Nucleotide Sequence of the Cell Wall Proteinase Gene of
_Streptococcus cremoris Wg2_

JAN KOK, KEES J. LEENHOUTS, ALFRED J. HAANDRIKMAN, AAT M. LEDEBOER, and GERARD VENEMA

Institute of Genetics, University of Groningen, Kerklaan 30, 9751 NN Haren, and Unilever Research Laboratory, Olivier van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands

Received 26 June 1987/Accepted 22 October 1987

A 6.5-kilobase HindIII fragment that specifies the proteolytic activity of _Streptococcus cremoris Wg2_ was sequenced entirely. The nucleotide sequence revealed two open reading frames (ORFs), a small ORF1 with 295 codons and a large ORF2 containing 1,772 codons. For both ORFs, there was no stop codon on the HindIII fragment. A partially overlapping PstI fragment was used to locate the translation stop of the large ORF2. The entire ORF2 contained 1,902 coding triplets, followed by an apparently rho-independent terminator sequence. The inferred amino acid sequence would result in a protein of 200 kilodaltons. Both ORFs have their putative transcription and translation signals in a 345-base-pair Cial fragment. ORF2 is preceded by a promoter region containing a 15-base-pair complementary direct repeat. Both the truncated 33- and the 200-kilodalton proteins have a signal peptide-like N-terminal amino acid sequence. The protein specified by ORF2 contained regions of extensive homology with serine proteases of the subtilisin family. Specifically, amino acid sequences involved in the formation of the active site (viz., Asp-32, His-64, and Ser-221 of the subtilisins) are well conserved in the _S. cremoris Wg2_ proteinase. The homologous sequences are separated by nonhomologous regions which contain several inserts, most notably a sequence of approximately 200 amino acids between the His and Ser residues of the active site.

Because of their importance in the development of flavor and texture in a wide variety of fermented foods throughout the world, the proteolytic enzymes of lactic acid bacteria have been the subject of extensive research during the last decade. These studies have revealed the existence of an astonishingly complex system of proteinases and peptidases. Until now, attempts to unravel the complexity of the proteolytic systems have concentrated on the localization and biochemical characterization of these activities. Intracellular proteinases, as well as several different proteolytic activities associated with the cell wall, have been reported for _Streptococcus cremoris_ and _S. lactis_ (for reviews, see references 16 and 35). Exterkate (6) distinguished three proteolytic activities, on the basis of pH and temperature optima, which were present in different combinations in the cell walls of different _S. cremoris_ strains. In _S. lactis_, multiple proteinases have been demonstrated by a zymogram staining technique on lysozyme-treated cells (3). In an attempt to probe proteolytic activities to separate enzymes, Hugenholtz et al. (12) have used antibodies against the purified proteolytic systems. Characteristic combinations of protein peaks in crossed immunoelectrophoresis experiments were the basis of a new classification of the proteolytic systems of _S. cremoris_ strains. The overall impression of the biochemical data available is that the cell wall bound proteinases of lactic acid streptococci are very large enzymes (with molecular weights of 130,000 or more) which require Ca$^{2+}$ ions for stabilization in an active configuration. Inhibition studies show that they are serine proteinases (8, 9, 13).

In a previous paper, we reported on the cloning and expression of the genetic information of the proteolytic activity of _S. cremoris Wg2_ (14). A 6.5-kilobase HindIII fragment from the proteinase plasmid pWV05 of this strain complemented the proteinase deficiency in _S. lactis_ (Prt$^-$). It specified two proteins, A and B, of the proteolytic system of _S. cremoris Wg2_ in _S. lactis_ as well as in _B. subtilis_. Here we report on the nucleotide sequence of the gene that specifies the cell wall-bound proteinase of _S. cremoris Wg2_ and on some characteristics of this sequence, and of the enzyme as deduced from the predicted amino acid sequence.

MATERIALS AND METHODS

Bacterial strains, plasmids, and media. Growth and maintenance of bacteria and selective conditions have been previously described (14). Plasmid pGKV500 (14) and its derivatives were constructed and maintained in _B. subtilis_ PSL1 (21). _Escherichia coli_JM101 (44) was used as the host for M13 and its derivatives.

Molecular cloning techniques. Plasmid DNA was isolated as described previously (14). Restriction nuclease enzymes, T4 DNA ligase, and the Klenow fragment of _E. coli_ DNA polymerase I were purchased from Boehringer Mannheim Biochemicals and used as recommended by the manufacturer. General procedures for cloning and DNA manipulations were essentially as described by Maniatis et al. (18). Competent cells of _E. coli_ were transformed as described by Mandel and Higa (17). Protoplasts of _B. subtilis_ were transformed as described by Chang and Cohen (2).

DNA sequence analysis. Subfragments of the 6.5-kb HindIII fragment of pGKV500 were cloned in both orientations in phage M13 by using M13 mp10, mp11, mp18, and mp19 (44). The deoxyribonucleotide sequencing method of Sanger et al. (25) was used with buffer gradient gels and [α-35S]dATP (1). Synthetic 17-mer primers were prepared on a model 380A DNA synthesizer (Applied Biosystems, Foster City, Calif.) and purified on 20% polyacrylamide gels. To confirm the nucleotide sequence around the restriction enzyme sites used for cloning in M13, a sequence reaction was performed on pGKV500, which was made single stranded in the region of interest by cutting with an appropriate restric-
tion enzyme and subsequent treatment with *E. coli* exonuclease III (Bethesda Research Laboratories, Inc., Gaithersburg, Md.) as advised by the manufacturer. Exonuclease III-treated DNA (1.5 to 2 μg) was used in a standard sequencing reaction. Nucleotide sequences were stored, matched, and processed by using the computer programs of Staden (29–31).

RESULTS

M13 cloning and DNA sequencing. Figure 1 shows a 6.5-kb *HindIII* fragment of the proteinase-specifying plasmid pWV05 of *S. cremoris* Wg2 and an overlapping *BamHI* fragment of 7.5 kb. The *HindIII* fragment, introduced in *S. lactis* (Pr*) on pGKV500, complemented the proteinase deficiency of this strain. The *BamHI* fragment was cloned in pACYC184, resulting in pGD4 (14). All of the subfragments of the 6.5-kb *HindIII* fragment shown in Fig. 1 were cloned in both orientations in phage M13 mp10 and mp11 (44). Initially, the 345-base-pair (bp) *ClaI* fragment was found in a single mp10 clone, but it was lost upon subculturing. To determine the DNA sequence of the fragments, the two M13 clones of each fragment were sequenced in a cascade sequencing strategy with synthetic primers. After part of the sequence of a fragment had been determined by the dideoxynucleotide method (25), two primers were synthetically prepared. One primer was used to extend the nucleotide sequence, while a reversed primer was used to confirm the sequencing data by sequencing the opposite strand. To confirm the nucleotide sequence around the restriction enzyme sites used for subcloning in M13, pGKV500 was digested with an appropriate restriction enzyme and treated with exonuclease III to produce single-stranded DNA in the region of interest (26). This DNA was the template in a dideoxynucleotide sequencing reaction using one of the synthetic primers near the site to be sequenced. The exonuclease III strategy was also used to determine the nucleotide sequence of the 345-bp *ClaI* fragment. In this way, the nucleotide sequence of both strands of the entire *HindIII* fragment was obtained.

Codon preference analysis. Codon preference analysis (31) of the DNA sequence revealed two high-probability reading frames, one on each strand, orientated in opposite directions (data not shown). Both open reading frames (ORFs), one containing 295 codons (ORF1) and one with 1,772 codons (ORF2), had their endpoints outside the *HindIII* fragment. To extend the sequence of the large ORF2, a partially overlapping 3.5-kb *PstI* fragment, isolated from pGD4 (Fig. 1), was cloned in M13 mp18 and mp19. With synthetic primers, both strands of the left-hand part of this fragment were sequenced. A detailed restriction enzyme map deduced from this sequence and the position of the two ORFs are presented in Fig. 2. The *HindIII* fragment originally cloned in pGKV500 is shown shaded in gray. A third *HindIII* site, located 16 bp downstream of the second one, was not present in pGKV500. The first stop codon after ORF2 was found 380 bp downstream of the third *HindIII* site.

Nucleotide sequence. The nucleotide sequence of the *HindIII* fragment, extended with part of the DNA sequence of the *PstI* fragment, is presented in Fig. 3. In Fig. 4, the 345-bp *ClaI* fragment containing the putative promoter region of both ORF1 and ORF2 is shown in more detail. ORF2 starts with an ATG start codon at position 1,206, and the first stop codon (TAG) is located at position 6,912, giving it a total length of 5,706 bp or 1,902 coding triplets. It has the potential to synthesise a protein of 200 kilodaltons (kDa).

Upstream of the ATG start codon, around nucleotide 1,196, a Shine-Dalgarlo sequence (GGAGG) similar to those reported for *B. subtilis* (11) is present, having a window of 10 bases and a free energy of binding of −14.4 kcal/mol (36). Although a second in-frame ATG codon is present at position 1,176, it is unlikely to be the start codon because it is not preceded by a reasonable ribosome-binding site. Starting at position 1,186, there are several potential promoter regions (20, 24). A continuous sequence of TTGAATTTGTTC contains two putative −35 sequences. With a spacing of 16 and 15 bases, respectively, the two −35 sequences are followed by two overlapping consensus −10 regions (TATAATAAT, starting at position 1,106). The region from position 1,106 to 1,141 contains several other partially overlapping Prinbo boxlike sequences. Upstream of the −35 region, there is an AT-rich region (86% AT over the first 50 bases), with several alternating stretches of A's and T's which resemble the signals known to enhance transcription in *B. subtilis* (4). Actually, the whole region between the two *ClaI* sites, 345 bp in length, is AT rich (73% A's and T's). In the promoter region (from position 1,084 to 1,145), two long direct repeats of 15 and 13 bases are present. The promoter region further contains a long complementary inverted repeat starting at position 1,104. A hypothetical stem-loop structure with a calculated free energy of −10.2 kcal/mol (36)
is depicted in Fig. 5A. The Pri-bow box-rich region is completely buried in the proposed hairpin structure, thereby leaving the two −35 regions without their respective −10 regions.

The Clal site at position 884 is located in the ATG start codon of ORF1 (Fig. 3). Nine bases upstream of this start codon, the sequence GAGGAGA constitutes a possible ribosome-binding site (11, 34). It is less clear-cut, however, to assign a promoter region upstream of this ribosome-binding site. There are several candidate −35 sequences, but only two of them have −10 regions which conform reasonably well to the consensus −10 sequence for E. coli and B. subtilis (20, 24). These are indicated by the leftward-directed arrows (at positions 1,155 and 1,132 and at positions 959 and 936) in Fig. 4. Promoter region 1,155/1,132 overlaps with the putative promoter for ORF2, and its −10 region is occupied in the stem of the proposed hairpin structure in this region (Fig. 5A).

The nucleotide sequence indicates that the codon usage in S. cremoris is quite different from that in E. coli. S. cremoris resembles B. subtilis in that it tends to distribute the codons for its amino acids more evenly (23).

**Terminator structure downstream of ORF2.** In the nucleotide sequence approximately 6,000 bp from the start of ORF2, a region of dyad symmetry is present between nucleotides 7,045 and 7,080, 130 nucleotides downstream of the TAG stop codon. It has all of the features of a rho-independent terminator sequence (24) consisting of two complementary inverted repeats which can form a stem of 15 bp (with seven G-C pairs and two mismatches). The hairpin structure is followed by a run of several T’s and has a ΔG of −24.6 kcal/mol (36; Fig. 5B).

**Putative signal peptides.** The protein specified by ORF2 starts with a sequence of amino acids which closely resembles a typical signal peptide (39). Four positively charged amino acids are followed by a run of hydrophobic residues (Fig. 3). By the rules of von Heijne (40) for processing probability, a putative signal sequence cleavage site is situated between Ala-33 and Ala-34 in the canonical Ala-X-Ala-Ala sequence. Cleavage at this site would result in a signal peptide of 33 amino acids, which is in the size range reported for signal peptides of other Gram-positive exoproteins. The 33-kDa protein coded for by ORF1 also contains a putative signal sequence structure with 32 amino acids.

**Homology comparison.** Of the proteins present in the National Biomedical Research Foundation protein data bank in October 1986, four showed homology with the S. cremoris Wg2 ORF2 protein. All four were bacterial serine proteases of the subtilisin family, and the overlaps are shown in Fig. 6. These subtilisins, produced by bacilli only, can be divided into two groups on the basis of structural and functional comparisons, including amino acid composition and sequence analysis, enzymatic activities, and immunological properties (22). Subtilisin Carlsberg and BPN’ exemplify the two groups. At the amino acid sequence level, these two enzymes are approximately 70% homologous (22, 27). Obviously, ORF2 specifies a proteinase of the subtilisin type. One region of the S. cremoris proteinase, extending over 34 amino acids (amino acids 599 to 632), showed 50 to 56% homology with a region in the different subtilisins containing the reactive Ser-221. The corresponding serine in the S. cremoris proteinase, Ser-620, is contained in a stretch of seven amino acids with complete homology. A second region of homology, with 38 to 46% matches over a stretch of 117 amino acids, is found between amino acids 276 and 393 of the S. cremoris proteinase, corresponding with amino acids 59 to 166 in the subtilisins. This region in the subtilisins includes the amino acids involved in the formation of the S1 specificity crevice (residues 125 to 127 and 152 to 154) and His-64, which, together with Ser-221 and Asp-32, constitutes the charge relay system crucial for enzymatic activity (15).

Asp-32 of the subtilisins is also found in a smaller region of homology with the S. cremoris proteinase. A stretch of seven amino acids around Asp-32, conserved in the subtilisins, is found around Asp-217 of the S. cremoris proteinase. In Fig. 6B, the results of the homology comparison are summarized and drawn to scale on a linear map. No homologies between the proteins present in the National Biomedical Research Foundation data bank and the truncated protein specified by ORF1 were found.

**DISCUSSION**

We sequenced over 7,000 bp of a region of the proteinase plasmid pWV05 of S. cremoris Wg2, which was shown to specify proteolytic activity. A 6,519-bp HindIII fragment contained within this sequence restored the proteolytic deficiency in S. lactis (Prt−) (14). From the two incomplete ORFs found on the fragment, only the largest was sequenced to its end. The first stop codon was located 5,706 bp downstream of the ATG start. The proposed transcription- and translation-regulatory sequences of this lactic acid streptococcal gene closely resemble those reported for B. subtilis and E. coli (11, 20, 24, 34) and are in agreement with the sequences determined by van der Vossen et al. (36a).

Metabolic regulation of proteinase synthesis in lactic acid streptococci has been observed (6, 7, 12, 16). The occurrence of a 36-bp region of dyad symmetry in the promoter region is suggestive of a regulatory region and might be a binding site for a regulatory protein (24). A similar region of dyad symmetry has been reported in front of the sprE gene that encodes the B. subtilis subtilisin E protease, a gene which is under catabolite repression (42).

From the homology comparison with the subtilisins, it is obvious that ORF2 specifies a serine proteinase. This finding is in accordance with the results of inhibition studies on the purified enzyme showing its sensitivity to the serine proteinase inhibitor phenylmethylsulfonyl fluoride (8, 9, 13). The three most-conserved regions include the triplet Arg-32, His-64, and Ser-221 of the reactive center of subtilisin (Arg-217, His-281, and Ser-620 in the S. cremoris proteinase). In the four enzymes compared, the three regions Asp-32/Asp-217, His-64/His-281, and Ser-221/Ser-620 share 50, 43, and 53% identical residues, respectively (Fig. 6A). When the percent match with at least one of the subtilisins was calculated and the most conservative amino acid replacements were regarded as identical residues (5), the homology increased to 83, 55, and 68%, respectively. The stretch of 107 amino acids around His-64 in the subtilisins also includes two sequences involved in the formation of the S1 specificity crevice (Ser-125–Leu-126–Gly-127 makes up one side of this pocket, and the side chains of Ala-152–Ala-153–Gly-154 form the other side [15]). The analogous sequences in the S. cremoris proteinase are Ser-349–Leu-350–Gly-351 and Ser-380–Ala-381–Gly-382. The latter region is part of a longer stretch of complete homology and includes the highly conserved Asn-155 (Asn-383 in the S. cremoris proteinase), which is important for stabilization of the reaction intermediate formed during proteolysis (15). The most striking feature of the S. cremoris proteinase in comparison with the subtilisins, however, is the presence of several stretches of amino acids not found in the subtilisin sequences (Fig. 6B). The distance
FIG. 3. Nucleotide sequence and inferred amino acid sequence of the S. cremoris Wg2 proteinase gene and its flanking regions. For both ORFs, the sequence of the nontranscribed DNA strand is presented. Numbering of the nucleotides is from the leftmost HindIII site. Amino acid numbering is shown under the sequence. The small untranslated region from position 866 to 1,205 contains the putative −35 and −10 sequences for ORF1 and ORF2 and is shown in more detail in Fig. 4. The putative transcriptional start sites are shown by small vertical arrows. Asp-217, His-281, and Ser-620 are boxed. At the 3′ end of the nucleotide sequence, the putative terminator of ORF2 is indicated. For details, see the text.
between Asp-32 and His-64 in subtilisin is doubled to 64 amino acids, whereas His-64 and Ser-221 are spaced by an extra 182 amino acids in the S. cremoris proteinase (from 157 in subtilisin to 339 in the streptococcal proteinase). In the region where subtilisin has a small exterior loop (Gly-160 to Asn-163), the S. cremoris proteinase contains a stretch of approximately 180 amino acids not found in subtilisin. Because the spatial relationship among the amino acids of the active center, the S1 specificity crevice, and Asn-383 are kept intact, we may speculate that this large insert (and perhaps some of the smaller ones) can be envisaged as protruding from a subtilisinlike core.

The predicted amino acid sequence gives the S. cremoris proteinase a calculated molecular weight of 200,000. This value cannot be easily reconciled with the results of Hugenholz et al. (13). These investigators showed that the proteolytic system of S. cremoris Wg2 consists of two proteinases, A and B, with estimated molecular weights of 140,000 each, based on sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Part of the difference in size can be explained by the assumption that the S. cremoris proteinase is synthesized as a preproenzyme, as are the subtilisins (33, 41). At the N terminus, there is a signal peptide-like sequence of 33 amino acids. The tentative cleavage site is separated from Asp-217 by 184 residues. If we assume that 30 to 40 amino acids are required for proper folding of the mature enzyme at the N terminus (in subtilisin, this number is 32), approximately 130 to 140 residues would remain, which might constitute a pro-region (sizes of gram-positive pro-sequences range from 77 to about 194 amino acids (37)). From 16 to 17 kDa could be split off in this way from the N terminus. Recently, the gene for the extracellular serine protease of Serratia marcescens was cloned and sequenced (43). The mature protease is formed by processing of a proenzyme at the N terminus, as well as at the C-terminal part. The mature enzyme contains 388 amino acids (size, 41 kDa). The C-terminal peptide split off contains another 637 residues, with an approximate size of 70 kDa. Similarly, the S. cremoris proteinase might be processed at the C terminus, and this,

### FIG. 5. Hypothetical stem-and-loop structures flanking the S. cremoris Wg2 proteinase gene. (A) Hairpin structure in the promoter region of ORF2. The ~35 and ~10 sequences of ORF2 are indicated by thin arrows. Part of the sequence is presented double stranded to show the possible promoter region, 1,155/1,132, of ORF1 (thick arrows). (B) Terminator structure 130 bases downstream of the TAG stop codon of ORF2.

together with the putative processing steps at the N terminus, might result in a mature enzyme of 140 kDa.

The observation that, upon prolonged incubation, the purified enzyme is subject to self-digestion might offer an alternative explanation. Because low (1 mM) concentrations of Ca²⁺ ions activate the similar S. cremoris AC1 proteinase (9), it is conceivable that, under the isolation conditions used, the streptococcal proteinase is released from the cell wall by a self-digestion step, resulting in the purification of a truncated protein of 140 kDa. Indeed, under certain conditions, proteinase activity can be isolated in protein bands with molecular weights as low as 60,000 (J. Erkelens, personal communication). Interestingly, at least one of these self-digestion sites, which have to be postulated to explain these observations, may be identified in the C terminus of the proteinase. The amino acid sequence Leu-1434 to Ser-1437 is identical to one of the digestion sites of the S. cremoris AC1 and S. lactis NCDO763 proteinase in β-casein (A. Geis and W. Bockelmann, personal communication; 19). This hypothesis is also in agreement with the genetic finding that the cloned HindIII fragment specifies a proteinase lacking 130 amino acids at the C terminus which still can complement proteinase deficiency. Moreover, a deletion in the gene removing the C-terminal 343 residues still specified an active enzyme (13a), showing that at least part of the C-terminal region can be deleted without severely affecting enzyme activity. This finding is in contrast with the situation for the extracellular proteolytic activity of S. marcescens in E. coli, which is lost upon introduction of frame shifts in the C-terminal part of the gene (43). Both processing and self-digestion might also offer an explanation for the localization of the genetic information for both proteins A and B (each with a size of 140 kDa) on the HindIII fragment in pGBK500 (14). Possibly, one of the proteins is a processing or breakdown product of the other which still exhibits proteinase activity. To match this with the crossed immunoelectrophoresis results, one would have to postulate the exposure of completely different antigenic determinants in A and B as a result of one of these digestion steps.

All S. cremoris proteinases are extremely specific and degrade only β-casein (9, 10, 19, 38), except the S. cremoris AM1 and SK11 proteinases, which also hydrolyze α-casein (38). A protein like bovine serum albumin, readily degraded by the subtilisins, is not hydrolyzed by S. cremoris proteinases (9). A differentiation between lactic acid streptococcal proteinases exists in the production of bitter peptides during
cheese production. This major flavor defect in cheese is thought to be related to proteinase (over) activity (16, 32). It will be interesting to learn whether the inserts or the long C terminus found in the S. cremoris Wg2 proteinase are involved in this specificity. Deletion analysis experiments are in progress to answer these questions and to find out whether the long C terminus plays a role in cell wall association, as suggested by the self-digestion hypothesis. We believe that the elucidation of the complete nucleotide sequence reported here is important in at least three respects. (i) It provides a basis for the construction of efficient expression and secretion vectors for lactic acid streptococci. (ii) It is essential for future research aimed to determine which parts of the enzyme are involved in its specificity. (iii) The nucleotide sequence is basic to research aimed at changing the properties of the enzyme to make it more suitable for dairying and, perhaps, other purposes.

ACKNOWLEDGMENTS

This work was supported by the Programme Committee on Biotechnology of the Netherlands and the Biomolecular Engineering Programme of the Commission of the European Communities. We thank Beike Leegte for typing the manuscript and Henk Mulder for photography and preparation of the figures.

LITERATURE CITED


