Genomics 100 (2012) 370–379

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Protein domain repetition is enriched in Streptococcal cell-surface proteins

I-Hsuan Lin [a,b], Ming-Ta Hsu [c,d], Chuan-Hsiung Chang [a,e,*]

[a] Institute of BioMedical Informatics, National Yang-Ming University, Taipei, Taiwan
[b] Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
[c] VGH Yang-Ming Genome Research Center, National Yang-Ming University, Taipei, Taiwan
[d] Institute of Biochemistry and Molecular Biology, National Yang-Ming University, Taipei, Taiwan
[e] Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 4 February 2012
Accepted 7 August 2012
Available online 22 August 2012

Keywords:
Domain repeats
Domain repetition
Protein subcellular localization
Protein structure modeling
Virulence
Streptococcus

ABSTRACT

Tandem repetition of domain in protein sequence occurs in all three domains of life. It creates protein diversity and adds functional complexity in organisms. In this work, we analyzed 52 streptococcal genomes and found 3748 proteins contained domain repeats. Proteins not harboring domain repeats are significantly enriched in cytoplasm, whereas proteins with domain repeats are significantly enriched in cytoplasmic membrane, cell wall and extracellular locations. Domain repetition occurs most frequently in *S. pneumoniae* and least in *S. thermophilus* and *S. pyogenes*. DUF1542 is the highest repeated domain in a single protein, followed by Rib, CW_binding_1, G5 and HemolysinCabind. 3D structures of 24 repeat-containing proteins were predicted to investigate the structural and functional effect of domain repetition. Several repeat-containing streptococcal cell surface proteins are known to be virulence-associated. Surface-associated tandem domain-containing proteins without experimental functional characterization may be potentially involved in the pathogenesis of streptococci and deserve further investigation.

1. Introduction

Protein domains are viewed as the basic building blocks of a protein and often contribute to specific tertiary structure and function. Most proteins are composed of one or a combination of several domains. Through domain shuffling, an organism can gain novel proteins with complex domain organizations and diverse functions. One type of the domain architectures is the tandem repetition of a domain family within a protein sequence. Domain repeats occur in all three kingdoms of life but are particularly common in multicellular eukaryotes [1–3]. It was believed that the formation of domain repeats allow eukaryotes to expand its proteome repertoire with increased functional versatility. The precise genetic mechanisms responsible for domain duplication are not thoroughly understood, however homologous recombination or site-specific recombination between tandem repeats may be involved in the tandem insertion/deletion of domain repeats [4].

The diverse variation in the number and sequence composition of protein domain repeats results in structures capable of binding to repetitive surfaces of a variety of binding partners. The C2H2-type zinc finger, leucine-rich repeat and WD40 repeats are the most common repeated domains in multicellular eukaryotes [2]. They function as interaction modules that bind to DNA, RNA, proteins, or small molecules. These motifs are present in proteins involved in diverse biological processes, including transcription, translation, protein folding, signal transduction, cell adhesion, RNA processing, apoptosis, and the immune response. Although much work has focused on eukaryotic organisms, there is little systematic attempt to analyze tandem domains in prokaryotes.

Several surface proteins from pathogenic microorganisms, including streptococci [5–7], *Clostridium difficile* [7], *Leuconostoc mesenteroides* [6], *Burkholderia pseudomallei* [8], *B. mallei* [8], *Mycoplasma hominis* [9] and *Haemophilus* cryptic genospecies [10] have been found to possess domain repeats. Variations in the number of repeats have been shown to generate antigenic and functional diversity among some bacterial surface proteins [9–13]. However, there has been a lack of systematic discovery of such domain architecture in bacteria. In this study, we focused on the computational identification of tandem domains in 52 streptococcal proteomes. *Streptococcus* is the most sequenced bacterial genus to date. Members of this genus include beneficial bacteria used in food fermentation, normal flora in animals and humans, and important pathogenic bacteria capable of infecting humans and a wide range of animals. The surface-associated proteins and secreted factors of streptococci play important roles in the interaction with host tissue and blood components [14–20]. In the present study, we describe a comprehensive analysis of domain repetition in streptococcal proteomes. We also explored the effect of such repeat architecture on protein structure based on computational modeling. Analysis of the results reveals that streptococcal proteins that contain domain repeats are significantly enriched in extracytoplasmic locations and proteins that contained highly repeated sequences are mostly virulence-associated. Our results also

* Corresponding author at: Institute of Biomedical Informatics, National Yang-Ming University, 155, Sec. 2, Linong St., Taipei 11221, Taiwan. Fax: +886 2 28206754.
E-mail address: cchang@ym.edu.tw (C.-H. Chang).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.08.001

provide the basis to further identify this type of domain architecture in other bacterial genera.

2. Materials and methods

2.1. Sequence data

The predicted proteomes of 52 streptococcal genomes, representing 14 species, were used in this study (see Supplementary data 1). The protein sequences were collected from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

2.2. Protein domain prediction

The protein sequences were compared with the hidden Markov models (HMM) of the Pfam-A database (version 25.0) [21] using the pfam_scan.pl script (version 1.3) available at the Pfam FTP site (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/) and the HMMER hmmscan program (version 3.0) [22]. Domain hits that have scores and e-values equal or greater than the 'gathering' (GA) threshold, individually defined by the Pfam curators for each domain, were reported. In this study, domain repetition is defined as at least two adjacent domains of the same Pfam ID and no more than 110 unassigned amino acid residues between the domains.

2.3. Protein subcellular localization prediction

Putative subcellular localization of proteins was determined using PSORTb (version 3.0.3) by selecting prediction specific for Gram-positive bacteria [23]. Normal output format from the prediction displays both the final prediction results and associated scores. The Gram-positive localization sites predicted include: cytoplasm (CYT), cytoplasmic membrane (CPM), cell wall (CWL), extracellular space (EXC), multiple sites (MUL) and unknown sites (UKN).

2.4. Statistical enrichment analysis

This analysis assesses whether a certain number of repeats occurs more frequently in a particular subcellular localization than it would be expected based on the frequency of its occurrence within the defined 'protein universe'. The test for overrepresentation was achieved by calculating the hypergeometric probability using the phyper function in the R statistical package. The cutoff point of $P<0.00001$ was used to determine if a certain number of repeats was significantly enriched in a specific subcellular location.

2.5. Protein 3D structure modeling

The I-TASSER server, ranked as the number 1 server for protein structure prediction in CASP7 (2006), CASP8 (2008) and CASP9 (2010) experiments, was used to perform automated *ab initio* protein tertiary structure prediction [24]. The qualities of the models were assessed using PROCHECK via PDBsum [25]. Hydrogen atoms were added to the predicted protein structure by PDB2PQR (version 1.7) [26]. Partial atomic charges and van der Waals radii were assigned according to PARSE force field parameters. The electrostatic surface potentials were calculated by solving the linearized PoissoneBoltzmann equation (PBE) with the APBS package (version 1.3) [27]. 3D structures with surface potentials produced by APBS were visualized using Jmol viewer (version 12.0) [28]. The molecular hydrophobic and hydrophilic properties of the predicted models were calculated by the PLATINUM web-server based on the concept of empirical Molecular Hydrophobicity Potential [29].

3. Results

3.1. Significant occurrence of domain repetition in surface-associated proteins

The presence of Pfam protein domain in the 52 streptococcal proteomes was identified using the Perl script "pfam_scan.pl" supplied by Pfam. Of the 104140 total streptococcal proteins, the "protein universe", 85496 proteins (82.1%) contain at least one Pfam domain. Of these, 81748 proteins (95.6%) harbor no repeated domain, 3446 proteins (4.0%) have two to four repeats and 302 proteins (0.4%) contain domain repeated five or more times. The protein subcellular localizations were predicted using PSORTb. The number of proteins in each of the six localization sites are as follows: 53707 (51.6%) in cytoplasm (CYT), 27499 (26.4%) in cytoplasmic membrane (CPM), 1319 (1.3%) in cell wall (CWL), 1534 (1.5%) in extracellular space (EXC), 244 (0.2%) in multiple sites (MUL) and 19837 (19.0%) with unknown subcellular localizations (UKN). Information about the subcellular localization and protein domain annotation is summarized in Fig. 1. The Pfam domain and subcellular localization information of the 3748 proteins containing domain repeats is provided in Supplementary data 2.

Hypergeometric distribution was used to statistically assess the correlation of each repeat number-subcellular localization pair. As shown in Supplementary data 3, proteins with no repeated domains were significantly enriched only in CYT. On the other hand, proteins containing tandem domains are significantly enriched CWL and EXC. Cell wall proteins tend to have fewer repeats (between two and seven), whereas extracellular proteins have more repeats (between seven and twelve). Interestingly, there are also a significant number of membrane proteins, extracellular proteins and proteins with unknown subcellular localizations with no assigned Pfam domain, possibly due to the lack of functional studies of these proteins. Together, these results indicate that domain-containing proteins that do not harbor repeating domain are generally cytoplasmic proteins, whereas tandem domain sequences that may form repeating 3D folds is a favored architecture in surface-associated proteins.

3.2. Domain repetition occurs most frequently in S. pneumoniae

The 14 different streptococcal species analyzed in this study showed a varied percentage of proteins with domain repeats (see Table 1). The genomes of *S. thermophilus*, *S. agalactiae* and *S. mutans* have the fewest percentages of proteins containing domain repeats (an average of 3.67%, 3.81% and 3.89% respectively). The sequenced *S. pyogenes* strains generally have low percentages, an average of 3.86%, with four strains that are over 4%. The *S. pneumoniae*, *S. mitis* and *S. gallolyticus* have higher percentages (over 4.5%). Fig. 2 shows the numbers of proteins with domain repeats broken down by repeat count (raw data in Supplementary data 4). *S. pneumoniae* and *S. mitis* have a high number of proteins with five or more repeats than other streptococcal species. Most of these proteins are members of the choline-binding protein (ChBP) family. *S. pyogenes* strains generally have a low number of proteins with repeated domain. However, five of ten *S. pyogenes* serotypes (M1, M4, M12, M28 and M49) have a plasminogen-binding protein Epf with extremely high number of DUF1542 repeats. Fig. 2 also shows *S. pneumoniae*, *S. pyogenes* and *S. sanguinis* are the only analyzed species that have proteins containing more than ten repeats. The proteins that have ten or more repeats are summarized in Table 2.

3.3. Top five Pfam domains with most numbers of tandem repeats

Five Pfam domains were found in streptococcal proteins that have 10 or more domain repetitions (see Table 2). The distributions of the number of repetitions of these five domains in the analyzed proteins

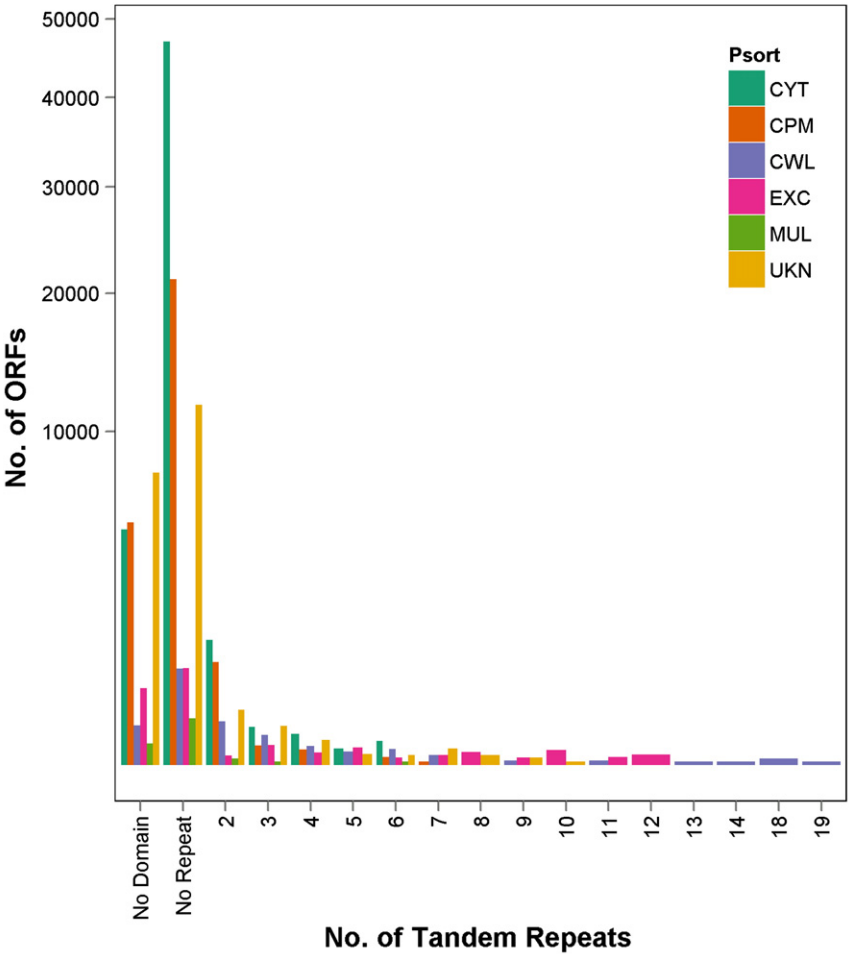

Fig. 1. Distributions of proteins with varying numbers of tandem repeats in different subcellular localizations. The six subcellular locations are cytoplasmic (CYT), cytoplasmic membrane (CPM), cell wall (CWL), extracellular (EXC), multiple sites (MUL) and unknown location (UKN).

are presented in Fig. 3. Detailed results of individual domains are presented in the following paragraphs.

3.3.1. DUF1542 (PF07564)

The DUF1542 is the most repeated domain identified in *Streptococcus*. Proteins that contain repeated DUF1542 are presented in Supplementary data 5. This domain is repeated 18 to 19 times in the virulence-associated extracellular protein factor (EF or Epf) of *S. pyogenes* M1, M12 and M28 serotypes. The Epf proteins in *S. pyogenes* M4 and M49 have eleven and seven repeats respectively. The remaining *S. pyogenes* serotypes (M2, M3, M5, M6 and M18) do not have Epf or other proteins with this domain. The DUF1542 domain is known to present in cell surface proteins that are involved in antibiotic resistance [30] and/or cellular adhesion [31]. The purified Epf protein from *S. pyogenes* serotype M49 can bind to immobilized plasminogen [32]. In the highly virulent *S. suis* 98HAH33 and 05ZYH33 isolates, the EF proteins also contain the DUF1542 domain. The domain is repeated two times; therefore the deduced peptide length of EF is 154 aa, which is much shorter than the 2107 aa-long Epf protein encoded by M28_Spy0539 from *S. pyogenes* MGAS6180.

3.3.2. Rib (PF08428)

The Rib domain is named after the Rib protein of *S. agalactiae*, a surface protein that confers protective immunity [33]. The α, Rib, Alp2 and R28/Alp3 proteins are members of the Alp protein family. The Rib protein is found in *S. agalactiae* serotype V, α protein in serotype Ia, Alp2 protein in serotype III, and R28 protein in *S. pyogenes* serotypes M2 and M28 (see Supplementary data 6). The gene encoding R28 is located within the RD2 of *S. pyogenes* MGAS6180 (M28) genome, a region that has been suggested to be involved in the pathogenesis and adaptation to specific environments, and with the subset of *S. pyogenes* strains epidemiologically associated with maternal–fetal infections [34]. The same inserted region in M28, flanked by MGAS10270_Spy1378 and MGAS10270_Spy1411, can also be found in *S. pyogenes* MGAS10270 (M2). The *S. pyogenes* M2 and M28 most likely obtained the RD2 through horizontal gene transfer from *S. agalactiae* [34,35]. In *S. pyogenes*, R28 is a known virulence factor that increases attachment of *S. pyogenes* to human cervical cells [35].

3.3.3. G5 (PF07501)

The G5 domain was first recognized in 2004 as a potential N-acetylglucosamine recognition domain that may be involved in biofilm formation [168]. G5 domain is widely distributed in Gram-positive bacteria and all G5-containing streptococcal protein (in singular or repeats) are predicted to localize in CWL. Based on the annotations of 35 tandem G5 domain-containing proteins, the protein can be broadly categorized into three main functions: glycosidases, zinc metalloproteinases and cell surface proteins with no known function (see Supplementary data 7). The G5 domains are generally found in the N-terminus of zinc metalloproteinase such as IgA proteases, ZmpB and ZmpC zinc metalloproteinases. The ZmpC proteins of *S. sanguinis*, *S. gordonii* and *S. suis* are high G5 repeat-containing, whereas the ZmpB proteins and IgA proteases of *S. sanguinis*, *S. gordonii* and *S. mitis* have only one to two G5 repeats. Besides zinc metalloproteinases, the *S. gordonii* cell surface protein SGO_2004 also contains many G5 repeats. The domain organization of SGO_2004

Table 1
Summary of repeated Pfam domain distribution in streptococcal genomes.

Organism	%RepPro[a]	%RepDom[b]	Most Common Pfam ID		
Streptococcus agalactiae 2603V/R	3.72	6.86	CBS (7)	DUF161 (3)	Alpha-amylase (3)
Streptococcus agalactiae A909	3.99	7.21	CBS (7)	Alpha-amylase (3)	PRD (2)
Streptococcus agalactiae NEM316	3.73	6.55	CBS (7)	PRD (3)	DUF161 (3)
Streptococcus dysgalactiae subsp. equisimilis GGS_124	4.09	7.06	CBS (6)	PRD (4)	DUF161 (4)
Streptococcus equi subsp. equi 4047	4.09	7.28	Collagen (9)	CBS (4)	DUF161 (4)
Streptococcus equi subsp. zooepidemicus H70	4.94	8.61	Collagen (10)	PRD (4)	CBS (4)
Streptococcus equi subsp. zooepidemicus MGCS10565	4.90	8.53	Collagen (9)	PRD (4)	CBS (4)
Streptococcus gallolyticus ATCC 43143	5.06	9.32	PRD (8)	GBS_Bsp-like (6)	MatE (6)
Streptococcus gallolyticus ATCC BAA-2069	4.86	8.79	PRD (8)	CBS (5)	GBS_Bsp-like (5)
Streptococcus gallolyticus UCN34	4.98	8.72	PRD (8)	MatE (6)	CBS (5)
Streptococcus gordonii str. Challis substr. CH1	4.27	8.42	CBS (6)	PRD (4)	CW_binding_1 (4)
Streptococcus mitis B6	5.18	10.99	CW_binding_1 (16)	CBS (5)	EamA (4)
Streptococcus mutans NN2025	3.90	7.00	CBS (6)	CW_binding_1 (4)	PRD (3)
Streptococcus mutans UA159	3.88	7.00	CBS (6)	CW_binding_1 (4)	PRD (3)
Streptococcus pasteurianus ATCC 43144	4.23	6.92	PRD (8)	CBS (5)	DUF161 (3)
Streptococcus pneumoniae 670-6B	5.19	10.89	CW_binding_1 (14)	PRD (5)	CBS (5)
Streptococcus pneumoniae 70585	5.11	10.93	CW_binding_1 (13)	PRD (5)	CBS (5)
Streptococcus pneumoniae AP200	4.62	9.85	CW_binding_1 (13)	CBS (5)	EamA (3)
Streptococcus pneumoniae ATCC 700669	5.09	10.18	CW_binding_1 (11)	CBS (5)	PhoU (4)
Streptococcus pneumoniae CGSP14	5.17	10.94	CW_binding_1 (15)	CBS (5)	PRD (4)
Streptococcus pneumoniae D39	5.27	11.12	CW_binding_1 (12)	CBS (5)	PRD (4)
Streptococcus pneumoniae G54	5.03	10.24	CW_binding_1 (13)	CBS (5)	Strep_his_triad (4)
Streptococcus pneumoniae Hungary19A-6	5.06	10.71	CW_binding_1 (13)	CBS (5)	PRD (4)
Streptococcus pneumoniae JJA	5.04	10.27	CW_binding_1 (15)	CBS (5)	EamA (3)
Streptococcus pneumoniae P1031	4.92	10.02	CW_binding_1 (10)	CBS (5)	Strep_his_triad (5)
Streptococcus pneumoniae R6	5.12	10.87	CW_binding_1 (12)	CBS (5)	Strep_his_triad (5)
Streptococcus pneumoniae Taiwan19F-14	5.33	11.01	CW_binding_1 (12)	PRD (5)	CBS (5)
Streptococcus pneumoniae TCH8431/19A	5.03	10.54	CW_binding_1 (12)	CBS (5)	PRD (4)
Streptococcus pneumoniae TIGR4	5.29	11.01	CW_binding_1 (14)	PRD (5)	CBS (5)
Streptococcus pyogenes M1 GAS SF370	3.84	7.21	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes Manfredo	3.76	6.35	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS10270	3.87	6.68	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS10394	3.66	6.40	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS10750	4.09	7.26	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS2096	3.93	7.71	CBS (4)	DUF161 (4)	Fn_bind (4)
Streptococcus pyogenes MGAS315	3.63	6.28	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS5005	3.77	7.00	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS6180	4.20	8.54	CBS (4)	DUF161 (4)	Fn_bind (4)
Streptococcus pyogenes MGAS8232	3.71	6.30	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes MGAS9429	4.01	7.78	CBS (4)	DUF161 (4)	Fn_bind (4)
Streptococcus pyogenes NZ131	4.12	7.41	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus pyogenes SSI-1	3.64	6.27	CBS (4)	PRD (3)	DUF161 (3)
Streptococcus sanguinis SK36	4.07	8.18	CBS (7)	PRD (3)	Cna_B (3)
Streptococcus suis 05ZYH33	4.13	7.14	PRD (4)	CBS (4)	Methylase_S (4)
Streptococcus suis 98HAH33	3.91	6.83	CBS (4)	DUF161 (4)	PRD (3)
Streptococcus suis BM407	3.96	6.79	PRD (4)	CBS (4)	DUF161 (3)
Streptococcus suis P1/7	4.27	7.32	PRD (4)	CBS (4)	DUF161 (3)
Streptococcus suis SC84	4.12	7.06	PRD (4)	CBS (4)	DUF161 (3)
Streptococcus thermophilus CNRZ1066	3.51	5.93	CBS (5)	EamA (3)	DUF161 (2)
Streptococcus thermophilus LMD-9	3.85	6.32	CBS (4)	HTH_Tnp_1 (4)	EamA (3)
Streptococcus thermophilus LMG 18311	3.66	6.04	CBS (5)	HTH_Tnp_1 (5)	Hexapep (4)
Streptococcus uberis 0140J	4.11	6.73	CBS (5)	PRD (4)	DUF161 (4)

[a] Percentage of Pfam domain-containing proteins that have domain repeats.
[b] Percentage of Pfam domains that is repeated.

resembles the *Staphylococcus aureus* SasG and Aap cell surface proteins. In *S. aureus*, the tandem G5 domains-containing Aap and SasG proteins have a functional role in biofilm formation; in addition, SasG has a masking effect on the adherence of *S. aureus* to the immobilized ligands [36,37].

3.3.4. HemolysinCabind (PF00353)

The SSA_1099 of *S. sanguinis* is the only streptococcal protein that contains HemolysinCabind domain. The protein has ten HemolysinCabind repeats, and the predicted product is the repeats-in-toxin (RTX) hemolysin HlyA. Tandem repeats of this domain are known to present in the RTX exoproteins of Gram-negative bacteria. In *Escherichia coli*, the *hlyCABD* operon encodes the toxin activation protein (HlyC), the RTX hemolysin (HlyA), the ABC transporter (HlyB) and the membrane fusion protein (HlyD) to allow the secretion of the cytolytic toxin via type I secretion system [38]. The immediate downstream of SSA_1099 are two genes (SSA_1100 and SSA_1101) predicted to encode the HlyB and HlyD proteins. A recent review stated that there are only a handful of Gram-positive bacteria that were predicted to harbor the predicted RTX proteins and *S. sanguinis* SK36 is the only genome to have a complete *hlyABD* locus [39]. It is unclear how *S. sanguinis* SK36 came to possess this set of genes, also, without encoding the toxin activation protein HlyC, the activity of HlyA and its implication in *S. sanguinis* pathogenesis remain to be addressed in future experiments.

3.3.5. CW_binding_1 (PF01473)

As seen in Fig. 3, the cell wall binding repeat (CWBR) is the most abundant domain repeat in streptococcal proteins (n = 214). Twenty-two genomes analyzed in this study have proteins that contain this domain repeat, and are especially rich in *S. pneumoniae* and the closely related *S. mitis* (n = 195, average 13 copies per genome) (see Supplementary data 8). In *S. pneumoniae* and *S. mitis*, proteins

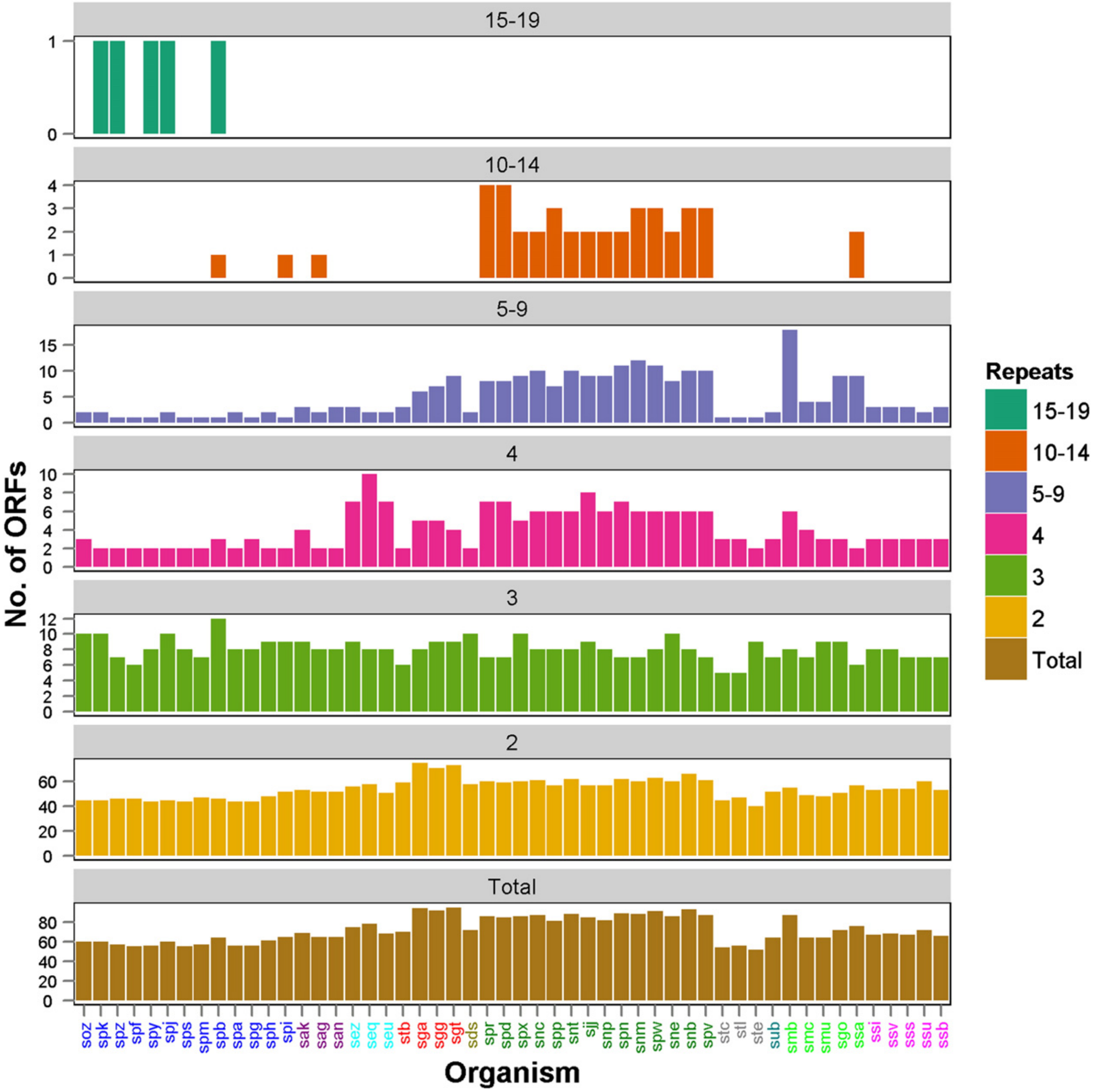

Fig. 2. Distribution of repeated domains of different tandem repetitions in streptococci. The KEGG organism codes were used to label the x-axis and the full bacterial names can be found in Supplementary data 1.

with tandem CWBRs are known as choline-binding proteins (ChBPs). The domain repeats are high in glycine and aromatic residues that allow recognition and interaction of phosphorylcholine residues on the cell surface. The endo-beta-N-acetylglucosaminidases (LytB) from different strains of *S. pneumoniae* have CWBRs ranging from eight to twelve. This enzyme is responsible for daughter cell separation [40]. Other members of the choline-binding protein family, harbor two to twelve CWBRs, have also been implicated in virulence. Autolysin (LytA) was implicated in blocking phagocytosis of live pneumococci and preventing the triggering of phagocyte-activating cytokines [41]. The pneumococcal surface protein A (PspA) inhibits complement deposition and complement activation [42]. Choline-binding protein A (CbpA/PspC) is an adhesin that binds secretory IgA and complement activation regulatory protein factor H [43]. Mutagenesis studies of *cbpD* and *cbpE* showed major loss of nasopharyngeal colonization in animal models [44]. The multifunctional CbpG is a serine protease that can cleave host extracellular matrix and also participates in bacterial adherence [45]. In oral streptococci and *S. gallolyticus*, the CWBRs are found in cell-surface associated glucosyltransferases and glucan-binding proteins (see Supplementary data 9). The CWBRs in these proteins are spaced more widely apart than the CWBRs in choline-binding proteins. The aromatic residues in the repeat units are believed to be important for the protein–glucose unit interaction [6]. Overall, the CWBRs function as an interacting domain for the repeat-containing proteins. The varying spacing between the repeats may generate slight different folding and thus form binding pockets that interact with different molecules, i.e. choline and glucose.

3.4. Tertiary structural modeling of domain repeats

In order to understand the influence of the domain repeats on the protein structure, a list of proteins containing the five most repeated domains were used in structural modeling. Using computational modeling with I-TASSER, that utilizes a combination of comparative modeling, threading, and *ab initio* modeling, it is possible to generate reliable full-length 3D protein structural predictions when little or no

Table 2
List of streptococcal ORFs that have high numbers of repeated Pfam domain (≥10).

Organism	Serotype	Locus ID	Pfam domain name	No. of repeats
Streptococcus agalactiae 2603V/R	V	SAG0433	Rib	14
Streptococcus pneumoniae P1031	1	SPP_0185	CW_binding_1	10
Streptococcus pneumoniae P1031	1	SPP_2242	CW_binding_1	10
Streptococcus pneumoniae P1031	1	SPP_0971	CW_binding_1	12
Streptococcus pneumoniae D39	2	SPD_0126	CW_binding_1	10
Streptococcus pneumoniae D39	2	SPD_1965	CW_binding_1	10
Streptococcus pneumoniae D39	2	SPD_2017	CW_binding_1	10
Streptococcus pneumoniae D39	2	SPD_0853	CW_binding_1	12
Streptococcus pneumoniae R6	2	spr0121	CW_binding_1	10
Streptococcus pneumoniae R6	2	spr1945	CW_binding_1	10
Streptococcus pneumoniae R6	2	spr1995	CW_binding_1	10
Streptococcus pneumoniae R6	2	spr0867	CW_binding_1	12
Streptococcus pneumoniae TIGR4	4	SP_0117	CW_binding_1	10
Streptococcus pneumoniae TIGR4	4	SP_0965	CW_binding_1	10
Streptococcus pneumoniae 70585	5	SP70585_2316	CW_binding_1	10
Streptococcus pneumoniae 70585	5	SP70585_0197	CW_binding_1	11
Streptococcus pneumoniae 70585	5	SP70585_1005	CW_binding_1	11
Streptococcus pneumoniae CGSP14	14	SPCG_0120	CW_binding_1	10
Streptococcus pneumoniae CGSP14	14	SPCG_0941	CW_binding_1	10
Streptococcus pneumoniae CGSP14	14	SPCG_2158	CW_binding_1	11
Streptococcus pneumoniae JJA	14	SPJ_0906	CW_binding_1	10
Streptococcus pneumoniae JJA	14	SPJ_2217	CW_binding_1	10
Streptococcus pneumoniae AP200	11A	SPAP_0173	CW_binding_1	12
Streptococcus pneumoniae AP200	11A	SPAP_2234	CW_binding_1	12
Streptococcus pneumoniae Hungary19A-6	19A	SPH_2388	CW_binding_1	10
Streptococcus pneumoniae Hungary19A-6	19A	SPH_0232	CW_binding_1	11
Streptococcus pneumoniae Hungary19A-6	19A	SPH_1067	CW_binding_1	12
Streptococcus pneumoniae TCH8431/19A	19A	HMPREF0837_10423	CW_binding_1	10
Streptococcus pneumoniae TCH8431/19A	19A	HMPREF0837_11523	CW_binding_1	12
Streptococcus pneumoniae G54	19F	SPG_0121	CW_binding_1	11
Streptococcus pneumoniae G54	19F	SPG_0889	CW_binding_1	12
Streptococcus pneumoniae Taiwan19F-14	19F	SPT_0163	CW_binding_1	10
Streptococcus pneumoniae Taiwan19F-14	19F	SPT_1238	CW_binding_1	12
Streptococcus pneumoniae ATCC 700669	23F	SPN23F_08900	CW_binding_1	10
Streptococcus pneumoniae ATCC 700669	23F	SPN23F_22240	CW_binding_1	10
Streptococcus pneumoniae 670-6B	6B	SP670_2336	CW_binding_1	10
Streptococcus pneumoniae 670-6B	6B	SP670_0200	CW_binding_1	11
Streptococcus pneumoniae 670-6B	6B	SP670_1354	CW_binding_1	12
Streptococcus pyogenes M1 GAS SF370	M1	SPy_0737	DUF1542	18
Streptococcus pyogenes MGAS5005	M1	M5005_Spy_0561	DUF1542	18
Streptococcus pyogenes MGAS2096	M12	MGAS2096_Spy0622	DUF1542	18
Streptococcus pyogenes MGAS9429	M12	MGAS9429_Spy0613	DUF1542	18
Streptococcus pyogenes MGAS6180	M28	M28_Spy1336	Rib	13
Streptococcus pyogenes MGAS6180	M28	M28_Spy0539	DUF1542	19
Streptococcus pyogenes MGAS10750	M4	MGAS10750_Spy0643	DUF1542	11
Streptococcus sanguinis SK36	–	SSA_1099	HemolysinCabind	10
Streptococcus sanguinis SK36	–	SSA_1018	G5	11

primary sequence similarity to a protein of known structure can be found. The results of which are summarized in Supplementary data 10. The C-score is a confidence score for estimating the quality of predicted models by I-TASSER. It is typically in the range of −5 to 2, where a C-score of higher value indicates a model with a high confidence and vice-versa. The template modeling score (TM-score) and root mean square deviation (RMSD) value of the predicted model relative to the native structure are predicted based on the C-score. A TM-score more than 0.5 indicates a model of correct topology and a TM-score less than 0.17 means random similarity. Using these parameters, the 24 predicted models have C-scores and TM-scores in the range of −3.44 to 0.11 and 0.35 to 0.73 respectively. Eight models have TM-scores below 0.5 and C-scores below −1.8, but are still within the limit of accuracy. Among those submitted are the ChBPs of *S. pneumoniae* TIGR4 that have five or more CW_binding_1 domain repeats to represent the diverse distribution of repetition patterns. The predicted models of ChBPs showed evidence of modular organization (see Fig. 4). The repeating CW_binding_1 domains form the C-terminal choline-binding domain whereas the N-terminus forms a biologically active domain, with the exception of SP_0965 (LytB) and SP_1573 (LytC) that have reversed arrangement. Repeating DUF1542, Rib, G5 and HemolysinCabind domains showed a less conspicuous modular design (see Fig. 5). In Epf and EF proteins that are almost solely made up by α-helix-rich DUF1542 domains, two folding conformations were predicted. The MGAS2096_Spy0622 has 18 DUF1542 repeats, and was predicted to fold into a tight "bundle of bundles" that may mediate protein–protein interaction at cell–cell and/or cell–matrix junction (see Fig. 5A). On the other hand, Epf proteins that have less repeats were predicted to form a ring-like molecule (see Figs. 5B,C). This structure resembles the eukaryotic karyopherins that are involved in the transport of macromolecules across the nuclear envelope. The predicted model of *S. suis* EF protein was shown in Fig. 5D. It contains two DUF1542 repeats that formed a bundle of five helices. This structure appears to be the basic building block for the larger Epf proteins. The G5 and Rib domains are both rich in β-strands. I-TASSER have modeled G5 and Rib domain-containing proteins into structures that contain several immunoglobulin (Ig)-like fold domains forming a flat and large "valley" and produced a possible binding cavity (see Figs. 5E–K). The C-terminal domain is rich in α-helices and loops and likely formed another binding area. The SSA_1099 is the only streptococcal protein to contain the HemolysinCabind domain. The predicted model, based on the second half of the protein that contains the tandem repeats, showed the formation of a β-roll sandwich that has two parallel β-roll domains stacked

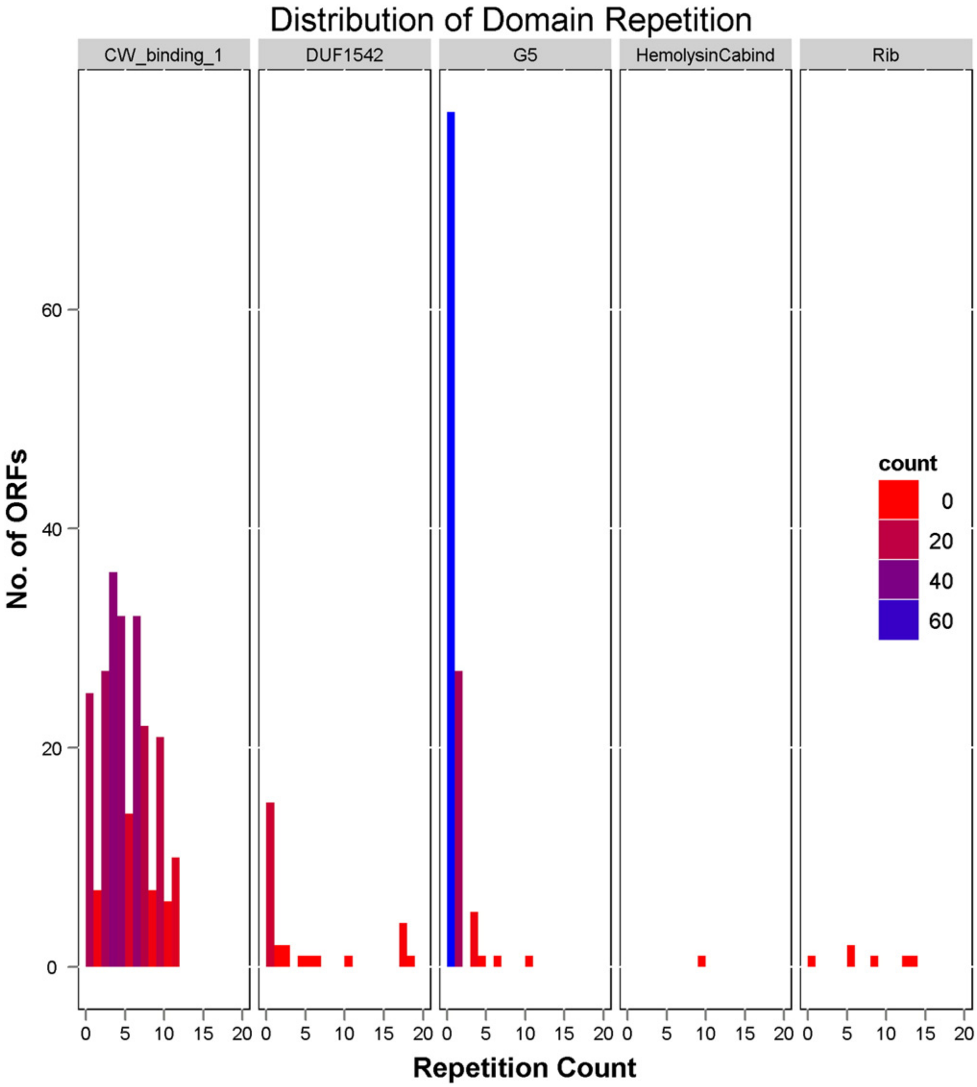

Fig. 3. Distribution of the top five repeated domains in streptococci. The distribution of CW_binding_1, DUF1542, G5, HemolysinCabind and Rib were individually presented in separate panels.

closely together (see Fig. 5L). The β-roll folds form the calcium-binding motif. This motif is commonly found in lipases, hemolysins and some proteases.

4. Discussion

This study focused on the discovery of proteins containing domain repeats and highly repeated protein domains in *Streptococcus*. The percentages of proteins containing domain repeats in *Streptococcus* ranged from 3.4% to 5.3%, which is slightly lower than the 5% calculated from seven bacterial and seven archaeal proteomes by Björklund and coworkers [2]. It has been suggested that repeats are evolutionary events because eukaryotic genomes encode far more proteins with domain repeats than prokaryotic and archaeal genomes. The functions of these eukaryotic domain repeat-containing proteins have little similarity to those of the microbial proteins. We showed streptococcal domain repeat-containing proteins are significantly enriched in extracytoplasmic locations further supports the notion of adaptive evolution. In response to the dynamic interactions with the habitat and host immune and defense systems, bacterial cell surface proteins are often under stronger selection pressures than proteins that resides in cytoplasm that perform core functions such as metabolism, DNA replication, transcription and translation [46–49]. Many of the surface-associated proteins are experimentally verified bona fide virulence-associated factors, such as ChBPs and zinc metalloproteases of *S. pneumoniae*, glucan-binding proteins and glucosyltransferases of oral streptococci, Alp proteins of *S. agalactiae* and extracellular factor of *S. suis*. The repeating protein domains can form functionally important structures, therefore the extracytoplasmic subcellular localization and the presence of domain repeats of a protein may be a strong indicator of involvement in virulence. Understanding the functions of hypothetical proteins with these two properties could assist in our understanding of pathogenesis of bacterial pathogens.

Unlike simple mutations that involves single or short nucleotide and/or amino acid changes, contraction and expansion of repeating unit in an array of domain repeats will bring about structural changes. The changes may be subtle, like those observed in the choline binding domains of the *S. pneumoniae* ChBPs. In the case of large domain such as DUF1542 that has an average of 66 amino acids, we showed the *S. pyogenes* Epf protein that contains many repeats has a different predicted 3D structure than that has fewer repeats (see Figs. 5A,B). The Epf proteins are made up of variable numbers of DUF1542 domain. The Epf proteins from different strains are considered orthologs based on standard sequence analysis. However, not all Epfs are structural homologs and thus likely to have different binding targets. Hence, when automated sequence-based functional annotation was

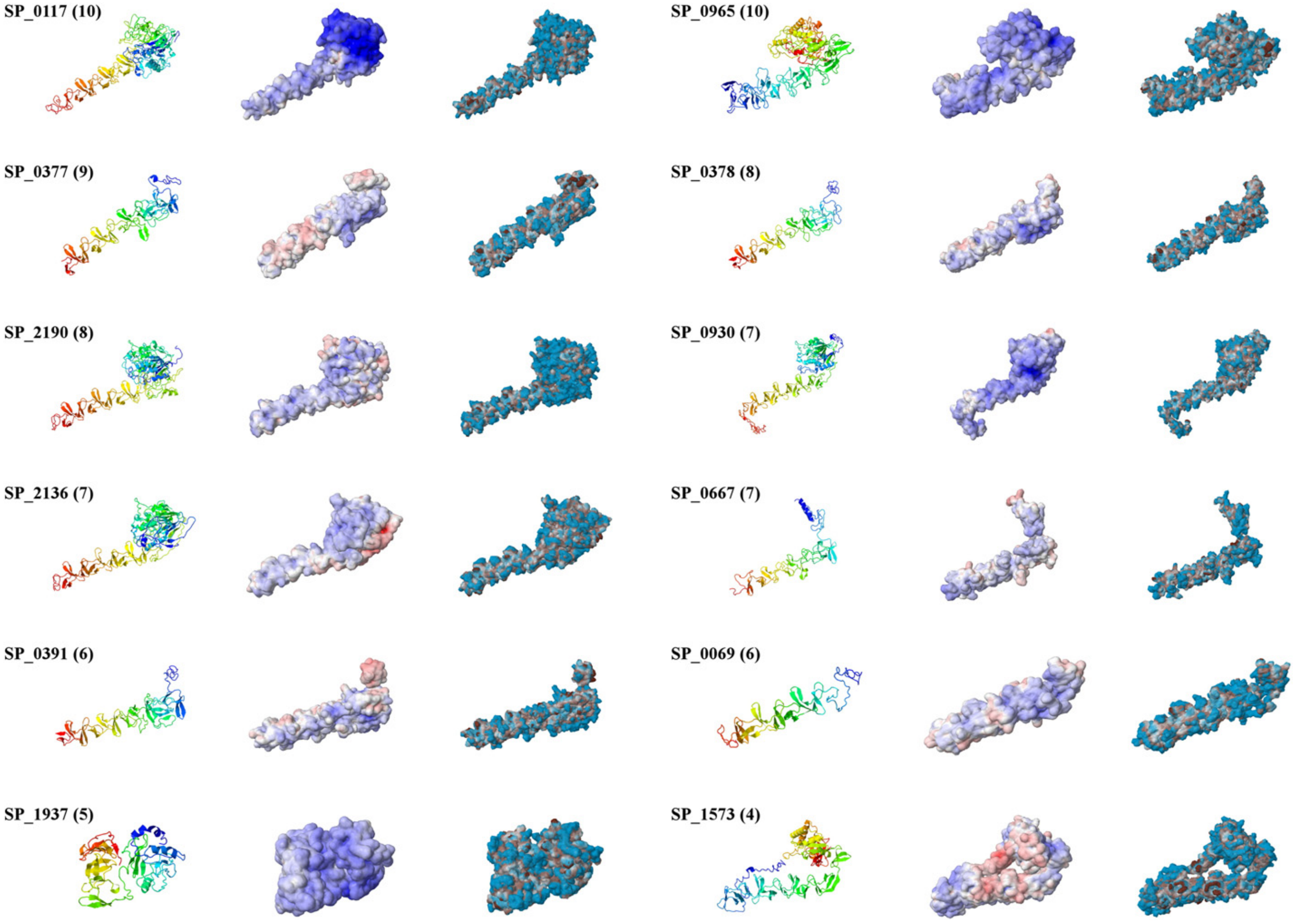

Fig. 4. Structural predictions by I-TASSER for the various *S. pneumoniae* choline binding proteins. The number of CW_binding_1 domain repeats present in each protein was provided in round brackets. For each protein, three structural representations were provided. On the left is the protein depicted as a continuous rainbow colored cartoon (N-terminus = blue and C-terminus = red). The electrostatic surface potential of the protein is provided in the middle (negative potential = blue; positive potential = red; contoured from −15 to 15 kT/e). On the right is the hydrophobic and hydrophilic properties of the protein (hydrophilic = blue and hydrophobic = brown).

employed on predicted genes, additional attention should be taken to prevent misannotation that may compromise future functional studies and interpretation of results.

Besides affecting the protein structure, changes in the number of repeats will also bring about alteration in conformational epitopes of antigens. This in turns elicits different repertoires of antibodies. Earlier studies showed streptococcal proteins with variable number of repeats were found to exhibit antigenic variations [11]. In the case of members of the Alp family, the C-terminal hydrophobic membrane anchor domains help embedding the proteins in the bacterial cell wall. The extracellular exposed Rib domain repeating regions were predicted to have the Ig-like fold and thus likely to have a function in molecular recognition. The R28 protein of *S. pyogenes* M28, a member of the Alp family, is nearly double in length and has twice the amount of Rib domain repeats than the R28 protein of *S. pyogenes* M2 (see Figs. 5H,J). There are also significant differences in the electrostatic surface potential and the electrostatic hot spots of both R28 variants, which may have affected the recognition and binding affinity of the protein. The contraction and expansion of repeats have been shown to affect the binding affinity of many cell surface carbohydrate-binding proteins [6,7,50]. Sequence analysis showed choline binding protein A, choline binding protein C, PcpA, LytB and pneumococcal surface protein A of the pneumococci ChBPs exhibit noticeable polymorphism in the number of CW_binding_1 repeats in different *S. pneumoniae* strains (see Supplementary data 8). On the other hand, the remaining ChBPs, such as LytA and LytC, have a conserved number of CW_binding_1 repeats. As both groups of ChBPs include well-characterized major pneumococcal virulence factors, the rationales and the effect on bacterial fitness behind having polymorphic repeats in certain ChBPs are not known at present.

This study identified the statistical correlation between protein subcellular localization and tandem repetition of protein domain in 52 streptococcal genomes. The findings reported in this study foster the understanding of the effect of domain repeats on protein structure and function. While crystallization of most of the domain repeats mentioned in this study is unavailable, *ab initio* modeling was employed to predict structural models of several repeat-containing proteins, some for the first time. This provides new insight to these proteins and hopefully inspires further research to understand the structural and functional dynamics of these repeats and proteins. Using the same strategy, it will be interesting to determine the proteins containing repeats and the types of domain repeats in other pathogenic bacteria, and examine the contributions of the repeats to protein function. The modular organization of domain repeats, for example the choline-binding module in ChBPs, is also a useful feature in protein design and synthetic biology.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.08.001.

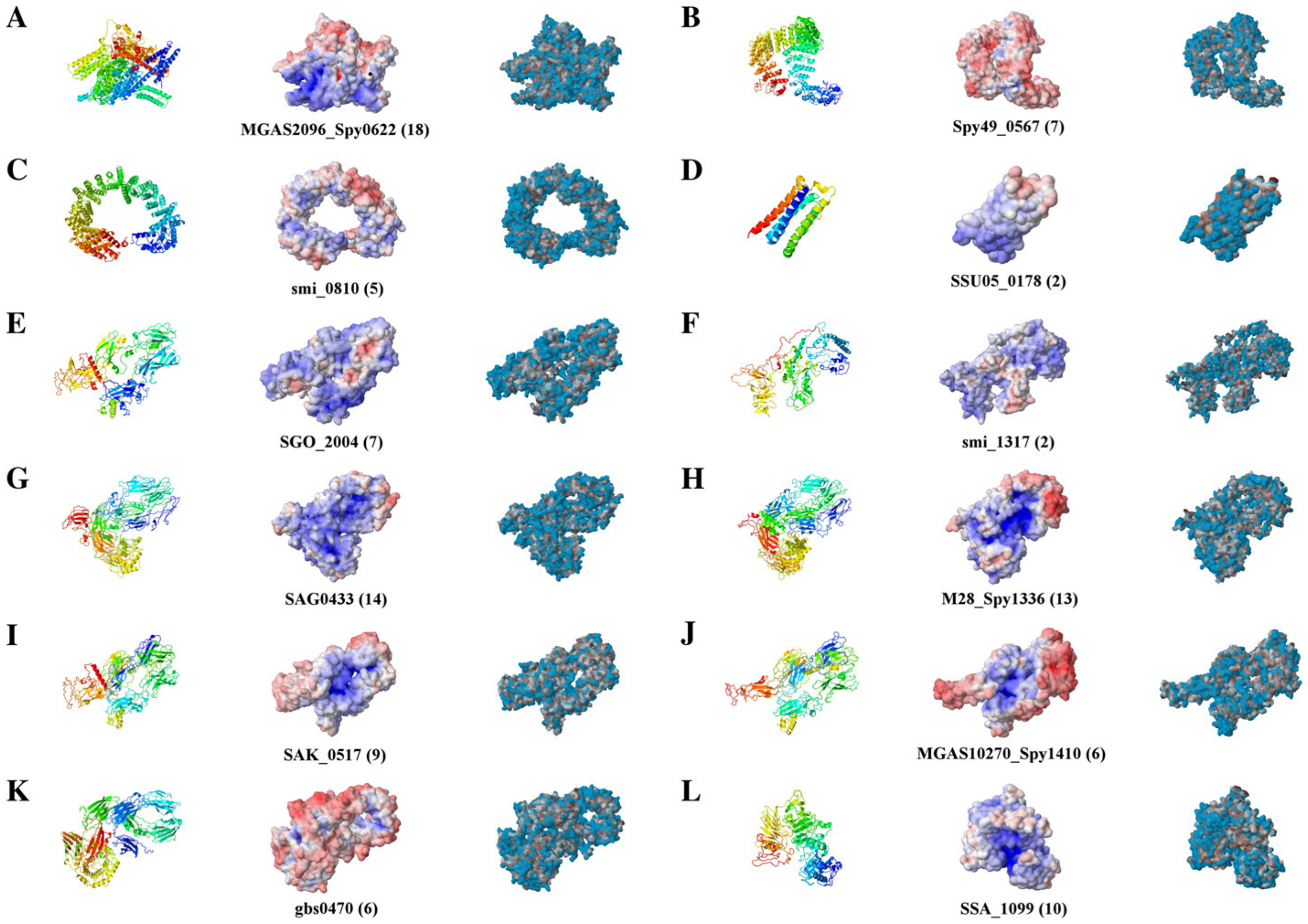

Fig. 5. Structural predictions by I-TASSER for the DUF1542, G5, Rib and HemolysinCabind domain-containing proteins. A to D are DUF1542-containing proteins, E and F are G5-containing proteins, G to K are Rib-containing proteins and L is the HemolysinCabind domain-containing protein. The number of domain repeats in each protein was provided in round brackets. The three structural representations are the same as Fig. 4, except the electrostatic surface potential of Rib- and HemolysinCabind-containing proteins were contoured from −25 to 5 kT/e.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IHL, MTH and CHC conceived and designed the study. IHL performed the analysis. IHL wrote the original manuscript. MTH and CHC contributed to revisions of the manuscript. All authors read and approved the final version.

Acknowledgments

This study was supported by the National Science Council, Taiwan (NSC 100-2319-B-010-002) and the Ministry of Education, Taiwan, Aim for the Top University Plan.

References

[1] G. Apic, J. Gough, S.A. Teichmann, Domain combinations in archaeal, eubacterial and eukaryotic proteomes, J. Mol. Biol. 310 (2001) 311–325.
[2] A.K. Bjorklund, D. Ekman, A. Elofsson, Expansion of protein domain repeats, PLoS Comput. Biol. 2 (2006) e114.
[3] D. Ekman, A.K. Bjorklund, J. Frey-Skott, A. Elofsson, Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions, J. Mol. Biol. 348 (2005) 231–243.
[4] M.W. van der Woude, A.J. Baumler, Phase and antigenic variation in bacteria, Clin. Microbiol. Rev. 17 (2004) 581–611.
[5] C.S. Lachenauer, R. Creti, J.L. Michel, L.C. Madoff, Mosaicism in the alpha-like protein genes of group B streptococci, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 9630–9635.
[6] D.S. Shah, G. Joucla, M. Remaud-Simeon, R.R. Russell, Conserved repeat motifs and glucan binding by glucansucrases of oral streptococci and *Leuconostoc mesenteroides*, J. Bacteriol. 186 (2004) 8301–8308.
[7] C. von Eichel-Streiber, M. Sauerborn, H.K. Kuramitsu, Evidence for a modular structure of the homologous repetitive C-terminal carbohydrate-binding sites of *Clostridium difficile* toxins and *Streptococcus mutans* glucosyltransferases, J. Bacteriol. 174 (1992) 6707–6710.
[8] R. Tiyawisutsri, et al., Burkholderia Hep_Hag autotransporter (BuHA) proteins elicit a strong antibody response during experimental glanders but not human melioidosis, BMC Microbiol. 7 (2007) 19.
[9] Q. Zhang, K.S. Wise, Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent vaa genes, Infect. Immun. 64 (1996) 2737–2744.
[10] A.J. Sheets, J.W. St Geme III, Adhesive activity of the haemophilus cryptic genospecies cha autotransporter is modulated by variation in tandem peptide repeats, J. Bacteriol. 193 (2011) 329–339.
[11] C. Gravekamp, D.S. Horensky, J.L. Michel, L.C. Madoff, Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes, Infect. Immun. 64 (1996) 3576–3583.
[12] M. Hijnen, F.R. Mooi, P.G. van Gageldonk, P. Hoogerhout, A.J. King, G.A. Berbers, Epitope structure of the *Bordetella pertussis* protein P.69 pertactin, a major vaccine component and protective antigen, Infect. Immun. 72 (2004) 3716–3723.
[13] Y.R. Ho, et al., Variation in the number of tandem repeats and profile of surface protein genes among invasive group B Streptococci correlates with patient age, J. Clin. Microbiol. 45 (2007) 1634–1636.
[14] M. Moschioni, W. Pansegrau, M.A. Barocchi, Adhesion determinants of the Streptococcus species, Microb. Biotechnol. 3 (2010) 370–388.
[15] A.H. Nobbs, R.J. Lamont, H.F. Jenkinson, Streptococcus adherence and colonization, Microbiol. Mol. Biol. Rev. 73 (2009) 407–450.

[16] R. Hakenbeck, A. Madhour, D. Denapaite, R. Bruckner, Versatility of choline metabolism and choline-binding proteins in Streptococcus pneumoniae and commensal streptococci, FEMS Microbiol. Rev. 33 (2009) 572–586.

[17] M.J. Jedrzejas, Unveiling molecular mechanisms of bacterial surface proteins: *Streptococcus pneumoniae* as a model organism for structural studies, Cell. Mol. Life Sci. 64 (2007) 2799–2822.

[18] C.G. Baums, P. Valentin-Weigand, Surface-associated and secreted factors of *Streptococcus suis* in epidemiology, pathogenesis and vaccine development, Anim. Health Res. Rev. 10 (2009) 65–83.

[19] G. Lindahl, M. Stalhammar-Carlemalm, T. Areschoug, Surface proteins of *Streptococcus agalactiae* and related proteins in other bacterial pathogens, Clin. Microbiol. Rev. 18 (2005) 102–127.

[20] T. Areschoug, F. Carlsson, M. Stalhammar-Carlemalm, G. Lindahl, Host-pathogen interactions in *Streptococcus pyogenes* infections, with special reference to puerperal fever and a comment on vaccine development, Vaccine 22 (Suppl. 1) (2004) S9–S14.

[21] R.D. Finn, et al., The Pfam protein families database, Nucleic Acids Res. 38 (2010) D211–D222.

[22] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.

[23] N.Y. Yu, et al., PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26 (2010) 1608–1615.

[24] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinforma. 9 (2008) 40.

[25] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283–291.

[26] T.J. Dolinsky, et al., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations, Nucleic Acids Res. 35 (2007) W522–W525.

[27] N.A. Baker, D. Sept, S. Joseph, M.J. Holst, J.A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 10037–10041.

[28] E.L. Willighagen, M. Howard, Fast and Scriptable Molecular Graphics in Web Browsers without Java3D, Nat. Precedings (2007).

[29] T.V. Pyrkov, A.O. Chugunov, N.A. Krylov, D.E. Nolde, R.G. Efremov, PLATINUM: a web tool for analysis of hydrophobic/hydrophilic organization of biomolecular complexes, Bioinformatics 25 (2009) 1201–1202.

[30] S.W. Wu, H. De Lencastre, Mrp–a new auxiliary gene essential for optimal expression of methicillin resistance in *Staphylococcus aureus*, Microb. Drug Resist. 5 (1999) 9–18.

[31] S.R. Clarke, L.G. Harris, R.G. Richards, S.J. Foster, Analysis of Ebh, a 1.1-megadalton cell wall-associated fibronectin-binding protein of *Staphylococcus aureus*, Infect. Immun. 70 (2002) 6680–6687.

[32] B. Kreikemeyer, et al., The *Streptococcus pyogenes serotype* M49 Nra-Ralp3 transcriptional regulatory network and its control of virulence factor expression from the novel eno ralp3 epf sagA pathogenicity region, Infect. Immun. 75 (2007) 5698–5710.

[33] M. Stalhammar-Carlemalm, L. Stenberg, G. Lindahl, Protein rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections, J. Exp. Med. 177 (1993) 1593–1603.

[34] N.M. Green, et al., Genome sequence of a serotype M28 strain of group a streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity, J. Infect. Dis. 192 (2005) 760–770.

[35] M. Stalhammar-Carlemalm, T. Areschoug, C. Larsson, G. Lindahl, The R28 protein of *Streptococcus pyogenes* is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells, Mol. Microbiol. 33 (1999) 208–219.

[36] D.G. Conrady, C.C. Brescia, K. Horii, A.A. Weiss, D.J. Hassett, A.B. Herr, A zinc-dependent adhesion module is responsible for intercellular adhesion in staphylococcal biofilms, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 19456–19461.

[37] J.A. Geoghegan, et al., Role of surface protein SasG in biofilm formation by *Staphylococcus aureus*, J. Bacteriol. 192 (2010) 5663–5673.

[38] W. Wagner, M. Vogel, W. Goebel, Transport of hemolysin across the outer membrane of *Escherichia coli* requires two functions, J. Bacteriol. 154 (1983) 200–210.

[39] I. Linhartova, et al., RTX proteins: a highly diverse family secreted by a common mechanism, FEMS Microbiol. Rev. 34 (2010) 1076–1112.

[40] P. Garcia, M.P. Gonzalez, E. Garcia, R. Lopez, J.L. Garcia, LytB, a novel pneumococcal murein hydrolase essential for cell separation, Mol. Microbiol. 31 (1999) 1275–1281.

[41] A. Martner, S. Skovbjerg, J.C. Paton, A.E. Wold, Streptococcus pneumoniae autolysis prevents phagocytosis and production of phagocyte-activating cytokines, Infect. Immun. 77 (2009) 3826–3837.

[42] B. Ren, et al., The virulence function of *Streptococcus pneumoniae* surface protein A involves inhibition of complement activation and impairment of complement receptor-mediated protection, J. Immunol. 173 (2004) 7506–7512.

[43] S. Dave, S. Carmicle, S. Hammerschmidt, M.K. Pangburn, L.S. McDaniel, Dual roles of PspC, a surface protein of *Streptococcus pneumoniae*, in binding human secretory IgA and factor H, J. Immunol. 173 (2004) 471–477.

[44] K.K. Gosink, E.R. Mann, C. Guglielmo, E.I. Tuomanen, H.R. Masure, Role of novel choline binding proteins in virulence of *Streptococcus pneumoniae*, Infect. Immun. 68 (2000) 5690–5695.

[45] B. Mann, et al., Multifunctional role of choline binding protein G in pneumococcal pathogenesis, Infect. Immun. 74 (2006) 821–829.

[46] S.L. Chen, et al., Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 5977–5982.

[47] L. Petersen, J.P. Bollback, M. Dimmic, M. Hubisz, R. Nielsen, Genes under positive selection in *Escherichia coli*, Genome Res. 17 (2007) 1336–1343.

[48] A. Muzzi, M. Moschioni, A. Covacci, R. Rappuoli, C. Donati, Pilus operon evolution in *Streptococcus pneumoniae* is driven by positive selection and recombination, PLoS One 3 (2008) e3660.

[49] Z. Xu, H. Chen, R. Zhou, Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae*, BMC Evol. Biol. 11 (2011) 203.

[50] J.G. Ho, A. Greco, M. Rupnik, K.K. Ng, Crystal structure of receptor-binding C-terminal repeats from *Clostridium difficile* toxin A, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 18373–18378.