***emm* and *sof* gene sequence variation in relation to serological typing of opacity-factor-positive group A streptococci**

Bernard Beall,[1] Giovanni Gherardi,[1]† Marguerite Lovgren,[2]
Richard R. Facklam,[1] Betty A. Forwick[2] and Gregory J. Tyrrell[2]

Author for correspondence: Bernard Beall. Tel: +1 404 639 1237. Fax: +1 404 639 3123.
e-mail: beb0@cdc.gov

[1] Centers for Disease Control and Prevention,
Respiratory Diseases Branch, 1600 Clifton Rd,
Mailstop C02, Atlanta,
GA 30333, USA

[2] National Centre for Streptococcus, Provincial Laboratory of Public Health for Northern Alberta, 8440-112 St,
Edmonton, Alberta,
Canada T6G 2J2

Approximately 40–60 % of group A streptococcal (GAS) isolates are capable of opacifying sera, due to the expression of the *sof* (serum opacity factor) gene. The *emm* (M protein gene) and *sof* 5′ sequences were obtained from a diverse set of GAS reference strains and clinical isolates, and correlated with M serotyping and anti-opacity-factor testing results. Attempts to amplify *sof* from strains with M serotypes or *emm* types historically associated with the opacity-factor-negative phenotype were negative, except for *emm12* strains, which were found to contain a highly conserved *sof* sequence. There was a strong correlation of certain M serotypes with specific *emm* sequences regardless of strain background, and likewise a strong association of specific anti-opacity-factor (AOF) types to *sof* gene sequence types. In several examples, M type identity, or partial identity shared between strains with differing *emm* types, was correlated with short, highly conserved 5′ *emm* sequences likely to encode M-type-specific epitopes. Additionally, each of three pairs of historically distinct M type reference strains found to share the same 5′ *emm* sequence, were also found to share M serotype specificity. Based upon *sof* sequence comparisons between strains of the same and of differing AOF types, an approximately 450 residue domain was determined likely to contain key epitopes required for AOF type specificity. Analysis of two Sof sequences that were not highly homologous, yet shared a common AOF type, further implicated a 107 aa portion of this 450-residue domain in putatively containing AOF-specific epitopes. Taken together, the serological data suggest that AOF-specific epitopes for all Sof proteins may reside within a region corresponding to this 107-residue sequence. The presence of specific, hypervariable *emm*/*sof* pairs within multiple isolates appears likely to be a reliable indicator of their overall genetic relatedness, and to be very useful for accurate subtyping of GAS isolates by an approach that has relevance to decades of past M-type-based epidemiological data.

Keywords: *emm* gene sequences, *sof* variable gene sequences, *Streptococcus pyogenes*, opacity factor

INTRODUCTION

Roughly 40–45 % of the group A streptococcal (GAS) invasive isolates from the Centers for Disease Control (CDC) population-based surveillance within the United States are found to opacify sera (see http://www.cdc.gov/ncidod/biotech/strep/strepindex.html), due to the presence of serum opacity factor (Sof). This is in reasonable agreement with previous results in which the *sof* gene was found in 43 % of invasive GAS isolates and in 56 % of isolates recovered from non-sterile sites (Kreikemeyer *et al.*, 1995). GAS Sof is an approximately 1000 residue cell-surface-bound apoproteinase named

†**Present Address:** Libera Università Campus Bio-Medica, Via E. Longoni 83, 00155 Rome, Italy.

Abbreviations: AOF, anti-opacity factor; CDC, Centers for Disease Control, GAS, group A streptococci, NT, nontypable, OF, opacity factor.

for its property of rendering various sera opaque (Krumwiede, 1954, Kreikemeyer *et al.*, 1995; Rakonjac *et al.*, 1995; Courtney *et al.*, 1999). Serum opacity is generated by Sof-mediated apoprotein AI cleavage of high-density lipoprotein, which causes subsequent high-density lipoprotein aggregation (Saravani & Martin, 1990). Sof also has fibronectin-binding activity that resides in a relatively short C-proximal domain (situated N-terminal of its consensus wall-attachment motif) that is distinct from the large opacity-factor (OF)-conferring segment (Rakonjac *et al.*, 1995; Courtney *et al.*, 1999). Sof is a virulence factor of unknown mechanism in an intraperitoneal mouse model (Courtney *et al.*, 1999), however the roles of the enzymic and fibronectin-binding activities of Sof in this model are unknown.

M protein serotyping has served as a subtyping standard for GAS for much of the 20th century. It has long been known that GAS strains within certain M surface-virulence-protein serotypes are associated with the opacity-factor-positive (OF+) phenotype (Gooder, 1961; Widdowson *et al.*, 1970). Of 86 known M-protein serotypes and provisional serotypes, 36 of these historically correlate with the OF+ phenotype (Fraser & Colman, 1985; Johnson & Kaplan, 1993; Facklam *et al.*, 1999), and these strains are commonly found in sterile- and nonsterile-site infections (Colman *et al.*, 1993). Antisera against these OF+ GAS strains have been reported to inhibit the OF+ reaction only in strains of the same M serotype (Maxted *et al.*, 1973). This observation of anti-OF (AOF) specificity is consistent with previous observations that the *sof* locus is quite variable between strains of different M serotypes (Rakonjac *et al.*, 1995). In fact, much of the entire N-terminal 80% of the Sof protein sequence appears to be hypervariable, with interspersed small conserved regions (Rakonjac *et al.*, 1995; Courtney *et al.*, 1999). Most of this large, variable domain has been found to be essential for OF activity, which complicates the determination of epitopes targeted by AOF sera.

Although specific M serotypes have been shown to be conferred by epitopes at the mature M protein N terminus (see Fischetti, 1989 for review of M protein structure), for unknown reasons many OF+ strains have always been very difficult to M serotype. Instead, in many studies the M serotype has been inferred based upon AOF specificity. This is a fundamentally illogical inference, since the *emm* and *sof* genes are situated at least 15 kb apart (Rakonjac *et al.*, 1995), and horizontal gene transfer events do occur in GAS (Bessen & Hollingshead, 1994; Whatmore *et al.*, 1994). M-protein gene (*emm*) sequences have been documented in some instances to be identical between strains of differing genetic lineages (Whatmore *et al.*, 1994). The differing strain backgrounds within specific *emm* types are often reflected by differing serological specificities of the poorly defined T antigens (Beall *et al.*, 1997), although each of the commonly occurring *emm*/M types are represented primarily by a closely related T agglutination pattern, suggesting overall genetic relatedness within many *emm* types (Johnson & Kaplan, 1993; Beall *et al.*, 1998). Recently we replaced M typing at the CDC with *emm* sequence typing, since limited M-typing data indicated that *5′ emm* sequence can be correlated very well to M serological data (Beall *et al.*, 1996, 1997). Furthermore, we have found that isolates within the same *emm* type that share similar or identical T agglutination patterns are usually genetically highly related on the basis of genomic-restriction-digest pattern analysis (unpublished observations).

For only a minority of *emm* types has it been shown that the *emm* gene specifically encodes the M serotype (for examples see Hollingshead *et al.*, 1986; Robbins *et al.*, 1987; Miller *et al.*, 1988; Mouw *et al.*, 1998; Dale *et al.*, 1993). Since many GAS strains, including most OF+ strains, have additional '*emm*-like' genes in addition to *emm* (described as the single gene amplified by a specific primer set; Whatmore *et al.*, 1994) at the *vir* locus (Hollingshead *et al.*, 1993; Whatmore *et al.*, 1995) these other *emm*-like genes potentially contribute to the M serotype since they are likely to be present in crude M antigen extracts. In this work, using a set of highly geographically and temporally diverse OF+ GAS strains, we have found additional circumstantial evidence that *5′ emm* sequences dictate M-serotype specificity. We also present data demonstrating that although the sequence of the first 190–240 codons of *sof* is generally highly predictive of AOF type, a 100–450 residue region within the previously defined enzymic domain (Rakonjac *et al.*, 1995; Courtney *et al.*, 1999) appears likely to dictate AOF type specificity. We show several instances where *sof* types are not predictive of M type or the corresponding *emm* sequence type, and in several instances the combination of *sof* and *emm* type appears to be highly predictive of genetically related strain sets.

METHODS

Serology. T agglutination patterning, AOF determination and M serotyping were performed as previously described by Johnson & Kaplan (1996). Antisera were produced in guinea pigs or rabbits. M typing and AOF sera were prepared against validated reference strains for M types 2, 4, 13L, 22, 25, 28, 44, 48, 49, 58–63, 66, 73, 75, 76, 77, 79, 81, ST2967, 87, 89 and 90. The latter three types were previously recognized as provisional types PT2841, PT4245 and PT4931, respectively (Facklam *et al.*, 1999). 13L refers to Lancefield serotype M13, found in the Lancefield reference strain in Table 1. The source of the *emm13W* type strain, referred to as *emm13* in earlier work (Whatmore *et al.*, 1994), was the Public Health Laboratory Service, Colindale, UK (Table 1). M-typing sera, but not AOF-typing sera, were prepared from the OF+ reference strains for M types 13L, 27L and 68. M-typing sera were prepared against OF− reference strains for M types 1, 3, 5, 6, 14–19, 23, 24, 26, 29–43, 46, 47, 50–57, 64, 65, 71, 72, 74, 80 and 83 (formerly provisional M type PT2110). AOF-typing sera, but not M-typing sera, were prepared from reference strains for M types 9, 11, 27G, 78 and 92 (formerly provisional type PT5110).

Strains. The CDC collection of M type reference strains was used, many of which originated from Dr Rebecca Lancefield's original M type collection (Beall *et al.*, 1996; Facklam *et al.*, 1999). US isolates from California, Oregon, Minnesota, Georgia, Tennessee, Connecticut and Maryland were obtained from normally sterile sites through the Emerging Infections Program/Active Bacterial Core Surveillance (see http://www.cdc.gov/ncidod/dbmd/abcs/gas98.pdf) during 1995–1999. Isolates from other states within the US and from other countries were usually from sterile sites, however a small percentage were from unknown and nonsterile sites.

Of the 86 recognized M serotypes, 36 have been consistently associated with the OF+ phenotype (Fraser & Colman, 1985; Johnson & Kaplan, 1993; Colman *et al.*, 1993; Facklam *et al.*, 1999). For two of these serotypes historically associated with the OF+ phenotype, M13 and M27, two distinct reference *emm* sequence types exist (Facklam *et al.*, 1999). These *emm* types are shown in Table 1 as M13L/*emm13L* and M13W/*emm13W,* and M27L/*emm27L* and M27G/*emm27G*. Additionally, it has been found that isolates containing the commonly occurring *emm* sequence types *st2967* (M. Lovgren & G. Tyrrell, unpublished data) and *pt5118* (M92) (Facklam *et al.*, 1999) represent unique M and/or AOF sero-specificities which brings the total of OF+ GAS *emm* types associated with known serological correlates to 39.

For each of the *emm* types featured in this study, the 5′ *sof* sequence was obtained from a CDC reference strain. CDC reference strains for many M nontypable (NT) strains with new *emm* sequence types were also subjects of this study. To maximize potential strain variability within *emm* types, strains with unusual T pattern/*emm* type associations and strains from diverse geographic locations were examined.

Sequence analysis. Sequence analysis was carried out using the Wisconsin package version 10.0. Signal-sequence predictions were carried out as described at the web site http://www.cbs.dtu.dk/services/SignalP/ (Nielsen *et al.*, 1997), using the N-terminal 22 aa from the GenBank accession AF019890 (or U02290 and X88303, which are identical) plus the first 48 aa deduced from primer F-based sequence.

***emm* typing.** *emm* and *sof* gene-specific PCR was performed using standard protocols described for the Boehringer Mannheim Hi Fidelity system. *emm* sequence typing and criteria defining *emm* type designations have been previously described (Beall *et al.*, 1998; CDC, 1999b). All *emm* sequences used for this study are available at http://www.cdc.gov/ncidod/biotech/strep/strepindex.html) and were independently obtained in the CDC streptococcal laboratory from CDC reference strains. All of these *emm* sequences were in close agreement with the given GenBank accession numbers in Table 1, except that in some instances longer sequences were generated for purposes of sequence comparisons. Sequences of a designated *emm* type shared 97–100% sequence identity over at least 252 bases of the corresponding CDC reference strain *emm* sequence encoding the mature protein. The two *emm68.1* isolates were deleted of M68 mature protein codons 3–9 and had three single-base changes resulting in conservative substitutions.

***sof* amplification and sequencing.** Conserved primer sets were based upon comparison of the *sof* gene with GenBank accession nos U02290/X88303 and AF019890, which represented the only two GAS *sof* sequences in GenBank at the time of this work. Primers F2 (5′-GTATAAACTTAGAAAGTT-ATCTGTAGG-3′) and R3 (5′-GGCCATAACATCGGCAC-CTTCGTCAATT-3′) were used to generate approximately 560–700 bp fragments from all strains that encompassed *sof* sequence encoding the mature protein plus 22 residues of signal sequence. Primer F (5′-GGGCTCGTCTCCGTCGG-AACGATGCTG-3′) was used for sequencing the *sof* 5′ region encoding 7–31 signal-sequence residues and up to 270 mature-protein residues. For many strains, primers F2+R5 (5′-GTAAAGGATGCTTCACGTTTGTCTCCAG-3′) were used to amplify most of the *sof* structural gene. F3 (5′-GAAG/CAAATTGACGAAGGTGCCGATGT-3′) was another universal primer used for sequence analysis and PCR. Various other conserved or nonconserved *sof* primers were also used for amplification and sequencing reactions.

PFGE. Selected strains were typed by PFGE of chromosomal digests using *Sma*I. Isolates differing by only 1–6 bands from a common reference strain for each group were assigned a common type. More than six bands of difference from subtype 1 of each type were considered unrelated isolates and assigned a different PFGE type (Tenover *et al.*, 1995).

RESULTS

emm sequence types of M-typable strains correspond to their M type specificity

Table 1 shows the M serotyping data for 104 of the strains included in this study. Of these 104 strains, 66 were M typable with the available M-typing sera. The serotypes M4, M9, M11, M12, M13L, M22, M44, M48, M49, M58, M60, M61, M62, M63, M64, M66, M68, M73, M76, M77, M79, M81, M87 and M89 were only found in reference strains and/or clinical isolates with the corresponding *emm* sequence type. In total, 64 of the 68 total M-typable strains were of the M types predicted by previously obtained 5′ *emm* gene sequence designations (Whatmore *et al.*, 1994; Podbielski *et al.*, 1991; Facklam *et al.*, 1999).

Ten of the 57 M nontypable (NT) strains were found to be within the *emm* types 22, 25, 28, 48, 73, 75, 81 and 27/77, for which corresponding M-typing sera was available. At present, the basis for this nontypability is not known. In contrast, only one of the strains shown in this study, 1588-96, was not *emm* sequence typable since an *emm*-specific amplicon could not be generated. Although this strain was also M NT, it is likely that altered primer-annealing site(s), rather than the absence of the *emm* gene, prevented a successful PCR reaction since this strain multiplies in the indirect bactericidal assay (data not shown; see Johnson & Kaplan, 1996 for assay).

In one example, previous M-typing results recorded many years earlier were in disagreement with our *emm*-sequencing results. Strain D734, the source of the first *sof* gene to be sequenced, was previously recorded as an M type 22 strain (Rakonjac *et al.*, 1995). We found it to be M NT and to have the *emm* sequence type *pt2233* (Table 1). It must be noted that D734 was a strain from Dr Lancefield's collection (http://www.rockefeller.edu/vaf) that was serotyped long before type PT2233 was documented (Fraser & Colman, 1985) and may have cross-reacted with M type 22 antisera.

Sequence correlates of new *emm* sequence types with classical M serotypes

Four M-typable strains were found to have new *emm* gene sequences that differed significantly from the

Table 1. *emm* and *sof* sequence types of M type reference strains and clinical isolates correlated with M types, AOF types and T agglutination patterns

ND, Not done.

sof type/ accession no. of reference strain or comparison of *sof* to reference sequence*	AOF type	*emm*†	M type	T type	CDC strain-year	Source or where isolated‡
2/AF157555	2	2	2	2	633-66	RL (M2)
2/identical	2	2	2	8/25/Imp19	826-97	Brazil
2/identical	ND	2	ND	2	1899-97, 4313-97	Bulgaria, Argentina
2/identical	OF−	NT§	NT	2/28	1588-96	Chile
4/AF137607	4	4	4	4	SS470-54	PHLS (M4)
4/5 aa insert (58–62)	4	4	4	4	2066-99	Georgia, USA
4/5 aa insert (58–62), 7 aa frameshift (108–114)	NT	4	4	4	SS81-49	RL (M4)
4/5 aa insert (58–62)	ND	4	ND	2/28	489-97	Minnesota, USA
8/AF138790	NT	8	NT	8	634-66	RL (M8)
8/S161 → R	NT	8	NT	8/25/Imp19	194-96, 970-97	Brazil, Poland
8/S161 → R	NT	*st3018*	NT	6	SS1468-97	Malaysia
9/AF174430‖	9	9	NT	9	SS129-66	RL (M9)
9/identical	ND	9	ND	9/18/14	SS650-66	RL (M9)
9/identical	ND	9	ND	9	191-96, 715-97	Georgia, USA; Colombia
11/AF141140	11	11	NT	11/12	SS68-67	RL (M11)
11/identical	11	11	11	11/12	3137-99	Georgia, USA
11/identical	ND	11	ND	11	17-95	Minnesota, USA
12/AF138792	OF−	12	12	12	SS635-51, 3179-99	RL (M12), Georgia
12/identical	OF−	12	ND	12	835-97, 4621-97, 6156-99	Brazil, Korea, Georgia, USA
13L/AF138793	NT	13L	13L	3/13/B3264	SS636-51	RL (M13)
13W/AF138794	NT	13W	NT	3/B3264	SS1475-97	PHLS
13W/V180 → G	NT	13W	NT	3/B3264	2938-97	Chile
13W/Δ14–22, V180 → G	NT	13W	NT	3/B3264	2328-99	Maryland, USA
22/AF138791	22	22	NT	22	SS638-68	RL (M22)
22/identical	22	22	22	12/3/B3264	4020-98, 3167-99	Argentina; Georgia, USA
22/identical	ND	22	ND	12	2820-99	Illinois, USA
22/identical	ND	22	ND	NT	195-96, 3030-97	Brazil, Malaysia
25/AF138795	OF−	25	NT	8/25/Imp19	SS639-66	RL (M25)
25/K203 → E	25	11	11	11/12	4808-96	Hawaii, USA
27L/AF138796	NT	27L/77¶	77	5/27/44	SS132-49	RL (M27L)
27L/R18 → S	NT	27L/77¶	77	5/12/27/44	1707-97	Argentina
27G/AF177978	OF−	27G	NT	5/12/27/44	SS582-66	PHLS (M27G)
27G/identical	27G	*st4935*	NT	6	73-97	California, USA
27G/identical	ND	27G	NT	5/12/27/44	4653-97	Chile
27G/Δ92–93	27G	27G	ND	5/27/44	4624-97	California, USA
28/AF138797	28	28	NT	28	SS789-68, 4613-97	RL, Korea
28/identical	ND	28	ND	28	3971-98, 2323-99	Brazil; Maryland, USA
28/identical					3135-99, 3138-99	California, USA
44/AF138798‖	44	44/61#	44 + 61††	5/12/27/44	SS511-55, 1764-97	RL (M44), Colombia
48/AF138799	48	48	NT	4	SS737-67	RL (M48)
48/identical	48	48	48	4	5304-98	Georgia, USA
49/AF138800	OF−	49	49	14	SS702	RL (M49)
49/identical	49	49	49	8/14	4956-96	India
49/identical	ND	49	ND	14	1111-96	California, USA
49/T95 → S, Δ97T	ND	49	ND	14	NZ131-90	NZ (M49)
58/AF138801	58	58	58	25	SS872-69, 6038-99	PHLS (M58), Czech Republic
58/identical	58	58	58	8/25/Imp19	1883-99	Minnesota, USA
59/AF138802	59	59	59	12	SS1454-69, SS913-69	PHLS (M59), HD (M59)
59/identical	59	59	59	11/12	1229-95	Georgia, USA
60/AF138803	60	60	60	4	SS874-69	HD (M60)
60/A121 → P	60	60	60	4	4534-96	Malaysia
60/A121 → P	ND	60	ND	4	180-91	France
61/AF138804	61	44/61#	61	11/12	SS875-69, 1312-95	HD; Georgia, USA
61/identical	ND	44/61#	ND	11/12	656-98	Virginia, USA
62/AF133805	62	62	62	12	SS984-70	PHLS (M62)
62/identical	62	62	62	12/3/B3264	966-97	Poland
63/AF133806	63	63	63	4	SS985-71, 21-96	PHLS (M63); California, USA
sof PCR negative	OF−	none	64	3	2594-97	Chile
66/AF138807	66	66	66	12	SS1037-73	HD (M66)
66/identical	66	66	66	12/3/B3264	1637-95	Georgia, USA
68/AF138808	68	68	68	1	SS1095-71, 2367-97	Egypt (M68); California, USA
73/AF138809	73	73	73	3/13/B3264	SS1145-76	PHLS (M73)
73/identical	73	73	NT	3/13/B3264	2368-97	California, USA
73/F65 → S, R167 → S	73	73	NT	3/13/B3264	5102-98	California, USA
75/AF139736	75	75	NT	8/25/Imp19	SS1147-76	PHLS (M75)
75/identical	75	75	75	8/25/Imp19	6033-99	Czech Republic
75/identical	ND	75	ND	8/25/Imp19	3134-99	Georgia, USA
75/identical	ND	75	ND	25	4020-99	Connecticut, USA
75/identical	75	84	NT	25	SS1449-97	PHLS (M84)
75/identical	NT	84	NT	8/25	D734-79	VF
75/identical	NT	25	NT	8/25/Imp19	246-95	Georgia, USA
75/identical/86L_S, 87V_S, 89	75	*st1815*	NT	8/25/Imp19	SS1479-97	North Carolina, USA
76/AF139734	76	76	76	12	SS1148-76, 209-96	PHLS (M76), Brazil
76/identical	ND	76	ND	8/25/Imp19	1685-95	Georgia, USA
76/identical	76	85	NT	3/B3264	SS1447-97, 261-96	PHLS, Hawaii
77/AF138810	77	27L/77¶	NT	13/28	SS149-76, 2099-97	PHLS (M77); Massachusetts, USA
77/identical	ND	27L/77¶	ND	13/28	4156-95, 6200-99	Maryland, USA
78/AF139739	78	78	NT	11	SS1150-76	PHLS (M78)
78/identical	78	78	NT	11/12	4321-97, 6034-97	Argentina, Czech Republic
78/identical	ND	78	ND	11	179-91	France

Table 1 (*cont.*)

sof type/ accession no. of reference strain or comparison of *sof* to reference sequence*	AOF type	*emm*†	M type	T type	CDC strain-year	Source or where isolated‡
79/AF192473	79	79	79	11/12	SS1151-76	PHLS (M79)
81/AF138811	81	81	81	NT	SS1173-78	PHLS (M81)
81/identical	81	81	81	3/13/B3264	4329-97	Argentina
81/Δ76–89	81	81	81	3/B3264	SS1452-96	PHLS (M81)
82/AF139753	NT	82	NT	5/12/27/44	SS1402-96, 1394-95	PHLS (M82); Georgia, USA
87/AF139744	87	87	87	28	SS1399-96, 6035-99	PHLS (M87), Czech Republic
87/identical	87	87	87	12/28	4431-95	Denmark
88/AF139752	61	88	NT	NT	SS1455-97	PHLS (M88)
89/AF139750	89	89	89	11/12	SS1397-96, 5001-98	PHLS (M89); Vermont, USA
89/identical	ND	89	ND	3/13/B3264	6039-99	Czech Republic
89/Δ26–27	89	89	89	11/12	817-97	Brazil
90/AF139740	90	90	90	3/13/B3264	SS1396-96	PHLS (M90)
90/Δ110–114	90	*st833*	NT	3/B3264	SS1444-96	Brazil
90/Δ110–114, S38 → T, S81 → P, N161 → D	ND	*st6735*	ND	11/12	6735-99	Brazil
92/AF139748	92	92	NT	8/25/Imp19	SS1460-94, 1135-95	NZ (M92); Georgia, USA
92/identical	ND	92	ND	8/25/Imp19	2974-95, 2109-98	California, USA; Oregon, USA
2967/AF139749	2967	*st2967*	2967	12	SS1357-95	California, USA
st2697/identical	2967	*st2967*	2967	NT	134-98	California, USA
st2697/identical	1967	*st1160*	2	11/12	1160-99, 2141-99	Egypt
213/AF139743	NT	*st213*	NT	4	SS1408-97	Brazil
436/AF192769	NT	*st436*	NT	12/27	SS1363-95	Connecticut, USA
448/AF191036	NT	*st448*	NT††	3/13/B3264	SS1364-95, 1191-98	Connecticut, USA
1207/AF191035	2967	*st1207*	NT	B3264	SS1457-97	Minnesota, USA
1482/AF177977	61	88	NT	8/9	1482-97	Brazil
1658/AF154330	NT	81	81	11/12	SS1401-96	PHLS (M81)
1881/AF139755	NT	*st4935*	NT	B3264	1881-97	Bulgaria
1965/AF192474	NT	81	81	4	1965-92	Ethiopia
2034/AF139742	OF–	*st2034*	NT	NT	SS1379-92	New Guinea
2034/identical	NT	*st2034*	NT	B3264	3019-97	Malaysia
2034/identical	ND	*st2034*	ND	13/B3264	4821-96, 841-97	Hawaii, Brazil
2034/identical	ND	*st2034*	ND	NT	76-97	California, USA
2147/AF178681	NT	*st2147*	59	8/25/Imp19	2147-99	Egypt
2904/AF139757	NT	*st2904*	NT	3/B3264	SS1471-97	Brazil
2920/AF139756	NT	4	4	8/25/Imp19	2920-97	Brazil
3894/AF191037	NT	*st448*	NT††	6	3894-98	Brazil
3930/AF178533	NT	44/61#**	61	11/12	3930-98	Brazil
4438/AF191034	NT	68.1	NT‡‡	3/13/B3264	4438-98	Georgia, USA
4438/identical	ND	68	ND	3/13/B3264	6615-99	Brazil
4470/AF179217	NT	68.1	NT‡‡	3/13	4470-96	Connecticut, USA
4532/AF192475	NT	*st4532*	76	5/27/44	SS1416-96	Malaysia
4539/AF139745	NT	87	87	11/12	4539-96	Malaysia
4835/AF139751	NT	89	89	13	4835-96	Hawaii, USA
4835/Δ58–71	NT	89	89	13	1090-96	California, USA
4935/AF139754	NT	*st4935*	NT	4	SS1422-96	India
4958/AF153315	NT	25.1	NT	NT	4958-96	India
4958/identical	ND	75	ND	6	6733-99	Brazil
A207/AF139747	NT	*stA207*	NT	3	SS1413-96	MB
A207/Δ15–23	NT	*stA207*	NT	3	2441-96	Georgia, USA
ns14x/AF145351	NT	*stns14x*	NT	12	SS1437-97	Australia
ns14x/identical	NT	*stns14x*	NT	12/3/B3264	675-99	Maryland, USA

* GenBank accession numbers for *sof* gene sequences encompassing 567–1400 bases of 5′ sequence for most of the indicated reference strains with the exceptions of *sof9*, *sof44*, *sof61*, *sof3875*, *sof1482* and *sof81* for which sequences of bases 2564–2768 were obtained. The top row of each *sof* type refers to the reference strain. Deletions refer to amino acid numbers.

† See http://www.cdc.gov/ncidod/biotech/strep/strains/emmtypes.html for GenBank accession numbers of all *emm* types shown and descriptions of indicated strains.

‡ M types determined by sources are shown in parentheses. Abbreviations: RL, Dr Rebecca Lancefield (Lancefield, 1962); PHLS, Streptococcal Reference Laboratory, Public Health Laboratory Service, Colindale, UK (supplied by Dr Androulla Efstraiou); HD, Dr H. Dillon (Dillon & Dillon, 1974); MB, Dr Michael Boyle (Pack & Boyle, 1995), AEK, Dr A. M. el-Kholy (el-Kholy *et al.*, 1973), VF, Dr Vincent Fischetti (Rakonjac *et al.*, 1995); NZ, New Zealand. Strains outside of the US were provided by Drs K. S. Sriprakash (Australia), Lucia Teixera (Brazil), Antoaneta Detcheva (Bulgaria), Rosa Bustos Vasquez (Chile), Elizabeth Castaneda (Colombia), Paula Kriz (Czech Republic), M. P. LePennee (France), Kwangtun Lee (Korea), Diana Martin (New Zealand), Deborah Lehuman (New Guinea) and Waleria Hryniewics (Poland).

§ We were unable to amplify an *emm* amplicon from this isolate.

‖ Although the first 678 5′ *sof9* and *sof44* bases obtained with primer sofF were identical, the *sof9* and *sof44* genes diverge after base 1026 and are readily distinguished by characteristic RFLP profiles of amplicons obtained with the sofF3 + sofR5 pair.

¶ The sequences of at least the N-terminal 124 residues of mature M27L and M77 deduced proteins are identical.

The sequences of at least the N-terminal 86 residues of mature M44 and M61 deduced proteins are identical.

** The strain typed as both M44 and as M61.

†† Upon testing with M49 antiserum, M extracts from these strains reacted specifically, but non-identically, with M49 extracts (see text for discussion).

‡‡ Upon testing with M68 antiserum, M extracts from these strains reacted specifically, but non-identically, with M68 extracts (see text for discussion).

(a)

```
M2     NSKNP.. .....VPVKK EA...KLSEA ELHDKIKNLE EEKAELFEKL DKVEEEHKKV EEE
ST1160 NSKTPAP AP..AVPVKK EATKSKLSEA ELHDKIKNLE EEKAELFEKL DKVEEEHKKV EEE
M73    DNQSPA. ......PVKK EAK..KLNEA ELYNKIQELE EGKAELFDKL EKVEEENKKV KEE
ST2967 NSKNPAP APASAVPVKK EAT..KLSEA ELYNKIQELE EGKAELFDKL EKVEEENKKV KED
```

(b)

```
               10        20        30        40        50        60
ST2147 EEASPKNGQLTLQQKYDALTNENKSLRKERDNYLNYLYEKEELEKKNKELHSELASVTETL
       |:|: :||:||||||||||||||||||:|||||||||||||||||||||| |::|:    |
M59    EQAKNNNGELTLQQKYDALTNENKSLRRERDNYLNYLYEKEELEKKNKELDSQVAG----L
               10        20        30        40        50

               70        80        90       100        110
ST2147 TSVTEADDKKIKDLTDRDK-ISSNLIGNAKDQINKLTTEKDKLAEKAKKLEE
        :|:|:|::: |    |:| :  ::: ::|||:|:||:||| ||||||||||
M59    IGVVESDEEEAK----RSKNMYETFLKQSKDQVNELTAEKDTLAEKAKKLEE
        60            70        80        90       110
```

(c)

```
              10        20        30        40
M68   EEANKKAEEVKKAEESESKSAAKMWEDMYKELDRDYSLLEKTVENMSLE
      ||       |||||||||||||||||:||||||||||||||||||||||
M68.1 EE-------VKKAEESESKSAAKMWENMYKELDRDYSLLEKTVENMSLE
                     10        20        30        40

      50        60        70        80        90
M68   NMEKLDKLSKENQGKLEKLELDYLKKLDHEHKEHQKEQQEQEERQKNQE
      |||:||||:||||||||||||||||||||||||||||||||||||||||
M68.1 NMENLDKLNKENQGKLEKLELDYLKKLDHEHKEHQKEQQEQEERQKNQE
            40        50        60        70        80
```

(d)

```
                  10            20        30        40
ST448 AEKKV----EVADSNASSVAK----LYNQIADLTDKNGEYLERIEELEERQ
      |||||    |||::|:||||:    ||:|||||||||||||||| ||||||
M49   AEKKVEAKVEVAENNVSSVARREKELYDQIADLTDKNGEYLERIGELEERQ
              10        20        30        40        50

           50        60        70        80        90
ST448 KNLEKLERQSQVAADKHYQEQVKKHQEYKQEQEERQKNLEELERQNKREIDKR
      |||||||:|||||||||||||:|||||||||||||||| |:|||:: ||::||
M49   KNLEKLEHQSQVAADKHYQEQAKKHQEYKQEQEERQKNQEQLERKYQREVEKR
             60        70        80        90       100
```

Fig. 1. (a) Alignment of the mature N-terminal sequences of the deduced M2, ST1160, M73 and St2967 proteins. The M serotypes of each are indicated. M2 residues 1–35, known to elicit type-specific opsonic antibody, is in bold type (Dale *et al.*, 1993). The sequence of ST1160 residues 21–43, which is identical to M2 residues 13–35, is underlined. (b) Comparison of the mature N-terminal regions of ST2147 and M59. A 38-residue sequence nearly identical between both proteins that possibly dictates the M59 serotype is in bold font. (c) Comparison of the M68 and M68.1 proteins which putatively share partial M serotype identity. (d) Comparison of the M proteins ST448 and M49 which putatively share partial M serotype identity. A highly homologous 60-residue sequence, which possibly provides the basis of the cross-reactivity between these two proteins, is in bold.

sequences predicted by their M types. Still, for each of these four strains, clear correlations with the M type *emm* sequences were found. Two type *st1160* isolates were found to type as serologically identical to M2 in gel-diffusion tests. This observation correlated to 23 residues of identical sequence shared between ST1160 and a portion of the 35 N-terminal M2 residues known to elicit opsonic type-specific antibodies (Dale *et al.*, 1996) (Fig. 1a). This 23-residue sequence is not perfectly conserved between any other two known M sequences. It is interesting that while the N-terminal ends of the predicted ST1160, M73 and ST2967 proteins all have highly related sequences with similar overall homology to M2, strains with the *emm* sequence *st1160* are serotype M2, while the *emm73* and *st2967* sequences are correlated with the specific serotypes M73 and ST2967 respectively (Table 1).

A similar situation to that found with *emm* type *st1160* isolates was found with an *emm* type *st2147* strain, which serotyped as M59. The closest match to the deduced ST2147 is in fact M59, with nearly identical sequence (38/39) between the N-terminal residues 7–49 of ST2147 and M59 (Fig. 1b). The next closest known match to this sequence is M63, with 80 % sequence identity over mature residues 12–41 (data not shown).

The remaining M serotypable strain, SS1416 (*st4532*), typed as M76, even though the deduced ST4532 50 N-terminal residues share only 54·2 % sequence identity with the corresponding M76 sequence. However, besides the M27G sequence, M76 represents the closest match to the ST4532 N terminus (and we do not have anti-M27G typing sera). A 40-residue segment of the mature ST4532 protein consisting of residues 27–66 is consistent with the serological M76 result, since other than seven substitutions (six conservative), it is identical to the corresponding M76 segment (data not shown).

Serological non-identical M cross-reactivity correlates with *emm* sequence type

In two examples where partial M serotype identity was found in gel-diffusion tests, clear sequence correlations were also found. In one example, M extracts from two strains carrying an *emm68* allele (*emm68.1*) deleted of mature codons 3–9 and containing three conservative substitutions (Fig. 1c) were found to specifically cross-react only with M68 antiserum and showed partial identity against reference M68 strain extracts.

All three *st448* strain extracts, representing two different genetic backgrounds on the basis of *sof* sequence types and T agglutination patterns, also showed partial identity with M49 extracts when tested with M49 antisera. Significantly, the closest sequence match to the ST448 protein is M49 and the two deduced proteins share marked similarity between M49 residues 26–90 and the corresponding N-terminal residues 18–81 of ST448 (Fig. 1d).

Distinct classical M serotypes corresponding with identical 5′ *emm* sequences overlap in M type specificity

For unknown reasons, classical M type reference strains for M27L, M77, M44, M61, M81 and PT1658 have been reported to have distinct M types, even though the strain pairs for M27L/M77, M44/M61, M44/M61 and M81/PT1658 have recently been found to share *emm* gene sequence types (Whatmore *et al.*, 1994; Beall *et al.*, 1996). Our data were in partial disagreement with classical M-typing data in that we observed that two of four strains sharing the *emm44/61* sequence, including the CDC Lancefield M44 type strain SS511, were found

Table 2. Data indicating that conserved *sof* and *emm* sequences are indicative of overall genetic relatedness

Strain (*emm* type, T type)	*sof* type	PFGE type common to *emm* type?	No. strains within *emm* examined by PFGE (T types encountered)*
1899-97(*emm2*, T2)	2	Yes	20 (18 T2, 2 T28)
826-97 (*emm2*, T8/25/Imp19)	2	Yes	20 (18 T2, 2 T28)
2066-99 (*emm4*, T4)	4	Yes	23 (22 T4, 1 T2/28)
489-97 (*emm4*, T2/28)	4	Yes	23 (22 T4, 1 T2/28)
2920-97 (*emm4*, T8/25/Imp19)	2920	No	23 (22 T4, 1 T2/28)
835-97 (*emm12*, T12)	12	Yes	34 (T12)
195-96 (*emm22*, T NT)	22	Yes	16 (9 T12, 7 T11/12)
3971-98(*emm28*, T28)	28	Yes	34 (T28)
817-97 (*emm89*, T11/12)	89	Yes	19 (10 T11, 8 T11/12, 1 NT)
4835-96 (*emm89*, T13)	89	No	19 (10 T11, 8 T11/12, 1 NT)
2109-98 (*emm92*, T NT)	92	Yes	8 (6 T8/25/Imp19, 2 TImp19)
134-98 (*st2967*, T NT)	2967	Yes	8 (4 T11/12, 3 TNT, 1 T11)

* Within each *emm* type, the same PFGE type (differing by 0–4 bands; except for 826-97, which differed by 6 bands) was encountered among all of the indicated randomly selected isolates.

in this study to M type as both M44 and as M61 (Tables 1 and 2). Why only T pattern 5/27/44, *sof44*, *emm44/61* (M44 + 61) strains, but not T pattern 11/12, *emm44/61*, *sof61* (M61) strains displayed this dual M type specificity is unknown (Table 1).

A simple example of distinct M type strains sharing the same M type specificity was found with the classical Lancefield M27 reference strain, SS132 (T5/27/44) and the more recent M77 reference strain, SS1149 (T13). The *emm27L* allele from this strain has only one nucleotide difference in a 372 base overlap with the *emm77* allele from the M77 reference strain SS1149, and their partial deduced M protein sequences are identical over their entire 124 residue overlap (Beall *et al.*, 1996; http://www.cdc.gov/ncidod/biotech/strep/strepindex.html). This is consistent with the observation that one clinical isolate and SS132 (both of which were *emm27L*/*emm77*, T5/27/44, *sof27L*) were M type 77. Although the M77 reference strain [SS1149 (T13, AOF77, *sof77*)] was M NT, the *emm27L/77*, T13/28, *sof77* clinical isolate 2099-97 was M type 77. We are unable to explain why no positive results were obtained using anti-M27L typing sera.

Another example of distinct M type reference strains with identical M serotypes and corresponding identical *emm* sequence types were the PT1658 (in Table 1 with *sof1658*) and M81 reference strains (Whatmore *et al.*, 1994; http://www.cdc.gov/ncidod/biotech/strep/strepindex.html). All five strains with the *emm81* sequence, accounting for T NT, T4 or T3/13/B3264-related agglutination patterns and three distinct *sof* sequence types, were found to be M type 81, and this result was recently confirmed at the Public Health Laboratory Service, Colindale, UK (Table 1).

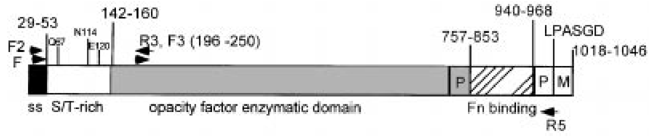

Fig. 2. Representative Sof protein. The approximate annealing locations of the four 'universal' *sof* primers used for this work are shown in relation to the deduced Sof protein sequence. Different predicted signal sequences of 29–53 residues in length are indicated by ss. Three conserved residues within the variable serine and threonine (S/T)-rich region are indicated. The putative enzymic domains lying between approximately residues 130–150 and 757–853, including a variable-length proline-rich region demarcating the C terminus of the putative enzymic domains (left-hand P), represent our sequence comparisons to previous results (Rakonjac *et al.*, 1995; Courtney *et al.*, 1999). The locations of the N-terminal fibronectin-repeat region, LPASGD cell-wall attachment motif, proline-rich conserved putative cell-wall spanning region (right-hand P), membrane-associated region (M) and C terminus (1018–1046) are inferred from previous publications (Kreikemeyer *et al.*, 1995, Rakonjac *et al.*, 1995, Courtney *et al.*, 1999).

5′ *sof* sequences

All of the OF+ reference strains and various OF+ isolates shown yielded approximately 580–750 bp *sof*-specific PCR products with the primer pair F2+R3, which anneal to sequences encoding signal sequence and a conserved amino-proximal region respectively (Fig. 2). This conserved amino-proximal region represents one of several highly conserved short sequences previously seen to be distributed along the length of the enzymic domain (Courtney *et al.*, 1999). Additionally, all strains tested yielded approximately 3·0 kb amplicons with the F2+R5 primer pair, with R5 annealing to a conserved sequence overlapping the wall-attachment-motif encoding sequence (Fig. 2). With the exception of *sof61*, *sof3875*, *sof1482*, *sof2967*, *sof1207*, *sof9* and *sof44*, for which protein sequences of 872–922 amino acids were deduced, the 62 different *sof* designations shown in Table 1 represent a remarkably variable set of related partial 190–470 residue Sof proteins that share 50–70 % sequence identity. The *fnbA* product from *Streptococcus dysgalactiae* (Lindgren *et al*., 1993) which also functions as an OF (Courtney *et al*., 1999), shared approximately 35–40 % sequence identity over this range (data not shown).

In several instances mosaic-like structures were evident, with distinct segments shared between *sof* segments from other strains. For example, *sof2841* was nearly identical over bases 1–213 and 352–555 to the corresponding sequence from *sof79*, with bases 78–116 nearly identical to *sof3894*. Such instances may be reflective of horizontal transfer events between GAS strains.

Each of the *sof* sequences determined was predicted to encode a membrane export signal peptide, with the first 10 residues (corresponding to amino acids 23–32 of *sof* proteins encoded by the sequences with accession nos U02290/X88303 and AF019890) totally conserved in the majority of the isolates. Based upon sequence differences and predicted cleavage sites, there was a total of 12 different predicted signal peptides of 29–53 residues in length.

One striking feature found in all of the various Sof peptides was an abundance of N-terminal serine and threonine residues, which comprised about 50 % of the first 100–150 residues. This region lies outside of the putative enzymic domains of these proteins and displays a remarkable degree of sequence diversity. It is also interesting that three residues (corresponding to Sof2967 Q67, N114 and E120; see accession no. AF139749) are totally conserved among all of the known Sof protein sequences. The functional significance of these shared features of the Sof N-terminal region lying amino-proximal to known Sof functional domains remain to be determined (Fig. 2).

The only classically OF− M/*emm* type examined that yielded *sof*-specific PCR products (both F2+R3 and R2+R5 generated) was M12/*emm12*. This is consistent with previous data demonstrating the presence of sequences in an M type 12 strain hybridizing under high stringency to a *sof* gene probe (Rakonjac *et al*., 1995). We have not explored the basis of the OF− phenotype in these strains. We were unable to amplify *sof* gene sequences from reference strains and/or clinical isolates corresponding to the *emm*/M types 1, 3, 5, 6, 15, 18, 33, 41, 43, 56, 64, 69 and 86, which are all commonly associated with an OF− phenotype (Fraser & Colman, 1985; Podbielski *et al*., 1991; Colman *et al*., 1993; Whatmore *et al*., 1994; http://www.cdc.gov/ncidod/biotech/strep/strepindex.html). We found *sof12* amplicons from strains isolated in the US and South America that had the identical 5′ sequence as the *sof* amplicon from the CDC M12 reference strain (Tables 1 and 2). *Dde*I digests of the 3 kb F2+R5-generated amplicon from 30 random *emm12* isolates in the CDC collection shared an identical six-band profile (data not shown), further demonstrating that the entire *sof12* gene is highly conserved among *emm12* isolates.

sof 5′ sequence types usually predict AOF types

Of 113 strains that we attempted to type for AOF, 75 were typable. Nine of the 160 strains included in this study failed to produce detectable OF and could not be tested. AOF specificities of 73 of the 75 AOF-typable strains were directly predictable by *sof* gene sequences that were identical or nearly identical to the M type reference strain (Table 1). The AOF types 2, 4, 9, 11, 22, 25, 27G, 28, 44, 48, 49, 58, 59, 60, 61, 62, 63, 66, 68, 73, 75, 76, 77, 78, 81, 87, 89, 92 and ST2967 were only found in reference strains and/or clinical isolates with the corresponding designated *sof* sequence types (Table 1). In four instances, AOF NT strains were found within a given *sof* type for which corresponding AOF typing sera was available. These included the only *sof13L* strain, one of three *sof4* strains and two of five *sof75* strains. The basis for this nontypability is not known, although among the *sof75* strains it did not often appear to be due to gross alterations of the *sof* structural gene, since *Dde*I restriction digest profiles of F2+R5-generated 3 kb amplicons showed identical seven-band profiles.

In four instances, identical AOF types were found between strains with different *emm* types and/or M serotypes. The only strains that typed as AOF75 were of the *emm* types *emm75*, *emm84* and *st1815* (Table 1), which correlated with perfect or nearly perfect sequence identity over the 5′ 621 bases. Similarly, identical *sof* sequence and *Dde*I digestion data found for all five of the *emm76* and *emm85* strains (four of which were AOF type 76) indicates the identity or near identity of the *sof* genes among these strains. A third example of multiple *emm* types sharing the same AOF type and 5′ *sof* sequence was found with *emm* type *st1160* and *st2967* strains (Table 1). Finally, the *emm* type *st833* isolate was AOF type 90, in agreement with its partial *sof* product differing from the corresponding Sof90 sequence by only a five-residue deletion in the serine-rich region amino-proximal to the enzymic domain (Table 1), and the indistinguishable *Dde*I profiles shared between the two *sof* amplicons (data not shown). Nearly identical sequence to *sof90* was also found in a strain with the *emm* type *st6735* (Table 1).

Sof9 and Sof44 confer distinct AOF types, but share identity over their N-terminal 43 %

In only one instance were identical 5′ *sof* sequences for the first 180–270 codons found between strains belonging to two distinct AOF types. The *sof9* and *sof44* sequences were identical over their 5′ 342 codons. However, in agreement with the clearly distinct AOF specificity between AOF type 9 and 44 strains, the *sof9* and *sof44* genes were found to abruptly diverge after

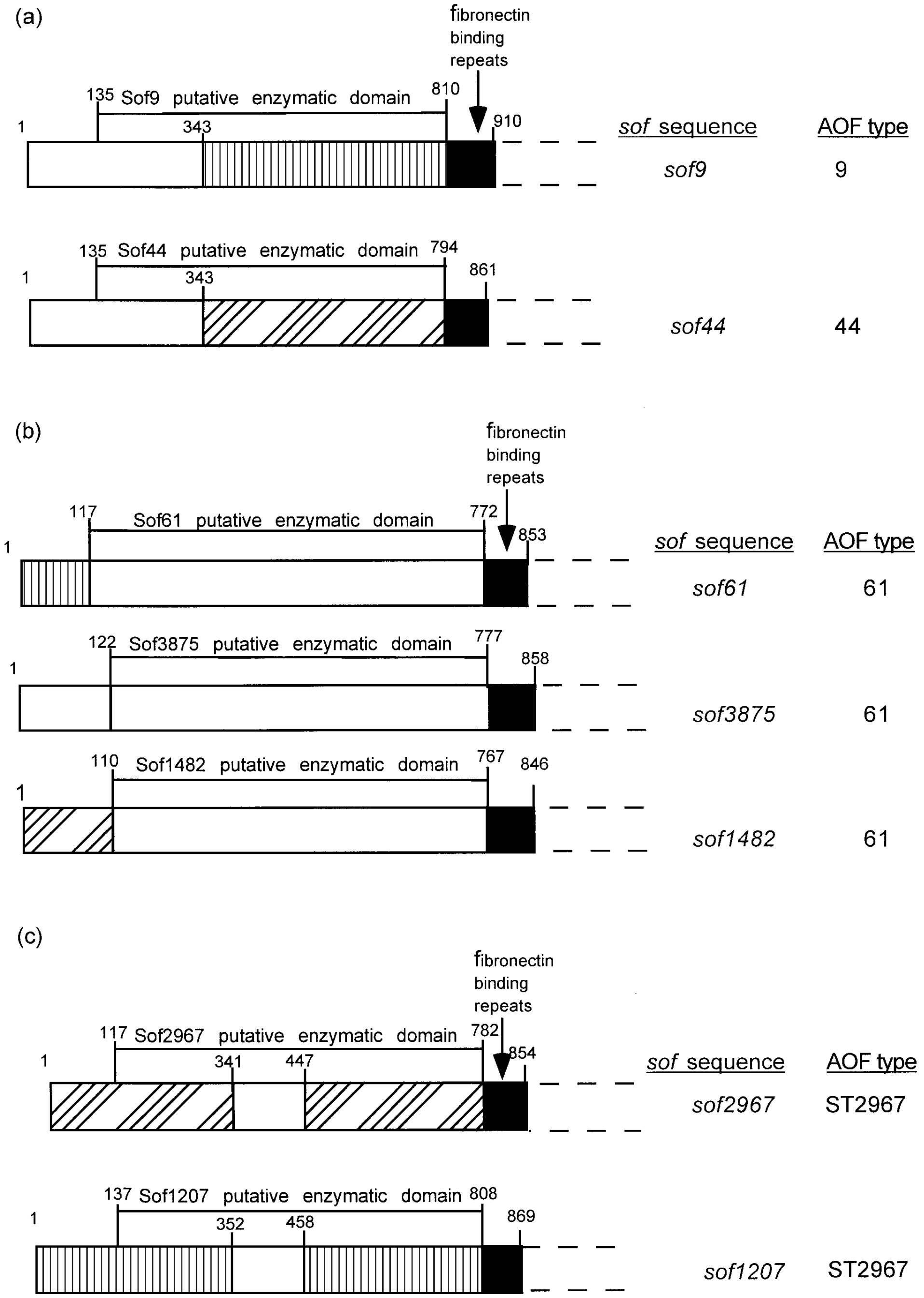
(a)
fibronectin binding repeats
135 Sof9 putative enzymatic domain 810
1
343
910
sof sequence
AOF type
sof9
9
135 Sof44 putative enzymatic domain 794
1
343
861
sof44
44
(b)
fibronectin binding repeats
117 Sof61 putative enzymatic domain 772
1
853
sof sequence
AOF type
sof61
61
122 Sof3875 putative enzymatic domain 777
1
858
sof3875
61
110 Sof1482 putative enzymatic domain 767
846
1
sof1482
61
(c)
fibronectin binding repeats
117 Sof2967 putative enzymatic domain 782
1
341
447
854
sof sequence
AOF type
sof2967
ST2967
137 Sof1207 putative enzymatic domain 808
1
352
458
869
sof1207
ST2967

Fig. 3. (a) Depiction of the relationship between Sof9 and Sof44 mature proteins. The conserved regions are indicated by white and black portions. Residues corresponding to the Sof2 enzymic domain (Courtney *et al.*, 1999) are indicated. Nonconserved regions are indicated with vertical and diagonal stripes. The dashed lines indicate that the entire sequences have not been obtained. (b) Depiction of the relationship between the three Sof proteins putatively conferring AOF type 61. The nonconserved N-terminal residues are indicated by the rectangles with different shading. The totally conserved putative enzymic domain (117–772, 122–777 and 110–767) is indicated in white and the conserved fibronectin-binding repeat regions are black. (c) Depiction of the relationship between the two Sof proteins putatively conferring AOF type ST2967. The white rectangle represents an area of high localized homology that is not shared among 16 other Sof proteins for which this sequence is available (see text).

residue 343 of the predicted mature Sof9 product (Fig. 3a). The sequence of Sof9 residues 343–810, corresponding to the C-terminal 445 residues of the 695 aa Sof2 enzymic domain (Courtney *et al.*, 1999), was distinct from the equivalent region of Sof44 (residues 343–794). The distinct AOF types conferred by Sof9 and

Sof44 suggests that type-specific AOF epitopes may only reside in the C-terminal 450 residues of the enzymic domain.

Sequence relationships of heterologous *sof* genes among strains with the same AOF type

Of the three AOF type results that were not in direct agreement with CDC reference strain *sof* sequence results, it was of interest that strains SS1455 (*emm88*, *sof88*) and 1482-97 (*emm88*, *sof1482*) were AOF type 61. Although the deduced Sof61 partial product from strain SS875 (*sof61*) shared only 54–63 % sequence identity over its first 116 residues with the corresponding sequences of Sof3875 and Sof1482, mature residues 117–771 of Sof61 were found to be unique and almost totally conserved between the three proteins (Fig. 3b). Significantly, these residues almost exactly correspond to the minimal region of Sof2 found to be essential for enzymic activity (Courtney *et al.*, 1999). The carboxy-proximal portion of the AOF type 61-specific sequence encompasses the corresponding *sof*9- and *sof44*-specific sequences that are apparently required for the AOF9 and AOF44 reactions respectively (residues 343–810 and 343–794 respectively, Fig. 3a, b).

Analogous to the AOF type 61 situation, SS1457 (*sof1207*) was found to be AOF type ST2967, while the Sof2967 and Sof1207 mature N-terminal residues (1–782 and 1–808 respectively) were only 68 % identical over their entire overlap. Closer analysis revealed striking similarity (91 % identity) between between the two proteins over a 107-residue region within their putative enzymic domains (Fig. 3c). In contrast, for 15 other available GAS Sof proteins for which this sequence was available, their corresponding 107-residue regions shared only 20–58 % identity. These Sof proteins included four from previous studies (Rakonjac *et al.*, 1995; Courtney *et al.*, 1999) and 12 from the present study, including Sof11, Sof12, Sof28, Sof44, Sof61, Sof77, Sof81, Sof82, Sof87, Sof88, Sof1482 and Sof4539 (see accession nos in Table 1). Additionally, Sof9 was found to share 84 % identity with Sof2967 over this 107-residue region (corresponding to Sof9 residues 356–462, compare Fig. 3a and 3c), however the region also included seven nonconservative substitutions dispersed along the length of the Sof2967/Sof9 overlap in contrast to only two nonconservative subtitutions found in the Sof2967/Sof1207 overlap. At this point it appears logical to speculate that all strains within a given AOF type may share strong homology in the region corresponding to this 107-residue domain associated with AOF type 2967.

A portion of the region putatively encoding the fibronectin-binding repeats (Fig. 2), in the seven *sof* genes for which these longer sequences (2·7–2·8 kb) were obtained (*sof61*, *sof1482*, *sof3875*, *sof9*, *sof44*, *sof1207* and *sof2967*), was as expected, highly conserved with the corresponding regions of other known *sof* products (Kreikemeyer *et al.*, 1995; Rakonjac *et al.*, 1995; Courtney *et al.*, 1999).

Concordance between *sof* and *emm* types

For the majority of strains, the 5′ 189–258 codon *sof* gene sequence was either identical or highly similar to the corresponding *sof* sequence found in the reference strain of the same *emm* type (Table 1). For example, for both *emm22* and *emm28* strains, 6/6 strains contained identical 5′ 657–696 bp *sof* sequences (data not shown, depicted as identical amino acid sequences in Table 1). For the majority of the specific *emm* type reference strains shown in Table 2, dating from as long ago as 1949, their deduced Sof amino acid sequences were >99 % identical to 1–5 recent clinical isolates with the corresponding *emm* type (Table 1). Analogous to what has been seen with M types and corresponding *emm* sequence types (Whatmore *et al.*, 1994), for the majority of strains within specific AOF types there appears to be surprisingly little allelic variation of *sof* genes within the common 5′ variable region analysed. For the majority of given *sof* gene comparisons, identical amino acid sequences corresponded to identical nucleotide sequences, with few examples of silent base substitutions. Although any base substitution was uncommon within the various *sof* gene designations assigned in Table 1, there were more base substitutions resulting in missense mutations than in silent substitutions. Additionally, deletions/insertions of 1–14 codons, or two single base deletions resulting in short frame-shifts were not common (Tables 1 and 2). The observed deletions were often associated with tandem homologous repeats, analogous to those seen in the GAS *emm* and *sic* deletion alleles (Hollingshead *et al.*, 1997; Mejia *et al.*, 1997).

PFGE profiles from strains with *sof*/*emm* combinations of the same designations were very similar to PFGE profiles from the majority of randomly selected strains within the same *emm* type, indicating that these particular *emm* types are comprised mainly of highly genetically related strains (Table 2). Not surprisingly, strains with unusual *sof* gene associations for a given *emm* type also differed in their PFGE patterns from the major pattern observed within an *emm* type (Table 2, see strains 2920-97 and 4835-96). Strains sharing both highly conserved *sof* and *emm* genes also usually shared related T agglutination patterns, although exceptions are evident in Table 1 (for example, see strain 826-97 compared to other *emm2*, *sof2* strains; 6039-99 compared to other *emm89*, *sof89* strains, and the two *emm59*, *sof59* strains SS1454 and SS913).

Table 3 summarizes the number of isolates within various *sof*-positive *emm* types that we have identified during the last 3 years. In general, for the *emm* types with 10 or more isolates listed in Table 3, this reflects their relative isolation frequency compared to the other *sof*-positive *emm* types in our ongoing population-based GAS surveillance within the US (Beall *et al.*, 1997, 1998; Zurawski *et al.*, 1998; http://www.cdc.gov/ncidod/biotech/strep/strepindex.html). It is evident that there is usually a specific 5′ *sof* sequence type most commonly associated with a given frequently occurring *emm* type

Table 3. *sof*-positive *emm* types representing two or more isolates encountered in CDC US surveillance and/or miscellaneous studies during 1995–1998 in order of isolate frequency

ND, Not done.

emm **type**	**Total no. isolates**	***sof*** **types of** ***emm*****-type reference strain(s)**	***sof*** **types found in clinical isolates (no. examined)**
emm12	266	*sof12*	*sof12* (4)
emm28	244	*sof28*	*sof28* (5)
emm11	151	*sof11*	*sof11* (2), *sof25* (1)
emm89	120	*sof89*	*sof89* (3), *sof4835* (2)
emm22	83	*sof22*	*sof22* (5)
emm75	67	*sof75*	*sof75* (3)
emm27L/77	74	*sof77*, *sof27L*	*sof77* (3), *sof27L* (1)
st2967	63	*sof2967*	*sof2967* (1)
emm92	55	*sof92*	*sof92* (2)
emm76	52	*sof76*	*sof76* (2)
emm58	49	*sof58*	*sof58* (2)
emm2	48	*sof2*	*sof2* (3)
emm59	42	*sof59*	*sof59* (1)
emm87	42	*sof87*	*sof87* (2), *sof4539* (1)
emm82	35	*sof82*	*sof82* (1)
emm73	34	*sof73*	*sof73* (2)
st2034	26	*sof2034*	*sof2034* (4)
emm13w	33	*sof13w*	*sof13w* (2)
emm4	25	*sof4*	*sof4* (3), *sof2920* (1)
emm44/61	25	*sof44*, *sof61*	*sof61* (2), *sof44* (1), *sof3930* (1)
emm85	21	*sof76*	*sof76* (1)
emm78	18	*sof78*	*sof78* (3)
emm81	7	*sof81*	*sof81* (2), *sof1965* (1), *sof1658* (1)
stns14x	16	*sofns14x*	*sofns14x* (1)
emm66	15	*sof66*	*sof66* (1)
emm49	11	*sof49*	*sof49* (3)
emm9	10	*sof9*	*sof9* (3)
st448	8	*sof448*	*sof448* (1), *sof3894* (1)
emm68	7	*sof68*	*sof68* (1), *sof4470* (1), *sof4438* (1)
emm60	7	*sof60*	*sof60* (2)
st4935	7	*sof4935*	*sof4935* (1), *sof1881*(1), *sof27G* (1)
emm88	6	*sof88*	*sof1482* (1)
emm25	6	*sof25*	*sof4958* (1), *sof75* (1)
emm8	6	*sof8*	*sof8* (2)
st3018	6	*sof8*	ND
emm48	4	*sof48*	*sof48* (1)
st213	4	*sof213*	none examined
emm63	3	*sof63*	*sof63* (1)
emm27G	3	*sof27G*	*sof27G* (2)
st1160	3	*sof2967*	*sof2967* (1)
st1815	3	*sof75*	ND
st436	3	*sof436*	ND
emm62	2	*sof62*	*sof62* (1)
stA207	2	*sofA207*	*sofA207* (1)
st4532	2	*sof4532*	ND

and it is also evident that these *sof* genes are sometimes conserved between strains of different genetic backgrounds (reflected by different *emm* types and T agglutination patterns).

Sequence similarities shared between the 5′ sequences of certain highly related, yet distinct, *emm* sequence types of some strains were reflected by identity or a high degree of similarity between 5′ *sof* gene sequences, suggesting that these strains may have evolved from a recent common ancestor. Examples of such pairs included *st2967* (*sof2967*)/*st1160* (*sof2967*), *emm90* (*sof90*)/*st833* (*sof90*), *emm8* (*sof8*)/*st3018* (*sof8*), *emm79* (*sof79*)/*emm87* (*sof87*), *emm27G* (*sof27G*)/*st4935* (*sof27G*) and *emm61* (*sof61*)/*st436* (*sof436*). The possible overall relatedness of such strains is further suggested by the total conservation of signal-sequence-encoding regions from both the *sof* and *emm* gene pairs depicted above, which are presumably under no selective pressure. However, most often there was little apparent correlation between homologous *5′ emm* product pairs and their corresponding *5′ sof* product pairs. With the exception of *sof1482* and *sof3875*, which occurred in strains sharing the *emm* sequence type *emm88* (Table 1), the overlaps of identical *sof* sequences between strains within the AOF61 sets and strains within the AOF9/44 *sof* sets are probably not indicative of overall strain relatedness, but most likely reflect inter-strain recombination events that occurred relatively recently.

Distinct 5′ *sof* sequences found within the same *emm* type

In several examples strains within highly conserved *5′ emm* sequence types were characterized by having distinct AOF types and/or *5′ sof* sequence types. These types included the *emm27L/77* and *emm44/61* strains described above, which were further distinguishable by differing T agglutination types (Table 1). A third *sof* sequence found in an *emm44/61* strain was the unique *sof3930* sequence, which correlated with AOF non-typability. The other examples of non-concordant *emm/sof* associations were found within the *emm* types *emm4* (AOF4/*sof4* and AOF NT/*sof2920*), *emm11* (AOF11/*sof11* and AOF25/*sof25*), *emm25* (*sof25*, *sof4958* and *sof75*), *emm68* (AOF68/*sof68*, AOF NT/*sof4470* and AOF NT/*sof4438*), *emm81* (AOF81/*sof81*, AOF NT/*sof1658*, AOF NT/*sof1965*), *emm88* (*sof88* and *sof1482*), *emm89* (AOF89/*sof89*, AOF NT/*sof4835*), *st4935* (*sof4935*, *sof1881*, AOF27G/*sof27G*) and *st448* (*sof448*, *sof3894*). When strains of the same *emm* type were of different *5′ sof* sequence types, it is probable that this would correlate with dissimilar PFGE profiles, indicating divergent genetic lineages as in the two examples shown in Table 2.

DISCUSSION

Epidemiological study of GAS has been primarily based on M serotyping for much of this century. To relate current and past trends of GAS epidemiology, it is therefore logical to develop sequence-based subtyping systems that have a high predictive value for M type specificity. The importance of such a strategy may become more evident during the development of M-type-specific vaccines (Dale, 1999) and possibly for subsequent work analysing the efficiency of different GAS vaccines (Dale *et al.*, 1997). Possibly it will become important for future vaccine formulations to identify potential M epitopes shared between heterologous *emm* sequence types that are capable of eliciting common protective antibody.

It is important to note that although *emm* is often referred to as the M protein gene, it is sometimes referred to as any one of up to three *emm*-like genes that lie at the *vir* locus. In this work *emm* refers to the specific gene amplified by primers 1 and 2 (Whatmore *et al.*, 1994) and it is this specific gene that has been shown in several strains to encode the protein that evokes M-type-specific antibodies (Hollingshead *et al.*, 1986; Robbins *et al.*, 1987; Miller *et al.*, 1988; Mouw *et al.*, 1988; Dale *et al.*, 1993, 1996). Nonetheless, at present it has not been established that the *emm* gene provides the basis of M type specificity in all GAS strains. This is especially true of OF+ strains in which *emm* is usually flanked by two additional '*emm*-like' genes situated at the vir locus (Haanes *et al.*, 1992; Hollingshead *et al.*, 1993; Podbielski, 1993). This study may be the first fairly extensive analysis of the circumstantial coincidence of specific *emm* sequences and M serotypes in OF+ clinical isolates.

This work provides further strong circumstantial evidence that it is the *emm* gene that encodes M type specificity. The first observation is that in a diverse group of GAS including 64 of 68 M-typable strains, the specific M type correlated with a highly specific *emm* sequence type. Second, in many instances the same M and corresponding *emm* type was found among strains judged to be of differing genetic backgrounds on the basis of differing *sof* genes and T agglutination patterns (Table 1). Third, in three circumstances, identical M types and corresponding identical *emm* sequences were found within distinct M type reference strains (see M44/M61, M27L/M77 and PT1658/M81). In each of these three examples, one serotype was established many years prior to the later one, suggesting that inadvertently the later M type reference strain may not have been exhaustively tested against all previous M typing sera. The M27L/M77 and PT1658/M81 results are straightforward and expected on the basis of identical *emm* genes. The basis of the apparent contradiction of *emm44/61* (*sof44* T5/27/44) strains having dual M44/M61 specificity while the *emm44/61* (*sof61* or *sof3930*, T11/12) strains had solely M61 type specificity is unknown. Since the M type 44 and 61 *emm* genes have not been entirely sequenced, it is possible that M44 contains epitopes that are not present in M61, or that even another *emm*-like gene encodes M44-specific epitopes. It should be noted that the M-type reference strains for M44 and M61, besides having identical *5′ emm* sequences encoding at least their first 85 mature residues, also have identical *5′ enn* sequences which map immediately downstream of *emm* (Whatmore *et al.*, 1995). The fourth and perhaps strongest line of circumstantial evidence provided here that *emm* encodes the M-serotype-specific epitopes came from four sets of strains that shared M type identity or partial identity with M type reference strains (Fig. 1, see *st1160* and M2, *st448* and M49, *st2147* and M59, M68.1

and M68). Although these strains had 5′ *emm* gene sequences with significant differences from the M-type reference *emm* sequences, these *emm* genes obviously shared with them identical potential epitope-encoding sequences.

There have been few studies concerning the genetic diversity of strains within GAS M serotypes. Historically, M serotypes have been treated as indicative of strain types. An earlier study documented the identical *emm* types shared among various reference strains and demonstrated their differing genetic backgrounds (Whatmore *et al.*, 1994). This study also clearly demonstrates that presenting M serotypes as strain types is an oversimplification. Besides the M type reference strains discussed above, we have found current clinical isolates with identical M serotypes that vary in their *sof* gene sequences and associated AOF types, *emm* gene sequences and PFGE profiles (Tables 1–3). There have been multiple studies that have assumed M serotypes based upon AOF types (for one example see Colman *et al.*, 1993). This approach is possibly valid for the majority of isolates obtained in developed countries, although at this point it is not possible to be certain. Between roughly August 1999 and November 1999 we analysed more than 80 additional clinical isolates from patients within the US, including two or more isolates within the frequently encountered *emm* types 2, 4, 11, 12, 22, 27L/77 (T13), 28, 48, 58, 75, 82, 87 and 89. In each isolate, there was perfect agreement of *emm* and *sof* sequence designations as determined either by direct sequencing or by comparison of the highly conserved *emm* and *sof* restriction profiles. However, in this study we found strains within the M types 2, 11 and 61 that correlate with AOF types ST2967, 25 and 44 respectively, clearly indicating that M serotypes should not be inferred from AOF typing (Table 1). To our knowledge, this is the first report of strains with M types associated with more than one AOF type. These data are corroborated in strains 1160-99 (M2, AOF2967), 4808-96 (M11, AOF25) and SS511 (M44/61, AOF44), by the presence of distinct *sof* and *emm* gene sequences that are nearly identical to the corresponding serotyping-reference-strain gene sequences. It must be re-emphasized that this remarkable diversity of *sof* types within defined *emm* types would not be expected from a random study of strains within given *emm* types, but is a direct result of our attempts to include genetically diverse GAS strains through examining strains with unusual T type/*emm* associations and from developing countries, where we have previously found a large degree of strain diversity (Jamal *et al.*, 1999; Facklam *et al.*, 1999).

It appears that continued sequence-based analysis of heterologous *sof* and *emm* gene sets that confer identical serological specificity may aid in the identification of the specific epitopes responsible for M type and AOF type specific reactions. The identical Sof 655-residue sequence immediately N-terminal to the fibronectin-binding repeats shared between the three AOF61 strains, as well as the highly conserved 107-residue region shared between *sof1207* and *sof2967* are totally consistent with the previously mapped Sof2 enzymic domain (Fig. 3a, c). This indicates the liklihood that critical type-specific AOF epitopes of these proteins reside within Sof regions corresponding to Sof9 residues 343–810. Since Sof9 residues 356–462 correspond to the apparently critical Sof2967 residues 341–447, it is possible that this 107-residue region represents the sole region determining AOF type specificity. Further work, possibly involving the use of purified protein fragments and site-directed mutagenesis, is required for further elucidation. It is also possible that additional short regions that are conserved among all Sof proteins contain epitopes critical for the AOF reaction. Work involving simple AOF sera absorptions with heterologous Sof proteins should address this possibility.

This work clearly indicates that the probable basis of the OF− phenotype in most classically OF− *emm*/M types is simply the absence of the *sof* gene, which is consistent with previous work demonstrating the absence of *sof*-hybridizing sequences in several OF− M types (Rakonjac *et al.*, 1995) and the absence of *sof* sequences in the type M1 GAS genome (see http://www.genome.ou.edu/strep.html). We were unable to amplify *sof* sequences from various OF− strains (with the exception of *emm12* strains) although the possibility exists that the primers used do not anneal with *sof* sequence types present in some OF− strains. Due to the variability of OF activity within certain strains, we find that *sof* amplification is much more reliable than OF testing for the general deduction of whether an isolate has a classical OF+ or OF− *emm* type (Johnson & Kaplan, 1993). Typically, OF− strains, including M/*emm* types 1, 3, 5, 6, 12, 18, 33 and 56 which were *emm* types found in some of the *sof* negative strains referred to in this study, are designated class I GAS due to their M protein reactivity with defined monoclonal antibodies associated with class I M proteins (Bessen *et al.*, 1989). These strains typically have *emm* and *emm*-like gene arrangements at the *vir* locus categorized as one of the patterns A, B or C based upon their number and their peptidoglycan-spanning-domain sequence (Bessen *et al.*, 1996). The *emm12* isolates are the only class I and/or pattern A–C strains known at this time that have been associated with a specific *sof* sequence.

The results shown in this work indicate that *sof* or *emm* sequence-based analysis is generally more discriminating than serological analysis for subtyping strains. For example, strains within M serotypes 2 and 59, and within AOF types 61 and ST2967, could be further subdivided by *emm* and *sof* sequence differences. There are many examples of AOF NT and M NT strains listed in Table 1, although all were *sof* and *emm* typable (with the single exception of the *emm* NT *sof2* strain 1588-99). Even the *sof9* and sof44 amplicons, which share the identical sequence over their first 1026 bases, are readily distinguishable by further sequence comparison that can be easily obtained with universal *sof* sequencing primers. Also, the apparent full-length conservation among many *sof* genes should allow their identification through conserved *sof* amplicon restriction profiles.

ACKNOWLEDGEMENTS

We thank each of many different investigators for supplying bacterial strains. We are grateful to Harry Courtney for sharing unpublished information with us. We also thank Zhongya Li for her superb technical assistance. G.G. was supported as a recipient of a Career Award from the Italian Fondazione Cariverona Progetto Sanità.

REFERENCES

Beall, B., Facklam, R. & Thompson, T. (1996). Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. *J Clin Microbiol* **34**, 953–958.

Beall, B., Facklam, R., Hoenes, S. & Schwartz, B. (1997). A survey of *emm* gene sequences from systemic *Streptococcus pyogenes* infection isolates collected in San Francisco, California; Atlanta, Georgia; and Connecticut in 1994 and 1995. *J Clin Microbiol* **35**, 1231–1235.

Beall, B., Facklam, R. R., Elliott, J. A., Franklin, A. R., Hoenes, T., Jackson, D., Laclaire, L., Thompson, T. & Viswanathan, R. (1998). Streptococcal *emm* types associated with T agglutination types and the use of conserved *emm* gene restriction fragment patterns for subtyping group A streptococci. *J Med Microbiol* **47**, 893–898.

Bessen, D. E. & Hollingshead, S. K. (1994). Allelic polymorphism of *emm* loci provides evidence for horizontal gene spread in group A streptococci. *Proc Natl Acad Sci USA* **91**, 3280–3284.

Bessen, D., Jones, K. F. & Fischetti, V. A. (1989). Evidence for two distinct classes of streptococcal M protein and their relationship to rheumatic fever. *J Exp Med* **169**, 269–283.

Bessen, D. E., Sotir, C. M., Readdy, T. L. & Hollingshead, S. K. (1996). Genetic correlates of throat and skin isolates of group A streptococci. *J Infect Dis* **173**, 896–900.

Colman, G., Tanna, A., Efstratiou, A. & Gaworzewska, E. T. (1993). The serotypes of *Streptococcus pyogenes* present in Britain during 1980–1990 and their association with disease. *J Med Microbiol* **39**, 165–178.

Courtney, H. S., Hasty, D. L., Li, Y., Chiang, H. C., Thacker, J. L. & Dale, J. B. (1999). Serum opacity factor is a major fibronectin-binding protein and a virulence determinant of M type 2 *Streptococcus pyogenes*. *Mol Microbiol* **32**, 89–98.

Dale, J. B. (1999). Group A streptococcal vaccines. *Infect Dis Clin North Am* **13**, 227–243.

Dale, J. B., Chiang, E. Y. & Lederer, J. W. (1993). Recombinant tetravalent group A streptococcal M vaccine. *J Immunol* **151**, 2188–2194.

Dale, J. B., Simmons, M., Chiang, E. C. & Chiang, E. Y. (1996). Recombinant, octavalent group A streptococcal M protein vaccine. *Vaccine* **14**, 944–948.

Dale, J. B., Cleary, P. P., Fischetti, V. A., Kasper, D. L., Musser, J. M. & Zabriskie, J. B. (1997). Group A and group B streptococcal vaccine development: a round table presentation. *Adv Exp Med Biol* **418**, 863–868.

Dillon, H. C. & Dillon, M. S. A. (1974). New streptococcal serotypes causing pyoderma and acute glomerulonephritis types 59, 60, and 61. *Infect Immun* **9**, 1070–1078.

Facklam, R., Beall, B., Efstratiou, A. & 13 other authors (1999). Demonstration of *emm* typing and validation of provisional M types for group A streptococci. *Emerg Infect Dis* **5**, 247–253.

Fischetti, V. A. (1989). Streptococcal M protein: molecular design and biological behavior. *Clin Microbiol Rev* **2**, 285–314.

Fraser, C. A. M. & Colman, G. (1985). Some provisional types among *Streptococcus pyogenes*. In *Recent Advances in Streptococci and Streptococcal Diseases: Proceedings of the IX Lancefield Symposium on Streptococci and Streptococcal Diseases*, pp. 35–36. Edited by Y. Kimura, S. Kotami & Y. Shiokaswa. Bracknell, UK: Reedbooks.

Gooder, H. (1961). Association of a serum opacity reaction with serological type in *Streptococcus pyogenes*. *J Gen Microbiol* **25**, 347–352.

Haanes, E. J., Heath, D. J. & Cleary, P. P. (1992). Architecture of the *vir* regulons of group A streptococci parallels opacity factor phenotype and M protein class. *J Bacteriol* **174**, 4967–4976.

Hollingshead, S. K., Fischetti, V. A. & Scott, J. R. (1986). Complete nucleotide sequence of type 6 M protein of the group A streptococcus. *J Biol Chem* **261**, 1677–1686.

Hollingshead, S. K., Readdy, T. L., Yung, D. L. & Bessen, D. E. (1993). Structural heterogeneity of the *emm* gene cluster in group A streptococci. *Mol Microbiol* **8**, 707–717.

Hollingshead, S. K., Fischetti, V. A. & Scott, J. R. (1997). Size variation in group A streptococcal M protein is generated by homologous recombination between intragenic repeats. *Mol Gen Genet* **207**, 196–203.

Jamal, F., Pit, S., Facklam, R. & Beall, B. (1999). New *emm* (M protein gene) sequences obtained from group A streptococci isolated from Malaysian patients. *Emerg Infect Dis* **5**, 182–183.

Johnson, D. R. & Kaplan, E. L. (1993). A review of the correlation of T-agglutination patterns and M-protein typing and opacity factor production in the identification of group A streptococci. *J Med Microbiol* **38**, 311–315.

Johnson D. R. & Kaplan, E. L. (1996). *Laboratory Diagnosis of Group A Streptococcal Infections*. Bahrain: World Health Organization.

el-Kholy, A. M., Sorour, A. H., Rotta, J. & Guirguirs, N. (1973). Group A beta hemolytic streptococci in skin lesions among an Egyptian school children population. *J Hyg Epidemiol Microbiol Immunol* **17**, 316–322.

Kreikemeyer, B., Talay, S. R. & Chhatwal, G. S. (1995). Characterization of a novel fibronectin-binding surface protein in group A streptococci. *Mol Microbiol* **17**, 137–145.

Krumwiede, E. (1954). Studies on a lipoproteinase of group A streptococci. *J Exp Med* **100**, 629–638.

Lancefield, R. C. (1962). Current knowledge of the type specific M antigens of group A streptococci. *J Immunol* **89**, 307–313.

Lindgren, P. E., McGavin, M. J., Signas, C., Guss, B., Gurusiddappa, S., Hook, M. & Lindberg, M. (1993). Two different genes coding for fibronectin-binding proteins from *Streptococcus dysgalactiae*: the complete nucleotide sequences and characterization of the binding domains. *Eur J Biochem* **214**, 819–827.

Maxted, W. R., Widdowson, J. P. M., Fraser, C. A., Ball, L. & Bassett, D. C. J. (1973). The use of the serum opacity reaction in the typing of group A streptococci. *J Med Microbiol* **6**, 83–90.

Mejia, L. M., Stockbauer, K. E., Pan, X., Cravioto, A. & Musser, J. M. (1997). Characterization of group A *Streptococcus* strains recovered from Mexican children with pharyngitis by automated DNA sequencing of virulence-related genes: unexpectedly large variation in the gene (*sic*) encoding a complement-inhibiting protein. *J Clin Microbiol* **35**, 3220–3224.

Miller, L., Gray, L., Beachey, E. & Kehoe, M. (1988). Antigenic variation among group A streptococcal M proteins: nucleotide sequence of the serotype 5 M protein gene and its relationship with genes encoding types 6 and 24 M proteins. *J Biol Chem* **263**, 5668–5673.

Mouw, A. R., Beachey, E. H. & Burdett, V. (1988). 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*. *J Bacteriol* **170**, 676–684.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1–6.

Pack, T. D. & Boyle, M. D. (1995). Characterization of a type II'o group A streptococcal immunoglobulin-binding protein. *Mol Immunol* **32**, 1235–1243.

Podbielski, A. (1993). Three different types of organization of the *vir* regulon in group A streptococci. *Mol Gen Genet* **237**, 287–300.

Podbielski, A., Melzer, B. & Lutticken, R. (1991). Application of the polymerase chain reaction to study the M protein(-like) gene family in beta-hemolytic streptococci. *Med Microbiol Immunol* **180**, 213–227.

Rakonjac, J. V., Robbins, J. C. & Fischetti, V. A. (1995). DNA sequence of the serum opacity factor of group A streptococci: identification of a fibronectin-binding repeat domain. *Infect Immun* **63**, 622–631.

Robbins, J. C., Spanier, J. G., Jones, S. J., Simpson, W. J. & Cleary, P. P. (1987). *Streptococcus pyogenes* type 12 M protein gene regulation by upstream sequences. *J Bacteriol* **169**, 5633–5640.

Saravani, G. A. & Martin, D. R. (1990). Opacity factor from group A streptococci is an apoproteinase. *FEMS Microbiol Lett* **56**, 35–39.

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelson, P. A., Murray, B. E., Persing, D. H. & Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33**, 2233–2239.

Whatmore, A. M., Kapur, V., Sullivan, D. J., Musser, J. M. & Kehoe, M. A. (1994). Non-congruent relationships between variation in *emm* gene sequences and the population genetic structure of group A streptococci. *Mol Microbiol* **14**, 619–631.

Whatmore, A. M., Kapur, V., Musser, J. M. & Kehoe, M. A. (1995). Molecular population genetic analysis of the *enn* subdivision of group A streptococcal *emm*-like genes: horizontal gene transfer and restricted variation among *enn* genes. *Mol Microbiol* **15**, 1039–1048.

Widdowson, J. P., Maxted, W. R. & Grant, D. L. (1970). The production of opacity in serum by group A streptococci and its relation with the presence of the M antigen. *J Gen Microbiol* **61**, 343–353.

Zurawski, C. A., Bardsley, M. S., Beall, B., Elliott, J. A., Facklam, R., Schwartz, B. & Farley, M. M. (1998). Invasive group A streptococcal disease in metropolitan Atlanta: a population-based assessment. *Clin Infect Dis* **27**, 150–157.

Received 19 October 1999; revised 6 January 2000; accepted 14 January 2000.