Analysis of Structure and Function of Putative Surface-Exposed Proteins Encoded in the *Streptococcus pneumoniae* Genome: A Bioinformatics-Based Approach to Vaccine and Drug Design

Daniel J. Rigden,[1] Michael Y. Galperin,[2] and Mark J. Jedrzejas[3]*

[1]National Center of Genetic Resources and Biotechnology, Cenargen/Embrapa, Brasília, Brazil, D.F. 70770–900; [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA; [3]Children's Hospital Oakland Research Institute, Oakland, California 94609, USA

* Correspondence should be addressed to: Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland, CA 94609. Phone 510–450–7932. Fax 510–450–7910. Email mjedrzejas@chori.org

TABLE OF CONTENTS

1040-9238/03/$.50

ABSTRACT: *Streptococcus pneumoniae* is the most common cause of fatal community-acquired pneumonia, middle ear infection, and meningitis. The prevention and treatment of this infection have become a top priority for the medical-scientific community. The present polysaccharide-based vaccine used to immunize susceptible hosts is only ~60% effective and is ineffective in children younger than 2 years of age. The new conjugate vaccine, based on the engineered diphtheria toxin coupled to polysaccharide antigens, is approved only for use in children under 2 years of age to treat invasive disease. While penicillin is the drug of choice to treat infections secondary to *S. pneumoniae,* increasing numbers of bacterial strains are resistant to penicillin as well as to broad spectrum antibiotics such as vancomycin. Thus, there is a need to identify new strategies to prevent and treat diseases caused by to *S. pneumoniae*.

In this article, we summarize the utilization of the recently available *S. pneumoniae* genomic information in order to identify and characterize novel proteins likely located on the surface of this Gram-positive pathogenic bacterium. Because only a limited number of surface proteins of *S. pneumoniae* have been characterized to date, this information provides new insights into the pathogenesis of this organism as well as highlights possible avenues for its treatment and/or prevention in the future. The review is divided into two sections.

First, we briefly summarize current information about known surface-exposed proteins of *S. pneumoniae*. This is followed by the illustration of procedures for the identification of new putative surface-exposed proteins. These have signal peptides required for their extra-cytoplasmic transport and/or additional signature sequences. Some of these will be *S. pneumoniae* virulence factors. The signature sequences we have chosen are those leading to protein binding to choline present on the bacterial surface, attachment to peptidoglycan of the cell wall, or anchoring to lipids of the cytoplasmic membrane. All these signatures are indicative of binding of proteins to the surface of this organism.

Secondly, we illustrate the application of bioinformatics and modeling tools to these selected proteins in order to provide information about their likely functions and preliminary three-dimensional structure models. The focal point of the analysis of these proteins, their sequences, and structures is the evaluation of their antigenic properties and possible roles in pathogenicity. The information obtained from the genome analysis will be instrumental in the development of a more effective prophylactic and/or therapeutic agents to prevent and to treat infections due to *S. pneumoniae*.

I. INTRODUCTION

A. Surface-Exposed Bacterial Proteins

The objective of this review is to (**1**) summarize the information regarding cell-surface proteins of *Streptococcus pneumoniae* that has become available from the analysis of the sequence of the pneumococcal genome (Tettelin *et al*., 2001; Dopazo *et al*., 2001; Hoskins *et al*., 2001), (**2**) evaluate and, where possible, update functional assignments of these proteins, and (**3**) analyze and, if feasible, model their likely three-dimensional structures. We hope that summarizing this information will lead to a better overall understanding of the pathogen and its interactions with the human host. The results of such studies may help in the development of new therapeutic agents such as antibiotics and/or prophylactic agents such as vaccines. For example, the information summarized here may foster the development of specific inhibitors for the enzymes encoded in the genome and/or identification of important epitopes of antigenic proteins for a vaccine development.

The improvement of the understanding of the processes of protein secretion and cell wall biogenesis in Gram-positive bacteria has resulted in the identification of several distinct groups of surface-exposed proteins: (**1**) those covalently bound to cross-bridges of the peptidoglycan structures of the cell wall, (**2**) those electrostatically bound to the choline moieties of (lipo)teichoic acid, and (**3**) those directly anchored in the cytoplasmic membrane of the bacterium. Signature sequence motifs for each of these groups of proteins have been identified, which allowed computational analysis of putative surface-exposed proteins on the genome scale. In this review, we discuss the bioinformatics methods used for the identification of surface-exposed proteins and specifically demonstrate their use in the detailed analysis of the peptidoglycan-bound proteins in pneumococci.

B. *Streptococcus pneumoniae* as a Pathogen

S. pneumoniae is the most common cause of fatal community-acquired pneumonia, middle ear infection, and meningitis (Mufson, 1990). Disease rates are especially high in young children, the elderly, and immuno-compromised individuals with predisposing conditions such as asplenia or AIDS (Gray *et al*., 1980; Johnston, 1991; Musher, 1991). The mortality rate in the U.S. alone is approximately 40,000 per year, higher than that caused by any other bacterial disease (Alexander *et al*., 1994; Anon., 1985; Fedson and Musher, 1994). Therefore, prevention and treatment of this infection has become a top priority for the medical-scientific community working in this field (Cohen, 1994). One additional reason for such a high priority is the emergence of antibiotic resistance in *S. pneumoniae* (Novak *et al*., 1999).

The present polysaccharide-based vaccine used to immunize susceptible hosts is only ~60% effective and is ineffective in children less than 2 years of age. This vaccine contains 23 purified capsular polysaccharide antigens and is still produced by Merck and Company, Inc. and Lederle Laboratories. In 1983 it replaced an earlier 14-valent vaccine. In the year 2000 another vaccine was introduced to market that is designed for the use in toddlers and children to prevent invasive pneumococcal diseases, including bacteremia and meningitis. It is the first pneumococcal vaccine approved for use in children under 2 years of age (U.S. Food and Drug Administration, Annual Report FY2000) and is manufactured by the Lederle Laboratories Division of American Cyanamid Company. It is also the first pneumococcal conjugate vaccine made from the polysaccharides coupled to a nontoxic, engineered form of the diphtheria toxin protein. Its tradename is Prevnar and the toxin-conjugated polysaccharides consist of seven of most common capsular polysaccharide antigens, together accounting for approximately 80% of the invasive disease in infants. The U.S. Food and Drug Administration Annual Report FY2000 clearly indicates that 'this vaccine is not indicated for use in adults or as a substitute for other pneumococcal polysaccharide vaccines approved for high-risk children over the age of 2'.

While penicillin is the drug of choice to treat infections secondary to *S. pneumoniae,* increasing numbers of bacterial strains are resistant to penicillin as well as to broad spectrum antibiotics such as vancomycin. Thus, despite the availability of two vaccines and additional ones being under development, there is still a need to identify new strategies to prevent and to treat diseases due to *S. pneumoniae*.

C. Sequence of *Streptococcus pneumoniae* Genome

The complete genome sequence of the capsular serotype 4 isolate of *S. pneumoniae* (designated TIGR4) and nonencapsulated strain R36A derived from the capsular type 2 clinical isolate strain D39 (designated R6) have been completed recently (Tettelin *et al*., 2001; Hoskins *et al*., 2001) and are publicly available in GenBank, for example, at the NCBI Microbial Genomes web site (www.ncbi.nih.gov/PMGifs/Genomes/micr.html). Yet another nearly completed genomic sequence of strain G54 of type 19 clinical isolate is available at www.gwer.ch/pneumo (Dopazo *et al*., 2001). The genome sequence of *S. pneumoniae* TIGR4 strain (available also at the TIGR web site, www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=bsp) consists of a single circular chromosome with a G+C content of ~40%. In the original analysis of the TIGR4 genome, biological roles were assigned to 64% of the predicted proteins, 16% of predicted proteins matched proteins of unknown function, and 20% had no database match (Tettelin *et al*., 2001). The analysis of the R6 and G54 genomic sequences afforded similar results to those of the TIGR4 genome (Dopazo *et al*., 2001; Hoskins *et al*., 2001).

S. pneumoniae has a wide substrate utilization range for sugars and substituted nitrogen compounds and a correspondingly large number of membrane transporters, including ATP-dependent ones (ABC transporters). The extracellular enzymes of *S. pneumoniae* include a variety of glycosidases that metabolize polysaccharides and hexosamines, providing sources of carbon and nitrogen for the bacterium. There is also a group of secreted hydrolases (hyaluronidases, neuraminidases, and endoglycosidases) that likely degrade the host polymers (mucins, glycolipids, hyaluronan) and probably the organism's own capsule (M. J. Jedrzejas, unpublished results). The availability of genomic sequences for the three pneumococcal strains, encapsulated TIGR4, nonencapsulated R6, and G54, provides an unique opportunity to compare the sequence data among the three different strains.

The preliminary identification of the pneumococcal surface antigens was initiated recently by computational analysis of the genomic sequences of *S. pneumoniae* (Hoskins *et al.,* 2001; Tettelin *et al*., 2001) and continued in several subsequent studies (Wizemann *et al*., 2001; Glass *et al*., 2002). These analyses are largely based on the functional data obtained in the pregenomic era, when only a few surface-exposed pneumococcal proteins were identified and characterized in significant detail (Jedrzejas, 2001).

II. EXPERIMENTALLY CHARACTERIZED PNEUMOCOCCAL SURFACE PROTEINS AND THEIR SIGNATURE SEQUENCES

A. Surface of *Streptococcus pneumoniae*

In addition to the polysaccharide capsule, *S. pneumoniae* displays on its surface numerous proteins, the majority of which are virulence factors that contribute to the pathogenesis of this organism. Such proteins participate in specific interactions with human host tissues, thereby facilitating bacterial survival, helping the organism spread within host tissues, and concealing the bacterial surface from the host's defense mechanisms. Due to these properties, many surface-exposed proteins are potential targets for the design of prophylactic agents such as vaccines and some of them, the enzymes, could serve as targets for the design of therapeutic drugs.

The structure and function of the better understood surface molecules of pneumococci were reviewed recently (Table 1) (Jedrzejas, 2001). Surface-exposed proteins that have already been identified using classic, non-genome-based methods include hyaluronate lyase (Hyl) (Jedrzejas, *et al*., 2002; Li *et al*., 2000; Berry *et al*., 1994), pneumolysin (Ply) (Kelly and Jedrzejas, 2000a, b; Feldman *et al*., 1992), two neuraminidases (NanA and NanB) (Berry *et al*., 1996; Camara *et al*., 1994; Lock *et al*., 1988), major autolysin (LytA) (Medrano *et al*., 1996; Usobiaga *et al*., 1996), choline binding protein A (CbpA)/pneumococcal surface protein C (PspC) (Brooks-Walter *et al*., 1999; Rosenow *et al*., 1997; Cundell *et al*., 1995), pneumococcal surface antigen A (PsaA) (Lawrence *et al*., 1998; Berry and Paton, 1996; Sampson *et al*., 1994), and pneumococcal surface protein A (PspA) (Jedrzejas *et al*., 2001; McDaniel *et al*., 1991) (Plate 1a)*. These proteins can be grouped together based on their mechanism of attachment to pneumococci and their corresponding sequence signatures. The groups are (1) choline-binding proteins, (2) those covalently attached to peptidoglycan, (3) proteins directly attached to lipid of the cytoplasmic membrane, and (4) histidine triad family macromolecules (Plate 1b) (Jedrzejas, 2001; Wizemann *et al*., 2001).

B. Choline-Binding Proteins

These proteins are attached to the pneumococcal cell via terminal choline residues of the teichoic/lipoteichoic acids that are present on the surface of this bacterium (Plates 1a,b). In pneumococci, surface-exposed proteins of this group contain at their C-termini multiple (usually around 10) copies of a repeated segment of around 20 amino acid residues, originally described as glucan-binding domain (Banas *et al*., 1990; Wren, 1991) and now usually referred to as choline-binding domain (Fernandez-Tornero *et al*., 2001). This repeat region is typically connected to the N-terminal part of the protein by a proline-rich flexible linker. Known examples of choline-binding proteins include PspA, PspC, CbpA, and several others.

In the 744 amino acid long PspA protein (SP0117, accession number AAK74303), for example, the leader peptide consists of 31 amino acid residues and is followed by a 400 aa coiled-coil domain (Plate 2a), which is most likely responsible for pneumococcal anticomplementary properties (Jedrzejas *et al*., 2000). This functional module of PspA is followed by a 80 aa proline-rich linker. Finally, the carboxy-terminal region of PspA contains 10 copies of a 20 aa repeat region (choline-binding domain, Plate 2b):

^{525}KTGWKQENGM WYFYNTDGSM^{544}
^{545}AIGWLQNNGS WYYLNANGAM^{564}
^{565}ATGWVKDGDT WYYLEASGAMK^{585}
^{586}ASQWFKVSDK WYYVNSNGAM^{605}
^{606}ATGWLQYNGS WYYLNANGDM^{625}
^{626}ATGWLQYNGS WYYLNANGDM^{645}
^{646}ATGWAKVNGS WYYLNANGAM^{665}
^{666}ATGWAKVNGS WYYLNANGSM^{685}
^{686}ATGWVKDGDT WYYLEASGAMK^{706}
^{707}ASQWFKVSDK WYYVNGLGAL^{726}

Although first sequence analyses and alignments of the choline-binding domain appeared more than 10 years ago (Banas *et al*., 1990; Wren, 1991), its three-dimensional structure was solved only recently (Fernandez-Tornero *et al*., 2001; 2002). It turned out that a stable choline-binding unit is a boomerang-like structure, consisting of two monomers, each of which is formed by six beta-hairpins, corresponding to individual 20 aa choline-binding repeats (Plate 2) (Fernandez-Tornero *et al*., 2001;2002). The "putative cell wall binding repeat" (PF01473) entry in a recent release of the Pfam database (www.sanger.ac.uk/Software/Pfam, Bateman *et al*., 2002), a collection of protein multiple sequences alignments and profile hidden Markov models, contains an alignment of 1373 sequences of the choline-binding domains found in proteins from various low G+C Gram-positive bacteria and their phages. Searching the choline-binding protein segment against the Pfam database is an easy and convenient way to identify potential new choline-binding domains.

C. Proteins Covalently Attached to Peptidoglycan

Attachment of streptococcal proteins to the peptidoglycan is catalyzed by sortase, an extracellular

* Plates appear following page 147.

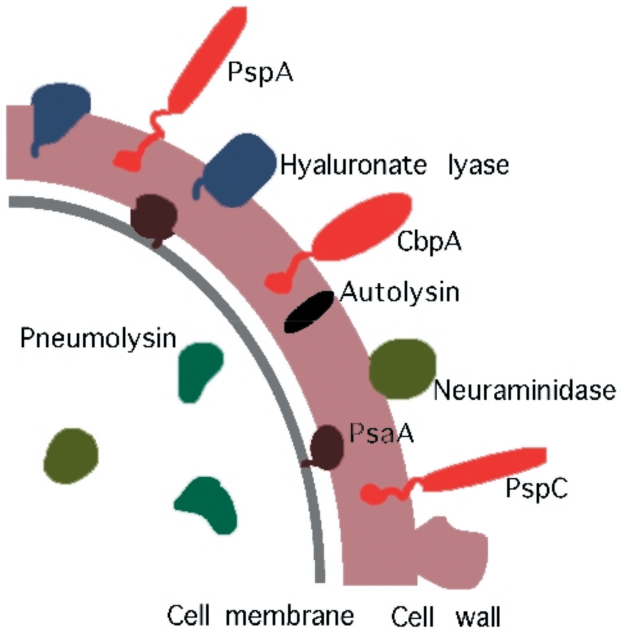

PLATE 1A.

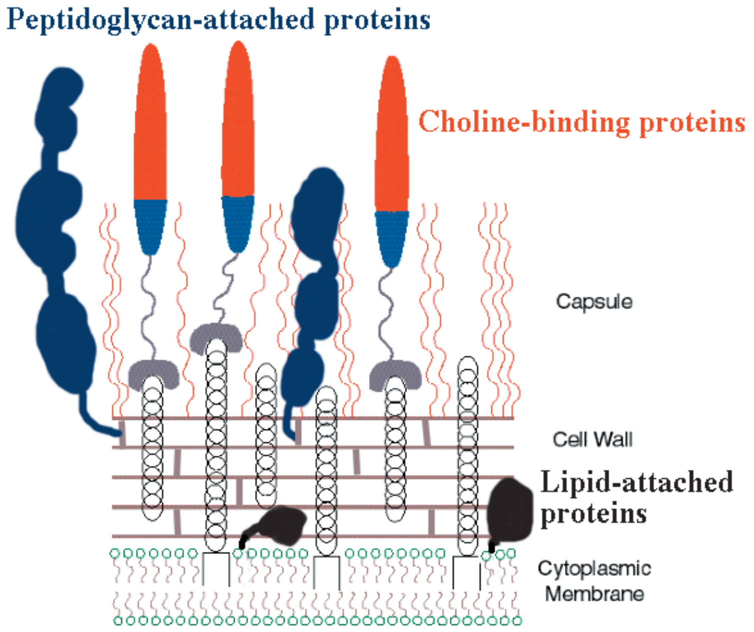

PLATE 1B.

PLATE 1. Pneumococcal surface proteins. **(A)** Schematic diagram of selected known virulence factors. As time progresses more of these pneumococcal virulence factors are discovered. Several of these proteins, such as PspA and PsaA, are used as antigens in novel pneumococcal vaccines that are currently the process of development. The antibodies to these proteins often provide cross-strain protection to humans, including young children and the elderly. The capsule, containing polysaccharide components used in the current vaccine(s) and in the conjugate vaccines under development, extends outside of the cell wall and is not depicted. **(B)** Modes of attachment of surface proteins. Proteins attached through utilization of direct peptidoglycan covalent linkage (in blue color), choline-binding domain (in red/blue color), and direct lipid linkage (in brown color) are schematically illustrated. The cell wall composed of cytoplasmic bilayer type membrane, peptidoglycan structures linked by peptide crossbridges, teichoic and lipoteichic acid structures, and the capsule are also shown.

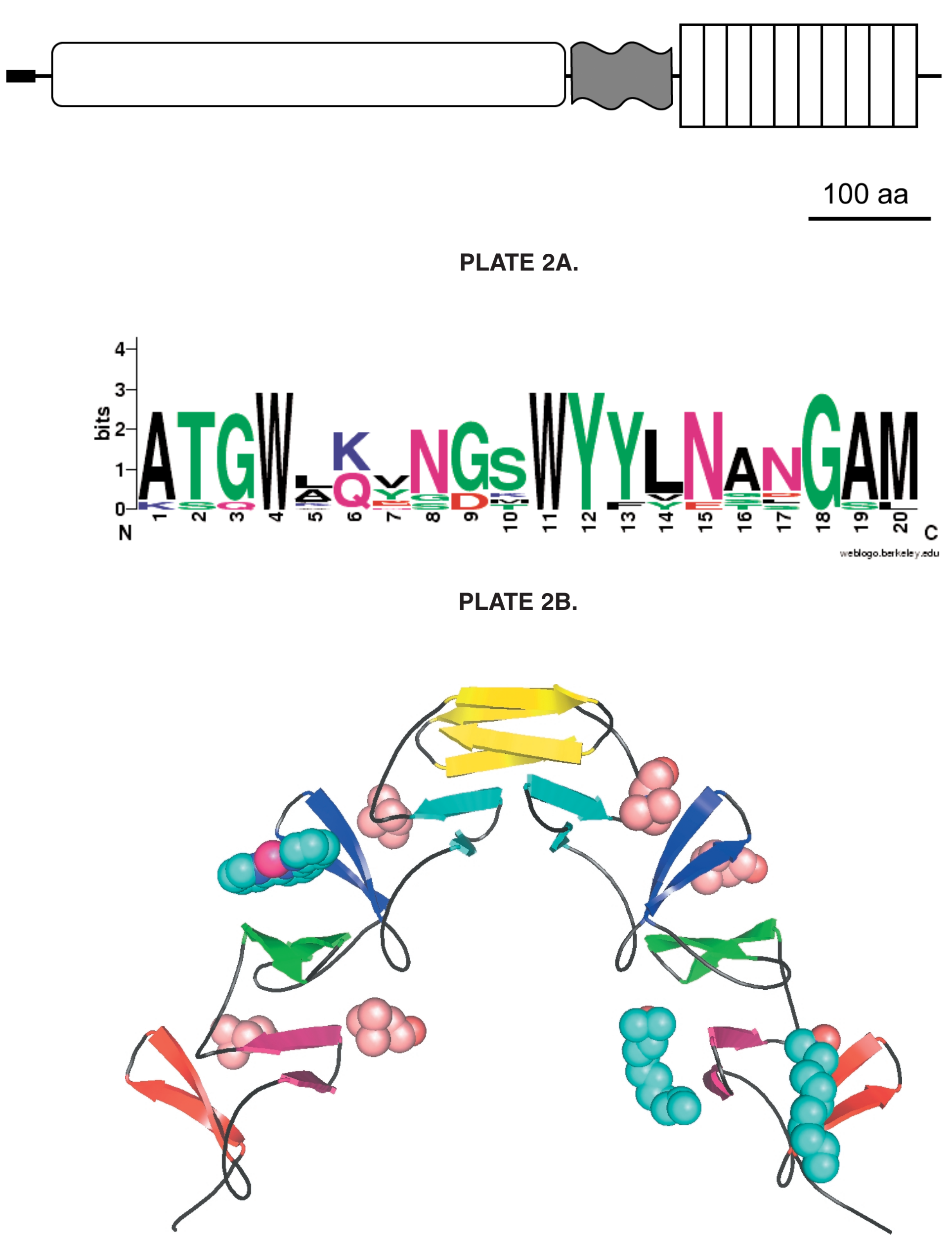

PLATE 2A.

PLATE 2B.

PLATE 2C.

FIGURE 2. Domain organization of choline-binding proteins and the molecular structure of the choline-binding domain. **(A)** A SMART (http://smart.embl-heidelberg.de) diagram of the domain organization of the PspA (SP0117) protein. The small black box on the left indicates the likely 31 aa signal peptide, predicted by SignalP program (Nielsen *et al.*, 1997). The rounded box indicates a 406 aa coiled coil region, recognized by the Coils2 program (Lupas, 1996); the wave-like shape indicates aa 80 aa Pro-rich region, identified by the SEG program (Wootton and Federhen, 1996) as a segment of low complexity. The 10 boxes on the right signify 10 20 aa choline-binding repeats. The 100 aa bar is included as an indication of scale. **(B)** Consensus sequence of the choline-binding repeat drawn in the SeqLogo format using the WebLogo tool (http://weblogo.berkeley.edu, Crooks *et al.*, 2003). The height of each letter indicates the degree of its conservation, the total height of each column represents the statistical importance of the given position. **(C)** Structure of the choline-binding domain from *S. pneumoniae* LytA (Fernandez-Tornero *et al.*, 2001). The two subunits (left and right) form a boomerang-shaped dimer. Each subunit contains six beta-hairpins, here colored red, magenta, green, blue, cyan and yellow from the *N*- to C-termini. The choline (mainly pink) and other bound ligands (mainly cyan) are shown in space-filling representation. These ligands occupy the choline-binding sites, four per monomer, which lie between the beta-hairpins and stabilize the fold.

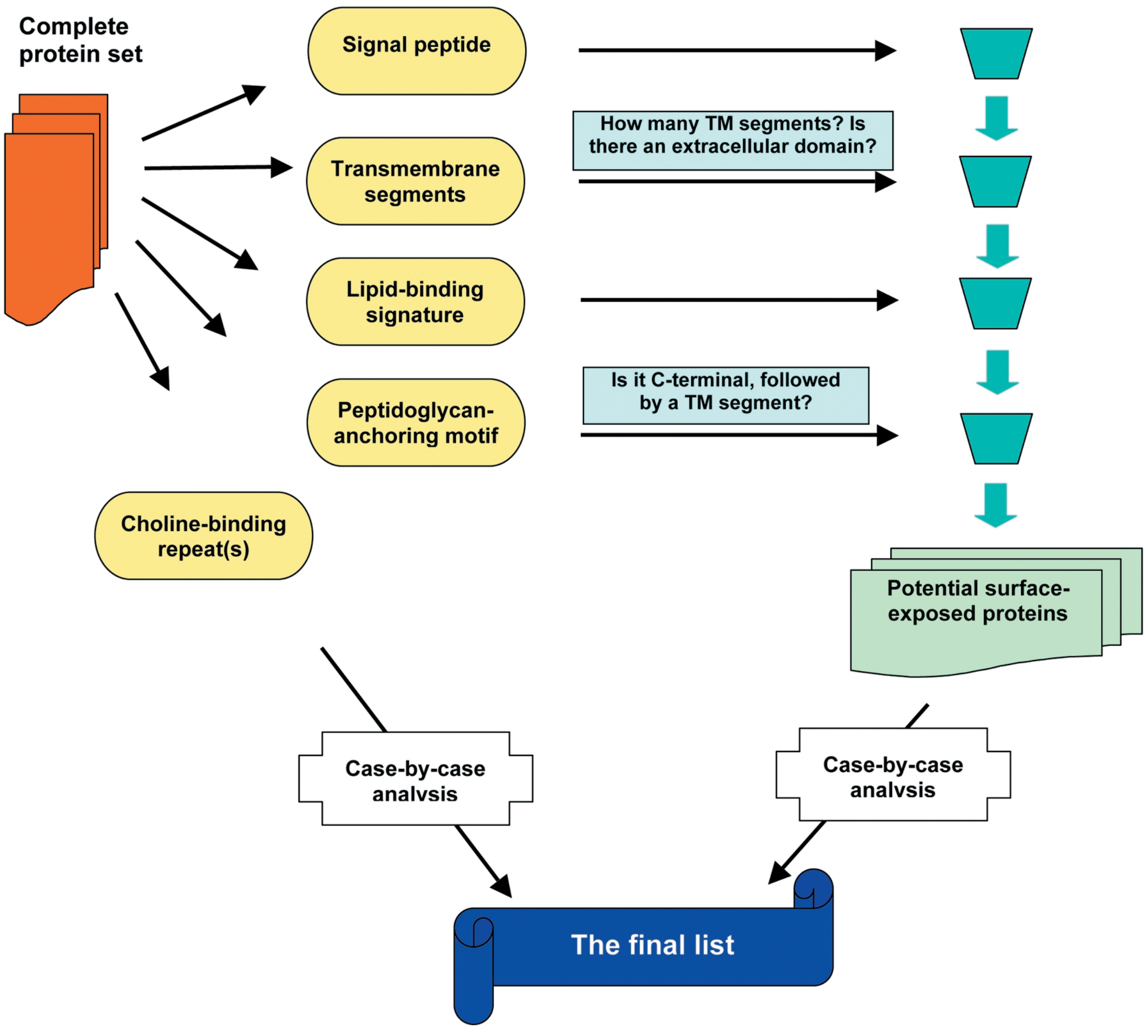
Complete protein set
Signal peptide
Transmembrane segments
Lipid-binding signature
Peptidoglycan-anchoring motif
Choline-binding repeat(s)
How many TM segments? Is there an extracellular domain?
Is it C-terminal, followed by a TM segment?
Potential surface-exposed proteins
Case-by-case analvsis
Case-by-case analvsis
The final list

PLATE 3. A schematic of the search for potential surface-exposed proteins encoded in the pneumococcal genome. Automated sequence comparisons and analysis of sequence motifs are necessarily supplemented by case-by-case analysis in order to eliminate false-positives.

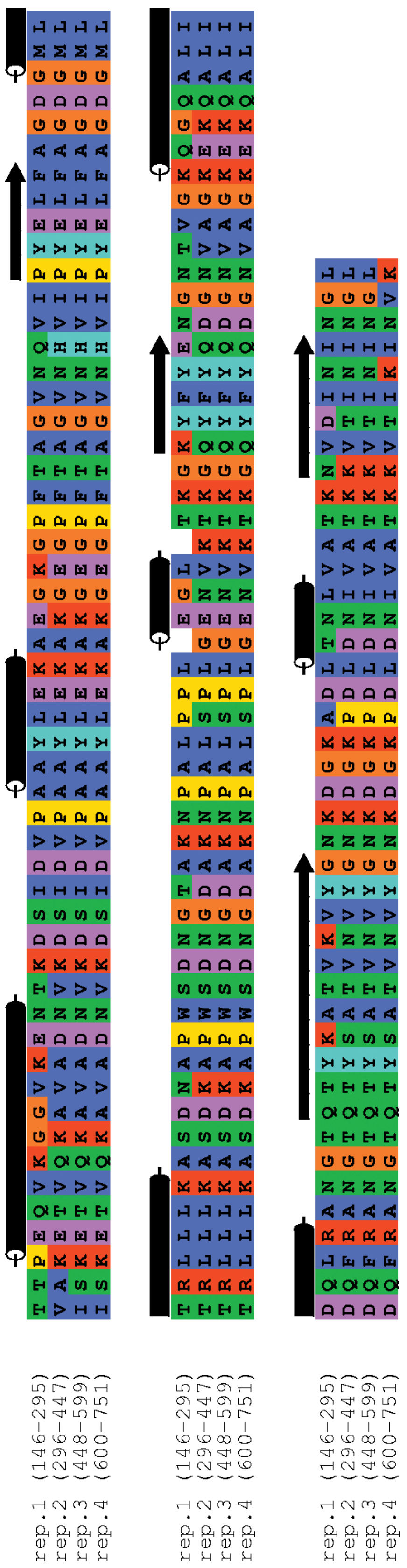

PLATE 4. Sequence alignment of the four repeats identified in SP0082, along with the PSIPRED (Jones, 1999) secondary structure prediction for the third repeat. The existence of sequence repeats allows for accurate determination of domain boundaries, while the failure of fold recognition methods to associate the sequence with known structures suggests that SP0082 may contain a novel fold.

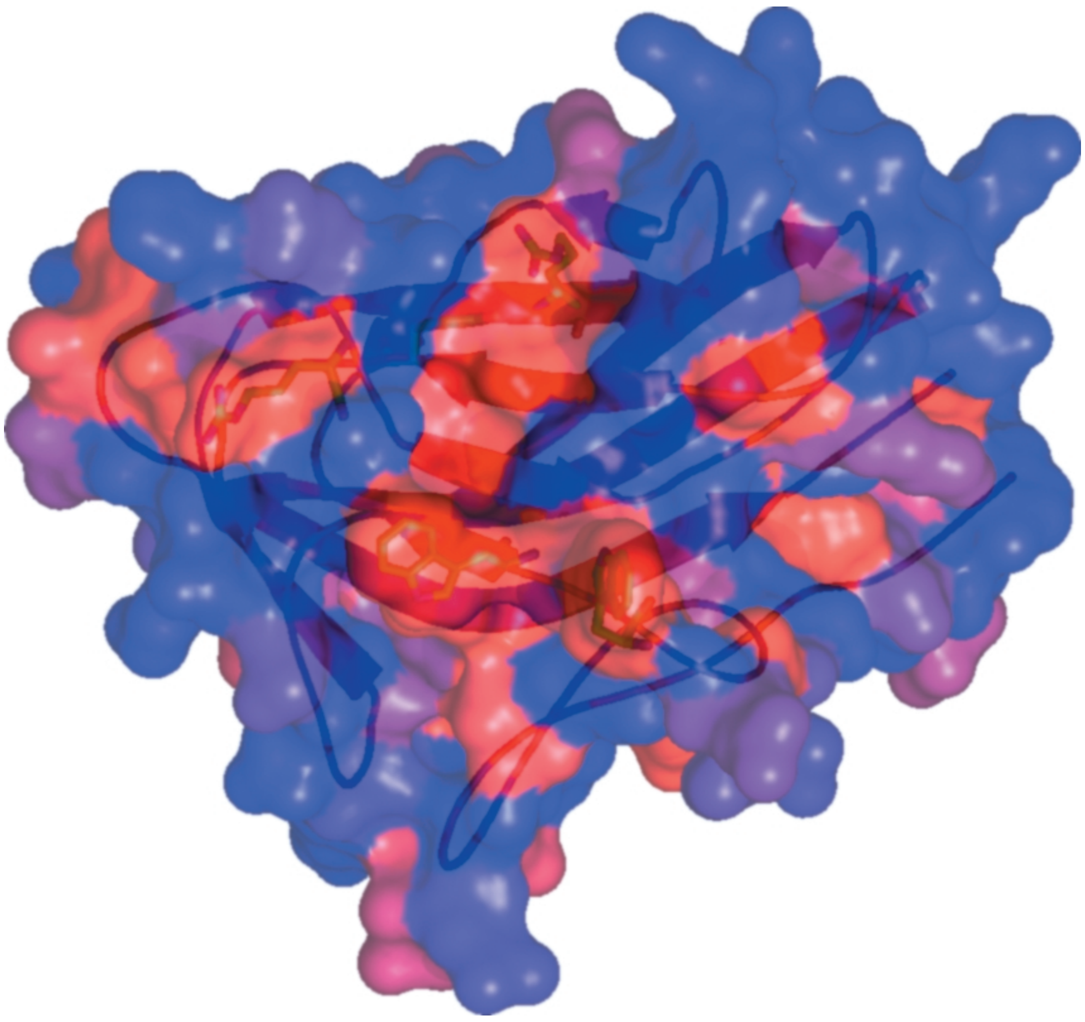

PLATE 5. Preliminary model of the N-terminal domain of hyaluronate lyase, SP0314. Mapping of sequence conservation within streptococcal hyaluronidases onto a molecular surface (red conserved, blue not conserved) reveals a potential hyaluronan-binding surface. Two completely conserved Trp residues, two completely conserved Args and a highly conserved Tyr are putative substrate contacts. Trp and Tyr commonly bind to the hydrophobic faces of carbohydrates in carbohydrate-binding proteins (Quiocho and Vyas, 1999), while the Arg residues may interact electrostatically with the negatively charged substrate.

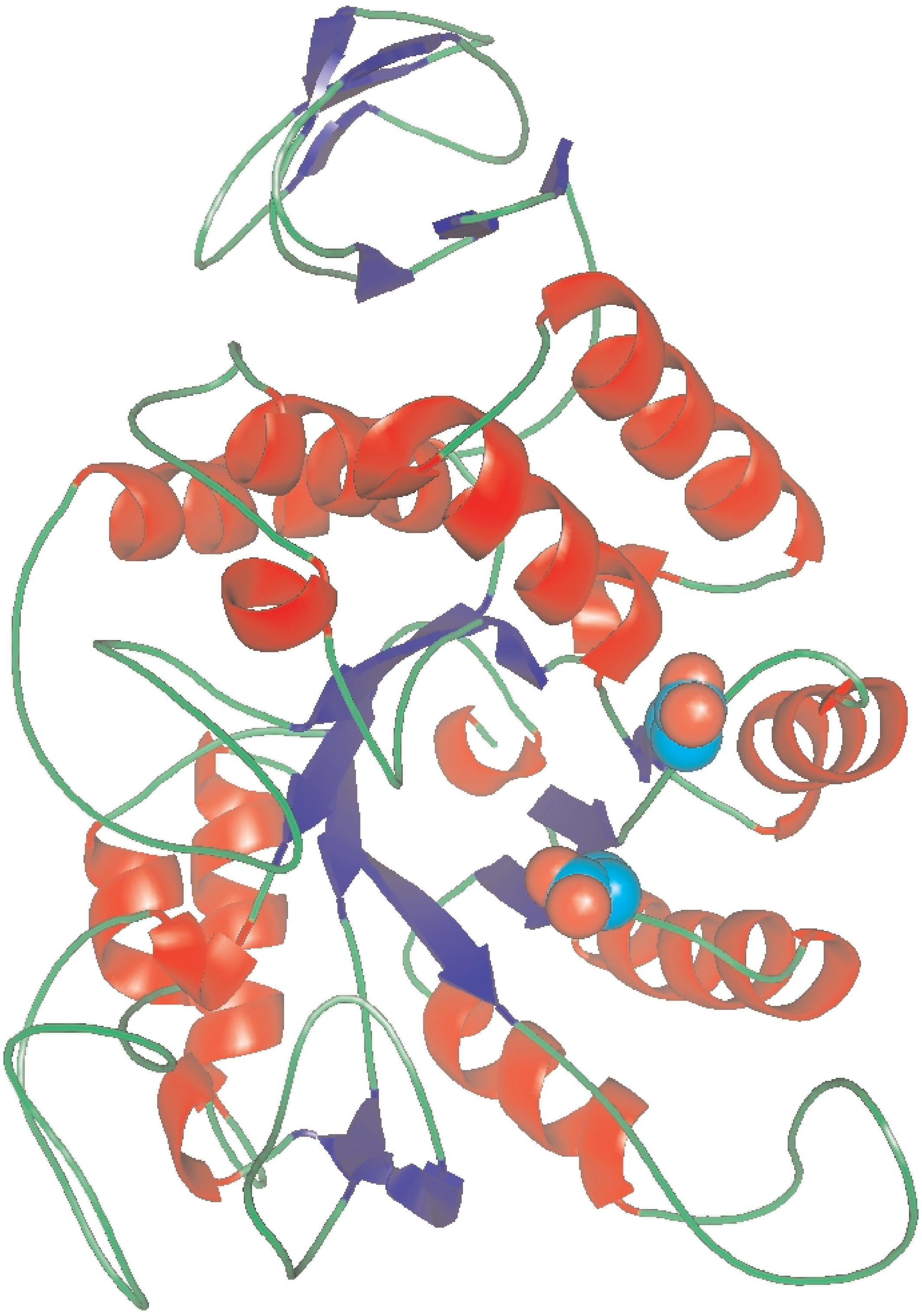

PLATE 6. Cartoon diagram of a rough model of the SP0498 molecule. The model is based on the β-amylase TIM barrel, with putative catalytic residues, Glu337 and Asp374, shown in a space-filling representation. These residues are highlighted by their conservation, location at the C-terminal end of the TIM barrel and suitable spacing to act together in glucoside hydrolysis (Nagano *et al.*, 2001).

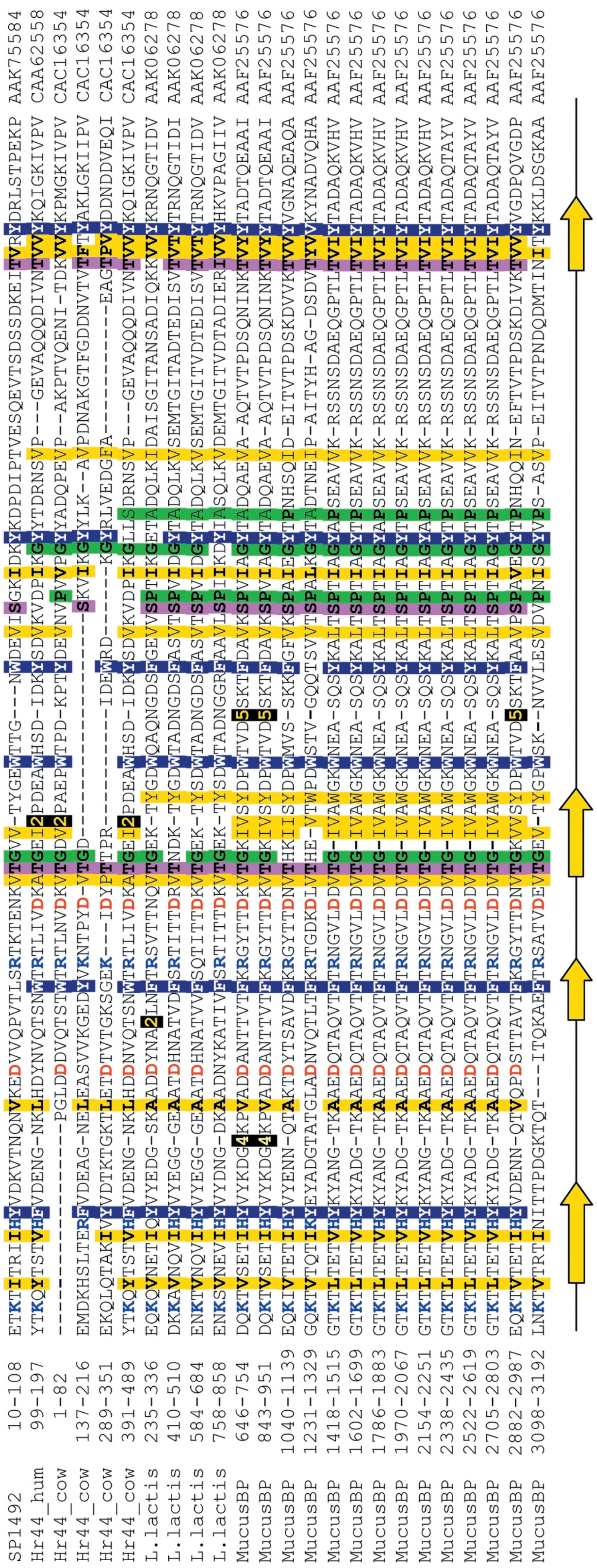

PLATE 7. Sequence alignment of SP1492 and mucin-binding domains in other organisms. Conserved residues are shown in bold and colored as follows: acidic — red; positively charged — blue; aromatic — white on blue background; aliphatic — yellow background; hydroxyl group-containing (Thr and Ser) — purple background; small (Gly, Ser, Ala) and Pro — green background. GenBank accession numbers for all the protein are shown in the right column. MucusBP indicates experimentally characterized mucin-binding protein from *Lactobacillus reuteri* (Roos and Jonsson, 2002). The PSI-PRED (Jones, 1999) secondary structure prediction is shown beneath the alignment.

TABLE 1
Surface-Exposed Proteins of *Streptococcus pneumoniae* with Known Three-Dimensional Structures

Protein	Function	Signature sequence	Type of 3D structure	Citation
Hyaluronate lyase	Degradation of hyaluronan	LPqTG	X-ray	Li *et al.*, 2000
Pneumococcal surface adhesin A	Metal transport and possible adhesin	Lipoprotein	X-ray	Lawrence *et al.*, 1998
Pneumococcal surface protein A	Anti-complement properties	Choline-binding repeats	Model	Jedrzejas *et al.*, 2000
Pneumolysin	Host cell attachment and lysis	None	X-ray (of homologous enzyme)	Rossjohn *et al.*, 1997
Neuraminidase A	Attachment and release from the host cell surface by interacting with sialic acid	LPeTG	X-ray (of homologous enzyme)	Crennell *et al.*, 1993
Autolysin LytA	Degradation of peptidoglycan	Choline-binding repeats	X-ray structure of the choline-binding domain	Fernandez-Tornero *et al.*, 2001

protease/transpeptidase that recognizes the conserved sequence motif LPxTG (the middle residue type, x, is unimportant) in an extracellular protein, cleaves the Thr-Gly bond, and covalently links the Thr residue to a penta-glycine crossbridge in the peptidoglycan (Mazmanian *et al*., 1999; Navarre and Schneewind, 1999; Ton-That *et al*., 2000). In addition to the LPxTG motif, various sortases can apparently recognize VPxTG, IPxTG, YPxTG, and some other motifs; their exact specificities have not been defined (see Navarre and Schneewind, 1999; Pallen *et al*., 2001). It appears, however, that additional sequence requirements for sortase action include a stretch of hydrophobic residues (a putative transmembrane segment) after the LPxTG motif, followed by positively charged residues. It seems very likely that these hydrophobic and positively charged residues interact, respectively, with the hydrophobic tails and the negatively charged head groups of the membrane phospholipids, helping to properly orient the substrate protein with respect to the cytoplasmic membrane and the cell wall. Such properly positioned substrate protein makes possible its covalent attachment to the peptidoglycan crossbridges. In accordance with the "positive inside" rule (von Heijne, 1989), the positively charged residues are more likely to be located on the cytoplasmic side of the membrane.

In most cases, the LPxTG-like motif is located near the C-terminus of the protein and sortase cleaves only ~30 C-terminal residues, leaving the rest of the protein exposed to the extracellular space. An example of such protein is hyaluronate lyase, which degrades hyaluronan the main component of extracellular matrix of the host tissues hyaluronan (Jedrzejas, 2000, 2001, and 2002). In some cases, however, the LPxTG-like motif is at the N-terminus, indicating a potentially different mechanism of peptidoglycan anchoring (see below). Given the relatively well-defined recognition motif, a search for potential peptidoglycan-anchored proteins is relatively straightforward and can be done either by simple text matching using Perl scripts such as GREF (Walker and Koonin, 1997), or by more sophisticated motif-searching programs such as hidden Markov model search (HMMmer, Durbin *et al*., 1998), Position-Specific Iterative BLAST (PSI-BLAST), and Pattern-Hit Initiated BLAST (PHI-BLAST, Altschul *et al*., 1997).

D. Proteins Directly Attached to the Lipid of the Cytoplasmic Membrane

These proteins have yet another signature sequence at their N-terminus immediately past the signal sequence (Plates 1a,b). This motif consists of a stretch of amino acids Lx_1x_2C (where x_1 is usually A, S, V, Q, or T and x_2 is G or A) or similar sequences (Tam and Saier, 1993). In this attachment mode the protein forms a covalent linkage between the Cys residue and the diacylglyceryl of the lipid bilayer. This attachment process is followed by cleavage of the protein's signal peptide. After this process the modified Cys residue is present at the N-terminus of the surface attached protein. The high variability of sequences in the very N-terminal region of such proteins suggests somewhat flexible structural properties of this peptide block with a range of variable three-dimensional structures that most likely are disordered. The flexible nature of such linker peptide likely facilitates the binding of the major part of protein to the cytoplasmic membrane of *S. pneumoniae* as well as attachment to bacterial surface (Berry and Paton, 1996). An example of such protein is pneumococcal surface adhesin A, PsaA, that, according to some reports, in addition to its apparent function as a component of an ABC-type metal transporter, also takes part in the pneumococcal adhesion to the host cells (Lawrence *et al*., 1998; Dintilhac *et al*., 1997; Sampson *et al*., 1994). The structure of PsaA is known (Lawrence *et al*., 1998) and confirms the disordered, highly flexible nature of the membrane linking peptide at its N-terminus. This part of PsaA structure was not visualized in the crystal structure but may be inferred to protrude from the remainder of the protein with which it likely has few interactions (Jedrzejas, 2001).

E. Other Surface Proteins, Including the Histidine Triad Family

The three signature sequences described above are certainly not the only motifs to recognize surface proteins on *S. pneumoniae* or other Gram-positive bacteria (Munoa *et al*., 1991; Inouye *et al*., 1982). For example, recently four pneumococcal surface-exposed proteins containing putative hydrophobic leader sequences (Pht family) were identified in the genome, described, and characterized immunologically in mice

(Adamou *et al*., 2001). Each of these Pht family members, termed PhtA, PhtB, PhtD, and PhtE, had four or five copies of a novel signature motif, HxxHxH, called "the histidine triad". Such hydrophobic leader sequences are characteristic of proteins exported across the cytoplasmic membrane (Munoa *et al*., 1991; Inouye *et al*., 1982). Although the biological functions of these (apparently metal-binding) proteins are still unknown, they proved to be immunogenic and elicited protection against pneumococcal infection in mice, specifically protective against sepsis and death (Adamou *et al*., 2001). In addition, the identification of proteins with sequence motifs associated with transport such as signal peptidase I or II, or type IV prepilin signal sequences is needed. Furthermore, other pneumococcal surface proteins could be identified by homology-based sequence or structure searches to known virulence factors and surface proteins in other bacteria. As time progresses, additional proteins perhaps with new signature sequences will likely be identified and described. However, for the purpose of this review we focus on the first three groups of proteins that constitute the majority of pneumococcal surface proteins: choline-binding proteins, proteins covalently attached to peptidoglycan, and lipid-anchored proteins as described above.

III. RELEVANT BIOINFORMATICS TOOLS

The bioinformatics analysis of the putative surface-exposed proteins includes their identification in pneumococcal genome, sequence-motif based analysis of their mode of attachment, followed by the elucidation of their structural properties, prediction of their functions, and finally obtaining their three-dimensional model structures. The details of such bioinformatics analyses are described in more detail below.

A. Identification of Putative Surface-Exposed Proteins

Although the above-listed signature sequences for choline binding and peptidoglycan attachment are fairly specific, they could potentially be found also in cytoplasmic proteins. For example, *S. pneumoniae* protein SP0439 is peptide chain release factor 3, a GTPase that participates in translation, but contains a ^{322}LPRTG^{326} sequence that is followed by some hydrophobic and positively charged residues. Nevertheless, it is a typical cytoplasmic protein that is not exported to the cell surface. The lipid-binding signature Lx[GA]C is even more common, being present in 85 pneumococcal proteins, many of which are cytoplasmic. For example, SP0280, a 16S rRNA pseudouridylate synthase (RsuA), contains the pattern ^{7}LVAC^{10} on its N-terminus, whereas SP0933, a pyrroline-5-carboxylate reductase (ProC), contains the pattern ^{168}LAGC^{171} in the middle of its sequence. Both these proteins are close homologs of well-characterized cytoplasmic enzymes and show no indication of being surface exposed. To avoid such false-positive hits, one has to combine search for signature sequences with searches for potential signal peptides and transmembrane segments and predictions of globular and nonglobular regions (Plate 3).

This process can be visualized as a series of filtration steps. Initially, possible signal peptides in pneumococcal proteins are sought using the SignalP program (Nielsen *et al*., 1997), preferably trained on sequences from Gram-positive bacteria. With signal peptides identified and duly masked, one can search for the potential transmembrane segments using, for example, the PhDhtm program (Rost *et al*., 1995). It is also useful to check for the presence of specialized signal sequences, such as those recognized by the twin-arginine (Tat) and prepilin secretion systems (Berks *et al*., 2000; Tjalsma *et al*., 2000; van Dijl *et al*., 2002). Those proteins that lack any type of the signal peptide and have no predicted transmembrane segments are most likely to be cytoplasmic and can be safely removed from further consideration.

Another important filter is the search for the presence of low-complexity regions using the SEG program (Wootton, 1994; Wootton and Federhen, 1996). Low-complexity regions, found in many surface-exposed proteins, may adopt nonglobular structures, but usually form no stable structure at all. On the sequence level, such regions are identified based on statistically nonrandom (i.e., biased) distribution of certain amino acid residues. Using the SEG program with various sets of parameters allows one to delineate various low-complexity segments, from those almost certain to form nonglobular fragments to those that display just moderately biased composition. However, one should be aware that due to the biased sequences of most all-β domains (the prevalence of hydrophobic residues in the β-strands and of Gly, Ser, Ala, and Pro residues in the turn regions), such domains are often recognized by SEG as low-complexity segments; this does not mean they have disordered structures (Koonin and Galperin, 2002). One common type of low-complexity fragments is the coiled-coil that can be identified using COILS2 program (Lupas, 1996).

Although the first approach to sequence analysis of the putative surface-exposed proteins includes running the above-mentioned programs in batch with minimal human intervention, the results of these searches have to be carefully manually inspected on a case-by-case basis. We have already mentioned false-positive hits that can arise in the search for signature sequences. On the opposite side, the failure of SignalP to recognize a signal peptide would cause false-negative hits, leading one to overlook perfectly legitimate surface-exposed proteins. The only way to ensure low levels of both false-positives and false-negatives is manual examination of all the proteins that have been identified in either type of screens shown in Plate 3. A useful tool for such manual examination of extracellular localization is the PSORT set of programs (Nakai and Horton, 1999; Nakai, 2001).

B. Sequence Analysis and Structural Predictions

Clearly, the deeper the understanding of structure and function of a protein that can be obtained, the more precisely future experiments can be guided. For example, structural genomics has as key criteria for ranking of targets novel structure and novel function, which run alongside considerations of pathogenic importance.

The most direct route to understanding structure and function of a particular protein is through demonstration of its homology — its sharing of a common evolutionary ancestor — with another protein already characterized. This assures that the two proteins have a similar overall fold, although differences proportional to the time elapsed since their evolutionary divergence are to be expected (Chung and Subbiah, 1996). The inference of functional similarity is more complicated (Wilson *et al*., 2000; Devos and Valencia, 2000; Todd *et al*., 2000), with serious consequences for genome annotation (Devos and Valencia, 2001). At high levels of sequence identity, two proteins can be assumed to share the same function. However, in the twilight zone of 25% sequence identity and below, the proteins' functions frequently turn out to be substantially different than those of the known homologs. The consideration of protein structure, including comparative molecular modeling, provides a route to function verification and prediction in these more difficult cases (Aloy *et al*., 2001).

For inference of homology, a range of sequence comparison methods are available, from the rapid but less sensitive BLAST (Altschul *et al*., 1997) to the slower but more sensitive Hidden Markov Models (Karplus *et al*., 1998) and PSI-BLAST (Altschul *et al*., 1997). An alternative approach to demonstrating directly an evolutionary relationship between proteins A and B is to demonstrate independently relationships between proteins A and C and between proteins B and C. In principle, any number of intermediate proteins can be used to form a chain of inferences. This approach is the basis of the Intermediate Sequence Search (Li *et al*., 2000) and the FASTA walk (Holm and Sander, 1997) methods. In the most difficult cases, an analysis of inferred characteristics, such as predicted secondary structure and solvent exposure, can help match a sequence to an existing fold. This finding is the basis of fold recognition or threading techniques that can match sequences to folds even in the absence of significant sequence similarity between the protein analyzed and the known structure. A comparison of these methods shows there to be no individual method that performs best in all cases (Fischer *et al*., 2001). This prompted the development of consensus fold recognition methods that reanalyze the results of several individual methods, giving high scores to folds that appear independently in the results of several different methods (Lundstrom *et al*., 2001). An ongoing web comparison of fold recognition methods confirms that these consensus methods outperform the individual algorithms (Bujnicki *et al*., 2001a). For fold recognition, the MetaServer (Bujnicki *et al*., 2001b) is a particularly powerful resource, conveniently providing access to many of the best performing fold recognition methods (Fischer *et al*., 2001). In most of these cases, the shared fold indicates an evolutionary relationship, but the possibility of convergent global evolution (structural analogy) also exists.

Fold recognition methods work best on single-domain sequences that have neither additional nor missing amino acids (e.g., Rigden, 2002). Probably the most generally applicable domain identification tool is PASS (Prediction of Autonomously folding units base on Sequence Similarities; Kuroda *et al*., 2000), which post-processes BLAST results. Interest in the area has resulted in the appearance of several other methods, but each suffers drawbacks; current applicability to two-domain proteins alone (Rigden, 2002), large computational demands (George and Herenga, 2002), or an inability to favor one domain definition above several others of similar ranking (Wheelan *et al*., 2000).

For the prediction of nonglobular, low-complexity, and transmembrane regions, the PEDANT server (Protein Extraction, Description and Analysis Tool; Frishman *et al*., 2001) is the most convenient resource, providing access to precomputed predictions for all *S. pneumoniae* open reading frames. RADAR (Rapid Automatic Detection and Alignment of Repeats; Heger and Holm, 2000) can be utilized for the analysis of protein repeats.

The most commonly used program for model construction is MODELLER (Sali and Blundell, 1993). In cases of low sequence identity between target (the protein to be modeled) and the template(s) (the existing protein structure(s)), it is important that a rigorous methodology be adopted involving the construction of different sets of models based on variant alignments (e.g., Rigden and Carneiro, 1999; Rigden *et al*., 2000, 2001a, 2001b, 2002). Statistics-based protein structure quality measurements (Sippl, 1993) can then be used to determine the most probable of the alignments. An iterative cycle of alignment improvements then follows until the final, best available alignment is obtained. When the optimal alignment has been reached, analysis of stereochemical quality with PROCHECK (Laskowski *et al*., 1993) and, in particular, the quality of the Ramachandran plot help to pinpoint probable local errors.

C. Inference of Function from Structure

A developing set of methodologies for the inference of function from structure is available (Norin and Sundstrom, 2002). This set is applicable to crystal structures of proteins of unknown function, as well as to proteins that could be modeled on the basis of existing crystal structures, but whose function was not clear.

First, structures may be scanned against the PROCAT database of potential catalytic sites (Wallace *et al*., 1997). These are three-dimensional arrangements of particular sets of residues conferring a certain catalytic activity that may evolve independently several times, for example, the hydrolase (protease/lipase) catalytic triad. Second, the surface of the structure may be scanned for the largest pockets with, for example, PASS (Putative Active Sites with Spheres; Brady and Stouten, 2000). These correspond to binding sites with surprising frequency (Laskowski *et al*., 1996; Brady and Stouten, 2000). Third, the electrostatic characteristics of the model can be analyzed and visualized with GRASP (Graphical Representation and Analysis of Structural Properties; Nicholls *et al*., 1991). Strongly positive electrostatic potential associated with helical motifs, for example, is suggestive of DNA binding ability (e.g., Rigden and Carneiro, 1999; Rigden *et al*., 2002). Significant polarization near predicted binding sites also gives clues as to the nature of possible substrates (e.g., Rigden *et al*., 1998). Fourth, when the structure has a significant number of sequence homologues, spatial analysis of sequence conservation is a powerful tool. Thus the sequences may be aligned T-COFFEE (Notredame *et al*., 2000), for example, and ESPRIPT (Easy Sequencing in Postscript; Gouet *et al*., 1999) used for mapping sequence conservation onto structures. In the ESPRIPT result, the degree of conservation is output in the B-factor column of the PDB file, enabling its ready color-coded visualization in molecular modeling programs such as PYMOL (Delano, 2002). Additional case-specific information may be available. For example, tryptophan residues are very commonly found at carbohydrate-binding sites (Quiocho and Vyas, 1999), an observation with useful predictive value (Rigden and Jedrzejas, 2003). When catalytic activity is suspected, consideration of the residues most commonly found in catalytic sites will often be useful (Bartlett *et al*., 2002). Frequently, a particular site will be highlighted by several different indicators, for example, high sequence conservation in combination with significant electrostatic potential. The greater the number of independent analyses pointing to a given site, the greater the confidence in its annotation as functionally significant. Analyses have also shown statistically different residue compositions for various types of protein-protein interface (permanent vs. transitory, heterooligomer vs. homooligomer, etc.) (Ofran and Rost, 2003). Such information can be used for predictions of surfaces involved in protein-protein binding (Zhou and Shan, 2001).

Further information can be extracted when the new structure is similar to another already determined. This will be most often the case for modeled structures, which will be compared to their corresponding templates. The models can be analyzed in order to determine if active sites, localized by previous work for example, remain conserved and that access for substrate or other interacting molecule, remains unimpeded by changes elsewhere in the protein. A survey of the structural differences responsible for absence of catalytic activity in families largely composed of enzymes has been published recently (Todd *et al*., 2002). In some cases, the catalytic machinery will be conserved, but other changes in binding sites may lead to significantly different catalyzed reactions (e.g., Rigden *et al*., 2001b). An automated method for this purpose, analyzing the spatial positioning of conserved hydrophilic residues, is available (Aloy *et al*., 2001). Model analysis, even in the absence of closely homologous sequences, can also reveal key conserved positions and interactions that suggest evolutionary relationships (e.g., Rigden *et al*., 2001a).

IV. SURFACE-EXPOSED PROTEINS IDENTIFIED FROM THE *STREPTOCOCCUS PNEUMONIAE* GENOME SEQUENCE

The initial analysis of the TIGR4 genomic sequence of *S. pneumoniae*, reported by Tettelin *et al.* (2001), identified 69 genes coding for proteins that were likely to be exposed on the surface of the pneumococcal cell. This protein set included 19 predicted proteins with the cell wall surface anchor family sequence motif LPxTG (or similar; likely substrates of sortases), 15 predicted proteins with putative choline-binding motifs, 36 proteins with putative lipid attachment motif (predicted lipoproteins), and in addition 60 proteins with predicted N-terminal signal peptides (Tettelin *et al*., 2001). In 62 cases out of the total of 69, there were two or more independent indications of the surface localization of the predicted protein. This set of 62 proteins comprises the most obvious surface-exposed proteins. Among other putative surface proteins, Tettelin *et al.* identified a putative sortase (SP0468, GenBank accession no. AAK74628), the enzyme that covalently attaches extracellular proteins to the peptidoglycan (Mazmanian *et al*., 1999; Ton-That *et al*., 2000). The nuclear magnetic resonance (NMR) structure of sortase from *Staphylococcus aureus* has been solved recently, revealing an unusual β-barrel structure (Ilangovan *et al*., 2001). Remarkably, *S. pneumoniae* strain TIGR4 encodes four sortase-like proteins (Pallen *et al*., 2001, and our own observations), the closest of which shares only 32% identity with the *S. aureus* sortase. Other attractive surface protein targets from this list include choline-binding proteins, including choline-binding protein A (CbpA), C (CbpC), D (CbpD), E (CbpE), F (CbpF), G (CbpG), I (CbpI), and J (CbpJ), which belong to the vast family of PspA-like pneumococcal surface proteins with choline attachment motif; predicted proteases SP0071, SP0641, SP0664, and SP1154; autolysis-related proteins SP0965, SP1573, and SP1937; and number of uncharacterized (putative) lipoproteins.

In a recent pneumococcal genome-based study, out of 130 identified ORFs with secretion motifs or homology to other predicted virulence factors, 108 were tested as vaccine candidates in mice (Wizemann *et al*., 2001). It was shown that choline-binding proteins LytB (SP0965) and LytC, histidine triad family protein SP1175, and peptidoglycan-anchored serine proteinase PrtA (SP0641) all afforded reasonable protection in mice in the lethal sepsis model, providing convincing proof of the concept that computational predictions are a valuable resource for the development of new vaccines and for vaccine research in general. Remarkably, the number of predicted cell wall-anchored proteins in that study, 34 proteins, was substantially higher than the figure reported by Tettelin *et al.* (2001), which identified only 19 such macromolecules. There were additional discrepancies, which prompted us to take a careful look at these numbers.

Our own analysis of the putative surface proteins of *S. pneumoniae*, using the SignalP program trained on proteins from Gram-positive bacteria, identified 124 candidate secreted and membrane proteins, almost twice the number reported by Tettelin *et al.* (2001). Similarly, we identified several additional proteins with peptidoglycan-anchoring motifs, including the relatively well-characterized sialidase–neuraminidase A (NanA) (see Camara *et al*., 1994), which contains the LPeTG motif. In addition, we searched for proteins with transmembrane segments as some of them are likely to have substantial extracytoplasmic domains. In this group, 567 predicted proteins were found to contain from 3 to 12 transmembrane segments. Twenty more proteins with two predicted transmembrane segments were also likely to be membrane anchored. A case-by-case analysis of these proteins revealed the extent of their surface exposure.

The following section illustrates selected results of the genome analysis of the pneumococcal surface-exposed proteins that has been performed using the methodology and the programs listed above. This analysis was followed by manual inspection and analysis of each of the identified proteins. For several selected cases of higher interest, we offer a more detailed description and, where possible, three-dimensional molecular models. Such analysis lead to often striking and important conclusions about the functional properties of analyzed proteins and about the extent of their likely involvement in processes directly related to pneumococcal pathogenesis.

V. Bioinformatics Analysis of Peptidoglycan-Attached Surface Proteins

A. General Overview of Properties of Identified Surface Proteins

Summary information regarding the three sets of putative surface proteins is presented in Table 2. After due consideration of motifs specifying surface location and the presence or absence of the various kinds of signal peptide, we place 124 proteins on the surface of *S. pneumoniae*. This corresponds to ~6% of the total number of genome predicted ORFs in *S. pneumoniae*.

Two striking conclusions are immediately evident from the analysis of data in Table 2. First, even among the proteins whose likely significant medical importance has stimulated intensive study, many proteins have no useful functional and structural annotation. They are either annotated only as hypothetical proteins, or their annotation simply reflects their predicted cellular location, for example, cell wall surface anchor family protein. Secondly, all surface protein classes contain many large proteins with mean sizes in each class ranging from 345 to 1495 amino acid residues. An analysis has shown that protein domains, the building blocks of protein structure, are most commonly around 100 to 120 residues long, with the largest single domain yet observed having around 550 residues (Wheelan *et al*., 2000). Thus, most of the proteins analyzed here will likely consist of more than one domain. This simple conclusion has significant consequences for their analysis, both by bioinformatics methods, and experimentally by X-ray crystallography or NMR, especially due to the limited performance of current domain identification methods (see above). The sensitive fold recognition methods that are required for reliable identification of cases of distant homology are well known for working most effectively on individual domain sequences — extraneous or missing sequence will readily hamper the identification of homologous structures (e.g., Rigden, 2002). Similar considerations apply to the experimental determination of their three-dimensional structures. The size limitations of the NMR technique require that large proteins be addressed through cleavage into domains. X-ray crystallography is capable of dealing with large proteins, but multidomain proteins remain problematical since the frequent flexibility of inter-domain linkers leads to conformational heterogeneity, thereby complicating crystallization.

As the first step toward the systematic modeling and structure determination of *S. pneumoniae* proteins, detailed analysis has been carried out on the peptidoglycan-attached protein set, as summarized in Table 3. An intriguing mixture of cases is apparent, ranging from the single protein, the majority of whose structure has been determined experimentally (Li *et al*., 2000), to 7 ORFs annotated simply as “cell wall surface anchor family protein”. In two other cases experimental validation of annotated activity is available — SP0057 is a β-*N*-acetylhexosaminidase (Clarke *et al*., 1995) and SP0648 is a β-galactosidase (Zahner and Hakenbeck, 2000). However, it is notable that not a single protein could be said to be completely understood — the hyaluronate lyase was crystallized, for practical reasons, without *N*- and C-terminal domains, while the identity of the catalytic TIM barrel(s) in the β-*N*-acetylhexosaminidase that, unusually, contains two of these domains, is not known. In the case of the β-galactosidase, mysteries remain to be solved regarding the function(s) of the 1600 residues following the clear catalytic TIM barrel.

Just as these cases represent a spectrum from well- to poorly understood proteins, the application of bioinformatics yields insights of varying degrees. In some cases, clear structural assignments predicting activity can be made. In others, extensive analysis reveals nothing more than a secondary structure prediction and allocation of protein fold class, not even

TABLE 2
General Characteristics of the Initial Predicted Sets of Surface Proteins

Protein class	Motif of attachment	Number of representatives	Number with annotated function (%)	Largest (residues)	Smallest (residues)	Mean size (residues)
Choline binding	Characteristic repeat (PspA-like)	13	3 (23%)	744	211	458
Peptidoglycan-attached	LPxTG-like	20	11 (55%)	4776	202	1350
Lipid-attached	Lx[GA]C-like	33	23 (70%)	661	236	370

providing reliable domain boundaries to facilitate future structural determination. The examples given below illustrate some of the useful conclusions that can be drawn and the ways in which they guide further experiments.

B. Structural and Functional Description of Selected Peptidoglycan-Anchored Proteins

1. SP0082 — A Fourfold Domain repeat with Probable Novel Fold

One of the simplest ways in which a domain boundary can be reliably determined is through the presence of repeats. A good example was found in SP0082, which contains four near perfect repeats of around 150 residues (Plate 4). With the structural domain well defined, the lack of any strong fold recognition results is a good indicator of a substantially novel fold, given the strong performance of modern fold recognition methods (Sippl *et al*., 2001). In this case the structural domain appears to be mixed α/β-type. Despite a lack of strong fold recognition results, the presence of EtxxK motifs toward the start of three of the four repeats is suggestive of a possible role in binding to an extracellular matrix protein. The motif has been implicated in integrin binding in several other proteins (Deivanayagam *et al*., 2000).

2. SP0314 — Completing the Fold Description for Hyaluronidase

S. pneumoniae hyaluronidase (hyaluronate lyase) has been studied extensively crystallographically, yielding a detailed picture of the catalytic process. However, these studies were carried out, for reasons of instability of the complete protein, with a truncated form of the enzyme, lacking segments of around 285 and 70 amino acid residues at the N- and C-termini, respectively. This raised the possibility of further, uncharacterized domains being present in the complete native form of the enzyme, particularly in the region N-terminal to that crystallized. The structure of a homologous enzyme from another member of *Streptococcus* species, *Streptococcus agalactiae,* hyaluronidase revealed the presence of an additional small, 74 residue domain to the N-terminus of the *S. pneumoniae* hyaluronidase crystal structure, but still around 200 residues past the identified signal peptide remained to be structurally accounted for. BLAST and PSI-BLAST searches did not reveal homologous sequences outside streptococcal hyaluronidases, but fold recognition methods were remarkably unanimous in suggesting a structural correspondence with cellulose binding domains of known NMR structure previously observed in the β-1,4-glucanase from *Cellumonas fimi* (PDB codes 1cx1, 1ulo and 1ulp; Johnson *et al*., 1996; Brun *et al*., 2000). These structures were top scoring by all the specialized fold recognition methods implemented at the MetaServer (Bujnicki *et al*., 2001a), leading to consensus scores (Lundstrom *et al*., 2001) well in excess of the best scoring false-positives (Bujnicki *et al*., 2001b). This structural correspondence was further strongly supported by the sharing of a carbohydrate-modifying activity between hyaluronate lyases, the modular structure of carbohydrate active enzymes (Coutinho and Henrissat, 1999a, b), and the various specificities exhibited by single families of carbohydrate binding domain (Coutinho and Henrissat, 1999c). Even the simple coloring of a surface of the preliminary model of this domain by sequence conservation reveals a potential carbohydrate-binding surface with completely conserved aromatic and positively charged residues suitable for interacting with the hydrophobic sugar faces and negative side chains, respectively, of the substrate hyaluronan (Plate 5) (Rigden and Jedrzejas, 2003).

3. SP0498 — Location of a Catalytic TIM Barrel and Probable Catalytic Residues

S. pneumoniae appears to have two peptidoglycan-attached β-*N*-acetylhexosaminidase enzymes (Table 3). These are of potential pathogenic importance since several host cell-surface molecules contain GlcNAcβ1-linked residues. One of these, SP0057, has been studied experimentally (Clarke *et al*., 1995) and exhibits two large repeated domains, each with clear homology to *Streptomyces plicatus* β-*N*-acetylhexosaminidase whose X-ray structure is known (Mark *et al*., 2001). In contrast, SP0498, grouped in the CAZY database (Coutinho and Henrissat, 1999c) into glycosyl hydrolase family 85, is relatively ill-understood. The application of the best-performing fold recognition methods (Fischer *et al*., 2001), accessed via the MetaServer (Bujnicki *et al*., 2001a), to four diverse members of the family enabled the clear identification of a TIM barrel domain in the N-terminal portion of SP0498. The predicted location of the TIM barrel coincided perfectly with the size of the

TABLE 3
Results of Structural Bioinformatics Analysis of Putative Peptidoglycan-Attached Proteins

Genome ID	Genbank ID	Functional annotation	Protein length, aa[b]			Structural description, comments	Reference
			Total	Exposed	Globular		
SP0057	14971522	Beta-N-acetyl-hexosaminidase	1312	1251	766	Internally duplicated catalytic TIM barrel followed by 3-fold, mixed α/β repeat of unknown structure.	Clarke *et al.*, 1995
SP0071	14971537	Immunoglobulin A1 protease	1856	1713	990	Mixed α/β structure of unclear domain boundaries	
SP0082	14971548	Cell wall surface anchor family protein	857	783	375	Contains 4 nearly identical repeated domains of 148 residues with mixed α/β structure, see Fig. 4	
SP0110[b,c]	14971577	Uncharacterized membrane protein	694	337	299	A 7TM protein with an N-terminal 150-aa extracellular domain and a 270-aa surface-exposed domain located between 4th and 5th TM segments; the LPxTG motif is in the middle of this second domain	
SP0268	14971739	Alkaline amylopullulanase	1280	1207	981	Catalytic amylase-like TIM barrel preceded by an Ig-like domain. Contains two N-terminal PUD domains and a mixed α/β domain at the C-terminus.	
SP0314	14971788	Hyaluronidase	1066	1011	954	Catalytic α/β toroid and all-β domain, preceded by all-β carbohydrate-binding domain, see Fig. 5	Li *et al.*, 2000
SP0368[d]	15457884	Cell wall surface anchor family protein	1767	1639	1439		
SP0453[b]	14971923	Amino acid ABC transporter, amino acid-binding protein/permease protein	521	274	274	A typical α/β domain with "periplasmic binding protein-like II" fold, anchored in the membrane; most likely not attached to peptidoglycan	
SP0462[e]	14971932	Cell wall surface anchor family protein	893	827	827	All-β domains flanking mixed α/β structure	
SP0463[e]	14971934	Cell wall surface anchor family protein	665	602	164	Contains several all-β domains	

SP0464[e]	14971935	Cell wall surface anchor family protein	393	342	342	Contains 3 all-β domains	
SP0498[b]	14971970	Endo-beta-N-acetylglucosaminidase	1659	1589	1017	Contains core TIM barrel along with PKD and other domains, see Fig. 6	
SP0641[c]	14972117	Cell wall-associated serine proteinase PrtA	2140	2090	1668	α/β catalyic domain along with probable Ig domains in C-terminal region	Bethe *et al.*, 2001
SP0648	14972122	Beta-galactosidase	2233	2156	1942	Core TIM barrel along with probable Ig domains in C-terminal region	Zahner and Hakenbeck, 2000
SP0664[c]	14972138	Zinc metalloprotease ZmpB	1881	1678	1148	Mixed α/β structure of unclear domain boundaries; N-terminal LPxTG motif	Novak *et al.*, 2000
SP0803[b,c]	14972280	Cell division protein RodA	407	107	95	A 10TM membrane protein of the FtsW/SpoVE family. Contains a 70-aa outside loop; the LPxTG motif is just before the last TM segment	
SP1154	14972632	Immunoglobulin A1 protease	2004	1849	1227	Mixed α/β structure of unclear domain boundaries	Wani *et al.*, 1996
SP1492[c]	14972981	Mucin-binding protein	202	170	136	Two domains, the first all-β, the second rich in proline; see Fig. 7	Roos and Jonsson, 2002
SP1772	14973269	Cell wall surface anchor family protein	4776	4709	154	A150-aa all-β domain of unknown fold preceding large unstructured Ser-rich domain. Attached to cell wall through a C-terminal LPxT/G motif and possibly also using a VPITG motif between the all-β and Ser-rich domains	
SP1833	14973334	Cell wall surface anchor family protein	708	642	555	Contains a central Amb-aII (right-handed β-helix) domain, found previously in pectate lyase. Could be involved in polysaccharide modification	Yoder et al., 1993
SP1992	14973500	Cell wall surface anchor family protein	221	191	176		

TABLE 3 (continued)

a - The three columns show the total length of the predicted protein, the length of its surface-exposed region (without the signal sequence, membrane-spanning segments and the fragment removed by sortase) and the total length of the exposed fragment(s) that is/are predicted to have globular structure (the surface-exposed part filtered for low-complexity regions). The latter was determined by using the SEG program with the following parameter set: trigger window length W = 45 residues, trigger complexity $K_2(1)$ =3.4 bits, and extension complexity $K_2(2)$ = 3.75 bits, as recommended by Wotton and Federhen (1996).

b - These proteins were not listed as peptidoglycan-anchored in the original study of Tettelin *et al.*, 2001.

c - The protein annotation has been modified from the original one

d - These proteins contain alternative sortase-cleavage motifs: IPRTG, IPQTG, and VPDTG (see Pallen *et al.*, 2001)

e - This protein has been omitted from *S. pneumoniae* TIGR4 genome annotation but translated in *S. pneumoniae* R6 genome.

smallest, human homologue. Secondary structure predictions clearly showed the alternating α- and β-structure elements expected of a TIM barrel. Most interestingly, however, strongly significant fold recognition matches were obtained to three of the four structurally differentiated families of Nagano *et al.* (2001). The failure of fold recognition to favor one or other of the known categories of TIM barrel leads to the conclusion that SP0498, and the CAZY glycoside hydrolase family 85 in general, contain a novel variant of the TIM barrel architecture. Nevertheless, two facts enabled the identification of likely catalytic residues: the knowledge that TIM barrel catalytic activity is invariably located at the C-terminal ends of the β-strands forming the barrel (Wierenga, 2001), and the independent information that, with very few exceptions, glycoside hydrolysis involves two or three conserved acidic residues (Nagano *et al.*, 2001). Glu337 and Asp374, perhaps along with Asp441, were highlighted by this analysis. Their positions on an approximate amylase TIM barrel-based model are shown in Plate 6.

4. SP1492 — A Mucin-Binding Protein

A recent study of mucin-binding properties of the bacterium *Lactobacillus reuteri* (Roos and Jonsson, 2002) found that this activity was conferred by a high-molecular-weight surface protein, anchored to the cell wall by a typical LPqTG motif. The authors identified in this protein two types of repeats, 183–197 and 184 amino acids residues long, respectively, both of which promoted adhesion the mucus. An analysis of the SP1492 sequence showed that it consists of a single copy of mucus-binding domain, closely related to both types of repeats in the *L. reuteri* protein (Plate 7), also followed by the peptidoglycan-anchoring signature. Although strict SEG filtering suggested that this protein is completely nonglobular, secondary structure analysis confidently predicts SP1492 to be an all-β protein. Remarkably, the same mucin-binding domain is found in human and cow protein hr44 (function unknown), indicating that SP1492 may promote adhesion of *S. pneumoniae* to mucosal cells without causing any immune response. Thus, despite its likely importance for adhesion, SP1492 is probably a poor vaccine candidate.

5. SP1833 — A Right-Handed β-Helical Protein, A Probable Polysaccharide Modifying Enzyme

This protein clearly has homologous sequences in *Streptomyces coelicolor* and *Staphylococcus aureus*. Beyond these, BLAST does not recognize any other significant sequence relationships so that SP1833 is currently annotated simply as a "cell wall surface anchor family protein". Nevertheless, the comparison of the SP1833 against the Pfam using its HMMer search tool or against the NCBI's Conserved Domain Database using RPS-BLAST (Marchler-Bauer *et al.*, 2002) suggest that the central part of this protein corresponds to the Amb-aII (PF00544) domain, found previously in pectate lyase (PDB entry 1qcx), which has a right-handed β-helix structure. While the E values reported by RPS-BLAST in the CDD search were fairly convincing (1e-06 and 2e-04), they were calculated using only a part of the domain alignment. In contrast, the HMM search of Pfam database returned a much longer alignment of SP1833 with pectate lyase, but with an unreliable E value of 0.056.

We carried out fold recognition experiments, again making use of the MetaServer (Bujnicki *et al.*, 2001a). Strongly significant results were obtained by several methods, indicating a structural correspondence between SP1833 and right-handed α-helical proteins. Although unequivocally establishing this relationship, the top scoring protein structures were not the same in each case, suggesting that SP1833 may be a member of a new family of such proteins. Indeed, for the fold recognition alignments, no conservation of catalytic residues between SP1883 and the highlighted enzymes was observed. Nevertheless, the structural limits of the α-helical domain could be clearly established using a variety of sources of information. These included the size of the significantly smaller *S. aureus* protein, the clearly amphiphilic predicted α-helix that commonly caps the N-terminal end of the α-helix (Jenkins *et al.*, 1998) and the nature and cross-method reliability of the secondary structure predictions. Additional help in threading the SP1883 sequence onto the α-helical architecture could be obtained through the identification of putative Asn-ladders (Jenkins *et al.*, 1998) — conserved Asn residues in adjacent turns of the α-helix whose stacking provides for favorable hydrogen bonding interactions. Unusually, the simple identification of fold, in this case, provides a strong hint as to function because most of the known enzymes containing right-handed α-helices act on polysaccharides (Jenkins *et al*., 1998), possessing hydrolase, lyase, or methylesterase activity. There seems to be a correlation between the presence of Asn ladders and lyase activity so that our predicted function for SP1883 is polysaccharide lyase. Although a function in formation or remolding of *S. pneumoniae*'s own polysaccharide coat cannot be ruled out, perhaps SP1883 most likely functions in the degradation of the host's extracellular polysaccharide matrix, thereby facilitating bacterial penetration into tissues. In this way it would be functionally analogous to *Flavobacterium heparinum* chondroitinase B, which also has the right-handed α-helical architecture.

VI. CONCLUSIONS

A. Bioinformatics Studies of Pneumococci

The five examples above are representative of the results that can be obtained by extensive bioinformatics analysis of *S. pneumoniae* proteins. They have in common the prediction of domain limits, through sequence repeats (SP0082), fold recognition (SP0314, SP0498, SP1833), size of homologues (SP0498, SP1492, SP1833) or secondary structure prediction (SP0498, SP1883). Such prediction is not only integral to the bioinformatics analysis, but essential for the structure determination by X-ray crystallography, a process that is now underway. Different conclusions were drawn in the four cases, ranging from the prediction of novel fold (SP0082), through the prediction of overall structural architecture (SP0498, SP1492, SP1833) to a clear identification of a specific related fold (SP0314). The simple identification of fold is enough in the case of SP1833 to predict functional category, while the determination of probable catalytic residues for SP0498 will undoubtedly help guide structure-function studies in glycoside hydrolase family 85. Thus, our ongoing bioinformatics analysis of *S. pneumoniae* surface proteins is providing numerous insights at various levels of protein structure and function, while guiding biophysical and biochemical experiments.

The goal of the study of *S. pneumoniae* surface proteins is to understand on the molecular level the role of these proteins in the mechanism of invasion of host tissues and the penetration of host defenses by pneumococci and other bacteria. Consequently, the results of such analyses followed by more studies shed new light on pathogen-host interactions and likely on bacterial pathogenesis in general. Such genome-derived results will likely induce further biochemical, functional, and pathogenesis studies to determine the importance of these molecules in pneumococcal pathogenesis.

B. Development of a Better Cure

The functions of all the above analyzed putative surface proteins facilitate significant aspects of pneumococcal colonization and/or invasion. Therefore, it follows that hindering their function will likely lead to compromised pathogenicity of *S. pneumoniae* (e.g., Berry and Paton, 2000). If investigation of the expression of these proteins by *S. pneumoniae*, confirmation of their surface-exposed character, and analysis of their immunogenic properties together confirm their pathogenic importance, these proteins can serve as targets for the development of novel cures against pneumococcal disease. Such further investigations should not be limited to one set of growth conditions but should include studies of various environments of *S. pneumoniae* such as those that can be modeled by (**1**) low iron, (**2**) high osmolarity, (**3**) growth on blood agar, (**4**) exposure to an atmosphere containing carbon dioxide, and (**5**) static temperature shift (Marra *et al*., 2002, Adamou *et al*., 2001; Weizmann *et al*., 2001). Some of the proteins discussed here might be utilized by the pathogen only during very specific conditions, while remaining crucial for the survival of *S. pneumoniae*. Two approaches may be envisaged. First, if the antibodies obtained against a given surface antigen are protective, then that protein becomes a candidate for incorporation into new vaccines. For example, PspA and PsaA are already being used to develop a novel vaccine and initial results are very promising (e.g., Jedrzejas, 2001; Nabors *et al*., 2000). The individual contributions of selected known virulence factors of *S. pneumoniae* to the pathogenesis of this organism, and hence their potential for the development of novel pneumococcal vaccines have been investigated recently (Berry and Paton, 2000). Secondly, some of these proteins could have their function compromised or totally abolished by small molecules binding most likely in their active sites. Therefore, these proteins are possible targets for the development of potent new drugs (e.g., Jedrzejas, 2001; Galperin and Koonin, 1999).

The availability of reliable functional annotation and modeled or experimental (X-ray or NMR) three-dimensional structural information (as shown in Tables 1 and 3) for pneumococcal proteins will certainly facilitate the elucidation of their detailed function and mechanism. Such knowledge will aid in the development of treatment strategies for pneumococcal disease as well as further scientific understanding in general. However, structural information needs to be accompanied by an increased understanding of the expression, surface character, and the role of such proteins during various stages of pathogenesis in animals, and ultimately in humans.

Polyvalent vaccines based on purified capsular polysaccharides, like the available pneumococcal vaccine, are limited in their potency because of their poor immunogenicity, especially in susceptible groups of patients like young children and the elderly (Stansfield, 1987). The poor immunogenicity of polysaccharide vaccines is primarily due to a poor antibody response elicited by these vaccines and because the T-cell independence of the response fails to induce memory. In addition, the available pneumococcal vaccines comprise only a limited number of serotypes out of over 90 known. The development of conjugated vaccines by coupling the polysaccharides with protein carriers increases the potency of the vaccine, as it is the case for Prenvar, but also limits the serotypes that can be included in such conjugate mixtures (only seven types of conjugated polysaccharides are included in Prenvar). The combination of polysaccharides with a protein has been shown to significantly increase immunogenicity and immunological memory to polysaccharide antigens. The protein carrier(s) of conjugated vaccines induce additional antibodies, thereby improving the degree of protection offered. Such additional protection might also be independent of serotype as it is the case for several pneumococcal surface immunogenic proteins. Therefore, the development of a two-component vaccine comprising a polysaccharide and a nonpolysaccharide part, such as the proteins discussed above, is a promising approach (Alexander *et al*., 1994; Nabors *et al*., 2000; Paton, 1998; Tuomanen, 1999). More studies are, however, needed to assess the usefulness of various antigens or their mixtures in various modes of pneumococcal challenge. The bioinformatics-based mining of genomes for identifying and characterizing possible candidate proteins for further studies shows excellent potential for the acceleration of the development of better treatments for pneumococcal and other diseases.

ACKNOWLEDGMENTS

Work in the authors' laboratories was supported by grants AI44079 from the National Institutes of Health (NIH; MJJ) and N66001-01-C-8013 from the Defense Advanced Research Projects Agency (DARPA; MJJ) and by the Conselho Nacional de Desenvovimento Cientifico e Tecnologico (CNPq; DJR) and the Projeto de Apoio ao Desenvolvimento de Tecnologia Agropecuária para o Brasil (PRODETAB; DJR).

REFERENCES

Adamou, J. E., Heinrichs, J. H., Erwin, A. L., Walsh, W., Gayle, T., Dormitzer, M., Dagan, R., Brewah, Y. A, Barren, P., Lathigra, R., Langermann, S., Koenig, S., and Johnson, S. 2001. Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. *Infect. Immun.* **69**: 949–958.

Alexander, J. E., Lock, R. A., Peeters, C. C., Poolman, J. T., Andrew, P. W., Mitchell, T. J., Hansman, D., and Paton, J. C. 1994. Immunization of mice with pneumolysin toxoid confers a significant degree of protection against at least nine serotypes of *Streptococcus pneumoniae*. *Infect. Immun.* **62**: 5683–5688.

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**: 395–408.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.

Anon. 1985. Acute respiratory infections in underfives. 15 million deaths a year. *Lancet* **2**: 699–701.

Banas, J. A., Russell, R. R., and Ferretti, J. J. 1990. Sequence analysis of the gene for the glucan-binding protein of *Streptococcus mutans* Ingbritt. *Infect. Immun.* **58**: 667–673.

Bartlett, G.J., Porter, C.T., Borkakoti, N., and Thornton, J.M. 2002. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**: 105–121.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30**: 276–280.

Berks, B. C., Sargent, F., and Palmer, T. 2000. The Tat protein export pathway. *Mol. Microbiol.* **35**: 260–274.

Berry, A. M. and Paton, J. C. 2000. Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. *Infect. Immun.* **68**: 133–140.

Berry, A. M. and Paton, J. C. 1996. Sequence heterogeneity of PsaA, a 37–kilodalton putative adhesin essential for virulence of *Streptococcus pneumoniae*. *Infect. Immun.* **64**: 5255–5262.

Berry, A. M., Lock, R. A., and Paton, J. C. 1996. Cloning and characterization of nanB, a second *Streptococcus pneumoniae* neuraminidase gene, and purification of the NanB enzyme from recombinant *Escherichia coli*. *J. Bacteriol.* **178**: 4854–4860.

Berry, A. M., Lock, R. A., Thomas, S. M., Rajan, D. P., Hansman, D., and Paton, J. C. 1994. Cloning and nucleotide sequence of the *Streptococcus pneumoniae* hyaluronidase gene and purification of the enzyme from recombinant *Escherichia coli*. *Infect. Immun.* **62**: 1101–1108.

Bethe, G., Nau, R., Wellmer, A., Hakenbeck, R., Reinert, R. R., Heinz, H. P., and Zysk, G. 2001. The cell wall-associated serine protease PrtA: a highly conserved virulence factor of *Streptococcus pneumoniae*. *FEMS Microbiol. Lett.* **205**: 99–104.

Brady, G. P. and Stouten, P. F. W. 2000. Fast prediction and visualization of protein binding pockets with PASS. *J. Comp. Aided. Mol. Des.* **14**: 383–401.

Brooks-Walter, A., Briles, D. E., and Hollingshead, S. K. 1999. The *pspC* gene of *Streptococcus pneumoniae* encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infect. Immun.* **67**: 6533–6542.

Brun, E., Johnson, P. E., Creagh, A. L., Tomme, P., Webster, P., Haynes, C. A., and McIntosh, L. P. 2000. Structure and binding specificity of the second N-terminal cellulose-binding domain from *Cellulomonas fimi* endoglucanase C. *Biochemistry* **39**: 2445–2458.

Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski L. 2001a. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins* **45** (Suppl 5): 184–191.

Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. 2001b. Structure prediction meta server. *Bioinformatics* **17**: 750–751.

Camara, M., Boulnois, G. J., Andrew, P. W., and Mitchell, T. J. 1994. A neuraminidase from *Streptococcus pneumoniae* has the features of a surface protein. *Infect. Immun.* **62**: 3688–3695.

Chung, S. Y. and Subbiah, S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure* **4**: 1123–1127.

Clarke, V. A., Platt, N., and Butters, T. D. 1995. Cloning and expression of the *N*-acetylglucosaminidase gene from *Streptococcus pneumoniae*. Generation of truncated enzymes with modified aglycon specificity. *J. Biol. Chem.* **270**: 8805–8814.

Cohen, J. 1994. Bumps on the vaccine road. *Science* **265**: 1371–1373.

Coutinho, P. M. and Henrissat, B. 1999a. Carbohydrate-active enzymes: an integrated database approach. **In:** *Recent Advances in Carbohydrate Bioengineering*. pp. 3–12. Gilbert, H.J., Davies, G., Henrissat, B., and Svensson, B., Eds., The Royal Society of Chemistry, Cambridge.

Coutinho, P. M. and Henrissat, B. 1999b. The modular structure of cellulases and other carbohydrate-active enzymes: An integrated database approach. **In:** *Genetics, Biochemistry and Ecology of Cellulose Degradation*. pp. 15–23. Ohmiya, K., Hayashi, K., Sakka, K., Kobayashi, Y., Karita, S., and Kimura, T., Eds., Uni Publishers, Tokyo,

Coutinho, P.M. and Henrissat, B. 1999c. Carbohydrate-active Enzymes server at URL: http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html

Crennell, S. J., Garman, E. F., Laver, W. G., Vimr, E. R., and Taylor, G. L. 1993. Crystal structure of a bacterial sialidase (from *Salmonella typhimurium* LT2) shows the same fold as an influenza virus neuraminidase. *Proc. Natl. Acad. Sci. U.S.A.* **90**: 9852–9856.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. 2003. WebLogo: A sequence logo generator. *Genome Res.* **13**: (in press).

Cundell, D. R., Gerard, N. P., Gerard, C., Idanpaan-Heikkila, I., and Tuomanen, E. I. 1995. *Streptococcus pneumoniae* anchor to activated human cells by the receptor for platelet-activating factor. *Nature* **377**: 435–438.

Deivanayagam, C. C., Rich, R. L., Carson, M., Owens, R. T., Danthuluri, S., Bice, T., Hook, M., and Narayana, S. V. 2000. Novel fold and assembly of the repetitive B region of the *Staphylococcus aureus* collagen-binding surface protein. *Structure* **8**: 67–78.

DeLano, W.L. 2002. The PyMOL Molecular Graphics System on World Wide Web http://www.pymol.org.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41**: 98–107.

Devos, D. and Valencia, A. 2001. Intrinsic errors in genome annotation. *Trends Genet.* **17**: 429–431.

Dintilhac, A., Alloing, G., Granadel, C., and Claverys, J. P. 1997. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. *Mol. Microbiol.* **25**: 727–739.

Dopazo, J., Mendoza, A., Herrero, J., Caldara, F., Humbert, Y., Friedli, L., Guerrier, M., Grand-Schenk, E., Gandin, C., de Francesco, M., Polissi, A., Buell, G., Feger, G., Garcia, E., Peitsch, M., and Garcia-Bustos, J.F. 2001. Annotated draft genomic sequence from a *Streptococcus pneumoniae* type 19F clinical isolate. *Microb. Drug Resist.* **7**: 99–125.

Durbin, R., Eddy, S., Krogh. A., and Mitchison, A. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Fedson, D. S. and Musher, D. M. 1994. Pneumococcal vaccine. **In:** *Vaccines*. 2nd ed., pp. 517–563. Plotkin, S. A. and Mortimer, E. A., Jr., Eds.,WB Saunders, Philadelphia, PA.

Feldman, C., Munro, N. C., Jeffery, P. K., *et al.* 1992. Pneumolysin induces the salient histologic features of pneumococcal infection in the rat lung *in vivo*. *Am. J. Respir. Cell. Mol. Biol.* **5**: 416–423.

Fernandez-Tornero, C., Garcia, E., Lopez, R., Gimenez-Gallego, G., and Romero, A. 2002. Two new crystal forms of the choline-binding domain of the major pneumococcal autolysin: Insights into the dynamics of the active homodimer. *J. Mol. Biol.* **321**: 163–173.

Fernandez-Tornero, C., Lopez, R., Garcia, E., Gimenez-Gallego, G., and Romero, A. 2001. A novel solenoid fold in the cell wall anchoring domain of the pneumococcal virulence factor LytA. *Nat. Struct. Biol.* **8**: 1020–1024.

Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. R., and Dunbrack, R. L., Jr. 2001. CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins* **45** (Suppl 5): 171–83.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., and Mewes, H. W. 2001. Functional and structural genomics using PEDANT. *Bioinformatics* **17**: 44–57.

Galperin, M.Y. and Koonin, E.V. 1999. Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.* **10**: 571–578.

George, R. A. and Heringa, J. 2002. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.* **316**: 839–851.

Glass, J. I., Belanger, A. E., and Robertson, G. T. 2002. *Streptococcus pneumoniae* as a genomics platform for broad-spectrum antibiotic discovery. *Curr. Opin. Microbiol.* **5**: 338–342.

Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. 1999. ESPript: Multiple sequence alignments in PostScript. *Bioinformatics* **15**: 305–308.

Gracy, J. and Argos, P. 1998. DOMO: A new database of aligned protein domains. *Trends Biochem Sci.* **23**: 495–497.

Gray, B. M., Converse, G. M., 3rd, and Dillon, H. C., Jr. 1980. Epidemiologic studies of *Streptococcus pneumoniae* in infants: Acquisition, carriage, and infection during the first 24 months of life. *J. Infect. Dis.* **142**: 923–933.

Heger, A. and Holm, L. 2000. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**: 224–237.

Holm, L. and Sander, C. 1997. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins* **28**: 72–82.

Hoskins, J., Alborn, W. E. Jr, Arnold, J., Blaszczak, L. C., Burgett, S., DeHoff, B. S, Estrem, S. T., Fritz, L., Fu, D. J, Fuller, W., Geringer, C., Gilmour, R., Glass, J. S., Khoja, H., Kraft, A. R., Lagace, R. E., LeBlanc, D. J., Lee, L. N., Lefkowitz, E. J., Lu, J., Matsushima, P., McAhren, S. M., McHenney, M., McLeaster, K., Mundy, C. W., Nicas, T. I., Norris, F. H., O'Gara, M., Peery, R. B., Robertson, G. T., Rockey, P., Sun, P. M., Winkler, M. E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C. A., Baltz, R. H., Jaskunas, S. R., Rosteck, P. R. Jr, Skatrud, P. L., and Glass J. I. 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J. Bacteriol.* **183**: 5709–5717.

Ilangovan, U., Ton-That, H., Iwahara, J., Schneewind, O., and Clubb, R. T. 2001. Structure of sortase, the transpeptidase that anchors proteins to the cell wall of *Staphylococcus aureus*. *Proc. Natl. Acad. Sci. USA* **98**: 6056–6061.

Inouye, S., Soberon, X., Franceschini, T., Nakamura, K., Itakura, K., and Inouye, M. 1982. Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. *Proc. Natl. Acad. Sci. USA*. **79**: 3438–3441.

Jedrzejas, M. J. 2002. Three-dimensional structures of hyaluronate lyases from *Streptococcus* species and their mechanism of hyaluronan degradation. **In:** *Science of Hyaluronan Today*, Hascall, V.C. and Yanagishita, M. Eds., Glycoforum, www.glycoforum.gr.jp/science/hyaluronan.

Jedrzejas, M. J. 2001. Pneumococcal virulence factors: Structure and function. *Microbiol. Mol. Biol. Rev.* **65**: 187–207.

Jedrzejas, M. J. 2000. Structural and functional comparison of polysaccharide-degrading enzymes. *Crit. Rev. Biochem. Mol. Biol.* **35**:221–251.

Jedrzejas, M. J., Mello, L. V., de Groot, B. L., and Li., S. 2002. Mechanism of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase. Structures of complexes with the substrate. *J. Biol. Chem.* **277**: 28287–28297.

Jedrzejas, M. J., Lamani, E., and Becker, R. S. 2001. Biophysical characterization of selected strains of pneumococcal surface protein A. *J. Biol. Chem.* **276**: 33121–33128.

Jedrzejas, M. J., Hollingshead, S. K., Lebowitz, J., Chantalat, L., Briles, D. E., and Lamani, E. 2000. Production and characterization of pneumococcal surface protein A. *Arch. Biochem. Biophys.* **373**: 116–125.

Jenkins, J., Mayans, O., Smith, D., Worboys, K., and Pickersgill, R. W. 2001. Three-dimensional structure of *Erwinia chrysanthemi* pectin methylesterase reveals a novel esterase active site. *J. Mol. Biol.* **305**: 951–960.

Johnson, P. E., Joshi, M.D., Tomme, P., Kilburn, D.G., and McIntosh, L. P. 1996. Structure of the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **35**: 14381–14394.

Johnston, R. B., Jr. 1991. Pathogenesis of pneumococcal pneumonia. *Rev. Infect. Dis.* **13**: S509–S517.

Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **29**: 195–202.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**: 846–856.

Kelly, S. J. and Jedrzejas, M. J. 2000a. Structure and molecular mechanism of a functional form of pneumolysin: a cholesterol-dependent cytolysin from *Streptococcus pneumoniae*. *J. Struct. Biol.* **132**: 72–81.

Kelly, S. J. and Jedrzejas, M. J. 2000b. Crystallization and preliminary X-ray analysis of a functional form of pneumolysin, A virulence factor from *Streptococcus pneumoniae*. *Acta Cryst.* **D56**: 1452–1455.

Koonin, E. V. and Galperin, M. Y. 2002. *Sequence — Evolution — Function. Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, Boston.

Kuroda, Y., Tani, K., Matsuo, Y., and Yokoyama, S. 2000. Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics. *Protein Sci.* **9**, 2313–2321.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**: 283–291.

Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. 1996. Protein clefts in molecular recognition and function. *Protein Sci.* **5**: 2438–2452.

Lawrence, M. C., Pilling, P. A., Epa, V. C., Beery, A. M., Ogunniyi, A. D., and Paton, J. C. 1998. The crystal structure of pneumococcal surface antigen PsaA reveals a metal-binding site and a novel structure for a putative ABC-type binding protein. *Structure* **6**: 1553–1561.

Li, S., Kelly, S. J., Lamani, E., Ferraroni, M., and Jedrzejas, M.J. 2000. Structural basis of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase. *EMBO J.* **19**: 1228–1240.

Li, W., Pio, F., Pawlowski, K., and Godzik, A. 2000. Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics* **16**: 1105–1110.

Lock, R. A., Paton, J. C., and Hansman, D. 1988. Purification and immunological characterization of neuraminidase produced by *Streptococcus pneumoniae*. *Microb. Pathog.* **5**: 461–467.

Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. 2001. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* **10**: 2354–2362.

Lupas, A. 1996. Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**: 513–525.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., and Bryant, S. H. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**: 281–283.

Mark, B. L., Vocadlo, D. J., Knapp, S., Triggs-Raine, B. L., Withers, S. G., and James, M. N. 2001. Crystallographic evidence for substrate-assisted catalysis in a bacterial beta-hexosaminidase. *J. Biol. Chem.* **276**: 10330–10337.

Marra, A., Asundi, J., Bartilson, M., Lawson, S., Fang, F., Christine, J., Wiesner, C., Brigham, D., Schneider,

W. P., and Hromockyj, A. E. 2002. Differential fluorescence induction analysis of *Streptococcus pneumoniae* identifies genes involved in pathogenesis. *Infect. Immun.* **70**: 1422–1433.

Mazmanian, S. K., Liu, G., Ton-That, H., and Schneewind, O. 1999. *Staphylococcus aureus* sortase, an enzyme that anchors surface proteins to the cell wall. *Science* **285**, 760–763.

McDaniel, L. S., Sheffield, J. S., DeLucchi, P., and Briles, D. E. 1991. PspA, a surface protein of *Streptococcus pneumoniae*, is capable of eliciting protection against pneumococci of more than one capsular type. *Infect. Immun.* **59**: 222–228.

Meddrano, F. J., Gasset, M., Lopez, Zumel, C., Usobiaga, P., Garcia, J. L., and Menendez, M. 1996. Structural characterization of the unligated and choline-bound forms of the major pneumococcal autolysin LytA amidase. Conformational transitions induced by temperature. *J. Biol. Chem.* **271**: 29152–29161.

Mufson, M. A. 1990. *Streptococcus pneumoniae*. **In:** *Principles and Practice of Infectious Diseases*. pp. 1539–1550, Mandell, G. L., Douglas, R. G. Jr., and Bennett, J. E., Eds., Churchill Livingstone, New York.

Munoa, F. J., Miller, K. W., Beers, R., Graham, M., and Wu, H. C. 1991. Membrane topology of *Escherichia coli* prolipoprotein signal peptidase (signal peptidase II). *J. Biol. Chem.* **266**: 17667–17672.

Musher, D. M. 1991. Infections caused by *Streptococcus pneumoniae*: Clinical spectrum, pathogenesis, immunity, and treatment. *Clin. Infect. Dis.* **14**: 801–807.

Nabors, G. S., Braun, P. A., Herrmann, D. J., Heise, M. L., Pyle, D. J., Gravenstein, S., Schilling, M., Ferguson, L. M., Hollingshead, S.K., Briles, D.E., and Becker, R.S. 2000. Immunization of healthy adults with a single recombinant pneumococcal surface protein A (PspA) variant stimulates broadly cross-reactive antibodies to heterologous PspA molecules. *Vaccine* **18**: 1743–1754.

Nagano, N., Porter, C. T., and Thornton, J. M. 2001. The (betaalpha)(8) glycosidases: sequence and structure analyses suggest distant evolutionary relationships. *Protein Eng.* **14**: 845–855.

Nakai, K. 2001. Prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* **134**: 103–116.

Nakai, K. and Horton, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**: 34–36.

Navarre, W. W. and Schneewind, O. 1999. Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* **63**: 174–229.

Nicholls, A., Sharp, K. A., and Honig, B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**: 281–296.

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

Norin, M. and Sundstrom, M. 2002. Structural proteomics: Developments in structure-to-function predictions. *Trends Biotechnol.* **20**: 79–84.

Notredame, C., Higgins, D. G., and Heringa, J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**: 205–217.

Novak R., Charpentier E., Braun J. S., Park E., Murti S., Tuomanen E., and Masure R. 2000. Extracellular targeting of choline-binding proteins in *Streptococcus pneumoniae* by a zinc metalloprotease. *Mol. Microbiol.* **36**: 366–376.

Novak, R., Henriques, B., Charpentier, E., Normark, S., and Tuomanen, E. 1999. Emergence of vancomycin tolerance in *Streptococcus pneumoniae*. *Nature* **399**: 590–593.

Ofran, Y. and Rost, B. 2003. Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**: 377–387.

Pallen, M. J. Lam, A. C., Antonio, M., and Dunbar, K. 2001. An embarrassment of sortases — a richness of substrates? *Trends Microbiol.* **9**: 97–102.

Paton, J C. 1998. Novel pneumococcal surface proteins: role in virulence and vaccine potential. *Trends Microbiol.* **6**: 85–87.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**: 63–98.

Quiocho, F. A. and Vyas, N. K. 1999. *Carbohydrates*. **In:** *Bioinorganic Chemistry*. pp. 441–457, Oxford University Press, New York.

Rigden, D. J. 2002. Covariance analysis and prediction of structural domains from boundaries from multiple protein sequence alignments. *Protein Eng.* **15**: 65–77.

Rigden, D. J. and Carneiro, M. 1999. A structural model for the rolA protein and its interaction with DNA. *Proteins* **37**: 697–708.

Rigden, D. J., Alexeev, D., Phillips, S. E., and Fothergill-Gilmore, L. A. 1998. The 2.3 Å X-ray crystal structure of *S. cerevisiae* phosphoglycerate mutase. *J. Mol. Biol.* **276**: 449–459.

Rigden, D. J., Bagyan, I., Lamani, E., Setlow, P., and Jedrzejas M. J. 2001b. A cofactor-dependent phosphoglycerate mutase homolog from *Bacillus stearothermophilus* is actually a broad specificity phosphatase. *Protein Sci.* **10**: 1835–1846.

Rigden, D. J., Mello, L. V., and Bertioli, D. J. 2000. Structural modeling of a plant disease resistance gene product domain. *Proteins* **41**: 133–143.

Rigden, D. J., Monteiro, A. C., and Grossi de Sa, M. F. 2001a. The protease inhibitor chagasin of *Trypanosoma cruzi* adopts an immunoglobulin-type fold and may have arisen by horizontal gene transfer. *FEBS Lett.* **504**: 41–44.

Rigden, D. J., Setlow, P., Setlow, B., Stein, R. A., and Jedrzejas, M. J. 2002. PrfA protein of *Bacillus* species: Prediction and demonstration of DNA nuclease activity. *Protein Sci.* **11**: 2370–2381.

Rigden, D. J. and Jedrzejas, M.J. 2003. Genome-based identification of a carbohydrate binding module in *Streptococcus pneumoniae* hyaluronate lyase. *Proteins*, in press.

Roos, S. and Jonsson, H. 2002. A high-molecular-mass cell-surface protein from *Lactobacillus reuteri* 1063 adheres to mucus components. *Microbiology* **148**: 433–442.

Rosenow, C., Ryan, P., Weiser, J. N., Johnson, S., Fontan, P., Ortqvist, A., and Masure, H. R. 1997. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*. *Mol. Microb.* **25**: 819–829.

Rossjohn, J., Feil, S.C., McKinstry, W.J., Tweten, R.K., and Parker, M.W. 1997. Structure of a cholesterol-binding, thiol-activated cytolysin and a model of its membrane form. *Cell* **89**: 685–692.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**: 521–533.

Sali, A. and Blundell, T. L. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**: 779–815.

Sampson, J. S., O'Connor, S. P., Stinson, A. R., Tharpe, J. A., and Russell, H. 1994. Cloning and nucleotide sequence analysis of psaA, the *Streptococcus pneumoniae* gene encoding a 37–kilodalton protein homologous to previously reported *Streptococcus* sp. adhesins. *Infect. Immun.* **62**: 319–324.

Sippl, M. J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**: 355–362.

Sippl, M. J., Lackner, P., Domingues, F. S., Prlic, A., Malik, R., Andreeva, A., and Wiederstein, M. 2001. Assessment of the CASP4 fold recognition category. *Proteins* **45** (Suppl 5): 55–67.

Stansfield, S. K. 1987. Acute respiratory infections in the developing world: Strategies for prevention, treatment and control. *Pediatr. Infect. Dis.* **6**: 622–629.

Tam, R. and Saier, M. H. Jr. 1993. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev.* **57**: 320–346.

Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., Heidelberg, J., DeBoy, R. T., Haft, D. H., Dodson, R. J., Durkin, A. S., Gwinn, M., Kolonay, J. F., Nelson, W. C., Peterson, J. D., Umayam, L. A., White, O., Salzberg, S. L., Lewis, M. R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A. M., Utterback, T. R., Hansen, C. L., McDonald, L. A., Feldblyum, T. V., Angiuoli, S., Dickinson, T., Hickey, E. K., Holt, I. E, Loftus, B. J., Yang, F., Smith, H. O., Venter, J. C., Dougherty, B. A., Morrison, D. A., Hollingshead, S. K., and Fraser, C. M. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**: 498–506.

Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S., and van Dijl, J. M. 2000. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.* **64**: 515–547.

Todd, A. E., Orengo, C. A., and Thornton, J. M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**: 1113–1143.

Todd, A. E., Orengo, C. A., and Thornton, J. M. 2002. Sequence and structural differences between enzyme and nonenzyme homologs. *Structure* **10**: 1435–1451.

Ton-That, H., Liu, G., Mazmanian, S. K., Faull, K. F., and Schneewind, O. 1999. Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of *Staphylococcus aureus* at the LPXTG motif. *Proc. Natl. Acad. Sci. USA* **96**: 12424–12429.

Tuomanen, E. 1999. Molecular and cellular biology of pneumococcal infection. *Curr. Opin. Microb.* **2**: 35–39.

Usobiaga, P., Medrando, F. J., Gasset, M., Garcia, J. L., Saiz, J. L., Rivas, G., Laynez, J., and Menendez, M. 1996. Structural organization of the major autolysin from *Streptococcus pneumoniae*. *J. Biol. Chem.* **271**: 6832–6838.

van Dijl, J. M., Braun, P. G., Robinson, C., Quax, W. J., Antelmann, H., Hecker, M., Muller, J., Tjalsma, H., Bron, S., and Jongbloed, J. D. 2002. Functional genomic analysis of the *Bacillus subtilis* Tat pathway for protein secretion. *J. Biotechnol.* **98**: 243–54.

von Heijne, G. 1989. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341**: 456–458.

Walker, D. R. and Koonin, E. V. 1997. SEALS: a system for easy analysis of lots of sequences. *ISMB* **5**: 333–339.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. 1997. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. *Protein Sci.* **6**: 2308–2323.

Wani, J. H., Gilbert, J. V., Plaut, A. G., and Weiser, J. N. 1996. Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of *Streptococcus pneumoniae*. *Infect. Immun.* **64**: 3967–3974.

Wheelan, S. J., Marchler-Bauer, A., and Bryant, S. H. 2000. Domain size distributions can predict domain boundaries. *Bioinformatics* **16**: 613–618.

Wierenga, R. K. 2001. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**: 193–198.

Wilson, C. A., Kreychman, J., and Gerstein, M. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**: 233–249.

Wizemann, T. M., Heinrichs, J. H., Adamou, J. E., Erwin, A. L., Kunsch, C., Choi, G. H., Barash, S. C., Rosen, C. A., Masure, H. R., Tuomanen, E. et al. 2001. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* **69**: 1593–1598.

Wootton, J. C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18**: 269–285.

Wootton, J. C. and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**: 554–571.

Wren, B. W. 1991. A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. *Mol. Microbiol.* **5**: 797–803.

Yoder, M. D., Lietzke, S. E., and Jurnak, F. 1993. Unusual structural features in the parallel beta-helix in pectate lyases. *Structure* **1**: 241–251.

Zahner, D. and Hakenbeck, R. 2000. The *Streptococcus pneumoniae* beta-galactosidase is a surface protein. *J. Bacteriol.* **182**: 5919–5921.

Zhou, H.X. and Shan, Y. 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44**: 336–343.