Structure-based group A streptococcal vaccine design: Helical wheel homology predicts antibody cross-reactivity among streptococcal M protein-derived peptides

Received for publication, September 27, 2019, and in revised form, February 5, 2020. Published, Papers in Press, February 6, 2020, DOI 10.1074/jbc.RA119.011258

Michelle P. Aranha, Thomas A. Penfound, Jay A. Spencer, Rupesh Agarwal, Jerome Baudry, James B. Dale, and Jeremy C. Smith

From the ¹Department of Biochemistry and Cellular and Molecular Biology and the ²Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, Tennessee 37996, the ³University of Tennessee/Oak Ridge National Laboratory Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, the ⁴Department of Medicine, Division of Infectious Diseases, University of Tennessee Health Science Center, Memphis, Tennessee 38163, and the ⁵Department of Biological Sciences, University of Alabama in Huntsville, Huntsville, Alabama 35899, United States

Edited by Peter Cresswell

Group A streptococcus (Strep A) surface M protein, an α-helical coiled-coil dimer, is a vaccine target and a major determinant of streptococcal virulence. The sequence-variable N-terminal region of the M protein defines the M type and also contains epitopes that promote opsonophagocytic killing of streptococci. Recent reports have reported considerable cross-reactivity among different M types, suggesting the prospect of identifying cross-protective epitopes that would constitute a broadly protective multivalent vaccine against Strep A isolates. Here, we have used a combination of immunological assays, structural biology, and cheminformatics to construct a recombinant M protein–based vaccine that included six Strep A M peptides that were predicted to elicit antisera that would cross-react with an additional 15 nonvaccine M types of Strep A. Rabbit antisera against this recombinant vaccine cross-reacted with 10 of the 15 nonvaccine M peptides. Two of the five nonvaccine M peptides that did not cross-react shared high sequence identity (≥50%) with the vaccine peptides, implying that high sequence identity alone was insufficient for cross-reactivity among the M peptides. Additional structural analyses revealed that the sequence identity at corresponding polar helical-wheel heptad sites between vaccine and nonvaccine peptides accurately distinguishes cross-reactive from non–cross-reactive peptides. On the basis of these observations, we developed a scoring algorithm based on the sequence identity at polar heptad sites. When applied to all epidemiologically important M types, this algorithm should enable the selection of a minimal number of M peptide–based vaccine candidates that elicit broadly protective immunity against Strep A.

Streptococcus pyogenes, or group A streptococcus (Strep A), causes noninvasive infections such as pharyngitis, impetigo, and cellulitis, as well as invasive and life-threatening infections such as bacteremia, streptococcal toxic shock syndrome, and necrotizing fasciitis (1). Poststreptococcal sequelae that may follow Strep A infections include acute rheumatic fever, rheumatic heart disease, and glomerulonephritis (2). Rheumatic heart disease and invasive Strep A infections are estimated to cause over 500,000 deaths/year, with the vast majority of the disease burden in low- and middle-income countries where the mortality rate is also disproportionately high (3).

To reduce the health burden and the economic impact of these infections, efforts to develop an effective vaccine against Strep A have been ongoing (4). There are a number of potential vaccine candidates (4–8), and in particular, there has been considerable progress in the preclinical and clinical development of multivalent M protein–based vaccines (9–13). These vaccines have been formulated using recombinant fusion proteins containing multiple peptides from the variable N terminus of epidemiologically prevalent M types of Strep A (10, 11). One obstacle to this approach has been the perception that protection against Strep A infections is type-specific and that the diversity of M types (>200) may prevent the development of broadly efficacious M protein–based vaccines. However, there is recent evidence that N-terminal M peptides evoke antibodies that cross-react with heterologous M types of Strep A (11), and natural infection with Strep A elicits cross-opsonic antibodies (14). This is supported by the observation that the majority of M proteins are members of structurally and functionally related clusters, indicating that immunity may be a combination of cluster-specific and type-specific antibody responses (15).

We previously showed that the relatedness of M peptides within a single M cluster could be exploited by using a computational structure-based approach to select five N-terminal M peptides that elicited antibodies that reacted with all 17 M peptides in the cluster and opsonized 15 of 17 M types of Strep A (16). In the present study we examined 117 M proteins, respon-
sible for 92% of Strep A infections globally (15, 17). The N-terminal hypervariable region (HVR) of the M protein (residues 1–50) contains epitopes that elicit antibodies with the greatest bactericidal activity and is least likely to elicit antibodies that cross-react with host tissues (4, 9–11, 18). Therefore, we confine our analysis to residues 1–50 and have redefined the sequence-based M clusters (15) to include seven N-terminal clusters (NTCs) containing all 117 M peptides. In this initial study, we focused on a cluster of 21 peptides (NTC6) that shared significant sequence similarity. Using a structure-based computational approach, we designed an NTC6 vaccine that elicited antibodies in rabbits that cross-reacted with 10 of 15 nonvaccine peptides. A posteriori analysis of the nonreactive versus cross-reactive peptides revealed that sequence identity within the polar heptad sites of the predicted α-helical domains within the N-terminal region is a strong predictor of cross-reactivity. The application of this new approach to the structure-based design of multivalent vaccines may result in more broadly cross-reactive and efficacious M protein vaccines.

Results

Sequence-based clustering

117 M types were divided into M peptide NTCs (Fig. 1) by constructing a phylogenetic sequence-based tree of the N-terminal 50 amino acids of the mature proteins that define the HVR region (Geneious, version 9.1.6). The seven N-terminal clusters were designated based on the calculated common branches. The overall phylogenetic relationships of the N-terminal peptides bear some resemblance to the previous description of M clusters based on the whole M sequences (15). In this study we limited the analysis to the NTC6 cluster, which contains 21 different M types that collectively accounted for 33% of all Strep A isolates from children with pharyngitis in North America (19), many of which are prevalent globally (17).

Subclusters of immunologically similar M peptides

A functional matrix of antibody binding and cross-reactivity among the NTC6 peptides, which describes the inhibition of antibody binding to 12 NTC6 M peptides by all 21 peptides in the cluster, was developed by performing ELISA inhibition experiments. The relational matrix of experimentally obtained antibody binding between NTC6 peptides (Table S1) was sub-clustered using $k$ means into seven immunologically related peptide groups (Fig. 2). To resolve the optimal number of clusters, $k$ was varied from 2 to 9, and the maximal average silhouette coefficient was obtained for $k = 7$ (Fig. S1). The silhouette coefficient is considered as measure of quality of the structure of a cluster; in other words, it informs us how closely related objects in a cluster are and how distinct or well-separated a cluster is from other clusters (20). Clusters with high silhouette
coefficients are well-separated and were considered to contain M peptides more likely to cross-react than clusters with low-silhouette coefficients. For example, from Fig. 2, M84 and M89 belonging to FC2 \((s = 0.46)\) would be predicted to cross-react with greater probability than M1, M9, and M227 belonging to FC1 \((s = 0.19)\).

**Clusters of structurally and immunologically similar M peptides**

The structures of the 21 M peptides were calculated using the *de novo* computational framework, PEP-FOLD3 (21) Fig. S2A). To identify structural features that are most relevant to correctly predicting antibody binding, we used the 44 structure-based protein descriptors from the Molecular Operating Environment program (22) as independent variables to describe the PEP-FOLD3 models and the columns of the relational matrix as the dependent variables in a multiple regression analysis with feature ranking. The feature ranking resulted in a subset of 20 top-ranked descriptors from the supervised regression approach that were then used in \(k\)-means clustering of the NT C6 M peptides to obtain experimentally informed structure-based clusters (Fig. 3). An overlay of the models that have been grouped together for each cluster shows that models belonging to the same cluster share greater structural similarity than models belonging to different clusters (Fig. S2B).

There was considerable overlap between the experimentally informed clusters and antibody-binding function-based clusters (Rand index = 0.77). FC3, FC4, FC6, and FC7 in Fig. 2 resembled and respectively corresponded to experimentally informed clusters 1, 5, 4, and 7 in Fig. 3. This demonstrates that the experimentally informed structure-based top 20 descriptors can adequately detect and isolate different M peptides with similar antibody-binding function.

Next, the experimentally informed clusters (Fig. 3) were considered together with the functional data (Fig. 2) to select a minimal number of peptides predicted to elicit broad cross-reactivity against the remaining 15 M peptides in NTC6. The \(s\) coefficient was used as the main criterion in this selection. However, in Figs. 2 and 3, for some of the clusters the values of \(s\) do not indicate strong clustering of data. Therefore, additional information was taken into account in selecting the final vaccine candidates. In the case of the experimentally informed clusters, if three or more PEP-FOLD3 models of different M types each shared a cluster, they were considered to be more likely to cross-react than those with fewer than three models in a same cluster. As an example, in Fig. 3 we observed that cluster 6 contained only one model each of M114, M112, M73, M50, and M49, and these M types were considered unlikely to be immunologically related. In contrast, M1, M238, and M239 had all five models placed in cluster 4 and were considered likely to cross-react. Based on these considerations and also taking into account epidemiological prevalence, six M types were predicted to elicit cross-reactive antibodies against the remaining 15 M peptides in NTC6 (Table 1). The vaccine construct is shown in Fig. 4A.

**Hexavalent NTC6 vaccine evoked antibodies against vaccine and nonvaccine peptides**

In three rabbits immunized with the hexavalent NTC6 vaccine, immune sera contained significant levels of antibodies against all vaccine peptides as well as significant levels of antibodies against 10 of the 15 nonvaccine peptides (Fig. 4B). All preimmune sera resulted in antibody titers of 100 against all 21 peptides. For this analysis, the nonvaccine peptides were considered to be cross-reactive when two of the three immune sera displayed antibody titers of at least 800, which is an 8-fold
increase over preimmune antibody levels. Using these criteria, the NTC6 vaccine antisera were only marginally cross-reactive with five nonvaccine peptides: M15, M50, M112, M183, and M238 (Fig. 4B).

Cross-reactivity and sequence identity were correlated but not without exceptions

Given the experimental results in Fig. 4B, it was of interest to conduct an a posteriori analysis of factors that distinguish antibody cross-reactivity with nonvaccine peptides. To understand the extent of correlation between sequence identity and the antibody cross-reactivity, we performed a pairwise sequence alignment between the 21 M peptides and calculated their mutual sequence identities using the EMBOSS Needle program (23) (Table 2). The maximum pairwise sequence identities between nonvaccine peptides and vaccine peptides extracted from Table 2 are listed in Table 3.

Five nonvaccine M types (M84, M124, M114, M175, and M232) have pairwise sequence identities greater than 60% with at least one or more vaccine M peptides and exhibited cross-reactivity with the antibodies elicited by the vaccine. The nonvaccine M types with low pairwise sequence identity (<40%) with one or more vaccine peptides were M15 and M50, and these M peptides displayed marginal cross-reactivity with antibodies elicited by the vaccine. Thus, there is a moderate positive correlation between the pairwise sequence identity among the vaccine peptides and the nonvaccine peptides and the cross-reactivity, as indicated by the Spearman correlation ($\rho = 0.56, p$ value = 0.05) (Fig. 5). However, when nonvaccine peptides shared between 40 and 60% sequence identity with any of the vaccine peptides, sequence identity could not be used to infer whether the antibodies raised by the vaccine peptide would cross-react with the nonvaccine peptide. For instance, M238, M112, and M183 share 45–60% sequence identities with vaccine peptides and were only marginally cross-reactive with the vaccine antisera. On the other hand, nonvaccine peptides M9, M49, M102, M227, and M239 share a similar degree of sequence identity with at least one of the vaccine M peptides and did cross-react with the NTC6 vaccine antisera. Thus, although significant sequence identity is a useful consideration for antibody cross-reactivity, it alone cannot reliably predict cross-reactivity.

Coiled-coil heptad repeat sequence identity

Although overall sequence identity was correlated with cross-reactivity, one might expect the three-dimensional structure of the antigen in vivo to also be of importance. The monomer structures predicted by PEP-FOLD3 have extensive helical content, and in many cases the peptide was predicted to fold on itself in a manner resembling coiled-coil structures. The $\alpha$-helical coiled-coil structure is also evidenced by available crystal structures (24, 25) of M1, M2, M22, M28, and M49. Therefore in a further analysis, we assume that this coiled-coil extends into the N terminus. Heptad repeats form coiled coils, and the positions in the heptad repeat are labeled a–g. The core-forming positions of the coiled-coil (a and d) are usually occupied by hydrophobic residues whereas the remaining, solvent-exposed positions (b, c, e, f, and g) are dominated by hydrophilic residues.

The probability of shared epitopes between vaccine and nonvaccine M types increases with increased sequence identity. Therefore, instead of simply comparing overall sequence identity, we considered only the region within the N-terminal 1–50 residues that is predicted to be significantly coiled-coil (MARCOIL (26) assigned coiled-coil probability $\geq 2\%$) and calculated the sequence identity between each of the corresponding heptad sites of vaccine and nonvaccine M types. The heptad repeat projected on a helical wheel generated using DRAWCOIL 1.0 (27) for M types that share high overall sequence identity with vaccine type M1 (>40%) is shown in Fig. 6.

Empirical scoring scheme for predicting cross-reactivity

We developed an empirical scoring scheme that penalizes low sequence identity and rewards high pairwise sequence identity at the corresponding heptad sites between the vaccine and nonvaccine M types. Pairwise alignments of the vaccine and nonvaccine peptides at the heptad positions was performed. In calculating the empirical pair score between vaccine and nonvaccine M types, we left out the hydrophobic positions (a and d) of the heptad. This was done in part because little sequence variation is found at the hydrophobic core sites of the heptad (a and d) that are restricted in terms of the types of residue that can occupy those positions, favoring hydrophobic residues such as leucine, isoleucine, valine, and alanine. Heptad positions a and d are therefore conserved among many M types and do not serve as good discriminants between cross-reactive and non–cross-reactive types. Most of the NTC6 M types are leucine zippers, i.e. they contain an abundance of leucines at the d site. Another reason for not taking a and d into account is that on the surface of the bacterium, and possibly in the vaccine, these residues are buried and are presumably least accessible for antibody contact and recognition.

For the remaining polar heptad positions, the following scores were given when comparing vaccine and nonvaccine peptides: 20 $\leq$ sequence identity $< 30$, score: 1.0; 30 $\leq$ sequence identity $< 40$, score: 1.5; 40 $\leq$ sequence identity $< 50$, score: 2.0; 50 $\leq$ sequence identity $< 60$, score: 3.0; sequence identity $\geq 60$, score: 3.5; and sequence identity $< 20$, score: $-1.0$. An empirical

| Vaccine type | Nonvaccine type predicted to be covered | Coverage predicted based on FC or SC | M types covered |
|--------------|----------------------------------------|-------------------------------------|----------------|
| M49          | M84                                   | FC                                 | 5              |
| M232         | M175                                  | SC                                 | 3              |
| M227         | M238                                  | FC/SC                              | 3              |
| M239         | M114                                  | SC/FC                              | 4              |
| M239         | M124                                  | FC                                 | 3              |
| M114         | M102                                  | SC/FC                              | 6              |
| M112         | M124                                  | SC                                 | 3              |
| M112         | M15                                  | SC                                 | 6              |
| M112         | M175                                  | SC                                 | 3              |
| M175         | M183                                  | SC                                 | 6              |
| M175         | M49                                  | SC                                 | 3              |
| M49          | M9                                   | SC                                 | 3              |
| M49          | M50                                  | SC                                 | 3              |
| M50          | M238                                 | SC                                 | 3              |
| M238         | M118                                 | SC                                 | 3              |
| M118         | M73                                 | SC                                 | 3              |
| M118 + M73   | M118                                 | SC                                 | 3              |

**Table 1**

Selected vaccine M peptides and predicted antibody cross-reactivity against nonvaccine M peptides

FC stands for antibody binding function-based cluster described in Fig. 3, and SC stands for experimentally informed structure-based clusters using the 20 top-ranked protein features described in Fig. 4.
score was then calculated by summing up the scores at each of the polar heptad positions. As an example, the pairwise sequence identity of nonvaccine M238 with vaccine peptides at each of the heptad positions is shown in Table 4. The empirical pairwise score for M238 with vaccine peptide M1 is thus calculated:

\[
\frac{2}{100} + \frac{1}{100} + \frac{0}{100} + \frac{0}{100} + \frac{1}{100} + \frac{1}{100} = 6.
\]

Low sequence identity at heptad positions is correlated with low antibody cross-reactivity against M peptides that share high overall sequence identity

Empirical pairwise scores based on the sequence identity at the heptad sites between all vaccine and nonvaccine peptides for NTC6 are given in Table S2. Table 5 shows the maximum...
Heptad identity predicts cross-reactivity among M proteins

Table 3
Maximum pairwise sequence identity between NTC6 nonvaccine peptides and vaccine peptides

| Nonvaccine peptide | Maximum pairwise sequence identity (%) | Corresponding vaccine peptide |
|--------------------|----------------------------------------|-------------------------------|
| M84                | 73                                     | M89                          |
| M124               | 71                                     | M2                           |
| M114               | 67                                     | M73                          |
| M175               | 67                                     | M2                           |
| M232               | 62                                     | M89                          |
| M49                | 57                                     | M118                         |
| M239               | 56                                     | M1                           |
| M102               | 55                                     | M77                          |
| M238s             | 54                                     | M1                           |
| M112s             | 50                                     | M77                          |
| M9                 | 47                                     | M118                         |
| M227               | 46                                     | M1                           |
| M183s             | 46                                     | M118                         |
| M15s              | 35                                     | M118                         |
| M50s              | 25                                     | M2                           |

* The M types against which marginal cross-reactivity was observed.

empirical pairwise score between nonvaccine and vaccine M types for the NTC6 peptides based on heptad homologies. The score distinguished between cross-reactive and marginally cross-reactive M peptides. When the score for a pair was ≥10.5, antibody cross-reactivity was observed. Conversely, when a score was <10.5, cross-reactivity was not observed. Fig. 7 shows the geometric mean antibody titers obtained after immunizing three rabbits with the NTC6 vaccine plotted against the maximum heptad identity score of each nonvaccine peptide. This indicates that a score of ≥10.5 predicted all five marginally cross-reactive peptides in the NTC6 cluster. Spearman’s rank-order correlation is (ρ = 0.75, p value = 0.001).

An illustration of the use of heptad identity is the correct prediction of lack of cross-reactivity of M238 and vaccine peptide M1, despite high overall sequence identity between the two (54%). These results indicate that mutual heptad identity between the vaccine and nonvaccine M types is an important indication of the degree of shared epitopes, as well as an important determinant of antibody cross-reactivity among sequence-similar M peptides.

Discussion

The M protein of Strep A is a major protective antigen and a leading vaccine target. A significant challenge to the development of M protein–based Strep A vaccines has been the number of different M types (>200) identified to date. Each emm type is defined by the 5’ sequence that encodes the variable N terminus of the mature protein (28). Recently it has been shown that the majority of M proteins can be clustered based on similar structural and functional characteristics (15), leading to a new paradigm suggesting that M antibody responses may be M cluster–specific as well as type-specific. The overall goal of structure-based approaches to vaccine design is to focus on the structural and functional similarities of M proteins to identify the fewest number of M peptides needed to formulate vaccines that will elicit immune responses against the majority of epidemiologically important M types of Strep A. In the present study, we have used both 3D structure-based and sequence-based approaches to analyze the antibody cross-reactivity among Strep A N-terminal M peptides from one sequence-related cluster. Because our subunit vaccines contain peptides from the N-terminal regions of the M proteins (11), we have redefined the M clusters using only the first 50 amino acid residues, as opposed to our previous studies that clustered the sequences of the entire protein (15).

In principle, if a complete antibody-binding matrix were available, it could simply be clustered to obtain immunologically related M types. However, the Table S1 is incomplete because the human intravenous immunoglobulin (IVIG) antibodies did not react with all peptides, so we identified structural features that correlated with immunological data and then used those features to cluster M types. The end result was the identification of clusters of peptides that shared immunologically relevant structural features. This also allowed us to test the hypothesis that M types that share similar structural features are immunologically related.

The starting point for the current structure-based analysis was sequence similarities among M peptides within the same cluster. However, because sequence similarity does not directly take into account conformational similarity of putative epitopes, we then extended the analysis to determine 3D structures of the peptides and to compare their shape-dependent properties using cheminformatics methods. All 21 peptides within the NTC6 cluster were then subclustered based on similarities among 3D models and functional antibody inhibition studies. Six peptides, each representing one of the subclusters, were selected to construct the recombinant NTC6.1 vaccine. The vaccine elicited significant levels of cross-reactive antibodies against 10 of the 15 nonvaccine peptides within NTC6.

In an a posteriori analysis of the cross-reactivity results, we found that a method distinguishing cross-reactive and non–cross-reactive pairs that only considers the coiled-coil domain
within the N-terminal region and that calculates the homology between the residues at the corresponding polar heptad sites was more discriminating than predicting cross-reactivity from calculated 3D monomer structures. This method makes the assumption that the peptide epitopes are in a helical conformation. Although coiled-coil structures are strongly predicted for the M proteins, there is some evidence that the N-terminal residues are less likely to be coiled coils (24). The heptad identity–based scoring scheme incorporates simple structural data into sequence-based considerations and is a computationally efficient and effective way of predicting antibody cross-reactive immunogenicity among Strep A M peptides.

Two factors that must be considered while determining antibody cross-reactivity are that the primary requirement for antibody binding or antigenicity is surface accessibility (29, 30) and that 80% of all naturally occurring antibody epitopes studied so far are discontinuous in sequence, as expected because the antibody sees a contiguous surface. The heptad identity–based

Table 4
Example of heptad position identity scoring between nonvaccine and vaccine M types

The nonvaccine type was M238. The vaccine types were M1, M2, M77, M118, M73, and M89.

| Heptad position | M238–M1 | M238–M2 | M238–M73 | M238–M77 | M238–M89 | M238–M118 |
|-----------------|---------|---------|----------|----------|----------|----------|
| a               | 50      | 10      | 10       | 29       | 0        | 7        |
| d               | 43      | 42.9    | 43       | 38       | 25       | 29       |
| b               | 67      | 11.1    | 9        | 29       | 13       | 9        |
| c               | 33      | 11.1    | 0        | 11       | 9        | 14       |
| e               | 29      | 9.1     | 22       | 14       | 11       | 8        |
| f               | 14      | 0       | 13       | 8        | 8        | 9        |
| g               | 25      | 11.1    | 11       | 29       | 25       | 43       |
| Sequence identity range | Number of polar heptad sites within the sequence identity range |
| ≥20 and <30    | 2       | 0       | 1        | 2        | 0        | 1        |
| ≥30 and <40    | 1       | 0       | 0        | 0        | 1        | 0        |
| ≥40 and <50    | 0       | 0       | 0        | 0        | 0        | 0        |
| ≥50 and <60    | 0       | 0       | 0        | 0        | 0        | 0        |
| ≥60             | 1       | 0       | 0        | 0        | 0        | 0        |
| <20             | 1       | 5       | 4        | 3        | 4        | 4        |
| Empirical pairwise score | 6       | −5      | −3       | −1       | −2       | −3       |

Figure 6. MARCOIL heptad position assignment of the N-terminal regions of M1, M238, M227, and M239 projected onto a helical wheel generated using DRAWCOIL 1.0. Only heptad repeat regions with probability percentages of >2% were considered for helical wheel representation. The view is from the N-terminal region to the C-terminal region. Heptad repeat positions are labeled a–g. The following color scheme is followed: polar positive, blue; polar negative, red; polar neutral, orange; and nonpolar aliphatic, gray.
Heptad identity predicts cross-reactivity among M proteins

Experimental procedures

Vaccine design strategy

The workflow for the selection of vaccine candidates is presented in Fig. 8. Sequence-based clustering of the N-terminal 50 residues was performed, and one cluster, NTC6, containing 21 M types was selected for study. We aimed to select a minimum number of M peptides that would elicit antibody responses against the majority of streptococcal M peptides within NTC6. The sequences of the HVRs of the 21 M types are provided in Table S3, and their pairwise sequence identities were obtained using the EMBOSS program Needle (Tables 2 and 3). 3D structural models were created of monomeric forms of these peptides upon which cheminformatic analysis was performed. Partial data on experimental antibody binding were also derived (antibody cross-reactivity and the development of a functional matrix). Combining the experimental antibody-binding data with the cheminformatic results involved multiple regression calculations and resulted in the generation of experimentally informed cheminformatics descriptors. The candidate experimental vaccine was chosen based on the experimental and computational results.

Antibody cross-reactivity and the development of a functional matrix—A functional matrix of antibody binding and cross-reactivity among the NTC6 peptides was developed using naturally acquired human antibodies in commercially available IVIG (Gammagard liquid; Baxalta US Inc., Westlake Village, CA). IVIG contains purified IgG that is pooled from the serum of multiple donors and contains antibodies against many of the common M antigens. Functional relatedness was assessed by performing ELISA-inhibition experiments performed in duplicate (16) using all 21 NTC6 peptides as inhibitors of IVIG antibody binding to 12 of the NTC6 peptides (Table S1). The 12 peptides were selected based on direct ELISA results, indicating that the IVIG contained levels of antibodies sufficient to perform inhibition assays. Relative functional “distances” were expressed as 100 − (percentage of inhibition), where 0 = complete functional identity and 100 = no functional activity, i.e. no inhibition. k-means clustering of antibody binding between NTC6 peptides was used to identify “subclusters” of functionally related peptides, and the optimal number of clusters was determined by a silhouette analysis (20).

Structural models of monomeric M peptides—In designing the vaccine we used our previous approach (16) of modeling the structures of the linear peptides in monomeric form in aqueous solution and identifying physicochemical similarities in the resulting computed structures as an indication of cross-reactivity. PEP-FOLD3 (21) is a computational framework that allows de novo structure prediction for linear peptides between 5 and 50 amino acids. PEP-FOLD3 was used to predict the 3D stuc-
tures of the N-terminal regions (50 residues long) of all 21 NTC6 M peptides as monomers (Fig. S2A). Five top-scoring conformations per M peptide were retained for calculation of structural descriptors. The Molecular Operating Environment program package (22) was used to calculate structure-based physicochemical descriptors (44 nonzero variance descriptors) such as peptide surface area, patches of excess charge, hydrophobicity, volume, and shape of the 1–50 residue 3D models of M peptides that were obtained using PEP-FOLD3 with default parameters. For initial quality check, clusters based on all the nonzero descriptors are shown in Fig. S3.

Machine learning to identify clusters of structurally and immunologically related M types—Instead of using all of the nonzero variance descriptors as in our previous study (16), we chose only the top descriptors that correlated the PEP-FOLD3–predicted structures with antibody binding. We assumed that the use of these descriptors would improve the probability of identifying shared structural similarities and shared epitopes. We used supervised machine learning employing multiple regression to rank and identify the descriptors that contributed the most to the M peptide structure–activity relationship. The experimental inhibition of antibody binding to each of the 12 NTC6 M types in the relational matrix (Table S1) was used as a response or dependent variable, and the 44 scaled protein descriptors were used as predictors/independent variables in univariate multiple regression models. The univariate regression allowed us to rank the 44 descriptors per regression model and identify a subset of higher-ranked descriptors that appeared more frequently than others in the 12 regression models.

Following the multiple regression analysis, the subset of the 20 top-ranked cheminformatic descriptors from the supervised regression approach was used in k-means clustering of the NTC6 M peptides to obtain subclusters. The new subclusters thus obtained were compared with those that were obtained from clustering experimental antibody-binding data, i.e. the clusters based on the 12 response variables. We observed some correlation between the subgroups of M peptides defined by the k-means clustering based on the 20 physicochemical descriptors and those based on immunological relatedness. We were thus able to use antibody-binding data in a supervised learning approach to identify important descriptors that were then used in an unsupervised k-means clustering scheme to identify groups of M peptides that were predicted to be both structurally and functionally related.

NTC6 vaccine construction, immunization of rabbits, and detection of antibodies—Six NTC6 peptides containing residues 1–50 of M1, M77, M89, M2, M73, and M118 were selected as vaccine components based on the structural and antibody-binding predictions described above. A recombinant hybrid protein vaccine containing the six M peptides joined in tandem without linkers was produced from extracts of Escherichia coli containing pUC57, into which was inserted a chemically synthesized hybrid gene (Genscript, Piscataway, NJ), using methods previously described (11). The synthetic gene was also designed to encode an upstream T7 promoter and a 3’ polyhistidine motif followed by a stop codon. Three rabbits were immunized with 200 µg of protein on alum via the intramuscular route at time 0, 4 weeks, and 8 weeks. Serum was obtained prior to immunization and 2 weeks after the final injection. Antibody levels in preimmune and immune sera were assayed using the 21 NTC6 M peptides as solid-phase antigens, as previously described (16). Assays were repeated once to confirm the results of the initial ELISA. Antibody titer was defined as the reciprocal of the last serum dilution resulting in an A450 of ≥0.2. All research involving animals was reviewed and approved by the University of Tennessee Health Science Center Institutional Animal Care and Use Committee.
Heptad identity predicts cross-reactivity among M proteins

Prediction of coiled-coil regions of M peptides to determine helical wheel homology among M types

After the vaccine design and testing phase, we searched for 3D structural correlates distinguishing cross-reactive from nonreactive peptide pairs. A simple approach makes use of the predominantly α-helical coiled-coil dimeric structure of the M protein (18, 31). A standard, canonical coiled-coil structure consists of two α-helices twisting around each other with their side chains interlocking in a “knobs” into “holes” packing. The regular meshing of knobs into holes requires recurrence of the side-chain residue types every seven residues along the helix interface (32). Various tools exist to predict the structures of coiled-coil regions in protein to a level of detail that permits the assignment of the individual residues to the positions of the heptad repeat. One such tool is MARCOIL, which calculates posterior probabilities of hidden Markov models and has been reported to offer the best combination of sensitivity and speed compared with other similar tools (26, 33). The knowledge of the heptad register using tools like MARCOIL can eliminate the need for homology-based methods for modeling coiled-coil proteins, because the structure of coiled coil, unlike almost any other known protein fold, can be computed from parametric equations if the heptad assignment is known.

We confirmed the MARCOIL prediction of coiled-coil domains and the assignment of individual residues in a sequence to the heptad by comparing it with the knobs-in-holes interactions in experimental 3D X-ray crystallographic structures recognized by the SOCKET program (34). The SOCKET program can both identify the repeating knobs-into-holes structural motif and can use this information to assign oligomer order (number of helices), orientation (parallel, anti-parallel, and mixed), and heptad register for the coiled-coil. We found the prediction by MARCOIL and the actual heptad assignment calculated by SOCKET for the three available M dimer crystal structures (PDB entries 2OTO, 5HZP, and 5HYT) to be identical, thus validating further use of MARCOIL (Table S4 and Fig. S4, A–C). Therefore, the heptad assignment of the HVR of all 21 M types was made using MARCOIL. Table S5 provides the sequence predicted to be coiled-coil within the 1–50 residues, along with their starting heptad register and the sequence length. Additionally, the sequences at the heptad sites for each of the M peptides are provided in Table S6.

References

1. Lamagni, T. L., Darenberg, J., Luca-Harari, B., Siljander, T., Efstratiou, A., Henriques-Normark, B., Vuopio-Varkila, J., Bouvet, A., Creti, R., Ekelund, K., Koliou, M., Reinert, R. R., Stathi, A., Strakova, L., Ungureanu, V., et al. (2008) Epidemiology of severe Streptococcus pyogenes disease in Europe. J. Clin. Microbiol. 46, 2359–2367 CrossRef Medline

2. Haggar, A., Nerlich, A., Kumar, R., Abraham, V. I., Brahmadathan, K. N., Ray, P., Dhanda, V., Joshua, J. M., Mehra, N., Bergmann, R., Chbatwal, G. S., and Norrby-Teglund, A. (2012) Clinical and microbiologic characteristics of invasive Streptococcus pyogenes infections in north and south India. J. Clin. Microbiol. 50, 1626–1631 CrossRef Medline

3. Carapetis, J. R., Steer, A. C., Mulholland, E. K., and Weber, M. (2005) The global burden of group A streptococcal diseases. Lancet Infect. Dis. 5, 682–694 CrossRef Medline

4. Dale, J. B., Batzloff, M. R., Cleary, P. P., Courtney, H. S., Good, M. F., Grandi, G., Halperin, S., Margarit, I. Y., McNeil, S., and Pandey, M. (2016) Current approaches to group A streptococcal vaccine development. In Streptococcus pyogenes: Basic Biology to Clinical Manifestations, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma

5. Vekemans, J., Gouveia-Feis, S., kim, J. H., Excler, J.-L., Smeesters, P. R., O’Brien, K. L., Van Beneden, C. A., Steer, A. C., Carapetis, J. R., and Kaslow, D. C. (2019) The path to group A streptococcal vaccines: World Health Organization research and development technology roadmap and preferred product characteristics. Clin. Infect. Dis. 69, 877–883 CrossRef Medline

6. Davies, M. R., McIntyre, L., Mutreja, A., Lacey, J. A., Lees, J. A., Towers, R. J., Duchêne, S., Smeesters, P. R., Frost, H. R., Price, D. J., Holden, M. T. G., David, S., Giffard, P. M., Worthing, K. A., Seale, A. C., et al. (2019) Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. Nat. Genet. 51, 1035–1043 CrossRef Medline

7. Fritzner, A., Sen, B. M., Min, D. B., Hanner, M., Gelmann, D., Noiges, B., Henics, T., Schulze, K., Guzman, C. A., Goodacre, I., von Gabain, A., Nagy, E., and Meinke, A. L. (2010) Novel conserved group A streptococcal proteins identified by the antigenome technology as vaccine candidates for a non–M protein–based vaccine. Infect. Immun. 78, 4051–4067 CrossRef Medline

8. Steer, A. C., Carapetis, J. R., Dale, J. B., Fraser, J. D., Good, M. F., Guilherme, L., Moreland, N. J., Mulholland, E. K., Schofel, D., and Smeesters, P. P. (2016) Status of research and development of vaccines for Streptococcus pyogenes. Vaccine 34, 2953–2958 CrossRef Medline

9. Dale, J. B. (1999) Multivalent group A streptococcal vaccine designed to optimize the immunogenicity of six tandem M protein fragments. Vaccine 17, 193–200 CrossRef Medline

10. Hu, M. C., Walls, M. A., Stoop, S. D., Reddish, M. A., Beall, B., and Dale, J. B. (2002) Immunogenicity of a 26-valent group A streptococcal vaccine. Infect. Immun. 70, 2171–2177 CrossRef Medline

11. Dale, J. B., Penfound, T. A., Chiang, E. Y., and Walton, W. J. (2011) New 30-valent M protein–based vaccine evokes cross-opsinetic antibodies against non-vaccine serotypes of group A streptococci. Vaccine 29, 8175–8178 CrossRef Medline

12. Kotloff, K. L., Corrett, M., Palmer, K., Campbell, J. D., Reddish, M. A., Hu, M. C., Wasserman, S. S., and Dale, J. B. (2004) Safety and immunogenicity of a recombinant multivalent group A streptococcal vaccine in healthy adults: phase 1 trial. JAMA 292, 709–715 CrossRef Medline

13. McNeil, S. A., Halperin, S. A., Langley, J. M., Smith, B., Warren, A., Sharratt, G. P., Baxendale, D. M., Reddish, M. A., Hu, M. C., Stoop, S. D., Linden, I., Fries, L. F., Vink, P. E., and Dale, J. B. (2005) Safety and immunogenicity of 26-valent group A streptococcal vaccine. Infect. Immun. 73, 1114–1122 CrossRef Medline

14. Frost, H. R., Laho, D., Sanderson-Smith, M. L., Licciardi, P., Donath, S., Curtis, N., Kado, J., Dale, J. B., Steer, A. C., and Smeesters, P. R. (2017) Immune cross-opsonization within emm clusters following group A streptococci skin infection: broadening the scope of type-specific immunity. Clin. Infect. Dis. 65, 1523–1531 CrossRef Medline

15. Sanderson-Smith, M., De Oliveira, D. M., Guglielmini, J., McMillan, D. J., Vu, T., Holen, J. K., Henningham, A., Steer, A. C., Bessen, D. E., Dale, J. B.,...
Heptad identity predicts cross-reactivity among M proteins

Curtis, N., Beall, B. W., Walker, M. J., Parker, M. W., Carapetis, J. R., et al. (2014) A systematic and functional classification of Streptococcus pyogenes that serves as a new tool for molecular typing and vaccine development. J. Infect. Dis. 210, 1325–1338 CrossRef Medline

Dale, J. B., Smeesters, P. R., Courtney, H. S., Penfound, T. A., Hohn, C. M., Smith, J. C., and Baudry, J. Y. (2017) Structure-based design of broadly protective group a streptococcal M protein–based vaccines. Vaccine 35, 19–26 CrossRef Medline

Steer, A. C., Law, L., Matatolu, L., Beall, B. W., and Carapetis, J. R. (2009) Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. Lancet Infect. Dis. 9, 611–616 CrossRef Medline

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

Shulman, S. T., Tanz, R. D., Dale, J. B., Beall, B., Kabat, W., Kabat, K., Cederlund, E., Patel, D., Rippe, J., Li, Z., Sakota, V., and North American Streptococcal Pharyngitis Surveillance Group (2009) Seven-year surveillance of North American pediatric group A streptococci isolates. Clin. Infect. Dis. 49, 78–84 CrossRef Medline

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 CrossRef Medline

Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., and Tuffery, P. (2016) PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Res. 44, W449–454 CrossRef Medline

Vilar, S., Cozza, G., and Moro, S. (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr. Top. Med. Chem. 8, 1555–1572 CrossRef Medline

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., and Lopez, R. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47, W636–W641 CrossRef Medline

McNamara, C., Zinkernagel, A. S., Macheboeuf, P., Cunningham, M. W., Nizet, V., and Ghosh, P. (2008) Coiled-coil irregularities and instabilities in group A Streptococcus M1 are required for virulence. Science 319, 1405–1408 CrossRef Medline

Buffalo, C. Z., Bahn-Suh, A. J., Hirakis, S. P., Biswas, T., Amaro, R. E., Nizet, V., and Ghosh, P. (2016) Conserved patterns hidden within group A Streptococcus M protein hypervariability recognize human C4b-binding protein. Nat. Microbiol. 1, 16155–16155 CrossRef Medline

Delorenzi, M., and Speed, T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18, 617–625 CrossRef Medline

Grigoryan, G., and Keating, A. E. (2008) Structural specificity in coiled-coil interactions. Curr. Opin. Struct. Biol. 18, 477–483 CrossRef Medline

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

Benjamin, D. C., Berzofsky, J. A., East, I. J., Gurd, F. R., Hannum, C., Leach, S. J., Margoliash, E., Michael, J. G., Miller, A., Prager, E. M., Reichlin, M., Sercaz, E. E., Smith-Gill, S. J., Todd, P. E., and Wilson, A. C. (1984) The antigenic structure of proteins: a reappraisal. Annu. Rev. Immunol. 2, 67–101 CrossRef Medline

Novotný, J., Handschumacher, M., Haber, E., Brucoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A., and Rose, G. D. (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc. Natl. Acad. Sci. U.S.A. 83, 226–230 CrossRef Medline

Smeesters, P. R., McMillan, D. J., and Srirakachar, K. S. (2010) The streptococcal M protein: a highly versatile molecule. Trends Microbiol. 18, 275–282 CrossRef Medline

Lupas, A. N., and Gruber, M. (2005) The structure of α-helical coiled-coils. In Advances in Protein Chemistry, pp. 37–38, Elsevier Science Publishing Co., Inc., New York

Gruber, M., Söding, J., and Lupas, A. N. (2006) Comparative analysis of coiled-coil prediction methods. J. Struct. Biol. 155, 140–145 CrossRef Medline

Walshaw, J., and Woolfson, D. N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. 307, 1427–1450 CrossRef Medline