genera have turned out to be exclusively levivirus-like. This would then put the loss of the A1 protein back to a very long time ago, before ssRNA phage diversification on all of the different bacterial hosts took place, as independent riddance of the protein in each of the lineages seems less probable. Of course, it might well be just a matter of looking and since only a single representative from each of the non-F pili-specific phage groups has been sequenced, it is completely possible that a non-F pili-specific allolevivirus might be isolated some day. Until then, it seems more probable that the A1 protein evolved once in a separate branch of the ssRNA phages. The appearance of the read-through domain by genome expansion seems easier to explain as it would require a single insertion event and there is evidence that Qβ and MS2 replicases can cause RNA recombination (Biebricher and Luce, 1992; Olsthoorn and van Duin, 1996). Gradual growth of the extension by incremental nucleotide additions to the C-terminus of the coat protein gene seems implausible, since it would have taken a long time for the extension to bring any benefit, and the phage would quickly optimize the genome by deleting it. For an entirely new gene to evolve, an in-frame insertion of a longer protein-coding piece of RNA at the end of the coat protein gene, either by duplication of a sequence from phage RNA or by recombination with some bacterial mRNA, seems like a better option as it would have been harder for the phage to quickly get rid of the entire sequence. Since Qβ coat protein molecules with even short C-terminal extensions appear to be unable to form normal particles by themselves (Vasiljeva et al., 1998), the second necessary event would have needed to be the appearance of a leaky termination codon at the end of the coat protein, which, given the high error rate of ssRNA phage replicases, does not seem unlikely at all. This way, only a few copies of the extension per capsid would have been present that would not impair their assembly, and even if initially the extension brought no benefit for the phage, this might have given some time for it to evolve to something useful.

## 4.2. The coat protein – RNA interaction

The more one studies the ssRNA phages, the more it becomes clear that to refer to the "structure of the ssRNA phages" just as the three-dimensional structure of their proteins means to tell just part of the story. Function of three out of the four phage proteins – replicase, maturation and coat – is intricately linked with specific RNA structures that they recognize and bind to at some point during the viral life cycle. As a result, the three-dimensional organization of the phage genome is just as important as that of the proteins, and over time, the structure of phage proteins and RNA co-evolve to fulfill their function together. A prime example for this is the specific interaction between the coat protein and the RNA operator of the replicase gene, where changes at the nucleotide level in the RNA hairpin have to be complemented with corresponding changes of the RNA-binding surface of the coat protein to maintain the interaction.

Including the Qβ structure presented in this thesis, the three-dimensional structures of coat protein-operator complexes from four different ssRNA phages have now been determined, and despite some profound differences, a number of common themes that all of the complexes share can be recognized. An essential feature for all of the phage operator structures appears to be that some of the nucleotide bases in the loop stack with bases in the helical stem. In MS2, PRR1 and Qβ, the nucleotide stack further extends to the aromatic side chain of a conserved tyrosine residue, whereas in PP7, a van der Waals interaction with a valine residue takes place. The aromatic stacking is likely important for constraining the loop nucleotides in an appropriate position to bind to the protein and is therefore conserved during evolution. Extensive structural studies of MS2 coat protein complexes with operator variants also showed that the stacking itself and not the identity of the bases is of the greatest importance for the protein – RNA interaction to take place (Grahn et al., 2000, 2001; Helgstrand et al., 2002). Another RNA recognition strategy shared between all phages involves sequence-specific interactions between nucleotide bases and the RNA-binding surface of the protein. None of these interactions are universally conserved, but in all of the studied phages, binding of an adenine base in the loop into an adenine-recognition pocket in the coat protein is critical for the operator-coat protein interaction, although the pockets of MS2/PRR1/Qβ and those of PP7 are very different. As the coat protein dimer and the RNA-binding surface have a twofold rotational symmetry, there are two symmetrical adenine-binding pockets in the protein, and in MS2, PRR1 and PP7 two adenine bases from the operator hairpin bind to those pockets, while in Qβ, only a single pocket is occupied. The importance of other base-specific contacts for maintaining the protein – RNA interaction varies. For the PP7 coat protein, sequence-specific interactions are fundamental and besides the two adenines, two more bases make contact with the protein, while the sugar-phosphate backbone does not contribute significantly to the

binding. In MS2, in total three bases specifically bind to the protein; however, the RNA backbone also makes significant interactions with the protein in the stretch between the bulged adenosine and the loop. In Qβ, the adenosine in the hairpin loop is the only nucleotide that makes base-specific contacts with the coat protein while the majority of interactions between the protein and RNA involve the sugar-phosphate backbone. Despite the smaller amount of sequence-specific information, the Qβ coat protein is still able to discriminate its cognate operator, which demonstrates how co-evolution of the protein and RNA can result in a highly specific interaction based on the conformation of the phosphate backbone rather than numerous sequence-specific contacts with bases. The three very different modes of accommodating an unpaired base in PP7, MS2 and Qβ further demonstrate the notable flexibility of protein-RNA interactions in the ssRNA phages.

Nevertheless, the overall binding mode of the Qβ coat protein to its operator is clearly similar to those of MS2 and PRR1, which suggests that this particular mechanism is conserved among the conjugative plasmid-dependent *Leviviridae* phages. It has been shown that just a couple of amino acid substitutions can result in MS2 coat protein mutants able to bind the Qβ operator much better than the wild-type (Spingola and Peabody, 1997) and vice versa (Lim et al., 1996), and that the MS2 coat protein can bind RNA hairpins with three-nucleotide loops or no bulged adenosine (Hirao et al., 1999). While it is impossible to know whether the ancestral coat protein and operator were MS2-like, Qβ-like or something intermediate, it is not hard to envision a step-by-step transition between the two types of protein-RNA interactions while maintaining the binding.

Outside the plasmid-specific ssRNA phage group, the coat protein of *Pseudomonas* phage PP7 is the only one that still has some traces of sequence identity with MS2 and Qβ, but its RNA recognition mechanism is very distinct. When the PP7 coat protein sequence is structurally aligned with those of the plasmid-specific ssRNA phages (Persson et al., 2008), only five residues are universally conserved, and three of those are involved in formation of the adenine-binding pocket in the plasmid-specific phages. Only one of the conserved pocket-forming residues participates in RNA binding in PP7, while the pockets themselves are non-functional because a valine residue that forms one side of the pocket in the plasmid-specific phages is replaced by an arginine in PP7. The other two of the conserved pocket-forming residues do not seem to have a particular purpose in PP7 and it is puzzling why those have been preserved as well. It might be just a coincidence, but it could be speculated that at the point when the lineages leading to the modern *Pseudomonas*-infecting and plasmid-specific phages separated, the ancestral PP7-like phage employed an adenine-recognition mode similar to the one the plasmid-specific phages use today, and only later the PP7-like lineage

switched to the other RNA-binding mode. The fact that most of the "old" adenine-binding pocket is still present would imply that this was a relatively recent event, but the vastly different RNA-binding mechanisms seem to contradict that idea. On the other hand, the very high mutation rates of ssRNA phages can result in rapid changes under the right selective conditions, and the evolution of a new RNA-binding mechanism could be facilitated by the fact that the specific coat protein – RNA interaction does not appear to be critical in the life cycle of the phage. This was best demonstrated by an MS2 pseudorevertant that lacks the operator sequence completely but still produces virions with titers only fivefold lower than the wild type (Licis et al., 2000). Thus even if a mutation completely disrupts the operator binding, the phage, although at a fitness disadvantage, still remains viable which might give an opportunity for novel RNA-binding mechanisms to evolve.

The question of the importance of the coat protein – operator interaction is also raised when considering phages distantly related to PP7 and the plasmid-specific phages. Up to now, three such phages have been sequenced that are remarkably different from the rest: *Acinetobacter* phage AP205, *Caulobacter* phage φCb5 and the marine RNA phage MB. For phage AP205, a putative operator hairpin at the beginning of the replicase gene has been identified, which unlike other phages, has a bulged uridine located on the 3' side of the stem (Klovins et al., 2002). Hairpin-like structures at the respective position could be identified also in the φCb5 and MB genomes. Regarding the φCb5 coat protein and the putative operator hairpin, a standard RNA-binding assay failed to show an interaction between the two (Paper III), raising the question whether such interaction exists at all. The three-dimensional structure of the φCb5 virion revealed strong electron density for RNA bases between the dimers, which indicates a very different RNA binding mechanism (Plevka et al., 2009). Since all of the known distantly related phages nonetheless have a hairpin around the translation initiation site of the replicase gene, a specific protein – RNA interaction cannot be excluded, and further studies of protein-RNA interactions of the ssRNA phages have the potential to provide even more discoveries about the evolution of protein and RNA structure in these viruses.

## 4.3. The lysis genes

After the genomes of phages φCb5 and M had been sequenced, a somewhat unexpected feature that emerged was the different locations of the lysis gene in these genomes. Before these studies, two possible positions of the lysis gene had been described – the "classic" one in leviviruses where it overlaps with the coat and replicase ORFs in a different reading frame, and the one in phage AP205 where it is located at the very 5' end of the genome preceding the maturation gene. Both in φCb5 and M, the lysis

genes completely overlap the replicase gene in a different reading frame but are otherwise very different. The putative φCb5 lysis protein is 135 residues long and contains two predicted transmembrane helices while that of phage M is very short with just 37 residues and a single helix. Because of the great genetic distance to other phages, a novel location of the φCb5 lysis gene is not that surprising, but it is quite unexpected to find such phenomenon also in phage M that has several close relatives in the plasmid-specific levivirus group all with "canonically" located lysis genes. This suggests that the location of the lysis gene at this position is probably limited to the IncM plasmid-specific leviviruses or even to a smaller subgroup of these phages. Since M is the only IncM plasmid-specific RNA phage that has been isolated so far, it is not possible to address this question presently. The lysis proteins encoded by the differently located ORFs are all predicted to contain a transmembrane helix but share no sequence similarity. Thus it seems that the ssRNA phages have arrived at the same lysis mechanism independently and that it is apparently relatively easy for a short gene encoding a transmembrane helix that causes cell lysis to appear by random mutations.

For a functional lysis gene, not only the gene product has to be able to lyse the cells, but also the timing of its production has to be right in order not to destroy the cells prematurely. In MS2, the lysis gene cannot be translated independently but ribosomal termination of the coat gene occasionally leads to reinitiation at the upstream lysis start codon, resulting in slow accumulation of the lysis protein and cell lysis late in the infection cycle. Regulation of the other three types of lysis genes has not been studied, but, unlike MS2, they all have Shine-Dalgarno sequences upstream the coding sequence. Although in φCb5 and M, translational coupling of the replicase and lysis genes cannot be excluded, this seems unlikely since the replicase is a characteristic early gene product while the lysis protein is a late one. In both φCb5 and M, the initiation codon of the lysis gene is UUG that is rarely used in bacteria and is estimated to be about 10% as efficient as an AUG (Barrick et al., 1994). This might allow the phage to delay the accumulation of the lysis protein until an appropriate time, but the secondary structure of the genome might also play a role in the regulation. In phage AP205, the lysis gene has a strong Shine-Dalgarno sequence and an AUG initiation codon, and in this case, folding of the 5' end of the genome might control its translation in a similar way as the synthesis of the maturation protein is regulated in MS2. Finally, it might be that not all of the lysis proteins are equally effective, and some of them might be necessary in larger amounts than others, which in turn would require different regulation strategies.

## 4.4. Evolutionary history of the *Leviviridae* family

It is undoubtedly very intriguing to try to reconstruct the evolutionary history of the *Leviviridae* family, but it is not an easy task due to the high RNA mutation rate that

can sometimes lead to basically non-existent sequence identity. When the AP205 genome was sequenced, an evolutionary tree was proposed based on structures of the conserved 3' UTRs that unexpectedly placed MS2 closer to PP7 and AP205 than to Qβ (Klovins et al., 2002). However, by combining data from phylogenetic analyses based on protein and RNA alignments, considering the different genome organizations and conservation of RNA secondary structure elements, I have made an alternative attempt to reconstruct the chain of events that might have led to the diversity of the ssRNA phages we see today, presented in Figure 24.

Phylogenetic analysis places phages PP7, AP205 and φCb5 increasingly further apart from the plasmid-specific ones, and these phages also have very simple 3' UTRs with the smallest in φCb5 with just three stem-loops, therefore it seems reasonable to suggest that the ancestral phage presumably also had a very short, φCb5-like 3' UTR. Although the core genome organization of all of the known ssRNA phages is the same with the maturation-coat-replicase genes in the 3' to 5' direction, the lysis genes in PP7, AP205 and φCb5 are located differently in each phage and seem to have arisen independently from each other. This does not allow to conclude whether the predecessor of PP7, AP205 and φCb5 had the lysis gene in any of these positions or yet somewhere else. Although it has previously been suggested that the ancestral phage might have had a bi-functional maturation/lysis protein like the alloleviviruses have today (Klovins et al., 2002), this assumption is speculative, and, as I have discussed above, it might be that a maturation protein cannot perform the additional lysis function without handing some of its other functions over to the A1 protein. The most probable time when the "classic" lysis gene position overlapping the coat and replicase ORFs might have appeared is after the *Caulobacter* and *Acinetobacter*-specific phage lineages had separated. This also appears to be an appropriate time when the MS2/Qβ-like adenine recognition pockets and the coat-protein operator interaction might have evolved since all phages further down the tree have conserved residues that make up the pocket. The lineage leading to phage PP7, however, evolved a distinct RNA binding mechanism along the way.

Now it should be noted that although all *Leviviridae* phages use pili for attachment, there is a marked difference between the types of pili they utilize. The type IV pili used by phages AP205, φCb5 and PP7 are produced via a genome-encoded type II secretion pathway (Peabody et al., 2003), whereas the plasmid-borne conjugative pili that the other phages utilize belong to a type IV secretion system (Lawley et al., 2003). Both systems share some functional similarities, like a retractable pilus and a membrane pore, but are thought to have evolved independently (Hazes and Frost, 2008). Therefore a jump from one to the other type of pili had to occur at some point in the *Leviviridae* history. The phylogenetic analysis suggests that the ancestral phage infected cells via

type IV pili, like phages AP205, φCb5 and PP7 are doing today and a PP7-like virus then might have evolved the ability to bind to some kind of conjugative pili and still sustain infectivity. Since it was presumably after this point when the F pili-specific levivirus and allolevivirus lineages separated, it is possible that the first conjugative pili that the ssRNA phages could bind were some kind of proto-F pili, however, the levivirus and allolevivirus maturation protein sequences are so different that these phages might have arrived at the ability to bind to F pili via independent evolutionary paths.

After the phages were able to use conjugative plasmid-encoded pili for infection, the next event in the diversification of the *Leviviridae* family appears to be the separation of a lineage that led to the emergence of alloleviviruses. In this branch some rather profound changes took place like the evolution of the lytic function of the maturation protein and the appearance of the A1 protein. In the meantime, in the other branch that led to the leviviruses an expansion of the 3' UTR took place and the further diversification of the phages seems to be best explained if it is presumed that two extra hairpins prior to the 3' domain appeared at this time. From this configuration, a slight rearrangement leads to the 3' UTR seen in phage GA and a somewhat different rearrangement and addition of two hairpins to that of phage MS2. In the branch leading to the IncM/C/H/P-specific phages, an early event appears to be the replacement of a hairpin in the 3' domain with a loop. The IncM plasmid-specific phage lineage was the first to separate from this ancestor, and lost and re-invented the lysis gene at some point. There is not a clear branching order of the IncC, IncH and IncP plasmid-specific phage lineages, and they appear to have radially diverged from a similar ancestral sequence. These phages are rather similar to each other, except that IncP plasmid-specific phage PRR1 has a slightly smaller 3' UTR that can be explained by a loss of a hairpin prior to the 3' domain.

Although the proposed tree is admittedly speculative, probably has errors and might not have the lowest number of evolutionary events possible, I think it fits both the sequence alignment and secondary structure conservation data reasonably well, and as more *Leviviridae* genomes are sequenced and other research is done, I am curious to see of what of this tree holds true and what does not.

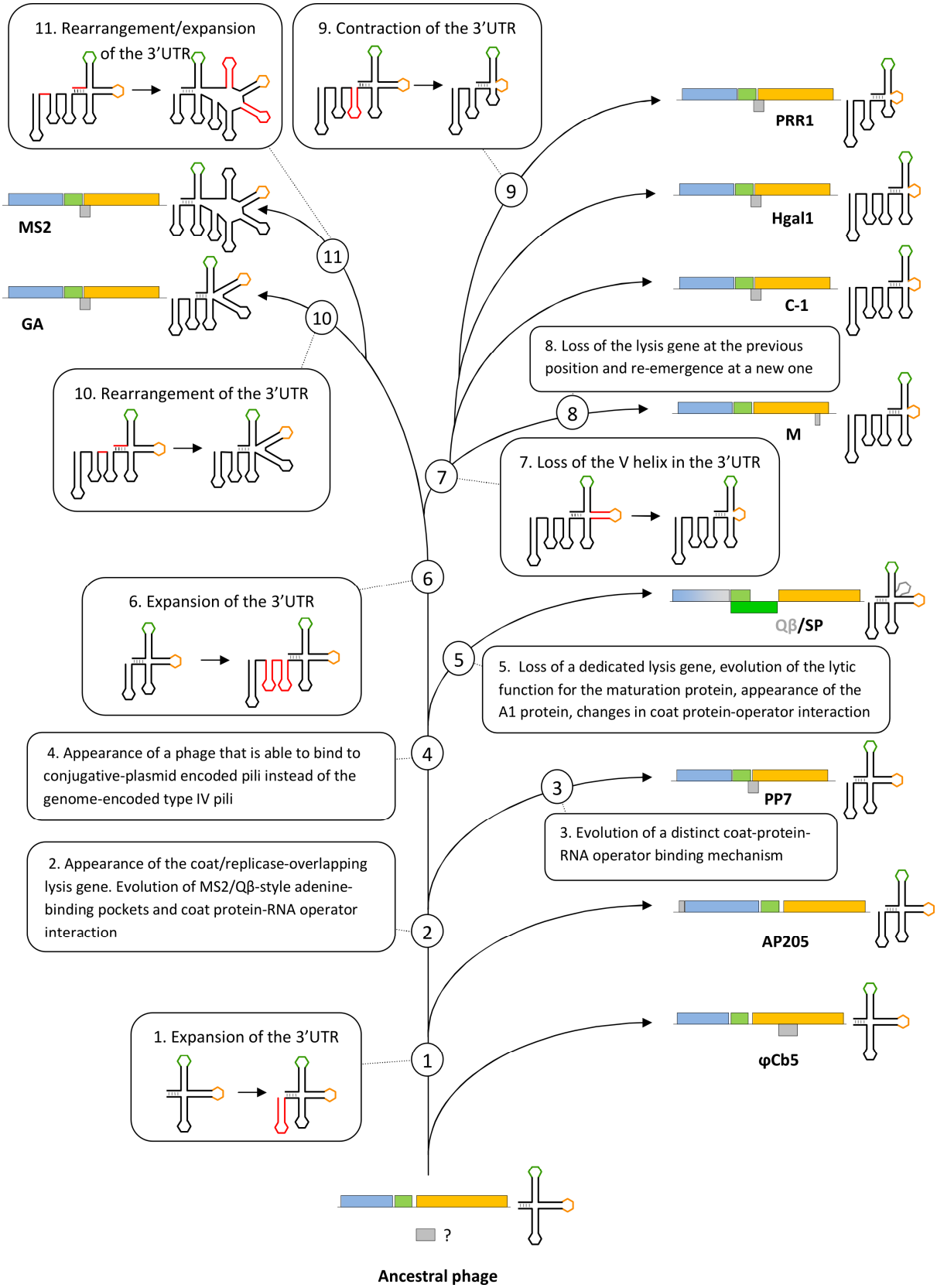


**Figure 24.** Proposed evolutionary tree of the *Leviviridae* family. Maturation genes are represented in blue, coat in green, A1 in bright green and replicase in orange. The 3' UTRs are not drawn to scale, and branch lengths do not represent actual evolutionary distances but just the order of events.

# 5. PROSPECTS FOR FUTURE WORK

Some 55 years after the first-ever ssRNA phage was isolated, the studies on the "small RNA phages", as they are sometimes called, has shown that they might be small in appearance but not simple in how they are built and how they function. Although there have been significant advances in both structural and genome research on these viruses in the recent years, there is still a lot of unknown about them and no shortage of things to do.

From a structural biologist's point of view, the biggest unsolved mystery of the ssRNA phages is the structure of the maturation protein. The protein sequence is not similar to any other proteins, and the structure is also probably unique to these phages. The infection and especially genome penetration stages are perhaps the biggest "white spots" in ssRNA phage biology, and a high-resolution structure of the maturation protein has the potential to provide some answers about how it recognizes and binds RNA, incorporates into the capsid, binds to the pilus and guides the genome into the cell. The maturation proteins are notoriously hard to work with as their usual state *in vitro* outside of the capsid appears to be an insoluble precipitate, but recent advances in our laboratory with the Qβ A2 protein have provided some hopes that determination of the three-dimensional structure of the maturation protein might be achievable after all. Another structure that has eluded determination is that of phage AP205 capsids. The AP205 coat protein sequence is not similar to those of other ssRNA phages, and although the overall fold of the protein is probably the same, some interesting and unexpected features are almost certainly present there. Although AP205 capsids have been crystallized, the crystals always diffracted very poorly and were unsuitable for structure determination. Currently, work in our laboratory involving AP205 coat protein mutants is ongoing that has a potential to provide some structural information about this phage. Finally, probably the highest achievement in ssRNA phage structural biology would be the determination of the high-resolution structure of the entire phage genome and understanding of how it interacts with the phage proteins. Although crystallography might not seem as the best possible method for studying these presumably flexible and structurally non-homogeneous molecules, binding of phage proteins such as replicase or the maturation protein might constrain them and the idea of crystallizing the genome might not seem absolutely crazy. Alternatively, recent advances in cryoEM have allowed the determination of structures at an increasing resolution and speed, and currently this is perhaps the best option to study the structure of phage genomes.

Some twenty years ago, the F pili-specific *E.coli* phages were the only ones that had been sequenced and those were neatly classified in two genera, leviviruses and

allolevivirus, each further consisting of two serological groups. During the last two decades, genome sequences of four non-F plasmid-specific ssRNA phages as well as those of the increasingly distinct PP7, AP205 and φCb5 viruses have been determined. This has considerably expanded the horizon of what we know about ssRNA phage evolution, but during this work it increasingly felt like this might still be only the tip of an iceberg of the entire diversity of the ssRNA phages that are out there. Regarding the conjugative plasmid-specific phages, of the many incompatibility groups only five have been covered, and except for the F pili-specific phages, only a single representative has been sequenced from each of those groups. There is no reason to doubt that like with the F pili-specific phages, several genogroups exist also within the phages that have specialized on other conjugative plasmids. The same is almost certainly true for phages infecting the *Caulobacter*, *Acinetobacter* and *Pseudomonas* hosts, and ssRNA phages specific for other bacterial genera are probably out there, too. It would be even more exciting to find an RNA phage infecting bacteria outside the *Proteobacteria* group altogether which might provide invaluable information about the evolution of RNA bacteriophages and RNA viruses in general. Therefore in further studies I would be delighted also to go out and search for new RNA phages and decipher what more secrets these fascinating viruses have concealed inside their capsids.

# CONCLUSIONS

1. The read-through part of the Qβ A1 protein is a separate domain that can be recombinantly produced, purified and crystallized.
2. The three-dimensional structure of the Qβ A1 protein is not similar to any other known protein. The domain consists of a five-stranded β-barrel, a β-haipin and several short α- and $3_{10}$-helices. There is a long polyproline type II helix at the N-terminal part of the domain.
3. In the allolevivirus A1 read-through domains, there are several conserved amino acid stretches around residues 207-219 and 228-238, as well as in the polyproline helix and at the C-terminus. These regions map on one side of the read-through domain, suggesting that it is the most important for performing its function.
4. The overall binding mode of Qβ coat protein to the RNA operator of the replicase gene is similar to that of the widely studied phage MS2.
5. An adenine base in the Qβ operator hairpin loop makes sequence-specific contacts with the coat protein. The Qβ coat protein uses a stacking interaction with a tyrosine side chain to accommodate a bulged adenine base in the hairpin stem and the EF loops of the protein make contact with the lower part of the RNA stem.
6. *Caulobacter* phage φCb5 and IncM plasmid-specific phage M have a levivirus-like core genome organization of maturation, coat and replicase genes in the 5' to 3' direction, but differently located lysis genes that completely overlap with the replicase gene in another reading frame. Both lysis genes encode proteins with predicted transmembrane helices like those of other leviviruses.
7. The genome of bacteriophage φCb5 has very low sequence identity to the other known RNA phages and the simplest known 3' untranslated region with just three hairpins.
8. Bacteriophage M is closely related to the other known leviviruses, but has an atypical location of the lysis gene. Phage M is more similar to IncP, IncC and IncH, but not IncF plasmid-specific leviviruses.

# THESIS FOR DEFENSE

1. The read-through domain of bacteriophage Qβ A1 protein adopts a protein fold not seen in other proteins.
2. The coat protein of bacteriophage Qβ recognizes an RNA hairpin at the beginning of the replicase gene based primarily on RNA backbone conformation instead of many sequence-specific interactions.
3. Lysis genes encoding small proteins with transmembrane helices have arisen independently several times in the *Leviviridae* family.
4. The IncM plasmid-specific RNA phage lineage branched off from other leviviruses early in the course of RNA phage specialization on different conjugative pili.
5. The modern *Caulobacter*-infecting RNA phage lineage represents the oldest known separation event from the common RNA phage ancestor.

# ACKNOWLEDGEMENTS

The project “Support for Doctoral Studies at University of Latvia” from the European Social Fund provided the financial support for much of the time I was working on my thesis. It offered a great deal of stability and allowed me to focus on my work, especially during times when the situation looked really bleak for Latvian science. Without it, the life would have been much harder indeed.

I would first and foremost like to thank my supervisor Kaspars Tārs for giving me the opportunity to work in his then-newly founded lab at BMC and for giving me a great freedom to pursue the studies of the RNA phages that I found so fascinating. I really appreciate all your interest and support for my ideas and I will never forget those synchrotron trips together and the first structure I solved. I am also very grateful to Andris Kazāks for his constant support, all of the protein purification you did for me as well as the opportunity to regularly “borrow” all of your restriction enzymes and other reagents which we did not have. I also thank Ināra Akopjana for your help with growing bacteria and the always-competent cells you prepared as well as the rest of the “3rd floor people” for providing everything one would ever need for growing bacteria, producing proteins and working with phages. I am also thankful to all of the other students in our lab, past and present, for all of your help, the discussions we had and for making our lab such a nice place to work. And finally, I am very grateful to my family for their patience, understanding and never-ending support through all of these years that took me to finish my thesis.

# REFERENCES


Achtman, M., Willetts, N., and Clark, A.J. (1971). Beginning a genetic analysis of conjugational transfer determined by the F factor in Escherichia coli by isolation and characterization of transfer-deficient mutants. J. Bacteriol. *106*, 529–538.

Adhin, M.R., and Van Duin, J. (1990). Scanning model for translational reinitiation in eubacteria. J. Mol. Biol. *213*, 811–818.

Argetsinger, J.E., and Gussin, G.N. (1966). Intact ribonucleic acid from defective particles of bacteriophage R17. J. Mol. Biol. *21*, 421–434.

August, J.T., Cooper, S., Shapiro, L., and Zinder, N.D. (1963). RNA Phage Induced RNA Polymerase. Cold Spring Harb. Symp. Quant. Biol. *28*, 95–97.

Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence, C.E., Gold, L., and Stormo, G.D. (1994). Quantitative analysis of ribosome binding sites in E.coli. Nucleic Acids Res. *22*, 1287–1295.

Beckett, D., Wu, H.N., and Uhlenbeck, O.C. (1988). Roles of operator and non-operator RNA sequences in bacteriophage R17 capsid assembly. J. Mol. Biol. *204*, 939–947.

Beekwilder, J., Nieuwenhuizen, R., Poot, R., and Van Duin, J. (1996). Secondary structure model for the first three domains of Qβ RNA. Control of A-protein synthesis. J. Mol. Biol. *256*, 8–19.

Beekwilder, M.J., Nieuwenhuizen, R., and Van Duin, J. (1995). Secondary structure model for the last two domains of single-stranded RNA phage Qβ. J. Mol. Biol. *247*, 903–917.

Bendis, I., and Shapiro, L. (1970). Properties of Caulobacter ribonucleic acid bacteriophage phiCb5. J. Virol. *6*, 847–854.

Beremand, M.N., and Blumenthal, T. (1979). Overlapping genes in RNA phage: a new protein implicated in lysis. Cell *18*, 257–266.

Berisio, R., Loguercio, S., De Simone, A., Zagari, A., and Vitagliano, L. (2006). Polyproline helices in protein structures: A statistical survey. Protein Pept. Lett. *13*, 847–854.

Berkhout, B., De Smit, M.H., Spanjaard, R.A., Blom, T., and Van Duin, J. (1985). The amino-terminal half of the MS2-coded lysis protein is dispensable for function:implications for our understanding of coding region overlaps. EMBO J. *4*, 3315–3320.

Bernhardt, T.G., Roof, W.D., and Young, R. (2000). Genetic evidence that the bacteriophage φX174 lysis protein inhibits cell wall synthesis. Proc. Natl. Acad. Sci. U. S. A. *97*, 4297–4302.

Bernhardt, T.G., Wang, I.N., Struck, D.K., and Young, R. (2001). A protein antibiotic in the phage Qβ virion: diversity in lysis targets. Science *292*, 2326–2329.

Biebricher, C.K., and Luce, R. (1992). In vitro recombination and terminal elongation of RNA by Qβ replicase. EMBO J. *11*, 5129–5135.

Blumenthal, T., Landers, T.A., and Weber, K. (1972). Bacteriophage Qβ replicase contains the protein biosynthesis elongation factors EF Tu and EF Ts. Proc. Natl. Acad. Sci. U. S. A. *69*, 1313–1317.

Bollback, J.P., and Huelsenbeck, J.P. (2001). Phylogeny, genome evolution, and host specificity of single-stranded rna bacteriophage (family Leviviridae). J. Mol. Evol. *52*, 117–128.

Bradley, D.E. (1966). The structure and infective process of a Pseudomonas aeruginosa bacteriophage containing ribonucleic acid. J. Gen. Microbiol. *45*, 83–96.

Bradley, D.E., Taylor, D.E., and Cohen, D.R. (1980). Specification of surface mating systems among conjugative drug resistance plasmids in Escherichia coli K-12. J. Bacteriol. *143*, 1466–1470.

Bradley, D.E., Coetzee, J.N., Bothma, T., and Hedges, R.W. (1981). Phage t: a group T plasmid-dependent bacteriophage. J. Gen. Microbiol. *126*, 397–403.

Butcher, S.J., Grimes, J.M., Makeyev, E. V, Bamford, D.H., and Stuart, D.I. (2001). A mechanism for initiating RNA-dependent RNA polymerization. Nature *410*, 235–240.

Carey, J., Cameron, V., De Haseth, P.L., and Uhlenbeck, O.C. (1983a). Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. Biochemistry *22*, 2601–2610.

Carey, J., Lowary, P.T., and Uhlenbeck, O.C. (1983b). Interaction of R17 coat protein with synthetic variants of its ribonucleic acid binding site. Biochemistry *22*, 4723–4730.

Caspar, D.L.D., and Klug, A. (1962). Physical principles in the construction of regular viruses. Cold Spring Harb. Symp. Quant. Biol. *27*, 1–24.

Cavalli, L.L., Lederberg, J., and Lederberg, E.M. (1953). An Infective Factor Controlling Sex Compatibility in Bacterium coli. J. Gen. Microbiol. *8*, 89–103.

Chao, J.A., Patskovsky, Y., Almo, S.C., and Singer, R.H. (2008). Structural basis for the coevolution of a viral RNA-protein complex. Nat. Struct. Mol. Biol. *15*, 103–105.

Chetverin, A.B., and Spirin, A.S. (1995). Replicable RNA vectors: prospects for cell-free gene amplification, expression, and cloning. Prog. Nucleic Acid Res. Mol. Biol. *51*, 225–270.

Chetverina, H. V., and Chetverin, A.B. (1993). Cloning of RNA molecules in vitro. Nucleic Acids Res. *21*, 2349–2353.

Clarke, M., Maddera, L., Harris, R.L., and Silverman, P.M. (2008). F-pili dynamics by live-cell imaging. Proc. Natl. Acad. Sci. U. S. A. *105*, 17978–17981.

Coetzee, J.N., Bradley, D.E., and Hedges, R.W. (1982). Phages Iα and I2-2: IncI plasmid-dependent bacteriophages. J. Gen. Microbiol. *128*, 2797–2804.

Coetzee, J.N., Bradley, D.E., Hedges, R.W., Fleming, J., and Lecatsas, G. (1983). Bacteriophage M: an incompatibility group M plasmid-specific phage. J. Gen. Microbiol. *129*, 2271–2276.

Coetzee, J.N., Bradley, D.E., Lecatsas, G., Du Toit, L., and Hedges, R.W. (1985a). Bacteriophage D: an IncD group plasmid-specific phage. J. Gen. Microbiol. *131*, 3375–3383.

Coetzee, J.N., Bradley, D.E., Fleming, J., Du Toit, L., Hughes, V.M., and Hedges, R.W. (1985b). Phage pilHα: a phage which adsorbs to IncHI and IncHII plasmid-coded pili. J. Gen. Microbiol. *131*, 1115–1121.

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res. *36*, W197–201.

Convery, M.A., Rowsell, S., Stonehouse, N.J., Ellington, A.D., Hirao, I., Murray, J.B., Peabody, D.S., Phillips, S.E., and Stockley, P.G. (1998). Crystal structure of an RNA aptamer-protein complex at 2.8 Å resolution. Nat. Struct. Biol. *5*, 133–139.

Crawford, E.M., and Gesteland, R.F. (1964). The adsorption of bacteriophage R17. Virology *22*, 165–167.

Danziger, R.E., and Paranchych, W. (1970). Stages in phage R17 infection. III. Energy requirements for the F-pili mediated eclipse of viral infectivity. Virology *40*, 554–564.

Davis, J.E., Strauss, J.H., and Sinsheimer, R.L. (1961). Bacteriophage MS2: Another RNA phage. Science *134*, 1427.

Dent, K.C., Thompson, R., Barker, A.M., Hiscox, J.A., Barr, J.N., Stockley, P.G., and Ranson, N.A. (2013). The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release. Structure *21*, 1225–1234.

Van Duin, J., and Tsareva, N. (2006). Single-Stranded RNA Phages. In The Bacteriophages, R. Calendar, ed. (New York: Oxford University Press), pp. 175–196.

Engelhardt, D.L., and Zinder, N.D. (1964). Host-dependent mutants of the bacteriophage f2. III. Infective RNA. Virology *23*, 582–587.

Feary, T.W., Fisher, E., and Fisher, T.N. (1964). Isolation and preliminary characteristics of three bacteriophages associated with a lysogenic strain of Pseudomonas aeruginosa. J. Bacteriol. *87*, 196–208.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Vandenberghe, A., et al. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507.

Firth, N., and Skurray, R. (1992). Characterization of the F plasmid bifunctional conjugation gene, traG. Mol. Gen. Genet. *232*, 145–153.

Fronzes, R., Schäfer, E., Wang, L., Saibil, H.R., Orlova, E. V, and Waksman, G. (2009). Structure of a type IV secretion system core complex. Science *323*, 266–268.

Frost, L.S., and Paranchych, W. (1988). DNA sequence analysis of point mutations in traA, the F pilin gene, reveal two domains involved in F-specific bacteriophage attachment. Mol. Gen. Genet. *213*, 134–139.

Frost, L.S., Ippen-ihler, K., and Skurray, R.A. (1994). Analysis of the Sequence and Gene Products of the Transfer Region of the F Sex Factor. Microbiol. Rev. *58*, 162–210.

Furuse, K. (1987). Distribution of coliphages in the environment: general considerations. In Phage Ecology, S.M. Goyal, C.P. Gerba, and G. Bitton, eds. (New York: John Wiley & Sons), pp. 87–123.

Garwes, D., Sillero, A., and Ochoa, S. (1969). Virus-specific proteins in Escherichia coli infected with phage Qβ. Biochim. Biophys. Acta *186*, 166–172.

Goessens, W.H.F., Driessen, A.J.M., Wilschut, J., and Van Duin, J. (1988). A synthetic peptide corresponding to the C-terminal 25 residues of phage MS2 coded lysis protein dissipates the protonmotive force in Escherichia coli membrane vesicles by generating hydrophilic pores. EMBO J. *7*, 867–873.

Golmohammadi, R., Valegård, K., Fridborg, K., and Liljas, L. (1993). The refined structure of bacteriophage MS2 at 2.8 Å resolution. J. Mol. Biol. *234*, 39.

Golmohammadi, R., Fridborg, K., Bundule, M., Valegård, K., and Liljas, L. (1996). The crystal structure of bacteriophage Qβ at 3.5 Å resolution. Structure *4*, 343–354.

Gomis-Ruth, F.X., Moncalian, G., Perez-luque, R., Gonzalez, A., Cabezon, E., De La Cruz, F., and M, C. (2001). The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase. Nature *409*, 637–641.

Gott, J.M., Wilhelm, L.J., and Uhlenbeck, O.C. (1991). RNA binding properties of the coat protein from bacteriophage GA. Nucleic Acids Res. *19*, 6499–6503.

Grahn, E., Stonehouse, N.J., Murray, J.B., Van den Worm, S., Valegård, K., Fridborg, K., Stockley, P.G., and Liljas, L. (1999). Crystallographic studies of RNA hairpins in complexes with recombinant MS2 capsids: implications for binding requirements. RNA *5*, 131–138.

Grahn, E., Stonehouse, N.J., Adams, C.J., Fridborg, K., Beigelman, L., Matulic-Adamic, J., Warriner, S.L., Stockley, P.G., and Liljas, L. (2000). Deletion of a single hydrogen bonding atom from the MS2 RNA operator leads to dramatic rearrangements at the RNA-coat protein interface. Nucleic Acids Res. *28*, 4611–4616.

Grahn, E., Moss, T., Helgstrand, C., Fridborg, K., Sundaram, M., Tars, K., Lago, H., Stonehouse, N.J., Davis, D.R., Stockley, P.G., et al. (2001). Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coat-protein complex. RNA *7*, 1616–1627.

Gralla, J., Steitz, J.A., and Crothers, D.M. (1974). Direct physical evidence for secondary structure in an isolated fragment of R17 bacteriophage mRNA. Nature *248*, 204–208.

Groeneveld, H., Thimon, K., and Van Duin, J. (1995). Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? RNA *1*, 79–88.

Harris, R.L., Hombs, V., and Silverman, P.M. (2001). Evidence that F-plasmid proteins TraV, TraK and TraB assemble into an envelope-spanning structure in Escherichia coli. Mol. Microbiol. *42*, 757–766.

Haruna, I., and Spiegelman, S. (1965a). Specific template requirments of RNA replicases. Proc. Natl. Acad. Sci. U. S. A. *54*, 579–587.

Haruna, I., and Spiegelman, S. (1965b). Autocatalytic Synthesis of a Viral RNA in vitro. Science *150*, 884–886.

Haruna, I., Nozu, K., Ohtaka, Y., and Spiegelman, S. (1963). An RNA “replicase” induced by and selective for a viral RNA: isolation and properties. Proc. Natl. Acad. Sci. U. S. A. *50*, 905–911.

Hasson, M.S., Muscate, A., McLeish, M.J., Polovnikova, L.S., Gerlt, J.A., Kenyon, G.L., Petsko, G.A., and Ringe, D. (1998). The crystal structure of benzoylformate decarboxylase at 1.6 Å resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes. Biochemistry *37*, 9918–9930.

Hazes, B., and Frost, L. (2008). Towards a systems biology approach to study type II/IV secretion systems. Biochim. Biophys. Acta *1778*, 1839–1850.

Helgstrand, C., Grahn, E., Moss, T., Stonehouse, N.J., Tars, K., Stockley, P.G., and Liljas, L. (2002). Investigating the structural basis of purine specificity in the structures of MS2 coat protein RNA translational operator hairpins. Nucleic Acids Res. *30*, 2678–2685.

Higgins, D.G., and Sharp, P.M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene *73*, 237–244.

Hirao, I., Spingola, M., Peabody, D., and Ellington, A.D. (1999). The limits of specificity: an experimental analysis with RNA aptamers to MS2 coat protein variants. Mol. Diversity *4*, 75–89.

Hirsh, D., and Gold, L. (1971). Translation of the UGA triplet in vitro by tryptophan transfer RNA's. J. Mol. Biol. *58*, 459–468.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. *31*, 3429–3431.

Hofstetter, H., Monstein, H.J., and Weissmann, C. (1974). The readthrough protein A1 is essential for the formation of viable Qβ particles. Biochim. Biophys. Acta *374*, 238–251.

Hohn, T. (1969). Role of RNA in the assembly process of bacteriophage fr. J. Mol. Biol. *43*, 191–200.

Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res. *38*, W545–549.

Hooker, J.M., Kovacs, E.W., and Francis, M.B. (2004). Interior surface modification of bacteriophage MS2. J. Am. Chem. Soc. *126*, 3718–3719.

Horiuchi, K., Webster, R.E., and Matsuhashi, S. (1971). Gene products of bacteriophage Qβ. Virology *45*, 429–439.

Horn, W.T., Convery, M.A., Stonehouse, N.J., Adams, C.J., Liljas, L., Phillips, S.E. V, and Stockley, P.G. (2004). The crystal structure of a high affinity RNA stem-loop complexed with the bacteriophage MS2 capsid: further challenges in the modeling of ligand-RNA interactions. RNA *10*, 1776–1782.

Horn, W.T., Tars, K., Grahn, E., Helgstrand, C., Baron, A.J., Lago, H., Adams, C.J., Peabody, D.S., Phillips, S.E., Stonehouse, N.J., et al. (2006). Structural basis of RNA binding discrimination between bacteriophages Qβ and MS2. Structure *14*, 487–495.

Hung, P.P., and Overby, L.R. (1969). The reconstitution of infective bacteriophage Qβ. Biochemistry *8*, 820–828.

Inokuchi, Y., Takahashi, R., Hirose, T., Inayama, S., Jacobson, A.B., and Hirashima, A. (1986). The complete nucleotide sequence of the group II RNA coliphage GA. J. Biochem. *99*, 1169–1180.

Kaerner, H.C. (1970). Sequential steps in the in vivo assembly of the RNA bacteriophage fr. J. Mol. Biol. *53*, 515–529.

Kamen, R. (1975). Structure and function of the Qβ replicase. In RNA Phages, N. Zinder, ed. (New York: Cold Spring Harbor Laboratory), pp. 203–234.

Kamen, R., Kondo, M., Romer, W., and Weissmann, C. (1972). Reconstitution of Qβ Replicase Lacking Subunit alpha with Protein-Synthesis-Interference Factor i. Eur. J. Biochem. *31*, 44–51.

Kannoly, S., Shao, Y., and Wang, I.-N. (2012). Rethinking the evolution of single-stranded RNA (ssRNA) bacteriophages based on genomic sequences and characterizations of two R-plasmid-dependent ssRNA phages, C-1 and Hgal1. J. Bacteriol. *194*, 5073–5079.

Karnik, S., and Billeter, M. (1983). The lysis function of RNA bacteriophage Qβ is mediated by the maturation (A2) protein. EMBO J. *2*, 1521–1526.

Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., and Leberman, R. (1996). The structure of the Escherichia coli EF-Tu·EF-Ts complex at 2.5 Å resolution. Nature *379*, 511–518.

Kay, B.K., Williamson, M.P., and Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J. *14*, 231–241.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature *181*, 662–666.

Kidmose, R.T., Vasiliev, N.N., Chetverin, A.B., Andersen, G.R., and Knudsen, C.R. (2010). Structure of the Qβ replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins. Proc. Natl. Acad. Sci. U. S. A. *107*, 10884–10889.

Klovins, J., and Van Duin, J. (1999). A long-range pseudoknot in Qβ RNA is essential for replication. J. Mol. Biol. *294*, 875–884.

Klovins, J., Berzins, V., and Van Duin, J. (1998). A long-range interaction in Qβ RNA that bridges the thousand nucleotides between the M-site and the 3' end is required for replication. RNA *4*, 948–957.

Klovins, J., Overbeek, G.P., Van den Worm, S.H.E., Ackermann, H.-W., and Van Duin, J. (2002). Nucleotide sequence of a ssRNA phage from Acinetobacter: kinship to coliphages. J. Gen. Virol. *83*, 1523–1533.

Köhler, S.D., Weber, A., Howard, S.P., Welte, W., and Drescher, M. (2010). The proline-rich domain of TonB possesses an extended polyproline II-like conformation of sufficient length to span the periplasm of Gram-negative bacteria. Protein Sci. *19*, 625–630.

Koning, R., Van den Worm, S.H., Plaisier, J.R., Van Duin, J., Abrahams, J.P., and Koerten, H. (2003). Visualization by cryo-electron microscopy of genomic RNA that binds to the protein capsid inside bacteriophage MS2. J. Mol. Biol. *332*, 415–422.

Koonin, E. V, and Dolja, V. V (1993). Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. *28*, 375–430.

Krab, I.M., and Parmeggiani, A. (1998). EF-Tu, a GTPase odyssey. Biochim. Biophys. Acta *1443*, 1–22.

Krahn, P.M., O'Callaghan, R.J., and Paranchych, W. (1972). Stages in phage R17 infection. VI. Injection of A protein and RNA into the host cell. Virology *47*, 628–637.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kuzmanovic, D.A., Elashvili, I., Wick, C., O'Connell, C., and Krueger, S. (2003). Bacteriophage MS2: molecular weight and spatial distribution of the protein and RNA components by small-angle neutron scattering and virus counting. Structure *11*, 1339–1348.

Lanka, E., and Wilkins, B.M. (1995). DNA processing reactions in bacterial conjugation. Annu. Rev. Biochem. *64*, 141–169.

Lawley, T., Wilkins, B.M., and Frost, L.S. (2004). Bacterial conjugation in Gram-negative bacteria. In Plasmid Biology, B.E. Funnell, and G.J. Phillips, eds. (Washington, D.C.: ASM Press), pp. 203–226.

Lawley, T.D., Klimke, W.A., Gubbins, M.J., and Frost, L.S. (2003). F factor conjugation is a true type IV secretion system. FEMS Microbiol. Lett. *224*, 1–15.

Lee, M.H., Kosuk, N., Bailey, J., Traxler, B., and Manoil, C. (1999). Analysis of F factor TraD membrane topology by use of gene fusions and trypsin-sensitive insertions. J. Bacteriol. *181*, 6108–6113.

Licis, N., Balklava, Z., and Van Duin, J. (2000). Forced retroevolution of an RNA bacteriophage. Virology *271*, 298–306.

Liljas, L., Fridborg, K., Valegård, K., Bundule, M., and Pumpens, P. (1994). Crystal structure of bacteriophage fr capsids at 3.5 Å resolution. J. Mol. Biol. *244*, 279–290.

Lim, F., and Peabody, D.S. (2002). RNA recognition site of PP7 coat protein. Nucleic Acids Res. *30*, 4138–4144.

Lim, F., Spingola, M., and Peabody, D.S. (1996). The RNA-binding site of bacteriophage Qβ coat protein. J. Biol. Chem. *271*, 31839–31845.

Lim, F., Downey, T.P., and Peabody, D.S. (2001). Translational repression and specific RNA binding by the coat protein of the Pseudomonas phage PP7. J. Biol. Chem. *276*, 22507–22513.

Ling, C.M., Hung, P.P., and Overby, L.R. (1969). Specificity in self-assembly of bacteriophages Qβ and MS2. Biochemistry *8*, 4464–4469.

Ling, C.M., Hung, P.P., and Overby, L.R. (1970). Independent assembly of Qβ and MS2 phages in doubly infected Escherichia coli. Virology *40*, 920–929.

Lodish, H.F., Horiuchi, K., and Zinder, N.D. (1965). Mutants of bacteriophage f2. V. On the production of noninfectious phage particles. Virology *27*, 139–155.

Loeb, T. (1960). Isolation of a bacteriophage specific for the F plus and Hfr mating types of Escherichia coli K-12. Science *131*, 932–953.

Loeb, T., and Zinder, N.D. (1961). A bacteriophage containing RNA. Proc. Natl. Acad. Sci. U. S. A. *47*, 282–289.

Low, H.H., Gubellini, F., Rivera-Calzada, A., Braun, N., Connery, S., Dujeancourt, A., Lu, F., Redzej, A., Fronzes, R., Orlova, E. V, et al. (2014). Structure of a type IV secretion system. Nature *508*, 550–553.

Lowary, P.T., and Uhlenbeck, O.C. (1987). An RNA mutation that increases the affinity of an RNA-protein interaction. Nucleic Acids Res. *15*, 10483–10493.

Marvin, D.A., and Folkhard, W. (1986). Structure of F-pili: reassessment of the symmetry. J. Mol. Biol. *191*, 299–300.

Marvin, D.A., and Hoffmann-Berling, H. (1963). Physical and chemical properties of two small bacteriophages. Nature *197*, 517–518.

Van Meerten, D., Girard, G., and Van Duin, J. (2001). Translational control by delayed RNA folding: identification of the kinetic trap. RNA *7*, 483–494.

Meyer, F., Weber, H., and Weissmann, C. (1981). Interactions of Qβ replicase with Qβ RNA. J. Mol. Biol. *153*, 631–660.

Miranda, G., Schuppli, D., Barrera, I., Hausherr, C., Sogo, J.M., and Weber, H. (1997). Recognition of bacteriophage Qβ plus strand RNA as a template by Qβ replicase: role of RNA interactions mediated by ribosomal proteins S1 and host factor. J. Mol. Biol. *267*, 1089–1103.

Ng, K.K.S., Arnold, J.J., and Cameron, C.E. (2008). Structure-function relationships among RNA-dependent RNA polymerases. Curr. Top. Microbiol. Immunol. *320*, 137–156.

Nuttall, D., Maker, D., and Colleran, E. (1987). A method for the direct isolation of IncH plasmid-dependent bacteriophages. Lett. Appl. Microbiol. *5*, 37–40.

Olsen, R.H., and Shipley, P. (1973). Host range and properties of the Pseudomonas aeruginosa R factor R1822. J. Bacteriol. *113*, 772–780.

Olsen, R.H., and Thomas, D.D. (1973). Characteristics and purification of PRR1, an RNA phage specific for the broad host range Pseudomonas R1822 drug resistance plasmid. J. Virol. *12*, 1560–1567.

Olsthoorn, R.C.L., and Van Duin, J. (1996). Random removal of inserts from an RNA genome: selection against single-stranded RNA. J. Virol. *70*, 729–736.

Olsthoorn, R.C.L., Garde, G., Dayhuff, T., Atkins, J.F., and Van Duin, J. (1995). Nucleotide sequences of a single-stranded RNA phage from Pseudomonas aeruginosa: kinship to coliphages and conservation of regulatory RNA structures. Virology *206*, 611–625.

Overby, L.R., Barlow, G.H., Doi, R.H., Jacob, M., and Spiegelman, S. (1966). Comparision of two serologically distinct ribonucleic acid bacteriophages. I. Properties of the viral particles. J. Bacteriol. *91*, 442–448.

Paiva, W.D., Grossman, T., and Silverman, P.M. (1992). Characterization of F-pilin as an inner membrane component of Escherichia coli K12. J. Biol. Chem. *267*, 26191–26197.

Paranchych, W., and Graham, A.F. (1962). Isolation and properties of an RNA-containing bacteriophage. J. Cell. Comp. Physiol. *60*, 199.

Parrott, A.M., Lago, H., Adams, C.J., Ashcroft, A.E., Stonehouse, N.J., and Stockley, P.G. (2000). RNA aptamers for the MS2 bacteriophage coat protein and the wild-type RNA operator have similar solution behaviour. Nucleic Acids Res. *28*, 489–497.

Peabody, D.S. (1997). Role of the coat protein-RNA interaction in the life cycle of bacteriophage MS2. Mol. Genet. Genomics *254*, 358–364.

Peabody, C.R., Chung, Y.J., Yen, M.-R., Vidal-Ingigliardi, D., Pugsley, A.P., and Saier, M.H. (2003). Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. Microbiology *149*, 3051–3072.

Persson, M., Tars, K., and Liljas, L. (2008). The Capsid of the Small RNA Phage PRR1 Is Stabilized by Metal Ions. J. Mol. Biol. *383*, 914–922.

Persson, M., Tars, K., and Liljas, L. (2013). PRR1 coat protein binding to its RNA translational operator. Acta Crystallogr. D Biol. Crystallogr. *69*, 367–372.

Pickett, G.G., and Peabody, D.S. (1993). Encapsidation of heterologous RNAs by bacteriophage MS2 coat protein. Nucleic Acids Res. *21*, 4621–4626.

Plevka, P., Kazaks, A., Voronkova, T., Kotelovica, S., Dishlers, A., Liljas, L., and Tars, K. (2009). The structure of bacteriophage phiCb5 reveals a role of the RNA genome and metal ions in particle stability and assembly. J. Mol. Biol. *391*, 635–647.

Radloff, R.J., and Kaesberg, P. (1973). Electrophoretic and other properties of bacteriophage Qβ: the effect of a variable number of read-through proteins. J. Virol. *11*, 116–128.

Reed, C.A., Langlais, C., Kuznetsov, V., and Young, R. (2012). Inhibitory mechanism of the Qβ lysis protein A2. Mol. Microbiol. *86*, 836–844.

Romaniuk, P.J., Lowary, P., Wu, H.N., Stormo, G., and Uhlenbeck, O.C. (1987). RNA binding site of R17 coat protein. Biochemistry *26*, 1563–1568.

Rosenberg, A.H., Lade, B.N., Dao-shan, C., Lin, S.-W., Dunn, J.J., and Studier, F.W. (1987). Vectors for selective expression of cloned DNAs by T7 RNA polymerase. Gene *56*, 125–135.

Rowsell, S., Stonehouse, N.J., Convery, M.A., Adams, C.J., Ellington, A.D., Hirao, I., Peabody, D.S., Stockley, P.G., and Phillips, S.E. (1998). Crystal structures of a series of RNA aptamers complexed to the same protein target. Nat. Struct. Biol. *5*, 970–975.

Ruokoranta, T.M., Grahn, A.M., Ravantti, J.J., Poranen, M.M., and Bamford, D.H. (2006). Complete genome sequence of the broad host range single-stranded RNA phage PRR1 places it in the Levivirus genus with characteristics shared with Alloleviviruses. J. Virol. *80*, 9326–9330.

Schmidt, J.M., and Stanier, R.Y. (1965). Isolation and characterization of bacteriophages active against stalked bacteria. J. Gen. Microbiol. *39*, 95–107.

Schuppli, D., Miranda, G., Qiu, S., and Weber, H. (1998). A branched stem-loop structure in the M-site of bacteriophage Qβ RNA is important for template recognition by Qβ replicase holoenzyme. J. Mol. Biol. *283*, 585–593.

Shiba, T., and Miyake, T. (1975). New type of infectious complex of E. coli RNA phage. Nature *254*, 157–158.

Shiba, T., and Suzuki, Y. (1981). Localization of A protein in the RNA-A protein complex of RNA phage MS2. Biochim. Biophys. Acta *654*, 249–255.

Silverman, P.M. (1997). Towards a structural biology of bacterial conjugation. Mol. Microbiol. *23*, 423–429.

Sirgel, F.A., Coetzee, J.N., Hedges, R.W., and Lecatsas, G. (1981). Phage C-1: an IncC group plasmid-specific phage. J. Gen. Microbiol. *122*, 155–160.

Skripkin, E.A., Adhin, M.R., De Smit, M.H., and Van Duin, J. (1990). Secondary structure of bacteriophage MS2. Conservation and biological significance. J. Mol. Biol. *211*, 447–463.

De Smit, M.H., and Van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc. Natl. Acad. Sci. U. S. A. *87*, 7668–7672.

Sørensen, M.A., Fricke, J., and Pedersen, S. (1998). Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. J. Mol. Biol. *280*, 561–569.

Spingola, M., and Peabody, D.S. (1997). MS2 coat protein mutants which bind Qβ RNA. Nucleic Acids Res. *25*, 2808–2815.

Stockley, P.G., Stonehouse, N.J., and Valegård, K. (1994). Molecular mechanism of RNA phage morphogenesis. Int. J. Biochem. *26*, 1249–1260.

Stockley, P.G., Rolfsson, O., Thompson, G.S., Basnak, G., Francese, S., Stonehouse, N.J., Homans, S.W., and Ashcroft, A.E. (2007). A Simple, RNA-Mediated Allosteric Switch Controls the Pathway to Formation of a T=3 Viral Capsid. J. Mol. Biol. *369*, 541–552.

Strauss, E.G., and Kaesberg, P. (1970). Acrylamide gel electrophoresis of bacteriophage Qβ: Electrophoresis of intact virions and of viral proteins. Virology *42*, 437–452.

Takamatsu, H., and Iso, K. (1982). Chemical evidence for the capsomeric structure of phage Qβ. Nature *298*, 819–824.

Takeshita, D., and Tomita, K. (2010). Assembly of Qβ viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. Proc. Natl. Acad. Sci. U. S. A. *107*, 15733–15738.

Takeshita, D., and Tomita, K. (2012). Molecular basis for RNA polymerization by Qβ replicase. Nat. Struct. Mol. Biol. *19*, 229–237.

Takeshita, D., Yamashita, S., and Tomita, K. (2012). Mechanism for template-independent terminal adenylation activity of Qβ replicase. Structure *20*, 1661–1669.

Takeshita, D., Yamashita, S., and Tomita, K. (2014). Molecular insights into replication initiation by Qβ replicase using ribosomal protein S1. Nucleic Acids Res. *42*, 10809–10822.

Tars, K., Bundule, M., Fridborg, K., and Liljas, L. (1997). The crystal structure of bacteriophage GA and a comparison of bacteriophages belonging to the major groups of Escherichia coli leviviruses. J. Mol. Biol. *271*, 759–773.

Tars, K., Fridborg, K., Bundule, M., and Liljas, L. (2000). The three-dimensional structure of bacteriophage PP7 from Pseudomonas aeruginosa at 3.7-Å resolution. Virology *272*, 331–337.

Taylor, D.E., Gibreel, A., Lawley, T.D., and Tracz, D.M. (2004). Antibiotic resistance plasmids. In Plasmid Biology, B.E. Funnell, and G.J. Phillips, eds. (Washington, D.C.: ASM Press), pp. 473–491.

Tomita, K. (2014). Structures and Functions of Qβ Replicase: Translation Factors beyond Protein Synthesis. Int. J. Mol. Sci. *15*, 15552–15570.

Toropova, K., Basnak, G., Twarock, R., Stockley, P.G., and Ranson, N.A. (2008). The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. J. Mol. Biol. *375*, 824–836.

Ugarov, V.I., and Chetverin, A.B. (2008). Functional circularity of legitimate Qβ replicase templates. J. Mol. Biol. *379*, 414–427.

Uhlenbeck, O.C., Carey, J., Romaniuk, P.J., Lowary, P.T., and Beckett, D. (1983). Interaction of R17 coat protein with its RNA binding site for translational repression. J. Biomol. Struct. Dyn. *1*, 539–552.

Valegård, K., Liljas, L., Fridborg, K., and Unge, T. (1990). The three-dimensional structure of the bacterial virus MS2. Nature *345*, 36–41.

Valegård, K., Murray, J.B., Stockley, P.G., Stonehouse, N.J., and Liljas, L. (1994). Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature *371*, 623–626.

Valegård, K., Murray, J.B., Stonehouse, N.J., Van den Worm, S., Stockley, P.G., and Liljas, L. (1997). The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. J. Mol. Biol. *270*, 724–738.

Valentine, R.C., and Strand, M. (1965). Complexes of F-pili and RNA bacteriophage. Science *148*, 511–513.

Vasiljeva, I., Kozlovska, T., Cielens, I., Strelnikova, A., Kazaks, A., Ose, V., and Pumpens, P. (1998). Mosaic Qβ coats as a new presentation model. FEBS Lett. *431*, 7–11.

Vollmer, W., Blanot, D., and De Pedro, M.A. (2008a). Peptidoglycan structure and architecture. FEMS Microbiol. Rev. *32*, 149–167.

Vollmer, W., Joris, B., Charlier, P., and Foster, S. (2008b). Bacterial peptidoglycan (murein) hydrolases. FEMS Microbiol. Rev. *32*, 259–286.

Wahba, A.J., Miller, M.J., Niveleau, A., Landers, T.A., Carmichael, G.G., Weber, K., Hawley, D.A., and Slobin, L.I. (1974). Subunit I of Qβ replicase and 30 S ribosomal protein S1 of Escherichia coli. Evidence for the identity of the two proteins. J. Biol. Chem. *249*, 3314–3316.

Watanabe, I. (1964). Persistent infection with an RNA bacteriophage. Nihon Rinsho *22*, 243–251.

Weber, H. (1976). The binding site for coat protein on bacteriophage Qβ RNA. Biochim. Biophys. Acta *418*, 175–183.

Weber, H., and Weissmann, C. (1970). The 3′-termini of bacteriophage Qβ plus and minus strands. J. Mol. Biol. *51*, 215–224.

Weber, K., and Konigsberg, W. (1975). Proteins of RNA phages. In RNA Phages, N. Zinder, ed. (New York: Cold Spring Harbor Laboratory), pp. 51–84.

Weiner, A.M., and Weber, K. (1971). Natural read-through at the UGA termination signal of Qβ coat protein cistron. Nat. New Biol. *234*, 206–209.

Weiner, A.M., and Weber, K. (1973). A single UGA codon functions as a natural termination signal in the coliphage Qβ coat protein cistron. J. Mol. Biol. *80*, 837–855.

Weissmann, C., Feix, G., Slor, H., and Pollet, R. (1967). Replication of viral RNA. XIV. Single-stranded minus strands as template for the synthesis of viral plus strands in vitro. Proc. Natl. Acad. Sci. U. S. A. *57*, 1870–1877.

Willetts, N., and Achtman, M. (1972). Genetic analysis of transfer by the Escherichia coli sex factor F, using P1 transductional complementation. J. Bacteriol. *110*, 843–851.

Williamson, M.P. (1994). The structure and function of proline-rich regions in proteins. Biochem. J. *297*, 249–260.

Winter, R.B., and Gold, L. (1983). Overproduction of bacteriophage Qβ maturation (A2) protein leads to cell lysis. Cell *33*, 877–885.

Witherell, G.W., and Uhlenbeck, O.C. (1989). Specific RNA binding by Qβ coat protein. Biochemistry *28*, 71–76.

Van den Worm, S.H., Stonehouse, N.J., Valegård, K., Murray, J.B., Walton, C., Fridborg, K., Stockley, P.G., and Liljas, L. (1998). Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments. Nucleic Acids Res. *26*, 1345–1351.

Van den Worm, S.H.E., Koning, R.I., Warmenhoven, H.J., Koerten, H.K., and Van Duin, J. (2006). Cryo electron microscopy reconstructions of the Leviviridae unveil the densest icosahedral RNA packing possible. J. Mol. Biol. *363*, 858–865.

Young, R. (2013). Phage lysis: do we have the hole story yet? Curr. Opin. Microbiol. *16*, 790–797.

Zinder, N.D. (1975). Preface. In RNA Phages, N.D. Zinder, ed. (New York: Cold Spring Harbor Laboratory), pp. v–vi.

# Paper I

# Crystal structure of the read-through domain from bacteriophage Qβ A1 protein

**Janis Rumnieks* and Kaspars Tars**

Latvian Biomedical Research and Study Centre, Ratsupites 1, Riga LV-1067, Latvia

Received 31 May 2011; Revised 13 July 2011; Accepted 13 July 2011
DOI: 10.1002/pro.704
Published online 29 July 2011 proteinscience.org

**Abstract: Bacteriophage Qβ is a small RNA virus that infects *Escherichia coli.* The virus particle contains a few copies of the minor coat protein A1, a C-terminally prolonged version of the coat protein, which is formed when ribosomes occasionally read-through the leaky stop codon of the coat protein. The crystal structure of the read-through domain from bacteriophage Qβ A1 protein was determined at a resolution of 1.8 Å. The domain consists of a heavily deformed five-stranded β-barrel on one side of the protein and a β-hairpin and a three-stranded β-sheet on the other. Several short helices and well-ordered loops are also present throughout the protein. The N-terminal part of the read-through domain contains a prominent polyproline type II helix. The overall fold of the domain is not similar to any published structure in the Protein Data Bank.**

**Keywords: *Leviviridae*; allolevivirus; small RNA phages; bacteriophage Qβ; minor coat protein; read-through protein; polyproline helix**

## Introduction

Bacteriophages of the *Leviviridae* family are among the smallest and simplest known viruses. They have a single-stranded, positive-sense RNA genome, which is about 3500–4200 nucleotides long and encodes a maturation protein, a coat protein, and a subunit of the replicase complex.[1] The capsid is built from 90 dimers of coat protein that assemble in an icosahedral shell with $T = 3$ symmetry.[2] In addition to the coat protein, each virion contains a single copy of the maturation protein.[3] The maturation protein is bound to the genomic RNA[4] and mediates the attachment of the phage to the sides of bacterial pili,[5] which is the cellular receptor for all known *Leviviridae* phages. After attachment to the pili, the RNA-maturation protein complex leaves the capsid and enters the cell through an unknown mechanism.

Many of the known *Leviviridae* phages are further divided into two genera, leviviruses and alloleviviruses. A marked difference between the two genera is how the phages achieve cell lysis: leviviruses encode a small lysis protein that overlaps with coat and replicase genes in a different reading frame, whereas alloleviviruses mediate lysis using the maturation protein.[6,7] The other unique feature of alloleviviruses is the presence of a minor coat protein species A1 in their capsid. The A1 protein is produced when ribosomes occasionally read-through the leaky UGA termination codon of the coat protein gene[8] and translation continues for another 600 nucleotides, resulting in a C-terminal extension of the coat protein. The A1 protein is incorporated in 3–10 copies per virion[1] and is essential for producing infectious virus particles,[9] but its precise function is not known. To gain new insights about this protein, we solved the crystal structure of the read-through extension from bacteriophage Qβ A1 protein.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: European Social Fund; Grant number: 1DP/1.1.1.2.0/09/APIA/VIAA/150; Grant sponsor: Latvian Council of Science; Grant number: 09.1294; Grant sponsor: European Social Fund (Support for Doctoral Studies at University of Latvia); Grant number: 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004 (J.R.).

**Correspondence to:* Janis Rumnieks, Latvian Biomedical Research and Study Centre, Department of Protein Engineering, Ratsupites 1, Riga LV-1067, Latvia. E-mail: j.rumnieks@biomed.lu.lv

## Results and Discussion

### Structure determination and quality of the models

Because of the low number and presumed random orientation in the capsid, the read-through extensions were not visible in the crystal structure of bacteriophage Qβ.[10] The A1 protein alone is insoluble and cannot assemble into particles without the assistance of the coat protein,[11] and the amount of A1 protein that can be incorporated into the particles seems to be limited to about 15%.[1] To make the A1 protein amenable to structural analysis, we expressed the read-through domain separately. The complete read-through extension starting from the end of the coat protein was largely insoluble (data not shown), but a hexahistidine tagged variant starting 11 amino acids away from the coat protein part (residues 144–328 of full-length A1 protein) was highly soluble, could be readily purified, and was chosen to proceed with crystallization. The protein was crystallized in two crystal forms, monoclinic and hexagonal, which diffracted to 1.8 and 2.9 Å resolutions, respectively. The structure of the monoclinic form was solved by multiple isomorphous replacement with anomalous scattering using two derivatives. Except for the expression tag and the first two residues of the crystallized domain, the polypeptide chain could be traced unambiguously, without breaks, from residue 146 (the numbering of residues is as of full-length A1 protein) to the end of the chain. In the hexagonal form, another seven N-terminal residues could not be located in the electron density, and the chain was traced starting from residue 153. The domain adopts an almost identical conformation in the two crystal forms, with an rms deviation of 0.76 Å for the main chain atoms.

### Overall structure

The overall fold of the read-through domain [Fig. 1(A)] is not similar to any other published structure in the Protein Data Bank, according to the DALI server.[13] Except for the N-terminal region, the domain has a compact, roughly globular shape with a mixed α/β architecture. The core of the domain is built of β-sheets: strands β2, β3, β6, β7, and β8 form a heavily deformed, five-stranded β-barrel on one side of the protein, whereas β1 and β4 and β5, β9, and β10 form two antiparallel sheets on the other side. There are three α-helices and two $3_{10}$-helices in the protein, which are all short and are located predominantly on the surface. A remarkably long loop (23 residues) connects the first $3_{10}$-helix and strand β5, but it is well ordered and kept in place by extensive hydrogen bonding involving main chain and side chain atoms. Eight of the first 15 residues that are visible in the electron density map are prolines. These residues form a polyproline type II helix that stretches for about 45 Å before turning 90° toward the rest of the protein [Fig. 1(B)]. The polyproline helix is held in position by two crystal contacts with the globular part of neighboring molecules in the monoclinic crystal form but not in the hexagonal form. Consequently, the distant part of the helix is not visible in the hexagonal form, which suggests that it is flexible in solution. It should be noted that, although poplyproline type II helices are not uncommon in proteins, the vast majority of them are shorter than six residues[14] and long helices are rare. The polyproline helix is A1 is quite remarkable in this aspect, since, according to a statistical survey of polyproline helices in protein structures in 2006,[14] the longest such helix observed in a crystal structure was that of the benzoylformate decarboxylase from *Pseudomonas putida*[15] (PDB ID 1BFD), which is 14 residues long and contains three prolines. The helix connects two subdomains of the enzyme but otherwise does not seem to have a specific function.

Currently, there is no structural information about residues 133–145, which separate the coat and read-through domains. Secondary structure prediction by JPred[16] suggests that this region is unstructured except for the coat protein-proximal six residues, which, together with the last three residues of the coat protein, may be involved in a short α-helix.

### Conserved regions

On the basis of phylogenetic and serological criteria, allolevivirus cluster into two groups denoted III and IV.[17,18] Up to date, there are 15 allolevivirus genome sequences available, of these eight are from Group III and seven from Group IV. When all of the sequences are aligned, coat proteins are the most conserved (~64% sequence identity), followed by the replicase (~44% identity) and maturation proteins (~29% identity). When sequences of all of the known A1 extensions are aligned, the total identity is only 26%, making them the most divergent part of all phage proteins. However, in a sequence alignment of A1 extensions from representative phages from Group III (Qβ and MX1) and Group IV (FI and SP) several conserved regions emerge [Fig. 2(A)]. First, in the N-terminal part (residues 146–159), ~50% of the residues are prolines in all alloleviviruses, suggesting that the polyproline helix is present in all allolevivirus A1 proteins and is probably important for their function. A short stretch of amino acids immediately following the helix is also conserved. The most prominent conserved regions are located at residues 207–219 and 228–238, which form part of the long loop between helix $3_{10}1$ and β5 and extend to strand β5 and the beginning of helix α2. The C-terminal region of the domain is also relatively conserved. Interestingly, the majority of conserved residues cluster on one side of the protein closer to the polyproline helix [Fig. 2(B)], suggesting that this part of the domain is the most critical for performing its function.

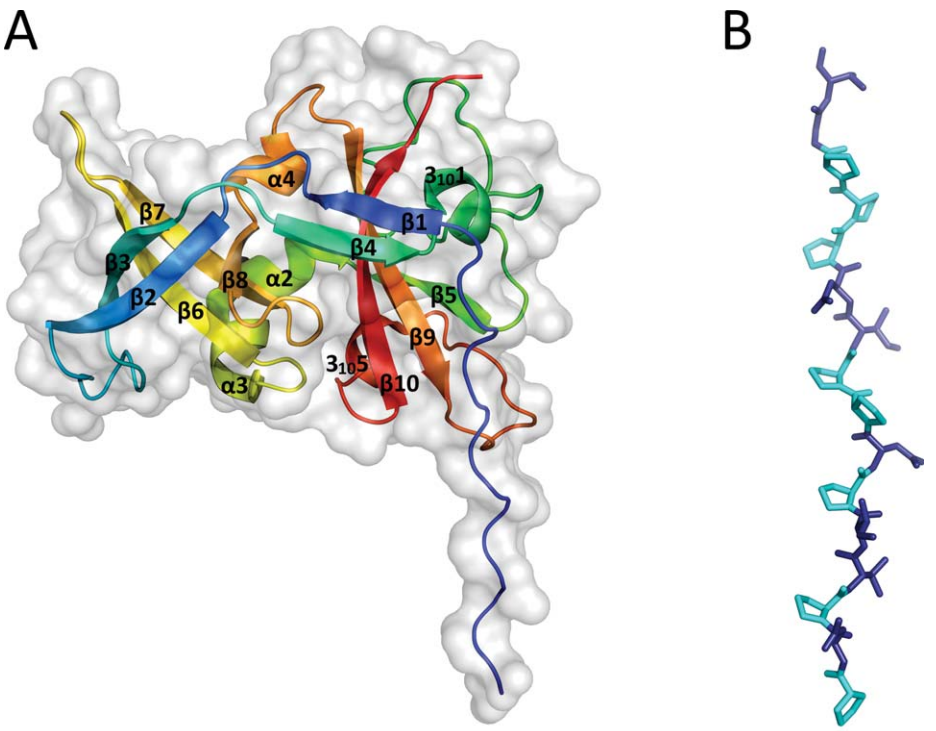


**Figure 1.** Structure of the read-through domain. (A) Overall structure of the domain. The protein is represented as a cartoon model rainbow-colored blue (N-terminus) to red (C-terminus) and overlaid with a surface representation of the domain (light grey). (B) A detailed view of the polyproline helix. In the first 16 residues of the model, prolines are represented in cyan and other residues in deep blue. Figures 1(A,B) and 2(B) were prepared using PyMol.[12]

### *Possible function of the A1 protein*

The actual function of the read-through domain has remained enigmatic. The amino acid sequence and the three-dimensional (3D) structure of the A1 extension are not similar to other known proteins, leaving no clues about its evolutionary origin. The A1 protein is a landmark of the rather small group of alloleviviruses, which all infect *Escherichia coli*,

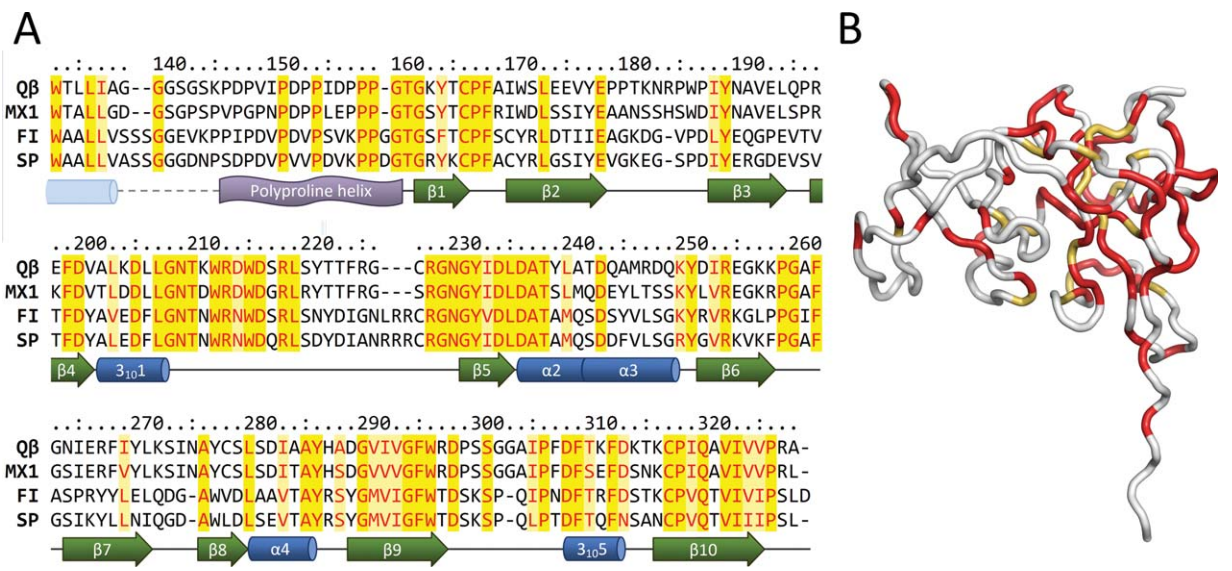


**Figure 2.** Conserved regions of the read-through domains. (A) Sequence alignment of the read-through domains from different alloleviviruses. Conserved residues are colored red; of these, identical residues are shaded yellow and nonidentical light yellow. Assigned secondary structure elements are presented below the alignment. A dashed line represents the portion for which no experimental data are available; the α-helix from secondary structure prediction is drawn as a pale blue cylinder. (B) Mapping of the conserved regions on the three-dimensional structure of the read-through domain. Identical and nonidentical but conserved residues as of Figure 2(A) are colored red and yellow-orange, respectively.

whereas all other known *Leviviridae* phages suffice with just the coat and maturation proteins in the virion. However, the A1 protein is essential for producing viable phage particles, as shown by *in vitro* virus reassembly assays[9] and *in vivo* plasmid complementation studies.[19] The C-termini of coat proteins with some minor structural rearrangements could reach both the inner and outer surface of the capsid. However, current evidence suggests that as structural components of the virion, the read-through extensions are located on the exterior of the capsid. First, Qβ virions form a diffuse band in native polyacrylamide gel electrophoresis,[20] and their mobility in the gel and in sucrose density gradients depends on how many copies of the A1 protein are present in the capsid.[21] Additionally, when recombinant Qβ capsids contained A1 extensions with an engineered internal epitope tag from hepatitis B virus preS1 region, the tags were accessible to antibodies in an ELISA assay and immunogold electron microscopy confirmed that the antibodies were indeed bound to the capsid surface.[11] The five-residue tag was inserted after residue 204, which is now known to be located in the short $3_{10}1$-helix on the surface of the protein and likely did not disturb the structure of the domain.

An interesting feature of the A1 protein undoubtedly is the long polyproline type II helix at the N-terminal part of the read-through domain. Polyproline helices and proline-rich regions in general are relatively abundant in proteins and have different functions,[22] but they frequently serve as ligands for various protein–protein interaction domains, resulting in formation of protein complexes that are often involved in signaling and regulatory pathways in eukaryotic cells (reviewed in Ref. [23]). In other proteins, proline-rich regions have a structural role and act as relatively rigid spacers to keep protein domains apart. For example, a 68 residue long proline-rich segment of the bacterial protein TonB was recently shown to adopt a polyproline II conformation that spans the periplasm.[24]

The linker between the coat and read-through domains would stretch for estimated 35 Å, and is then followed by the 45-Å-long polyproline helix, which is apparently also somewhat flexible. The logical explanation for such a long linker is that the read-through domain in the virion is positioned far away from the viral quasi-threefold symmetry axis (relating the three quasi-equivalent subunits A, B, and C) where the C-termini of coat proteins are located. A recent study localized the maturation protein from the distantly related phage MS2 on one of the viral fivefold symmetry axes and suggested that the Qβ maturation protein is localized similarly.[25] Because both maturation and A1 proteins are required for infectivity, it seems possible that the two proteins might interact with each other and that the long linker would allow the read-through domain to reach viral fivefold and threefold symmetry axes that are ~45 Å away from the C-termini of coat proteins. Experiments to test the association of the read-through domain with the maturation protein are underway in our laboratory.

## Conclusions

We have shown that the read-through domain of Qβ A1 protein adopts a previously unseen protein fold and has some intriguing structural features, such as a 15 residue-long polyproline type II helix which is one of the best examples of this kind of helix in globular proteins for which the 3D structures have been determined. Although the structure of the read-through domain does not provide immediate answers about its function, it gives a good starting point for further studies that could eventually lead to the understanding of the molecular mechanism by which the small RNA phages infect the bacterial host.

## Materials and Methods

### *Cloning, expression, and purification*

The coding sequence of Qβ A1 extension was amplified from plasmid pQβ10[26] using forward primer 5′-TACCATGGGGCACCATCATCACCATCATTCAAACCCGATCCGGTTATTCC-3′ and reverse primer 5′-ATCTGCAGTTAAGCACGAGGAACGACTATCACG-3′. The resulting fragment, encoding an N-terminally 6xHis-tagged A1 extension (denoted His-A1 hereafter) was cloned into a modified pBAD/Thio vector (Invitrogen). For protein production, the plasmid was transformed in *E.coli* strain TOP10, and cells were grown in LB medium containing 50 μg/mL ampicillin at 37°C until $OD_{590}$ of the culture reached 1.0. Arabinose was then added to a final concentration of 0.2%, and cells were grown for another 4 h and harvested by centrifugation.

Cells were resuspended in a lysis buffer containing 40 m*M* Tris-HCl pH 8.0, 200 m*M* NaCl, 20 m*M* $MgSO_4$, 0.1% Triton-X100, 0.1 mg/mL DNAse, and 1 mg/mL lysozyme and lysed by three freeze-thaw cycles. The lysate was centrifuged, supernatant was loaded on a HIS-Select cartridge (Sigma), and bound His-A1 protein was eluted with buffer containing 40 m*M* Tris-HCl pH 8.0, 300 m*M* NaCl, and 300 m*M* imidazole. The sample buffer was then exchanged to 20 m*M* Tris-HCl pH 8.0 and 50 m*M* NaCl using Amicon spin filters (Millipore), and the preparation was applied to a HiPrep 16/10 Q FF ion exchange column (GE Healthcare), which was equilibrated with the same buffer. The His-A1 protein did not bind to the column under these conditions, whereas the majority of contaminants did. Finally, fractions containing His-A1 were pooled, concentrated, and loaded on a Superdex 200 10/300 GL gel filtration column (GE Healthcare), which was equilibrated

**Table I.** *Crystallographic data collection, scaling, and refinement statistics*

| Dataset | Monoclinic | Hexagonal |
|---|---|---|
| Data collection and scaling | | |
| Beamline | ESRF ID29 | MAX-Lab I911–2 |
| Spacegroup | $P2_1$ | $P6_322$ |
| Cell parameters | $a = 44.01$ Å<br>$b = 49.12$ Å<br>$c = 44.26$ Å<br>$\beta = 118.41°$ | $a = 69.11$ Å<br>$c = 167.30$ Å |
| Wavelength (Å) | 0.9762 | 1.0387 |
| Resolution (Å) | 49.15–1.76 | 40.79–2.90 |
| Highest resolution bin (Å) | 1.86–1.76 | 3.06–2.90 |
| $R_{merge}$ | 0.078 (0.334) | 0.098 (0.543) |
| Total number of observations | 55321 | 54994 |
| Number of unique reflections | 16414 | 5773 |
| $I/\sigma(I)$ | 10.0 (3.2) | 17.5 (3.8) |
| Completeness (%) | 99.2 (99.5) | 100.0 (100.0) |
| Multiplicity | 3.4 (3.3) | 9.5 (10.0) |
| Refinement | | |
| $R_{work}$ | 0.174 | 0.213 |
| $R_{free}$ | 0.242 | 0.297 |
| Average B factor (Å$^2$) | 17.153 | 41.116 |
| Number of atoms | | |
| Protein | 1463 | 1411 |
| Solvent | 154 | 19 |
| RMS deviations from ideal | | |
| Bond lengths (Å) | 0.023 | 0.012 |
| Bond angles (°) | 1.925 | 1.368 |
| Ramachandran plot[27] | | |
| Residues in favored regions (%) | 97.8 | 95.4 |
| Residues in allowed regions (%) | 100.0 | 99.4 |

Values in parentheses are given for the highest resolution shell.

with 20 m*M* Tris-HCl pH 8.0. Fractions containing His-A1 were pooled, concentrated to 10 mg/mL, and stored at −20°C until use.

### *Crystallization and data collection*

The His-A1 protein was initially crystallized using the sitting drop vapor diffusion technique by mixing 1 μL of the protein solution (10 mg/mL) with 1μL of the well solution (0.1*M* Tris-HCl pH 8.5, 40% PEG 300). Plate-shaped crystals (the monoclinic form) appeared after 3–6 days at room temperature (298 K) and reached maximum dimensions of 0.3 × 0.1 mm. For data collection, crystals were flash-frozen in liquid nitrogen without additional cryoprotectant.

After optimization of crystallization conditions, slightly thicker crystals were obtained using 0.1*M* Tris-HCl pH 8.5, 20% PEG 300, and 10% PEG 2000 MME as the well solution. These crystals were less fragile, had less anisotropic diffraction and were used for heavy atom compound soaks. To prepare the mercury derivative, crystals were soaked in a mother liquor containing 20 m*M* $Hg(NO_3)_2$ for 30 min, followed by backsoaking in the original mother liquor for 10 s. For iodine derivatization, mother liquor containing 0.1*M* $I_2$ in 0.1*M* KI was prepared, the undissolved iodine was removed by centrifugation, and the resulting iodine-saturated solution was used for soaking the crystals overnight. Crystals were flash-frozen without backsoaking.

When the structure of the A1 domain in the monoclinic crystal form was already solved, a hexagonal crystal form was discovered when any buffer was omitted from the crystallization drop (40% PEG 300 in water). Crystals appeared after 2–3 days at room temperature and grew bigger for about 1 week, reaching maximum size of 0.2 mm.

Crystal diffraction data were collected at European Synchrotron Radiation Facility (ESRF) and MAX-lab as indicated in Table I and Supporting Information Table 1.

### *Structure determination*

Data were indexed with MOSFLM[28] and scaled using SCALA.[29] For the monoclinic crystal form, native and derivative datasets were scaled with SCALEIT[30] and merged using CAD from the CCP4 suite.[31] The position of the first mercury atom was calculated manually from the strongest peak in the Harker section of the isomorphous difference Patterson map. The coordinates of the mercury atom were input into MLPHARE[32] and used to locate the remaining mercury and iodine atoms. Heavy atom refinement and phasing was performed in SHARP[33] and was followed by solvent flattening in SOLOMON.[34] From the resulting map, a partial model was built by BUCANEER[35] that was included to provide extra phase information in a second SHARP and SOLOMON run. The resulting map was used to build an improved model with BUCANEER that served as a starting point for manual model building in COOT[36] using the high-resolution native data. Refinement was performed using REFMAC.[37] The structure of the hexagonal form was solved by molecular replacement with MOLREP[38] using coordinates of the A1 domain in the monoclinic crystal form as a search model, followed by model building in COOT and refinement with REFMAC. Scaling and refinement statistics for native datasets are presented in Table I; detailed phasing statistics are given in Supporting Information Table 1.

Atomic coordinates were deposited in the Protein Data Bank under accession codes 3RLK (monoclinic crystal form) and 3RLC (hexagonal crystal form).


## Acknowledgments

The authors thank the staff at ESRF and MAX-Lab for their help during data collection and Anna Janson for collecting the mercury derivative datasets.

## References


1. Weber K, Konigsberg W, Proteins of the RNA phages. In: Zinder ND, Ed. (1975) RNA phages. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, pp 51–84.
2. Valegård K, Liljas L, Fridborg K, Unge T (1990) The three-dimensional structure of the bacterial virus MS2. Nature 345:36–41.
3. Steitz JA (1968) Identification of the A protein as a structural component of bacteriophage R17. J Mol Biol 33:923–936.
4. Kozak M, Nathans D (1971) Fate of maturation protein during infection by coliphage MS2. Nat New Biol 234: 209–211.
5. Brinton CC, Gemski P, Carnahan J (1964) A new type of bacterial pilus genetically controlled by the fertility factor of *E. coli* K 12 and its role in chromosome transfer. Proc Natl Acad Sci USA 52:776–783.
6. Winter RB, Gold L (1983) Overproduction of bacteriophage Qβ maturation (A2) protein leads to cell lysis. Cell 33:877–885.
7. Karnik S, Billeter M (1983) The lysis function of RNA bacteriophage Qβ is mediated by the maturation (A2) protein. EMBO J 2:1521–1526.
8. Weiner AM, Weber K (1971) Natural read-through at the UGA termination signal of Qβ coat protein cistron. Nat New Biol 234:206–209.
9. Hofstetter H, Monstein HJ, Weissmann C (1974) The readthrough protein A1 is essential for the formation of viable Qβ particles. Biochim Biophys Acta 374:238–251.
10. Golmohammadi R, Fridborg K, Bundule M, Valegård K, Liljas L (1996) The crystal structure of bacteriophage Qβ at 3.5 Å resolution. Structure 4:543–554.
11. Vasiljeva I, Kozlovska T, Cielens I, Strelnikova A, Kazaks A, Ose V, Pumpens P (1998) Mosaic Qβ coats as a new presentation model. FEBS Lett 431:7–11.
12. DeLano WL (2002) The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific.
13. Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucl Acids Res 38:W545–W549.
14. Berisio R, Loguercio S, De Simone A, Zagaria A, Vitaglianoa L (2006) Polyproline helices in protein structures: a statistical survey. Protein Pept Lett 13:847–854.
15. Hasson MS, Muscate A, McLeish MJ, Polovnikova LS, Gerlt JA, Kenyon GL, Petsko GA, Ringe D (1998) The crystal structure of benzoylformate decarboxylase at 1.6 Å resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes. Biochemistry 37: 9918–9930.
16. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucl Acids Res 36: W197–W201.
17. Bollback JP, Huelsenbeck JP (2001) Phylogeny, genome evolution, and host specificity of single-stranded RNA bacteriophage (family Leviviridae). J Mol Evol 52: 117–128.
18. Furuse K, Distribution of coliphages in the environment: general considerations. In: Goyal SM, Gerba CP, Bitton G, Eds. (1987) Phage ecology. New York: Wiley, pp 87–124.
19. Priano C, Arora R, Butke J, Mills DR (1995) A complete plasmid-based complementation system for RNA coliphage Qβ: three proteins of bacteriophages Qβ (group III) and SP (group IV) can be interchanged. J Mol Biol 249:283–297.
20. Strauss EG, Kaesberg P (1970) Acrylamide gel electrophoresis of bacteriophage Qβ: electrophoresis of the intact virions and of the viral proteins. Virology 42: 437–452.
21. Radloff RJ, Kaesberg P (1973) Electrophoretic and other properties of bacteriophage Qβ: the effect of a variable number of read-through proteins. J Virol 11: 116–128.
22. Williamson MP (1994) The structure and function of proline-rich regions in proteins. Biochem J 297: 249–260.
23. Kay BK, Williamson MP, Sudol M (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J 14:231–241.
24. Köhler SD, Weber A, Howard SP, Welte W, Drescher M (2010) The proline-rich domain of TonB possesses an extended polyproline II-like conformation of sufficient length to span the periplasm of Gram-negative bacteria. Protein Sci 19:625–630.
25. Toropova K, Stockley PG, Ranson NA (2011) Visualising a viral RNA genome poised for release from its receptor complex. J Mol Biol 408:408–419.
26. Kozlovska TM, Cielens I, Dreilinna D, Dislers A, Baumanis V, Ose V, Pumpens P (1993) Recombinant RNA phage Qβ capsid particles synthesized and self-assembled in *Escherichia coli*. Gene 137:133–137.
27. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21.
28. Leslie AGW (1992) Recent changes to the MOSFLM package for processing film and image plate data. Joint CCP4 + ESF-EAMCB. Newslett Protein Crystallogr 26.
29. Evans PR (1997) Scala. Joint CCP4 + ESF-EAMCB. Newslett Protein Crystallogr 33:22–24.
30. Howell L, Smith D (1992) Normal probability analysis. J Appl Cryst 25:81–86.
31. Collaborative Computational Project Number 4 (1994) The CCP4 suite: programs for protein crystallography (1994) Acta Crystallogr D Biol Crystallogr 50:760–763.
32. Otwinowski Z, Maximum likelihood refinement of heavy atom parameters. In: Wolf W, Evans PR, Leslie AGW, Eds. (1991) Proceedings of the CCP4 study weekend. Warrington, UK: Science and Engineering Research Council, Daresbury Laboratory, pp 80–85.
33. Bricogne G, Vonrhein C, Flensburg C, Schiltz M, Paciorek W (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr D Biol Crystallogr 59:2023–2030.
34. Abrahams JP, Leslie AG (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr D Biol Crystallogr 52:30–42.
35. Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr D Biol Crystallogr 62:1002–1011.
36. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66:486–501.
37. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53:240–255.
38. Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. J Appl Cryst 30: 1022–1025.

# Paper II


# Crystal Structure of the Bacteriophage Qβ Coat Protein in Complex with the RNA Operator of the Replicase Gene

**Janis Rumnieks and Kaspars Tars**

***Biomedical Research and Study Center,*** *Ratsupites 1, Riga LV1067, Latvia*

***Correspondence to Kaspars Tars:*** *Biomedical Research and Study Center, Ratsupites 1, Riga LV1067, Latvia.* *kaspars@biomed.lu.lv*

http://dx.doi.org/10.1016/j.jmb.2013.08.025

***Edited by F. Allain***

## Abstract

The coat proteins of single-stranded RNA bacteriophages specifically recognize and bind to a hairpin structure in their genome at the beginning of the replicase gene. The interaction serves to repress the synthesis of the replicase enzyme late in infection and contributes to the specific encapsidation of phage RNA. While this mechanism is conserved throughout the *Leviviridae* family, the coat protein and operator sequences from different phages show remarkable variation, serving as prime examples for the co-evolution of protein and RNA structure. To better understand the protein–RNA interactions in this virus family, we have determined the three-dimensional structure of the coat protein from bacteriophage Qβ bound to its cognate translational operator. The RNA binding mode of Qβ coat protein shares several features with that of the widely studied phage MS2, but only one nucleotide base in the hairpin loop makes sequence-specific contacts with the protein. Unlike in other RNA phages, the Qβ coat protein does not utilize an adenine-recognition pocket for binding a bulged adenine base in the hairpin stem but instead uses a stacking interaction with a tyrosine side chain to accommodate the base. The extended loop between β strands E and F of Qβ coat protein makes contacts with the lower part of the RNA stem, explaining the greater length dependence of the RNA helix for optimal binding to the protein. Consequently, the complex structure allows the proposal of a mechanism by which the Qβ coat protein recognizes and discriminates in favor of its cognate RNA.



## Introduction

For bacteriophages of the *Leviviridae* family, the single-stranded RNA genome does not merely encode phage proteins but also forms extensive secondary and tertiary structures that are critical for RNA replication, regulation of phage protein synthesis and assembly of virus particles [1]. The function of three out of the four phage proteins—replicase, maturation and coat—is intricately linked with specific RNA structures that they recognize and bind to at some point in the viral life cycle [2–4]. The *Leviviridae* coat protein adopts a fold observed only in this virus family with an N-terminal β hairpin, a five-stranded antiparallel β sheet and two C-terminal α helices [5]. The helices from two coat protein molecules interlock to form a very stable dimer with a continuous ten-stranded β sheet that lines the interior of the capsid and forms the RNA-binding surface of the protein. Although the primary role of the coat protein is to encapsulate the genome, it also acts as a translational repressor that regulates the synthesis of the replicase. The operator is an RNA sequence of approximately 20 nucleotides at the beginning of the replicase gene that folds into a stem–loop structure and comprises the initiation codon of the gene [6,7]. Specific binding of the coat protein to the RNA hairpin effectively shuts down the translation of the replicase when the coat protein accumulates in the infected cells [8] and marks the genome for packaging into capsids [9]. This regulatory mechanism is highly conserved within the *Leviviridae* family, but similarities in operator hairpins are limited to a stem structure of seven to eight base pairs with an unpaired base in it, whereas the number and identity of nucleotides in the loop as well as the position of the bulged nucleotide vary from phage to phage. The

bulged nucleotide is usually an adenosine and is located on the 5′ side of the stem, and the loops of all of the studied operators contain another adenosine that is critical for specific interactions with the coat proteins.

The interaction between the coat protein and the operator of phage MS2 has been extensively studied both biochemically [10–12] and structurally [13–17], making it one of the best characterized protein–RNA interactions to date. Three-dimensional structures of coat protein–operator complexes have also been solved for the phages PRR1 [18] and PP7 [19]. The MS2 and PRR1 operators differ primarily in the size of the loop, and the respective coat proteins utilize a very similar RNA binding mode. The recognition mechanism in both cases involves the binding of two adenine bases, namely, the unpaired one in the stem and another in the loop, to symmetrical adenine-recognizing pockets in the protein dimer. The complex is further stabilized by aromatic stacking that extends from the helical RNA stem via two bases in the hairpin loop to a conserved tyrosine side chain in the coat protein. The operator of phage PP7 is remarkably different from MS2 and uses a distinct RNA binding mode. Nonetheless, the PP7 coat protein also uses symmetrical pockets to bind two adenine bases in the bulge and the loop despite the fact that the pockets are very different from those found in MS2.

The bacteriophage Qβ is distantly related to MS2 with their coat proteins only about 20% identical. Both coat proteins preferentially bind their cognate translational operators, which are also rather different (Fig. 1). For strong binding to the MS2 coat protein, the operator helix needs to be at least five base pairs long and contain an unpaired purine nucleotide two base pairs prior to a four-nucleotide loop with adenosines as the first and last nucleotides and a pyrimidine nucleotide at the penultimate position [12]. For high-affinity binding to the Qβ coat protein, the operator requires a three-nucleotide loop and an eight-base-pair stem with a bulged nucleotide four base pairs from the loop [20]. The only critical nucleotide in the loop is an adenosine at the last position, whereas the unpaired adenosine in the stem can be mutated or removed altogether with a rather minor decrease in affinity [21]. Despite the differences, several facts suggest that the RNA binding modes of MS2 and Qβ coat proteins are nevertheless related. Although the overall sequence identity is low, the three-dimensional structure of the two proteins is very similar, and many of the residues that are involved in RNA binding in MS2 are conserved in Qβ [22]. Furthermore, MS2 and Qβ coat protein mutants that are able to tightly bind the operator of the other phage have been isolated [21,23], but analogous experiments were unsuccessful with PP7 [24].

The mechanism by which the MS2 coat protein discriminates between the MS2 and Qβ operators is well understood. Genetic studies have shown that amino acid changes at residues 87 and 89 of the MS2 coat protein confer an ability to bind the Qβ operator with high affinity [23]. The molecular mechanism for this discrimination has been elucidated by solving crystal structures of the mutant coat proteins bound to the Qβ operator [25]. In the wild-type MS2 coat protein, Asn87 forms a hydrogen bond with the –5 uracil base in the cognate operator, while the Qβ operator has the bulkier adenine base in the equivalent +7 position, which results in a steric clash with the asparagine side chain. Mutation of the asparagine to a serine or alanine decreases the affinity for the MS2 operator because the hydrogen

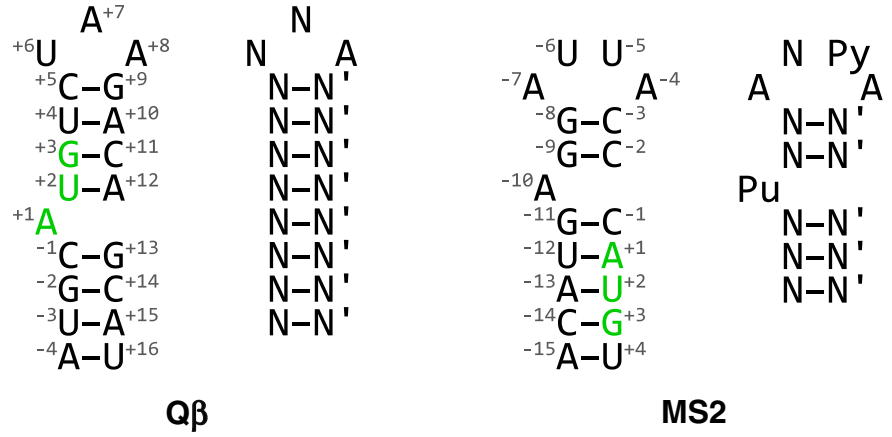


**Fig. 1.** Secondary structure of the Qβ and MS2 operators. For both phages, the wild-type operator sequence is shown on the left and the minimal sequence requirements for binding to the coat protein are shown on the right (Py, pyrimidine; Pu, purine; N, any nucleotide; N′, a nucleotide complementary to N). For the wild-type operators, the initiation codons of the replicase are marked in green, and nucleotide positions relative to the start of the replicase ORF are indicated as superscript numbers next to the bases.

bond interaction is lost but improves binding of the Qβ operator by allowing sufficient space to accommodate the adenine base. A second mutation of Glu89 to a lysine eliminates an unfavorable electrostatic repulsion with the phosphate backbone of the Qβ operator and instead provides an additional contact that further improves the binding.

Although a genetic study of the Qβ coat protein [21] demonstrated that the RNA binding modes of the Qβ and MS2 coat proteins are similar, the molecular mechanism that allows the Qβ coat protein to recognize and discriminate its cognate operator has remained unknown. To address this issue, we solved the crystal structure of the Qβ coat protein in complex with its operator hairpin, which we present here and compare to the coat protein–RNA complexes found in other RNA phages.

## Results and Discussion

### Design and structure determination of assembly-deficient Qβ coat protein in complex with RNA

Previous work with MS2 that led to numerous protein–RNA complex structures relied on the ability to soak small RNA hairpins into pre-crystallized capsids via pores that are present at their 3-fold and 5-fold symmetry axes. However, the same approach failed when applied to Qβ, which was attributed to the fact that the FG loops from neighboring Qβ coat protein dimers are covalently linked to each other with disulfide bonds that could in turn restrict RNA diffusion into capsids. To address this issue, we crystallized Qβ capsids assembled from modified coat proteins that had cysteines in the FG loop mutated to glycines and used these crystals for the RNA soaking experiments. Unfortunately, still no bound RNA was detected in the electron density maps, suggesting that the crystallization conditions (0.4 M NaCl at pH 7.5) could be suboptimal for RNA binding and that the approach of soaking capsid crystals with RNA would not be successful with Qβ.

The structure of the PP7 coat protein in complex with its operator was determined via a different approach, namely, by crystallizing the RNA together with coat protein dimers that were lacking the FG loops and therefore incapable of assembling into capsids. However, our initial attempts to truncate the FG loop in Qβ resulted in a largely insoluble protein; therefore, an approach was devised to introduce other amino acid changes into the coat protein that would prevent it from assembling into particles. Examination of the Qβ capsid structure suggested Asn129 as a good candidate for mutagenesis as its side chain forms two hydrogen bonds with the main chain of the adjacent dimer; thus, introduction of a bulkier side chain at this position would both destroy the bonding and cause a steric clash with the nearby chain. A similar situation was observed for Pro42 in the CD loop where substitution with a longer side chain would likely result in a collision with the neighboring dimer. Mutation of the two residues to arginines (Pro42Arg, Asn129Arg) in the cysteine-less mutant (Cys74Gly, Cys80Gly) indeed resulted in a protein that produced a highly soluble and homogenous dimeric species suitable for structural studies. The coat protein–RNA complex was obtained by mixing purified dimers and RNA in a molar ratio of 1:1.2, and the mixture was immediately subjected to crystallization. Crystals that diffracted to 2.4 Å resolution were obtained, and the structure was solved by molecular replacement.

### Quality of the model

The final model (Fig. 2a) contains one Qβ coat protein dimer (chains A and B) and one RNA molecule (chain R). There are no crystal contacts close to the protein–RNA interface, suggesting that the model represents a biologically relevant structure. The unassembled dimer adopts a conformation highly similar to that found in the crystallized phage capsids [22], with a root-mean-square deviation (rmsd) of $C^{\alpha}$ atoms of 0.8 Å. Notably, the EF loops of the assembly-deficient dimer make contacts with RNA and can be reliably modeled, whereas they were only partly visible in the virus structure. In contrast, the FG loops (residues 74–84 of chain A and residues 75–83 of chain B) are disordered in the unassembled dimer and were not included in the final model. Electron density for the whole RNA molecule (20 nucleotides) was clearly visible (Fig. 2b), and the complete hairpin was modeled without breaks. Interestingly, the stems of two neighboring RNA hairpins stack end-to-end in the crystal in a somewhat similar manner as in the PP7 coat protein–RNA structure. This arrangement likely restricts their movement and contributes to the well-defined electron density observed for the RNA. The final model also includes six zinc ions from the crystallization solution. One of them is tetrahedrally coordinated between Asp102 and Glu103 of two adjacent dimers where it provides an important crystal contact, while the others are located in the proximity of the RNA.

### Structure of Qβ coat protein–operator complex

The experimentally observed structure of the RNA hairpin is consistent with the predicted secondary structure and consists of an eight-base-pair stem, a three-nucleotide loop and an unpaired adenosine in the stem. The stem adopts a canonical A-form helical conformation with ribose puckers in the C3′-*endo* conformation except for loop nucleotides

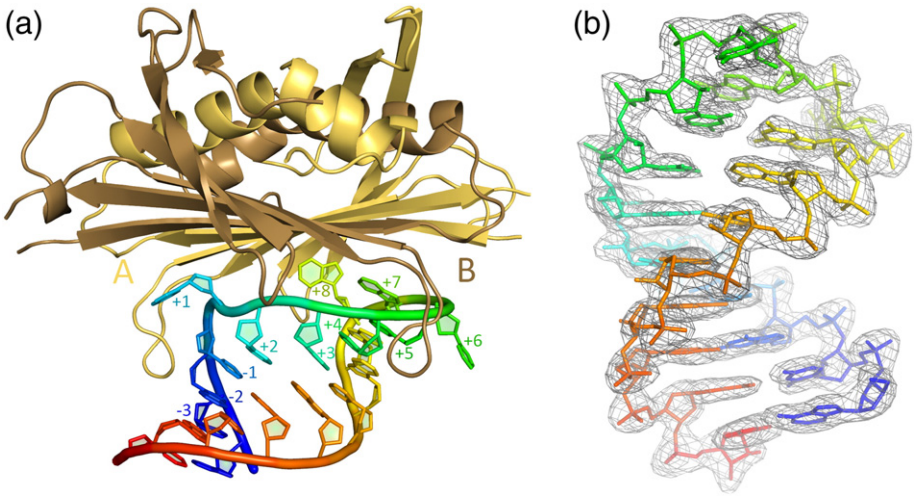


**Fig. 2.** Three-dimensional structure of the Qβ coat protein–operator complex. (a) Overall structure of the complex. The coat protein dimer is represented in light orange (monomer A) and light brown (monomer B), and the RNA is rainbow-colored blue (5′ end) to red (3′ end). Nucleotide positions relative to the first nucleotide of the replicase initiation codon are indicated next to the bases. (b) A close-up view of the RNA hairpin. The RNA is shown in a stick representation colored as in (a) and shown as modeled into a $2F_o - F_c$ electron density contoured at 1.1 σ. This figure and Figs. 3–5 were prepared using PyMOL [26].

A+7 and A+8, which adopted more of a C2′-*endo* conformation. The overall conformation of the operator hairpin (nucleotides +2 to +12) in the cognate complex is very similar to that observed for the Qβ operator in complex with the MS2 coat protein mutant [25] with an rmsd of 0.8 Å (Fig. 3). The majority of the contacts between the protein and RNA are sequence-independent interactions between the sugar-phosphate backbone of the RNA and the EF loop and β strand F of both coat protein monomers (Table 1). The adenine base of the A+8 nucleotide fits into an adenine-binding pocket formed by Val32, Thr49, Ser51, Gln65 and Lys67 of chain A in the coat protein dimer. The base of the A+7 nucleotide is stacked between C+5 in the stem and the aromatic side chain of Tyr89 of the A chain. In addition, the hydroxyl group of the tyrosine forms a stabilizing hydrogen bond with an oxygen atom in the phosphate backbone. In the crystallized complex, there is also a zinc ion from the crystallization solution coordinated between the OD2 oxygen of Asp91 of the A monomer, the N1 nitrogen of A+7 and two water molecules. This interaction is not physiologically relevant because both atoms would act as hydrogen acceptors at physiological pH, and the AspA91 side chain would not be able to form a hydrogen bond with the adenine base under these conditions. The base of the last loop nucleotide, U+6, points away from the protein and does not make any contacts with it. The unpaired A+1 nucleotide bulges out from the stem and stacks with Tyr89 in chain B of the coat protein. There seem to be no additional stabilizing interactions involving the base, but the phosphate oxygen of A+1 forms an electrostatic interaction with the side chain of Lys63 in

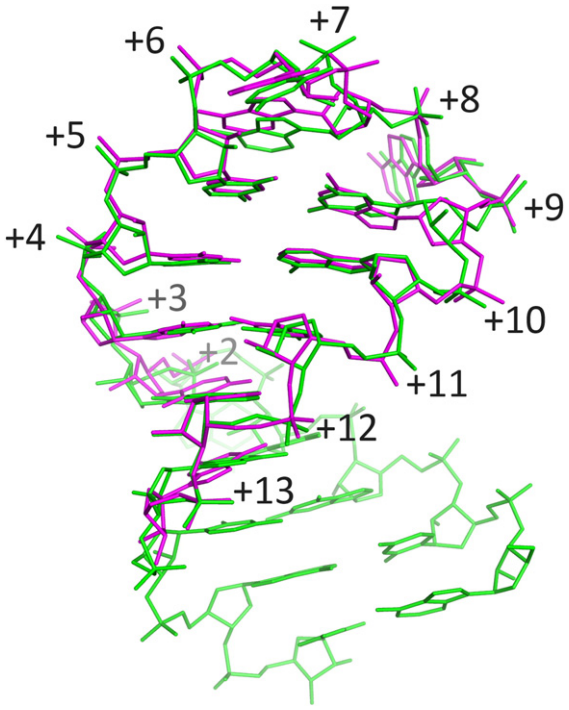


**Fig. 3.** Structure of the Qβ operator bound to the Qβ coat protein and the MS2 coat protein mutant. Although the upper part of the hairpin adopts a remarkably similar conformation in both cases, the lower part, including the bulged adenosine, is disordered in the complex with the MS2 mutant. Nucleotide numbers as of Fig. 1 are indicated next to the phosphates. The operator hairpin from the cognate Qβ complex is represented in green while that bound to the Asn87Ser MS2 coat protein mutant (PDB entry 1ZSE) is represented in magenta.

**Table 1.** Hydrogen bonds and electrostatic interactions between protein and RNA in the Qβ coat protein–RNA operator complex

| RNA | | Protein | | Distance (Å) |
|---|---|---|---|---|
| Residue | Atom | Residue | Atom | |
| U–3 | O2′ | AsnA58 | ND2 | 2.7 |
| G–2 | OP1 | ArgA59 | NH1 | 3.8 |
| C–1 | OP1 | LysA63 | NZ | 3.3 |
| | OP2 | LysA63 | NZ | 3.4 |
| A+1 | OP2 | LysA63 | NZ | 3.0 |
| G+3 | OP1 | LysB67 | NZ | 2.9 |
| U+4 | OP1 | ArgB59 | NE | 2.7 |
| | OP2 | LysB63 | NZ | 2.4 |
| C+5 | OP1 | LysB60 | N | 2.9 |
| | | AsnB61 | N | 2.8 |
| | OP2 | LysB63 | NZ | 2.8 |
| U+6 | OP1 | LysB60 | NZ | 2.9 |
| | OP2 | AsnB61 | ND2 | 2.8 |
| A+7 | OP2 | TyrA89 | OH | 2.4 |
| A+8 | O2′ | AsnA30 | OD1 | 3.0 |
| | N1 | SerA51 | OG | 2.8 |
| | N6 | ThrA49 | OG1 | 3.2 |
| | | GlnA65 | O | 2.9 |
| | N7 | ThrA49 | OG1 | 2.6 |

the A chain, and additional contacts with sugars and phosphates of C–1, G–2 and U–3 nucleotides in the lower part of the stem stabilize the hairpin in the observed orientation.

## Comparison of RNA binding between Qβ and MS2

The 970-Å$^2$ interface between the Qβ coat protein and its operator hairpin is close to the value reported for PP7 (950 Å$^2$) and slightly larger than that of MS2 (830 Å$^2$), but the overall structure of the complex is undoubtedly more similar to that of MS2. The top part of the Qβ hairpin that faces the protein (nucleotides +3 to +8) adopts a conformation that is remarkably similar to that of the MS2 operator (nucleotides –9 to –4, respectively) with an rmsd of 1.1 Å, which supports the hypothesis that the two proteins share a similar RNA binding mode. The number of hydrogen bonds and electrostatic interactions between the protein and RNA is similar in Qβ and MS2; however, in MS2, a higher proportion of the interactions involve contacts with the nucleotide bases rather than the sugar-phosphate backbone (Fig. 4a). The adenine-binding pocket of the Qβ coat protein is almost identical with that of MS2, and all of the base–protein interactions within the pocket are the same in the two phages. However, the nearby interaction between LysA43 and the phosphate backbone in MS2 is not preserved as the equivalent ArgA47 in Qβ is too far away from the RNA (4.4 Å) to make any significant contribution to the interaction. The similarities in RNA binding of the two proteins extend to the A+7 nucleotide, which in Qβ is stacked with TyrA89 while in MS2 an analogous interaction is found between U–5 and TyrA85, and a contact between the hydroxyl of the tyrosine and a phosphate of the RNA backbone is also conserved. Like U–6 of MS2, the U+6 in Qβ points away from the protein and does not make contacts with it. Finally, residues AsnB61 and LysB63, which make interactions with the sugar-phosphate backbone in Qβ, are conserved and provide the same function in MS2.

Away from the hairpin loop, the differences in protein–RNA interactions in the two phages become more pronounced. In the lower part of the hairpin, only a single electrostatic interaction exists between Arg49 of the A monomer and the –13 phosphate in MS2, but in Qβ the arginine residue is not conserved and interactions involving AsnA58, ArgA59 and LysA63 take place instead. The additional contacts are possible due to an extended EF loop that, in Qβ, is two residues longer than in MS2. However, the most profound difference between Qβ and other RNA phages involves the interaction with the bulged adenosine in the stem of the hairpin. In MS2, the bulged A–10 base fits into the same pocket as A–4 in the other monomer, albeit in a different orientation; however, in Qβ, the other adenine-binding pocket is empty, and the A+1 base is stacked with Tyr89 of the other monomer (Fig. 4b). This configuration has not been observed in any other coat protein–RNA complex and thus represents a novel mechanism for accommodating an unpaired base in the stem.

## RNA binding discrimination of Qβ coat protein

The conformation of the β sheet that makes up the RNA-binding surface of the coat protein is very similar in MS2 and Qβ with an rmsd of 0.9 Å when $C^\alpha$ atoms from strands D, E, F and G of the two proteins are superimposed. In the superimposed protein–RNA complexes, the A+8/A–4 bases, the adenine-binding pockets and other conserved RNA-binding residues align remarkably well. A possible RNA discrimination mechanism for the Qβ coat protein can therefore be modeled with some confidence by combining protein coordinates from the Qβ complex with RNA coordinates from the fitted MS2 complex.

In the modeled Qβ coat protein–MS2 operator complex, the A–10 and A–4 bases fit very well into the adenine-binding pockets of the Qβ coat protein, and many of the interactions with the RNA backbone in the upper stem seem to be preserved. AsnB61 and LysB63 in Qβ occupy positions equivalent to AsnB55 and LysB57 in MS2, and although LysB60 of the Qβ coat protein is not conserved in MS2, there is no reason to exclude an interaction with the MS2 operator. There appear to be some differences regarding the interactions involving Arg49, which is found in the wild-type MS2 complex but is not conserved in Qβ. In the wild-type MS2 complex, Arg49 in the A monomer forms a salt bridge with the –13 phosphate, but this interaction is lost with the Qβ

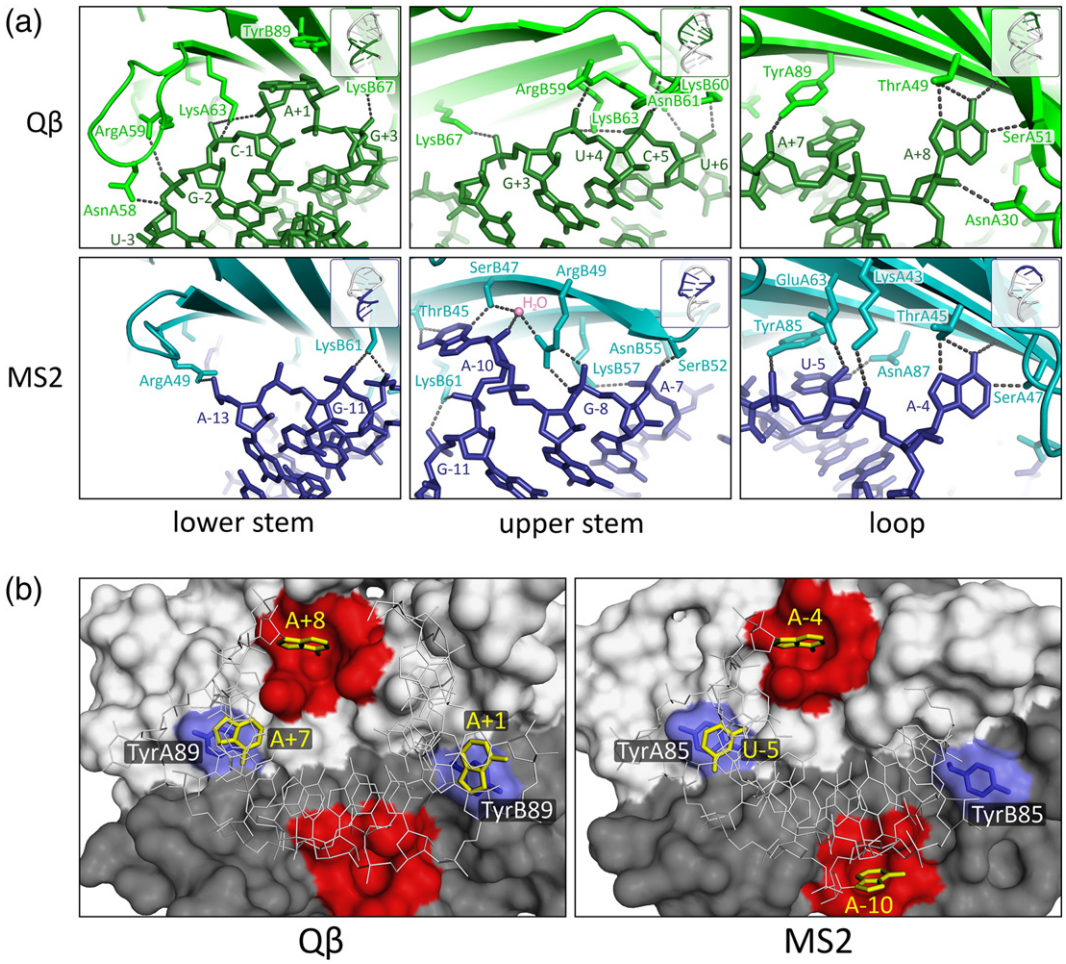


**Fig. 4.** Differences in binding of the Qβ and MS2 coat proteins to their cognate operators. (a) Close-up views of the protein–RNA interactions in Qβ and MS2. Hydrogen bonds and electrostatic interactions in the lower and upper parts of the stem and the hairpin loops are indicated as gray broken lines. Side chains of interacting amino acid residues and nucleotides are labeled as in Table 1. The insets on top right highlight the approximate region of the operator hairpin that is visible in the particular close-up. (b) Comparison of protein–RNA interactions in Qβ and MS2 involving the loop and the bulged adenosine. The solvent-accessible surfaces of Qβ and MS2 coat protein dimers are shown in different shades of gray as for A and B monomers. The adenine-binding pockets are shown in red, while the tyrosine residues that stack with RNA bases are colored blue. The RNA is shown in light gray as a stick model except for the bases that occupy the adenine-binding pockets or stack with the tyrosine side chains, which are shown in yellow. In Qβ, only one of the symmetrical adenine-binding pockets is occupied and tyrosines from both monomers participate in base stacking. In contrast, both pockets are occupied by adenine bases in MS2, while only a single tyrosine is involved in base stacking.

coat protein, which has a serine residue at the equivalent position. In the B monomer, Arg49 forms a salt bridge with the –8 phosphate and additionally coordinates a water molecule that forms a hydrogen bond with the O2′ atom of the A–10 ribose. In Qβ, the side chain of ArgB59 lies in approximately the same place as ArgB49 in MS2 and partly serves the same function by providing an electrostatic interaction with the phosphate of U+4. This interaction would likely be preserved in the complex with the MS2 operator, but the arginine side chain would be too far away from the A–10 nucleotide to allow interactions similar to those observed in the cognate MS2 complex. Consequently, this might contribute to the weaker binding of the MS2 operator to the Qβ coat protein.

Another reason for the poor binding of the MS2 operator likely involves the –5 uracil base in the loop. The side chain of TyrA89 that stacks with A+7 in Qβ is tilted by approximately 20° compared to TyrA85 in MS2. This orientation is observed both in complex with the RNA and in assembled capsids and is unlikely to switch to an MS2-like conformation due to the proximity of the GlnA69 and GlnA87 side chains. As a result, planes going through the U–5 base of the MS2 operator and the side chain of TyrA89 in Qβ coat protein would not be parallel, which could lead

to impaired binding of the RNA. In addition, the interaction between U–5 and AsnA87 that is present in the cognate MS2 complex is lost. The corresponding amino acid in Qβ is AspA89, and repulsion between the acidic side chain and the O2 carbonyl of the uracil base would prevent an analogous interaction with the Qβ coat protein. This is consistent with the observation that the interaction between an Asp91Asn Qβ coat protein mutant and the MS2 operator is 20 times stronger than with wild-type Qβ coat protein [21]. In contrast to aspartic acid, the asparagine side chain would permit formation of a hydrogen bond between the protein and RNA and result in the observed improvement in binding.

It is interesting to note that the Qβ coat protein is able to bind the operator of the closely related phage SP with the same affinity as the cognate one [27]. A notable difference between the two hairpins is that the SP operator contains a C-A base pair in the upper part of the stem. It was further demonstrated [27] that Qβ coat protein can tolerate several other base-pair mismatches in the stretch between the bulged adenosine and the loop, suggesting that the integrity of the upper part of the helix is not critical for high-affinity binding. Accommodation of a non-Watson–Crick base pair in an RNA hairpin has been visualized in the crystal structure of an RNA aptamer bound to the MS2 coat protein [28], which showed that a non-canonical G-A base pair does not result in the disruption of the helical stem. In the absence of similar experimentally determined structures for Qβ, it seems reasonable to assume that single-base-pair mismatches in the Qβ operator would be tolerated in a similar manner as in the MS2 aptamer. Apparently, the interactions between the protein and the RNA backbone on the 5′ side of the stem are sufficiently strong to hold the RNA in the protein-bound conformation and render perfect base pairing in the stem redundant.

### Effects of hairpin loop size and bulged nucleotides on RNA binding

In Qβ, the size of the hairpin loop plays an important role for optimal binding of the cognate operator, as the addition of an extra nucleotide in the loop severely reduces the affinity [20]. When the stems of the MS2 and Qβ operators are superimposed (residues +10 to +15 for Qβ and residues –3 to +3 in MS2), substantial differences in loop conformations are evident due to the extra base pair at the top of the Qβ hairpin (Fig. 5a). However, in the superimposed protein–RNA complexes of the two phages, the smallest conformational differences are observed in the region

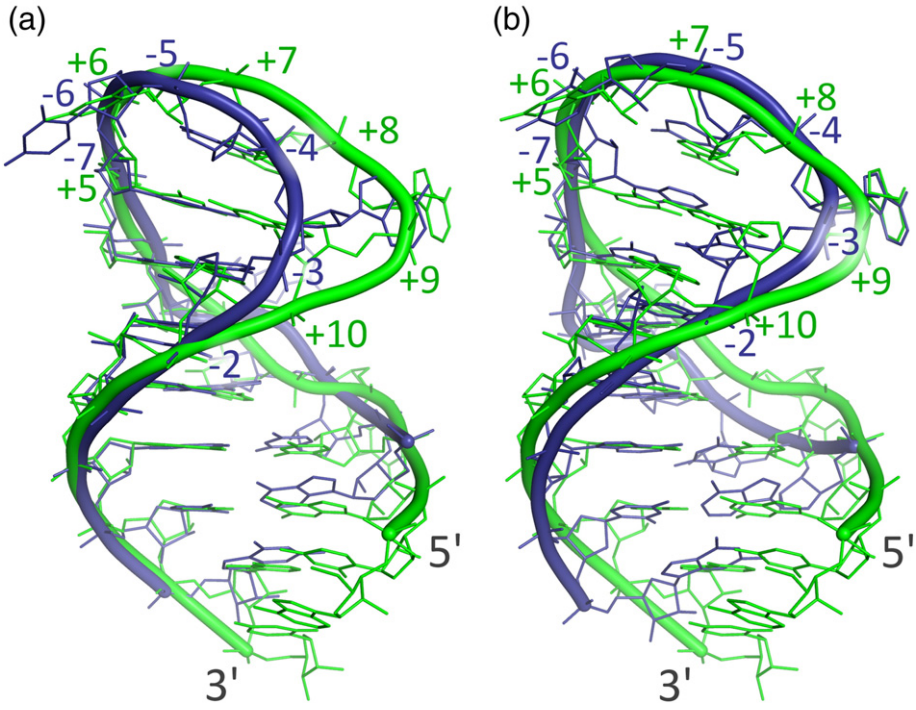


**Fig. 5.** Conformational differences of the Qβ and MS2 operators. Superimposition of the helical stems (a) demonstrates the differences in hairpin loop conformations of the two operators. Superimposition of the RNA-binding residues of the two cognate protein–RNA complexes (b) results in different relative orientations of the stems that, in turn, cause the phosphate backbones of the two RNAs to follow different paths. The Qβ (green) and MS2 (blue) operators are shown as stick models with the phosphate backbones represented by ribbon traces. Nucleotides in the loop and the upper part of stem are numbered as of Fig. 1 and indicated next to the phosphates.

comprising the loop and two preceding nucleotides and not in the stems (Fig. 5b). Consequently, the different-sized loops impose different relative orientations of the RNA stems that appear to play some role in optimal binding of the RNA. Biochemical studies have shown that the Qβ coat protein requires a longer RNA stem than MS2 for high-affinity binding [20], which is likely necessary to compensate for the lack of some of the interactions in the upper stem. The length dependence is explained by the EF loops, which are longer in the Qβ coat protein than their MS2 counterparts and make contacts with the lower stem; however, binding of a hairpin with a three-nucleotide loop would position the phosphate backbone in a more favorable orientation regarding the interactions than the binding of a four-nucleotide loop. The conformation with a three-nucleotide loop also restricts the ability to accommodate bulged nucleotides in the stem except those at a position four nucleotides prior to the loop; in this case, an additional stacking interaction with the protein that further stabilizes the complex is formed. However, the unpaired adenosine is not critical for binding and results in only 1.5- to 5-fold reduction in affinity when absent [20,21]. Removal of the bulged adenosine would eliminate only a single stacking interaction since there are no additional contacts between the protein and the base and would indeed result in a rather minor decrease in affinity. The lack of the unpaired base apparently does not impose significant conformational changes to the stem and still permits the EF loop to bind the lower part of the RNA hairpin, although the interactions are probably somewhat different from those in the wild-type complex. For the MS2 operator bound to the Qβ coat protein, the combined effects of binding a four-nucleotide loop and the requirement to accommodate the −10 adenine in its binding pocket would cause the lower stem to adopt an orientation that is not optimal for interacting with the EF loop.

Interactions with the lower stem are also impaired for the Qβ operator bound to the MS2 coat protein mutant because the lower part of the hairpin, including the bulged adenosine, was disordered and not visible in the three-dimensional structure of the complex [25] (Fig. 3). In this case, the size of the hairpin loop does not seem to play a very important role because the MS2 coat protein can bind a three-nucleotide loop almost as well as a four-nucleotide one given that the distance between the −10 and −4 adenosines is preserved [29]. In the Qβ operator, however, the distance is greater by one nucleotide, which would prevent the bulged adenosine from being accommodated in the MS2 pocket and would not allow favorable stacking interactions with the tyrosine side chain. We believe that this, together with the shorter EF loop found in the MS2 coat protein that cannot interact with the lower part of the stem, explains the observed lack of interactions with the lower part of the Qβ operator.

## RNA recognition mechanisms among *Leviviridae* phages

Including the Qβ structure presented here, the three-dimensional structures of coat protein–operator complexes are now known for four different RNA phages. Despite some profound differences, a number of common themes can also be recognized. One such feature that has been observed in all phage operator structures is that some of the nucleotide bases in the loop stack with bases in the helical stem. In MS2, PRR1 and Qβ, the nucleotide stack further extends to the aromatic side chain of a conserved tyrosine residue, whereas in PP7, a van der Waals interaction with a valine residue takes place. The aromatic stacking is likely important for constraining the loop nucleotides in an appropriate position to bind the protein and is therefore conserved during evolution. Another RNA recognition strategy shared between all phages involves sequence-specific interactions between nucleotide bases and a complementary RNA-binding surface of the protein. In all of the studied phages, binding of an adenine base in the loop into an adenine-recognition pocket in the coat protein is critical for the operator–coat protein interaction, but the importance of other base-specific interactions varies. For the PP7 coat protein, base-specific interactions play a fundamental role in operator recognition and involve four nucleotides in the loop and the bulge, while the sugar-phosphate backbone does not make any contacts with the protein outside of these regions. In MS2, the situation is somewhat similar in that three bases make direct contact with the protein; however, the RNA backbone also makes significant interactions with the protein in the stretch between the bulged adenosine and the loop. In Qβ, the loop adenine is the only nucleotide that makes base-specific contacts with the coat protein while the majority of interactions between the protein and RNA involve the sugar-phosphate backbone. Despite the smaller amount of sequence-specific information, the Qβ coat protein is still able to discriminate its cognate operator, which demonstrates how co-evolution of the protein and RNA can result in a highly specific interaction based on the conformation of the phosphate backbone rather than numerous sequence-specific contacts with bases. The three very different modes of accommodating an unpaired base in PP7, MS2 and Qβ further demonstrate the notable flexibility of protein–RNA interactions in evolutionarily related viruses.

Nevertheless, the overall binding mode of the Qβ coat protein to its operator is clearly similar to those of MS2 and PRR1, which suggests that this particular mechanism is conserved among the conjugative plasmid-dependent *Leviviridae* phages. Outside this group, the PP7 coat protein is the only one that still has some traces of sequence identity with MS2 and Qβ,

but its RNA recognition mechanism is very different. Two other *Leviviridae* phages that are remarkably different from the rest have been identified and sequenced: *Acinetobacter* phage AP205 and *Caulobacter* phage ϕCb5. Their coat protein sequences share no recognizable similarities with those of MS2, Qβ, PP7 or each other. For phage AP205, a putative operator hairpin at the beginning of the replicase gene has been identified, which, unlike other phages, has a bulged uridine located on the 3′ side of the stem [30]. An operator hairpin could not be reliably identified in the genome of phage ϕCb5, raising the question of whether it exists at all [31]. The three-dimensional structure of the ϕCb5 virion revealed strong electron density for RNA bases between the dimers, which indicates a very different RNA packaging and recognition mechanism [32]. Therefore, further studies on protein–RNA interactions of the small RNA phages have the potential to provide even more discoveries about the biology, evolution and structure of these fascinating viruses.

## Materials and Methods

### Preparation of coat protein and RNA

Plasmid p205 encoding the Qβ coat protein with cysteines in the FG loop mutated to glycines was kindly provided by Dr. Indulis Cielēns. Using p205 as a template, we PCR-amplified the coat protein coding sequence with forward primer 5′-CAGGATCCATGGCAAAATTAGAGACTGTTAC-3′ and reverse primer 5′-TATGAAGCTTAATACGCTGGGCGCAGCTGATCAA-3′ to introduce the Asn129Arg amino acid substitution and cloned it into the pET28a expression vector (Novagen). The resulting plasmid was used as a template for site-directed mutagenesis by PCR using primers 5′-CAAGCGGGTGCAGTTCGTGCGCTGGAGAAGCGT-3′ and 5′-ACGCTTCTCCAGCGCACGAACTGCACCCGCTTG-3′ to introduce the additional Pro42Arg mutation. The resulting plasmid was named pET28-Qβ150 and used to produce the assembly-deficient coat protein dimer for crystallization.

For protein production, *Escherichia coli* BL21(DE3) cells containing pET28-Qβ150 were grown in LB medium supplemented with 30 μg/ml kanamycin with aeration at 37 °C. When the $OD_{590}$ of the culture reached 0.5, IPTG was added to a final concentration of 1 mM, and the bacteria were grown for another 4 h and harvested by centrifugation. To purify the protein, we resuspended cells in buffer containing 40 mM Tris–HCl (pH 8.0), 200 mM NaCl, 20 mM $MgSO_4$, 0.1% Triton X-100, 0.1 mg/ml DNase and 1 mg/ml lysozyme and lysed them by three freeze–thaw cycles. The lysate was clarified by centrifugation at 18,500*g*, and the supernatant was loaded on a 1-ml HiTrap SP FF column (GE Healthcare) equilibrated with buffer A [20 mM Tris–HCl (pH 8.0) and 200 mM NaCl]. After extensive washing with buffer A, we eluted bound proteins with a 10-ml gradient of 0–100% buffer B [20 mM Tris–HCl (pH 8.0) and 1 M NaCl] and collected them in 1-ml fractions. Individual fractions containing coat protein were diluted to 5 ml with buffer A and loaded on a Mono S 5/50 GL column (GE Healthcare) equilibrated with buffer A. Bound proteins were eluted with a 15-ml gradient of 0–50% buffer B, corresponding to 200–600 mM NaCl. Fractions containing coat protein and no major contaminants were pooled, concentrated to 500 μl with an Amicon Ultra 10K spin unit (Millipore) and loaded on a Superdex 200 10/300 GL gel-filtration column (GE Healthcare) equilibrated with buffer C [50 mM 4-morpholineethanesulfonic acid (pH 6.0) and 50 mM NaCl]. Fractions containing coat protein were pooled, concentrated and stored at 4 °C until crystallization.

An HPLC-purified RNA oligonucleotide with the sequence 5′-AUGCAUGUCUAAGACAGCAU-3′ corresponding to the wild-type Qβ translation operator was purchased from Metabion AG.

**Table 2.** Crystallographic data collection, scaling and refinement statistics

| | |
|---|---|
| *Data collection and scaling* | |
| Space group | $P6_522$ |
| Cell parameters (Å) | |
| *a* | 75.84 |
| *c* | 303.49 |
| Wavelength (Å) | 1.0000 |
| Resolution (Å) | 38–2.40 |
| Highest-resolution bin (Å) | 2.53–2.40 |
| $R_{merge}$ | 0.09 (0.61) |
| Total number of observations | 73,407 |
| Number of unique reflections | 21,316 |
| $I/\sigma(I)$ | 9.7 (2.2) |
| Completeness (%) | 99.9 (100.0) |
| Multiplicity | 3.4 (3.5) |
| *Refinement* | |
| Number of reflections in work set | 20,180 |
| Number of reflections in test set | 1089 |
| $R_{work}$ | 0.25 |
| $R_{free}$ | 0.29 |
| *B*-factor (Å$^2$) | |
| Protein atoms | 33.1 |
| RNA atoms | 39.8 |
| From Wilson plot | 34.9 |
| Number of atoms | |
| Protein | 1854 |
| RNA | 422 |
| Solvent | 93 |
| rmsd from ideal | |
| Bond lengths (Å) | 0.016 |
| Bond angles (°) | 1.670 |
| Ramachandran plot (%) | |
| Residues in favored regions | 96.2 |
| Residues in allowed regions | 100.0 |

Values in parentheses are given for the highest-resolution shell.

### Crystallization and data collection

The concentration of coat protein was quantified spectrophotometrically assuming that one absorption unit at λ = 280 nm corresponds to a protein concentration of 2.37 mg/ml, as calculated with the ProtParam utility on the ExPASy server [33]. To quantify the RNA, we used data provided by the supplier. The coat protein (10.5 mg/ml in buffer C) and RNA (20 mg/ml in diethylpyrocarbonate-

treated water) were mixed immediately before crystallization at a molar ratio of coat protein dimer to RNA operator of 1:1.2, corresponding to a final concentration of 9.2 mg/ml of protein and 2.48 mg/ml of RNA. The complex was crystallized using the sitting-drop method by mixing 1 μl of the coat protein–RNA complex with 1 μl of reservoir solution [0.1 M sodium acetate (pH 4.5), 0.2 M zinc acetate and 9% polyethylene glycol 3000] and incubating at room temperature (293 K). Hexagonal bipyramid-shaped crystals appeared overnight and grew for a few days, reaching maximum dimensions of 0.15 mm. Prior to data collection, the crystals were cryoprotected by briefly soaking them in a reservoir solution containing 30% ethylene glycol and flash-frozen in liquid nitrogen. Data were collected at MAX-Lab beamline I911-3 (Lund, Sweden). The crystal parameters and data collection statistics are presented in Table 2.

### Structure determination

Diffraction data were indexed using MOSFLM [34] and scaled using Scala [35] from the CCP4 suite [36]. Molecular replacement was performed with MOLREP [37] using the coordinates of a coat protein dimer in the AB conformation from the crystal structure of Qβ bacteriophage (PDB entry 1QBE) as the search model. The solution was further refined using REFMAC [38]. Examination of the resulting electron density map revealed a region of unassigned density below the RNA-binding surface of the coat protein dimer that could be readily interpreted as RNA. To facilitate the modeling of the RNA stem–loop, we performed another round of molecular replacement in MOLREP using the coat protein dimer as the fixed input model and the partial Qβ operator (chain R from PDB entry 1ZSE) as the search model. The resulting model was subjected to several rounds of model building in Coot [39], refinement in REFMAC and validation using the MolProbity server [40]. Refinement and validation statistics are shown in Table 2.

### Analysis and superimposition of atomic coordinates

The interface areas between the protein and RNA were calculated with PISA [41]. The rmsd values were calculated using the program LSQMAN [42]. The Qβ coat protein dimer in the AB conformation from the capsid structure (PDB entry 1QBE) was compared to that in the unassembled state by superimposing the $C^{\alpha}$ atoms of the two dimers with a distance cutoff of 3.5 Å. This resulted in the fit of 238 out of the 239 corresponding atoms, namely, residues 1–73 and 85–132 of chain A and residues 1–55, 61–74 and 85–132 of chain B in the unassembled dimer to the equivalent residues of monomers B and A in the capsid structure. To superimpose the RNAs, we used the backbone atoms C4′, P, C1′, C2′, C3′, O2′, O3′ and O4′ in all cases with residue ranges as indicated in the text. The cognate Qβ and MS2 coat protein–operator complexes were superimposed by explicitly fitting the Qβ coat protein $C^{\alpha}$ atoms of residues 33–37, 46–54, 62–71 and 87–97 to those of residues 30–34, 42–50, 56–65 and 83–93, respectively, in the MS2 coat protein–RNA complex (PDB entry 1ZDI).

### Accession numbers

The atomic coordinates and structure factors of the Qβ coat protein–operator complex have been deposited in the Protein Data Bank with the accession code 4L8H.

## Acknowledgments

We thank Dr. Andris Kazāks for help with protein purification and Ināra Akopjana for excellent technical assistance. We also thank the personnel at MAX-Lab for their help during our stay at the synchrotron. We are also grateful to Prof. Lars Liljas for reading and commenting on the manuscript. Our studies were supported by grant 09.1294 from the Latvian Research Council and grant 2DP/2.1.1.1.0/10/APIA/VIAA/052 from the European Regional Development Fund.

*Received 2 July 2013;*
*Received in revised form 29 August 2013;*
*Accepted 30 August 2013*
Available online 11 September 2013

***Keywords:***
Leviviridae;
allolevivirus;
protein–RNA interaction;
RNA recognition;
translational repression

## References

[1] van Duin J, Tsareva N. Single-stranded RNA phages. In: Calendar R, editor. The Bacteriophages. New York, NY: Oxford University Press; 2006. p. 175–96.
[2] Shiba T, Suzuki Y. Localization of A protein in the RNA-A protein complex of RNA phage MS2. Biochim Biophys Acta 1981;654:249–55.
[3] Bernardi A, Spahr PF. Nucleotide sequence at the binding site for coat protein on RNA of bacteriophage R17. Proc Natl Acad Sci USA 1972;69:3033–7.
[4] Meyer F, Weber H, Weissmann C. Interactions of Qβ replicase with Qβ RNA. J Mol Biol 1981;153:631–60.
[5] Valegård K, Liljas L, Fridborg K, Unge T. The three-dimensional structure of the bacterial virus MS2. Nature 1990;345:36–41.
[6] Gralla J, Steitz JA, Crothers DM. Direct physical evidence for secondary structure in an isolated fragment of R17 bacteriophage mRNA. Nature 1974;248:204–8.
[7] Weber H. The binding site for coat protein on bacteriophage Qβ RNA. Biochim Biophys Acta 1976;418:175–83.
[8] Nathans D, Oeschger MP, Polmar SK, Eggen K. Regulation of protein synthesis directed by coliphage MS2 RNA. I. Phage protein and RNA synthesis in cells infected with suppressible mutants. J Mol Biol 1969;39:279–92.

[9] Beckett D, Wu HN, Uhlenbeck OC. Roles of operator and non-operator RNA sequences in bacteriophage R17 capsid assembly. J Mol Biol 1988;204:939–47.
[10] Carey J, Cameron V, De Haseth PL, Uhlenbeck OC. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. Biochemistry 1983;22:2601–10.
[11] Uhlenbeck OC, Carey J, Romaniuk PJ, Lowary PT, Beckett D. Interaction of R17 coat protein with its RNA binding site for translational repression. J Biomol Struct Dyn 1983;1:539–52.
[12] Romaniuk PJ, Lowary P, Wu HN, Stormo G, Uhlenbeck OC. RNA binding site of R17 coat protein. Biochemistry 1987;26:1563–8.
[13] Valegård K, Murray JB, Stockley PG, Stonehouse NJ, Liljas L. Crystal structure of an RNA bacteriophage coat protein–operator complex. Nature 1994;371:623–6.
[14] Valegård K, Murray JB, Stonehouse NJ, Van den Worm S, Stockley PG, Liljas L. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein–RNA interactions. J Mol Biol 1997;270:724–38.
[15] van den Worm SH, Stonehouse NJ, Valegård K, Murray JB, Walton C, Fridborg K, et al. Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments. Nucleic Acids Res 1998;26:1345–51.
[16] Grahn E, Moss T, Helgstrand C, Fridborg K, Sundaram M, Tars K, et al. Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coat-protein complex. RNA 2001;7:1616–27.
[17] Helgstrand C, Grahn E, Moss T, Stonehouse NJ, Tars K, Stockley PG, et al. Investigating the structural basis of purine specificity in the structures of MS2 coat protein RNA translational operator hairpins. Nucleic Acids Res 2002;30:2678–85.
[18] Persson M, Tars K, Liljas L. PRR1 coat protein binding to its RNA translational operator. Acta Crystallogr Sect D Biol Crystallogr 2013;69:367–72.
[19] Chao JA, Patskovsky Y, Almo SC, Singer RH. Structural basis for the coevolution of a viral RNA–protein complex. Nat Struct Mol Biol 2008;15:103–5.
[20] Witherell GW, Uhlenbeck OC. Specific RNA binding by Qβ coat protein. Biochemistry 1989;28:71–6.
[21] Lim F, Spingola M, Peabody DS. The RNA-binding site of bacteriophage Qβ coat protein. J Biol Chem 1996;271:31839–45.
[22] Golmohammadi R, Fridborg K, Bundule M, Valegård K, Liljas L. The crystal structure of bacteriophage Qβ at 3.5 Å resolution. Structure 1996;4:343–54.
[23] Spingola M, Peabody DS. MS2 coat protein mutants which bind Qβ RNA. Nucleic Acids Res 1997;25:2808–15.
[24] Lim F, Peabody DS. RNA recognition site of PP7 coat protein. Nucleic Acids Res 2002;30:4138–44.
[25] Horn WT, Tars K, Grahn E, Helgstrand C, Baron AJ, Lago H, et al. Structural basis of RNA binding discrimination between bacteriophages Qβ and MS2. Structure 2006;14:487–95.
[26] Schrödinger L. The PyMOL Molecular Graphics System, version 1.5.0.1. at http://www.pymol.org; 2012.
[27] Spingola M, Lim F, Peabody DS. Recognition of diverse RNAs by a single protein structural framework. Arch Biochem Biophys 2002;405:122–9.
[28] Rowsell S, Stonehouse NJ, Convery MA, Adams CJ, Ellington AD, Hirao I, et al. Crystal structures of a series of RNA aptamers complexed to the same protein target. Nat Struct Biol 1998;5:970–5.
[29] Convery MA, Rowsell S, Stonehouse NJ, Ellington AD, Hirao I, Murray JB, et al. Crystal structure of an RNA aptamer–protein complex at 2.8 Å resolution. Nat Struct Biol 1998;5:133–9.
[30] Klovins J, Overbeek GP, Van den Worm SHE, Ackermann H-W, Van Duin J. Nucleotide sequence of a ssRNA phage from *Acinetobacter*: kinship to coliphages. J Gen Virol 2002;83:1523–33.
[31] Kazaks A, Voronkova T, Rumnieks J, Dishlers A, Tars K. Genome structure of caulobacter phage phiCb5. J Virol 2011;85:4628–31.
[32] Plevka P, Kazaks A, Voronkova T, Kotelovica S, Dishlers A, Liljas L, et al. The structure of bacteriophage phiCb5 reveals a role of the RNA genome and metal ions in particle stability and assembly. J Mol Biol 2009;391:635–47.
[33] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The Proteomics Protocols Handbook. Toyowa, NJ: Humana Press; 2005. p. 571–607.
[34] Leslie AGW. Recent changes to the MOSFLM package for processing film and image plate data. Joint CCP4 + ESF-EAMCB Newslett Protein Crystallogr, 26; 1992.
[35] Evans PR. Scala. Joint CCP4 + ESF-EAMCB Newslett Protein Crystallogr, 33; 1997. p. 22–4.
[36] Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. Acta Crystallogr Sect D Biol Crystallogr 1994;50:760–3.
[37] Vagin A, Teplyakov A. MOLREP: an automated program for molecular replacement. J Appl Crystallogr 1997;30:1022–5.
[38] Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr Sect D Biol Crystallogr 1997;53:240–55.
[39] Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr Sect D Biol Crystallogr 2010;66:486–501.
[40] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr 2010;66:12–21.
[41] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–97.
[42] Kleywegt GJ. Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallogr Sect D Biol Crystallogr 1996;52:842–57.

# Paper III

JOURNAL OF VIROLOGY, May 2011, p. 4628–4631 Vol. 85, No. 9
0022-538X/11/$12.00 doi:10.1128/JVI.02256-10


# Genome Structure of Caulobacter Phage phiCb5▽

Andris Kazaks, Tatyana Voronkova, Janis Rumnieks, Andris Dishlers, and Kaspars Tars*

*Biomedical Research and Study Center, Ratsupites 1, LV 1067 Riga, Latvia*

Received 28 October 2010/Accepted 3 February 2011

**The complete genome sequence of caulobacter phage phiCb5 has been determined, and four open reading frames (ORFs) have been identified and characterized. As for related phages, the ORFs code for maturation, coat, replicase, and lysis proteins, but unlike other *Leviviridae* members, the lysis protein gene of phiCb5 entirely overlaps with the replicase in a different reading frame. The lysis protein of phiCb5 is about two times longer than that of the distantly related MS2 phage and presumably contains two transmembrane helices. Analysis of the proposed genome secondary structure revealed a stable 5′ stem-loop, similar to other phages, and a substantially shorter 3′ untranslated region (UTR) structure with only three stem-loops.**

The small RNA phages belonging to the *Leviviridae* family have been extensively used as models to study various problems in molecular biology, including translational control, virus evolution, structure, and assembly. *Leviviridae* coliphages are divided into two genera, *Levivirus* and *Allolevivirus*. The levivirus genome (represented by phage MS2 in Fig. 1A) encodes four proteins: open reading frame 1 (ORF1) encodes a maturation or "A" protein (AP), responsible for the attachment of phages to bacterial F pili; ORF2 encodes the coat protein (CP); ORF3 encodes the replicase subunit (RP); and a fourth open reading frame partially overlaps with ORF2 and ORF3 and encodes a lysis protein (LP). Alloleviviruses (represented by phage Qβ in Fig. 1A) do not have a separate gene for the LP; instead, the AP is responsible for cell lysis. Capsids of alloleviviruses contain about 5% of A1 protein, which is a prolonged read-through variant of CP that has been shown to be necessary for infection (8). The genome organizations of the *Pseudomonas* phage PP7 (12) and broad-host-range phage PRR1 (14) are similar to that of leviviruses, while the *Acinetobacter* AP205 phage (10) has a slightly different genome organization (Fig. 1A).

The RNA phage φCb5, first isolated by Schmidt (16), infects bimorphic *Caulobacter crescentus* bacteria through adsorption to pili specific to swarmer cells (15). Phage φCb5 RNA was isolated and sequenced as previously described (13). The obtained sequence from several overlapping clones covered most of the phage genome, except the 5′ and 3′ ends. To determine the phage genome sequence of the 5′ end, its cDNA was tailed with dATP using terminal transferase, and PCR was carried out using the 5′-GCGCG$(T)_{18}$ primer and a primer complementary to nucleotides (nt) 257 to 275. The PCR products were cloned, and eight clones were sequenced. In all cases, we obtained the same 5′ sequence, with the exception of some shortened variants. To resolve the 3′ end, a poly(A) tail was added to the phage RNA using poly(A) polymerase. The cDNA was synthesized with the 5′-GCGCG$(T)_{18}$ primer, and PCR was carried out using the 5′-GCGCG$(T)_{18}$ primer and a primer complementary to nucleotides 3142 to 3161. The PCR fragment was cloned, and four clones were sequenced. All of the clones displayed the 3′ end of the RP gene followed by 82 additional nucleotides.

The genome of φCb5 is organized in a way similar to that of leviviruses (Fig. 1A). After a short 5′ untranslated region (UTR), ORF1 encodes AP, ORF2 encodes CP, and ORF3 encodes RP. However, the LP gene of φCb5 is placed differently and entirely overlaps with the RP gene in the (+1) reading frame.

The nucleotide sequence of the φCb5 genome and amino acid sequences of the individual proteins have very low homologies with their counterparts in other RNA phages. The only sequence which can be aligned unambiguously is the central part of RP (residues 295 to 537). The coat protein of φCb5 has low sequence similarity to other RNA phages, and none of the residues conserved among PRR1, PP7, and all coliphages are conserved in φCb5. However, reliable alignment based on known three-dimensional structures of coat proteins can be performed (13). In both cases, phylogenetic analysis suggests that φCb5 forms a distant branch among *Leviviridae* and does not belong to either *Levivirus* or *Allolevivirus* (Fig. 1B).

Like in other related phages, the capsid of φCb5 consists of 180 CP monomers. The crystal structure of the φCb5 capsid has been described in detail by Plevka et al. (13).

The CP of φCb5 is unusually short, only 122 amino acid residues, while the CPs of other *Leviviridae* phages have lengths ranging from 127 to 132 residues.

Although the most obvious initiation site for AP is the first AUG codon in the genome that also has a strong preceding Shine-Dalgarno (SD) sequence, mass spectrometry revealed the presence of a protein of a smaller mass than predicted from the sequence (data not shown). To establish the actual translation start site of the AP, the proteins of purified φCb5 virions were separated by SDS-PAGE, and the N-terminal sequence of the 40-kDa band was determined. The sequence was found to be ARIRN, corresponding to a nucleotide sequence 78 nucleotides from the 5′ end of the genome. The sequence is immediately preceded by a UUG codon, which can serve as an initiation codon in bacteria, which is probably the case for the AP of φCb5. However, we cannot exclude the possibility that the upstream AUG codon is in fact used for translational initiation and that proteolytic cleavage occurs later.

* Corresponding author. Mailing address: Biomedical Research and Study Center, Ratsupites 1, LV 1067 Riga, Latvia. Phone: (371) 270-76237. Fax: (371) 674-42407. E-mail: kaspars@biomed.lu.lv.

▽ Published ahead of print on 16 February 2011.

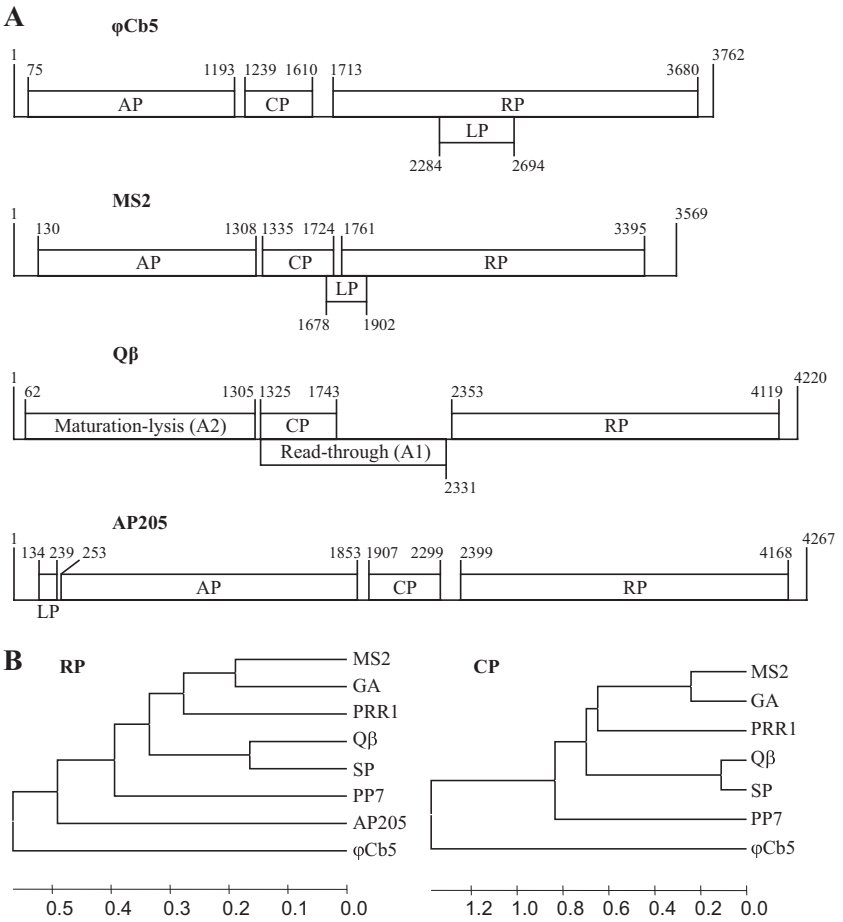


FIG. 1. (A) The genome organization of small RNA phages φCb5, MS2, Qβ, and AP205. Genes are drawn to their approximate scale. (B) Unweighted-pair group method using average linkages (UPGMA) tree of small RNA phages, based on sequence alignment of the RP central part and structure-based alignment of CP. AP205 is not included in the CP alignment-based tree, since its structure is unknown and the CP sequence itself shows no significant similarity to other phages. Trees were constructed by MEGA (17).

In alloleviviruses, lysis is achieved by the maturation protein that blocks MurA, an enzyme in the pathway of murein biosynthesis (4). However, this is not the case for φCb5, since overexpression of the AP gene of φCb5 in *Escherichia coli* did not cause cell lysis (Fig. 2A). Leviviruses have dedicated lysis proteins which form pores in the cellular membrane, leading to activation of autolysins and eventually cell lysis (6). The sequences, lengths, and locations in the genome of the LPs vary among different leviviruses, and their only conserved features seem to be the clustering of positively charged residues near the N terminus and a hydrophobic region near the C terminus, which has been demonstrated to form a transmembrane helix in case of MS2 (6). In MS2, the last 30 residues of LP are necessary and sufficient for cell lysis (3), suggesting that the presence of positively charged residues in the N-terminal region of the protein is not crucial.

In φCb5, no obvious ORF corresponding to the LP of leviviruses or to AP205 could be detected. Analysis of the translated genome sequence using the TMHMM 2.0 server (11) revealed in total three transmembrane helices that entirely overlapped with the RP gene in a different reading frame. The first helix, encoded by nucleotides 2098 to 2226, lacked any suitable upstream initiation codon, and when the respective sequence was cloned into an expression plasmid with an AUG initiation codon, no change in cell growth was observed upon induction (Fig. 2A). The other two transmembrane helices were found in a potential ORF that had a strong SD sequence but an unusual start codon, UUG. However, unusual LP start codons have been reported earlier for several other small RNA phages, like fr (1), PP7 (12), and PRR1 (14). Possibly, non-AUG start codons might help to limit expression of potentially dangerous proteins. The ORF was 135 residues long, considerably exceeding the LP lengths of other small RNA phages (Fig. 2B). Upon ORF cloning and expression, the optical density (OD) decreased (Fig. 2A) in a manner similar to that described for AP205 and PRR1 LPs (10, 14). The cell growth was not affected by a similar expression plasmid with an identical sequence except for a termination codon that was placed after the second codon of the putative LP gene (Fig. 2A). When both potential transmembrane helices were cloned and expressed separately, cell growth was halted, but no decrease in the OD value was observed (Fig. 2A). It should be noted that the above putative LP expression attempts were carried out in *E. coli*; therefore, we cannot exclude the possibility that analyzed ORFs would behave differently in φCb5 host *Caulobacter*.

RP is the most conserved protein among *Leviviridae* phages,

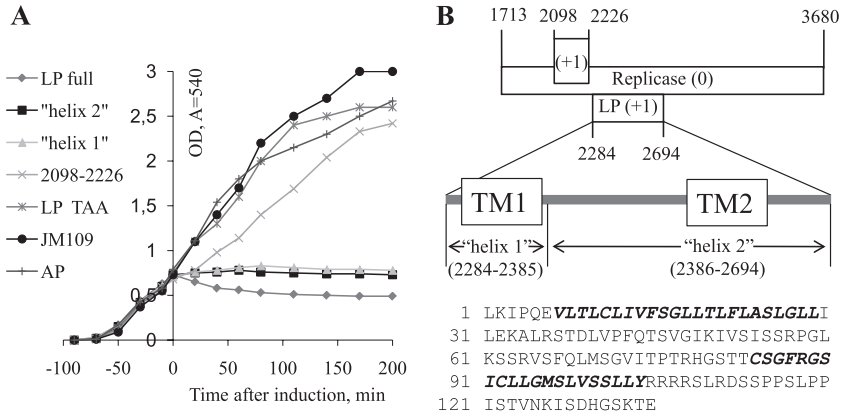


FIG. 2. Properties of LP. (A) *E. coli* cell growth upon expression of the proposed LP, the individual transmembrane helices 1 and 2 of LP, LP with early termination codon (LP TAA), AP, and the translated sequence of nucleotides 2098 to 2226 in the φCb5 genome. All the ORFs were cloned under arabinose-inducible promoter in pBAD plasmid (Invitrogen) and expressed in *E. coli* strain JM109. Growth of untransformed JM109 is shown as well. (B) Position of LP in the RP gene. Positions of predicted transmembrane regions are shown as "TM1" and "TM2" and actual expressed protein sequences as "helix 1" and "helix 2." The sequence of the proposed LP, with transmembrane regions in bold italic, is shown.

with the highest homology around the active site. Although the overall sequence identity is low, the φCb5 RP has the GDD and FRESCG motifs that are totally conserved among all known *Leviviridae* phages. There are several initiation codons in the beginning of the RP ORF. The first AUG also has a significant SD sequence upstream; therefore, we assume that RP is most likely translated from the first AUG codon.

The small RNA phages have a characteristic RNA stem-loop structure, including the RP start codon (Fig. 3A). This stem-loop serves as a binding site for a CP dimer, which represses the translation of the RP gene in late stages of infection. The putative secondary RNA structure of the region surrounding the RP initiation codon is shown in Fig. 3B. There are two more stem-loops close to the three downstream AUG codons that could potentially serve as repression sites (Fig. 3B). However, we failed to demonstrate binding of any of the three corresponding stem-loop RNA oligonucleotides (5′-UCAUCCCCUAGCUUUAUGAGGCUAAGAUGA, 5′-UAUCAGGCAGUUAUGAAAGACUACCUGAUG, 5′-AGGACGUUGAGCGUGACAUGUCACGCCUCCAACUCCU) to CP using a filter binding assay as described for phage R17 (5). As none of the conserved RNA-binding residues in related phages were identified in φCb5 CP (13), the interactions of φCb5 CP with RNA may be very different.

The small RNA phages have a characteristic stable stem-loop structure at their 5′ ends, believed to be necessary for strand separation during replication (2, 18). A similar loop is found near the 5′ end of the φCb5 genome (Fig. 4A).

3′ UTRs of small RNA phages are folded in a separate domain composed of four to nine stem-loops (Fig. 4B). φCb5 appears to have the simplest arrangement known so far, with just three stem-loops, including the RP termination codon containing the R1 loop. In coliphages and AP205, there is a conserved UGCUU sequence 15 to 17 nt from the 3′ end that in the case of phage Qβ has been shown to regulate replication via a long-distance interaction (9). The last stem-loop of φCb5 RNA is somewhat similar to the U1 loops in related phages, and it contains a UGCUG sequence 16 nt from the 3′ end. A

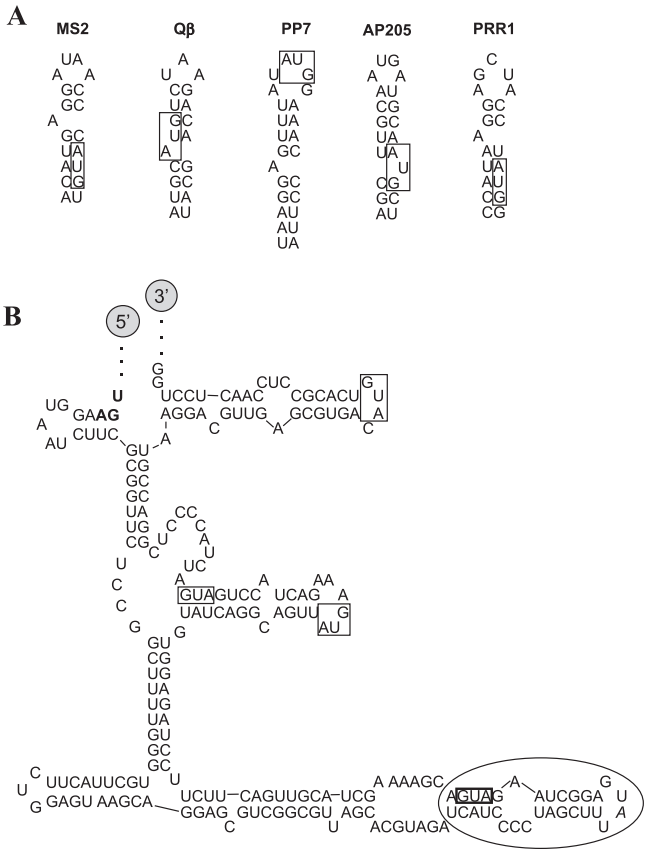


FIG. 3. (A) RP gene operator stem-loops of small RNA phages. RP initiation codons are boxed. Secondary structures were calculated using the RNAfold server (7). (B) Proposed secondary structure of the φCb5 genome region between the termination codon of the CP gene (shown in bold) and +124 in the RP gene. The first possible initiation codon of the RP gene is shown in a box with bold lines. The three downstream AUG codons in the same reading frame are boxed as well. The stem-loop structure containing the first initiation codon is circled.

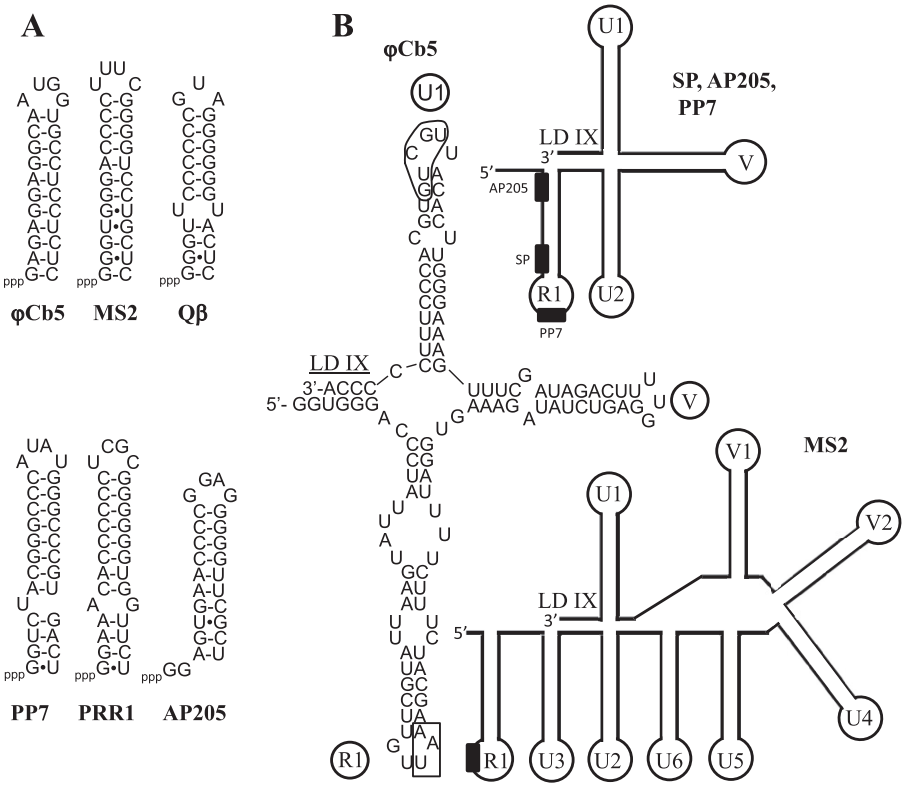


FIG. 4. Structure of terminal UTRs. (A) Hairpin loops near the 5′ ends of genomes of small RNA phages. (B) Structure of 3′ ends of small RNA phages. Sequence is displayed only for φCb5; for other phages, the schematic structures of stem-loops are shown. The RP termination codon in the R1 loop is boxed, and the conserved sequence in the U1 loop is circled. RP termination codons are shown as black boxes. Names of stem-loops and the long-distance interaction region (LD IX) are adopted from reference 10.

sequence complementary to UGCUG is found in two positions in the RP gene, but due to insignificant sequence similarity of φCb5 and Qβ genomes, it cannot be concluded whether a long-distance interaction takes place in φCb5 as well.

**Nucleotide sequence accession number.** The sequence has been deposited in GenBank under accession number HM066936.

This study was supported by ESF grant 1DP/1.1.1.2.0/09/APIA/VIAA/150, ERDF grant 2DP/2.1.1.1.0/10/APIA/VIAA/052, and grant 09.1294 from the Latvian Council of Science.

We thank Pavel Plevka for participation in the RNA-binding measurements. Lars Liljas and Janis Klovins are acknowledged for valuable discussions.

## REFERENCES

1. **Adhin, M. R., A. Avots, V. Berzin, G. P. Overbeek, and J. van Duin.** 1990. Complete nucleotide sequence of the group I RNA bacteriophage fr. Biochim. Biophys. Acta **1050:**104–109.
2. **Beekwilder, M. J., R. Nieuwenhuizen, and J. van Duin.** 1995. Secondary structure model for the last two domains of single-stranded RNA phage Qβ. J. Mol. Biol. **247:**903–917.
3. **Berkhout, B., M. H. de Smit, R. A. Spanjaard, T. Blom, and J. van Duin.** 1985. The amino-terminal half of the MS2-coded lysis protein is dispensable for function: implications for our understanding of coding region overlaps. EMBO J. **4:**3315–3320.
4. **Bernhardt, T. G., I. Wang, D. K. Struck, and R. Young.** 2001. A protein antibiotic in the phage Qβ virion: diversity in lysis targets. Science **292:**2326–2329.
5. **Carey, J., V. Cameron, P. L. de Haseth, and O. C. Uhlenbeck.** 1983. Sequence-specific interaction of R17 coat protein with its RNA binding site. Biochemistry **22:**2601–2610.
6. **Goessens, W. H. F., A. J. M. Driessen, J. Wilschut, and J. van Duin.** 1988. A synthetic peptide corresponding to the C-terminal 25 residues of phage MS2 coded lysis protein dissipates the protonmotive force in Escherichia coli membrane vesicles by generating hydrophilic pores. EMBO J. **7:**867–873.
7. **Gruber, A. R., R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker.** 2008. The Vienna RNA Websuite. Nucleic Acids Res. **36:**W70–W74.
8. **Hofstetter, H., H. Monstein, and C. Weissmann.** 1974. The readthrough protein A1 is essential for the formation of viable Qβ particles. Biochim. Biophys. Acta **374:**238–251.
9. **Klovins, J., and J. van Duin.** 1999. A long-range pseudoknot in Qβ RNA is essential for replication. J. Mol. Biol. **294:**875–884.
10. **Klovins, J., G. P. Overbeek, S. H. van den Worm, H. W. Ackermann, and J. van Duin.** 2002. Nucleotide sequence of a ssRNA phage from Acinetobacter: kinship to coliphages. J. Gen. Virol. **83:**1523–1533.
11. **Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305:**567–580.
12. **Olsthoorn, R. C. L., G. Garde, T. Dayhuff, J. F. Atkins, and J. van Duin.** 1995. Nucleotide sequences of a single-stranded RNA phage from Pseudomonas aeruginosa: kinship to coliphages and conservation of regulatory RNA structures. Virology **206:**611–625.
13. **Plevka, P., et al.** 2009. The structure of bacteriophage phiCb5 reveals a role of the RNA genome and metal ions in particle stability and assembly. J. Mol. Biol. **391:**635–647.
14. **Ruokoranta, T. M., A. M. Grahn, J. J. Ravantti, M. M. Poranen, and D. H. Bamford.** 2006. Complete genome sequence of the broad host range single-stranded RNA phage PRR1 places it in the *Levivirus* genus with characteristics shared with alloleviviruses. J. Virol. **80:**9326–9330.
15. **Schmidt, J. M.** 1966. Observations on the adsorption of Caulobacter bacteriophages containing ribonucleic acid. J. Gen. Microbiol. **45:**347–353.
16. **Schmidt, J. M., and R. Y. Stanier.** 1965. Isolation and characterization of bacteriophages active against stalked bacteria. J. Gen. Microbiol. **39:**95–107.
17. **Tamura, K., J. Dudley, M. Nei, and S. Kumar.** 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24:**1596–1599.
18. **Zamora, H., R. Luce, and C. K. Biebricher.** 1995. Design of artificial short-chained RNA species that are replicated by Qβ replicase. Biochemistry **34:**1261–1266.

# Paper IV

BMC Microbiology

RESEARCH ARTICLE Open Access

# Diversity of pili-specific bacteriophages: genome sequence of IncM plasmid-dependent RNA phage M

Janis Rumnieks* and Kaspars Tars

## Abstract

**Background:** Bacteriophages of the *Leviviridae* family are small RNA viruses with linear, positive-sense, single-stranded RNA genomes that encode only four proteins. All phages of this family require bacterial pili to attach to and infect cells. *Leviviridae* phages utilizing F-pili for this purpose have been extensively studied. RNA phages specific for conjugative plasmid-encoded pili other than that of plasmid F have been isolated, but are much less understood and their relation to the F-pili-specific phages in many cases is not known.

**Results:** Phage M has the smallest known *Leviviridae* genome to date and has the typical genome organization with maturation, coat and replicase genes in the 5′ to 3′ direction. The lysis gene is located in a different position than in other known *Leviviridae* phages and completely overlaps with the replicase gene in a different reading frame. It encodes a 37 residue long polypeptide that contains a transmembrane helix like the other known lysis proteins of leviviruses. Sequence identities of M proteins to those of other phages do not exceed 25% for maturation protein, 51% for coat protein and 41% for replicase. Similarities in protein sequences and RNA secondary structures at the 3′ untranslated region place phage M together with phages specific for IncP, IncC and IncH, but not IncF plasmid-encoded pili. Phylogenetic analysis using the complete genome sequences and replicase proteins suggests that phage M represents a lineage that branched off early in the course of RNA phage specialization on different conjugative plasmids.

**Conclusions:** The genome sequence of phage M shows that it is clearly related to other conjugative pili-specific leviviruses but has an atypical location of the lysis gene. It provides a better view on the remarkable diversification of the plasmid-specific RNA phages.

**Keywords:** *Leviviridae*, RNA phage, Pili-specific phage, IncM, Conjugative plasmid, Lysis

## Background

Bacteriophages of the *Leviviridae* family are small viruses that infect several genera of Gram-negative bacteria. They have linear, positive-sense, single-stranded RNA genomes about 3500 – 4200 nucleotides in length that encode only four proteins. All *Leviviridae* phages have three genes in common – maturation, coat and replicase [1]. The replicase cistron encodes the catalytic subunit of the RNA-dependent RNA polymerase complex, which is assembled together with several bacterial proteins [2,3] and replicates phage RNA. The coat protein forms dimers, 90 of which assemble in a *T*=3 icosahedral capsid about 27 nm in diameter and encapsidate the genome [4]. A single copy of the maturation protein binds to phage RNA [5] and gets incorporated into capsids along with it. It is required for infectivity of the virions – the maturation protein binds to bacterial pili, then leaves the capsid and enters the cell as an RNA-protein complex [6].

Many of the *Leviviridae* phages are divided in two genera – leviviruses and allolevivirus. The major distinction of allolevivirus is the presence of a minor coat protein A1 in their capsid which is produced by ribosomal read-through of a leaky termination codon of the coat gene [7]. The other difference is that the maturation protein of allolevivirus also triggers cell lysis [8,9], whereas leviviruses encode a dedicated small lysis polypeptide for this purpose [10-12].

* Correspondence: j.rumnieks@biomed.lu.lv
Biomedical Research and Study Centre, Ratsupites 1, Riga LV-1067, Latvia

The ssRNA phages that infect *Escherichia coli* cells by adsorbing to F plasmid-coded pili were the first isolates of the *Leviviridae* family [13,14], and to date these "male-specific" phages, with type species MS2 and Qβ, have been the most intensively studied and best characterized of this family. However, the F plasmid is just one of the many conjugative plasmids that are present in nature. These plasmids are often highly divergent from F and are most often grouped according to their mutual compatibility. In *Enterobacteriaceae*, the conjugative plasmids form more than 20 different incompatibility (Inc) groups which are denoted by capital Latin letters [15]. All these plasmids encode conjugative pili, but the pilin subunits often share no similarity.

Several ssRNA phages specific for conjugative pili other than that of plasmid F have been discovered. Phage PRR1 [16] which adsorbs specifically to IncP plasmid-encoded pili was the first such example, and later other phages specific for Inc group C [17], D [18], H [19,20], I [21], M [22] and T [23] plasmids followed. Phages PRR1, C-1 (IncC-specific) and Hgal1 (IncH-specific) have been sequenced [24,25] and phage PRR1 capsids have also been crystallized [26], but no research has been done on the other plasmid-specific phages since their isolation.

The IncM plasmid-specific RNA phage M [22] was isolated from sewage in Pretoria, South Africa in the beginning of the 1980s. IncM plasmids have a broad host range, code for rigid pili and transfer efficiently only when bacteria are growing on solid media [27]. Likewise, the phage is able to propagate in different strains of *Escherichia*, *Salmonella*, *Klebsiella*, *Proteus* and *Serratia*, provided they contain an IncM plasmid. To obtain more insight in plasmid-specific RNA phages, we determined the genome sequence of phage M and present here its analysis and comparison to the genomes of other RNA phages of the *Leviviridae* family.

## Results and discussion

### Overall structure of the genome

The genome of phage M is 3405 nucleotides long and follows the canonical *Leviviridae* genome organization with maturation, coat and replicase cistrons following each other in the 5′-3′ direction (Figure 1). An unusual feature of the genome is that the lysis gene appears to be located in a different position than in other leviviruses, as discussed below. It is also the smallest known *Leviviridae* genome to date, about 60 nucleotides shorter than that of the group II F-specific phage GA [28]. The protein coding regions of phage M are of similar length to those of phage GA, with maturation and coat genes being a bit longer and replicase somewhat shorter; the greatest savings in M's genome come from terminal untranslated regions (UTRs), the 5′ UTR being about 45 nucleotides and the 3′ UTR about 20 nucleotides shorter.

### Identification of the lysis gene

All members of the levivirus genus encode a short polypeptide that mediates cell lysis. Amino acid sequences of lysis proteins show great variation and their only unifying feature is the existence of a hydrophobic transmembrane helix within the protein [29]. Lysis proteins have been shown to accumulate in the bacterial membrane where they presumably form pores that lead to cell lysis [30]. In all of the known *Enterobacteria*-infecting leviviruses, the lysis gene overlaps with coat and replicase genes in a different reading frame and is translationally coupled with the coat gene [1]. However, in the genome of phage M, no candidate ORFs at this location could be identified: in the +2 frame relative to the coat gene there are no termination codons until the start of replicase and in the +1 frame only a 17 amino acid long ORF that would encode a non-hydrophobic peptide is found.

Up to now, there have been two reported cases in the *Leviviridae* family where the lysis gene in is in a different location: *Acinetobacter* phage AP205 has a short lysis gene preceding the maturation gene [31], while *Caulobacter* phage φCb5 codes for a longer, two-helix protein that completely overlaps with the replicase gene [32]. To test the possibility that phage M also has a non-canonical localization of the lysis gene, we utilized the fact that the pJET1.2 plasmid, where the cDNA copies of the genome were cloned for sequencing, contains a T7 promoter that can be used to transcribe the insert. Several clones with inserts in the correct orientation with respect to the T7

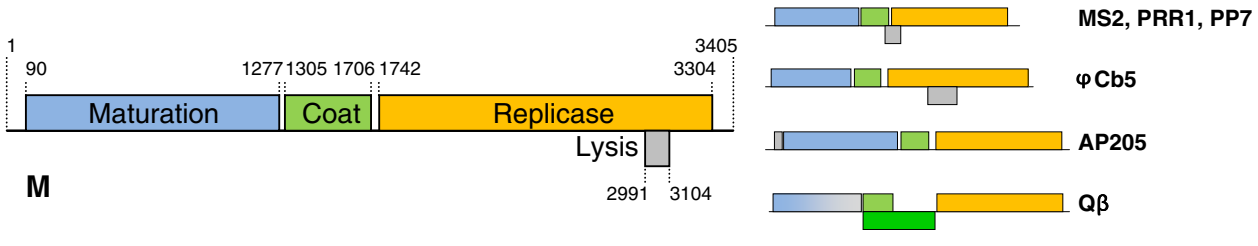


**Figure 1 Genome organization of phage M.** Start and end positions of phage genes are indicated. For comparison, the other known genome organizations of *Leviviridae* phages are represented on the right with genes color-coded as in the M genome. In phage Qβ, protein A1 (bright green) is an extended read-through variant of the coat protein and the lysis function is performed by the maturation protein.

promoter were selected and transformed to a T7 polymerase-producing *E.coli* strain. When the expression of T7 polymerase was induced, a clone containing an approximately 1000 nucleotide long fragment spanning nucleotides 2098-3129 of the phage genome resulted in a clear cell lysis. Examination of this sequence located a likely candidate for the lysis gene between nucleotides 2991-3104 (Figure 2A). This was based on several criteria: (1) it was the only ORF in the fragment with a significant length (37 amino acids; the shortest known *Leviviridae* lysis protein is that of phage AP205 with 34 amino acids); (2) according to the TMHMM server [33], the ORF-encoded protein was predicted to contain a transmembrane helix with over 95% probability; (3) although the ORF had an unusual initiation codon UUG, there was a rather strong Shine-Dalgarno (SD) sequence GAGG nine nucleotides upstream; (4) RNA secondary structure prediction using the RNAfold server [34] revealed that the initiation codon of the ORF is located on top of an AU-rich stem-loop that would presumably have sufficiently low thermodynamic stability to promote the initiation of translation [35] (Figure 2B). To verify the lytic function of the gene, the ORF together with the original SD sequence and UUG initiation codon was cloned in an inducible protein expression vector. Induction resulted in almost complete cell lysis some 45 minutes after (Figure 2C), thus demonstrating that the approximately 150 nucleotide long stretch is sufficient to encode a functional lysis protein. The above-mentioned evidence therefore lets us suggest with some confidence that this is the actual lysis gene of phage M.

### Protein similarities to other phages

The maturation proteins are very variable in *Leviviridae* phages, which is unsurprising given the vast diversity of pili they have evolved to bind. The maturation protein of phage M is most similar to those of the other plasmid-specific RNA phages, but the sequence identity is only 24.5% to phage PRR1, around 22% to C-1, Hgal1, GA and MS2 and drops to 17% when compared to allolevi-viruses SP and Qβ. The coat proteins are more conserved and here M groups clearly with phages PRR1, C-1 and Hgal1 with amino acid identities of 48-51%. The identity with F-specific phages is significantly lower and ranges from 27.1% for group II levivirus KU1 to 19% for group IV allolevivirus NL95. Notably, M coat protein shares 24.6% amino acids with that of *Pseudomonas* phage PP7, which is the only plasmid-independent phage for which the sequences could be reasonably aligned. For replicase, the trend is similar as for the maturation protein: the replicase of phage M most resembles that of PRR1 with 41% amino acid identity, followed by other plasmid-dependent phages C-1, Hgal1, MS2 and GA (33-37% identity) and alloleviviruses (27-29% identity). Again, M replicase turns out to be more closely related to that of phage PP7 (25.5% identity) than to the other plasmid-independent phages AP205 and $\varphi$Cb5 (17.7 % identity).

### Conserved RNA secondary structures

With the growing number of *Leviviridae* genomes that have been sequenced it has become clear that besides encoding proteins, the secondary and tertiary structure of the RNA itself is also very important. The complex structure of RNA provides binding sites for phage proteins [36-38], regulates their translation [1] and promotes genome packaging in capsids [39]. In many cases where nucleotide stretches from different phage genomes show no sequence similarity, the secondary

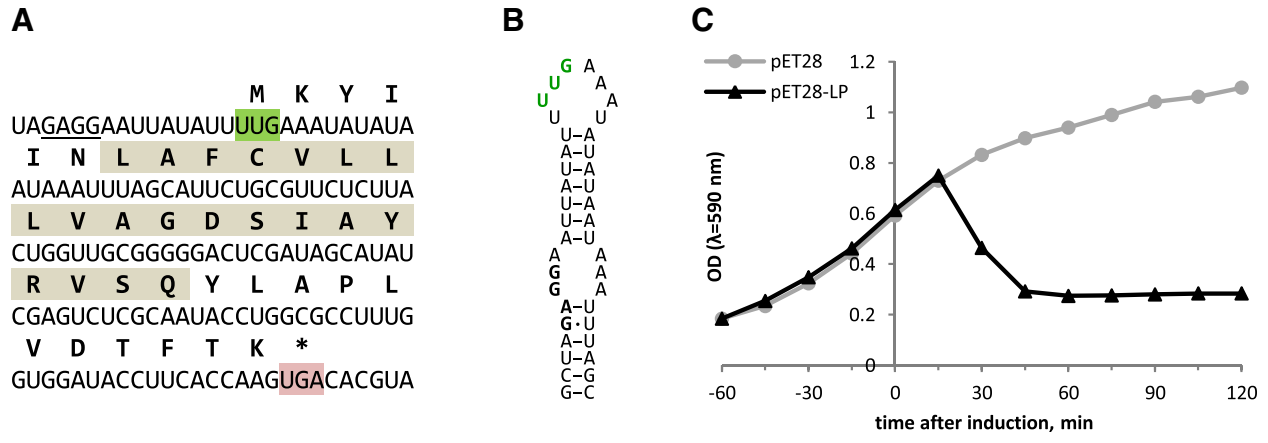


**Figure 2 Lysis protein of phage M.** (**A**) The lysis gene. The Shine-Dalgarno sequence is underlined and initiation and termination codons are indicated by green and pink shading, respectively. The translated amino acid sequence is given above the RNA sequence and the putative transmembrane helix is shaded gray. (**B**) An RNA hairpin around the initiation codon of the lysis gene. The initiation codon and the Shine-Dalgarno sequence are indicated. (**C**) Verification of the lysis gene. Growth of *E.coli* cells harboring either empty vector (pET28) or a plasmid with the cloned lysis gene (pET28-LP) before and after the induction of protein synthesis is shown.

structures they fold into are nevertheless well preserved. One such example lies at the very 5′ end of all of the sequenced ssRNA phage genomes, where there is a stable GC-rich hairpin that has been suggested to play an important role in phage RNA replication [40]. Phage M is no exception (Figure 3A). Another important RNA structure lies around the initiation codon of replicase. This approximately 20-nucleotide-long stretch folds into a hairpin structure that specifically binds the phage coat protein. This interaction acts as a translational operator to repress synthesis of replicase when enough coat protein accumulates [37] and has been suggested to play also a role in initiating specific encapsidation of the genomic RNA [41]. When the operator hairpin of phage M is compared to those of other ssRNA phages, it is evident that it groups with the conjugative pili-dependent phages PRR1, C-1, Hgal1 and MS2 (Figure 3B). An adenine residue in the loop four nucleotides upstream of the replicase initiation codon and an unpaired purine residue in the stem which are critical for RNA-protein binding in phages MS2 [42], GA [43] and PRR1 [44] are preserved also in phage M, therefore the mechanism of interaction is probably similar.

It is also interesting to take a look at the 3′ untranslated region of the phage genome. The configurations of 3′ UTRs vary between different phages, but nevertheless some similarities exist. In all known *Leviviridae* phages a long-distance interaction designated ld IX bridges the very 3′ terminus with a complementary nucleotide stretch upstream, forming the 3′ terminal domain [45]. The domain usually consists of at least three hairpins, denoted U1, U2 and V. In phage M, the 100-nucleotide-long 3′ UTR is made up from four hairpins U4, U3, U2 and U1 (Figure 3C). In all ssRNA phages the 3′-terminal helix U1 has a remarkably conserved nucleotide sequence in the loop: UGCUU in phages as diverse as MS2, SP and AP205, UGCUG in φCb5 and CGCUC in PP7. In the case of Qβ, this loop forms a long-distance pseudoknot with a complementary sequence approximately 1200 nucleotides upstream that is essential for phage replication [47]. In phage M, the sequence of the U1 loop is AUUGCUAUG. It has not been experimentally verified that phages other than Qβ have the pseudoknot, but in M genome a sequence AGCAA is found in the replicase gene some 1215 nucleotides upstream that could potentially basepair with UUGCU in the loop. The other notable feature of the 3′ domains, although less pronounced, is hairpin V (designated V2 in some phages) which in phages MS2, Qβ, SP and AP205 contains a large, adenine-rich loop. There is some evidence that in MS2 this might be one of the sites where the maturation protein binds to the RNA [36]. In phage φCb5, however, the candidate hairpin V lacks analogous features and in phages PRR1, C-1 and Hgal1 it does not seem to exist at all; instead, there is a stretch of unpaired nucleotides (UAUAAACA in PRR1, UAUA in Hgal1 and

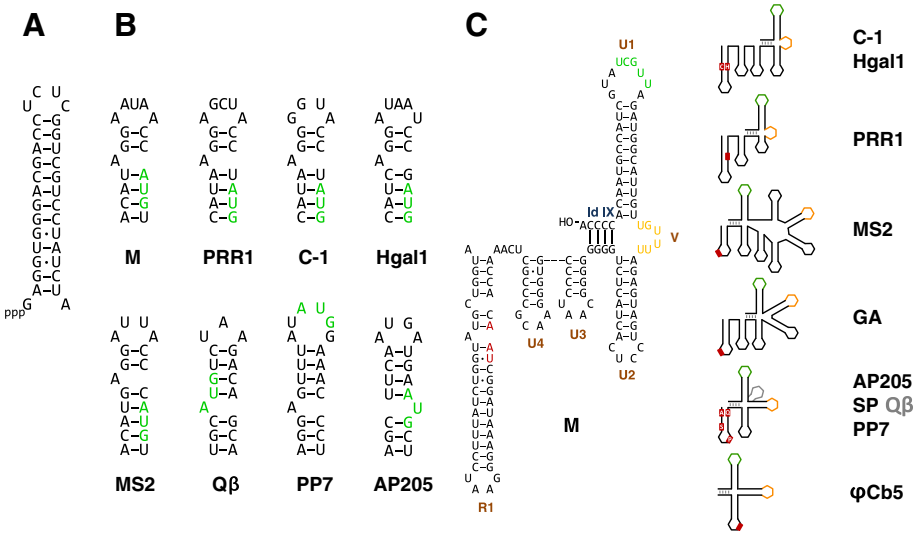


**Figure 3 RNA secondary structures in M genome.** (**A**) A stable hairpin at the very 5′ end of the genome important for phage RNA replication. (**B**) The operator hairpin around the initiation codon of replicase. The analogous hairpins from other *Leviviridae* phages are shown for comparison. Start codons of the replicase gene are colored green. (**C**) Structure of the 3′ untranslated region. The termination codon of replicase is colored dark red, the unpaired stretch corresponding to loop V or V2 in other phages in orange and the conserved nucleotide sequence in the loop of hairpin U1 that potentially forms a long-distance pseudoknot in green. On the right, schematic representations of 3′ UTRs from other phages based either on published data [31,32,45,46] or RNA secondary structure predictions are given for comparison. The 3′ UTR of phage Qβ is closely similar to that of phage SP except for a short extra helix which is depicted in gray. The locations of replicase gene termination codons are represented as red boxes. RNA secondary structures were predicted by the RNAfold server [34].

UUAAU in C-1) that connects hairpins U2 and U1 and might serve the same function as hairpin V in other phages. In phage M the situation is similar, but the loop sequence is UUUUGU and contains no adenine residues. When the overall structures of 3′ UTRs from different phages are compared (Figure 3C, right), it is evident that in the distantly related phages φCb5, AP205, PP7 and SP the 3′ domain is remarkably simple with just three hairpins, while it is considerably expanded in the plasmid-specific leviviruses, culminating in seven hairpins in phage MS2. In this respect, phages M, C-1, Hgal1 and PRR1 form their own group where the 3′ UTR adopts a characteristic fold of only two hairpins between the ld IX, a stretch of unpaired nucleotides instead of hairpin V and one or two hairpins between the terminal replicase hairpin R1 and ld IX.

### Evolutionary considerations

In many aspects, phage M is a typical representative of the *Leviviridae* family that is clearly related to other conjugative pili-dependent RNA phages. The feature that makes it unique though is the unusual location of its lysis gene. Although there are precedents of this in the distantly related phages AP205 and φCb5, it is a bit surprising to find such phenomenon also within a group of otherwise rather closely related phages. Apparently, it is relatively easy for a short ORF encoding a transmembrane helix that causes cell lysis to appear by random mutations, as several phages have arrived at the same mechanism independently. It would also suggest that the location of the lysis gene at this position is probably limited to the IncM plasmid-specific leviviruses or even to a smaller subgroup of these phages. Since M is the only IncM plasmid-specific RNA phage that has been isolated, it is not possible to address this question presently.

The high mutation rates and resulting sequence variability in RNA viruses makes reconstruction of their evolutionary history not a trivial task. Based on similarities between maturation and replicase proteins, phage M seems more related to phage PRR1, while coat protein sequences and structures of the 3′ UTRs suggest that it might be closer to phages C-1 and Hgal1. To further address this question we conducted a phylogenetic analysis of 15 representative *Leviviridae* phages using both the complete genome sequences and also the replicase protein sequences since the RNA-dependent RNA polymerases are the most conserved proteins of all positive-sense RNA viruses [48]. Both trees (Figure 4) confirm that phage M is more closely related to the IncC, IncH and IncP than to the IncF plasmid-dependent phages but they show differences in the clustering of the non-F plasmid specific phages. Although phylogenetic analysis of the coat proteins (not shown) gives the same (M(C-1 (Hgal1,PRR1))) clustering as the replicase, low bootstrap values for the IncC, IncH and IncP branches indicate that confidence in that particular branching order is not high and suggest that phages C-1, Hgal1 and PRR1 have radially diverged from a similar ancestral sequence. In both trees phage M represents a lineage that branched off early in the course of specialization on different plasmids after the separation of the IncF lineage had occurred but before the diversification on IncC, IncH and IncP plasmids took place. Both trees also support the idea that the allolevivirus lineage separated from the leviviruses before the specialization on different conjugative pili had occurred and that these phages arrived at the ability to bind to F-pili via an independent evolutionary path.

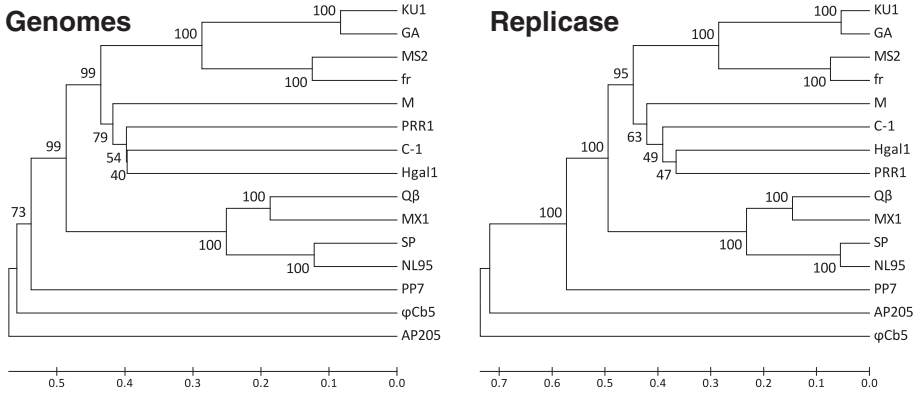


**Figure 4 Phylogeny of RNA phages.** The phylogenetic analysis was based on the complete genomic RNA sequences (left) and amino acid sequences of the replicase (right) which is the most conserved of all ssRNA phage proteins. Trees were constructed by unweighted pair group method with arithmetic mean (UPGMA) and tested using the bootstrap method with 500 replicates. The bootstrap values are expressed as percentages next to the nodes. RNA and protein sequences were aligned using MUSCLE [49] and the phylogenetic trees were constructed in program MEGA5 [50].

Although all *Leviviridae* phages use pili for attachment, there is a marked difference between the types of pili they utilize. The type IV pili used by phages AP205, φCb5 and PP7 are produced via a genome-encoded type II secretion pathway [51], whereas the plasmid-borne conjugative pili that the other phages utilize belong to a type IV secretion system [52]. Both systems share some functional similarities, like a retractable pilus and a membrane pore, but are thought to have evolved independently [53]. Therefore a jump from one to the other type of pili had to occur at some point in the *Leviviridae* history. Our phylogenetic analysis suggests that the ancestral phage infected cells via type IV pili, like phages AP205, φCb5 and PP7 are doing today and a PP7-like virus then might have evolved the ability to bind to some kind of conjugative pili and still sustain infectivity. Consequently, all of the specialized plasmid-dependent RNA phages we know today would be descendants of this ancestral virus.

## Conclusions

We have determined and characterized the genome sequence of IncM plasmid-dependent phage M and shown that it resembles the plasmid-specific leviviruses in many ways but has an atypical location of the lysis gene. It is a valuable addition to the growing number of sequenced *Leviviridae* genomes and provides a better view on the diversity and evolution within this phage family.

## Methods

### Phage propagation and purification

Bacteriophage M and its host *E.coli* J53(RIP69) were obtained from Félix d'Hérelle Reference Center for bacterial viruses, Laval University, Quebec, Canada (catalog numbers HER218 and HER1218, respectively). J53 (RIP69) cells were grown in LB medium containing 6 μg/ml tetracycline overnight at 37 °C without agitation. To propagate the phage, 0.5 ml of the host cell suspension and 10 μl of phage lysate (approximately $10^{10}$ pfu/ml) were spotted on 1.5% LB agar plates, overlaid with 15-20 ml of molten 0.7% LB agar cooled to 45 °C, mixed by swirling and incubated overnight at 30 °C. The next morning, top agar layers from several plates were scraped off, transferred to centrifuge tubes and centrifuged for 20 minutes at 18500 g. Supernatant was collected and phage particles were precipitated by addition of sodium chloride and PEG 6000 to concentrations of 0.5M and 10%, respectively, and incubation for 30 minutes at 4 °C. After centrifugation for 10 minutes at 18500 g, the supernatant was discarded and the pellet was resuspended in a small volume of distilled water. The phage preparation was then layered on top of a preformed five-step cesium chloride gradient (equal volumes of CsCl solutions in 20 mM Tris-HCl pH 7.5 with densities of 1.7, 1.6, 1.5, 1.4 and 1.3 g/ml) and centrifuged for 17 hours in a SW 40Ti rotor at 24000 rpm. 0.5 ml fractions were collected from the top of the gradient and the peak fractions containing phage were pooled and dialyzed against one liter of 20 mM Tris-HCl pH 7.5 overnight at 4 °C. The preparation was concentrated to 500 μl using Amicon Ultra 10K MW cutoff spin unit (Millipore) and used for RNA extraction.

### Isolation of genomic RNA and sequencing

500 μl of purified phage preparation was mixed with 500 μl of phenol and SDS was added to a final concentration of 0.5%. The mixture was vigorously vortexed for 60 s and centrifuged at 12000 g for 3 minutes. The aqueous phase was extracted two more times with a 1:1 phenol/chloroform mixture and once with chloroform. The RNA in the final aqueous phase was precipitated with ethanol, centrifuged and the pellet redissolved in a small volume of DEPC-treated water.

4 μg of the purified RNA was reverse-transcribed with RevertAid Premium reverse transcriptase (Fermentas) using primer 5′-GCAAATTCTGTTTTATCAGACNNNNNN-3′. Reaction products were purified using GeneJet PCR purification kit (Fermentas) and eluted in 20 μl of water. The 3′ termini of the purified first strand cDNAs were dATP-tailed using terminal deoxynucleotidyl transferase (Fermentas). The reaction products were again purified using the PCR purification kit and used as a template for second-strand PCR with primers 5′-GCAAATTCTGTTTTATCAGAC-3′ and 5′-GCGCG$(T)_{18}$-3′ and Pfu DNA polymerase (Fermentas). Reaction products were separated in a 1% agarose gel and a slice corresponding to 1000 – 3000 base pair DNA fragments was cut out. The fragments were extracted using GeneJet gel extraction kit (Fermentas) and ligated in pJET1.2/blunt vector (Fermentas).

Insert-containing clones were sequenced on an ABI Prism 3100 Genetic Analyzer using BigDye Terminator v3.1 kit (Applied Biosystems). Based on the obtained sequence data, additional reverse transcription-PCRs were performed using specific primers to fill gaps and increase coverage. Since the initial cloning procedure already involved 3′-tailing of cDNAs, it was possible to determine the 5′ end of the genome from these clones. To determine the sequence of the 3′ end, phage RNA was tailed with *E.coli* Poly(A) polymerase (Ambion), followed by reverse transcription with primer 5′-GCGCG$(T)_{18}$-3′ and PCR using primers 5′-GCGCG$(T)_{18}$-3′ and 5′-CTGGCGCCTTTGGTGGATAC-3′ corresponding to nucleotides 3072-3091 of the phage genome. Genome assembly and ORF prediction was done with the program ContigExpress from the VectorNTI Suite (Invitrogen).

The genome sequence was deposited in GenBank with accession code JX625144.

## Cloning and expression of the lysis gene

The putative lysis gene was PCR-amplified from a suitable cDNA clone using primers 5′-ATATT-CTAGACGAAGGAACAACCATTGCCG-3′ and 5′-TATGAAGCTTACTTGGTGAAGGTATCCACC-3′, the fragment was digested with XbaI and HindIII and ligated into XbaI-HindIII-digested pET28a vector (Novagen), yielding plasmid pET28-LP. To test for the lytic function of the protein, pET28-LP-containing *E.coli* BL21 AI cells (Invitrogen) were grown in LB medium supplemented with 30 μg/ml kanamycin and protein production was induced by adding arabinose to a final concentration of 0.2% and IPTG to a final concentration of 1 mM.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JR propagated and purified the phage, sequenced the genome, cloned the lysis gene, analyzed the genome and wrote the paper. KT supervised the work, analyzed the genome sequence and wrote the paper. Both authors read and approved the final manuscript.

### Acknowledgements

This work was supported by grant 09.1294 from the Latvian Council of Science and grant 2DP/2.1.1.1.0/10/APIA/VIAA/052 from the European Regional development fund (ERDF). The publishing costs were covered by ERDF grant 2DP/2.1.1.2.0/10/APIA/VIAA/004.

Received: 13 September 2012 Accepted: 20 November 2012
Published: 24 November 2012

### References

1. Van Duin J, Tsareva N: **Single-stranded RNA phages**. In *The Bacteriophages*. 2nd edition. Edited by Calendar RL.: Oxford University Press; 2006:175–196.
2. Blumenthal T, Landers TA, Weber K: **Bacteriophage Qβ replicase contains the protein biosynthesis elongation factors EF Tu and EF Ts.** *Proc Natl Acad Sci USA* 1972, **69**:1313–1317.
3. Wahba AJ, Miller MJ, Niveleau A, Landers TA, Carmichael GG, Weber K, Hawley DA, Slobin LI: **Subunit I of Qβ replicase and 30 S ribosomal protein S1 of *Escherichia coli* Evidence for the identity of the two proteins.** *J Biol Chem* 1974, **249**:3314–3316.
4. Valegård K, Liljas L, Fridborg K, Unge T: **The three-dimensional structure of the bacterial virus MS2.** *Nature* 1990, **345**:36–41.
5. Kozak M, Nathans D: **Fate of maturation protein during infection by coliphage MS2.** *Nat New Biol* 1971, **234**:209–211.
6. Shiba T, Miyake T: **New type of infectious complex of *E.coli* RNA phage.** *Nature* 1975, **254**:157–158.
7. Weiner AM, Weber K: **Natural read-through at the UGA termination signal of Qβ coat protein cistron.** *Nat New Biol* 1971, **234**:206–209.
8. Winter RB, Gold L: **Overproduction of bacteriophage Qβ maturation (A2) protein leads to cell lysis.** *Cell* 1983, **33**:877–885.
9. Karnik S, Billeter M: **The lysis function of RNA bacteriophage Qβ is mediated by the maturation (A2) protein.** *EMBO J* 1983, **2**:1521–1526.
10. Model P, Webster RE, Zinder ND: **Characterization of Op3, a lysis-defective mutant of bacteriophage f2.** *Cell* 1979, **18**:235–246.
11. Atkins JF, Steitz JA, Anderson CW, Model P: **Binding of mammalian ribosomes to MS2 phage RNA reveals an overlapping gene encoding a lysis function.** *Cell* 1979, **18**:247–256.
12. Beremand MN, Blumenthal T: **Overlapping genes in RNA phage: a new protein implicated in lysis.** *Cell* 1979, **18**:257–266.
13. Loeb T, Zinder ND: **A bacteriophage containing RNA.** *Proc Natl Acad Sci USA* 1961, **47**:282–289.
14. Davis JE, Strauss JH, Sinsheimer RL: **Bacteriophage MS2: another RNA phage.** *Science* 1961, **134**:1427.
15. Taylor DE, Gibreel A, Lawley TD, Tracz DM: **Antibiotic resistance plasmids**. In *Plasmid biology*. Edited by Funnell BE, Philips GJ. Washington, D.C: ASM Press; 2004:473–491.
16. Olsen RH, Thomas DD: **Characteristics and purification of PRR1, an RNA phage specific for the broad host range *Pseudomonas* R1822 drug resistance plasmid.** *J Virol* 1973, **12**:1560–1567.
17. Sirgel FA, Coetzee JN, Hedges RW, Lecatsas G: **Phage C-1: an IncC group; plasmid-specific phage.** *J Gen Microbiol* 1981, **122**:155–160.
18. Coetzee JN, Bradley DE, Lecatsas G, du Toit L, Hedges RW: **Bacteriophage D: an IncD group plasmid-specific phage.** *J Gen Microbiol* 1985, **131**:3375–3383.
19. Coetzee JN, Bradley DE, Fleming J, du Toit L, Hughes VM, Hedges RW: **Phage pilHα: a phage which adsorbs to IncHI and IncHII plasmid-coded pili.** *J Gen Microbiol* 1985, **131**:1115–1121.
20. Nuttall D, Maker D, Colleran E: **A method for the direct isolation of IncH plasmid-dependent bacteriophages.** *Lett Appl Microbiol* 1987, **5**:37–40.
21. Coetzee JN, Bradley DE, Hedges RW: **Phages Iα and I2-2: IncI plasmid-dependent bacteriophages.** *J Gen Microbiol* 1982, **128**:2797–2804.
22. Coetzee JN, Bradley DE, Hedges RW, Fleming J, Lecatsas G: **Bacteriophage M: an incompatibility group M plasmid-specific phage.** *J Gen Microbiol* 1983, **129**:2271–2276.
23. Bradley DE, Coetzee JN, Bothma T, Hedges RW: **Phage t: a group T plasmid-dependent bacteriophage.** *J Gen Microbiol* 1981, **126**:397–403.
24. Ruokoranta TM, Grahn AM, Ravantti JJ, Poranen MM, Bamford DH: **Complete genome sequence of the broad host range single-stranded RNA phage PRR1 places it in the Levivirus genus with characteristics shared with Allolevivirus.** *J Virol* 2006, **80**:9326–9330.
25. Kannoly S, Shao Y, Wang IN: **Rethinking the evolution of single-stranded RNA (ssRNA) bacteriophages based on genomic sequences and characterizations of two R-plasmid-dependent ssRNA phages, C-1 and Hgal1.** *J Bacteriol* 2012, **194**:5073–5079.
26. Persson M, Tars K, Liljas L: **The capsid of the small RNA phage PRR1 is stabilized by metal ions.** *J Mol Biol* 2008, **383**:914–922.
27. Bradley DE, Taylor DE, Cohen DR: **Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12.** *J Bacteriol* 1980, **143**:1466–1470.
28. Inokuchi Y, Takahashi R, Hirose T, Inayama S, Jacobson AB, Hirashima A: **The complete nucleotide sequence of the group II RNA coliphage GA.** *J Biochem (Tokyo)* 1986, **4**:1169–1980.
29. Young R: **Bacteriophage lysis: mechanism and regulation.** *Microbiol Rev* 1992, **56**:430–481.
30. Goessens WH, Driessen AJ, Wilschut J, van Duin J: **A synthetic peptide corresponding to the C-terminal 25 residues of phage MS2 coded lysis protein dissipates the protonmotive force in *Escherichia coli* membrane vesicles by generating hydrophilic pores.** *EMBO J* 1988, **7**:867–873.
31. Klovins J, Overbeek GP, van den Worm SH, Ackermann HW, van Duin J: **Nucleotide sequence of a ssRNA phage from *Acinetobacter*: kinship to coliphages.** *J Gen Virol* 2002, **83**:1523–1533.
32. Kazaks A, Voronkova T, Rumnieks J, Dishlers A, Tars K: **Genome structure of *Caulobacter* phage phiCb5.** *J Virol* 2011, **85**:4628–4631.
33. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567–580.
34. Hofacker IL: **Vienna RNA secondary structure server.** *Nucl Acids Res* 2003, **31**:3429–3431.
35. de Smit MH, van Duin J: **Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis.** *Proc Natl Acad Sci USA* 1990, **87**:7668–7672.
36. Shiba T, Suzuki Y: **Localization of A protein in the RNA-A protein complex of RNA phage MS2.** *Biochim Biophys Acta* 1981, **654**:249–255.
37. Bernardi A, Spahr PF: **Nucleotide sequence at the binding site for coat protein on RNA of bacteriophage R17.** *Proc Natl Acad Sci USA* 1972, **69**:3033–3037.
38. Meyer F, Weber H, Weissmann C: **Interactions of Qβ replicase with Qβ RNA.** *J Mol Biol* 1981, **153**:631–660.
39. Basnak G, Morton VL, Rolfsson O, Stonehouse NJ, Ashcroft AE, Stockley PG: **Viral genomic single-stranded RNA directs the pathway toward a T=3 capsid.** *J Mol Biol* 2010, **395**:924–936.

40. Beekwilder J, Nieuwenhuizen R, Poot R, van Duin J: **Secondary structure model for the first three domains of Qβ RNA. Control of A-protein synthesis.** *J Mol Biol* 1996, **256**:8–19.
41. Beckett D, Wu HN, Uhlenbeck OC: **Roles of operator and nonoperator RNA sequences in bacteriophage R17 capsid assembly.** *J Mol Biol* 1988, **204**:939–947.
42. Carey J, Lowary P, Uhlenbeck OC: **Interaction of R17 coat protein with synthetic variants of its ribonucleic acid binding site.** *Biochemistry* 1983, **22**:4723–4730.
43. Gott JM, Wilhelm LJ, Uhlenbeck OC: **RNA binding properties of the coat protein from bacteriophage GA.** *Nucl Acids Res.* 1991, **19**:6499–6503.
44. Persson M, Tars K, Liljas L: **PRR1 coat protein binding to its RNA translational operator.** *Acta Crystallogr D Biol Crystallogr*, . in press.
45. Beekwilder MJ, Nieuwenhuizen R, van Duin J: **Secondary structure model for the last two domains of single-stranded RNA phage Qβ.** *J Mol Biol* 1995, **247**:903–917.
46. Olsthoorn RC, Garde G, Dayhuff T, Atkins JF, Van Duin J: **Nucleotide sequence of a single-stranded RNA phage from *Pseudomonas aeruginosa*: kinship to coliphages and conservation of regulatory RNA structures.** *Virology* 1995, **206**:611–625.
47. Klovins J, van Duin J: **A long-range pseudoknot in Qβ RNA is essential for replication.** *J Mol Biol* 1999, **294**:875–884.
48. Koonin EV, Dolja VV: **Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences.** *Crit Rev Biochem Mol Biol* 1993, **28**:375–430.
49. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucl Acids Res* 2004, **32**:1792–1797.
50. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731–2739.
51. Peabody CR, Chung YJ, Yen MR, Vidal-Ingigliardi D, Pugsley AP, Saier MH Jr: **Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella.** *Microbiology* 2003, **149**:3051–3072.
52. Lawley TD, Klimke WA, Gubbins MJ, Frost LS: **F factor conjugation is a true type IV secretion system.** *FEMS Microbiol Lett* 2003, **224**:1–15.
53. Hazes B, Frost L: **Towards a systems biology approach to study type II/IV secretion systems.** *Biochim Biophys Acta* 2008, **1778**:1839–1850.

doi:10.1186/1471-2180-12-277
**Cite this article as:** Rumnieks and Tars: **Diversity of pili-specific bacteriophages: genome sequence of IncM plasmid-dependent RNA phage M.** *BMC Microbiology* 2012 **12**:277.