Cloning and Sequencing Spiroplasma DNA: Gene Structure and Expression

J. M. Bové, J. Renaudin, M. C. Pascarel-Devilder, C. Saillard and C. Chevalier

ABSTRACT. Spiroplasma virus SpV4 is an isometric virus with single-stranded circular DNA. Virus multiplication results in cell lysis. The viral DNA, with only 4.4 kb, is one of the smallest viral genomes known and a good candidate for total base sequencing. We have cloned the replicative form (RF) in *E. coli* strain HB 101 and shown it to remain highly infectious after cloning with all viral genes present. We have established the complete base sequence of the viral DNA. In addition to the gene for the capsid protein (65,000 daltons) we have identified 8 other open reading frames. We have also identified the promoter sequences for RNA polymerase binding and the sequences of the ribosome binding sites. These sequences are remarkably similar to those known in bacteria and represent further evidence for the bacterial origin of mollicutes. This is the first time that the complete base sequence of mollicute genes has been established.

In bacteria and higher organisms, UGG is the only codon for tryptophan, but in *Mycoplasma capricolum* UGA is used in addition to UGG as a tryptophan codon (12). In bacteria and higher organisms UGA is a protein termination codon. The SpV4 capsid protein contains 9 UGA codons and 1 UGG codon and was not expressed at full length in *E. coli* since protein synthesis stopped at the first UGA codon encountered. In contrast, the gene for spiralin, the major membrane protein of *S. citri*, is expressed in the *E. coli* transformant because it contains no tryptophan, and no UGA codon. Our sequencing work shows that in *Spiroplasma*, as in *Mycoplasma*, UGA is a tryptophan codon, not a termination codon.

Index words. Spiroplasma virus, Transfection, Spiroplasma gene structure, Spiralin.

Recently we have shown that, in contrast to *Escherichia coli*, *Spiroplasma* species are insensitive to rifampicin (1). In addition, we have found that the spiroplasma RNA polymerase is made of three major polypeptides, with a subunit structure of the type $\beta\beta'\alpha_2$, i.e. similar to the structure of RNA polymerases of Eubacteria, but different from those of Archaebacteria. For this reason, it might be expected that the promoter sequences recognized by RNA polymerase in spiroplasmas, as well as other remarkable sequences, are closely related to those in Eubacteria. We have developed two systems to verify these hypotheses.

The first system concerns *Spiroplasma melliferum* virus SpV4 (6, 7). The isometric capsid of the virus is made of one major protein of 65,000 daltons (Da) and contains a circular single-stranded DNA of 4.4 kb. This small genome would be expected to contain a limited number of genes. The similar coliphages G4 and ϕ x 174 genomes code for 11 proteins. Hence, sequencing the SpV4 genome would provide information on the structure and expression of several genes.

The second system represents the gene for spiralin (30,000 Da), the major membrane protein of *Spiroplasma citri*. We have recently cloned this gene and shown it to be expressed in the relevant *E. coli* transformant (4). The spiralin gene is contained in a genomic DNA fragment which has been reduced by subcloning to 4.5 kb.

Both the cloned DNA fragment of *S. citri*, and the SpV4 genome contain well-defined genes, namely the gene for spiralin, the genes for virus replication, and the gene for the 65,000 Da capsid protein. These DNAs thus provide suitable models to investigate the structure and expression of spiroplasma genes by nucleic acid sequencing and other techniques.

Based on the complete sequence of the SpV4 genome, the data presented in this report show a remarkable similarity between bacterial and spiroplasmal gene structures and regulatory sequences. Furthermore, evidence is given that in spiroplasmas as

in *Mycoplasma capricolum* (12), UGA codes for tryptophan, explaining the fact that the gene for the tryptophan-free spiralin protein is expressed in *E. coli*, but not the UGA-containing gene for the capsid protein of SpV4.

RESULTS AND DISCUSSION

Cloning SpV4 Replicative Form. We have first cloned the double-stranded replicative form (RF) in *E. coli* (5). The SpV4 RF DNA was purified from infected *Spiroplasma melliferum* cells by alkaline lysis. The RF DNA was further purified by a preparative agarose gel electrophoresis. A partial restriction map has been established. There is only one Cla 1 restriction site.

The SpV4 RF was cloned in *E. coli* using plasmid pBR 328 as a vector. The plasmid confers resistance to chloramphenicol, ampicillin and tetracycline. It contains a unique Cla 1 restriction site in the promoter of the gene responsible for tetracycline resistance. The SpV4 RF was linearized by cleavage with Cla 1 and it was ligated with the Cla 1-cleaved, dephosphorylated vector. The ligation products were used to transform HB101 *E. coli* competent cells.

Since the integration of DNA at the Cla 1 site inactivates the tetracycline gene, transformants that expressed resistance to chloramphenicol were further screened for their sensitivity to tetracycline. Among the recombinant clones, those containing full-size RF DNA as determined by gel electrophoresis, were further selected by *in situ* hybridization using a SpV4-specific probe, followed by analysis of plasmid content.

Analysis of a selected recombinant plasmid showed that upon cleavage with Cla 1 two fragments were obtained: one of 4.4 kbp corresponding to the linear RF, the other of 4.9 kbp corresponding to the vector. As expected, only the 4.4 kbp fragment hybridized with the SpV4 specific probe.

The inserted RF DNA was recovered by preparative agarose gel electrophoresis of the Cla 1 digested recombinant plasmid. The cloned RF was made circular by ligation and the ligation products were used to transfect *Spiroplasma melliferum* (strain G1) cells. Transfection was performed by the PEG method (10) as previously described for SpV4 virion DNA (7). The results showed that the cloned RF is as infectious as the purified virion DNA (5). Transfection efficiency is close to 10^3 PFU/μg of DNA. This result clearly demonstrates that full-size SpV4 RF was cloned and that its biological activity was not affected by cloning in *E. coli*.

Sequencing the cloned SpV4 Replicative Form. The cloned SpV4 DNA was fully sequenced by the dideoxy chain termination technique (9) using the M13 cloning and sequencing system (3). The SpV4 DNA was subcloned in coliphage M13 mp8 as follows. The recombinant plasmid, containing full-size SpV4 DNA was first sonicated. The DNA fragments were treated with the "Klenow" fragment of DNA polymerase I to restore blunt ends. They were ligated with the Sma I-cleaved, dephosphorylated M13 mp8 vector.

TG1 *E. coli* cells were made competent by the Hanahan technique (2) and were transformed by the ligation mixture. Recombinant phages (producing colourless plaques) containing SpV4 DNA were selected by *in situ* hybridization with a SpV4-specific probe. Single-stranded DNA of each recombinant phage was used as a template in the sequencing reaction using the M13 universal primer and S^{35} d ATP as the labeled nucleotide. Over 90% of the SpV4 genome was sequenced on both strands.

Organization of SpV4 genome. The genome of SpV4 is made of 4,421 nucleotides with a base composition of 34.0% A, 11.8% C, 20.2% G and 33.9% T. Only when TGA is taken as a non-termination codon is it possible to identify coding open reading frames (ORF). Nine such ORFs have been located on the genome. They all have a Shine Dalgarno ribosome-binding se-

quence upstream to the initiation codon. Eight ORFs begin with ATG, one with GTG. The termination codons are TAA seven times, TAG once. The ORFs are distributed within all three reading frames and there are three regions where ORFs overlap.

Three putative transcription promoter sequences have been found at position 544, 1292 and 3854 (position 1 corresponds to nucleotide C of the unique ClaI restriction sequence 5′ ATCGAT 3′). These sequences are as expected from the eubacterial structure of spiroplasma RNA polymerase (1).

At position 3930, an inverted repeat sequence is followed by a stretch of 7 T. Its RNA transcript is able to form a hairpin and probably represents a transcription terminator of the type that is independent on protein factor rho. A second inverted repeat sequence with 5 G-C base pairs in a row was found at position 530. This sequence could easily lead to the formation of a stable hairpin structure in the single-stranded circular virion DNA. Such a sequence is also characteristic of the single-stranded DNA coliphages and probably represents the initiation site for complementary DNA strand synthesis.

The sequence at position 530, in addition to leading to a hairpin structure on the single-stranded virion DNA, is also part of a transcription promoter on the double-stranded RF DNA.

Gene expression in spiroplasmas. As to gene expression in spiroplasmas, the ORF of the 65,000 d capsid protein contains 554 codons, is initiated by ATG and terminated by TAA. Codon usage analysis shows that A and T terminated codons are by far the preferred ones in agreement with the low G + C content of the genome. The capsid protein ORF contains 9 TGA codons but also one TGG triplet, both coding very probably for tryptophan. The 9 TGA codons are almost uniformly distributed throughout the ORF. Since TGA is a nonsense codon in *E. coli*, this probably explains why the capsid protein could not be expressed in *E. coli* regardless of the cloning strategy.

However, we have previously reported that spiralin was fully expressed in *E. coli* (4). Spiralin does not contain any tryptophan residues (11). The spiralin gene has now been fully sequenced and was found free of TGA and TGG codons, and thus can be fully expressed in *E. coli*.

CONCLUSION

SpV4 is the first mollicute virus whose genome has been totally sequenced. The capsid protein gene is the first full-length gene for which the base sequence has been established; the gene contains 9 TGA triplets.

Except for the finding that TGA is probably not a termination codon, but codes for tryptophan, our sequencing data show that, in spiroplasmas, DNA replication signals, transcription promoters and terminators as well as ribosome binding sites are closely similar to those of eubacteria and *E. coli* phages. These results fully support the bacterial origin of spiroplasmas (8).

LITERATURE CITED

1. Gadeau, A. P., C. Mouchès, and J. M. Bové
1986. Probable insensitivity of mollicutes to rifampicin and characterization of spiroplasmal DNA dependant RNA polymerase. J. Bacteriol. 166: 824-838.
2. Hanahan, D.
1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166: 557-580.
3. Messing, J., and J. Vieira
1982. A new pair of M113 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19: 269-276.

4. Mouchès. C., T. Candresse, G. Barroso, C. Saillard, H. Wroblewski, and J. M. Bové
1985. Gene for spiralin, the major membrane protein of the helical mollicute *Spiroplasma citri:* cloning and expression in *Escherichia coli*. J. Bacteriol. 164: 1094-1099.
5. Pascarel-Devilder, M. C., J. Renaudin, and J. M. Bové
1986. The spiroplasma Virus 4 replicative form cloned in *Escherichia coli* transfects spiroplasmas. Virology 151: 390-393.
6. Renaudin, J., M. C. Pascarel, M. Garnier, P. Carle-Junca, and J. M. Bové
1984. SpV4, a new spiroplasma virus with a circular, single-stranded DNA. Ann. Virol. 135E: 343-361.
7. Renaudin, J., M. C. Pascarel, M. Garnier, P. Carle, and J. M. Bové
1984. Characterization of spiroplasma virus 4 (SV4). Isr. J. Med. Sci. 20: 797-799.
8. Rogers, M. J., J. Simmons, R. T. Walker, W. G. Weisburg, C. R. Woese, R. S. Tanner, I. M. Robinson, D. A. Stahl, G. Olsen, R. H. Leach, and J. Maniloff
1985. Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data. Proc. Natl. Acad. Sci. USA 82: 1160-1164.
9. Sanger, F., S. Nicklen, and A. R. Coulson
1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA 74: 5463-5467.
10. Sladek, T. L. and J. Maniloff
1983. Polyethylene glycol-dependent transformation of *Acholeplasma laidlawii* with mycoplasma virus L2 DNA. J. Bacteriol. 155: 734-741.
11. Wroblewski, H., D. Robic, D. Thomas, and A. Blanchard
1984. Comparison of the aminoacid compositions and antigenic properties of spiralins purified from the plasma membranes of different spiroplasmas. Ann. Microbiol. (Inst. Pasteur) 135A: 73-82.
12. Yamao. F., A. Muto, Y. Kawauchi, M. Iwani, S. Iwagani, Y. Azumi, and S. Osawa
1985. UGA is read as tryptophan in *Mycoplasma capricolum*. Proc. Natl. Acad. Sci. USA 82: 2306-2309.