# Cloning and Nucleotide Sequence of Type 3 M Protein Gene (emm3) Consisting of an N-Terminal Variable Portion and C-Terminal Conserved C Repeat Regions: Relation to Other Genes of Streptococcus pyogenes

Streptococcal Diseases Study Group (Chief: Kiyoshi HOSHINA) Chihiro KATSUKAWA

Laboratory of Microbiology, Osaka Prefectural Institute of Public Health (Received: November 12, 1993) (Accepted: February 7, 1994)

Key words: Streptococcus pyogenes, M proteingenes, PCR, emm3, emm12

## Abstract

The structural gene for type 3 M protein of *Streptococcus pyogenes*, which consists of an N-terminal variable portion and C-terminal conserved repeat regions, has been cloned by the polymerase chain reaction (PCR) with two primers (K-1 and K-2). They were selected from the best conserved region of the leader sequences and of the C-terminal portion near the Hexapeptide (LPSTGE) sequence found in different M proteins. From the nucleotide sequences of the product, 1645 base pairs were determined, including 32 amino acids of the leader sequences, the complete N-terminal variable region and the conserved C repeat regions. Analysis of the deduced amino acids of the sequence revealed the existence of two major repeat regions, the B and C repeat regions. Comparison of the C-repeat regions among M3 and other M proteins showed them to be more than 90% identical. The two B repeat blocks in M3 protein are also similar to those in M12 protein. Predictive secondary structure analysis of M3 protein reveals a strong alpha-helical potential. The algorithm also shows that the beta-sheet and turn potential for region 23–42 in M3 protein are similar to those for region 28–50 in M12 protein. The results indicate that M3 protein is closely related to M12 protein.

# Introduction

Streptococcus pyogenes is responsible for a wide variety of human diseases, the most common of which are nasopharyngitis and impetigo<sup>1)</sup>. Moreover, streptococcal pharyngeal infection in humans may develop into rheumatic fever or glomerulonephritis<sup>1)</sup>. The principal virulence factor of *S. pyogenes* is a cell wall constituent known as M protein that gives the organism the ability to resist phagocytosis<sup>2)</sup>. This virulence factor displays antigenic diversity within its amino terminal region<sup>3)</sup>. 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<sup>2)</sup>. M protein is thought to inhibit alternative C3 convertase formation to restrict deposition of C3 on the streptococci and also to inhibit the classical C5 convertase formation in order to interfere with efficient complement receptor-mediated phagocytosis<sup>4)</sup>. However, the relationship between M protein antigenic diversity and its antiphagocytic activity is not understood, and neither is the genetic basis for M protein antigenic diversity or the structural basis for

別刷請求先: (〒537)大阪市東成区中道1-3-69 大阪府立公衆衛生研究所 勝川 千尋

functions common to all M protein serotypes. In order to characterize them in more detail, the nucleotide sequences of genes encoding a number of different M protein serotypes need to be cloned, sequenced and compared. In this report, we describe the cloning and sequencing of emm3 from the *S. pyogenes* type 3 M protein gene which has not been known to date and compare this gene with others reported previously.

# **Materials and Methods**

Bacterial strains, plasmid and media: *S. pyogenes*, the M<sup>+</sup> (type 3) strain C203 (ATCC 12384) was used in this study. *Escherichia coli* SJ2, harboring plasmid pJRS42.50, consists of an Xba I-Pvu II fragment including the emm6 gene from *S. pyogenes* D471 cloned into pUC19; it was a gift from Dr. June R. Scott<sup>5)</sup>. *E. coli* JM109 was used as the recipient for plasmid transformation and for phage M13 propagation. The plasmid vector was pUC118 obtained from Takara Shuzo Co., Ltd, Kyoto, Japan. *S. pyogenes* was grown in Todd-Hewitt broth (Difco Laboratories, Detroit, MI, U.S.A.); *E. coli* strains were grown in LB broth.

Isolation of DNA: Chromosomal DNA from *S. pyogenes* strain C203 was prepared according to a procedure reported previously<sup>6</sup>. Briefly, the cultured strain C203 (500 ml) was centrifuged at 8,000 rpm for 30 min and the resulting pellet was lysed in 25 ml of 10 mM Tris-HCl, 1 mM EDTA, pH 8.0 (TE), 1.25 ml of 10% SDS and 0.125 ml of a 20 mg/ml solution of proteinase K. The resulting mixture was incubated at 37°C for 45 min. To the lysate, 4.74 ml of 5 M NaCl was added with thorough mixing, and then 4 ml of CTAB/NaCl solution (10% hexadecyl trimethyl ammonium bromide in 0.7 M NaCl) was added to the mixed lysate and this mixture was incubated at 65°C for 20 min. The CTAB-treated lysate was extracted with an equal volume of phenol/chloroform/isoamyl alcohol to remove CTAB-protein/polysaccharide complexes. The aqueous phase was transferred to a fresh tube and then 0.6 volume of isopropanol was added. The resulting precipitate was washed once with 70% ethanol. The washed DNA was suspended at 100 ng of genomic DNA per ml in TE.

*E. coli* SJ2 cultured in LB broth containing  $50 \mu g$  of ampicillin per ml was collected by centrifugation at 8,000 rpm for 30 min, and the plasmid in the bacteria was isolated with a Qiagen column (QIAGEN-tip 100) (Diagen Inc. Chatsworth, CA, U.S.A.), according to the protocol of the manufacturer. The isolated plasmid was digested with Msp I and Pvu II restriction enzymes. After electrophoresis of the restricted plasmid in a 0.8% gel, the fragment, which contained only the region encoding the M6 protein and lacked 32 bases at the 5' end and 38 bases at the 3' end of the gene, was purfied with a Gene Clean Kit (Bio 101, Inc., La Jolla, CA, U.S.A.).

Oligonucleotides: Two oligonucleotides were synthesized with a 380 B automatic DNA synthesizer (Applied Biosystems Inc., Foster City, CA, U.S.A.) and used as specific primers for the required extension and amplification reactions. They were 31-mer (specific primer K-1; 5'-CCG<u>GGATCCTATTCGCTTA-GAAAATTAAAAA-3'</u>) and 30-mer (specific primer K-2; 5'-CCG<u>GTCGACAAGTTCTTCAGCTTGTT-TCGC-3'</u>). The K-1 and K-2 primers contained Bam HI and Sal I recognition sequences at the 5' end respectively. This made it easy to insert the amplified fragment into the vector.

Cloning of emm3 gene by the polymerase chain reaction: PCR was performed in a Hybaid Thermal Reactor (Hybaid Ltd., U.K.). The reaction mixture contained 10  $\mu$ l of 100 ng/ $\mu$ l genomic DNA, 25 pmol of each phosphorylated primer, 0.2 mM deoxynucleotide triphosphate mix, 0.01% (w/v) gelatin, 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub> and 2 U of Taq polymerase (Takara) in a total volume of 100  $\mu$ l, and the reaction mixture was overlaid with 2 drops of mineral oil (Sigma Chemical Co., St. Louis, MO, U.S.A.). PCRs (30 cycles) were performed with each cycle consisting of denaturation (94°C, 1 min), annealing (55°C, 2 min) and extension (72°C, 2 min). The PCR product was precipitated with ethanol and digested with Bam HI and Sal I restriction enzymes. After electrophoresis, it was extracted from the preparative agarose gel and cloned into the cloning vector pUC118 which had been digested with Bam HI and Sal I.

## Chihiro KATSUKAWA

DNA sequence: Plasmid pUC118 harboring the emm3 gene transformed in *E. coli* JM109 was grown in L-broth containing 50  $\mu$ g of ampicillin per ml and purified with Econopack Q (Pharmacia LKB Biotechnology Inc., Piscataway, N.J., U.S.A.). The purified emm3 gene inserted into the vector was digested with Bam HI, Sal I and Bgl II (fragments of about 440 and 1030 base pairs), with Bam HI, Sal I and Sca I (fragments of about 880 and 590 base pairs) and with Bam HI, Sal I and Stu I (fragments of about 1200 and 270 base pairs). Each fragment was purified with a Sephaglas® Band Prep Kit (Pharmacia) and inserted into M13 phage. Single-strand DNA was purified from the supernatant of the cultured phage with a Sephaglas® Phage Prep Kit (Pharmacia). The purified single-strand DNA was sequenced with an Auto Read® Sequencing Kit (Pharmacia) by A.L.F. DNA Sequencer (Pharmacia).

Hybridization: DNA samples were denatured by heating for 10 min at 95°C and then chilled rapidly on ice for dot blot hybridization. DNA samples also were denatured in NaOH for Southern hybridization before application to membrane filters (Schleicher and Schuell, Dassel, Germany) which had been soaked in distilled water and then in  $20 \times SSC$  buffer ( $1 \times SSC$  is 0.15 M sodium chloride, 0.15 M sodium citrate, pH 7.0). Hybridization of the restriction fragments was conducted by the method of Southern. The DNA samples were bound to a nitrocellulose membrane by baking for 2 hr in a vacuum at 80°C. The DNA bound to the filter was incubated for 2 hr at 68°C in prehybridization solution (0.25% powdered skim milk in  $2 \times SSC$ ). Hybridization was carried out by overnight incubation at 65°C with labeled probe DNA diluted to a 5 ml final volume of prehybridization solution. The hybridized filters were washed twice in  $2 \times SSC$ , 0.1% SDS for 15 min at room temperature and twice at 68°C in 0.1 × SSC, 0.1% SDS for 15 min 68°C. Probes were labeled with digoxigenin-11-UTP (Boehringer Corp. Ltd., Sussex, U.K.) by random hexanucleotide primers according to the protocol supplied by the manufacturer. Digoxigenin-labeled probes were detected by using an anti-digoxigenin antibody alkaline phosphatase conjugate (Boehringer) and the substrates of BCIP (5-bromo-4-chloro-3-indolyl phosphate) and NBT (nitroblue tetrazolium).

Nucleotide sequence accession number. The 1465 base pair nucleotide sequence of emm3 gene is available from DDBJ, EMBL and GenBank Nucleotide Sequence Databases under accession number

Fig. 1 Analysis of the amplification product. (A) Ethidium bromide picture of the agarose gel (0.8%) of the amplified product. Lane 1, PCR product with K-1 and K-2 primers; lane 2, marker. (B) Dot blot hybridization analysis with the Msp I-Pvu II fragment of the emm6 gene probe. Serial five-fold dilutions of 6 ng of the Msp I-Pvu II fragment of the emm6 gene (lane a) and 5 ng of DNA amplified with K-1 and K-2 primers (lane b) were spotted onto nitrocellulose and probed with the digoxigenin-labeled Msp I-Pvu II fragment. (C) Southern blot analysis with the Msp I-Pvu II fragment of the emm6 probe. Lane 1, DNA amplified with K-1 and K-2 primers; lane 2, Msp I-Pvu II fragment of emm 6 gene. DNA hybridized with the probe is shown by the arrows.



700

D14415.

# Results

Cloning and nucleotide sequence of the emm3 gene from S. *pyogenes* type 3 strain C203 with PCR: Much information is available on nucleotide sequences from various strains<sup>7</sup>)-12). We tried to obtain the

Fig. 2 Nucleotide and deduced amino acid sequences of the emm3 gene. The DNA strand is located at 5' to 3' and its nucleotides are numbered above each line. Amino acid residues are presented as single letters below each line. B repeat blocks and C repeat blocks are indicated by underlining.

Signal peptide →							
10		0 40			70	80	90
TATTCGCTTAGAAAATI YSLRKL							
Mature prof		r s v A	V V L	1 1 1 6		VAGQ	1 V
•	110 12	0 130	) 140	0 150	160	170	180
AAGGCAGATGCTAGGAG							
KADARS	VNGE	FPRH	IVKL	KNEI	ENL	LDQV	тQ
190	200 21	0 220	230	240	250	260	270
ттататаатаасатаа	TAGTAATTACCA	ACAATATAGTG	CACAAGCTGO	CAGACTTGACC	TGAGACAAAAG	GCTGAATATCT	AAAAGGC
LYNKHN	S N Y Q	Q Y S A	Q A G	RLDI	RQK	AEYL	K G
280	290 30	0 310	) 320	) 330	340	350	360
CTTAATGATTGGGCTGA							
LNDWAE							
274							
370 TTAGAAAAGGAGGTTAA	380 39 AGAACTTAAGAA				430 ACTTAGATAAA	440 GATTTTGACTT	450 3600 AAA
LEKEVK							
460 CAGGGGTATGTTTTATC	470 48				520	530	540
Q G Y V L S							
-							•
	560 57				610	620	630
ATTAGCGAAGAGCTAGA ISEELE							
10101	1	K V 5 3	1 1 9	DLIE	кQи	K V S Q	5 5
	650 66				700	710	720
CAAGAATTAGCTACTAC							
QELATT	KQNA	KEDF	ELA	ALAN	AAD	кокг	EA
730	740 75	760	770	780	790	800	810
AAGATTGCCGATTTAGA	AACAAAACTAAA	GAGGCAAAGG	AAGATTTTGA	ACTAGCAGCAT	PAGGTCACCAA	CATGCTCATAA	GAGTAT
	AACAAAACTAAA	GAGGCAAAGG	AAGATTITGA DFE	ACTAGCAGCAT L A A L	PAGGTCACCAA	CATGCTCATAA	GAGTAT
AAGATTGCCGATTTAGA K I A D L E	AACAAAACTAAA	GAGGCAAAGG E A K E	AAGATTTTGA DFE B1-re	ACTAGCAGCAT L A A L spaet	PAGGTCACCAA	CATGCTCATAA	GAGTAT
AAGATTGCCGATTTAGA K I A D L E 820 CAAGCAAAACTAGCAGA	AACAAAACTAAA T K L K 830 844 AAAAGATGATCAJ	GAGGCAAAGG E A K E 850 ATTAAACAAC	AAGATTTTGA DFE B1-re 860 TAGAA <u>GAGCA</u>	ACTAGCAGCAT LAAL spaet 870 AAAACAAATCC	TAGGTCACCAA G H Q 880 TAGATGCTAGC	CATGCTCATAAT H A H N 890 CGTAAAGGTACA	NGAGTAT E Y 900 NGCAAGA
AAGATTGCCGATTTAGA K I A D L E 820	AACAAAACTAAA T K L K 830 844 AAAAGATGATCAJ	GAGGCAAAGG E A K E 850 ATTAAACAAC	AAGATTTTGA DFE B1-re 860 TAGAA <u>GAGCA</u>	ACTAGCAGCAT LAAL spaet 870 AAAACAAATCC	TAGGTCACCAA G H Q 880 TAGATGCTAGC	CATGCTCATAAT H A H N 890 <u>CGTAAAGGTACA</u> R K G T	IGAGTAT E Y 900 IGCAAGA A R
AAGATTGCCGATTTAGA. K I A D L E 820 CAAGCAAAACTAGCAGAI Q A K L A E	AACAAAACTAAA T K L K 830 844 AAAAGATGATCAJ	GAGGCAAAGG E A K E 850 ATTAAACAAC I K Q L	ANGATTTTGA DFE B1-re 860 TAGAA <u>GAGCA</u> EEQ	ACTAGCAGCAT LAAL Spaet 870 AAAACAAATCC KQIL	TAGGTCACCAA G H Q 880 TAGATGCTAGC	CATGCTCATAAT H A H N 890 <u>CGTAAAGGTACA</u> R K G T	IGAGTAT EY 900 IGCAAGA AR -repeat
AAGATTGCCGATTTAGA. K I A D L E 820 CAAGCAAAACTAGCAGAI Q A K L A E	AACAAAAACTAAAA T K L K 830 844 AAAAAGATGATCAA K D D Q 920 934	GAGGCAAAGG, E A K E B50 ATTAAACAAC I K Q L 940	AAGATTTTGA DFE B1-re 860 TAGAA <u>GAGCA</u> EEQ 950	ACTAGCAGCAT L A A L spaet 870 <u>AAAACAAATCC</u> K Q I L 960	GHQ B80 TAGATGCTAGC DAS 970	CATGCTCATAAT H A H N 890 CGTAAAGGTACP R K G T B2 980	IGAGTAT EY 900 IGCAAGA AR -repeat 990
AAGATTGCCGATTTAGA. K I A D L E 820 CAAGCAAAACTAGCAGA. Q A K L A E 910	АЛСЛАЛАСТАЛА Т К L К 830 844 АЛАЛАДАТСАЈ К D D Q 920 931 <u>ССАЛОСТАЛА</u> ЛАЈ	GAGGCAAAGG. E A K E B 50 ATTAAACAAC I K Q L 9 940 GCTACGGAAGC	AAGATTTTGA DFE B1-re 860 TAGAA <u>GAGCA</u> EEQ 950 CTGAATTAAA	АСТАGCAGCAT L A A L spaet 870 <u>АЛААСАААТСС'</u> К Q I L 960 СААССТСАААС	G H Q B80 TAGATGCTAGC D A S 970 CAGAGCTTGCA	CATGCTCATAAN H A H N B90 <u>CGTAAAGGTACA</u> R K G T B2 980 AAAGTTACA <u>GAA</u>	E Y 900 <u>SCAAGA</u> A R -repeat 990 <u>CAAAAA</u>
AAGATTGCCGATTTAGAA K I A D L E 820 CAAGCAAAACTAGCAGAA Q A K L A E 910 <u>SACCTTGAAGCTGTTCGG</u> D L E A V R	АЛСАЛЛАСТЛАЛА Т К L К 830 844 АЛЛАДЗТGАТСАЛ К D D Q 920 936 <u>ССЛЛЯСТАЛЛ</u> АЛА Q A К К	GAGGCAAAGG, E A K E 0 850 ATTAAACAAC' I K Q L 0 940 GCTACGGAAGG A T E A	AAGATTTTGA D F E B1-re 860 TAGAA <u>GAGCA</u> E E Q 950 CTGGAATTAAA E L N	ACTAGCAGCAT L A A L spaet 870 <u>AAAACAAAATCC'</u> K Q I L 960 CAACCTCAAAGC N L K A	FAGGTCACCAA G H Q 880 FAGATGCTAGC D A S 970 CAGAGCTTGCA E L A	CATGCTCATAAN H A H N B90 <u>CGTAAAGGTACA</u> R K G T B2 980 AAAGTTACA <u>GAA</u> K V T E	NGAGTAT E Y 900 IGCAAGA A R -repeat 990 ICCAAAAA Q K
AAGATTGCCGATTTAGAA K I A D L E 820 CAAGCAAAACTAGCAGAA Q A K L A E 910 <u>SACCTTGAAGCTGTTCGG</u> D L E A V R	AACAAAACTAAAA    T  K  L  K    B30  B4i    AAAAGATGATCAI  K  D  Q    920  93i  CAAAGCTAAAA  Q    CCAAGCTAAAAA  Q  A  K  K    Q  A  K  K  D  1026	GAGGCAAAGG, E A K E 0 850 ATTAAACAAC I K Q L 0 940 GCTACGGAAGG A T E A 1030	AAGATTTTGA D F E B1-re 8600 TAGAA <u>GAGCA</u> E E Q 950 CTGAATTAAA E L N 1040	ACTAGCAGCAT L A A L spact 870 <u>ANANCANATOC'</u> K Q I L 960 CAACCTCANAG N L K A 1050	TAGGTCACCAA G H Q 880 TAGATGCTAGC D A S 970 CAGAGCTTGCA E L A S 1060	CATGCTCATAAN H A H N B90 CGTAAAGGTACA R K G T B2 980 AAAGTTACAGAA K V T E 1070	NGAGTAT E Y 900 <u>GCAAGA</u> A R -repeat 990 <u>CCAAAAA</u> Q K 1080
AAGATTGCCGATTTAGAA K I A D L E 820 Q A K L A E 910 GACCTGAACCTGTACG D L E A V R 1000 10	АЛСАЛААСТААА. Т К L К 830 844 АЛААДАТСАТСАЛ К D D Q 920 933 <u>ССААССТААА</u> ААЛ <u>Q A K K</u> 010 1020	GAGGCAAAGG,    E  A  K  E    O  850    ATTAAACAAC  I  S    I  K  Q  L    O  940  GCTACGGAAGC  A    A  T  E  A  I  I    O  1030  GCAAGAGATC <sup>2</sup> A  R  D  L	ANGATTTTGA D F E B1-re 860 TAGAA <u>GACCA</u> E E Q 950 CTGAATTAAA E L N 1040 <u>TTGAAGCAGT</u> E A V	ACTAGCAGCAT L A A L spact 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAGC N L K A 1050 TCGCCAAGCAA	TAGGTCACCAA G H Q 880 TAGATGCTAGC D A S 970 CAGAGCTTGCA E L A 1060 AAGCACAAGTT	CATGCTCATAAN H A H N B90 CGTAAAGGTACA R K G T B2 980 ANAGTTACAGAA K V T E 1070 SAAGCTGCTCCC	NGAGTAT E Y 900 NGCAAGA A R -repeat 990 CCAAAAA Q K 1080 CAAACAA
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  Q  A  L  A  E  910  S    910  S  SACCTGAACCTGTAGCG  D  L  B  V  R    1000  III  CAAATCTTAGAACTATGCAGS  Q  I  L  D  A  S	АЛСАЛЛАСТАЛАІ Т К L К 830 844 АЛАЛАGАТGАТСАІ К D D Q 932 <u>ССАЛБСТАЛА</u> ЛАІ <u>Q A K K</u> 010 1024 <u>ССБТАЛАGБТАСІ</u> R K G T	GAGGCAAAGG,    E  A  K  E    0  850    ATTAAAAAAC    I  K  Q    940  940    GCTACGGAAGC    A  T  E    0  1030    GCAAGAGATCT  A  R  D    C  C  C	ANGATTTTGA D F E B1-re 860 TAGAA <u>GAGCA</u> E E Q 950 CTGAATTAAA E L N 1040 TTGAACCAGT E A V 1-repeat	ACTAGCAGCAT L A A L paet 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAGC N L K A 1050 TCGCCAAGCAAJ R Q A K	TAGGTCACCAA G H Q 880 TAGATGCTAGC D A S 970 CAGAGCTTCCA E L A 1 1060 AAGCACAAGTTT A Q V 1	CATGCTCATAAN H A H N 890 CGTAAAGGTACA R K G T 980 ANAGTTACAGAA K V T E 1070 SAAGCTGCTCTC E A A L	КАЗТАТ Е У 900 <u>IGCAAGA</u> A R -repeat 990 <u>IGCAAAAA</u> Q K 1080 <b>IORO</b> <b>IORO</b> <b>IORO</b> <b>IORO</b> <b>IORO</b>
ANGATTISCOGATTINGAL    K  I  A  D  L  E    820  I  C  B20  I  I  C    CANGCANANCTNOCAGAN  Q  A  K  L  A  E    910  S  GACCTTGANGCGGTTOCCO  D  L  E  A  V  R    1000  11  I  D  A  S  I  I  D  A  S    1000  11  I  D  A  S  I <td< td=""><td>AAACAAAACTAAAA    T  K  L  K    B30  B40  B40    AAAAGATGATGATGA  K  D  Q    920  933  G  CCAGGTAAAAA    Q  A  K  K    010  102  CCGTAAAGTACAT    R  K  G  T    100  1110  1110</td><td>GAGGCAAAGG,    E  A  K  E    0  850    ATTAAAAAAC  I    I  K  Q  L    0  940  GCTACGGAAGC  A    A  T  E  A  I  I    0  940  GCTACGGAAGC  GCAAGAGAGACC  I</td><td>ANGATTTTGA D F E B1-re 860 TAGAA<u>GAGCA</u> E E Q 950 CTGAATTAAA E L N 1040 <u>TTGAAGCAGT</u> E A V 1-repeat 1130</td><td>АСТЛЭСЛЭСАЭС L A A L spaet 870 <u>АЛАСЛАЛТСС'</u> K Q I L 960 САССТСАЛАЭС N L K A 1050 <u>ТСЭССЛАЭССАЛ</u> R Q A K 1140</td><td>TAGGTCACCAA G H Q 880 <u>PAGATGCTAGC</u> D A S 970 CAGAGCTTGCAI E L A 1060 AAGCACAAGTTT A Q V 1 1150</td><td>CATGCTCATAAN H A H N B90 CGTAAAGGTACA R K G T 980 ANGTTACAGAA K V T E 1070 GAAGCTGCTCTC E A A L 1160</td><td>Калатат Е У 900 КССЛАСА А R -repeat 990 КСЛАЛАА Q K 1080 КАЛАСАА К Q 1170</td></td<>	AAACAAAACTAAAA    T  K  L  K    B30  B40  B40    AAAAGATGATGATGA  K  D  Q    920  933  G  CCAGGTAAAAA    Q  A  K  K    010  102  CCGTAAAGTACAT    R  K  G  T    100  1110  1110	GAGGCAAAGG,    E  A  K  E    0  850    ATTAAAAAAC  I    I  K  Q  L    0  940  GCTACGGAAGC  A    A  T  E  A  I  I    0  940  GCTACGGAAGC  GCAAGAGAGACC  I	ANGATTTTGA D F E B1-re 860 TAGAA <u>GAGCA</u> E E Q 950 CTGAATTAAA E L N 1040 <u>TTGAAGCAGT</u> E A V 1-repeat 1130	АСТЛЭСЛЭСАЭС L A A L spaet 870 <u>АЛАСЛАЛТСС'</u> K Q I L 960 САССТСАЛАЭС N L K A 1050 <u>ТСЭССЛАЭССАЛ</u> R Q A K 1140	TAGGTCACCAA G H Q 880 <u>PAGATGCTAGC</u> D A S 970 CAGAGCTTGCAI E L A 1060 AAGCACAAGTTT A Q V 1 1150	CATGCTCATAAN H A H N B90 CGTAAAGGTACA R K G T 980 ANGTTACAGAA K V T E 1070 GAAGCTGCTCTC E A A L 1160	Калатат Е У 900 КССЛАСА А R -repeat 990 КСЛАЛАА Q K 1080 КАЛАСАА К Q 1170
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  Q  A  L  A  E  910  S    910  S  SACCTGAACCTGTAGCG  D  L  B  V  R    1000  III  CAAATCTTAGAACTATGCAGS  Q  I  L  D  A  S	AAACAAAACTAAA    T  K  L  K    B30  84  AAAAGATGATGATGATGATGATGATGATGATGATGATGAT	GAGGCAAAGG,    E  A  K  E    0  850    ATTAAACAAC:  I  K  Q    I  K  Q  L    0  940  GCTACGGAAGG    A  T  E  A    1030  GCAAGAGATC:  C    C  1120  C    ACCCGTAAGGC  AGCCGTAAGGC  C	ANGATTTTGA D F E B1-re 860 TAGAAGAGCA E E Q 950 CTGAATTAAA E L N 1040 <u>TTGAAGCAGT</u> E A V 1-repeat 1130 <u>STCTTCGCCG</u>	ACTAGCAGCAT L A A L spact 870 ANACANATCC K Q I L 960 CAACCTCANAGA N L K A 1050 TCGCCAAGCAAI R Q A K 1140	TAGGTCACCAA  G  H  Q    880  880  RAGATGCTAGC    D  A  S  S    970  S  S  S    970  CAGAGCTTGCAA  S  S    1060  10600  S  S    1060  A  Q  V  S    1150  CATCACGTGAAM  S  S  S	CATGCTCATAAN H A H N 890 CGTAAAGGTACA R K G T 980 ANAGTTACAGAA K V T E 1070 SAAGCTGCTCTC E A A L 1160 SCTAAGAAGCAA	КАЛБТАТ Е У 900 (GCAAGA A R -repeat 990 (CAAAAA Q K 1080 СААААА К Q 1170 (GTTGAA
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  Q  I  K  L  E    910  I  I  G  I  G  I	AAACAAAACTAAAA    T  K  L  K    B30  B44  AAAAGATGATGATGA    K  D  D  Q    920  933  G  G    920  933  CCAAGCTAAAAA  Q    Q  A  K  K    010  102  CCGTAAAGTACA  R    R  K  G  T    100  1110  SATTTCAGAAGCA  I    I  S  E  A	GAGGCAAAGG E A K E ) 850 ATTAAACAAC' I K Q L ) 940 GCTACGGAAGG A T E A ) 1030 GCAAGAGATC' A R D L 1120 AGCCGTAAGGG S R K G	ANGATTITGA D F E B1-re 860 TAGAAAGCA E E Q 950 CTGAATTAAN E L N 1040 TTGAAGCAGT E A V 1-repeat 1130 3TCTTCCCCCG	ACTAGCAGCAT L A A L spact 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAGCAAA R Q A K 1140 TGACTTGGACGA D L D A	Indicate  Product Access    G  H  Q    880  880  Resource    D  A  S  S    970  SAGAGCTTGCA  E  L  A    1060  Indicate  A  Q  V    1150  S  R  E  I	CATGCTCATAAN H A H N 890 <u>CGTAAAGGTACA</u> R K G T 980 NAGTACAG <u>AA</u> K V T E 1070 GAAGCTGCTCTC E A A L 1160 <u>CCTAAGGAGCCAA</u> A K K Q C2-repea	KGAGTAT E Y 900 (GCAAGA A R -repeat 990 (CAAAAA Q K 1080 CAAAAA K Q 1170 (GTTGAA V E 1
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  G  I  I    Q  A  K  L  A  E  I    910  G  G  GACCTTGAAGCTGTTCGG  I  I  I  I    D  L  E  A  V  R  I  I    GACCTTGAAGCTGTTGCGG  I  I  I  I  I  I  I    Q  L  E  A  V  R  I	AAACAAAACTAAAA    T  K  L  K    B30  84i    AAAAACATGATGATGATGATGATGATGATGATGATGATGATGATGA	GAGGCAAAGG,    E  A  K  E    0  850    ATTAAACAAC'  I  K  Q    I  K  Q  L    0  940  GCTACGGAACAC    A  T  E  A    0  1030  GCAAGAGATC'    A  R  D  L    C  1120  AGCCGTAAGGC    S  R  K  G    1210  1210  1210	ANGATTITGA D F E B1-re 860 TAGAAAGAACA E E Q 950 CTGAATTAAA E L N 1040 TTGAACCAET E A V 1-repeat 1130 37CTTCOCCCE L R R 1220	ACTAGCAGCAT L A A L spaet 870 AAAACAAATCC K Q I L 960 CAACCTCAAAGA N L K A 1050 TCGCCAAGCAA R Q A K 1140 TGACTTGGACGG D L D A 1230	Indicate Calc  Indicate Calc    G  H  Q    880  B80  Regard Calc    D  A  S    970  CAGAGCTIGCA  E    L  A  S    1060  Indicate Calc  Regard Calc    L  V  S    1150  CATCACGTESAA  S    S  R  E    1240  Indicate Calc  Indicate Calc	CATGGTGATGATAAA H A H N B90 CGTAAAGGTACA R K G T B2 980 AAAGTTACAGAA K V T E 1070 SIAAGGTGGTGGTGGTGGT E A A L 1160 CCTAAGGAAGGAA A K K Q C 2-re pea 1250	КАЛСТАТ Е У 900 (GCAAGA A R -repeat 990 (CAAAAA Q K 1080 (AAACAA K Q 1170 (GTTGAA V E 1260
ANGATTGCCGATTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  Q  A  L  A  E    910  S  S  I  I  B  V  R    1000  II  E  A  V  R  I	RARCANANCTANAL    T  K  L  K    830  840    ANANGATGATGATGA  K  D  Q    920  933  CCANGCTANANAN  Q  Q  S    Q  A  K  K  D  Q  Q  S  S  CCANGCTANANAN  Q  Q  X  K  K  D  D  Q  Q  X  K  K  D  D  Q  A  K  K  D  D  Q  A  K  K  D  D  Q  A  K  K  D  D  D  D  D  D  D  Q  A  K  K  D	GAGGCAAAGG E A K E ) 850 ATTAAACAAC: I K Q L ) 940 GCTACGGAAG A T E A ) 1030 GCAAGAGATC: A R D L C 1120 AGCCGTAAGG S R K G 1210 GATAAGGTTAA	ANGATTITGA D F E B1-re 860 TAGAA <u>GACCA</u> E E Q 950 CTGGAATTAAN E L N 1040 <u>TTGAACCAGT</u> E A V 1-repeat 1130 <u>STCTTCGCCCG</u> L R R 1220 MAGAACAANA	ACTAGCAGCAT L A A L spact 870 AAAACAAATCCC K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAGCAA R Q A K 1140 D L D A 1230 ACAAATCTCAG	IAGGTCACCAA  G  H  Q    B80  B80  Ratactage    D  A  S  S    970  SGAGCTTCCA  S  S    1060  IAGGACCAACTTCA  IA  S    1060  IAGGACCAACTTCA  S  S    1150  IATCACGTGAAS  S  R  E    1240  ICGCAAGCCGTCC  ICGCAAGCCGTC  ICGCAAGCCGTC	CATGCTCATCAAA H A H N B90 CGTAAAGGTACA R K G T 980 AAAGTTACAGAA K V T E 1070 SAAGCTGCTCTC E A A L 1160 SCTAAGAAGCAA A K K Q C2-repea 1250 CAAGCTCTCCCC	Image: Signal and Sig
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  CAAGCAAAACTAGCAGA.  Q  I  E  I  I  E    910  I  GACCTTGAAGCTGTTCGC  I	AAACAAAACTAAAA    T  K  L  K    B30  844    AAAAAGATGATGATGAT  K  D  Q    K  D  D  Q  33    CCAAGCTAAAA  Q  Q  34  K    Q  A  K  D  Q  30    Q  A  K  D  1020    CGTAAGCTAAAAA  G  T  1000  1111    BATTTCAGAAGC7  I  S  E  A    100  1  S  E  A    100  I  S  E  A    190  1200  30  30  30    SACTGCTGGAACTT  T  A  E  L	GAGGCAAAGG,    E  A  K  E    0  850    A  T  E    0  940    GCTACGGAAGGA  A  T  E    0  1030    GCAAGAGAATCT  A  R  D  L    0  1120  AGCCGTAAGGCAAGGY  S  R  K  G    1200  AGCCGTAAGGCAAGGY  S  R  K  G  1120    AGCCGTAAGGT  S  R  K  G  1210    GATAAGGTTAA  K  V  K  K  K	ANGATTITGA D F E B1-re 860 TAGAAGAAGACA E E Q 9500 CTGAATTAAA E L N 1040 TTGAACCAET E A V 1-repeat 1130 37CTTCOCCC L R R 1220 MAGAAGAAAA E E K	ACTAGCAGCAT L A A L spact 870 AAAACAAATCCC K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAGCAA R Q A K 1140 D L D A 1230 ACAAATCTCAG	IAGGTCACCAA  G  H  Q    B80  B80  Ratactage    D  A  S  S    970  SGAGCTTCCA  S  S    1060  IAGGACCAACTTCA  IA  S    1060  IAGGACCAACTTCA  S  S    1150  IATCACGTGAAS  S  R  E    1240  ICGCAAGCCGTCC  ICGCAAGCCGTC  ICGCAAGCCGTC	CATGCTCATCAAA H A H N B90 CGTAAAGGTACA R K G T 980 AAAGTTACAGAA K V T E 1070 SAAGCTGCTCTC E A A L 1160 SCTAAGAAGCAA A K K Q C2-repea 1250 CAAGCTCTCCCC	Image: Signal and Sig
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  C  AGCCANAACTAGCAGA.  Q  A  K  L  A  E    910  S  S  S  G  S  S  S  S  S  S  S  S  S  S  I	AACANAACTAAA    T  K  L    K  L  K    830  840    AAAAGATGATGATGA  K    B30  930    920  933    CCAAGCTAAAAA    Q  A    K  D    Q  A    K  G    CCAAGCTAAAAA    Q  A    K  G    T  K    G  T    SACTACAGAACA    I  S    I  S    I  S    I  S    SACTGCTGAACTGAACT    T  A    E  L    280  1290	GAGGCAAAGG E A K E B 850 ATTAAACAAC: I K Q L 940 GCTACGGAAGA A T E A 1030 GCAAGAGAATC: A R D L 0 1120 AGCCGTAAGGC 1210 GATAAGGTAA D K V K 1300	ANGATTITGA D F E B1-repeat F E E Q 950 CTGAATGAAA E L N 1040 TTGAAGCAGT L R R 1220 ANGAAGAAAA E E K 1310	ACTAGCAGCAT L A A L spact 870 ANAACAAATCC K Q I L 960 CAACTCAAAG N L K A 1050 <u>TCGCCAAGCAAA</u> R Q A K 1140 D L D A 1230 ACAAATCTCAGA Q I S D 1320	IAGGTCACCAA  G  H  Q    B80  B80  R  R    D  A  S  S    970  SGAGCTTCCAA  S  R    1060  IAGCACAAGTT  A  Q  V    1150  S  R  E  1240    CACGCAAGCGGTT  A  S  R  C    1150  S  R  E  I  S  R  C  I	CATGCTCATAAN H A H N B90 CGTANAGGTACA R K G T B2 980 ANAGTTACAGAN K V T E 1070 SAAGCTGCTGCTCC E A A L 1160 SCTAAGAAGCAM A K K Q C2-repea 1250 CAAGGTCTTCGC 2 G L R 1340	NGAGTAT    E    Y    900    IGCAAGA    A    -repeat    990    (CAAAA    Q    I080    INANCAA    K    Q    I170    STTGAA    I1260    COTGRAC    R    D    1350
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  C  AGCATAGCAAAACTAGCAGA.  Q  A  K  L  A  E    910  I  G  G  G  G  G  G  G  G  G  G  G  I  I  D  L  E  A  V  R  I	ANCANANCTANAL    T  K  L    T  K  L  K    830  840  ANANGATGATGATGA    K  D  D  Q    920  933  G  CONSCIENTANA    920  930  Q  A  K    010  1020  CONSCIENTANAGTACH  R  K    010  1021  CONSCIENTANGGANCH  R  K  G    100  1110  SACTGCTAGAGACH  I  S  I  S    190  1200  SACTGCTGAGTACHACHA  I  Z  S  S  I  S  I  S  Z  S  Z  S  Z	GAGGCAAAGG E A K E ) 850 ATTAAAAAAC: I K Q L ) 940 GCTACGGAAGG A T E A ) 1030 GCAAGGAATC: A R D L 1120 AGCCGTAAGGC S R K G 1210 GATAAGGTAA D K V K 1300 GTTGAAAAAC	ANGATTITGA D F E B1-re 860 TAGAAAGCA E E Q 950 CTGAATTAAN E L N 1040 TTGAAGCAGT E A V 1-repeat 1130 3TCTTCCCCC L R R 1220 MAGAAGAAAAA E E K K 1310 CTTTAGAGAA	ACTAGCAGCAT L A A L spact 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAAGCAA R Q A K 1140 TGACTTGGACGA D L D A 1230 Q I S D 1320 AGCAAACAGCAA	INGGTCACCAN  G  H  Q    880  880  Rest  Rest    D  A  S  S    O  A  S  S    CAGAGCTTGCAL  F  L  A    E  L  A  S  S    1060  INGCACAAGTT  A  Q  V  S    1150  INTCACGTAAC  S  R  E  I    1240  CGCAAGCCGTT  A  S  R  C    1330  INTAGCTGCTCCACH  INTAGCTGCTCCACH  S  R  C	CATGCTCATAAN H A H N 890 CGTAAAGGTACA R K G T 980 NAGTACAGAA K V T E 1070 SAAGCTGCTCCTC E A A L 1160 CCTAAGGAGCAA A K K Q C2-repea 1250 CAA <u>GGTCTTCGC</u> Q G L R 1340 CTTGAAAAACTT	КАЛСТАТ Е У 900 GCAAGA A R -repeat 990 CCANANA Q K 1080 CANANA K Q 1170 GTTGAA K Q 1170 GTTGAA S 1170 GTTGAA S 1170 GTTGAA I I 1170 GTTGAA I I I 1170 GTTGAA I I I I I I I I I I I I I
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  I  C  AGCCANAACTAGCAGA.  Q  A  K  L  A  E    910  S  S  S  G  S  S  S  S  S  S  S  S  S  S  I	ANCANANCTANAL    T  K  L    T  K  L  K    830  840  ANANGATGATGATGA    K  D  D  Q    920  933  G  CONSCIENTANA    920  930  Q  A  K    010  1020  CONSCIENTANAGTACH  R  K    010  1021  CONSCIENTANGGANCH  R  K  G    100  1110  SACTGCTAGAGACH  I  S  I  S    190  1200  SACTGCTGAGTACHACHA  I  Z  S  S  I  S  I  S  Z  S  Z  S  Z	GAGGCAAAGG E A K E ) 850 ATTAAAAAAC: I K Q L ) 940 GCTACGGAAGG A T E A ) 1030 GCAAGGAATC: A R D L 1120 AGCCGTAAGGC S R K G 1210 GATAAGGTAA D K V K 1300 GTTGAAAAAC	ANGATTITGA D F E B1-re 860 TAGAAAGCA E E Q 950 CTGAATTAAN E L N 1040 TTGAAGCAGT E A V 1-repeat 1130 3TCTTCCCCC L R R 1220 MAGAAGAAAAA E E K K 1310 CTTTAGAGAA	ACTAGCAGCAT L A A L spact 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAAGCAA R Q A K 1140 TGACTTGGACGA D L D A 1230 Q I S D 1320 AGCAAACAGCAA	INGGTCACCAN  G  H  Q    880  880  Rest  Rest    D  A  S  S    O  A  S  S    CAGAGCTTGCAL  F  L  A    E  L  A  S  S    1060  INGCACAAGTT  A  Q  V  S    1150  INTCACGTAAC  S  R  E  I    1240  CGCAAGCCGTT  A  S  R  C    1330  INTAGCTGCTCCACH  INTAGCTGCTCCACH  S  R  C	CATGCTCATAAN H A H N 890 CGTAAAGGTACA R K G T 980 NAGTACAGAA K V T E 1070 SAAGCTGCTCCTC E A A L 1160 CCTAAGGAGCAA A K K Q C2-repea 1250 CAA <u>GGTCTTCGC</u> Q G L R 1340 CTTGAAAAACTT	КАЛСТАТ Е У 900 GCAAGA A R -repeat 990 CCANANA Q K 1080 CANANA K Q 1170 GTTGAA K Q 1170 GTTGAA S 1170 GTTGAA S 1170 GTTGAA I I 1170 GTTGAA I I I 1170 GTTGAA I I I I I I I I I I I I I
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  CAAGCAAAACTACCAGA.  Q  A  K  L  A  E    910  S  GACCTTGAAGCGGTTGCG  D  L  E  910  S    GACCTTGAAGCGGTTGCG  D  L  E  A  V  R    1000  11  CAMATCTTAGATGCTAGG  Q  I  L  D  A  S    1090  12  CTGAAGAACAAAACAAG  L  E  Q  N  R    1180  11  A  N  N  L  1270  12    TTGGACCCATCACCTGAAGAGCAACACAGCAGCAGCAGCAGCAGCAGCAGCAGCA	ANCANANCTANAL    T  K  L    T  K  L  K    830  840  ANANGATGATGATGA    K  D  D  Q    920  933  G  K    920  933  Q  A  K    0  D  Q  Q  K  K    010  102  IO2  IO3  IO2  IO3    100  102  IO3  IO3  IO3  IO3    100  102  IO3  IO3  IO3  IO3    100  101  IO2  IO3  IO3  IO3    100  110  IO3  IO3  IO3  IO3    100  120  IO3  IO3	GAGGCAAAGG E A K E ) 850 ATTAAACAAC: I K Q L ) 940 (GCTACGGAAGG A T E A ) 1030 GCAAGGAACC: A R D L 1120 AGCCGTAAGGC S R K G 1210 AGCCGTAAGGC A 1210 GTTGAAAAACC V E K A 1390	ANGATTITGA D F E B1-re 860 TAGAAAGCA E E Q 950 CTGAATTAAN E L N 1040 TTGAAGCAGT E A V 1-repeat 1130 37CTTCCCCCC L R R 1220 MAGAAGAAAAA E E K 1310 CTTTAGAGAA L E E 1400	ACTAGCAGCAT L A A L spact 870 AAAACAAATCC' K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAACAA R Q A K 1140 TGACTTGGACGA Q I S D 1230 Q I S D 1320 ACAAAATCCCAGCAA A N S K 1410	INGGTCACCAN  G  H  Q    880  880  RAGATGCTAGC    D  A  S  S    970  SAGAGCTTGCAL  E  L  A    1060  INGCACAAGTTA  A  Q  V  S    1150  INFCACGTAACGTTA  S  R  E  I    1240  IS30  S  R  C  I    1330  IATTAGCTCCTC  L  A  I    1420  I  I  I  I	CATGCTCATAAA H A H N 890 CGTAAAGGTACA R K G T 92 980 ANGTTACAGAA K V T E 1070 SAAGCTGCTCCTC E A A L 1160 CCTAAGAAGCAA A K K Q C2-re pea 1250 CAAGGTCTTCGC Q G L R 1340 CTTGAAAAACTT L E K L	KGAGTAT E Y 900 GGCAAGA A R -repeat 9900 CANANA Q K 1080 ANACAA K Q 1170 GTTGAA Y E 11 1260 CGTGAA CGTGAA N K 1350 AACAAA N K
ANGATTGCCCATTAGA,    K  I  A  D  L  E    820  I  C  E  E  E    Q  A  K  L  A  E  E    910	AACAAAACTAAA    T  K  L    K  L  K    830  840    AAAAAGATGATGATGATGATGATGATGATGATGATGATGA	GAGGCAAAGG E A K E B 850 ATTAAACAAC I K Q L GCTACGGAAG A T E A 1030 GCAAGAGATC A R D L C 1120 GCAAGAGATC S R K G 1210 GATAAGGTAA C 1210 GATAAGGTAAAG C 1210 GATAAGGTAAAAG C 1300 C 13	ANGATTITGA D F E B1-re 860 TAGAAGAAGACA E E Q 950 CTGAATTAAA E L N 1040 TTGAACCACT 1130 37CTTCACCG L R R 1220 ANGANGANAN E E K 1310 TTTHAGAGAA	ACTAGCAGCAT L A A L spact 870 AAAACAAATCCT K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAGCAAA R Q A K 1140 ACAAAATCTGG Q I S D 1320 ACAAAACAGCAA A N S K 1410 MGCAAAACTTGJ	IAGGTCACCAA  G  H  Q    B80  880  R    D  A  S  S    970  CAGAGCTIGCAL  E  L  A    1060  LAGCACAAGTTI  A  Q  V  S    1150  A  Q  V  S  L  A  S  CAGCACAAGTTI  A  Q  V  S  L  A  S  CAGCAAAGTTICAA  S  C  L  A  S  C  C  L  A  S  C  L  L  A  S  C  C  L  A  S  C  L  L  A  S  C  L  L  A  L  L  A  L  L  L  A  L <t< td=""><td>CATGGTCATAAA H A H N B90 CGTAAAGGTACA R K G T B2 980 ANAGTACAGAA K V T E 1070 SAAGGTGGTCTCCC E A A L 1160 SCTAAGAAGCAA A K K Q C 2-re pea 1250 CAAGGTCTTCCC 2 G L R 1340 CTTGAAAAACTT L E K L 1430 ANAGCACTCAAA</td><td>NGAGTAT    E    Y    9000    IGCANGA    A    -repeat    990    ICANANA    Q    K    1080    IANACAN    K    Q    1170    GTTGGAA    V    E    1260    CGTGGAC    R    D    1350    ANACANA    N    1440    GAACAA</td></t<>	CATGGTCATAAA H A H N B90 CGTAAAGGTACA R K G T B2 980 ANAGTACAGAA K V T E 1070 SAAGGTGGTCTCCC E A A L 1160 SCTAAGAAGCAA A K K Q C 2-re pea 1250 CAAGGTCTTCCC 2 G L R 1340 CTTGAAAAACTT L E K L 1430 ANAGCACTCAAA	NGAGTAT    E    Y    9000    IGCANGA    A    -repeat    990    ICANANA    Q    K    1080    IANACAN    K    Q    1170    GTTGGAA    V    E    1260    CGTGGAC    R    D    1350    ANACANA    N    1440    GAACAA
ANGATTGCCGATTTAGA.    K  I  A  D  L  E    820  CAAGCAAAACTACCAGA.  Q  A  K  L  A  E    910  S  GACCTTGAAGCGGTTGCG  D  L  E  910  S    GACCTTGAAGCGGTTGCG  D  L  E  A  V  R    1000  11  CAMATCTTAGATGCTAGG  Q  I  L  D  A  S    1090  12  CTGAAGAACAAAACAAG  L  E  Q  N  R    1180  11  A  N  N  L  1270  12    TTGGACCCATCACCTGAAGAGCAACACAGCAGCAGCAGCAGCAGCAGCAGCAGCA	AACAAAACTAAA    T  K  L    K  L  K    830  840    AAAAAGATGATGATGATGATGATGATGATGATGATGATGA	GAGGCAAAGG E A K E B 850 ATTAAACAAC I K Q L GCTACGGAAG A T E A 1030 GCAAGAGATC A R D L C 1120 GCAAGAGATC S R K G 1210 GATAAGGTAA C 1210 GATAAGGTAAAG C 1210 GATAAGGTAAAAG C 1300 C 13	ANGATTITGA D F E B1-re 860 TAGAAGAAGACA E E Q 950 CTGAATTAAA E L N 1040 TTGAACCACT 1130 37CTTCACCG L R R 1220 ANGANGANAN E E K 1310 TTTHAGAGAA	ACTAGCAGCAT L A A L spact 870 AAAACAAATCCT K Q I L 960 CAACCTCAAAG N L K A 1050 TCGCCAAGCAAA R Q A K 1140 ACAAAATCTGG Q I S D 1320 ACAAAACAGCAA A N S K 1410 MGCAAAACTTGJ	IAGGTCACCAA  G  H  Q    B80  880  R    D  A  S  S    970  CAGAGCTIGCAL  E  L  A    1060  LAGCACAAGTTI  A  Q  V  S    1150  A  Q  V  S  L  A  S  CAGCACAAGTTI  A  Q  V  S  L  A  S  CAGCAAAGTTICAA  S  C  L  A  S  C  C  L  A  S  C  L  L  A  S  C  C  L  A  S  C  L  L  A  S  C  L  L  A  L  L  A  L  L  L  A  L <t< td=""><td>CATGGTCATAAA H A H N B90 CGTAAAGGTACA R K G T B2 980 ANAGTACAGAA K V T E 1070 SAAGGTGGTCTCCC E A A L 1160 SCTAAGAAGCAA A K K Q C 2-re pea 1250 CAAGGTCTTCCC 2 G L R 1340 CTTGAAAAACTT L E K L 1430 ANAGCACTCAAA</td><td>NGAGTAT    E    Y    9000    IGCANGA    A    -repeat    990    ICANANA    Q    K    1080    IANACAN    K    Q    1170    GTTGGAA    V    E    1260    CGTGGAC    R    D    1350    ANACANA    N    1440    GAACAA</td></t<>	CATGGTCATAAA H A H N B90 CGTAAAGGTACA R K G T B2 980 ANAGTACAGAA K V T E 1070 SAAGGTGGTCTCCC E A A L 1160 SCTAAGAAGCAA A K K Q C 2-re pea 1250 CAAGGTCTTCCC 2 G L R 1340 CTTGAAAAACTT L E K L 1430 ANAGCACTCAAA	NGAGTAT    E    Y    9000    IGCANGA    A    -repeat    990    ICANANA    Q    K    1080    IANACAN    K    Q    1170    GTTGGAA    V    E    1260    CGTGGAC    R    D    1350    ANACANA    N    1440    GAACAA

1450 1460 1465 TTAGCGAAACAAGCTGAAGAACTTG L A K Q A E E L

平成6年5月20日

Fig. 3 Comparison of the leader sequences or of the deduced signal peptide of M types 3 and various M genes. (A) Leader sequences are subdivided according to structural domains of the resulting leader peptide. (B) Signal peptides are divided into basic (B), hydrophobic (H) and cleavage (C) regions. Unknown sequences and amino acids are indicated by asterisks.

#### (A)

	region	Primer			
M3	*****				
M12	ATGGCTAAAAATACCACGAATAG				
M1	ATGGCTAAAAATAACACGAATAG				
M6	ATGGCTAAAAATAACACGAATAG				
M5	ATGGCTAGAGAAAATACCAATAA				
M24	ATGACTAAAAACAACACGAATAG				
M49	ATGGCTAGAAAAGATACGAATAA				
M2	ATGGCTAGAAAAGATACGAATAA	ACAGTATICGCTTAGAAAATT	AAAAAGAGGIAGA		
Hydrop	phobic region				
М3	GCTTCAGTAGCGGTTGCTTTGAC	AGTTTTAGGGACAGGACTGGT	AGCA		
M12	GCTTCAGTAGCGGTTGCTTTAAC				
M1	GCTTCAGTAGCGGTAGCTTTGAC	TGTTTTAGGGGCAGGTTTTGC	GAAT		
M6	GCATCAGTAGCAGTGGCTTTGAGTGTAATAGGGGCAGGATTAGTTGTC				
M5	GCATCAGTAGCAGTAGCTTTGAG				
M24	GCTTCAGTAGCGGTAGCTTTGAC				
M49	GCATCCGTAGCGGTCGCTGTGGC				
M2	GCATCCGTAGCAGTCGCTGTGGC	TGTTTTAGGAGCAGGCTTTGC	AAAC		
Cleava	ge region				
M3	GGGCAGACAGTAAAGGCA				
M12	GGGCAGACAGTAAGAGCA				
M1	CAAACAGAGGTTAAGGCT				
M6	AATACTAATGAAGTTAGTGCA				
M5	AATACTAATGAAGTTAGTGCA				
M24	AATACTAATGAAGTTAGTGCA				
M49	CAAACAGAAGTTAAGGCT				
м2	CAAACAACAGTTAAGGCG				
(B)	_		•		
	B		С		
M3 :	*********YSLRKLKTGT	ASVAVALTVLGTGLVA	GQTVKA		
M12	MAKNTTNRHYSLRKLKTGT	ASVAVALTVVGAGLVA	GQTVRA QTEVKA		
M1 :	MAKNNTNRHYSLRKLKTGT	ASVAVALTVLGAGFAN			
M6 :	MAKNNTNRHYSLRKLKKGT	ASVAVALSVIGAGLVV	NTNEVSA NTNEVSA		
M5 :	MARENTNKHYWLRKLKKGT	ASVAVALSVLGAGLVV	NTNEVSA		
M24		ASVAVALTVLGAGLVV	OTEVKA		
M49	and the second sec	ASVAVAVAVLGAGFAN	OTTVKA		
M2 :	MARKDTNKQYSLRKLKTGT	ASVAVAVAVLGAGFAN	WIIVKA		

Fig. 4 Comparison of C-repeat regions found among M3, M6, M5, M24 and M2 proteins. Identified amino acid residues with deduced M3 protein are indicated by colons.

#### C1-repeat regions

₩3	NRISEASRKGLRRDLDASREAKKQVEKDLANLTAELDKVKE
M6	NKVSEASRKGLRRDLDASREAKKQVEKDLANLTAELDKVKE
	: :: ::::::::::::::::::::::::::::::::::
M5	NKISDASRKGLRRDLDASREAKKQLEAEHQKLEEQ
	: :::::: ::::::::::::::::::::::::::::::
M24	NKISEASRQSLRRDLDASREAKKQLE
	:::::::::::::::::::::::::::::::::::::::
м2	Q I SEASRKSLRRDLEASRAAKKD
C2-re	epeat regions
112	
M3	EKQISDASRQGLRRDLDASREAKKQVEK

- M6 EKQISDASRQGLRRDLDASREAKKQVEKALEE
- M5 NKISEASRKGLRRDLDASREAKKQLEAEQQKLEEQ
- M24 NKISEASRQSLRRDLDASREAKKQVEKALEE
- M2 QISEASRKSLRRDLEASRAAKKD

emm3 gene from the C203 strain by PCR. We selected a pair of forward and reverse primers, one (K-1; 31 mer) from the best conserved portion in leader sequences of seven different strains, and the other (K-2; 30 mer) from the C-terminal conserved portion<sup>5,7)~14</sup>). PCR was performed with these two primers. The amplified product showed a single band approximately 1.4–1.5 kilobases (Kb) in length in agarose gel electrophoresis (Fig. 1-A) and was hybridized with the digoxigenin-labeled Msp I-Pvu II fragment of the emm6 gene (Fig. 1-B and C).

The DNA sequence of 1465 Kb of the amplified product was determined with a DNA sequencer and is shown in Fig. 2. The sequence indicated that the oligonucleotide sequences of the two primers existed in the 5' and 3' end (forward primer, 5' (CCGGGATCC) TATTCGCTTAGAAAATTAAAAA 3'; reverse primer, 5' (CCGGTCGA) CAAGTTCTTCAGCTTGTTTCGC 3'), and also encoded 488 amino acids which contained a leader sequence (1-32; defect first 9 amino acids), B repeat region (B1, 286-309; B2, 328-351) and C repeat region (C1, 374-391; C2, 416-433).

Comparison of the amplified product with other M protein genes: When the amplified product was compared with previously reported M protein genes<sup>7)~13)</sup>, the leader sequence or C repeat regions showed

702

			B1 repeat (100%)			
	261	271	281	291	301	
M3	AKEDFELAAL	GHQHAHNEYQ	AKLAEKDDQI	KQLEEQKQIL	DASRKGTARD	
	:: ::::::					
	265	275	285	295	305	
M12	AKKDFELAAL	GHQHAHNEYQ	AKLAEKDGQI	KQLEEQKQIL	DASRKGTARD	
			B2 r	epeat (91.7%	b)	
	311	321	331	341	351	
MЗ	LEAVRQAKKA	TEAELNNLKA	ELAKVTEQKQ	ILDASRKGTA	RDLEAVRQAK	
	315	325	335	345	355	
M12	LEAVRQAKKA	TEAELNNLKA	ELAKVTEQKQ	ILDASRKGTA	RDLEAVRKSK	
			C1 repea	t (94.5%)		
	361	371	381	391	401	
MЗ	AQVEAALKQL	EEQNRISEAS	RKGLRRDLDA	SREAKKQVEK	DLANLTAELD	
			::::::::		::::::::::	
	366	376	386	396	406	
M12	QQVEAALKQL	EEQNKISEAS	RKGLRRDLDT	SREAKKQVEK	DLANLTAELD	
			eat (100%)			
	411	421	431	441	451	
MЗ		ASRQ <u>GLRRDL</u>				
	:::::::::					
	416		436	446	456	
M12	KVKEEKQISD	ASRQ <u>GLRRDL</u>	DASREAKKQV	<u>EK</u> ALEEANSK	LAALEKLNKD	
	461	471	481			
M3 -		EKAELQAKLE				
		********		::::::		
	466	476	486			
M12	LEESKKLTEK	EKAELQAKLE	AEAKALKEQL	AKQAEEL		

Fig. 5 Comparison of homologous regions in M3 and M12 proteins, and relationship between B and C repeats in the predicted amino acid sequences of M3 and M12 proteins. The residues from each protein being compared are indicated by the numbering system of Robbins et al.<sup>12)</sup> for M12 protein. B repeat blocks and C repeat blocks are indicated by underlining, and identified amino acid residues are indicated by colons. The numbers enclosed in parentheses indicate % homology.

high homology with the other emm genes (Fig. 3 and 4). While the N-terminal amino acid portion of the amplified product was variable, it was found to be identical to 96 of 98 nucleotides downstream of the leader peptide sequence of another emm3 gene<sup>15)</sup>. Therefore, the amplified product was characterized as the emm3 gene.

Comparison of homologous regions in M3 and M12 proteins: When the region of 252–488 deduced amino acids in M3 protein was compared with the region of 256–355 and 357–493 deduced amino acids in M12 protein, 96.6% homology was found between them (Fig. 5). Interestingly, the B repeat region showed high homology with only that of M12 protein (91.7 and 100%, respectively). However, the A repeat region in M12 protein was not present in the M3 protein. Predictive secondary structure analysis of M3 protein: From analysis of the predictive secondary structure of the amplified product by the algorithm of Robson<sup>16)</sup>, most of the product was found to exhibit strong alpha-helical potential. In addition, the beta-sheet and turn potential seen for region 23 to 42 in the M protein was similar to that seen for region 28 to 50 in M12 protein. The results suggest that M3 protein may be closely related to M12 protein.

## Discussion

DNA sequence analysis has made it clear that all M proteins studied to date<sup>7)~12)</sup> are structurally related and are therefore encoded by a family of genes. The regions of amino acid sequence homology in the protein include the signal sequences, the C repeat region in the central part of the protein chain and the carboxyl-terminal part. On the basis of available genetic information, we cloned the emm3 gene by using PCR and sequenced its DNA. The primers prepared originated from the best conserved leader sequence or the C-terminal portion of the emm genes<sup>7)~12)</sup>. The amplified product was hybridized with an emm6 gene probe (Fig. 1). Sequence analysis of the amplified product identified sequences complementary to both oligonucleotide primers (Fig. 2). In addition, the product had both B and C repeat regions. By comparing the amplified product with known emm genes, we found not only that the N-terminal portion is very variable, but also that the C-terminal region is conserved<sup>7)~12),14)</sup>.

Several streptococcal immunoglobulin-binding proteins have also been characterized as members of the M protein family and as having C repeats<sup>17,18</sup>. Our selected primers have similar regions which can be

#### Chihiro KATSUKAWA

screened. However, our product amplified by PCR from the type 3 strain C203 of *S. pyogenes* was a single DNA and had no homology with amino terminal regions of the streptococcal immunoglobulin-binding proteins. Furthermore, 96 of 98 nucleotides downstream of the leader peptide sequence of the amplified product were found to be identical to the corresponding sequence of another emm3 gene of type 3 M strain 3-3/317<sup>15</sup>). The evidence shows that the amplified product is the emm3 gene.

B and C repeat blocks that exist in M3 protein are similar to those in M12 protein (Fig. 5). Furthermore, predictive secondary structure analysis of M3 protein revealed that the majority of the products exhibit strong alpha-helical potential as found with other M protein structures<sup>19</sup>. The algorithm also showed that region 23-42 exhibits beta-sheet and turn potential with a pattern similar to that for region 28-50 found by predictive secondary analysis of M12 protein.

*S. pyogenes* can be divided into two major classes on the basis of their immune reactivity with monoclonal antibodies (mAbs) directed against epitopes which lie within the conserved half of M proteins<sup>20)</sup>. Class I serotype are defined as those which bind their mAbs, whereas class II isolates do not. Mainly the class I-specific mAb binding sites map to a region of C repeats within M proteins. Inasmuch as the C repeat region of our emm 3 gene represents more than 90% homology with the known emm genes, it belongs to the class I serotype. This agrees with the report of Bessen et al. who decided that M type 3 *S. pyogenes* had a class I protein<sup>20)</sup>. Furthermore, we found similarity between emm 3 and emm 12 genes in their B repeat regions and predictive secondary structure. Thus, there may exist a subclass of class I M proteins. Bessen et al.<sup>20)</sup> discriminated between serotypes sharing both B and C repeat region epitopes and those sharing only C repeat region epitopes by using only antibody probes directed to antigenic sites within the B and C repeat regions of the M protein molecules in class I serotype. We suggest that M3 and M12 proteins belong to a subclass of class I M proteins.

## Acknowledgements

I am grateful to Dr. Kyonsgu Hong of the Department of Bacteriology, Osaka University Medical School, Dr. Ikuya Yano of the Department of Bacteriology, Osaka City University Medical School, Dr. Masanao Makino and Dr. Yasuhiko Suzuki of Osaka Prefectural Institute of Public Health for technical advice and helpful discussions. In addition, I thank Dr. June R. Scott (Emory University) for providing the pUC19: M6 plasmid and Dr. Tatsuya Tanaka of the Central Laboratoey for Research and Education, Osaka University, Faculty of Medicine, for synthesizing the oligonucleotides. This work was supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan (C63570192).

## References

- 1) Bisno, A.: Group A streptococcal infections and acute rheumatic fever. N. Engl. J. Med. 325: 783-793, 1991.
- Lancefield, R.C.: Current knowledge of the type-specific M antigens of group A streptococci. J. Immunol. 89: 307-313, 1962.
- 3) Beachey, E.H., Seyer, J.M., Dale, J.B., Simpson, W.A. & Kang, A.H.: Type-specific protective immunity evoked by synthetic peptide of *Streptococcus pyogenes* M protein. Nature (London) 292: 457–459, 1981.
- 4) Hong, K., Kinoshita, T., Takeda, J., Kozono, H., Pramoonjago, P., Kim, Y.U. & Inoue, K.: Inhibition of the alternative C3 convertase and classical C5 convertase of complement by group A streptococcal M protein. Infect. Immun. 58: 2535-2541, 1990.
- 5) Scott, J.R., Pulliam, W.M., Hollingshead, S.K. & Fischetti, V.A.: Relationship of M protein genes in group A streptococci. Proc. Natl. Acad. Sci. USA 82: 1822–1826, 1985.
- Murry, M.G. & Thompson, W.F.: Rapid isolation of high-molecular-weight plant DNA. Nucleic Acids Res. 8: 4321– 4325, 1988.
- 7) Bessen, D.E. & Fischetti, V.A.: Nucleotide sequence 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. Infect. Immun. 60: 124-135, 1992.
- 8) Haanes, E.J. & Cleary, P.P.: Identification of a divergent M protein gene and an M protein-related gene family in

Streptococcus pyogenes serotype 49. J. Bacteriol. 171: 6397-6408, 1989.

- Hollingshead, S.K., Fischetti, V.A. & Scott, J.R.: Complete nucleotide sequence of type 6 M protein of the group A streptococcus: Repetitive structure and membrane anchor. J. Biol. Chem. 261: 1677–1686, 1986.
- 10) Miller, L., Gray, L., Beachey, E. & Kehoe, M.: Antigenic variation among group A streptococcal M proteins: Nucleotide sequence of the serotype 5 M protein gene and its relationship with genes encoding type 6 and 24 M proteins. J. Biol. Chem. 263: 5668-5673, 1988.
- Mouw, A.R., Beachey, E.H. & Burdett, V.: 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, 1988.
- 12) Robbins, J.C., Spanier, J.G., Jones, S.J., Simpson, W.J. & Cleary, P.P.: *Streptococcus pyogenes* type 12 M protein gene regulation by upstream sequences, J. Bacteriol. 169: 5633-5640, 1987.
- 13) Haanes-Fritz, E., Kraus, W., Burdett, V., Dale, J.B. & Beachey, E.H.: Comparison of the leader sequences of four group A streptococcal M protein genes. Nucleic Acids Res. 16: 4667-4677, 1988.
- 14) Hollingshead, S.K., Fischetti, V.A. & Scott, J.R.: A highly conserved region present in transcripts encoding heterologous M proteins of group A streptococci. Infect. Immun. 55: 3237-3239, 1987.
- 15) Podbielski, A., Baird, R. & Kaufhold, A.: The group A streptococcal M-type 3 protein gene exhibits a C terminus typical for class I M proteins. Med. Microbiol. Immunol. 181: 209–213, 1992.
- 16) Garnier, J., Osguthorpe, D.J. & Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120: 97-120, 1978.
- 17) Gomi, H., Hozumi, T., Hattori, S., Tagawa, C., Kishimoto, F. & Björck, L.: The gene sequence and some properties of protein H. J. Immunol. 144: 4046-4052, 1990.
- 18) Heath, D.G. & Cleary, P.P.: Fc-receptor and M-protein genes of group A streptococci are products of gene duplication. Proc. Natl. Acad. Sci. USA 86: 4741-4745, 1989.
- 19) Phillips, G.N., Flicker, P.F., Cohen, C., Manjula, B.N. & Fischetti, V.A.: Streptococcal M protein: α-helical coiled-coil structure and arrangement on the cell surface. Proc. Natl. Acad. Sci. USA 78: 4689–4693, 1981.
- 20) Bessen, D., Jones, K.F. & Fischetti, V.A.: Evidence for two distinct classes of streptococcal M protein and their relationship to rheumatic fever. J. Exp. Med. 169: 269-283, 1989.

PCR 法を用いてクローニングした A 群溶血レンサ球菌

M3蛋白遺伝子の解析および他菌型との比較

(レンサ球菌感染症研究会:会長 保科 清)大阪府立公衆衛生研究所

勝川千尋

(平成5年11月12日受付)(平成6年2月7日受理)

### 要 旨

A 群溶血レンサ球菌 M3蛋白遺伝子の N 末端 の高度型特異領域から C 未端の保存領域までの 部分を PCR 法を用いてクローニングを行った.

各菌型に共通なN末端のリーダーシクエンス 部分とC末端の保存領域部分をプライマーとし て用い,1465bpの遺伝子配列を決定し,他のM蛋 白遺伝子と比較検討した。その結果,N末端側の 100塩基から750塩基の範囲にM3型特異的な領域 を見いだすことができた。また、アミノ酸配列を 検討したところ, 2つの繰り返し配列を見いだし た(BリピートおよびCリピート).Cリピートは 現在知られている他の M 蛋白遺伝子の塩基配列 と非常に高い相同性を示した.これに対して,Bリ ピートは M12蛋白遺伝子の Bリピート配列との み高い相同性を示し,また二次構造の解析結果で もこの2菌型は構造が類似していた.これらの結 果より M3蛋白と M12蛋白は遺伝学的に非常に近 い蛋白であることが示唆された.

平成6年5月20日