Updated model of group A *Streptococcus* M proteins based on a comprehensive worldwide study

David J. McMillan 2, Pierre-Alexandre Drèze 1, Therese Vu 2, Debra E. Bessen 3, Julien Guglielmini 4,5, Andrew C. Steer 6,7, Jonathan R. Carapetis 8, Laurence Van Melderen 1, Kadaba S. Sriprakash 2, Pierre R. Smeesters 1,2,7 and the M Protein Study Group

1: Laboratoire de Génétique et Physiologie Bactérienne, Institut de Biologie et de Médecine Moléculaires, Faculté des Sciences, Université Libre de Bruxelles, Belgium
2: Bacterial Pathogenesis Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia
3: Department of Microbiology and Immunology, New York Medical College, Valhalla, New York, United States of America
4: Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, Paris, France
5: CNRS, UMR3525, F-75015 Paris, France
6: Centre for International Child Health, Department of Paediatrics, University of Melbourne, Royal Children’s Hospital, Melbourne, Australia
7: Murdoch Children Research Institute, Melbourne, Australia
8: Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, Perth, WA.

Running title: M protein global survey

Keywords: *Streptococcus pyogenes*, M protein, Virulence, Epidemiology, Typing, Vaccine.

Corresponding author

Pierre Smeesters
Laboratoire de Génétique et Physiologie Bactérienne, IBMM
Université Libre de Bruxelles

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1469-0691.12134
© 2012 The Authors Clinical Microbiology and Infection © 2012 European Society of Clinical Microbiology and Infectious Diseases
Abstract

**Background.** Group A *Streptococcus* (GAS) M protein is an important virulence factor and potential vaccine antigen, and constitutes the basis for strain typing (*emm*-typing). Although >200 *emm*-types are characterized, structural data were obtained from only a limited number of *emm*-types. We aim to evaluate the sequence diversity of near-full-length M proteins from worldwide sources and analyse their structure, sequence conservation and classification.

**Methods.** GAS isolates recovered from throughout the world during the last two decades underwent *emm*-typing and complete *emm* gene sequencing. Predicted amino acid sequence analyses, secondary structure predictions and vaccine epitope mapping were performed using MUSCLE and Geneious software.

**Results.** 1086 isolates from 31 countries were analysed, representing 175 *emm*-types. *emm*-type is predictive of the whole protein structure, independent of geographic origin or clinical association. Findings of an *emm*-type paired with multiple, highly divergent central regions were not observed. M protein sequence length, the presence or absence of sequence repeats, and predicted secondary structure was assessed in the context of the latest vaccine developments.
Conclusions. Based on these global data, the M6 protein model is updated to a three representative M protein (M5, M80, M77) model, to aid in epidemiological analysis, vaccine development and M protein-related pathogenesis studies.

Introduction

Amongst bacterial pathogens afflicting humans, group A streptococcus (GAS) is a leading cause of global morbidity and mortality [1]. Colonisation of the respiratory tract or skin by this organism can lead to diseases that manifest in different body sites, and require different modalities of treatment. Of these, Rheumatic Heart Disease (RHD) and serious streptococcal invasive diseases are associated with the greatest global mortality. Much of the GAS associated mortality occurs in low income regions and populations [2].

The M protein is a fibrillar coiled-coil dimer that extends from the bacterial cell wall, and is considered an archetypal Gram-positive surface protein [3]. The M protein is a key virulence factor and major target for GAS vaccine development. M protein inhibits phagocytosis of GAS in the absence of opsonising antibodies, promotes adherence to human epithelial cells and helps the bacterium overcome innate immunity [4]. The multifunctional nature of this protein is also evident from its interactions with numerous host proteins [4], occurring along the entire length of the surface exposed portion of M protein.

Most of the M protein sequence consists of heptad repeat motifs in which the first and fourth amino acids are typically hydrophobic, and are core stabilizing residues within the coiled coil [5]. Heterogeneity in the amino acid sequence of the N-terminal part of M protein, resulting in antigenic diversity, forms the basis of GAS serotyping which was used for many decades [6,
Serotyping has recently been superseded by nucleotide sequencing of the corresponding region, in a scheme called emm-typing [6, 8]. emm-type based surveillance studies show that the diversity of strains circulating in low income settings far exceeds that in high income settings [9, 10]. emm-typing relies on sequencing a small variable portion (10-15%) of the complete emm gene. As a consequence, emm-typing is not informative of the sequence, predicted conformational structure, or functional domains of the remainder of the M protein molecule, which remains largely uncharacterised at the global level.

Another typing method, called emm pattern-typing distinguishes distinct chromosomal architectures (patterns A-C, D and E) based on the presence and arrangement of emm and emm-like genes within the GAS genome [11]. Specific emm-types correlate well with specific emm patterns [12]. emm pattern also correlates well with tissue tropism, although several exceptions have been described [13]. Pattern A-C strains are usually associated with throat infections, pattern D strains are mainly recovered from superficial skin infection (impetigo), while pattern E represents a “generalist” group associated with both tissue sites. Although representing only a small proportion of emm-types, pattern A-C strains have been the most extensively studied [4]. Much of our understanding of M protein structure and function is based on early work on the M6 protein, an emm pattern A-C type [7]. The prototypical M6 protein contains several internal repeat sequences called ‘A’, ‘B’, ‘C’, and ‘D’ repeats. Much less is known of the structure of many other M proteins, particularly those belonging to emm patterns D and E [14].

Although there is increasing interest in GAS vaccine development by global health authorities, including the World Health Organisation, a GAS vaccine remains unavailable. Three M protein-based GAS vaccines are poised to enter, or are progressing through, human
clinical trials. One vaccine candidate incorporates amino terminal, M-type determinants from multiple M-proteins [15], while the others consist of more highly conserved sequence from the C repeat region (CRR) [16-19]. Given the clinical relevance of M protein in molecular epidemiology and GAS virulence, and its importance to vaccine development, a comprehensive unified view of M protein is needed. In this study we fill this knowledge gap by characterizing the complete surface-exposed portions of a large number of M proteins from strains recovered from geographical regions throughout the world.

Materials and Methods

Study profile

Globally distributed GAS isolates recovered during two recent decades (from 1987 to 2008) by the 25 partners of the M-protein study groups were included in the study. Each partner provided bacterial isolates, or genomic DNA representatives of each emm-type in their collection. Most isolates (n=835; 77%) are unique representatives of a particular emm-type per country of isolation. In some cases, two or three isolates of the same emm-type were included if they were collected in different regions of large countries such as USA, Canada, Brazil, and Australia. With one exception, continents and countries were classified according to the geoscheme created by the United Nations Statistics Division [20]; isolates recovered from Hawaii (USA) were artificially included in Oceania because of geographical proximity to the Pacific Islands with which Hawaii shares similar GAS epidemiology [9, 10]. Clinical data was also provided with most isolates. Eight emm-types could not be recovered during the
two last decades in our dataset (Figure 1; Table 1); the sequence of those particular \textit{emm}-types were obtained and described, but were not included in data analysis.

\textbf{Molecular typing}

PCR amplification and sequencing of \textit{emm} genes was performed as previously described [14]. The alignment of the forward and reverse \textit{emm} sequences was performed using the CodonCode Aligner® version 3.7 software with default parameters and were all manually checked. \textit{emm}-type was determined by BLASTn analysis using the CDC \textit{emm}-type database containing 223 \textit{emm}-types [21]. After translation, the predicted amino acid sequences of all M proteins were trimmed from the first amino acid (AA) of the predicted mature protein to the first AA of the D repeat near the COOH-terminal end [14]. The size of mature M proteins (from the first NH2-terminal residue to the Thr of the LPXTG sortase motif) was calculated by adding 54 or 73 residues respectively to the sequence we obtained for the M proteins of patterns E or A-C and D, as described previously [4]. The \textit{emm} pattern of at least one isolate of each of the 168 \textit{emm}-types was experimentally determined following the PCR mapping methodology previously described [22] or deduced from previous publications [12, 14].

\textbf{Bioinformatics}

Multiple alignments of trimmed amino acid sequences belonging to the same \textit{emm}-types were performed using the MUSCLE algorithm with default parameters. The presence of repeat sequences was detected by using T-reks with 3 different percentage similarity (Psim) thresholds (1, 0.9, and 0.7) and extensive manual analysis [23]. M protein annotation and structure prediction was performed with Geneious® 5.6 for one representative of each \textit{emm}-type.
Statistical analysis

Two-tailed student’s T-test were performed using Stata 12 software.

Results

Study population

The final dataset included 1086 GAS isolates representing 175 different emm-types recovered from 31 countries on six continents (Figure 1). Thus, this collection includes 78% of the emm-types described to date [21]. Twenty percent of the 175 emm-types belong to emm pattern A-C, while the remaining are distributed evenly among patterns D and E (Table 1). The number of isolates examined per emm-type ranged from 1 to 25 (mean 6.5) (Table S1). Clinical manifestations were reported for 1019 isolates: invasive diseases (n=365; 35.8%), pharyngitis (n=338; 33.2%) and skin infections (n=233; 22.9%; includes impetigo, wound infections and other unspecified skin infections). The remainder were associated with oropharyngeal carriage (n=46; 4.5%), non-suppurative sequelae (n=13; 1.3%) and other types of infections (n=24; 2.4%).

Updated structural model of M proteins

The size of the predicted mature form of M protein was highly heterogeneous, ranging from 229 to 509 residues. Importantly, M protein length was highly correlated with emm pattern.
proteins of pattern A–C were the longest (average 443 residues; 95% CI 427-463) followed
by pattern D (average 360 residues; 95% CI 353-368) while those of pattern E were the
shortest (average 316 residues; 95% CI 312-320) (Student’s T-test; for 2-way comparisons
among all pattern groups, t < 0.001).

emm sequence data, including detailed annotation of sequence repeats, for one representative
of each of 175 emm-types are available in GenBank (accession numbers JX028599-
JX028772, JX472406). The ‘A’ repeats are defined as amino acid sequence repeats beginning
within the first 50 amino-terminal residues of the mature protein (i.e., emm typing region).
Similarly, ‘B’ repeats are defined as sequence repeats starting between residue 51 and the
beginning of the CRR. The ‘C’ repeats are defined by their homology with a highly conserved
35-residue block (supplementary data S2). Data show that a majority (65%) of M proteins do
not possess ‘A’ repeat sequences. However, ‘A’ repeats are more frequent amongst the
pattern A-C group, whereby ~50% of M proteins have ‘A’ repeats, than amongst the D and E
(33 and 30% respectively). The presence of ‘B’ repeats also correlates with the emm pattern
groupings: 57, 51 and 15% of M proteins of patterns A-C, D and E, respectively, possess ‘B’
repeats. When present, 85% of the ‘B’ repeats consist of only two repeat units in tandem (size
range, 7 to 62 residues); higher numbers of ‘B’ repeat units were almost exclusively
associated with M proteins of the pattern A-C group. Both ‘A’ and ‘B’ repeat sequences
originating from different emm-types were rarely found to share extensive sequence
homology. On the contrary, all M proteins possess a CRR. The number of ‘C’ repeat units
ranges from 1 to 5, with the vast majority of M-types (90%) harboring 3 repeat units.

Based on the data obtained in this study, and on information from published literature [4, 9,
10, 13], we propose a new structural model with 3 representative M proteins (Figure 2). M5,
M80 and M77 proteins were selected as prototypes for the structural characteristics within each emm pattern group. This model provides the advantage of being far more representative of M proteins from organisms recovered worldwide.

**Sequence conservation within an emm-type**

In order to examine sequence heterogeneity from isolates of the same emm-type originating from different geographic regions, we identified all emm-types recovered from at least five locations. 80 emm-types encompassing 900 isolates fulfilled this criterion (Table S3). Sixty-five (81%) emm-types showed intra-emm-type differences in the size of M-proteins (Table S3). Within each emm-type, an average mean of 69% of isolates belonged to the most common size variant. The most prevalent size variant was used as the basis for comparison to other size variants within each emm-type. Comparisons of the 900 protein sequences revealed 408 insertions or deletions. Indels (i.e. insertions or deletions) were found in similar frequencies across all emm pattern groups (data not shown). As classically observed with coiled-coil proteins, 304 (75%) indels involved a sequence stretch that is a multiple of seven residues, and this heptad periodicity increases from the amino- to carboxy-terminal ends of the protein (Figure 3). These observations suggest that strong selective pressures preserve the coiled-coil structure at the carboxy-terminal end of M protein, whereas the amino-terminal extremity may better tolerate variation in its higher order structure.

M proteins assigned to the same emm-type are highly conserved across their surface exposed portions, despite differences in both geographical origins and clinical manifestations (Table S1 and S3). After excluding gaps, M protein sequences of the same emm-type are nearly
identical, with an average pair wise identity ranging from 88% to 100% (Supplementary data S3). The median pairwise identity is 99%. Only two M-types, emm14 (pattern A-C) and emm100 (pattern D) exhibit an average pairwise identity <90%. One or two isolates belonging to each of those two emm-types presented an atypical M protein sequence which was vastly different from the others (protein identity between atypical and typical variants ranging from 63 to 77%). Although many emm-types share highly homologous central regions spanning residue 51 to the CRR, it was extremely rare to find a given emm-type paired with multiple, highly divergent sub-N-terminal domains. Thus the emm-type of an M protein is largely predictive of the structure of the full-length protein, indicating that the emm-typing method is far more informative than previously appreciated.

**M protein conserved vaccine epitopes**

The highly conserved nature of the CRR signifies that CRR-based vaccines can potentially target a wide range of M proteins [16, 24]. One such vaccine candidate, J14, consists of 14 amino acid residues derived from CRR [17]. In this study, 42 J14 variants were identified, including 17 newly recognized variants (Supplementary data S4). Seven J14 variants accounted for 89% of all J14 variants recovered from the 1078 isolates (Supplementary data S5). To prevent bias due to over-representation of particular emm-types, we also examined the distribution of J14 variants in single representatives of each emm-type with similar results. Specific J14 variants clearly segregate with emm pattern. For example, M proteins belonging to pattern A-C almost exclusively contain variant J14.0 in their third C-repeat unit, pattern D proteins contain a mix of both variant J14.0 and J14.1 whereas J14.0 is absent from pattern E proteins (data not shown).
Discussion

This study is the most comprehensive analysis of globally distributed GAS M proteins undertaken. The data provide a significant increase in our understanding of the M protein structure as a whole, a new understanding of the biological relevance provided by older typing tools such as \textit{emm} typing and \textit{emm} pattern determination and insights for the development of future GAS vaccine formulations.

Approximately 75% of \textit{emm}-types belong to the pattern D and E groups. \textit{emm}-types of pattern D and E are also frequently recovered in epidemiologic settings where there is a high GAS-associated mortality burden and a very high diversity of circulating \textit{emm}-types [9, 10, 12, 14, 25]. Despite their epidemiologic relevance, these \textit{emm}-types have not been as extensively characterised as those of the pattern A-C group. The structure of M6 protein served well in the past as representative of M proteins. However with increased knowledge of the structure of additional M proteins, it has become evident that extrapolations based on M6 protein are limited. First, M6 is a pattern A-C \textit{emm}-type, which collectively account for only \textasciitilde 20% of \textit{emm}-types. Secondly, M6 protein has five ‘A’ and five ‘B’ repeats and is non-representative of even the pattern A-C group because half of pattern A-C \textit{emm} types lack ‘A’ repeats altogether and most possess fewer ‘B’ repeat units. Third, the size of M6 protein is smaller than most of the A-C pattern \textit{emm}-type. Our data also demonstrate that ‘A’ repeats are rarely found in the pattern D and E \textit{emm}–types while ‘B’ repeats are sometimes present, but usually as a single tandem repeat.

The \textit{emm}-typing region, despite its short length, is largely predictive of the whole M protein sequence independent of clinical association or geographical origin. This finding suggests that
emm-typing can be used to infer not only the N-terminal portion of the protein, but the entire surface exposed portion as well. M proteins are multifunctional, having roles in preventing phagocytosis, mediating adherence to host cells, and intracellular invasion, often through binding to human host products such as fibrinogen, factor H, albumin, IgA and IgG, plasminogen and others [26]. Many host proteins are bound to distinct regions or domains within M proteins. Therefore, the emm-typing system may be predictive of a unique array of biological functions for each emm-type.

In fungi, size variation generated by intragenic tandem repeats within surface protein genes allows for rapid adaptation to the environment and/or evasion of the host immune system [27]. In GAS, as previously described for M6 organisms [28], size differences in M protein from different isolates is also a common feature, largely a result of differences in the number of sequence repeat units. The M6 protein size mutants display heterogeneity in their antigenic and opsonogenic epitopes [28]. Our data show that intra-emm-type size variation occurs for most emm-types, and it is evenly distributed across the three emm pattern groups (data not shown). The significant differences in M protein length observed between the three emm pattern groups, and the close uniformity of M protein length within each emm pattern group, have not been previously described. Combined with knowledge that the emm pattern groups are associated with different clinical manifestations [13], it is tempting to speculate that M protein length underlies distinct functional attributes, perhaps related to a capacity to bind different subsets of host proteins. Future studies can assess whether M protein size is an attribute that impacts the virulence potential and clinical manifestations of GAS.

Our study is relevant to M protein-based vaccine design. A new multivalent antigen containing amino-terminal fragments from 30 M proteins showed unexpected in vitro cross
protection against isolates expressing M proteins not included in the immunizing antigen [15]. Extensive cross-protection was not demonstrated with a 26-valent antigen produced previously by the same laboratory. The primary difference in the composition of the two vaccines is a significant increase in the number (from 11 to 18) of sequences representing M proteins belonging to pattern E, in the 30-valent vaccine. Antibodies elicited by the 30-valent vaccine were tested against isolates belonging to 40 emm-types. Rates of cross protection, measured by an opsonophagocytosis assay, differed by pattern group (pattern A-C, 60%; pattern D, 45%; and pattern E, 84%) [15]. The underlying mechanism for cross protection against non-vaccine types is not yet understood. However, these data suggest that the emm-types belonging to the different pattern groups might differ in their ability to induce cross-protective antibodies.

Another vaccine approach utilizes protective epitopes within the highly conserved CRR, and aims to confer broad protection against all GAS strains [19, 24]. This and other studies have shown that there are many variants of J14 sequences [17, 29]. Our current study confirms that virtually all C3 repeat units harbor J14.0 or J14.1 variants and that a small number of J14 variants are predominant within a global collection of isolates.

M protein size and structure are characteristic of the emm pattern group to which they belong. The pattern classification, based on the content of emm gene forms and their chromosomal arrangement, was also recognized as a strong marker of preferred tissue sites of infection by GAS [13, 29, 30]. From this exhaustive study, we propose that three M proteins - M5, M80, M77 - belonging to the three emm pattern groups A-C, D and E, respectively, best represent the structures and possible host-pathogen interactions mediated by M proteins. This new model is likely to be a valuable tool for epidemiological, molecular and vaccine studies.

© 2012 The Authors Clinical Microbiology and Infection © 2012 European Society of Clinical Microbiology and Infectious Diseases
Footnote

This study was presented in part at the 29th annual meetings of the European Society for Pediatric Infectious Diseases in The Hague, The Netherlands in June 7-11, 2011; at the 7th World Congress of the World Society for Pediatric Infectious Diseases in Melbourne, Australia in November 16-19, 2011; at the Australian Society for Microbiology 2012 Annual Scientific Meeting in Brisbane, Australia in July 1–4, 2012 and at 52nd ICAAC meeting in San Francisco, USA in September 9-12, 2012.

Acknowledgements

The contributing members of the M protein study group (in addition to the authors of this paper) include Michael Batzloff and Rebecca Towers from Australia; Herman Goossens and Surhbi Malhotra-Kumar from Belgium; Luiza Guilherme and Rosangela Torres from Brazil; Donald Low and Allison Mc Geer from Canada; Paula Krizova from Czech Republic; Sawsan El Tayeb from Egypt; Joe Kado from Fiji; Mark van der Linden from Germany; Guliz Erdem from Hawai; Alon Moses and Ran Nir-Paz from Israel; Tadayoshi Ikebe and Haruo Watanabe from Japan; Samba Sow and Boubou Tamboura from Mali; Bard Kittang from Norway; José Melo-Cristino and Mario Ramirez from Portugal; Monica Straut from Romania; Alexander Suvorov and Artem Totolian from Russia; Mark Engel, Bongani Mayosi and Andrew Whitelaw from South Africa; Jessica Darenberg and Birgitta Henriques Normark from Sweden; Chuan Chiang Ni and Jiunn-Jong Wu from Taiwan; Aruni De Zoysa and Androulla Efstratiou from UK; Stanford Shulman and Robert Tanz from USA.

We also would like to sincerely acknowledge Bernard Beall for proofreading of the MS and for managing the very useful emm-typing database from CDC.
Conflict of interest

No conflict of interest.

Funding

This work was supported by the European Society for Clinical Microbiology and Infectious Diseases, European Society for Paediatric Infectious Diseases, Fonds National de la Recherche Scientifique (Belgium), Fonds Brachet and Fondation Van Buuren (Belgium), Australian National Health and Medical Research Council, National Institutes of Health (AI-065572). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Contributors

PRS was the primary coordinator of data collection, analysis, and writing. DJM, LVM, and KSS supervised data collection, analysis, and writing. TV and PAD were primarily involved in laboratory experiments and data collection. DB, JG, ACS, and JRC were primarily involved in data collection, analysis and writing. All authors contributed substantially to the preparation of the paper.

References


22 McDonald MI, Towers RJ, Fagan P, Carapetis JR, Currie BJ. Molecular typing of streptococcus pyogenes from remote aboriginal communities where rheumatic fever is
common and pyoderma is the predominant streptococcal infection. *Epidemiol Infect.* 2007; **135:** 1398-1405.


24 Bessen D, Fischetti VA. Synthetic peptide vaccine against mucosal colonization by group a streptococci. I. Protection against a heterologous m serotype with shared c repeat region epitopes. *J Immunol.* 1990; **145:** 1251-1256.

25 Smeesters PR, Dramaix M, Van Melderen L. The emm-type diversity does not always reflect the m protein genetic diversity--is there a case for designer vaccine against gas. *Vaccine.* 2010; **28:** 883-885.


28 Jones KF, Hollingshead SK, Scott JR, Fischetti VA. Spontaneous m6 protein size mutants of group a streptococci display variation in antigenic and opsonogenic epitopes. *Proc Natl Acad Sci U S A.* 1988; **85:** 8271-8275.


**Table 1: emm pattern groupings for 184 emm-types.**

<table>
<thead>
<tr>
<th>emm pattern</th>
<th>emm-types</th>
<th>Number of emm-types for set of 184 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>2, 4, 8, 9, 11, 13, 15, 22, 25, 27, 28, 44/61, 48, 49, 50/62, 58, 60, 63, 66, 68, 73, 75, 76, 77, 78, 79, 82, 84, 87, 88, 89, 90, 92, 94, 96, 102, 103, 104, 106, 107, 109, 110, 112, 113, 114, 117, 118, 124, st106M, st212, st213, st804, st833, st1207, st1389, st1731, st2147, st2460, st2463, st2904, st6735, st7406, st11014, stknb1, stMTH81, stNS292, stNS554, stNS555, st104</td>
<td>68 (37)</td>
</tr>
<tr>
<td>REA</td>
<td>st211, st1815</td>
<td>2 (1)</td>
</tr>
<tr>
<td>ND</td>
<td>31, st22, st1692, st1969, st9505, stil103, stpa57</td>
<td>7 (4)</td>
</tr>
</tbody>
</table>

Several identical emm-types were originally assigned two numbers: emm-type 44 is identical to emm-type 61 (emm-type 44/61), emm-types 50 and 62 (50/62), emm-types 65 and 69 (65/69), emm-types 38 and 40 (38/40). 223 emm-types are listed in the CDC database (September 18, 2012) [21]. REA, rearranged emm pattern (atypical amplification patterns). ND, not determined. *, emm-types from strains isolated prior to 1987. emm-types not included in this study, but whose emm pattern grouping was previously established [31], are indicated in italics. Isolates of emm-types 54 and st854 (underlined) are associated with more than one emm pattern group. Note that emm-types 7, 10, 16, 20, 21, 35, and 45 do not exist.
Figure legend

* List of countries with respective number of isolates in brackets: Argentina (5), Australia (137), Belgium (46), Brazil (105), Canada (69), Chile (5), Czech Republic (17), Germany (50), Egypt (39), Ethiopia (4), Fiji (55), India (51), Israel (67), Japan (12), Kenya (1), Malaysia (1), Mali (58), Mexico (7), New Zealand (1), Norway (19), Papua New Guinea (2), Portugal (21), Romania (22), Russia (15), South Africa (22), Sweden (45), Taiwan (37), The Gambia (1), United Kingdom (22), USA (138, including 83 in mainland and 55 in Hawaii), Venezuela (1). Geographical origin is unknown for 3 isolates. SDSE, *Streptococcus dysgalactiae* subspecies *equisimilus*. The eight *emm*-types recovered prior to 1987 are as follows: *emm*-types 17, 34, 37, 38, 46, 47, 51 and 72.

Figure 1: Study profile

Three representative M proteins model.

Three representative M proteins (M5, M80 and M77) were selected as prototypes for the structural characteristics within each *emm* pattern group. M protein length and the size of the repeat and non-repeat regions are drawn to scale. Pattern A-C *emm*-types represent the longest M proteins, with a (hyper)variable portion of about 230 residues. In comparison, pattern D and E proteins possess a (hyper)variable portion of ~150 and 100 residues, respectively. The ‘A’ repeats are absent from the vast majority of M proteins belonging to the pattern D and E groups. The ‘B’ repeats are present in most of the pattern A-C and D *emm*-types, but absent from most of the pattern E *emm*-types. Thirty-five conserved residues constitute the ‘C’ repeat unit. Consecutive ‘C’ repeat units are sometimes separated by a seven residue unit called ‘C’.
repeat linker (See supplementary data S2). Twenty percent of the M proteins (such as M80) do not possess non-helicoidal amino terminus. This proportion is 10%, 19% and 25% amongst the pattern A-C, D and E emm-types respectively. The portion of the protein considered by the emm-typing method is represented.

Figure 3: Insertion-deletion (indel) characteristics of M proteins belonging to the same emm-type.

Intra-emm-type alignments that include 900 M protein sequences of 80 emm-types reveal 408 indels. More than half of the indels (n=224; 55%) are located in the CRR (C Repeat Region). The remaining indels are equally distributed between regions corresponding to the 50 amino-terminus proximal residues (n=85; 21%; emm-type determinant) and from residue 51 to the beginning of the CRR (n=99; 24%; sub-N-terminal, central region). The number of indels having a heptad periodicity increases from the amino-terminal (36% of indels) to the carboxy-terminal (CRR; 99% of indels) regions of M proteins, whereas half (50%) of the indels from the central region of M protein involve multiples of seven residues.
Figure 1: Study profile

1278 representative GAS isolates received from 25 different epidemiological collections

200 excluded
64 emm sequence not obtained
123 discrepant results
5 not coding emm sequences
8 SDSE associated emm sequences

1078 contemporary GAS isolates included from 31 different countries*, representing 167 emm-types

8 emm-types recovered prior to 1987

175 emm-types
Figure 2: Three representative M proteins model.
Figure 3: Size variation amongst M proteins belonging to the same *emm*-types: the indels characteristics vary according to the portions of the protein.

- **No heptad periodicity**
- **Multiple of heptad periodicity**

Number of indels

- 50 amino-terminus residues
- From residue 51 to CRR
- CRR

Portions of the M protein