INFECTION AND IMMUNITY, Jan. 1992, p. 124–135
0019-9567/92/010124-12$02.00/0

Vol. 60, No. 1

Nucleotide Sequences of Two Adjacent M or M-Like Protein Genes of Group A Streptococci: Different RNA Transcript Levels and Identification of a Unique Immunoglobulin A-Binding Protein

DEBRA E. BESSEN* AND VINCENT A. FISCHETTI

Laboratory of Bacterial Pathogenesis and Immunology, The Rockefeller University, 1230 York Avenue, Box 276, New York, New York 10021-6399

Received 26 July 1991/Accepted 24 October 1991

M protein is a key virulence factor present on the surface of group A streptococci. M protein is defined by its antiphagocytic function, whereas M-like proteins, while structurally related to M proteins, lack an established antiphagocytic function. Group A streptococci can be divided into two main groups (class I and II) on the basis of the presence or absence of certain antigenic epitopes within the M and M-like molecules, and importantly, the two classes correlate with the disease-causing potential of group A streptococci. In an effort to better understand this family of molecules, a 2.8-kb region containing the two M protein-like genes from a class II isolate (serotype 2) was cloned and sequenced. The two genes lie adjacent to one another on the chromosome, separated by 211 bp, and have many structural features in common. The *emm*L2.1-derived product (ML2.1 protein) is immunoreactive with type-specific antiserum, a property associated with M proteins. The cloned product of the downstream gene, *emm*L2.2 (ML2.2 protein), is an immunoglobulin A (IgA)-binding protein, binding human myeloma IgA. Interestingly, the RNA transcript levels of *emm*L2.1 exceed that of *emm*L2.2 by at least 32-fold. Northern (RNA) hybridization and primer extension studies suggest that the RNA transcripts of *emm*L2.1 and *emm*L2.2 are monocistronic. The ML2.1 and ML2.2 proteins exhibit 53% amino acid sequence identity and differ primarily in their amino termini and peptidoglycan-spanning domains and in a Glu-Gln-rich region present only in the ML2.1 protein. However, the previously described M-like, IgA-binding protein from a serotype 4 isolate (Arp4) displays a higher level of amino acid sequence homology with the ML2.1 molecule than with the IgA-binding ML2.2 protein. Amino acid sequence alignments between all M and M-like proteins characterized to date suggest the existence of two fundamental M or M-like gene subclasses within class II organisms, represented by *emm*L2.1 and *emm*L2.2. In addition, IgA-binding activity can be found within both types of molecules.

Group A streptococci are responsible for a wide variety of human diseases, the most common of which are nasopharyngitis and impetigo. Nearly all clinical isolates have the antiphagocytic factor, M protein, on their surface. M proteins form alpha-helical coiled-coil fibrillar structures, within which there are regions consisting of tandemly arranged blocks of direct sequence repeats (8, 21, 33). This virulence factor displays extreme antigenic diversity within its amino-terminal region, which lies distal to the cell surface. The highly variable portions of M proteins form the basis of a serological typing scheme, and only antibodies directed to type-specific epitopes are capable of circumventing the antiphagocytic effect (8, 27). The definition of M protein as a type-specific, antiphagocytic factor was formulated in the 1930s, prior to our knowledge of its structural detail (27). In recent years, the sequences of eight M or M-like proteins have been reported (12, 14, 16, 21, 29, 30, 34). The M-like molecules are structurally similar to M proteins in that they exhibit significant levels of sequence homology; however, they are not considered to be M protein itself because an antiphagocytic property has not been formally demonstrated. Among the family of M or M-like molecules are immunoglobulin A (IgA)- and IgG-binding proteins (BP) (12, 14; also this report).

M and M-related proteins exhibit a moderate to high degree of sequence homology within their carboxy-terminal halves. Group A streptococci can be divided into two major classes partly on the basis of their immunoreactivity with a pair of monoclonal antibodies (MAbs) directed to epitopes which lie within the relatively conserved half of M proteins (2, 4). Class I isolates are defined as those binding one or both MAbs, whereas class II isolates do not bind either MAb. In addition, class II isolates produce opacity factor (4, 35), which turns serum opalescent and serves an unknown function for the organism. Conversely, class I isolates fail to exhibit opacity factor activity.

The class I-specific MAb binding sites map to a region of sequence repeats within M and M-like proteins, termed the C repeat domain, which is surface exposed and immediately adjacent to the cell wall (2, 26, 31). The degree of homology between the C repeat regions of any two molecules can be as low as 60% (2). The distinction between class I and II M or M-like proteins in the C repeat region appears to be due to only a small proportion of the amino acids that vary between molecules, suggesting a role for biological pressure in maintaining the class-specific epitopes (2). Most importantly, the class I and II groupings correlate with several pathogenic properties of these organisms (3, 4). For example, nearly all serotypes found in association with major outbreaks of rheumatic fever are class I. Among class I organisms, IgG-binding activity distinguishes between nasopharyngeal and impetigo isolates. IgA-binding activity is a class II-specific property.

The objective of this study is to increase our understanding of class II M and M-like protein structure and function.

* Corresponding author.

Two M or M-like protein genes were cloned from a single class II isolate (serotype 2). The product of the downstream gene (*emm*L2.2) exhibits IgA-binding activity and has strong homology to the predicted amino acid sequence of a previously described gene (*enn*X) of unknown function from a type 49 isolate. The *emm*L2.1 gene product reacts with M2 type-specific antibodies and displays the greatest homology to the previously described M49 protein (*emm*49) and type 4 IgA-BP (*arp*4), both of which originate from class II streptococci. The data suggests that two fundamental M or M-like gene subclasses exist within class II organisms and that IgA-binding activity can be found within both forms. In addition, the RNA transcript levels of the two genes differ significantly.

MATERIALS AND METHODS

Bacteria. Group A streptococcal strain T2/44/RB4/119 (class II) is the M2 typing strain from the Lancefield collection (The Rockefeller University). This culture contains a mixture of at least two distinct colony phenotypes when grown on blood agar plates at 6% CO_2: a low hemolytic colony type that is rich in M protein (T2/MR) and a high hemolytic colony type that is deficient in M protein (T2/MD). The M-rich, T2/MR isolate is designated as such on the basis of its survival in nonimmune whole human blood (data not shown) (27). Both the T2/MR and T2/MD isolates were characterized as group A streptococci on the basis of their positive precipitin reactions with group A carbohydrate antiserum. All isolates were derived from single colony picks; only the segregated T2/MR and T2/MD isolates were used in this study. The high hemolytic colony type exemplified by T2/MD represents approximately 0.1% or less of the CFUs in the original T2/44/RB4/119 culture. Additional streptococcal strains include D471 and S43/192 (type 6, class I) and 29452 (type 22, class II; Institute of Hygiene and Epidemiology, Prague, Czechoslovakia).

Antibodies. Antibodies were raised in rabbits to ColiM6 (the cloned gene product of *emm*6.1) and affinity purified with immobilized, pure ColiM6 protein (9). M2 typing sera was raised in rabbits to whole streptococci of the M2 serotype, and non-type-specific antibodies were removed following absorption to whole organisms of heterologous serotypes (27). Human myeloma IgA and human IgG-Fc fragment were obtained from Organon Teknika (Durham, N.C.). Immunoabsorbent materials were prepared by covalent linkage of lysates of *Escherichia coli* XL-1 Blue cells or S43/192 streptococcal cells to glutardialdehyde-activated beads (Boehringer Mannheim Corp., Indianapolis, Ind.). Subsaturating concentrations of beads, were incubated overnight with the immunoabsorbent beads, and unbound material was used for plaque screening and immunoblot analysis. Anti-ColiM6 was absorbed against the *E. coli* XL-1 lysate, whereas human myeloma IgA and human IgG-Fc fragment were absorbed against both XL-1 and S43/192 lysates. The synthetic peptide, SRKGLRRDLDASREAKKQVEK(C), was covalently linked to ovalbumin through a carboxy-terminal Cys of the peptide and injected into rabbits, and specific antibodies were affinity purified from the whole sera (4); the peptide corresponds in sequence to amino acid residues 240 to 260, found within the C repeat region of M6 protein (2, 21) (antisera and purified peptide were kindly provided by Institut Merieux).

Cloning and subcloning. Randomly sheared chromosomal DNA, derived from a low hemolytic colony of strain T2/44/RB4/119 termed T2/MR (type 2, class II), was cloned into λgt11 through *Eco*RI linkers and partially sequenced as previously published (2). Plaques were first screened with anti-ColiM6 and then with M2 typing sera. Two clones underwent further study: clones 9 and 11 (Fig. 1). Purified λgt11 replicative-form DNA containing inserts 9 and 11 was obtained from *E. coli* lysogens, and *Eco*RI-excised inserts were subcloned into both M13mp19 and pUC18 vectors. The insert from clone 11 was subcloned into pUC18 to generate pML2-11. The 1.6-kb *Pst*I-*Eco*RI fragment derived from pML2-11 was ligated into pUC18 to construct pML2-14.

RNA purification. T2/MR and T2/MD cells were grown at 37°C under aerobic conditions in Todd-Hewitt broth to an optical density at 600 nm of 0.1, at which time glycine was added. Total cellular RNA was isolated from streptococcal cultures upon reaching an optical density at 600 nm of 0.5. RNA was purified following centrifugation on CsCl gradients and stored in the presence of vanadyl ribonucleosides (20, 22, 34).

DNA and RNA sequencing. Foreign DNA cloned into M13 and pUC vectors was sequenced by the dideoxy-chain termination method, using Sequenase (United States Biochemical Corp., Cleveland, Ohio). Overlapping inserts were generated in M13mp19 by T4 polymerase digestion using the Cyclone I Biosystem (I.B.I., New Haven, Conn.). Streptococcal DNA cloned into pUC18 vectors was sequenced because of the inability to clone the sense strand of major portions of insert 11 into M13 bacteriophage. RNA sequencing was performed by using avian myeloblastosis virus reverse transcriptase (Life Sciences, St. Petersburg, Fla.). Sequencing primers included both the M13 universal primer and synthetic oligonucleotide primers (The Rockefeller University Sequencing Facility).

Southern and Northern hybridizations. For Southern hybridizations, T2/MR and T2/MD chromosomal DNA was digested with *Pst*I, electrophoresed on agarose gels, and transferred to Hybond-N+ nylon membranes (Amersham). The 123-bp DNA ladder (Life Technologies, Inc., Gaithersburg, Md.) was used to estimate the sizes of restriction fragments. Total cellular RNA was electrophoresed on formaldehyde-containing agarose gels (34) and transferred to nylon membranes; RNA molecular size markers were from Life Technologies, Inc. Linear, double-stranded DNA probes were purified following restriction digestion of plasmids, electrophoresis on low-gelling-temperature agarose, and extraction of agarose from excised fragments; they were radiolabeled by random primed labeling (Boehringer Mannheim Corp.). Synthetic oligonucleotides 5′-TCTAATTTC TCGAATAATTCTGCT-3′ and 5′-TTCTTTAGTAGTCTTA GCTAAAGTTGT-3′ (complementary to bp 232 to 255 and 1812 to 1838, respectively; Fig. 3) were end labeled with T4 polynucleotide kinase (Pharmacia LKB Biotechnology Inc., Piscataway, N.J.). pNC1 contains an internal fragment of the streptokinase gene (*skc*) from a group C streptococcus (kindly provided by Joseph Ferretti, University of Oklahoma Health Sciences Center) (24) and was radiolabeled by nick translation (Amersham). All hybridizations were performed in 6× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) at 55°C, except that 5× SSPE (1× SSPE is 0.18 M NaCl, 10 mM $NaPO_4$, and 1 mM EDTA [pH 7.7]) was substituted for double-stranded DNA probes in Southern hybridizations.

Primer extension of RNA templates. Primer extension reactions were performed as RNA sequencing, except that dideoxynucleoside triphosphates were omitted from the reaction. By using DNA templates of known sequence, sequencing reactions containing dideoxynucleoside triphos-

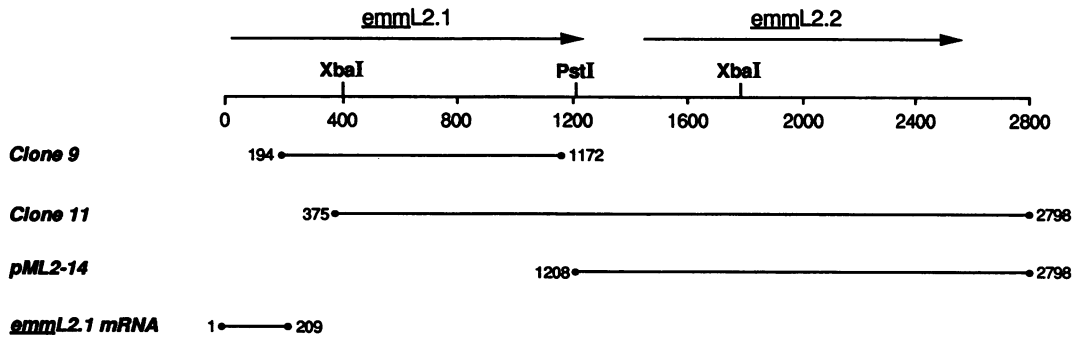

FIG. 1. Region of the T2/MR chromosome cloned and sequenced. Arrows indicate the coding regions of *emm*L2.1 and *emm*L2.2. Positions of streptococcal DNA that were cloned into the M13mp19 (inserts from clones 9 and 11) and pUC18 (pML2-14) vectors used for sequencing are shown. The region of the *emm*L2.1 mRNA which was sequenced is indicated. Restriction mapping of λgt11 clones 9 and 11 and T2/MR chromosomal DNA was performed in order to confirm that the insert DNA did not undergo rearrangement (data not shown).

phates were electrophoresed in parallel with the primer extension reactions in order to determine the transcriptional start sites.

Antibody absorption assay. Heat-killed streptococci were tested for nonimmune binding of human myeloma IgA, human polyclonal IgG, and human IgG-Fc fragment (Organon Teknika) by previously described methods (2–4).

Western blot (immunoblot) analysis. Western blot analysis was performed as previously described (2), except that the incubation buffer for human myeloma IgA and human IgG-Fc fragment consisted of 20 mM Tris-HCl (pH 7.5), saline, 0.5% Tween 20. IgA and IgG-Fc were used at a concentration of 1 μg/ml, and secondary antibodies were conjugated to alkaline phosphatase (Sigma Chemical Corp., St. Louis, Mo.). Streptococcal extracts obtained by treatment with the muralytic enzyme lysin were prepared by previously described methods (4, 10).

Sequence analysis. The Staden sequence analysis package and Protalign were used for analysis of nucleotide and amino acid sequences, respectively. The percent identity between two given sequences was calculated on the basis of the number of matched residues divided by the number of residues in the shortest sequence.

Nucleotide sequence accession number. The 2.8-kb nucleotide sequence encompassing genes *emm*L2.1 and *emm*L2.2 are available from EMBL/GenBank/DDJB under accession number X61276.

RESULTS

Characterization of cloned gene products. A λgt11 library was constructed with chromosomal DNA from T2/MR (2), which represents the predominant colony form of the M2 typing strain T2/44/RB4/119. Plaques resulting from lytic infection with λgt11 containing T2/MR DNA inserts were initially screened with anti-ColiM6 antibodies, in order to detect non-type-specific antigenic epitopes which are shared among M and M-like proteins. Phage from immunoreactive plaques generated by clones 9 and 11 were purified and rescreened with M2 typing sera; only clone 9 displayed type 2-specific immunoreactivity (Fig. 1). A previous report demonstrated that a β-galactosidase fusion product is expressed by λgt11 clone 9 lysogens following isopropyl-β-D-thiogalactopyranoside (IPTG) induction and is immunoreactive with both M2 typing sera and affinity-purified anti-ColiM6 on Western blots (2). While immunoreactivity with typing sera is a characteristic attributable to M proteins, determination of whether the ML2.1 protein is in fact the M2 protein requires functional tests of its antiphagocytic capacity.

A 1.6-kb *Pst*I-*Eco*RI fragment originally derived from clone 11 insert DNA was subcloned into pUC18 (pML2-14) in order to study the *emm*L2.2 gene product in the absence of *emm*L2.1 (Fig. 1). Whole-cell lysates of *E. coli* harboring pML2-14 bound both anti-ColiM6 and human myeloma IgA by Western blotting (Fig. 2A and B, lanes 4), and expression was not under IPTG control. The gene product expressed by pML2-14 displayed several bands, many of which are likely degradation products of the major band at 42 kDa. The *emm*L2.2 cloned gene product failed to bind human IgG-Fc fragment by Western blotting (Fig. 2C, lane 4).

Whole streptococci of the T2/MR parent strain bind both human myeloma IgA and human IgG-Fc fragment by a nonimmune mechanism (Table 1). The M-rich organism was

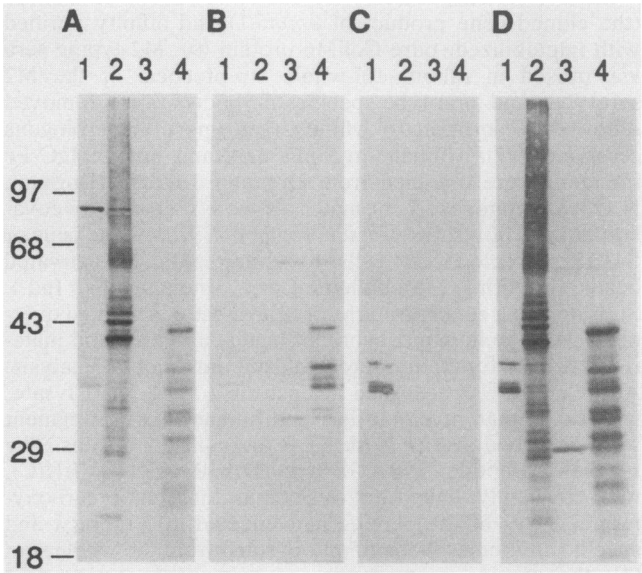

FIG. 2. Western blot analysis of the cloned *emm*L2.2 gene product. Lanes: 1, lysin extract of 29452; 2, lysin extract of T2/MR; 3, whole *E. coli* XL-1 cells; 4, whole *E. coli* XL-1 cells harboring pML2-14. (A) anti-ColiM6; (B) human myeloma IgA; (C) human IgG-Fc fragment; (D) anti-peptide (240 to 260) (preparation of antibodies is detailed in Materials and Methods). The positions of molecular size markers (in kilodaltons) are shown to the left.

TABLE 1. Nonimmune binding of Igs to whole group A streptococci

Strain	% Bound[a]		
	IgA	IgG	IgG-Fc
T2/MR	67	82	54
T2/MD	6	4	3
29452	86	85	60
D471	0	3	2

[a] Human myeloma IgA (IgA), human polyclonal IgG (IgG), and human IgG-Fc fragment (IgG-Fc).

digested with the muralytic enzyme lysin, and crude cellular extracts were examined by Western blotting (Fig. 2, lanes 2). Binding of human myeloma IgA could not be detected (Fig. 2B), despite strong binding by both anti-ColiM6 (Fig. 2A) and M2 typing sera (2), to major bands migrating at approximately 41 and 43 kDa. A multiple banding pattern is typical of lysin-extracted M proteins (10). In addition, an IgG-BP was not detected in lysin extracts of T2/MR (Fig. 2C, lane 2). In contrast, the IgG-BP but not the IgA-BP could be detected in lysin extracts of strain 29452 (Fig. 2B and C, lanes 1), despite the finding that 29452 strongly binds both IgA and IgG in the antibody absorption assay using whole organisms (Table 1). Thus, lysin treatment does not ensure extraction of intact Ig-BP from group A streptococci.

Nucleotide and deduced amino acid sequences. The DNA sequence of the inserts from clones 9 and 11 (Fig. 1) and the partial mRNA sequence of *emm*L2.1 is presented in Fig. 3. The open reading frame of the *emm*L2.1 gene is 1,221 bp long (bp 20 to 1240). The first 41 predicted amino acids are homologous to signal peptides of other M and M-like proteins (12, 14, 16, 17, 21, 29, 30, 34) (see Results). Therefore, the mature *emm*L2.1 gene product is expected to be 366 residues with a predicted molecular weight of 42,205; this size is in good accordance with the major 43-kDa band obtained by lysin extraction of the T2/MR parent strain (Fig. 2, lanes 2). A 7-amino-acid-residue repeat unit occurs twice near the amino terminus (amino acids 40 to 53), and four C repeat segments with intervening spacers are located at amino acids 117 to 244 (2) (Fig. 3).

A putative transcriptional terminator is indicated by the inverted repeat positioned downstream from the *emm*L2.1 stop codon (Fig. 3). Proposed promoter and ribosomal binding sites matching consensus sequences are found immediately upstream from the ATG start codon of *emm*L2.2. The noncoding region between *emm*L2.1 and *emm*L2.2 is 211 bp long.

The longest open reading frame for the *emm*L2.2 gene extends from nucleotide 1452 to 2567 (Fig. 3). The first 41 predicted amino acids are homologous to signal peptides of other M and M-like proteins (12, 14, 16, 17, 21, 29, 30, 34). Therefore, the mature *emm*L2.2 gene product is expected to be 331 residues with a predicted molecular weight of 36,769; this size is in reasonable accordance with the 42-kDa band expressed by *E. coli* harboring pML2-14 (Fig. 2, lanes 4). There is a single region of sequence repeats in the ML2.2 protein, consisting of three 23-residue C repeat segments (2), separated by two spacers (spanning amino acids 90 to 203). The *emm*L2.2 gene product fails to bind the class I-specific MAbs which recognize epitopes in the C repeat regions of class I molecules (data not shown). The predicted amino acid sequence of *emm*L2.2 has class II characteristics in that the amino acids critical to binding of the class I-specific MAbs are different (2); the class I Arg position is a Ser in ML2.2 (at amino acids 100, 142, and 191), and the class I Asp is a Glu in ML2.2 (at amino acids 104, 146, and 195). Immunoreactivity of the *emm*L2.1 (2) and *emm*L2.2 (Fig. 2D, lane 4) cloned gene products with affinity-purified antibodies directed to a synthetic peptide corresponding in sequence to part of the C repeat region of M6 protein, is in strong support of the open reading frames indicated in Fig. 3.

RNA transcription. It was of interest to determine whether both *emm*L2.1 and *emm*L2.2 are transcribed and whether *emm*L2.1 and *emm*L2.2 are part of polycistronic or monocistronic mRNAs. Oligonucleotide probes specific for each gene were tested for hybridization to dot blots containing total cellular RNA isolated from the T2/MR (M+) isolate (Fig. 4). The data reveals that the amount of *emm*L2.1 transcript exceeds that of *emm*L2.2 by at least 32-fold. Northern blots indicate transcript sizes for *emm*L2.1 and *emm*L2.2 of approximately 1.30 and 1.45 kb, respectively, suggesting that the mRNAs for each gene are monocistronic (Fig. 5). The stated binding specificities and specific activities of the oligonucleotide probes could be confirmed by Southern blot hybridization (Fig. 6). T2/MR chromosomal DNA was digested with *Pst*I, which recognizes a single site within the sequenced 2.8-kb region at bp 1206 (Fig. 3). The *emm*L2.1-specific oligonucleotide probe detected a DNA fragment migrating at 1.5 kb, whereas the *emm*L2.2-specific oligonucleotide probe bound to a 2.1-kb band (Fig. 6A and B, lanes 1). The nearly equivalent signals of the two oligonucleotide probes following hybridization to chromosomal DNA indicates that any purported differences in the specific activities of the probes do not account for the major difference in transcript levels observed by RNA dot blot hybridization.

In order to further substantiate that the relatively low level of RNA detected with the *emm*L2.2-specific probe is above background, RNA transcripts from the T2/MR isolate were compared with RNA derived from the M-deficient organism T2/MD. Both the T2/MR and T2/MD isolates originated from the T2/44/RB4 typing strain. The T2/MD organism was characterized as M-deficient on the basis of the following: (i) its failure to survive in a bactericidal assay under conditions whereby T2/MR resisted opsonophagocytosis and (ii) the lack of immunoreactive material in lysin extracts, when analyzed by Western blot using anti-ColiM6 antibody (data not shown). In addition, T2/MD lacked IgA-binding activity (Table 1). RNA dot blot hybridization revealed that the levels of both *emm*L2.1- and *emm*L2.2-specific RNAs derived from T2/MD (M−) are strikingly lower than those from the T2/MR (M+) isolate (Fig. 4). The difference in the *emm*L2.2 mRNA levels between T2/MR and T2/MD confirms that the *emm*L2.2 gene is transcribed in the T2/MR, IgA-binding isolate, despite its relatively low level compared with that of *emm*L2.1 mRNA.

Despite the apparent lack of detectable *emm*L2.1 and *emm*L2.2 mRNA in T2/MD, the *emm*L2.1 and *emm*L2.2 probes hybridize with 1.6- and 2.1-kb *Pst*I fragments derived from T2/MD chromosomal DNA by Southern blot (Fig. 6A to D, lanes 2), with signal intensities similar to those observed with the 1.5- and 2.1-kb fragments of T2/MR (lane 1). Since the *emm*L2.1-specific oligonucleotide probe corresponds to sequences located near the 5′ end of the *emm*L2.1 gene, where type-specific sequences are expected to reside, it is likely that T2/MD is also of serotype 2 origin. This is supported by the finding that the *emm*L2.1-specific oligonucleotide probe was restricted to type 2 isolates when tested for hybridization to DNA derived from organisms represent-

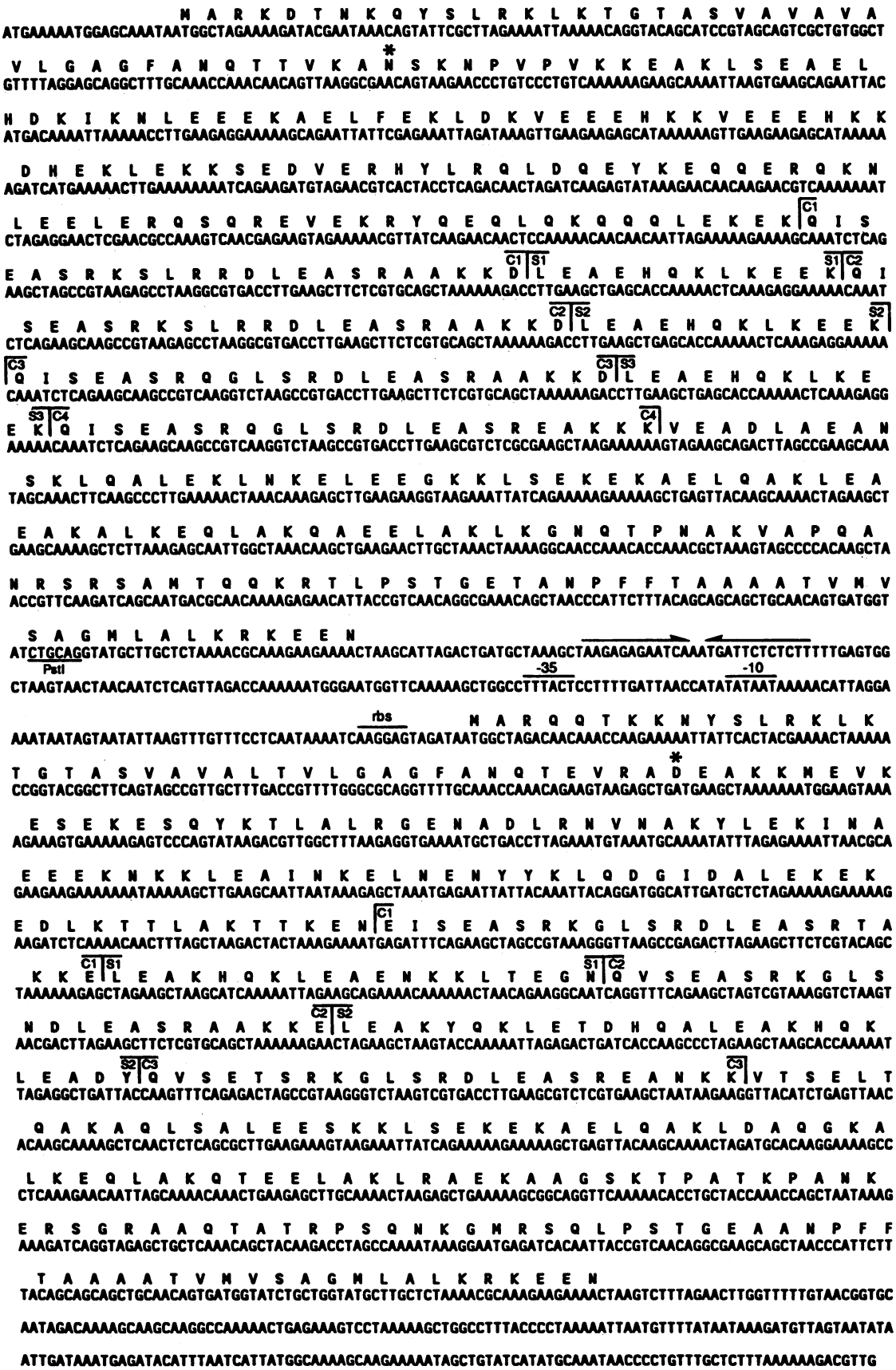
M A R K D T N K Q Y S L R K L K T G T A S V A V A V A
ATGAAAAATGGAGCAAATAATGGCTAGAAAAGATACGAATAAACAGTATTCGCTTAGAAAATTAAAAACAGGTACAGCATCCGTAGCAGTCGCTGTGGCT
V L G A G F A N Q T T V K A N S K N P V P V K K E A K L S E A E L
GTTTTAGGAGCAGGCTTTGCAAACCAAACAACAGTTAAGGCGAACAGTAAGAACCCTGTCCCTGTCAAAAAAGAAGCAAAATTAAGTGAAGCAGAATTAC
H D K I K N L E E E K A E L F E K L D K V E E E H K K V E E E H K K
ATGACAAAATTAAAAACCTTGAAGAGGAAAAAGCAGAATTATTCGAGAAATTAGATAAAGTTGAAGAAGAGCATAAAAAAGTTGAAGAAGAGCATAAAAA
D H E K L E K K S E D V E R H Y L R Q L D Q E Y K E Q Q E R Q K N
AGATCATGAAAAACTTGAAAAAAAATCAGAAGATGTAGAACGTCACTACCTCAGACAACTAGATCAAGAGTATAAAGAACAACAAGAACGTCAAAAAAAT
L E E L E R Q S Q R E V E K R Y Q E Q L Q K Q Q Q L E K E K Q I S
CTAGAGGAACTCGAACGCCAAAGTCAACGAGAAGTAGAAAAACGTTATCAAGAACAACTCCAAAAACAACAACAATTAGAAAAAGAAAAGCAAATCTCAG
E A S R K S L R R D L E A S R A A K K D L E A E H Q K L K E E K Q I
AAGCTAGCCGTAAGAGCCTAAGGCGTGACCTTGAAGCTTCTCGTGCAGCTAAAAAAGACCTTGAAGCTGAGCACCAAAAACTCAAAGAGGAAAAACAAAT
S E A S R K S L R R D L E A S R A A K K D L E A E H Q K L K E E K
CTCAGAAGCAAGCCGTAAGAGCCTAAGGCGTGACCTTGAAGCTTCTCGTGCAGCTAAAAAAGACCTTGAAGCTGAGCACCAAAAACTCAAAGAGGAAAAA
Q I S E A S R Q G L S R D L E A S R A A K K D L E A E H Q K L K E
CAAATCTCAGAAGCAAGCCGTCAAGGTCTAAGCCGTGACCTTGAAGCTTCTCGTGCAGCTAAAAAAGACCTTGAAGCTGAGCACCAAAAACTCAAAGAGG
E K Q I S E A S R Q G L S R D L E A S R E A K K K V E A D L A E A N
AAAAACAAATCTCAGAAGCAAGCCGTCAAGGTCTAAGCCGTGACCTTGAAGCGTCTCGCGAAGCTAAGAAAAAAGTAGAAGCAGACTTAGCCGAAGCAAA
S K L Q A L E K L N K E L E E G K K L S E K E K A E L Q A K L E A
TAGCAAACTTCAAGCCCTTGAAAAACTAAACAAAGAGCTTGAAGAAGGTAAGAAATTATCAGAAAAAGAAAAAGCTGAGTTACAAGCAAAACTAGAAGCT
E A K A L K E Q L A K Q A E E L A K L K G N Q T P N A K V A P Q A
GAAGCAAAAGCTCTTAAAGAGCAATTGGCTAAACAAGCTGAAGAACTTGCTAAACTAAAAGGCAACCAAACACCAAACGCTAAAGTAGCCCCACAAGCTA
N R S R S A M T Q Q K R T L P S T G E T A N P F F T A A A A T V M V
ACCGTTCAAGATCAGCAATGACGCAACAAAAGAGAACATTACCGTCAACAGGCGAAACAGCTAACCCATTCTTTACAGCAGCAGCTGCAACAGTGATGGT
S A G M L A L K R K E E N
ATCTGCAGGTATGCTTGCTCTAAAACGCAAAGAAGAAAACTAAGCATTAGACTGATGCTAAAGCTAAGAGAGAATCAAATGATTCTCTCTTTTTGAGTGG
PstI
-35
-10
CTAAGTAACTAACAATCTCAGTTAGACCAAAAAATGGGAATGGTTCAAAAAGCTGGCCTTTACTCCTTTTGATTAACCATATATAATAAAAACATTAGGA
rbs
M A R Q Q T K K N Y S L R K L K
AAATAATAGTAATATTAAGTTTGTTTCCTCAATAAAATCAAGGAGTAGATAATGGCTAGACAACAAACCAAGAAAAATTATTCACTACGAAAACTAAAAA
T G T A S V A V A L T V L G A G F A N Q T E V R A D E A K K N E V K
CCGGTACGGCTTCAGTAGCCGTTGCTTTGACCGTTTTGGGCGCAGGTTTTGCAAACCAAACAGAAGTAAGAGCTGATGAAGCTAAAAAAATGGAAGTAAA
E S E K E S Q Y K T L A L R G E N A D L R N V N A K Y L E K I N A
AGAAAGTGAAAAAGAGTCCCAGTATAAGACGTTGGCTTTAAGAGGTGAAAATGCTGACCTTAGAAATGTAAATGCAAAATATTTAGAGAAAATTAACGCA
E E E K N K K L E A I N K E L N E N Y Y K L Q D G I D A L E K E K
GAAGAAGAAAAAAATAAAAAGCTTGAAGCAATTAATAAAGAGCTAAATGAGAATTATTACAAATTACAGGATGGCATTGATGCTCTAGAAAAAGAAAAAG
E D L K T T L A K T T K E N E I S E A S R K G L S R D L E A S R T A
AAGATCTCAAAACAACTTTAGCTAAGACTACTAAAGAAAATGAGATTTCAGAAGCTAGCCGTAAAGGGTTAAGCCGAGACTTAGAAGCTTCTCGTACAGC
K K E L E A K H Q K L E A E N K K L T E G N Q V S E A S R K G L S
TAAAAAAGAGCTAGAAGCTAAGCATCAAAAATTAGAAGCAGAAAACAAAAAACTAACAGAAGGCAATCAGGTTTCAGAAGCTAGTCGTAAAGGTCTAAGT
N D L E A S R A A K K E L E A K Y Q K L E T D H Q A L E A K H Q K
AACGACTTAGAAGCTTCTCGTGCAGCTAAAAAAGAACTAGAAGCTAAGTACCAAAAATTAGAGACTGATCACCAAGCCCTAGAAGCTAAGCACCAAAAAT
L E A D Y Q V S E T S R K G L S R D L E A S R E A N K K V T S E L T
TAGAGGCTGATTACCAAGTTTCAGAGACTAGCCGTAAGGGTCTAAGTCGTGACCTTGAAGCGTCTCGTGAAGCTAATAAGAAGGTTACATCTGAGTTAAC
Q A K A Q L S A L E E S K K L S E K E K A E L Q A K L D A Q G K A
ACAAGCAAAAGCTCAACTCTCAGCGCTTGAAGAAAGTAAGAAATTATCAGAAAAAGAAAAAGCTGAGTTACAAGCAAAACTAGATGCACAAGGAAAAGCC
L K E Q L A K Q T E E L A K L R A E K A A G S K T P A T K P A N K
CTCAAAGAACAATTAGCAAAACAAACTGAAGAGCTTGCAAAACTAAGAGCTGAAAAAGCGGCAGGTTCAAAAACACCTGCTACCAAACCAGCTAATAAAG
E R S G R A A Q T A T R P S Q N K G M R S Q L P S T G E A A N P F F
AAAGATCAGGTAGAGCTGCTCAAACAGCTACAAGACCTAGCCAAAATAAAGGAATGAGATCACAATTACCGTCAACAGGCGAAGCAGCTAACCCATTCTT
T A A A A T V M V S A G M L A L K R K E E N
TACAGCAGCAGCTGCAACAGTGATGGTATCTGCTGGTATGCTTGCTCTAAAACGCAAAGAAGAAAACTAAGTCTTTAGAACTTGGTTTTTGTAACGGTGC
AATAGACAAAAGCAAGCAAGGCCAAAAACTGAGAAAGTCCTAAAAAGCTGGCCTTTACCCCTAAAAATTAATGTTTTATAATAAAGATGTTAGTAATATA
ATTGATAAATGAGATACATTTAATCATTATGGCAAAAGCAAGAAAAATAGCTGTATCATATGCAAATAACCCCTGTTTGCTCTTTAAAAAAGACGTTG

FIG. 3. Nucleotide and deduced amino acid sequences of *emm*L2.1 and *emm*L2.2. An inverted repeat is indicated by arrows. Putative promoter sequences (−35 and −10) and ribosomal binding site (rbs) are indicated. Amino acid residue numbers are given for both *emm*L2.1 and *emm*L2.2, beginning with what is believed to be the first residue of the mature forms (*), on the basis of sequence homologies with leader sequences from other M and M-like protein genes (see Results). C repeat segments (C) and intervening spacers (S) are indicated (2); a complete C repeat unit is defined as a C repeat segment plus the following spacer (2). DNA was sequenced in both orientations for bp 195 to 2663; for bp 1 to 209, *emm*L2.1 mRNA was sequenced. The vectors containing cloned streptococcal DNA that was sequenced are shown in Fig. 1. Oligonucleotides used for hybridization (Fig. 4 through 6) are complementary to bp 232 to 255 and bp 1812 to 1838; this corresponds to DNA encoding (K)AELFEKL(D) (*emm*L2.1) and TTLAKTTKE (*emm*L2.2), respectively.

ing 23 different serotypes (data not shown). It remains to be determined whether the 0.1-kb difference between the *emm*L2.1-specific, 1.6-kb *Pst*I fragment from T2/MD and the 1.5-kb fragment from T2/MR is related to decreased transcription of the T2/MD genes. In another control, hybridization with the pNC1 probe containing an internal fragment of the streptokinase gene, *skc*, confirms that equivalent amounts of RNA (Fig. 4) and DNA (Fig. 6E) were present for the T2/MR and T2/MD isolates.

Using the *emm*L2.1-specific oligonucleotide corresponding to bp 232 to 255, primer extension of RNA from T2/MR indicates a single major transcript beginning at approximately 68 bp upstream from the proposed start codon (bp 20 in Fig. 3) (data not shown). Based on primer extension results and the position of the putative transcriptional terminator (Fig. 3), the expected size of the *emm*L2.1 transcript is 1313 bp, which is in accordance with the estimated 1.30-kb band detected by Northern hybridization (Fig. 5). Primer extension using the *emm*L2.2-specific oligonucleotide corresponding to bp 1812 to 1838 indicates a single major transcriptional start site at approximately bp 1400 (data not shown); this estimated position is in accordance with the proposed −10 and −35 promoter sites indicated in Fig. 3. On the basis of the size of the *emm*L2.2 transcript as determined by Northern blot (Fig. 5), the transcriptional terminator would be located about 283 bp downstream from the stop codon; the sequence of this region was not determined in this report (Fig. 3).

A DNA probe corresponding to an internal fragment of the *emm*L2.1 gene (bp 195 to 1171), hybridized to two *Pst*I fragments from the M-rich T2/MR isolate, migrating at 1.5 and 2.1 kb by Southern blotting (Fig. 6C, lane 1). Similarly, a DNA probe corresponding in position to a major portion of the *emm*L2.2 gene and downstream sequences (bp 1785 to 2798) hybridized to only two bands also migrating at 1.5 and 2.1 kb (Fig. 6D, lane 1). Since no additional bands are detected with either of the large DNA probes, it is likely that the T2/MR isolate contains only two M or M-like genes within its chromosome.

Sequence identities. Amino acids 1 through 74 of mature ML2.1 protein and 1 through 71 of the mature ML2.2 protein sequence display only very limited homologies to one another or with other M and M-like molecules. This parallels the finding that the amino termini of M proteins are hypervariable (1, 25). According to maximal alignment of sequences by the Protalign algorithm, there is 53% amino acid sequence identity between ML2.1 and ML2.2 proteins, and the homology is located for the most part, within their carboxy-terminal two-thirds (data not shown). However, ML2.2 protein is more closely related to the deduced sequence of *enn*X (82% identity) than to ML2.1, whereas ML2.1 protein is most similar to M49 and Arp4 (77 and 70% identity, respectively) (12, 16). The two IgA-BPs, Arp4 (12)

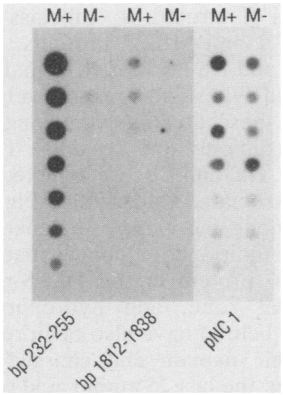

FIG. 4. Dot blot hybridization of total cellular RNA using *emm*L2.1- and *emm*L2.2-specific probes. Twofold dilutions of total cellular RNA from T2/MR (M+) and T2/MD (M−) were immobilized on nylon membranes (the highest concentration was 15 μg of RNA). RNA was hybridized with 1.6-pmol/ml ^{32}P-end-labeled oligonucleotide probes corresponding to bp 232 to 255 and 1812 to 1838 (specific activities of 2.05×10^5 and 2.57×10^5 cpm per pmol, respectively). Exposure times of autoradiograms and photographs are equivalent for both probes. The pNC1 probe containing the *skc* gene (control) confirms that equivalent amounts of RNA are present from the T2/MR and T2/MD isolates.

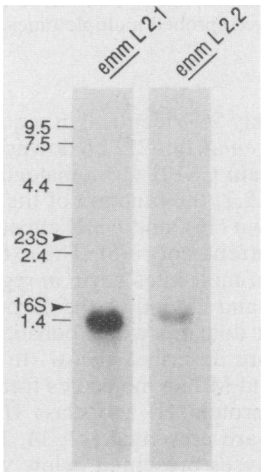

FIG. 5. Northern hybridization using *emm*L2.1- and *emm*L2.2-specific probes. Each lane contains 22.5 μg of total cellular RNA from T2/MR. Northern blots were probed with ^{32}P-end-labeled oligonucleotide probes corresponding to bp 232 to 255 (*emm*L2.1 specific) and 1812 to 1838 (*emm*L2.2 specific). The specific activities of the *emm*L2.1- and *emm*L2.2-specific probes were 1.6×10^5 and 6.8×10^5 cpm/pmol, respectively, and they were used at concentrations of 0.5 and 0.2 pmol/ml, respectively. Exposure times of autoradiograms and photographs are equivalent for both probes. The positions of RNA molecular size markers (in kilobases) are shown to the left.

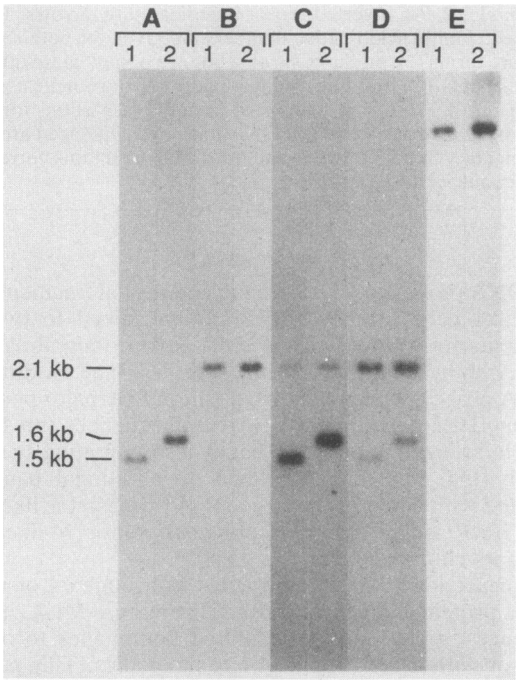

FIG. 6. Southern hybridization using *emm*L2.1- and *emm*L2.2-specific probes. Chromosomal DNA (approximately 1 μg) from T2/MR (lanes 1) and T2/MD (lanes 2) was cut with *Pst*I and probed with DNA corresponding to bp 232 to 255 (A lanes), 1812 to 1838 (B lanes), 195 to 1171 (λgt11 clone 9 insert DNA) (C lanes), and 1785 to 2798 (1-kb *Xba*I-*Eco*RI fragment of pML2-14) (D lanes). The pNC1 probe containing the *skc* gene (E lanes) confirms that equivalent amounts of DNA are present from T2/MR and T2/MD. The synthetic oligonucleotides (A and B lanes) were used at 0.2 pmol/ml and had specific activities of 1.37×10^6 and 1.13×10^6 cpm per pmol, respectively; exposure times of autoradiograms and photographs are equivalent for both oligonucleotide probes. T2/MR and T2/MD DNA were electrophoresed on a single agarose gel in nonadjacent lanes; DNA was simultaneously transferred to two sheets of Hybond, and blots were reprobed multiple times.

and ML2.2, are only 56% identical in sequence, despite their similar functions. *enn*X lies 212 bp downstream from *emm*49 in the type 49 strain CS101 (16), analogous to the genomic position of *emm*L2.2; the function of the *enn*X cloned gene product is unknown (16), and *enn*X appears to be transcriptionally silent in streptococci (6). The extensive homologies within both sets of molecules begin in regions preceding the C repeat domain and extend to the carboxy terminus. The distinct regions or domains which constitute the ML2.1 and ML2.2 proteins are described below. In addition, homologies to other M and M-like molecules from class I (M5, M6, M12, M24, and protein H) and class II (Arp4, M49, and *enn*X) organisms are presented (12, 14, 16, 21, 29, 30, 34); the Protalign analyses presented below were performed on individual domains rather than whole molecules. Extensive sequence alignments between M and M-like proteins have also been discussed by Haanes and Cleary (16).

(i) EQ-rich region. The Glu-Gln (EQ)-rich region is present in the M or M-like molecules, ML2.1, M49, Arp4, and protein H (ProtH, an IgG-BP) (Fig. 7A) (2, 12, 14, 16). This region is at least 41 amino acids long and lies adjacent to the amino terminus of the C repeat region, and its EQ content is 55 to 60%. In sharp contrast, the Asp-Asn content of this region is less than 3%. Maximally aligned sequence identities between any two of the four EQ-rich regions identified to date range from 62 to 93%.

While an analogous EQ-rich region is lacking in the ML2.2 and *enn*X sequences, a 14-residue stretch of almost complete identity (93%) is common to ML2.2 and the predicted *enn*X sequence (Fig. 7B). This short region of identity is located immediately adjacent to the amino-terminal side of the C repeat domain, at a position comparable to that of the EQ-rich region.

(ii) C repeat domain. Sequence homologies between the C repeat blocks of ML2.1 protein and other class I and II M and M-like proteins have been previously described in detail (2) and are estimated to be as low as approximately 60%. One spacer segment (S2; Fig. 3) of ML2.2 protein is distinct from other M and M-like molecules in that it is 26 residues long, in contrast to 12- or 19-amino-acid spacers observed elsewhere (2) (data not shown). *enn*X is analogous to ML2.2 protein in having an extended spacer segment beginning at bp 1981 (16) (data not shown).

(iii) Cell wall-associated helical zone. The cell wall-associated helical zone is a 54- or 61-amino acid stretch following the C repeat region which, in type 6 strain D471, is buried within the cell wall (probably within the group carbohydrate layer) (31). According to the algorithm of Garnier et al. (13) for predicting secondary structure, this region is 100% alpha-helical. The sequence alignments of ML2.1 and ML2.2 proteins (70% identical), ML2.1 and M6 proteins (90% identical), and ML2.2 and M6 proteins (70% identical) are illustrated in Fig. 8. However, if one focuses on the carboxy-terminal 40 amino acid residues of the helical zone, the homology between any two molecules exceeds 85%.

(iv) Proline-glycine-threonine-serine (PGTS)-rich domain. Carboxy terminal to the cell wall-associated helical zone is a domain that is largely beta-turn and beta-sheet in character (13). The PGTS-rich domain is believed to span the peptidoglycan (11, 21), and amino acids Pro, Gly, Thr, and Ser make up approximately 40% of its content. The data suggest there are at least three basic groups of PGTS-rich structures, one common to class I organisms and two found among class II organisms (Fig. 9). The identities among individual sequences within any one of these three groups approach 100%. The class II structures, subclass IIa and IIb, are typified by ML2.1 and ML2.2 proteins, respectively. The (sub)class I, IIa, and IIb PGTS-rich domains differ widely in length (55, 36, and 49 residues, respectively). Within the first three-fourths of these domains, the longest stretch of sequence homology is only four residues. These three groupings, (sub)class I, IIa, and IIb, are depicted in the alignments of the cell wall-associated helical zone (Fig. 8); in this region, the identities among individual sequences within a group approach 100%. Additional sequences that are highly homologous to the M6 protein in the PGTS-rich domain, cell-associated helical zone, and hydrophobic domain and charged tail (see below) have also been reported (23).

(v) Hydrophobic domain and charged tail. Unlike the PGTS-rich region, the last 25 amino acid residues of class II M and M-like proteins display complete identity (Fig. 10). This region is 76% identical to the carboxy-terminal 26 amino acid residues of M and M-like proteins from class I organisms.

(vi) Leader peptides. The sequence identity between the predicted leader peptides of ML2.1 and ML2.2 proteins (Fig. 3) is 80%. However, the leader sequence of ML2.2 protein is 100% identical to that predicted for *enn*X (16). Similarly, the identities between the leader peptides of the three subclass IIa molecules (ML2.1, M49, and Arp4) are 98 to 100% (12,

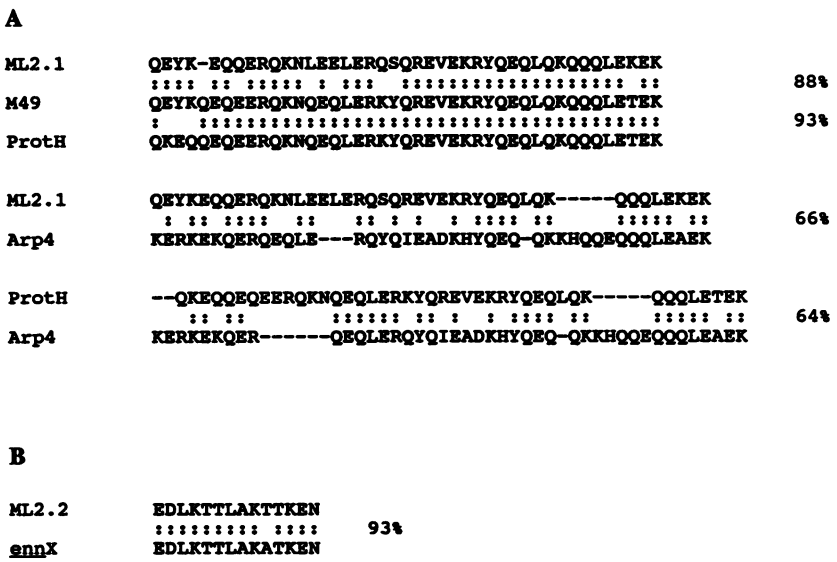

FIG. 7. Amino acid alignments of distinct regions or domains positioned on the amino-terminal side of the C repeat domain. Alignments of EQ-rich regions (A) are depicted for ML2.1 protein (amino acids [aa] 75 to 116; Fig. 3), M49 protein (bp 380 to 508) (16), Arp4 (aa 74 to 116) (12), and protein H (ProtH) (aa 97 to 159) (14). Note that the boundaries for the EQ-rich regions are not entirely clearcut, and additional EQ-rich sequences in M49 (eg., repeat block A.1) and Arp4 are excluded from this alignment. The alignment of sequences on the amino-terminal side of the C repeat region (B) in ML2.2 protein (aa 76 to 89; Fig. 3) and *enn*X (bp 1745 to 1785) (16) are shown. Matched amino acids (:) and percent identities are indicated. The sequence identity between ML2.1 and ProtH is 80%, and the sequence identity between M49 and Arp4 is 62% (alignments not shown).

16). Comparison of the leaders of these five class II M and M-like proteins to the five class I M proteins (serotypes 1, 5, 6, 12, and 24) indicates sequence identities ranging from 61 to 83% (17, 21, 29, 30, 34).

(vii) Noncoding region between M and M-like genes. The 211 bp that separate the *emm*L2.1 and *emm*L2.2 coding regions are almost completely identical to the noncoding regions immediately downstream from *emm*49 and *arp*4 (99 and 98% identity, respectively) (Fig. 11). Interestingly, one of the two differences between the type 2 and type 4 noncoding regions lies in a putative promoter site; a second difference between *emm*L2.2 and downstream *arp*4 sequences is in the proposed start codons (ATG and ACG, respectively). Sequence data strongly suggests that an M or M-like protein gene lies directly upstream from the ProtH gene (14); the 107-bp stretch immediately upstream from ProtH is 91% identical to nucleotides upstream from *emm*L2.2.

In summary, the sequence homology data suggest that there are two major subclasses of class II M or M-like proteins, represented by ML2.1 and ML2.2 proteins, respectively.

DISCUSSION

The objective of this study is to better understand the structure and expression of M and M-like genes and gene products of class II streptococci. Class I and II streptococci, which are structurally defined by antigenic epitopes within the C repeat domain of M and M-like proteins, parallel several other properties of group A streptococci. These properties include the potential to induce rheumatic fever among class I isolates and the nonimmune binding of human IgA and opacity factor production by class II organisms (2–4). In addition, IgG-binding activity distinguishes between nasopharyngeal and impetigo isolates within the class

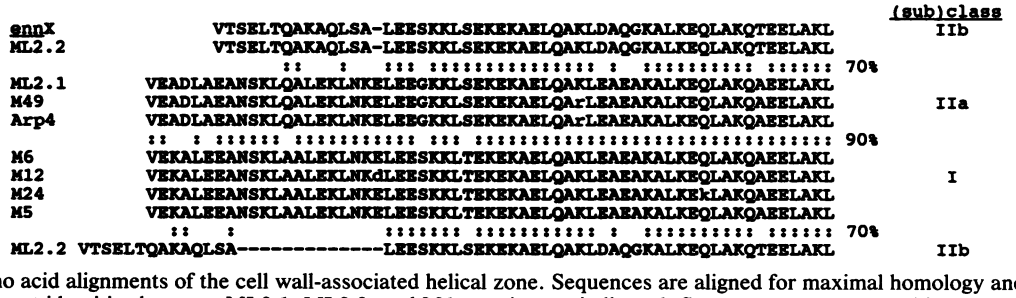

FIG. 8. Amino acid alignments of the cell wall-associated helical zone. Sequences are aligned for maximal homology and matched amino acids (:) and percent identities between ML2.1, ML2.2, and M6 proteins are indicated. Sequences are presented in three groupings: ML2.2 protein (*enn*X), ML2.1 protein (M49 and Arp4), and M6 protein (M5, M12, and M24). Amino acid differences from the consensus sequence within each group are indicated in lowercase type. Nucleotide and amino acid (aa) positions are as follows: ML2.1 (aa 245 to 305), ML2.2 (aa 204 to 257), M6 (aa 300 to 360) (21), *enn*X (bp 2131 to 2292) (16), M49 (bp 830 to 1012) (16), Arp4 (aa 224 to 284) (12), M5 (aa 351 to 411) (29), M24 (aa 398 to 458) (30), and M12 (bp 2582 to 2764) (34).

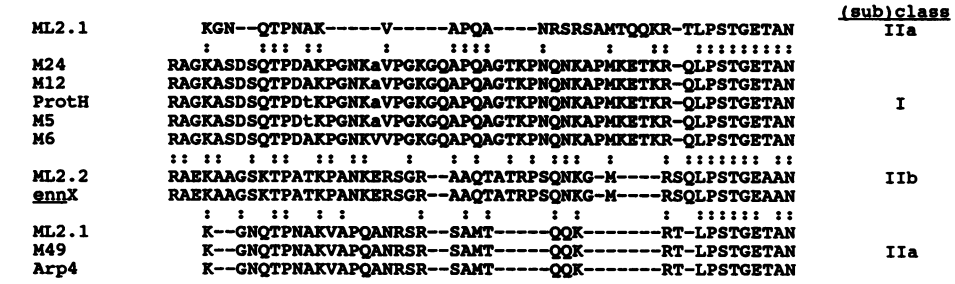

FIG. 9. Amino acid alignments of the PGTS-rich domain. Sequences are aligned for maximal homology and identities (:) between ML2.1, ML2.2, and M6 proteins are indicated. Sequences are presented in three groupings: ML2.2 protein (*ennX*), ML2.1 protein (M49 and Arp4), and M6 protein (M5, M12, M24, and ProtH). Amino acid differences from the consensus sequence within each group are indicated in lowercase type. Nucleotide and amino acid (aa) positions are as follows: ML2.1 (aa 306 to 341), ML2.2 (aa 258 to 306), M6 (aa 361 to 415) (21), *ennX* (bp 2293 to 2439) (16), M49 (bp 1013 to 1120) (16), Arp4 (aa 285 to 320) (12), M5 (aa 412 to 466) (29), M24 (aa 459 to 513) (30), M12 (bp 2765 to 2929) (34), and ProtH (aa 296 to 350) (14).

I grouping (3). In this report, by using a class II isolate, two adjacent genes displaying extensive sequence homology to other *emm* genes are characterized. While the ML2.1 and ML2.2 proteins have class II-like C repeat domains (2), they are significantly different from one another both in their PGTS-rich peptidoglycan-spanning domains and in the surface-exposed EQ-rich region positioned adjacent to the C repeat domain of ML2.1. On the basis of these structural differences, it is proposed that the *emm*L2.1 and *emm*L2.2 gene products represent two major subclasses of class II M and M-like proteins, subclasses IIa and IIb, respectively. IgA-binding activity can be attributed to a subclass IIb molecule in type 2 streptococci, whereas in type 4 streptococci (12) the IgA-BP is of subclass IIa structure. In the M protein-rich type 2 isolate under study, the RNA transcript level for the subclass IIa gene exceeds that of the subclass IIb gene by more than 30-fold.

The findings of this report raise several important questions regarding the function of M and M-like proteins. In an organism such as the M-rich T2/MR isolate, in which two M or M-like genes are expressed, what is the relative contribution of each M or M-like molecule to the organism's ability to resist nonimmune phagocytosis? Do both the *emm*L2.1 and *emm*L2.2 gene products contain type-specific immunodeterminants and furthermore, are the type-specific epitopes of ML2.1 and ML2.2 distinct from one another? Are the transcript levels significantly different for subclass IIa and IIb IgA-BP genes? Do subclass IIa and IIb IgA-BPs bind IgA via sites which are similar in structure, and are their binding affinities for IgA comparable?

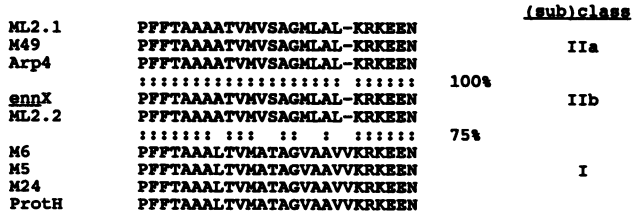

FIG. 10. Amino acid alignments of the hydrophobic domain and charged tail. Sequences are aligned for maximal homology and matched amino acids (:) and percent identities between ML2.1, ML2.2, and M6 proteins are indicated. Sequences are presented in three groupings: ML2.2 protein (*ennX*), ML2.1 protein (M49 and Arp4), and M6 protein (M5, M24, and ProtH). Amino acid differences from the consensus sequence within each group are indicated in lowercase type. Nucleotide and amino acid (aa) positions are as follows: ML2.1 (aa 342 to 366), ML2.2 (aa 307 to 331), M6 (aa 416 to 441) (21), *ennX* (bp 2440 to 2514) (16), M49 (bp 1121 to 1195) (16), Arp4 (aa 321 to 345) (12), M5 (aa 467 to 492) (29), M24 (aa 514 to 539) (30), and ProtH (aa 351 to 376) (14).

M proteins are viewed as alpha-helical coiled-coil fibrillar structures, with antigenically variable amino-termini that extend away from the cell surface (8). The amino-terminal regions contain type-specific epitopes, and only antibodies directed to type-specific epitopes are capable of overcoming the antiphagocytic effect of M proteins (1, 25). Secondary structure analyses (13) of the ML2.1 and ML2.2 molecules indicate that they are largely alpha-helical (data not shown). The seven-residue periodicity of non-polar amino acids characteristic of coiled-coil molecules (7) is evident throughout large portions of both ML2.1 and ML2.2 proteins, including their amino-terminal regions (data not shown). Thus, one would expect the ML2.1 and ML2.2 proteins to be somewhat similar in conformation to the coiled-coil dimeric fibers of M6 protein (11, 33). The data suggests that ML2.1 and ML2.2 proteins have distinct, hypervariable amino-terminal ends, since the predicted amino acid sequences of *emm*L2.1 and *emm*L2.2 are only weakly homologous to one another within this region. In addition, sequence homologies to the amino-terminal regions of other M and M-like molecules are low. The M2 typing sera is immunoreactive with the partial *emm*L2.1 product, but not with the *emm*L2.2 product. This may be a reflection of the immunodominance of the ML2.1 protein, perhaps as a result of its presence at higher quantities relative to ML2.2 protein or alternatively, a lack of type-specific epitopes within the ML2.2 molecule. The parent isolate expresses an antiphagocytic M protein, as evidenced by its survival in whole blood. Whether or not both ML2.1 and ML2.2 protein contribute to the organism's ability to resist opsonophagocytosis remains to be established.

The RNA transcript levels of the upstream gene, *emm*L2.1, exceed that of *emm*L2.2 by more than 30-fold. In the class I isolates D471 (type 6) and CS24 (type 12), transcription of the *emm*6 and *emm*12 genes is positively regulated by upstream regions termed *mry* (5, 32) and *virR* (6, 34, 39), respectively. Tn*916* insertional mutagenesis of *mry* leads to an approximately 50-fold decrease in both M6 protein and *emm*6 mRNA levels (5). In type 12 streptococci, M protein-deficient variants harboring deletions in *virR* have decreased levels of mRNA specific both to *emm*12 and to a downstream gene encoding for C5a peptidase; transcripts for each gene appear to be monocistronic (38, 39). The T2/MR

```
ML2.1      TAA GC-ATTAGACT GATGCTAAAG CTAAGAGAGA ATCAAATGAT TCTCTCTTTT
M49              T
Arp4             T

ML2.1           TGAGTGGCTA  AGTAACTAAC AATCTCAGTT AGACCAAAAA ATGGGAATGG
M49
Arp4

ML2.1           TTCAAAAAGC  TGGCCTTTAC TCCTTTTGAT TAACCATATA TAATAAAAAC
M49                                                G
Arp4                  T                                          C
ProtH           C T           A                                      T

ML2.1           ATTAGGAAAA  TAATAGTAAT ATTAAGTTTG TTTCCTCAAT AAAATCAAGG
M49
Arp4                                           A
ProtH                            C C       T   C    TT -

                            start
ML2.1           AGTAGATA    ATG
M49
Arp4                         C
ProtH
```

FIG. 11. Nucleotide sequence homologies in noncoding regions between M and M-like genes. Alignment of nucleotides beginning with the stop codons of the ML2.1, M49 (16), and Arp4 (12) protein genes and ending with the start codons or possible start codons of downstream genes (*emm*L2.2, *enn*X, and unknown gene from type 4, respectively). Included are 107 bp upstream from the ProtH gene start codon (14). Only nucleotides which differ from the type 2 sequence are indicated. Putative promoter and ribosomal binding sites are underlined (see Fig. 3).

class II isolate differs from the type 6 and 12 organisms in that it contains two adjacent *emm* or *emm*-like genes, rather than a single *emm* gene (37, 39). Similarly, the mRNA levels for both *emm*L2.1 and *emm*L2.2 from T2/MR are elevated relative to the M- and IgA-BP-deficient variant T2/MD. The differential in mRNA between T2/MR and T2/MD is in support of transcriptional control of *emm*L2.1 and *emm*L2.2 by elements analogous to *mry* or *virR*. However, it is still not known why the transcript level for the upstream gene, *emm*L2.1, exceeds that of *emm*L2.2 by more than 30-fold. Interestingly, the predicted *mry* gene product, which is structurally similar to receptor proteins of two-component regulatory systems, has two potential DNA-binding domains that perhaps regulate a different subset of genes (32).

There appears to be at least two structurally distinct types of IgA-BP in group A streptococci. *arp*4 is highly homologous to *emm*L2.1 and *emm*49 (12, 16) and is therefore considered to have subclass IIa structure. The noncoding sequences immediately downstream from these three subclass IIa genes are highly homologous as well. In contrast to the type 4 IgA-BP gene, the subclass IIb IgA-BP gene of the type 2 strain (*emm*L2.2) occupies the downstream position. IgA-binding activity is found among the majority of isolates of certain serotypes (types 4, 11, and 60), and among the minority of isolates of other serotypes (for example, types 2 and 22) (3, 36) (unpublished findings). To explain the close correlation of IgA-binding activity with certain serotypes, Lindahl et al. (28) propose that perhaps Arp4 is an M protein or alternatively that the genes for M protein and Ig-BPs are separate yet closely linked. In this report, the finding that the RNA transcript levels of two adjacent *emm*-like genes are different lead us to speculate that perhaps the type-specific molecule, as defined serologically (27), is simply that which is present in greatest abundance. It would follow that those IgA-BP genes which occupy the genomic position analogous to that of *emm*L2.1 would be transcribed at high levels and thereby contribute most significantly to the type specificity of the organism. Despite its structural similarity to Arp4, it is highly unlikely that ML2.1 binds IgA. The major lysin-extractable 41- and 43-kDa bands from T2/MR probably represent the *emm*L2.1 gene product on the basis of their immunoreactivity with both M2 typing sera (2) and anti-ColiM6. Yet these bands fail to bind IgA on Western blot, under conditions when the cloned *emm*L2.2 gene product binds IgA strongly.

The Ig-binding sites within M and M-like molecules have not been identified. The three M or M-like Ig-BP cloned and sequenced to date (Arp4, ProtH, and ML2.2) have in common class II-like C repeat regions (2) (data not shown). Whether subtleties within the class II C repeat influence Ig-binding capacity remains to be established. Interestingly, there is another group A streptococcal IgG-BP, FcRA76 (19), which is not considered to be M-like on the basis of its lack of an analogous C repeat region; FcRA76 displays only 30 to 36% overall homology with the three other Ig-BPs discussed (data not shown). Furthermore, the IgG-binding site of FcRA76 has been partially localized (18) to a region with even lower sequence homology to the other known group A streptococcal Ig-BPs, except for the cell wall-associated helical zone. Therefore, there is a relative lack of sequence homology between FcRA76 and other Ig-BPs of group A streptococci and between group A streptococcal Ig-BPs and those from other gram-positive bacteria (15, 40). Thus, it would not be unexpected for Arp4 and ML2.2 protein to bind IgA by structurally unrelated binding sites. Further studies are necessary in order to precisely identify the IgA-binding region of ML2.2 protein.

M and M-like proteins are composed of a series of discrete regions or domains, and some domains have variant forms. The C repeat region is an example of one such domain, in which the differences between variant forms (class I and II) are attributable to limited amino acid residues (2). On the basis of sequences currently available, there appears to be at least three major structural classes or subclasses, which we propose to designate class I (M5, M6, M12, and M24), subclass IIa (ML2.1, M49, and Arp4), and subclass IIb (ML2.2 and *enn*X). The main distinguishing factor between (sub)class I, IIa, and IIb molecules is the highly variant

PGTS-rich, peptidoglycan-spanning domain. In addition, the EQ-rich region is largely restricted to subclass IIa molecules. Differences in the hydrophobic domain fall precisely along class I and II lines. The ProtH molecule (derived from a class I isolate containing at least two M or M-like genes [14]), can be considered a hybrid between subclass IIa and class I molecules, with the transition point lying in the cell wall-associated helical zone; other hybrids may exist as well. Interestingly, of the 10 M and M-like molecules completely sequenced to date, all appear to have hypervariable amino termini, including those from organisms with two M or M-like molecules (types 2 and 49).

High-frequency, homologous intragenic recombination between repeat blocks has been demonstrated in type 6 organisms (22). It remains undetermined as to whether intergenic recombination is a common occurrence between two adjacent M or M-like genes, possibly generating a new mosaic of domains or hybrid molecules with unique functions. However, the non-type-specific portions of both the ML2.1 and ML2.2 proteins are more homologous to proteins from streptococci of different serotypes than they are to each other. This observation suggests that major portions of subclass IIa and IIb genes diverged from one another prior to the emergence of a wide array of distinct serotypes.

ACKNOWLEDGMENTS

We are grateful for the technical advice of Emil Gotschlich, Susan Hollingshead, Jeff Weiser, Tony Butler, and John Robbins. In addition, we thank Joseph Ferretti for pNC1 and Frank Butera and Chris Dickson for excellent technical assistance.

This work was supported by Public Health Service grant AI-28944 and a Grant-in-Aid from the American Heart Association (N.Y.C. affiliate) to D.E.B.

REFERENCES

1. **Beachey, E. H., J. M. Seyer, J. B. Dale, W. A. Simpson, and A. H. Kang.** 1981. Type-specific protective immunity evoked by synthetic peptide of *Streptococcus pyogenes* M protein. Nature (London) **292**:457–459.
2. **Bessen, D., and V. A. Fischetti.** 1990. Differentiation between two biologically distinct classes of group A streptococci by limited substitutions of amino acids within the shared region of M protein-like molecules. J. Exp. Med. **172**:1757–1764.
3. **Bessen, D., and V. A. Fischetti.** 1990. A human IgG receptor of group A streptococci is associated with tissue site of infection and streptococcal class. J. Infect. Dis. **161**:747–754.
4. **Bessen, D., K. F. Jones, and V. A. Fischetti.** 1989. Evidence for two distinct classes of streptococcal M protein and their relationship to rheumatic fever. J. Exp. Med. **169**:269–283.
5. **Caparon, M. G., and J. R. Scott.** 1987. Identification of a gene that regulates expression of M protein, the major virulence determinant of group A streptococci. Proc. Natl. Acad. Sci. USA **84**:8677–8681.
6. **Cleary, P. P., D. LaPenta, D. Heath, E. J. Haanes, and C. Chen.** 1991. A virulence regulon in *Streptococcus pyogenes*, p. 147–151. *In* G. M. Dunny, P. P. Cleary, and L. L. McKay (ed.), Genetics and molecular biology of streptococci, lactococci, and enterococci. American Society for Microbiology, Washington, D.C.
7. **Cohen, C., and D. A. D. Parry.** 1990. Alpha-helical coiled coils and bundles: how to design an alpha-helical protein. Proteins Struct. Funct. Genet. **7**:1–15.
8. **Fischetti, V. A.** 1989. Streptococcal M protein: molecular design and biological behavior. Clin. Microbiol. Rev. **2**:285–314.
9. **Fischetti, V. A., K. F. Jones, B. N. Manjula, and J. R. Scott.** 1984. Streptococcal M6 protein expressed in *Escherichia coli*. Localization, purification and comparison with streptococcal-derived M protein. J. Exp. Med. **159**:1083–1095.
10. **Fischetti, V. A., K. F. Jones, and J. R. Scott.** 1985. Size variation of the M protein in group A streptococci. J. Exp. Med. **161**:1384–1401.
11. **Fischetti, V. A., D. A. D. Parry, B. L. Trus, S. K. Hollingshead, J. R. Scott, and B. N. Manjula.** 1988. Conformational characteristics of the complete sequence of group A streptococcal M6 protein. Proteins Struct. Funct. Genet. **3**:60–69.
12. **Frithz, E., L.-O. Heden, and G. Lindahl.** 1989. Extensive sequence homology between IgA receptor and M protein in *Streptococcus pyogenes*. Mol. Microbiol. **3**:1111–1119.
13. **Garnier, J., D. J. Osguthorpe, and B. Robson.** 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120**:97–120.
14. **Gomi, H., T. Hozumi, S. Hattori, C. Tagawa, F. Kishimoto, and L. Bjorck.** 1990. The gene sequence and some properties of protein H. J. Immunol. **144**:4046–4052.
15. **Guss, B., M. Eliasson, A. Olsson, M. Uhlen, A.-K. Frej, H. Jornvall, J.-I. Flock, and M. Lindberg.** 1986. Structure of the IgG-binding regions of streptococcal protein G. EMBO J. **5**:1567–1575.
16. **Haanes, E. J., and P. P. Cleary.** 1989. Identification of a divergent M protein gene and an M protein-related gene family in *Streptococcus pyogenes* serotype 49. J. Bacteriol. **171**:6397–6408.
17. **Haanes-Fritz, E., W. Kraus, V. Burdett, J. B. Dale, and E. H. Beachey.** 1988. Comparison of the leader sequences of four group A streptococcal M protein genes. Nucleic Acids Res. **16**:4667–4677.
18. **Heath, D. G., M. D. P. Boyle, and P. P. Cleary.** 1990. Isolated DNA repeat region from fcrA76, the Fc-binding protein gene from an M-type 76 strain of group A streptococci, encodes a protein with Fc-binding activity. Mol. Microbiol. **4**:2071–2079.
19. **Heath, D. G., and P. P. Cleary.** 1989. Fc-receptor and M protein genes of group A streptococci are products of gene duplication. Proc. Natl. Acad. Sci. USA **86**:4741–4745.
20. **Hollingshead, S. K.** 1987. Nucleotide sequences that signal the initiation of transcription for the gene encoding type 6 M protein in *Streptococcus pyogenes*, p. 98–100. *In* J. J. Ferretti and R. Curtiss (ed.), Streptococcal genetics. American Society for Microbiology, Washington, D.C.
21. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1986. Complete nucleotide sequence of type 6 M protein of the group A streptococcus: repetitive structure and membrane anchor. J. Biol. Chem. **261**:1677–1686.
22. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1987. Size variation in group A streptococcal M protein is generated by homologous recombination between intragenic repeats. Mol. Gen. Genet. **207**:196–203.
23. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1987. A highly conserved region present in transcripts encoding heterologous M proteins of group A streptococci. Infect. Immun. **55**:3237–3239.
24. **Huang, T.-T., H. Malke, and J. J. Ferretti.** 1989. Heterogeneity of the streptokinase gene in group A streptococci. Infect. Immun. **57**:502–506.
25. **Jones, K. F., and V. A. Fischetti.** 1988. The importance of the location of antibody binding on the M6 protein for opsonization and phagocytosis of group A M6 streptococci. J. Exp. Med. **167**:1114–1123.
26. **Jones, K. F., S. A. Khan, B. W. Erickson, S. K. Hollingshead, J. R. Scott, and V. A. Fischetti.** 1986. Immunochemical localization and amino acid sequence of cross-reactive epitopes within the group A streptococcal M6 protein. J. Exp. Med. **164**:1226–1238.
27. **Lancefield, R. C.** 1962. Current knowledge of the type specific M antigens of group A streptococci. J. Immunol. **89**:307–313.
28. **Lindahl, G., B. Åkerström, L. Stenberg, E. Frithz, and L.-O. Hedén.** 1991. Genetics and biochemistry of protein Arp, an immunoglobulin A receptor from group A streptococci, p. 155–159. *In* G. M. Dunny, P. P. Cleary, and L. L. McKay (ed.), Genetics and molecular biology of streptococci, lactococci, and enterococci. American Society for Microbiology, Washington, D.C.

29. **Miller, L., L. Gray, E. H. Beachey, and M. A. Kehoe.** 1988. Antigenic variation among group A streptococcal M proteins: nucleotide sequence of the serotype 5 M protein gene and its relationship with genes encoding types 6 and 24 M proteins. J. Biol. Chem. **263**:5668–5673.
30. **Mouw, A. R., E. H. Beachey, and V. Burdett.** 1988. Molecular evolution of streptococcal M protein: cloning and nucleotide sequence of type 24 M protein gene and relation to other genes of *Streptococcus pyogenes*. J. Bacteriol. **170**:676–684.
31. **Pancholi, V., and V. A. Fischetti.** 1988. Isolation and characterization of the cell-associated region of group A streptococcal M6 protein. J. Bacteriol. **170**:2618–2624.
32. **Perez-Casal, J., M. G. Caparon, and J. R. Scott.** 1991. Mry, a *trans*-acting positive regulator of the M protein gene of *Streptococcus pyogenes* with similarity to the receptor proteins of two-component regulatory systems. J. Bacteriol. **173**:2617–2624.
33. **Phillips, G. N., P. F. Flicker, C. Cohen, B. N. Manjula, and V. A. Fischetti.** 1981. Streptococcal M protein: alpha-helical coiled-coil structure and arrangement on the cell surface. Proc. Natl. Acad. Sci. USA **78**:4689–4693.
34. **Robbins, J. C., J. G. Spanier, S. J. Jones, W. J. Simpson, and P. P. Cleary.** 1987. *Streptococcus pyogenes* type 12 M protein regulation by upstream sequences. J. Bacteriol. **169**:5633–5640.
35. **Saravani, G. A., and D. R. Martin.** 1990. Opacity factor from group A streptococci is an apoproteinase. FEMS Microbiol. Lett. **68**:35–40.
36. **Schalen, C.** 1980. The group A streptococcal receptor for human IgA binds via the Fc-fragment. Acta Pathol. Microbiol. Scand. Sect. **88C**:271–274.
37. **Scott, J. R., W. M. Pulliam, S. K. Hollingshead, and V. A. Fischetti.** 1985. Relationship of M protein genes in group A streptococci. Proc. Natl. Acad. Sci. USA **82**:1822–1826.
38. **Simpson, W. J., and P. P. Cleary.** 1987. Expression of M type 12 protein by a group A streptococcus exhibits phaselike variation: evidence for coregulation of colony opacity determinants and M protein. Infect. Immun. **55**:2448–2455.
39. **Simpson, W. J., D. LaPenta, C. Chen, and P. P. Cleary.** 1990. Coregulation of type 12 M protein and streptococcal C5a peptidase genes in group A streptococci: evidence for a virulence regulon controlled by the *virR* locus. J. Bacteriol. **172**: 696–700.
40. **Sjodahl, J.** 1977. Repetitive sequences in protein A from *Staphylococcus aureus*—arrangement of five regions within the protein, four being highly homologous and Fc-binding. Eur. J. Biochem. **73**:343–351.