JOURNAL OF BACTERIOLOGY, Feb. 1988, p. 676–684
0021-9193/88/020676-09$02.00/0
Copyright © 1988, American Society for Microbiology

Vol. 170, No. 2

Molecular Evolution of Streptococcal M Protein: Cloning and Nucleotide Sequence of the Type 24 M Protein Gene and Relation to Other Genes of *Streptococcus pyogenes*

ANDREA R. MOUW,[1] EDWIN H. BEACHEY,[2] AND VICKERS BURDETT[1]*

Department of Microbiology and Immunology, Duke University Medical Center, Durham, North Carolina 27710,[1] and Veterans Administration Medical Center, University of Tennessee, Memphis, Tennessee 38104[2]

Received 16 September 1987/Accepted 9 November 1987

The structural gene for the type 24 M protein of group A streptococci has been cloned and expressed in *Escherichia coli*. The complete nucleotide sequence of the gene and the 3′ and 5′ flanking regions was determined. The sequence includes an open reading frame of 1,617 base pairs encoding a pre-M24 protein of 539 amino acids and a predicted M_r of 58,738. The structural gene contains two distinct tandemly reiterated elements. The first repeated element consists of 5.3 units, and the second contains 2.7 units. Each element shows little variation of the basic 35-amino-acid unit. Comparison of the sequence of the M24 protein with the sequence of the M6 protein (S. K. Hollingshead, V. A. Fischetti, and J. R. Scott, J. Biol. Chem. 261:1677–1686, 1986) indicates that these molecules have are conserved except in the regions coding for the antigenic (type specific) determinant and they have three regions of homology within the structural genes: 38 of 42 amino acids within the amino terminal signal sequence, the second repeated element of the M24 protein is found in the M6 molecule at the same position in the protein, and the carboxy terminal 164 amino acids, including a membrane anchor sequence, are conserved in both proteins. In addition, the sequences flanking the two genes are strongly conserved.

The streptococcal M protein forms a fibrillar structure on the surface of group A streptococci, rendering the organism resistant to phagocytosis in the nonimmune host. Although protective immunity to group A streptococci is conferred by host antibodies directed against the surface protein, this protection is not absolute and reinfection can occur if the infecting strain elaborates a different antigenic serotype of M protein. Over 70 distinct serotypes of M protein have been identified, but in general only one serotype is expressed by each isolate (17).

Attempts to use M protein for development of an antistreptococcal vaccine have been complicated by the extensive variability of the M antigen and by the inability to demonstrate extensive cross-reactivity between the different serotypes. Furthermore, antibodies to some serotypes of M protein can be shown to cross-react with human heart tissue (9, 11, 12,); this effect may play a role in the development of rheumatic heart disease. One of the objectives of cloning and sequencing of M protein genes has been to identify regions of the M protein molecule shared among various serotypes which might have potential for vaccine production. Hybridization experiments have shown that sequences within the carboxy-terminal region of the gene are conserved among the different M protein serotypes (37). However, the exact extent of the homology and the arrangement of sequences within this region cannot be determined, except by knowing the DNA sequences of the molecules being compared. The carboxy-terminal region has recently been shown to contain antigenic determinants shared among a number of M protein serotypes, although the precise sequence of the determinant was not determined (22).

Despite the important role of M protein in streptococcal pathogenesis, detailed analyses of the structure of this class of protein have been hampered by the inability to isolate the intact molecule from the cell surface. Limited amounts of M protein fragments have been obtained after proteolysis of the cell surface, and partial amino acid sequence analysis of such material from three M protein species demonstrated a common periodicity in the placement of hydrophobic residues that are responsible for maintaining the alpha-helical character of the protein (32). Although primary sequences of these proteins are distinct, each contains tandemly repeated amino acid sequences. M protein from serotype 24 streptococci is particularly striking in this respect. This molecule, which is the subject of this paper, contains five near-perfect repeats of a 35-amino-acid sequence which is thought to define an alpha-helical domain (1). In this report we describe the isolation of the gene encoding type 24 M protein and determination of its DNA sequence, as well as the arrangement of sequences common to type 24 and type 6 M protein.

MATERIALS AND METHODS

Bacterial strains and plasmids. *Escherichia coli* DH1 (λ *c*I857 *S7*) (18) and M5248 (39) were used as plasmid hosts, and JM109 (29) was host for M13 phages. The original cloning was performed with pSCC31 (5), and fragments were subcloned into M13mp8-M13mp9 (29) or M13mp18-M13mp19 (47) for sequence analysis. Plasmid pSCC31 is a positive selection cloning vector containing the structural gene for *Eco*RI downstream from lambda p_L and is unable to transform *E. coli* unless the *Eco*RI endonuclease gene is inactivated (5), for example, by cloning into the *Bgl*II site within the endonuclease gene. The M24 protein gene was cloned from *Streptococcus pyogenes* A24 strain Vaughn, expressing M protein serotype 24. Streptococci were grown in brain heart infusion broth (Difco Laboratories, Detroit, Mich.), and *E. coli* strains were grown in LB broth (30). Solid media contained 1.4% Bacto-Agar (Difco).

DNA sequencing. DNA sequences were determined by the

* Corresponding author.

chain termination method of Sanger et al. (36) with [α-^{35}S]dATP (3). Because of the difficulties of sequence determination across repetitive regions, sequential sets of nested deletions were constructed for sequence analysis by the procedure of Dale et al. (13). M13 template DNAs were prepared essentially as described by Messing (29). The Klenow fragment of DNA polymerase and reagents for sequencing reactions were from United States Biochemical Corp., Cleveland, Ohio, and New England BioLabs, Inc., Beverly, Mass., respectively. Synthetic oligonucleotides were synthesized on a Biosearch instrument (Biosearch, Inc., San Rafael, Calif.).

Sequence analysis. The software packages developed by David Mount, University of Arizona, and commercially available software packages were used in the analysis of the DNA and deduced amino acid sequences. Nucleotide and amino acid sequences were aligned by using the NUCALN and PRTALN algorithms of Wilbur and Lipman (46).

Protein analysis. Strains to be analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (24) were grown at 30°C to 100 Klett Units. After 2 h at 42°C, 1-ml samples were centrifuged for 1 min, washed with 10 mM Tris hydrochloride (pH 8)–1 mM EDTA, suspended in 100 μl of sample buffer (18), and boiled for 5 min. Portions (10 to 15 μl) were run on an SDS–12.5% polyacrylamide gel. Western blots were performed (42) by using anti-pep M24 and anti-*Eco*RI antibodies and staphylococcal ^{125}I-labeled protein A (New England Nuclear Corp., Boston, Mass.). Molecular weight markers included phosphorylase (M_r 92,500), bovine serum albumin (M_r 66,200), ovalbumin (M_r 45,000), carbonic anhydrase (M_r 31,000), soybean trypsin inhibitor (M_r 21,500), and lysozyme (M_r 14,400).

Antibody preparation. Immune rabbit serum (anti-pep M24) was obtained from New Zealand White rabbits immunized with doses of M protein (pep M24) emulsified in complete Freund adjuvant as previously described (10). Rabbit antiserum to the *Eco*RI endonuclease was a gift of Paul Modrich, Department of Biochemistry, Duke University Medical Center.

RESULTS

Cloning of the *S. pyogenes* type 24 M protein determinant in *E. coli*. A library of *S. pyogenes* M protein serotype 24 was constructed in *E. coli* by using the positive selection vector pSCC31. Total cell DNA isolated from serotype 24 *S. pyogenes* Vaughn was sheared by sonication and ligated into the *Bgl*II site of pSCC31 following the addition of *Bgl*II linkers. Since the *Bgl*II site of pSCC31 is in the *Eco*RI gene, the only transformants obtained after ampicillin selection are the pSCC31 derivatives which contain insertions in this gene (5).

Ampicillin-resistant colonies recovered after transformation at 30°C were screened, after 4 h of further incubation at 42°C for the presence of M24 protein cross-reacting material by radioimmunoassay (42). Antiserum against a pepsin fragment of the M24 protein (pep M24) was used to identify expressing colonies. None of the positive colonies identified in this manner cross-reacted with preimmune serum or with antiserum raised against type 5 M protein. One isolate which reacted particularly strongly with the antiserum against M24 protein was chosen for further study. The recombinant plasmid in this isolate, pVB-L3, contained an insert of 2,425 base pairs.

A second library was constructed in the same vector by ligation of fragments generated by partial digestion of the chromosomal DNA with *Sau*3AI (*Mbo*I) into the *Bgl*II site of the vector by virtue of the common sticky ends. Several isolates were identified which cross-reacted with only the anti-pep M24 serum. One of these isolates, containing plasmid pVB41, reacted at 30°C as well as at 42°C, suggesting that it contained the functional streptococcal promoter for the M24 protein gene. The insert in pVB41 was about 1,000 base pairs, and the restriction map was found to overlap that of pVB-L3 (data not shown). The contiguous M24 protein gene was then reconstructed from pVB41 and pVB-L3 by taking advantage of the unique *Nru*I restriction endonuclease cleavage site contained within the overlapping region. The plasmid containing the contiguous M24 protein gene was designated pVB41-L3 and contained an insert of 2,900 base pairs.

For purposes of expression, pVB-L3, pVB41, and pVB41-L3 were transformed into strain M5248 and the p_L promoter on the plasmid was induced by temperature shift to 42°C. Total *E. coli* polypeptides were separated by electrophoresis on a sodium dodecyl sulfate–12.5% polyacrylamide gel and transferred to nitrocellulose, and the two halves of the blot were reacted with rabbit anti-pep M24 (Fig. 1A) or rabbit anti-*Eco*RI endonuclease (Fig. 1B). M5248(pVB41) (lanes 1) expresses a protein of M_r 22,000 which reacts with anti-pep M24 (a larger polypeptide can be seen which reacts with anti-*Eco*RI antibody [this and other experiments]). This M24 cross-reactive polypeptide is also detectable at 30°C in lower yield (data not shown). M5248(pVB-L3) (lanes 2) expresses a protein of M_r 75,000 which reacts with both antibodies, indicating that the pVB-L3 protein is a fusion product of the amino terminus of the *Eco*RI endonuclease and the M protein sequence. Because the sequence of the *Eco*RI endonuclease gene is known (31), 58,000 daltons of the fusion

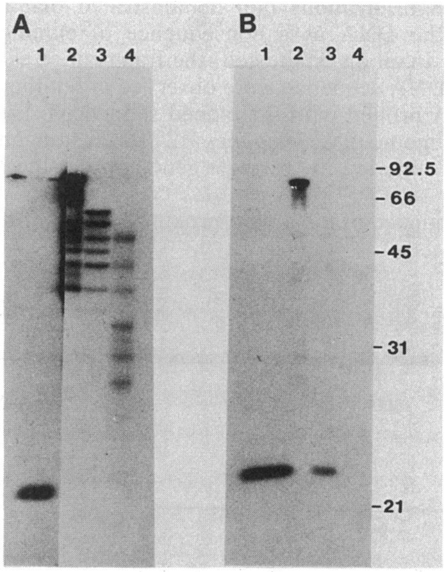

FIG. 1. Immunoblot analysis of proteins produced by *E. coli* strains carrying cloned M24 protein sequences. The proteins in whole-cell lysates of *E. coli* M5248 containing the indicated plasmids were analyzed as described in Materials and Methods. Lanes: 1, pVB41 containing the promoter for the M24 protein gene; 2, pVB-L3 encoding the fused polypeptides; 3, pVB41-L3 containing the reconstructed contiguous M24 protein gene; 4, hot-acid extract of *S. pyogenes* serotype 24. The proteins bound to the membrane were reacted with antiserum to pep M24 (A) or *Eco*RI (B), and bound antibody was detected by using ^{125}I-protein A followed by autoradiography.

polypeptide is the product of the cloned fragment. M protein extracted from *S. pyogenes* A24 strain Vaughn with hot HCl reacts strongly only with anti-pep M24 and not with anti-*Eco*RI (lanes 4). The M protein portion of the pVB-L3 polypeptide (lanes 2) is larger than the M protein species (M_r 50,000) isolated by hot-acid extraction, as well as the M protein isolated after pepsin digestion of whole streptococci (M_r 38,000). Plasmid pVB-L3 should therefore contain a substantial portion of the coding region for type 24 M protein, and pVB41 should encode the amino-terminal portion of this protein. Cells containing the contiguous M24 protein gene encoded by plasmid pVB41-L3 (lane 3) express a protein of M_r 58,000 which reacts with the antiserum against M24 protein (Fig. 1A, lane 3); a peptide which reacts with anti-endonuclease is a truncated *Eco*RI endonuclease peptide. Material which cross-reacts with anti-pep M24 is not detectable in cells in which the vector plasmid contains a random insert but which do produce the expected truncated endonuclease peptide (data not shown).

Sequence analysis of the type 24 M protein gene. Since the M protein sequences in pVB-L3 are expressed as an *Eco*RI fusion protein under the control of the lambda p_L promoter, the direction of transcription and reading frame is known. The sequence of the M24 protein-coding region was determined in two stages. First, the 2,425-base-pair insert of pVB-L3 was transferred into M13mp8 in both orientations, and defined fragments of the insert were subcloned. Preliminary sequence analysis of such fragments proved not to be straightforward owing to the repeated nature of the M24 sequence. Therefore, we adopted the approach of Dale et al. (13) and constructed a series of nested deletions which extended into the cloned fragment from either end. In this manner, we sequenced both strands in their entirety by using the sequencing strategy shown in Fig. 2. Moreover, Southern blot hybridizations (40) demonstrated that rearrangement of the DNA as a consequence of cloning had not occurred. In all digests tested, the fragment sizes predicted from the DNA sequence were observed in Southern blots of M24 DNA probed with the cloned sequences. In this manner, the genome of *S. pyogenes* A24 strain Vaughn was also shown to contain only a single gene copy of the M protein gene.

The sequence of the amino-proximal region of the protein encoded on pVB41 was obtained by cloning a fragment containing the streptococcal insert into M13mp18 in both orientations. In this way, the sequence of the relevant region of the insert could be obtained by using two synthetic oligonucleotides complementary to previously sequenced regions as primers for the sequencing reactions. To confirm that the M24 gene sequence encoded by pVB41-L3 was unaltered, we determined the sequence across the restriction endonuclease site used to join pVB41 and pVB-L3.

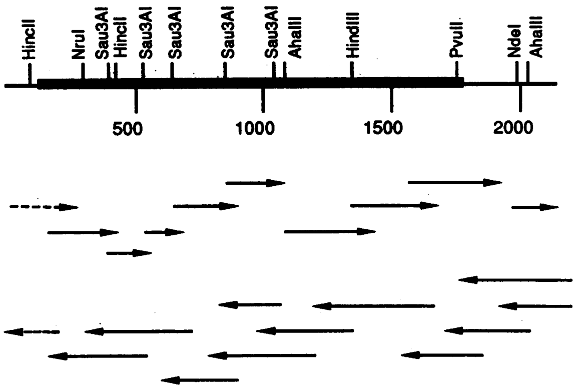

FIG. 2. Restriction maps of the insert of clone pVB41-L3. The coding region for the M24 protein is represented by a heavy line. The strategy for nucleotide sequencing is shown by the arrows (—►) below the restriction map, which indicate the extent and direction of each analysis. The dashed arrows (--►) are sequences obtained by using synthetic primers complementary to previously determined sequences.

The first ATG within the major open reading frame (Fig. 3, nucleotide 157) is presumed to represent the best translation initiation site. First, this first ATG begins an open reading frame of 1,617 nucleotides ending with a TAA codon which encodes a polypeptide (M_r 58,738) which can be aligned with peptides derived from pep M24 (see below). Second, this assignment is the most consistent with the size of the protein; the second ATG (nucleotide 556) would yield a protein which would be too small. Seven nucleotides upstream of the ATG is a putative ribosome-binding sequence (AAGGAG) which is complementary (6 of 11 bases) to the 3′ OH terminus of the 16S rRNA of *Bacillus subtilis* and *B. stearothermophilis* (3′-UCUUUCCUCCA-). The interval of seven spaces between the ribosome-binding site and the first ATG is consistent with the spacing for ribosome-binding sites identified in front of several other streptococcal genes (4, 25, 28, 38, 45). These unrelated genes show a spacing of 4 to 7 nucleotides. The open reading frame for the M24 protein gene is followed by an inverted repeat with the potential of forming a stable hairpin loop (underlined in Fig. 3) which could function in transcription termination (33).

Deduced sequence and organization of the M24 protein. We showed that the amino acid sequence encoded by the DNA sequence in Fig. 3 was the M24 protein-coding sequence by matching it with the previously determined sequences of several cyanogen bromide peptides (CB1, CB3, CB6, and CB7) of pep M24, the M protein polypeptide obtained from cells after treatment with pepsin (1, 2). The amino-terminal portion of the translation product following Val-43 is like pep M24 in that it begins with the same sequence as pep M24 and CB1.

The type 24 M protein isolated after pepsin hydrolysis of streptococcal cells (pep M24) is reported to contain five repeats of a highly conserved 35-amino-acid sequence (2), although the arrangement of these repeats was not previously determined. The DNA sequence presented here has allowed us to localize these repeats within the protein. Computer-assisted homology analysis (34) of the protein sequence indicates the position of the internal homologies. The first repeat region (A1 to A6.1) encodes a 35-amino-acid peptide tandemly reiterated 5.3 times in succession. Both the DNA and the deduced amino acid sequences observed in this region are aligned in Fig. 4. The DNA encoding this entire region has 105 nucleotides per repeat and contains only two changes per repeat. Only a single amino acid is altered in the A2 and A4 repeats, whereas two amino acids are altered in the A3 and A5 repeats. This first block of repeats (amino acids 118 to 301) overlaps repeats of the previously characterized cyanogen bromide peptides CB7 and CB6 in the order CB7-CB6-CB7-CB6 followed immediately by 27 of the 35 amino acids of CB3 (amino acid 134 to 301). It can be noted that the repeat unit A2 plus A3 is identical to the repeat unit of A4 plus A5 with no alterations in either the DNA or amino acid sequence.

The second class of repeats (B1 to B2.7) consists of a distinct 35-amino-acid segment which is reiterated 2.7 times. Close examination of the homology analyses suggests that

```
                                                                                            -35·P1
        ATT CAT CAT TAA TAG CAT TTA GGT CAA AAA GCT GGC AAA AGC TAA AAA AAC TGG TCT TTA CCT TTT GGC TTT TAT TAT TTA CAA    84
                                                                                    -35·P2                         -10·P2

        -10·P1                                                                          RBS
85      TAG AAT TAT TAG AGT TAA CCC CTG AAA ATG AGG GGT TTT TCC TAA AAA AAT GAT AAC ATA AGG AGC ATA AAA ATG ACT AAA AAC    168
                                                                                                        MET Thr Lys Asn

169     AAC ACG AAT AGA CAC TAT TCG CTT AGA AAA TTA AAA ACG GGA ACG GCT TCA GTA GCG GTA GCT TTG ACA GTT TTA GGG GCA GGA    252
        Asn Thr Asn Arg His Tyr Ser Leu Arg Lys Leu Lys Thr Gly Thr Ala Ser Val Ala Val Ala Leu Thr Val Leu Gly Ala Gly
                            10                                      20                                      30

253     TTA GTT GTC AAT ACT AAT GAA GTT AGT GCA GTC GCG ACT AGG TCT CAG ACA GAT ACT CTG GAA AAA GTA CAA GAA CGT GCT GAC    336
        Leu Val Val Asn Thr Asn Glu Val Ser Ala Val Ala Thr Arg Ser Gln Thr Asp Thr Leu Glu Lys Val Gln Glu Arg Ala Asp
                                    40                                      50                                      60

337     AAG TTT GAG ATA GAA AAC AAT ACG TTA AAA CTT AAG AAT AGT GAC TTA AGT TTT AAT AAT AAA GCG TTA AAA GAT CAT AAT GAT    420
        Lys Phe Glu Ile Glu Asn Asn Thr Leu Lys Leu Lys Asn Ser Asp Leu Ser Phe Asn Asn Lys Ala Leu Lys Asp His Asn Asp
                                            70                                      80

421     GAG TTA ACT GAA GAG TTG AGT AAT GCT AAA GAG AAA CTA CGT AAA AAT GAT AAA TCA CTA TCT GAA AAA GCT AGT AAA ATT CAA    504
        Glu Leu Thr Glu Glu Leu Ser Asn Ala Lys Glu Lys Leu Arg Lys Asn Asp Lys Ser Leu Ser Glu Lys Ala Ser Lys Ile Gln
            90                                     100                                     110

            A1
505     GAA(TTA GAG GCA CGT AAG GCT GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA    588
        Glu(Leu Glu Ala Arg Lys Ala Asp Leu Glu Lys Ala Leu Glu Gly Ala Met Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys
                    120                                     130                                     140

                                        A2
589     ACC TTA GAA GCA GAG AAA GCT GCT)(TTA GCG GCA CGT AAG GCT GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAC TTT TCA ACA    672
        Thr Leu Glu Ala Glu Lys Ala Ala)(Leu Ala Ala Arg Lys Ala Asp Leu Glu Lys Ala Leu Glu Gly Ala Met Asn Phe Ser Thr
                            150                                     160                                     160

                                                                    A3
673     GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT)(TTA GAG GCA CGC CAG GCT GAA CTT GAA AAA GCA TTA GAA    756
        Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala)(Leu Glu Ala Arg Gln Ala Glu Leu Glu Lys Ala Leu Glu
                                    180                                     190                                     200

                                                                                                A4
757     GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT)(TTA GCG GCA CGT AAG GCT    840
        Gly Ala Met Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala)(Leu Ala Ala Arg Lys Ala
                                            210                                     220

841     GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAC TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT    924
        Asp Leu Glu Lys Ala Leu Glu Gly Ala Met Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala
            230                                     240                                     250

            A5
925     GCT)(TTA GAG GCA CGC CAG GCT GAA CTT GAA AAA GCA TTA GAA GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA    1008
        Ala)(Leu Glu Ala Arg Gln Ala Glu Leu Glu Lys Ala Leu Glu Gly Ala Met Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys
                    260                                     270                                     280

                                        A6.1                                                                B1
1009    ACC TTA GAA GCA GAG AAA GCT GCT)(TTG GAG GCA GAG AAA GCT GAT CTT GAA)CAT CAA AGT CAA GTT TTA AAT GCT AAT(CGT CAA    1092
        Thr Leu Glu Ala Glu Lys Ala Ala)(Leu Glu Ala Glu Lys Ala Asp Leu Glu)His Gln Ser Gln Val Leu Asn Ala Asn(Arg Gln
                            290                                     300                                     310

1093    AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA CAA TTA GAA GCT GAA CAC CAA AAA CTA GAA GAA CAA AAC AAG    1176
        Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys Gln Leu Glu Ala Glu His Gln Lys Leu Glu Glu Gln Asn Lys
                                    320                                     330                                     340

                            B2
1177    ATT TCA GAA GCA AGC)(CGT CAA AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA CAA TTA GAA GCT GAA CAC CAA    1260
        Ile Ser Glu Ala Ser)(Arg Gln Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys Gln Leu Glu Ala Glu His Gln
                                            350                                     360

                                                        B2.7
1261    AAA CTA GAA GAA CAA AAC AAG ATT TCA GAA GCA AGC)(CGT CAA AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA    1344
        Lys Leu Glu Glu Gln Asn Lys Ile Ser Glu Ala Ser)(Arg Gln Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys
            370                                     380                                     390

                A6.2
1345    CAA GTT(GAA)AAA GCT TTA GAA GAA GCA)AAC AGC AAA TTA GCT GCT CTT GAA AAA CTT AAC AAA GAG CTT GAA GAA AGC AAG AAA    1428
        Gln Val(Glu)Lys Ala Leu Glu Glu Ala)Asn Ser Lys Leu Ala Ala Leu Glu Lys Leu Asn Lys Glu Leu Glu Glu Ser Lys Lys
                    400                                     410                                     420

1429    TTA ACC GAA AAA GAA AAA GCT GAG CTA CAA GCA AAA CTT GAA GCA GAA GCA AAA GCA CTC AAA GAA AAA TTA GCG AAA CAA GCT    1512
        Leu Thr Glu Lys Glu Lys Ala Glu Leu Gln Ala Lys Leu Glu Ala Glu Ala Lys Ala Leu Lys Glu Lys Leu Ala Lys Gln Ala
                            430                                     440                                     450

1513    GAA GAA CTT GCA AAA CTA AGA GCT GGA AAA GCA TCA GAC TCA CAA ACC CCT GAT GCA AAA CCA GGA AAC AAA GCT GTT CCA GGT    1596
        Glu Glu Leu Ala Lys Leu Arg Ala Gly Lys Ala Ser Asp Ser Gln Thr Pro Asp Ala Lys Pro Gly Asn Lys Ala Val Pro Gly
                                    460                                     470                                     480

1597    AAA GGT CAA GCA CCA CAA GCA GGT ACA AAA CCT AAC CAA AAC AAA GCA CCA ATG AAG GAA ACT AAG AGA CAG TTA CCA TCA ACA    1680
        Lys Gly Gln Ala Pro Gln Ala Gly Thr Lys Pro Asn Gln Asn Lys Ala Pro Met Lys Glu Thr Lys Arg Gln Leu Pro Ser Thr
                                            490                                     500

1681    GGT GAA ACA GCT AAC CCA TTC TTC ACA GCG GCA GCC CTT ACT GTT ATG GCA ACA GCT GGA GTA GCA GCA GTT GTC AAA CGC AAA    1764
        Gly Glu Thr Ala Asn Pro Phe Phe Thr Ala Ala Ala Leu Thr Val Met Ala Thr Ala Gly Val Ala Ala Val Val Lys Arg Lys
            510                                     520                                     530

1765    GAA GAA AAC TAA GCT ATC ACT TTG TAA TAC TGA GTG AAC ATC AAG AGA GAA CCA GTC GGT TCT CTC TTT TAT GTA TAG AAG AAT    1848
        Glu Glu Asn ***                                 <-------------------<       >--------------------->

1849    GAG GTT AAG GAG GTC ACA AAC TAA ACA ACT CTT AAA AAG CTG ACC TTT CTA ATA ATC GTC TTT TTT TTA TAA TAA GAT GTA ATA    1932
1933    ATA TAA TTG ATA AAT GAG ATA CAT TTA ATC ATT ATG ACA AAA GGC AAG GAA AAA TAG CTG TAT CAT ATG CAA ATA ACC CCT GTT    2016
2017    TGC TCT TTA AAA AAG ATG TTA TCC TTA TTT CTC TAC GCA CAG GTG AAC AGC TAG GAG AGA ATC GTT TGA TTC TCT CTT TTC TTA    2100
2101    ATG GTC ATA AAG ACA AAG TCT CTT CTC ATC A    2131
```

FIG. 3. Nucleotide sequence of the serotype 24 M protein gene. The nucleotide sequence of the noncoding DNA strand including the M24 protein gene is presented with the nucleotide positions numbered at the beginning and end of each line. Amino acids are numbered below the residue. The positions of possible promoter sequences are underlined (P1) or overlined (P2), as is a potential ribosome-binding site (RBS). Repeated regions within the M24 gene are enclosed in parentheses and labeled above the 5′-most base of the repeat block. A palindromic sequence 3′ to the coding region implicated in transcription termination is noted by opposing arrows. * * *, Translational stop codon.

```
A1    TTA GAG GCA CGT AAG GCT GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT   508  612
      Leu Glu Ala Arg Lys Ala Asp Leu Glu Lys Ala Leu Glu Gly Ala MET Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala ( 118  152 )

A2    TTA GcG GcA CGT AAG GCT GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAc TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT   613  717
      Leu Ala Ala Arg Lys Ala Asp Leu Glu Lys Ala Leu Glu Gly Ala MET Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala ( 153  187 )

A3    TTA GAG GCA CGC cAG GCT GAa CTT GAA AAA GCA TTA GAA GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT   718  822
      Leu Glu Ala Arg Gln Ala Glu Leu Glu Lys Ala Leu Glu Gly Ala MET Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala ( 188  222 )

A4    TTA GcG GCA CGT AAG GCT GAT CTT GAA AAA GCA TTA GAA GGC GCA ATG AAc TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT   823  927
      Leu Ala Ala Arg Lys Ala Asp Leu Glu Lys Ala Leu Glu Gly Ala MET Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala ( 223  257 )

A5    TTA GAG GCA CGC cAG GCT GAa CTT GAA AAA GCA TTA GAA GGC GCA ATG AAT TTT TCA ACA GCG GAT TCA GCT AAA ATC AAA ACC TTA GAA GCA GAG AAA GCT GCT   928 1032
      Leu Glu Ala Arg Gln Ala Glu Leu Glu Lys Ala Leu Glu Gly Ala MET Asn Phe Ser Thr Ala Asp Ser Ala Lys Ile Lys Thr Leu Glu Ala Glu Lys Ala Ala ( 258  292 )

A6.1  TTG GAG GCA GaG AAa GCT GAT CTT GAA                                                                                                         1033 1059
      Leu Glu Ala Glu Lys Ala Asp Leu Glu                                                                                                         ( 293  301 )

A6.2                                  GAA AAA GCt TTA GAA Gaa GCA                                                                                 1351 1371
                                      Glu Lys Ala Leu Glu Glu Ala                                                                                 ( 399  405 )


B1    CGT CAA AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA CAA TTA GAA GCT GAA CAC CAA AAA CTA GAA GAA CAA AAC AAG ATT TCA GAA GCA AGC   1087 1191
      Arg Gln Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys Gln Leu Glu Ala Glu His Gln Lys Leu Glu Glu Gln Asn Lys Ile Ser Glu Ala Ser ( 311  345 )

B2    CGT CAA AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA CAA TTA GAA GCT GAA CAC CAA AAA CTA GAA GAA CAA AAC AAG ATT TCA GAA GCA AGC   1192 1296
      Arg Gln Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys Gln Leu Glu Ala Glu His Gln Lys Leu Glu Glu Gln Asn Lys Ile Ser Glu Ala Ser ( 346  380 )

B2.7  CGT CAA AGT CTT CGT CGT GAC TTG GAC GCA TCA CGT GAA GCT AAG AAA CAA GTT GAA                                                                 1297 1353
      Arg Gln Ser Leu Arg Arg Asp Leu Asp Ala Ser Arg Glu Ala Lys Lys Gln Val Glu                                                                 ( 381  399 )
```

FIG. 4. Comparison of the repetitive units of M24 protein. The sequences of the repetitive regions have been aligned to achieve maximal homology with amino acids represented by standard three-character designations. The numbers on the right correspond to base numbers in Fig. 3; those enclosed in parentheses correspond to the amino acid residue number. Nucleotides which are different from the first unit of the repeat are written in lowercase, and alterations in the predicted amino acid are underlined.

the second block of repeats may have interrupted the last unit of the first repeat. A 7-amino-acid homology with block A (Fig. 3 and 4, A6.2) occurs after the last repeat in B. This alignment is best shown in Fig. 4.

Chou-Fasman calculations (6) of the secondary structure of the protein would predict that the amino-terminal three-quarters of the mature M24 protein is alpha-helical. This is consistent with predictions based on the 7-residue periodicity of nonpolar or hydrophobic residues throughout this region (27). The portion of the molecule containing the second repeat element is predominately alpha-helix interrupted periodically by random coil. The carboxy-terminal third of the protein is structurally more random and is disrupted by proline residues which would introduce breaks in the secondary structure (residues 469 to 514).

Calculations of the hydropathicity (23) of this protein indicate that there is an amino-terminal hydrophobic region which precedes Val-43 corresponding to the putative signal sequence. There is also a hydrophobic region near the carboxy-terminal end of the protein (amino acids 515 to 534) which has characteristics of membrane anchor sequences (44) and a high probability of forming an alpha-helix which could span the membrane anchored by five charged amino acids which complete the protein.

Relatedness of type 24 and type 6 M proteins. Hollingshead et al. (19) have reported the sequence of the type 6 M protein gene, and we have compared the sequence of the type 24 M protein gene with their sequence. Results of such a comparison at the amino acid level are displayed as a matrix homology plot in Fig. 5. Several regions are common to both the M24 protein and the M6 protein. First, the amino-terminal signal sequence regions, designated A in Fig. 5, are homologous over 38 of the 42 amino acids (Fig. 6, row A), including two conservative replacements. Immediately following the signal peptide, the sequences of the M proteins appear to be unrelated.

A second region of homology between M24 and M6 includes the second repeat region, designated B1 to B2.7, of the M24 protein. Similar repeats occur twice in the M6 protein and were designated C1 and C2 by Hollingshead et al. (19). These homologies are shown as elements B to G in Fig. 5 and 6. At the amino acid level these regions are 72 to 97% homologous, and most of the variations represent conservative amino acid replacements resulting from single base changes.

The third region of homology between the M24 protein and the M6 protein encompasses the carboxy-terminal third of the molecule. In this region 160 of 164 (97.6%) amino acids are identical and 3 of the 4 differences represent conservative replacements (Fig. 5 and 6, element G). The conserved region, amino acids 376 to 539 of M24 and amino acids 278 to 441 of M6, overlaps the common repeat shared by M24 (residues 372 to 402) and M6 (residues 277 to 301). All of the

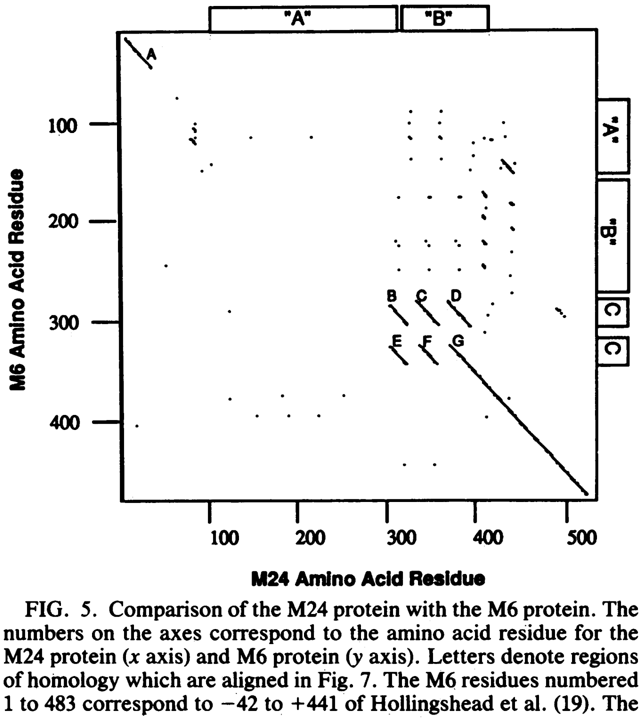

FIG. 5. Comparison of the M24 protein with the M6 protein. The numbers on the axes correspond to the amino acid residue for the M24 protein (*x* axis) and M6 protein (*y* axis). Letters denote regions of homology which are aligned in Fig. 7. The M6 residues numbered 1 to 483 correspond to −42 to +441 of Hollingshead et al. (19). The repeated elements within M24 (top) and M6 (right side) are indicated by the letters "A", "B", and C. The parameters used in this matrix were a range of 6 amino acids giving a window of 13 within which 75% homology was required to generate a signal.

```
                  residues                                                            % homology

      M24       1     42     MTKNNTNRHYSLRKLKTGTASVAVALTVLGAGLVVNTNGVSA               90.5
A:                           : ::::::::::::::: :::::::.:.::::::::::::::              (95.2)
      M6      -42     -1     MAKNNTNRHYSLRKLKKGTASVAVALSVIGAGLVVNTNEVSA


      M24      305 - 329          QVLNANRQSLRRDLDASREAKKQLE                          72.0
B:                                 :..: : .::::::::::::::.:                         (88.0)
      M6       235 - 259          KVSEASRKGLRRDLDASREAKKQVE


      M24      339 - 364         NKISEASRQSLRRDLDASREAKKQLE                          84.6
C:                               ::.::::: .::::::::::::::.:                         (96.2)
      M6       234 - 259         NKVSEASRKGLRRDLDASREAKKQVE


      M24      372 - 402         EQNKISEASRQSLRRDLDASREAKKQVEKAL                     83.9
D:                               : ::.::::: .::::::::::::::::: :                    (90.3)
      M6       232 - 262         EGNKVSEASRKGLRRDLDASREAKKQVEKDL


      M24      305 - 329           QVLNANRQSLRRDLDASREAKKQLE                         76.0
E:                                 :. .:.::.::::::::::::::.:                        (96.3)
      M6       277 - 301           QISDASRQGLRRDLDASREAKKQVE


      M24      339 - 364          NKISEASRQSLRRDLDASREAKKQLE                         92.0
F:                                   ::.::::.::::::::::::::::                       (96.0)
      M6       277 - 301           QISDASRQGLRRDLDASREAKKQVE


G:
                 380       390       400       410       420       430
      M24     ISEASRQSLRRDLDASREAKKQVEKALEEANSKLAALEKLNKELEESKKLTEKEKAELQA
              ::.::::.::::::::::::::::::::::::::::::::::::::::::::::::::::
      M6      ISDASRQGLRRDLDASREAKKQVEKALEEANSKLAALEKLNKELEESKKLTEKEKAELQA
               280       290       300       310       320       330

                 440       450       460       470       480       490
              KLEAEAKALKEKLAKQAEELAKLRAGKASDSQTPDAKPGNKAVPGKGQAPQAGTKPNQNK
              ::::::::::: :::::::::::::::::::::::::::::.::::::::::::::::::
              KLEAEAKALKEQLAKQAEELAKLRAGKASDSQTPDAKPGNKVVPGKGQAPQAGTKPNQNK
               340       350       360       370       380       390

                 500       510       520       530
              APMKETKRQLPSTGETANPFFTAAALTVMATAGVAAVVKRKEEN                  97.6
              ::::::::::::::::::::::::::::::::::::::::::::                 (99.4)
              APMKETKRQLPSTGETANPFFTAAALTVMATAGVAAVVKRKEEN
               400       410       420       430       440
```

FIG. 6. Comparison of homologous regions in M24 and M6 proteins. The optimal alignment of the regions of homology (A to G) identified as in Fig. 5 are shown. The residues from each protein being compared are indicated by the numbering system of Hollingshead et al. (19) for the M6 protein. Amino acid identities are represented by a colon; amino acids which would be considered conservative replacements are indicated by a period. The percent homology is tabulated to the right of each sequence. The first score excludes the conservative replacement, and the second score counts the conservative replacement as an identical residue.

homologies at the level of the protein are also reflected by similar DNA homologies which range from 78 to 98%. Since this region is so highly conserved, it may play an important role through interaction with the cell wall and membrane to anchor the M protein molecule to the cell surface. It has been observed that M6 protein obtained under certain isolation conditions has tightly bound cell wall material (16).

Several other surface proteins from gram-positive organisms have now been sequenced (14, 43). The proposed membrane-spanning regions of these molecules share strong homology with the hydrophobic anchor region of the M24 and M6 molecules.

The homologies between the M6 and M24 protein genes extend beyond the coding region. The sequences for at least 156 nucleotides upstream of the start codons share 96.2% homology, allowing for deletion-addition of nucleotides in this region. This upstream region is also homologous with sequences upstream of the M12 protein gene (35), with 94.2% homology. Downstream from the coding region for at least 108 nucleotides, the M24 and M6 sequences are 98.1% homologous, including the inverted repeat (nucleotides 1802 to 1838 of M24) which might serve as a factor-independent termination signal for transcription (33). These homologies are depicted in Fig. 7.

DISCUSSION

Previously available structural information on the M24 surface antigen has been based on amino acid sequence analysis of cyanogen bromide peptides derived from an M24 polypeptide fragment solubilized from cells following pepsin treatment (1, 2). Our work has localized these peptides

```
A.

                10        20        30        40        50        60
M6      attcatcattaatagcatttaggtcaaaa-g-tggcaaaagctaaaaaagctggtcttta
        ::::::::::::::::::::::::::::: : ::::::::::::::::: ::::::::::
M24     ATTCATCATTAATAGCATTTAGGTCAAAAAGCTGGCAAAAGCTAAAAAAACTGGTCTTTA
                                           :::::::::::::: ::::::::::
M12                                        caaaagctaaaaaagctggtcttta

                70        80        90       100       110       120
        ccttttggcttttattatttacaatagaattattagagttaacccctgaaaatgagggtt
        :::::::::::::::::::::::::::::::::::::::::::::::::::::::::: :
        CCTTTTGGCTTTTATTATTTACAATAGAATTATTAGAGTTAACCCCTGAAAATGAGGGGT
        ::::::::::: :::::::::::::::::::::::::::::: ::::::::::::::: :
        ccttttggcttatattatttacaatagaattattagagttaaaccctgaaaatgagggtt


               130       140       150
        ttt-cct-aaaaaatgataacataaggagcataaaa  Number of matches = 150/152
        ::: ::: ::::::::::::::::::::::::::::
        TTTTCCTAAAAAAATGATAACATAAGGAGCATAAAA
        ::::::::::: ::::::: :::::::::::::::
        ttttcctaaaa-aatgataacataaggagcataaac  Number of matches = 113/120


B.

              1780      1790      1800      1810      1820
M24     TAAGCTATCACTTTGTAATACTGAGTGAACATCAAGAGAGAACCAGTCGGTTCT
        :::::::::::::::::::::::::: :::::::::::::::::::::::::::
M6      taagctatcactttgtaatactgagtcaacatcaagagagaaccagtcggttct
          1620      1630      1640      1650      1660      1670

          1830      1840      1850      1860      1870      1880
        CTCTTTTATGTATAGAAGAATGAGGTTAAGGAGGTCACAAACTAAACAACTCTT
        :::::::::::::::::::::::: :::::::::::::::::::::::::::::
        ctcttttatgtatagaagaatgagattaaggaggtcacaaactaaacaactctt
                1680      1690      1700      1710      1720

                    Number of matches = 106/108
```

FIG. 7. (A) Comparison of the 5′ nontranslated regions of M6, M12, and M24 genes. Nucleotides that are identical with those of the M24 DNA are indicated by a colon. Any gaps in the sequence needed to achieve optimal alignment are indicated by a dash. (B) Comparison of the 3′ nontranslated regions of M24 and M6. Nucleotide identities are indicated as above.

within the total sequence of the M24 polypeptide and has demonstrated that the pepsin-solubilized fragment is derived from the amino-terminal portion of the protein. However, if our assignment of the initiating AUG is correct, then the initial methionine and the following 41 amino acids are not represented in the pepsin-solubilized polypeptide. This material may have been lost during proteolysis. However, these amino acids have a structure consistent with their being a signal sequence which may be processed biologically during export.

The amino acid sequence encoded by the 3′ terminal 40% of the gene is also not found in the pepsin-solubilized material. As discussed above, we postulate that the carboxy-terminal 25 amino acids constitute a transmembrane anchor. Hence, it is possible that the carboxy-terminal portion of the M24 protein is pepsin resistant at least in part because of protection by membrane and cell wall materials.

A putative pepsin-sensitive cleavage site is indicated in Fig. 8. As summarized in this figure, the central portion of the M24 polypeptide (residues 118 to 405) contains two regions of distinct repeated elements. Each of the units within the repeat element is 35 amino acids in length, and they are reiterated tandemly. Within each unit of the repeat, the DNA and amino acid sequence are highly conserved. These basic features are analogous to those reported by Hollingshead et al. (19) for the M6 protein, which is composed of three repeated elements. The most carboxy terminal of these repeats, C1 and C2, are closely related to the most carboxy-terminal repeats of M24. There is no homology between the more amino-proximal repeats of M24 and those of the M6 protein. These results suggest that the type specificity resides in the more amino-proximal region of the proteins, which includes the unrelated repeats. In contrast to the lack of homology in the amino-proximal repeats of M6 and M24, the amino-terminal signal sequences and the carboxy-terminal 140 amino acids are almost identical for the two proteins (Fig. 8). The amino-terminal sequence of the mature M5 protein has also been sequenced (26) and is unrelated to either the M24 and M6 sequences (data not shown).

This finding of homology between the M24 and M6 proteins is consistent with the observation of Scott et al. (37) that the genes for these proteins share conserved DNA sequences in the carboxy-terminal region on the basis of DNA hybridization analysis. However, the sequence data presented here provide the first direct evidence for extensive

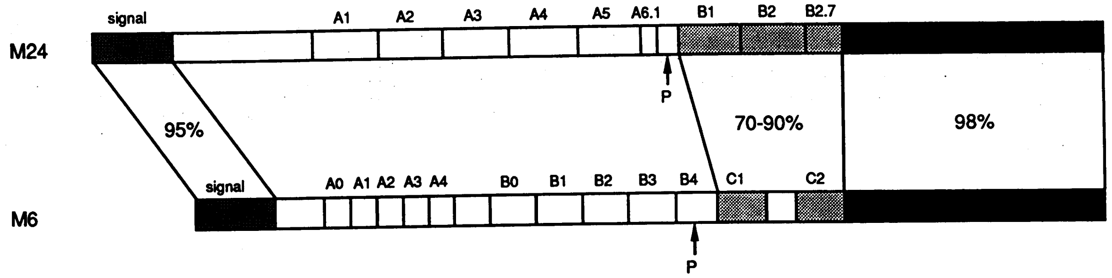

FIG. 8. Summary of the relationships between the type 24 and 6 M proteins. Sequences sharing homology are represented by corresponding shaded areas, and the approximate levels of homology are given by the numbers between the sequences. The repeated regions are denoted by the boxes representing the individual units and are labeled as described in the text. The unshaded areas are those for which no homology is found. P denotes the most probable pepsin cleavage site.

homology in this region. This might provide useful information for designing antistreptococcal vaccines. By using monoclonal antibodies developed against the purified M6 protein, it was found that M proteins isolated from a number of different serotypes possessed a common antigenic determinant thought to be located near the carboxy terminus of the M protein molecule (22).

This organization of the M24 and M6 proteins, a variable amino-terminal region followed by repeated elements with a conserved carboxyl terminus, may be a common feature of M proteins. The noncoding sequences 5′ and 3′ to the M24 gene are also highly conserved. The M24 sequences upstream of the gene are identical with the sequences upstream of the M6 (150 of 152 bases align) (19) and M12 (113 of 120 bases align) (35) genes. The sequences downstream of the M24 and M6 coding region are also related: 106 of 108 bases can be aligned. The lack of diversity in these regions suggests that they may be important in regulating M protein gene expression (7, 35) or that adjacent genes may also be conserved between these strains.

The presence of repeated domains within protein molecules, including M proteins, suggests that the genes encoding such proteins evolved by gene duplication and divergence. The introduction of new sequences into the middle of the M protein gene by translocation or recombination could yield any number of novel possible combinations. However, only combinations which would be able to function as an antiphagocytic determinant as part of an M protein molecule would ensure the survival of a strain expressing a new combination of antigenic determinants. The native M protein on the surface of the cell is thought to exist as an alpha-helical coiled coil, and sequences which assume this conformation would be favored (32). Exactly how these sequences arise and recombine to generate a polypeptide with the proper characteristics remains to be determined. It is known, however, that clinical isolates of group A streptococci are frequently lysogenized by transducing bacteriophages (20, 21), which could be responsible for generating new combinations of M protein sequences by recombination following transduction. In addition, accumulation of single mutations within the units of the repeats which are subject to further reassortment by homologous intragenic recombination.

In addition, one might anticipate that the presence of highly related repeated elements would be unstable to recombination. In fact, Cleary et al. (8) have observed that the M^+ phenotype of certain group A streptococcal strains is unstable, a finding which could be explained by the deletion of the strongly antigenic M^+ determinants from the structural gene. Small deletions can also result in the loss of virulence by *S. pyogenes* (7, 41), even when the deletions lie outside the M protein gene. Size heterogeneity of the M protein has been observed among isolates derived from a single strain (15, 16); this heterogeneity appears to be due to variation in the number of repeated elements.

ACKNOWLEDGMENTS

This work was supported by Public Health Service grants AI15619 (V.B.) and AI10085 (E.B.) from the National Institute of Allergy and Infectious Diseases.

LITERATURE CITED

1. **Beachey, E. H., J. M. Seyer, J. B. Dale, and D. L. Hasty.** 1983. Repeated covalent structure and protective immunogenicity of native and synthetic polypeptide fragments of type 24 streptococcal M-protein. J. Biol. Chem. **258:**13250–13257.
2. **Beachey, E. H., J. M. Seyer, and A. H. Kang.** 1978. Repeating covalent structure of streptococcal M-protein. Proc. Natl. Acad. Sci. USA **75:**3163–3167.
3. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80:**3963–3965.
4. **Cardineau, G. A., and R. Curtiss.** 1987. Nucleotide sequence of the *asd* gene of *Streptococcus mutans*. Identification of the promoter region and evidence for attenuator-like sequences preceding the structural gene. J. Biol. Chem. **262:**3344–3353.
5. **Cheng, S.-C., and P. Modrich.** 1983. A positive selection cloning vehicle useful for overproduction of hybrid proteins. J. Bacteriol. **154:**1005–1008.
6. **Chou, P. Y., and G. D. Fasman.** 1974. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry **13:**211–221.
7. **Cleary, P., S. Jones, J. Robbins, and W. Simpson.** 1987. Phase variation and M-protein expression in group A streptococci, p. 101–105. *In* J. J. Ferretti and R. Curtiss III (ed.), *Streptococcal genetics*. American Society for Microbiology, Washington, D.C.
8. **Cleary, P. P., Z. Johnson, and L. W. Wannamaker.** 1975. Genetic instability of M protein and serum opacity factor of group A streptococci: evidence suggesting extrachromosomal control. Infect. Immun. **12:**109–118.
9. **Dale, J. B., and E. H. Beachey.** 1982. Protective antigenic determinant of streptococcal M-protein shared with sarcolemmal membrane protein of human heart. J. Exp. Med. **156:**1165–1176.
10. **Dale, J. B., and E. H. Beachey.** 1984. Unique and common protective epitopes among different serotypes of group A streptococcal M proteins defined with hybridoma antibodies. Infect. Immun. **46:**267–269.
11. **Dale, J. B., and E. H. Beachey.** 1985. Multiple heart-cross reactive epitopes of streptococcal M proteins. J. Exp. Med. **161:**113–122.
12. **Dale, J. B., and E. H. Beachey.** 1985. Epitopes of streptococcal M protein shared with cardiac myosin. J. Exp. Med. **162:**583–

591.

13. **Dale, R. M. K., B. A. McClure, and J. P. Houchins.** 1985. A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18SrDNA. Plasmid **13**:31–40.
14. **Fahnestock, S. R., S. Alexander, M. Nagle, and D. Filipa.** 1986. Gene for an immunoglobulin-binding protein from a group G streptococcus. J. Bacteriol. **167**:870–880.
15. **Fischetti, V. A., M. Jarymowycz, K. F. Jones, and J. R. Scott.** 1986. Streptococcal M protein size mutants occur at high frequency within a single strain. J. Exp. Med. **164**:971–980.
16. **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.
17. **Fox, E. N.** 1974. M proteins of group A streptococci. Bacteriol. Rev. **38**:57–86.
18. **Hanahan, D.** 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166**:557–580.
19. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1986. Complete nucleotide sequence of type 6 M protein of group A *Streptococcus*. J. Biol. Chem. **261**:1677–1686.
20. **Hyder, S. L., and M. M. Streitfeld.** 1978. Transfer of erythromycin resistance from clinically isolated lysogenic strains of *Streptococcus pyogenes* via their endogenous phage. J. Infect. Dis. **138**:281–286.
21. **Joklik, W. K., H. P. Willett, and D. B. Amos (ed.).** 1980. Zinsser Microbiology, 17th ed. Appleton-Century-Crofts, New York.
22. **Jones, K. F., B. N. Manjula, K. H. Johnston, S. K. Hollingshead, J. R. Scott, and V. A. Fischetti.** 1985. Location of variable and conserved epitopes among the multiple serotypes of streptococcal M-protein. J. Exp. Med. **161**:623–628.
23. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
24. **Laemmli, U. K.** 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature (London) **227**:680–685.
25. **Malke, H., B. Roe, and J. J. Ferretti.** 1985. Nucleotide sequence of the streptokinase gene from *Streptococcus equisimilis* H46A. Gene. **34**:357–362.
26. **Manjula, B. N., A. S. Acharya, S. M. Mische, T. Tairwell, and V. A. Fischetti.** 1984. The complete amino acid sequence of a biologically active 197-residue fragment of M protein isolated from type 5 group A streptococci. J. Biol. Chem. **259**:3689–3693.
27. **Manjula, B. N. and V. A. Fischetti.** 1980. Tropomyosin-like seven residue periodicity in three immunologically distinct streptococcal M proteins and its implications for the antiphagocytic property of the molecule. J. Exp. Med. **151**:695–708.
28. **Martin, P., P. Trieu-Cuot, and P. Courvalin.** 1986. Nucleotide sequence of the *tetM* tetracycline resistance determinant of the streptococcal conjugative shuttle transposon Tn*1545*. Nucleic Acids Res. **14**:7047–7058.
29. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
30. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
31. **Newman, A. K., R. A. Rubin, S. H. Kim, and P. Modrich.** 1981. DNA sequences of structural genes for *Eco*RI DNA restriction and modification enzymes. J. Biol. Chem. **256**:2131–2142.
32. **Phillips, G. N., Jr., 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 **82**:4689–4693.
33. **Platt, T.** 1986. Transcription termination and the regulation of gene expression. Annu. Rev. Biochem. **55**:339–372.
34. **Pustell, J., and T. Kafatos.** 1982. A high speed, high capacity homology matrix: zooming through SV40 and polyoma. Nucleic Acids Res. **10**:4765–4782.
35. **Robbins, J. C., and P. Cleary.** 1987. Transcription studies of type 12 M protein phase variants, p. 109–111. *In* J. J. Ferretti and R. Curtiss III (ed.) Streptococcal genetics. American Society for Microbiology, Washington, D.C.
36. **Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen.** 1982. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. **162**:729–773.
37. **Scott, J. R., S. K. Hollingshead, and V. A. Fischetti.** 1986. Homologous regions within M protein genes in group A streptococci of different serotypes. Infect. Immun. **52**:609–612.
38. **Shaw, J., and D. B. Clewell.** 1985. Complete nucleotide sequence of the macrolide lincosamide streptogramin B resistance transposon Tn*917* in *Streptococcus faecalis*. J. Bacteriol. **164**:782–796.
39. **Shimatake, H., and M. Rosenberg.** 1981. Purified lambda regulatory protein cII positively activates promotors for lysogenic development. Nature (London) **292**:128–132.
40. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
41. **Spanier, J. G., S. J. C. Jones, and P. Cleary.** 1984. Small DNA deletions creating avirulence in *Streptococcus pyogenes*. Science **225**:935–938.
42. **Towbin, H., T. Staehelin, and J. Gordon.** 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc. Natl. Acad. Sci. USA **76**:4350–4354.
43. **Uhlen, M., B. Guss, B. Nilsson, S. Gatenbeck, L. Philipson, and M. Lindberg.** 1984. Complete sequence of the staphylococcal gene encoding protein A. A gene evolved through multiple duplications. J. Biol. Chem. **259**:1695–1702.
44. **Warren, G.** 1981. Membrane proteins: structure and assembly, p. 215–257. *In* J. B. Finean and R. H. Mitchell (ed.), Membrane structures. Elsevier/North Holland Biomedical Press, Amsterdam.
45. **Weeks, C. R., and J. J. Ferretti.** 1986. Nucleotide sequence of the type A streptococcal exotoxin (erythrogenic toxin) gene from *Streptococcus pyogenes* bacteriophage T12. Infect. Immun. **52**:144–150.
46. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
47. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.