JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 1996, p. 953–958 Vol. 34, No. 4
0095-1137/96/$04.00+0

Sequencing *emm*-Specific PCR Products for Routine and Accurate Typing of Group A Streptococci

BERNARD BEALL,* RICHARD FACKLAM,* AND TERRY THOMPSON

Centers for Disease Control and Prevention, Atlanta, Georgia

Received 23 October 1995/Returned for modification 5 January 1996/Accepted 22 January 1996

Rapid sequence analysis of specific PCR products was used to accurately deduce *emm* types corresponding to the majority of the known group A streptococcal (GAS) M serotypes. The study involved 95 M type reference GAS strains and a survey of 74 recent clinical isolates. A high percentage of agreement between M type serology and the previously published 5′ sequences of the *emm* genes of M type reference strains was noted. The 5′ sequences for six established M protein genes—the *emm-32*, *emm-34*, *emm-38*, *emm-40*, *emm-42*, and *emm-71* genes—were determined to supplement the existing *emm* sequence database. Rapid sequence analysis differentiated serologically M-nontypeable strains and was used to establish the probable clonal relationship between seven GAS isolates from one hospital outbreak.

The streptococcal group A *emm*-like genes are clustered together at the *vir* locus of the chromosome and are divided into the *mrp*, *emm*, and *enn* gene groups on the basis of differences in conserved 5′ and 3′ regions and relative positions within this region (reviewed in reference 9). The *emm* gene of the group A streptococcus (GAS) *Streptococcus pyogenes* encodes a major virulence factor of these important pathogens, the M protein. The surface-exposed amino termini of M proteins are heterogenous and appear to provide the basis for identifying more than 80 different serologic M types (1, 4, 7, 10, 12), although the contribution of other M-like proteins to M-type specificity has not been investigated (9).

Currently, identification of clinical isolates of GAS for surveillance and other epidemiologic studies often relies primarily on serologic typing of the surface M protein with available polyclonal sera. However, it is frequently difficult to detect M proteins in this way because typing reagents are not widely available and are difficult to prepare. It is also believed that many GAS isolates are nontypeable because of the lack of M expression or lack of reactivity of expressed M protein with available antisera (11). Additionally, for some isolates untypeability is undoubtedly due to the expression of a new M protein. Consequently, there is a need to correlate an alternative means of M type deduction with current serologic M typing.

The variable 5′ sequences of the majority of the *emm* genes which confer established distinguishable M serotypes to group A streptococci (GAS) have been previously determined by other laboratories. Recent work has shown that many different M serotypes could be correlated with hybridization to *emm* allele-specific oligonucleotides (8). While this technology is useful, it relies on oligonucleotides specific to known *emm* alleles, will not allow identification of new *emm* genes, and would be tedious for the identification of rarely occurring *emm* genes. In addition, such an identification system may not allow for the differentiation of potential hybrid *emm* genes that apparently arise through interstrain gene transfer (20). It is also questionable that for each known *emm* gene a given oligonucleotide probe of 30 bases would be useful in detecting all *emm* alleles conferring a given polyclonal serum-based M type (16). In this study we found that an oligonucleotide primer pair used in previous studies (19, 20) allowed amplification of the *emm* alleles of all of our available M-type reference strains and 74 of 77 GAS strains from a survey of recent clinical isolates. We show here that the use of one of these PCR primers for cycle sequencing these *emm*-specific PCR products is a practical and useful tool for subtyping clinical GAS isolates.

* Corresponding author. Mailing address: Childhood and Respiratory Diseases Branch, Division of Bacterial and Mycotic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Mailstop C02, 1600 Clifton Rd., NE, Atlanta, GA 30333. Phone for Bernard Beall: (404) 639-1237. Fax: (404) 639-3123. Electronic mail address: beb0@ciddbd2.em.cdc.gov. Phone for Richard Facklam: (404) 639-1379. Fax: (404) 639-3123.

MATERIALS AND METHODS

M type reference strains. The Centers for Disease Control and Prevention (CDC) identification numbers for the M typing reference strains are shown in Table 1. All reference strains for given M types were independent original isolates or multiply passaged derivatives of these original isolates. We used Lancefield type reference strains for M serotypes 1 through 51. Reference strains representing M types 52 through 81 and pt180 were received from individual investigators as provisional new M types over the years 1960 to 1978.

All clinical survey strains were 1995 isolates from normally sterile infection sites within a wide range of patients. All strains were taken in a random survey from throughout the United States, except for the 1995 Denmark isolates 4090d, 4093d, 4092, and 4094. Strains were confirmed as GAS by capillary precipitin grouping using CDC-prepared antiserum and Lancefield extraction procedures (13).

Serologic M types, T types, and opacity factor (OF) reactions were determined as previously described (13, 14).

PCR. The GAS were grown overnight on trypticase soy agar supplemented with 5% sheep blood (Becton Dickinson Inc., Cockeysville, Md.). One loopful of growth was resuspended in 300 μl of 0.8% NaCl and heated for 30 min at 60°C. Cells were centrifuged and resuspended in 100 μl of TE (10 mM Tris, 1 mM EDTA [pH 8]) containing 300 U of mutanolysin (Sigma, St. Louis, Mo.) per ml and 30 μg of hyaluronidase (Sigma) per ml for 30 min at 37°C. Samples were then heated at 100°C for 10 min and briefly centrifuged to pellet debris. Two microliters of supernatant was then used as template for each 100-μl PCR mixture.

Primers 1 and 2 were used in PCRs as previously described (20). PCR products were purified with Wizard columns (Promega, Madison, Wis.) as described by the manufacturer.

Sequence analysis. Approximately 60 ng of up to 36 PCR products was sequenced by using primer 1 with the dye terminator mix (Applied Biosystems, Foster City, Calif.) and subjected to automated sequence analysis on a 373 autosequenator (Applied Biosystems) as described by the manufacturer. The cycling parameters were 96°C for 30 s, 46.5°C for 1 s, and 60°C for 4 min. DNA sequences were subjected to homology searches against the bacterial DNA database with GCG software (Wisconsin package, version 8). Sequences for new *emm* gene GenBank entries were obtained from one strand of PCR fragments as described above, except that an average of three independent sequencing runs were used to compile the sequence. Sequences from this study that have not been submitted to GenBank are available upon request.

TABLE 1. Comparisons of 5′ *emm* allele sequences from CDC reference GAS M type strains with closest matches in GenBank

CDC strain[a]	Closest match(es)	% Identity[b]	Accession no.
745 (M1, T1, OF−)	*emm-1*, *emm-68*	98, 98	U11940, U11997
633 (M2, T2, OF+)	*emm-2*	97	U11958
1027 (M3, T3/13, OF−)	*emm-3*	96	U11945
90 (M3, T3, OF−)	*emm-3*	96	U11945
91 (M4, T4/8/14, OF+)	*emm-4*	98	X15198
470 (M4, T4, OF+)	*emm-4*	96	X15198
746 (M5, T5/27/44, OF−)	*emm-5*	95	M20374
93 (M6, T6, OF−)	*emm-6*	97	M11338
736 (M6, NT, OF−)	*emm-6*	96	M11338
412 (M8, T8/25/IMP, OF+)	*emm-8*	98	U12005
634 (M8, T8/14/25/IMP, OF+)	*emm-8*	100	U12005
650 (M9, T9/8/14, OF+)	*emm-9*	100	U12002
754 (M9, T9, OF+)	*emm-9*	97	U12002
68 (M11, T11, OF+)	*emm-11*	98	U11938
721 (M11, T11, OF+)	*emm-11*	96	U11938
392 (M12, T11, OF−)	*emm-12*	95	U11937
635 (M12, T12, OF−)	*emm-12*	99	U11937
622 (M14, T14/9, OF−)	*emm-14*	97	U11935
800 (M15, T23/8/14, OF−)	*emm-15*	98	U11934
637 (M17, T23/8/14, OF−)	*emm-17*	98	U11932
36 (M18, T9, OF−)	*emm-18*	96	S82057
101 (M19, NT, OF−)	*emm-19*	95	U11959
756 (M22, T22, OF+)	*emm-22*	98	U11955
638 (M22, T22, OF+)	*emm-22*	98	U11955
730 (M23, T23, OF−)	*emm-23*	98	U11953
987 (M24, T-NT, OF−)	*emm-24*	100	M19031
639 (M25, T8/25, OF+)	*emm-25*	98	U11952
728 (M26, T-ND, OF−)	*emm-26*	96	U11951
789 (M28, T28, OF+)	*emm-28*	97	U11948
988 (M29, NT, OF−)	*emm-29*	96	U11946
109 (M30, NT, OF−)	*emm-30*	97	U11944
901 (M31, NT, OF−)	*emm-31*	96	U11943
878 (M33, T3/13/B, OF−)	*emm-33*	100	U11942
873 (M36, T9, OF ND)	*emm-36*	99	U11941
53 (M37, NT, OF ND)	*emm-37*	99	U11970
731 (M39, NT, OF−)	*emm-39*	95	U11968
795 (M41, T3/13/B, OF−)	*emm-41*	97	U11967
380 (M43, NT, OF ND)	*emm-43*	99	U11965
511 (M44, T5/27/44, OF+)	*emm-44*, *emm-61*	96, 96	U11964, U11984
536 (M44, T5/12/27/44, OF+)	*emm-44*, *emm-61*	99, 99	U11964, U11984
642 (M46, T4, OF−)	*emm-46*	99	U11963
116 (M47, T23, OF−)	*emm-47*	96	U11962
737 (M48, T28, OF+)	*emm-48*	96	U11961
702 (M49, T14, OF+)	*emm-49*	96	S44880
456 (M50, NT, OF−)	*emm-50*, *emm-62*	99	U02465, U11983
643 (M51, T14, OF−)	*emm-51*	98	U11977
686 (M52, T3/13/B, OF)	*emm-52*	100	L27098
977 (M53, T1/13/B, OF−)	*emm-53*, *emm-67*	98, 98	L27099, U11998
725 (M54, T ND, OF−)	*emm-54*	95	U11974
934 (M55, T8/14/25/IMP)	*emm-55*	99	U11973
822 (M56, T3/13/28, OF−)	*emm-56*	100	U11972
790 (M57, T8/25, OF−)	*emm-57*	100	U11971
798 (M57, T8/25/Imp19, OF−)	*emm-57*	100	U11971
872 (M58, T8/14/25, OF+)	*emm-58*	96	U11988
966 (M58, T25, OF+)	*emm-58*	97	U11988
913 (M59, TIMP, OF+)	*emm-59*	100	U11987
874 (M60, T6w?, OF+)	*emm-60*	99	U11985
875 (M61, T11, OF+)	*emm-61*, *emm-44*	96, 97	U11984, U11964
984 (M62, T12, OF+)	*emm-62*, *emm-50*	96	U11983, U02465
985 (M63, T4, OF+)	*emm-63*	100	U11982
989 (M64, T8/14, OF−)	*emm-64*	99	U11981
1042 (M65, T8/14, OF−)	*emm-65*	98	U11980
1037 (M66, T12, OF+)	*emm-66*	95	U11999
1144 (M72, T12, OF−)	*emm-72*	96	U11982
1145 (M73, T13, OF+)	*emm-73*	99	U11995
1146 (M74, T9, OF−)	*emm-74*	95	U11994
965 (M75, T25, OF+)	*emm-75*	95	U11993
1147 (M75, T25, OF+)	*emm-75*	95	U11993
1148 (M76, T12, OF+)	*emm-76*	97	U11992
1149 (M77, T13, OF+)	*emm-77*	96	U11991
1150 (M78, T11, OF+)	*emm-78*	97	U11990
1152 (M80, T14, OF−)	*emm-80*	99	L27097
1173 (M81, NT, OF+)	*emm-81*	96	U12003
787 (MPT180, T5/27/44, OF+)	*emmPT180*	96	U11960

[a] ND, not done; NT, nontypeable.
[b] Of the first 160 bases of the reference strain *emm* sequence with the indicated GenBank sequence.

TABLE 2. Comparisons of M type reference strain 5′ *emm* sequences from this study with GenBank sequences

CDC strain	Closest match	% Identity[a]	New accession no.
48 (M32, T8/14/23)	*emm-36* (U11941)	85	L47325 (*emm-32*)
784 (M34, T3/13/28)	*enn-64/-14* (X72753)	98	L47324 (*emm-34*)
722 (M40, T NT,[b] OF−)	*emm-5* (M20374)	76	L46817 (*emm-40*)
138 (M40, T NT, OF−)	*emm-5* (M20374)	77	L46817 (*emm-40*)
306 (M38, T NT, OF−)	*emm-5* (M20374)	77	L46817 (*emm-40*)
54 (M38, T NT, OF−)	*emm-5* (M20374)	77	L46817 (*emm-40*)
641 (M42)	*emm-11* (X74138)	75	L46799 (*emm-42*)
1098 (M71, T28, OF−)	*emm-30* (U11944)	81	L46652 (*emm-71*)

[a] Of the first 160 bases of the reference strain *emm* sequence with the indicated GenBank sequence closest match.
[b] NT, nontypeable.

RESULTS

CDC reference strain *emm* designations and GenBank entry designations. The majority of our M reference strain 5′ *emm* sequences agreed with GenBank sequence entries with the same designations. For this study, we sequenced the 5′ *emm* sequences of one to five reference strains corresponding to 75 recognized M serotypes. In most instances strains of a given M type were of independent origin (not shown). The *emm* sequences for six of these M types had not been previously entered into the GenBank and were obtained in this study (Table 2). For 62 of these M types, 5′ *emm* sequences were found from one to three strains that had ≥95% identity over 160 bases with corresponding *emm* genes in GenBank (Table 1). Six other M types including types M34, M27, M69, M70, and M79 had *emm* sequences highly similar to *emm* or *emm*-like genes of different designations (Tables 2 and 4). For most of these *emm* genes, sequences of approximately 190 to 330 bases were actually obtained, depending upon the quality of a given sequencing run, and in no case did longer sequence comparison lead to a significant lowering of these identity values (data not shown).

Using *emm* typing for recent clinical isolates. We were able to successfully amplify *emm* sequences for determining the *emm* types of 74 of 77 recent clinical sterile-site GAS isolates taken randomly from a large collection for this study (Table 3). It is striking that the *emm* sequences of each of 69 of these strains were clearly matched to one of 25 different *emm* (or *emm*-like) genes previously sequenced and included in the database (Table 3). Each of 37 M-typeable clinical GAS strains randomly taken from our clinical collection had M types and *emm* sequences that correlated with 95 to 100% identity over 160 bases to one of 10 previously sequenced *emm* genes associated with their specific M types (Table 3). Sequences of approximately 200 to 500 bases were actually obtained from most of these strains with no significant lowering of these identity values (data not shown).

We defined as M nontypeable those strains that had *emm* sequences that correlated to known *emm* serotype genes with ≥95% sequence identity over at least the first 160 bases yet still did not react with any specific M-typing sera. The reason that these eight strains in the clinical survey were nontypeable is unknown. From the data presented in Tables 1 and 3, which show a perfect correlation between M types and *emm* types, we believe it is likely that the *emm* types shown in Table 3 that were found for nontypeable strains and strains for which no typing sera are available correspond to specific M specificities. This is further supported by the results obtained with the seven *emm-53* strains, six of which had identical T types (Table 3) and were isolated from the same hospital outbreak (data not shown). Three strains were M nontypeable, but all seven of their *emm* gene sequences were nearly identical to the GenBank *emm-53* entry, which in this case indicated a likely clonal origin (Table 3).

In the 8 nontypeable clinical isolates and 23 clinical isolates shown in Table 3 for which specific typing sera were not available the 5′ *emm* sequences that corresponded well to known M serotypes were in almost total agreement with the known general correlates of OF phenotype and T-antigen types to M specificities (3, 5, 13). Additionally, we found potentially strain-specific tags, as some sequence differences within *emm* alleles could allow differentiation of strains within a given M type (see the M27 strain in Table 3 and Discussion).

The utility of this method with clinical isolates carrying previously unsequenced *emm* genes was also evident. Identical 5′ *emm* sequences were found for two pairs of strains and one individual strain (the last three entries in Table 3). Together, these sequences clearly represent three new *emm* genes since

TABLE 3. Similarity of 5′ *emm* sequences of recent clinical GAS isolates to *emm* sequences in GenBank

Characteristics[a] (*n*)	Closest match(es)	% Identity[b]	Accession no.
M1, T1, OF− (11)	*emm-1*, *emm-68*	98–100	U11940
		98–100	U11997
M NT, T2, OF+ (1)	*emm-2*	95	U11958
M3, T3/13/B, OF− (7)	*emm-3*	97–99	U11945
M NT, T4, OF+ (1)	*emm-4*	99	X15198
M5, T5/27/44, OF− (2)	*emm-5*	96	M20374
M6, T6, OF− (1)	*emm-6*	98	M11338
M6, T NT, OF− (2)	*emm-6*	98–99	M11338
M NT, T6, OF− (1)	*emm-6*	97	M11338
M12, T12, OF− (3)	*emm-12*	97–99	U11937
M NT, T12, OF− (1)	*emm-12*	100	U11937
M18, T NT, OF− (2)	*emm-18*	99	S82057
M NS, T12, OF− (3)	*emm-22*	100	U11955
M27, T5/27/44, OF− (1)	*emm-27*	99[c]	U11949
M NS, T28, OF+ (2)	*emm-28*	99	U11948
M NT, T3/13/B, OF− (1)	*emm-33*	98	U11942
M41, T NT, OF− (1)	*emm-41*	100	U11967
M53, T3/13/B, OF− (4)	*emm-53*	98–99	L27099
M NT, T3/13/B, OF− (2)	*emm-53*	99–100	L27099
M NT, T NT, OF− (1)	*emm-53*	98	L27099
M NS, T NT, OF+ (1)	*emm-58*	98	U11988
M NS, T11/12, OF+ (2)	*emm-59*	99	U11987
M NS, T11/12, OF− (1)	*emm-66*	100	U11999
M NS, T13, OF+ (1)	*emm-73*	99	U11995
M NS, T8/25, OF+ (3)	*emm-76*	98–100	U11992
M NS, T13, OF+ (2)	*emm-77*	99–100	U11991
T NS, T NT, OF− (1)	*emm-80*	98	L27097
M NS, T6, OF− (3)	*emmpt64/4*	98–99	X72932
M NS, T NT, OF− (1)	*emmpt64/4*	99	X72932
M NS, T12, OF+ (2)	*emmpt4245*	98–100	U11966
M NS, T25, OF− (1)	*enn63*	95	U20842
M NS, T3, OF− (1)	*emmpt87/156*	98	L27096
M NA, T NT, OF+ (2)	*emm-73*	80[d]	U11995
M NA, TImp19, OF− (2)	*emm-27*	76[d]	U11949
M NA, T-NT, OF− (1)	*emm-32*	88[d]	L47325

[a] ND, not done; NA, nontypeable since no typing serum was made for this deduced new M type; NT, nontypeable for reasons unknown; NS, no specific typing sera available.
[b] Of the first 160 bases of the reference strain *emm* sequence with respect to the indicated GenBank sequence.
[c] This value does not account for a deletion of one of 2 nearly perfectly conserved direct 7 codon repeats contained in the GenBank *emm27* sequence.
[d] The 5′ *emm* sequence is not ≥90% identical over first 160 bases sequenced to any *emm* sequence in GenBank and is believed to represent a new *emm* gene.

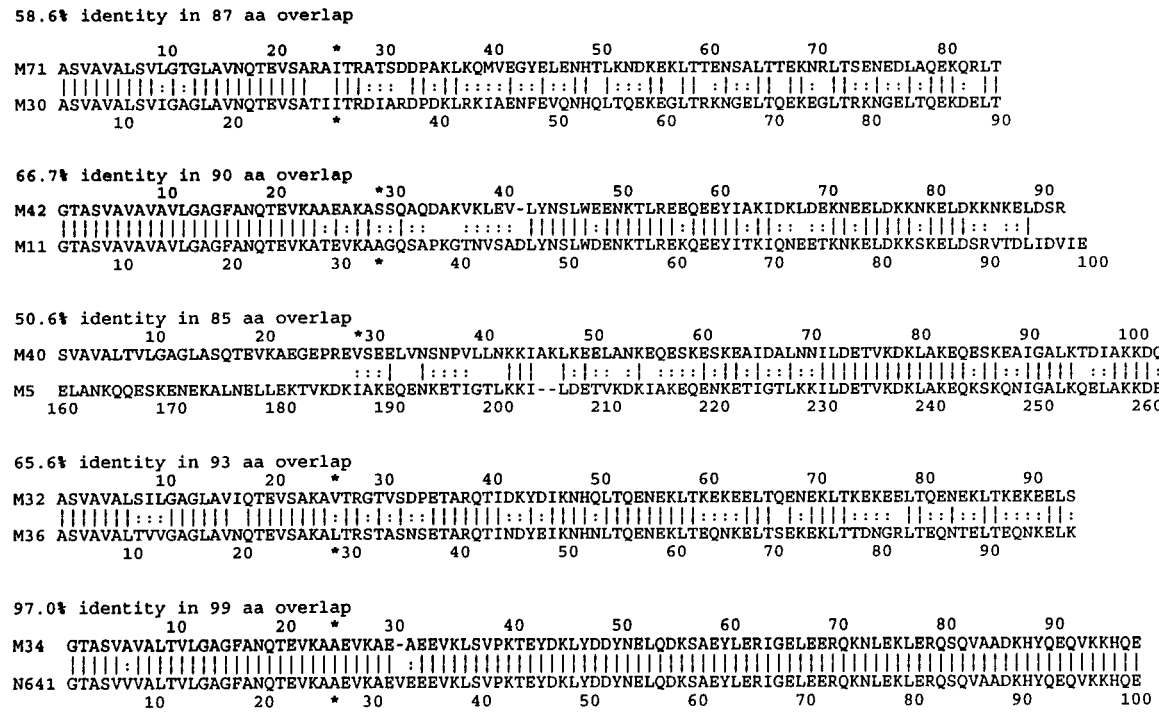

FIG. 1. Amino acid sequence alignment of the deduced N-terminal sequences of M71, M42, M40, M32, and M34 proteins determined in this study with the M30, M11, M5, M36, and *enn64/14* (N641) proteins, respectively. The accession numbers for the corresponding genes are given in Table 2. Amino acid 1 in all instances (except for M5) is predicted to occur at approximately residues 16 to 20 of the respective precursor protein leader sequence on the basis of homologies with other M and M-like proteins. The first residue of the mature forms (*) is predicted in the same manner.

their closest GenBank matches are only 76 to 88% identical over their first 160 bases (Table 3).

New 5′ *emm* sequences for established M types. Six new 5′ *emm* gene sequences were obtained in this study. As shown by their levels of identity with other *emm* genes, *emm-32*, *emm-34*, *emm-38*, *emm-40*, *emm-42*, and *emm-71* represent previously unsequenced *emm* genes (Table 2). We found that the 5′ *emm-34* sequence from our reference strain was nearly identical to the previously sequenced *enn64/14* gene (Table 2), with the deduced products differing over a 99-residue overlap only by the deletion of one residue and by two conservative substitutions (Fig. 1). It is doubtful that the *emm-34* sequence derived from strain 784 actually represents an *enn* gene sequence, since primer 1 used for PCR should not anneal to known 5′ *enn* gene regions under the conditions used because of four mismatches (20).

Comparisons of these sequences with the corresponding regions of other *emm* genes showed a high level of sequence identity within the deduced signal sequence region (of which only part is shown in Fig. 1), followed by a more variable region comprising the N terminus of the mature protein (Fig. 1). The *emm-40* gene also encodes a signal sequence conserved with other M proteins, although the *emm-40* product is similar to an N-terminal region of the *emm-5* product corresponding to M5 residues 188 to 272 (Fig. 1).

Unexpectedly, we found the 5′ *emm* sequence of two independent M38 strains, 306 (*emm-38*) and 54 (*emm-38*), to be identical to the corresponding sequence in strains 722 (*emm-40*) and 138 (*emm-40*). Lancefield extracts of the M38 and M40 strains reacted specifically with our corresponding M38 and M40 M-type precipitating antisera, and no cross-reactivity was observed. In this context it is interesting that the *emm-40* gene from strains 722 and 138 is approximately 100 bp larger than the *emm-38* gene from strains 54 and 306 (PCR fragment sizes are not shown). The basis of the lack of cross-reactivity between the M38 and M40 sera may be due to M structural differences, or it is possible that expression of other M-like proteins contributes to the serospecificity of either or both of these M types.

Discrepancies between GenBank entries and CDC reference strains. We found that for six M-type designations (M types 13, 27, 67, 68, 70, and 79) our reference strain 5′ *emm* sequences clearly differed from *emm* gene sequences in GenBank for the corresponding *emm* genes (Table 4). In order to help resolve these discrepancies, our results are described below.

CDC M13 5′ *emm* sequence. We found the 5′ *emm* sequence in five of our M-type 13 reference strains (including the original Lancefield M type 13 strain 376 and two animal-passaged derivatives 936 and 636) representing three independent original strains to have 96 to 99% sequence identity (Table 4). The closest match to these sequences was actually *emm-76*, which had only 79% sequence identity over 160 bases (Table 4).

CDC M27 5′ *emm* sequence. Our three M-type 27 strains representing two original M27 isolates contained *emm* genes that were closely matched or identical to the sequence previously reported for *emm-77* (Table 1) but were not similar (≤80% identical over 160 bases) to the previously entered *emm-27* sequence in the database. In our random clinical study we found an M27 serotype for isolate 1141d (Table 3). Unlike our M27 reference strains, this strain had a 5′ *emm* sequence very similar to the *emm-27* sequence in GenBank (Table 3) (19). At this point we are unable to explain why the CDC M27 reference strains as well as this clinical isolate M type as M27 with the same sera. Indeed, we are also unable to explain why our M77 strain does not type M27 with this sera.

CDC *emm-65*, *emm-67*, and *emm-69*. According to GenBank entries, the *emm-53* and *emm-67* genes have identical 5′ se-

TABLE 4. Discrepant results between GenBank entries and 5′ *emm* sequences of CDC M typing reference strains

CDC strain	Closest match	% Identity[a]	Accession no.
Not highly similar[b]			
376 (M13, T13, OF+)	*emm-76*	85	U11992
636 (M13, T13, OF+)	376	98	
936 (M13, T3/13, OF+)	376	98	
488 (M13, T ND, OF+)	376	98	
576 (M13, T3/13, OF+)	376	99	
1084 (M67, T3/13/B/28, OF−)	*emm-65*	92	U11980
1095 (M68, T1, OF+)	*emm-13*	75	U11936
Matching[c]			
627 (M27, T5/27/44, OF+)	*emm-77*	100	U11911
797 (M27, T5/27/44, OF+)	*emm-77*	98	U11911
132 (M27, T5/27/44/, OF+)	*emm-77*	99	U11911
1096 (M69, T3/13/B, OF−)	*emm-65*	99	U11980
1097 (M70, T28, OF−)	STBSB75	100	L27095
1151 (M79, T25/IMP/9, OF+)	*emm-80*	100	U12004

[a] Of the first 160 bases of the reference strain *emm* sequence with respect to the indicated GenBank sequence.

[b] *emm* sequences not highly similar over 160 bases to any GenBank *emm* entries. ND, not determined.

[c] *emm* sequences similar to *emm* genes not correlating to the reference strain serologic M type.

quences (19). However, our M type 67 strain (strain 1084, Bisno's PSC227) 5′ *emm* sequence was not closely related to the *emm-67* entry in GenBank (Table 4) which was reported to be determined from the same strain (19). Our M53 strain 5′ sequence was in close agreement with the previously submitted *emm-53* sequence (Table 1).

CDC reference strains for M types 65 and 69 (there is no GenBank entry available for *emm-69*) gave 5′ *emm* sequences nearly identical to the previously reported *emm-65* gene (19) (Table 4).

CDC *emm-68* sequence. The sequence of the 5′ *emm* sequence of our M68 reference strain was not highly similar to the GenBank *emm-68* entry (which is nearly identical to the *emm-1* GenBank entry) and clearly different from the most similar GenBank entries (Table 4).

CDC *emm-70* sequence. We found that the 5′ sequence for *emm-70* is identical to sequence STBSB75 in strain BSB75 (17) (Table 4). Strain BSB75 was reported to cross-react with M type 80 precipitating antisera but has a 5′ *emm* sequence that differs from *emm-80* (17) (Table 1). This strain was not tested with M type 70 precipitating antisera and the sequence for *emm-70* is not in GenBank. Furthermore, our M type 80 strain *emm* sequence correlates with an *emm-80* sequence in GenBank. These factors indicate that STBSB75 should be identified as *emm-70*.

CDC *emm-79* and *emm-80* sequences. Our M type 80 strain had a 5′ *emm* sequence identical to only one of the two sequences previously entered in GenBank as *emm-80* 5′ sequence (Table 1). The nonmatching *emm-80* sequence was identical to the *emm* sequence obtained with our *emm-79* strain (strain 1151 in Table 4). There is no sequence listed in the database for *emm-79*. These observations indicate a possible error in the previous entry for *emm-80* (accession no. U12004).

DISCUSSION

The purpose of this study was to establish the use of a rapid discriminating DNA sequence-based system for typing GAS in this laboratory. We also wished to expand and provide a reference for this *emm* sequence database. These aims were fully accomplished in that 95 serologic reference GAS strains were rapidly and accurately sequenced, with 81 of these sequences found to be identical (or almost identical) to specific *emm* sequences in GenBank. Additionally, five unique *emm* gene 5′ sequences (*emm-32*, *emm-34*, *emm-40*, *emm-42*, and *emm-71*) were added to GenBank for future reference. Our results show that only a relatively small number of apparent discrepancies concerning sequences corresponding to the 5′ ends of *emm-13*, *emm-27*, *emm-67*, *emm-68*, *emm-70*, and *emm-79* need to be resolved. We do not know at this point what the bases of these discrepancies are, but this accounting should hasten the clarification of these *emm* types.

Relatively few of our *emm* sequences from strains of a given M type were 100% identical to corresponding *emm* genes in GenBank. In some instances, the corresponding *emm* sequence diverged by as much as 5% over 160 bases (Table 1), but even in these instances the results were circumstantially strongly indicative of the corresponding serological type, since the next most homologous sequences usually had at best only about 70 to 80% identity. Divergence in some cases was undoubtedly partially due to errors in reading the sequences, since generally sequences were read only one time and most sequence reading errors could have been readily identified by closer examination. However, the major purpose of this study was to test a rapid *emm* typing scheme and to provide a baseline level of accuracy. In some cases differences were also due to a limited amount of divergence between strains within individual M types (15, 18, 20). Regardless, the level of divergence between our single-run *emm* sequences and the GenBank *emm* sequences did not interfere with the prediction of M serotypes for any of the clinical isolates of determinable M type (Table 3), and there was generally ≥95% agreement between *emm* allele sequences of M-type reference strains and GenBank *emm* sequences (Table 1).

Many new *emm* genes will likely become known through sequence analysis, as was evidenced by the lack of GenBank sequences corresponding to the *emm* sequences of our M-type reference strains M13, M32, M38, M40, M42, M67, M68, and M71 (Tables 1 and 3). This is even more apparent with the observation of three new *emm* gene sequences among our random survey strains (Table 2). Additionally, in a recent study of European, Asian, and African GAS isolates we have sequenced many more unique *emm* genes (2). Since the potential for intergenic and intragenic recombination between *emm*-like genes appears to be great, any attempts to keep abreast of *emm* gene diversity requires sequence analysis.

Recent studies established the feasibility of rapid *emm* gene hybridization-based typing (8). This technology depended upon hybridization of an 18- to 30-base DNA probe with its complementary sequence amplified from test GAS strains. Although this method is effective, disadvantages include a very limited target analysis, the need for a specific oligonucleotide probe for each *emm* allele, and the use of costly DNA labeling reagents. As previously noted (8, 19), DNA sequencing is much more discriminating in that it allows the rapid direct deduction of the sequence of up to 500 bases of the 5′ end of *emm* genes. This is much more specific and reliable than serologic M typing and should markedly improve surveillance. Although the DNA sequence analysis described here relied upon the use of an automated sequencing system using dye terminators in which the four dideoxynucleotides used for chain termination are differentially dye labeled, we have found that manual systems for thermo-cycle sequencing these *emm*-derived PCR frag-

ments using a ^{32}P-end-labeled sequencing primer are at least as effective (data not shown).

One example of the advantage of using *emm* typing for M-nontypeable strains was shown by the analysis of seven *emm53* strains (Table 3), which were obtained within a very short time from the same hospital outbreak of GAS infection. Although three of these strains were M nontypeable (Table 3), a combination of temporal, geographic, phenotypic (T-type and OF phenotype), and *emm* sequence data strongly suggested a clonal relation between these strains.

emm gene sequencing is also useful in distinguishing between strains within one serological M type. For example, in this study it was clear that one M6 strain *emm* gene sequence diverged from the almost identical *emm* sequences of the other three M6 strains after approximately 210 bases because of a different pattern of recombination between the 21-bp direct repeats in the *emm-6* repeat region (6; data not shown). Another example of an *emm* gene subtype was provided by the M type 27 strain (Table 3), which contained a deletion of 7 codons compared with the GenBank *emm-27* sequence. Variation within given *emm* genes, such as that within different *emm-6* and *emm-27* alleles, may prove very useful in surveillance of GAS strains of specific *emm* gene subtypes.

It must be stressed that *emm* gene characterization in itself is insufficient for the identification of specific GAS clones, since individual strains with the same M specificity can vary extensively in overall genetic relatedness (14, 17–19). For example, our results with *emm-65* and *emm-69* may be similar to the results found previously with the identical *emm-44* and *emm-61* 5′ sequences (19) (Table 1). Thus, even though these strains were clearly divergent genetically with respect to T types (Table 1) and multilocus electrophoretic types (19), as our M65 and M69 reference strains they have identical 5′ *emm* sequences and possibly confer identical M specificities. Similarly, it appears that our M type 77 reference strain (Table 1) and our M type 27 reference strains are also genetically distinct strains (as seen by their different T types; Table 1) that have identical or nearly identical 5′ *emm* sequences (Table 1). These results, as previously noted (18), indicate that selected GAS typing methods in addition to *emm* typing should be used to define clonal relatedness. It is obvious, however, that specific *emm* sequences in many circumstances could provide very sensitive supportive evidence of clonal disease associations.

ACKNOWLEDGMENTS

We are grateful to Theresa Hoenes and Anne Whitney for their patience and the use of their automated sequencing apparatus. We thank Gary Sanden for support and Kristine Walter for excellent technical assistance.

REFERENCES

1. **Beachey, E. H., J. M. Seyer, J. B. Dale, W. A. Simpson, and A. H. Kang.** 1981. Type-specific protective immunity evoked by synthetic peptide of *Streptococcus pyogenes* M protein. Nature (London) **292:**457–459.
2. **Beall, B., R. Facklam, and T. Hoenes.** Unpublished data.
3. **Coleman, G., A. Tanna, A. Efstratiou, and E. T. Gaworzewska.** 1993. The serotypes of *Streptococcus pyogenes* present in Britain during 1980–1990 and their association with disease. J. Med. Microbiol. **39:**165–178.
4. **Fischetti, V. A.** 1989. Streptococcal M protein: molecular design and biological behavior. Clin. Microbiol. Rev. **2:**285–314.
5. **Gooder, H.** 1961. Association of a serum opacity reaction with serological type in *Streptococcus pyogenes*. J. Gen. Microbiol. **25:**347–352.
6. **Hollingshead, S. K., V. A. Fischetti, and J. R. Scott.** 1986. Complete nucleotide sequence of type 6 M protein of the group A streptococcus. J. Biol. Chem. **261:**1677–1686.
7. **Jones, K. F., and V. A. Fischetti.** 1988. The importance of the location of antibody binding on the M6 protein for opsonization and phagocytosis of group A M6 streptococci. J. Exp. Med. **167:**1114–1123.
8. **Kaufhold, A., A. Podbielski, M. Blokpoel, and L. Schouls.** 1994. Rapid typing of group A streptococci by the use of DNA amplification and non-radioactive allele-specific oligonucleotide probes. FEMS Microbiol. Lett. **119:**19–26.
9. **Kehoe, M. A.** 1994. Cell wall-associated proteins in gram-positive bacteria, p. 217–261. *In* J. M. Ghuysen and R. Hakenbeck (ed.), Bacterial cell wall. Elsevier Science B. V., Amsterdam.
10. **Lancefield, R. C.** 1962. Current knowledge of the type specific M antigens of group A streptococci. J. Immunol. **89:**307–313.
11. **Maxted, W. R.** 1980. Disease association and geographical distribution of the M types of group A streptococcus, p. 763–777. *In* M. T. Parker (ed.), Streptococcal diseases and the immune response. Academic Press, Inc., New York.
12. **Maxted, W. R., and J. P. Widdowson.** 1972. The protein antigens of group A streptococci, p. 251–266. *In* L. W. Wannamaker and J. M. Matsen (ed.), Streptococci and streptococcal diseases. Academic Press, New York.
13. **Maxted, W. R., J. P. Widdowson, C. A. M. Fraser, L. C. Ball, and D. C. J. Bassett.** 1973. The use of serum opacity reactions in the typing of group A streptococci. J. Med. Microbiol. **68:**83–90.
14. **Moody, M. D., J. Padula, D. Lizana, and C. T. Hall.** Epidemiologic characterization of group A streptococci by T-agglutination and M-precipitation tests in the public health laboratory. Health Lab. Sci. **2:**149–162.
15. **Musser, J. M., V. Kapur, J. Szeto, X. Pan, D. S. Swanson, and D. R. Martin.** 1995. Genetic diversity and relationships among *Streptococcus pyogenes* strains expressing serotype M1 protein: recent intercontinental spread of a subclone causing episodes of invasive disease. Infect. Immun. **63:**994–1003.
16. **Penney, T. J., D. R. Martin, L. C. Williams, S. A. Demalmanche, and P. L. Berquist.** 1995. A single *emm* gene-specific oligonucleotide probe does not recognise all members of the *Streptococcus pyogenes* M type 1. FEMS Microbiol. Lett. **130:**145–149.
17. **Relf, W. A., D. R. Martin, and K. S. Sriprakesh.** 1994. Antigenic diversity within a family of M proteins from group A streptococci: evidence for the role of frameshift and compensatory mutations. Gene **144:**25–30.
18. **Single, L. A., and D. R. Martin.** 1992. Clonal differences within M-types of the group A streptococcus revealed by pulsed field gel electrophoresis. FEMS Microbiol. Lett. **91:**85–90.
19. **Whatmore, A. M., V. Kapur, D. J. Sullivan, J. M. Musser, and M. A. Kehoe.** 1994. Non-congruent relationships between variation in *emm* gene sequences and the population genetic structure of group A streptococci. Mol. Microbiol. **14:**619–631.
20. **Whatmore, A. M., and M. A. Kehoe.** 1994. Horizontal gene transfer in the evolution of group A streptococcal *emm*-like genes: gene mosaics and variation in Vir regulons. Mol. Microbiol. **11:**363–374.