Research Article

Cysteine-Free Proteins in the Immunobiology of Arthropod-Borne Diseases

J. Santiago Mejia, Erik N. Arthun, and Richard G. Titus

Department of Microbiology, Immunology and Pathology, Colorado State University, Central Receiving, 200 W. Lake Street, Campus Delivery 1619, Fort Collins, CO 80523, USA

Correspondence should be addressed to J. Santiago Mejia, jsmejia@colostate.edu

One approach to identify epitopes that could be used in the design of vaccines to control several arthropod-borne diseases simultaneously is to look for common structural features in the secretome of the pathogens that cause them. Using a novel bioinformatics technique, cysteine-abundance and distribution analysis, we found that many different proteins secreted by several arthropod-borne pathogens, including *Plasmodium falciparum, Borrelia burgdorferi*, and eight species of Proteobacteria, are devoid of cysteine residues. The identification of three cysteine-abundance and distribution patterns in several families of proteins secreted by pathogenic and nonpathogenic Proteobacteria, and not found when the amino acid analyzed was tryptophan, provides evidence of forces restricting the content of cysteine residues in microbial proteins during evolution. We discuss these findings in the context of protein structure and function, antigenicity and immunogenicity, and host-parasite relationships.

1. Introduction

Microbial pathogens transmitted by hematophagous arthropods are responsible for some of the most devastating human diseases and the design of vaccines to control them represents one of the main collective efforts in global health. These arthropod-borne diseases (ABDs) are caused by hundreds of species of viruses, bacteria, protozoa, and metazoan parasites and are transmitted by hematophagous mosquitoes, flies, fleas, lice, biting midges, kissing bugs, ticks, and mites [1]. Some ABDs are emerging, reemerging, or out-of-control as a result of high mobility of vertebrate and invertebrate hosts, environmental degradation, and global warming [2]. Further complicating the bleak scenario that ABDs represent, several of the pathogens that cause them could potentially be used as weapons [3–5]. Vaccination remains one of the most cost-effective interventions in public health but only a handful of effective vaccines are available for the control of ABDs [6, 7]. Central to attempts to control infectious diseases through use of vaccines is the idea that species-specific immunity is the best way to induce safe and effective protection. In light of the monumental task of the development of vaccines

directed to each species of arthropod-borne pathogen and/or their vectors, it is necessary to revisit the idea of cross-reactive immunity as a potential venue to protect humans and animals against several ABDs simultaneously. Thanks to advances in synthetic peptide libraries and phage expression libraries, evidence is accumulating that supports the concept of functional poly-specificity of antibodies [8, 9]. Such poly-specificity could be expected to be more apparent when antibodies interact with protein antigens that, due to the high degree of structural disorder and flexibility defined in its amino acid sequence, display a high tendency to engage in promiscuous interactions with other proteins [10, 11]. Some of these features can be identified in proteins using a variety of predictive algorithms that help in the identification of structural disorder [12, 13] and B-cell epitopes [14]. A complementary approach that could be used to identify such sequences is to scan proteins looking for the absence of structural features known to limit protein flexibility and that induce, through an effect on protein folding, the formation of conformational B-cell epitopes. One such structure, the disulfide bridge, is formed when a protein that has at least 2 cysteine residues is allowed to

fold under the oxidative conditions prevalent in extracellular compartments of tissues. It follows that secreted proteins incapable of forming disulfide bridges because of a lack of cysteine residues represent potential targets of cross-reactive vaccines directed to linear, nonconformational, and flexible epitopes. In order to understand both the biological significance that the absence of cysteine residues in pathogen proteins might have in the immunobiology of ABDs and the potential of cysteine-free proteins (CFPs) as vaccine targets, we conducted a cysteine-abundance and distribution analysis on secreted proteins from ten species of arthropod-borne pathogens that cause serious human pathologies, including malaria, Lyme disease, plague, tularemia, Q fever, rocky mountain-spotted fever, human granulocytic anaplasmosis, human monocytic ehrlichiosis, scrub typhus, and Carrion's disease. This analysis allowed for the identification of three major patterns of cysteine expression in secreted proteins of arthropod-borne pathogens. For comparison purposes, a similar abundance and distribution analysis was conducted for another amino acid residue of low abundance in proteins: tryptophan. The significance of the patterns described is discussed in the context of microbial evolution, host-parasite relationships, and prospects for the development of vaccines directed to control several ABDs simultaneously.

2. Materials and Methods

2.1. Genomes Analyzed. Ten arthropod-borne pathogens were selected for cysteine- and tryptophan-distribution analysis, including five species of Alphaproteobacteria (*Anaplasma phagocytophilum, Bartonella bacilliformis, Ehrlichia chaffeensis, Orientia tsutsugamushi,* and *Rickettsia rickettsii*), three Gammaproteobacteria (*Coxiella burnetii, Francisella tularensis,* and *Yersinia pestis*), one Spirochaetes (*Borrelia burgdorferi*), and one protozoan parasite (*Plasmodium falciparum*). In order to reveal potential links between the expression of secreted cysteine-free proteins (sCFPs) and/or secreted tryptophan-free proteins (sWFPs) with virulence and pathogenicity, the genomes of two nonpathogenic bacteria harbored by arthropods (*Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis* and *Wolbachia* endosymbiont of *Drosophila melanogaster*) were included in the analysis.

2.2. Bioinformatics Tools and Procedures Used. FTP files of the genomes of selected species of arthropod-borne microorganisms were downloaded from NCBI Reference Sequences (RefSeq) and transferred to word processing files for analysis of cysteine- and tryptophan-abundance and distribution patterns. We limited the analysis to proteins of at least 200 amino acid residues in length because cysteine and tryptophan are low-abundance amino acids (average content of 1.38% and 1.09%, resp., in UniProtKB/Swiss-prot protein knowledgebase, release 57.2), and thus the significance of their absence is highly dependent on sequence length. Incomplete sequences or those with sequence ambiguities were not included and care was taken to minimize the degree of redundancy of the list of proteins analyzed by

manually removing exact copies. The selected sequences were then analyzed using the Simple Modular Architecture Research Tool (SMART) [15] to detect the presence of domains and motifs that can be used to assign functionality and to track evolutionary trends via analysis of orthologs. SMART has a built-in SignalP 3.0 predictor that was used to identify proteins secreted through a classical signal-peptide-dependent pathway [16], which allowed for the identification of a subset of proteins with cysteine or tryptophan residues confined to the predicted signal peptide segment.

3. Results

3.1. Frequency of CFPs and WFPs in the Genomes of Arthropod-Borne Microorganisms. While it is possible to retrieve the sequence of CFPs from the UniProtKB/Swiss-Prot database with the ExPASy ScanProsite tool using the $\langle\{C\}^*\rangle$ syntax for pattern recognition, a very important subset of proteins is overlooked: those that have cysteine residues confined to an amino-terminal segment predicted to be excised upon secretion via the signal-peptide-dependent pathway. The precursor of those proteins would not be cysteine free but the mature-secreted protein would be (Figure 1). For that reason, all proteins of at least 200 residues with cysteine residues confined to a 70-amino acid NH2-terminal segment were analyzed for the presence of potential signal peptides. Similar attention was taken while screening for the presence of WFPs. The frequency of CFPs found in the genome of the 12 species of arthropod-borne microorganisms studied was found to be as low as 0.94% in *A. phagocytophilum* and as high as 16.8% in *B. burgdorferi*. For WFPs, the lowest percentage (3.86%) was found in *C. burnetii* and the highest (16.21%) in *B. burgdorferi* (Table 1). When only those proteins predicted to be secreted through the signal-dependent pathway were considered (sCFPs and sWFPs), a clear difference was detected with preferential secretion of WFPs by four species of Alphaproteobacteria (*A. phagocytophilum, E. chaffeensis, O. tsutsugamushi,* and *R. rickettsii*) and the only eukaryote included in the study, *P. falciparum*. In contrast, the other arthropod-borne pathogenic bacteria (*B. bacilliformis, C. burnetii, F. tularensis, Y. pestis,* and *B. burgdorferi*) preferentially secrete CFPs. Of interest, the two arthropod endosymbionts included in the study (*W. glossinidia* and *W.* endosymbiont of *D. melanogaster*) do not have preference for the secretion for either CFPs or WFPs (Table 1). When the ratio of sCFPs to sWFPs was plotted against the percentage of sCFPs, a clear separation of pathogenic and nonpathogenic microorganisms became evident in a logarithmic plot (Figure 2). *P. falciparum* and four species of pathogenic Alphaproteobacteria segregated into the lower left corner of the plot while *B. bacilliformis, B. burgdorferi,* and the three species of pathogenic Gammaproteobacteria segregated to the upper right corner of the plot. The two species of arthropod endosymbionts remained at the center of the plot.

3.2. CFPs and WFPs in Functionally Defined Proteins Secreted by Proteobacteria. In order to define the biological

>gi|110671084|ref|YP_667641.1| lipoprotein [*F. tularensis*]

MYFFANLNMRIILMKLNAKSTLLLSLLGVTLSSCSSLK ATKDNRDPFESYNRKMYAFNDKAYETLTPAANSYEKVVPDTLKSGIFNIFQNLAEPARVANDM
FQGEWDYAGDDGFRFLTNTTLGLAGYFDVADSWFNKPMRYHQSFAVTLHKWGVYKDNEASPYVVWPLIGPGTLEDITTGVDALFNPLTYIFFFAPVGSAIS
WGVSVGTTGAYYVNQGVSYLPSYSNLKEVSIDPYIAMRNAYLQNYDYGMAKVLKQQLSKDDATIQTDQAVLGVLGLDSDNVNAQIATTGGTKVKSSEPPVI
FKSNISTVNKASQVEEAFDQDYSNDSTALDLANVDNDDTTQSEASKLKTEIRDAKADLPGDADSPASAVQTLAEQG

>gi|108811011|ref|YP_646778.1| adhesin system component [*Y. pestis*]

MNSKLYKLIFCRRLGCLIA VGEFTRSYGRAFSSKGGQAGNNQRRAVGILSRLAMMTGLALGIFPLLVLAHPVLPVNGHVVIGQGMLDQQSSTLTVTQQTDK
LAINWDSFDIAHGHSVIYAQPGSQSIALNQVQGQSASQIYGRLQANGGVFLLNPRGILFGKEAQVNVGGLVASTKYMSNPEFLSGDYRLIGGESEGNIINQ
ANLRSAPGGYIALVGNRIDNQRSGSITTPQGNTVLAVGHSVTLNLDHGNLLGVQIQGETVAALIQNGGLIQADGGVIQLTAKGKDMLMDTVIDNTGILQAK
GLSAKNGAIYLDGGGEGVVSQMGTIDVNNQQGRGGRAVVEGKRIYLNKNSNIKAKGTAGGGTVLVGGGAGKAKIIRSEMPRLW

>gi|121602140|ref|YP_989591.1| surface antigen/outer membrane protein, OMP85 family [*B. bacilliformis*]

MICQAKVLLQSGILRTACLGLCLVFSFPQPLAA FEIFGIHLFGEKKSNAPSDGIQGAQKSYKVDVVAPPGAPLEGVKIVKTVSSLVADQDKSVSSSAGLLA
KARSNYREILSALYADGRYGGVISIKINGLEVADVSPMTELPDKSTIVITVDAGPQYVFNAVRINKASPLAKYKTGGMISVEDLGYKVGALAKSETILNAE
RWAIEGWRQRGYAKADVISRDIVADHARRIIDAQIVVDPKQKAYYGSINVRDVNKHPHIDPMYVVWMTGLKPGGQYDSEAVAKANKRLARLDVFRSIDVRE
ADTINSDGHLPLTVVVEERKLRRLGMGGSYSTLDGSGGEIYWTHSNLFGHAERFKIETKINSIGRRKEQSYHPKNFNYLLEVTFIKPGIITPDTDLGAELK
VQRDVLDSYITTAIKGKLSLTHVINDNLSGRVAMVITRGHLHDNHLGNRNFTVIGLQSDLIYDSRNNKLNATKGLYSEVILNPFYEMNLSSFMTKMTVEGR
SYWVLDTKNRFVFATRAKLGTIIGNNKILLPPDVLFFSGGGGSVRGYAYHNIGIKTENGIIGGRSLIEGSAELRFSLNDDIGLVSFVDGGRVEEKANFGFL
QKIKWGMGIGGRYMTGLGPVRFDVAWPLKREKGDPRIGIYVGIGQAF

>gi|124805602|ref|XP_001350486.1| blood stage antigen 41–3 precursor [*P. falciparum*]

MLLRHNSFCICIILVLCSVQRCLSD EQNINDWPIDFEYNSKSLPSIEVKLSPPENPLPQVSAEIKILESARLKLEEGMMQKLEDEYNKSLSMAKLKIKDTV
ENSLSIFNDPNILSSVISNSVKILKKKKNLRKIKETTDEEKTSDNVSQMYERKGGPLPPPELRKHTLFLEQNYLNKTIPSVKISLTEISEPSVLIKEKIEE
IEQYRTDEEVTMFETAISELDILTDITMLELEKQMQLQLNPFLVDKQIVHRSLEKELKEMGKAEQRENVNKSSQTQSSFLEQEENENTGNILNVKISQTDY
SYPTIDELVMQMQKKRDITEKLERQKILELQMKLLKAQSEMIKDALHFSISKVIAQYSPIVETLKLQTLKNF

FIGURE 1: Amino acid sequence of proteins secreted by four arthropod-borne pathogens (*F. tularensis, Y. pestis, B. bacilliformis,* and *P. falciparum*) with cysteine residues (shaded in gray) confined to the predicted signal peptide segment (boxed italics) of their precursors. Signal peptide excision sites were identified using the SMART online tool. The number of cysteine residues present in the signal peptide segments shown varies from one in *F. tularensis* lipoprotein to four in *P. falciparum* blood stage antigen. The mature proteins produced following the removal of signal peptide segments during the secretion process are free of cysteine residues.
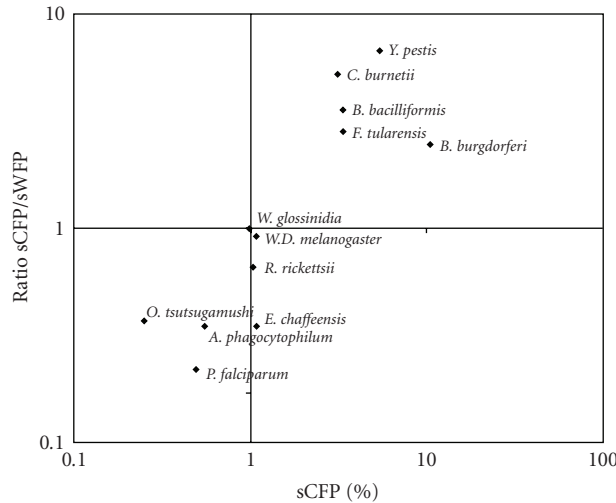

FIGURE 2: Frequency of sCFPs (percentage) plotted against a ratio of sCFP/sWFP in 12 arthropod-borne microorganisms. A logarithmic scale of the x and y axes was used to improve visual separation of microorganisms with low sCFP/sWFP ratios.

significance of the absence of cysteine or tryptophan residues in proteins secreted by arthropod-borne pathogens, it is necessary to determine the abundance of such residues in the corresponding orthologs secreted by other pathogenic and nonpathogenic microorganisms that are not transmitted by arthropods. We selected to conduct this analysis in Proteobacteria because it is a large taxonomic group

with 466 complete genomes sequenced, including many species of pathogenic bacteria with well-characterized virulence and pathogenicity determinants. Furthermore, the majority of the arthropod-borne microorganisms included in this study (10 out of 12) belong to the Proteobacteria phylum. We selected 28 different families of proteins secreted by Proteobacteria and belonging to
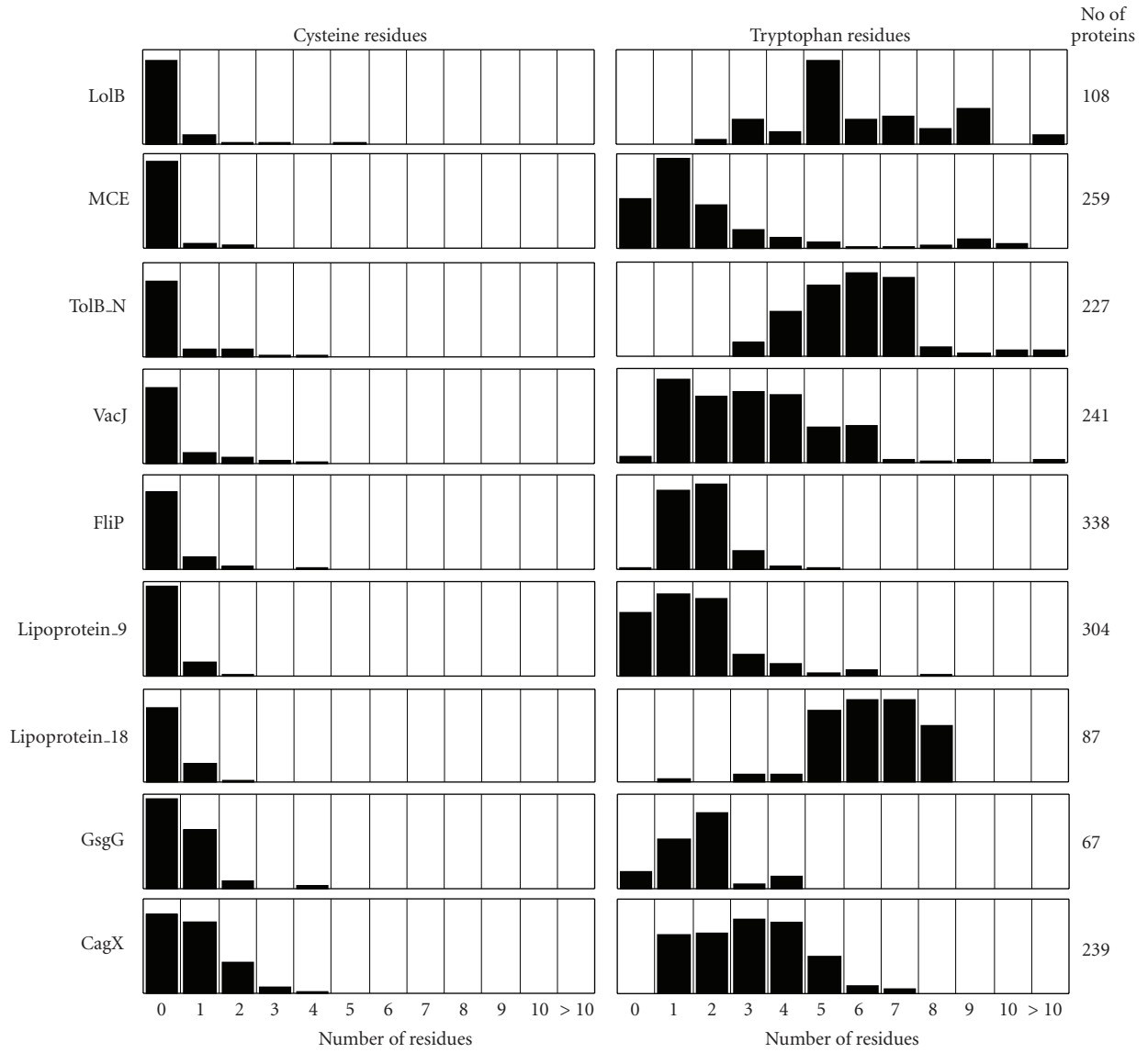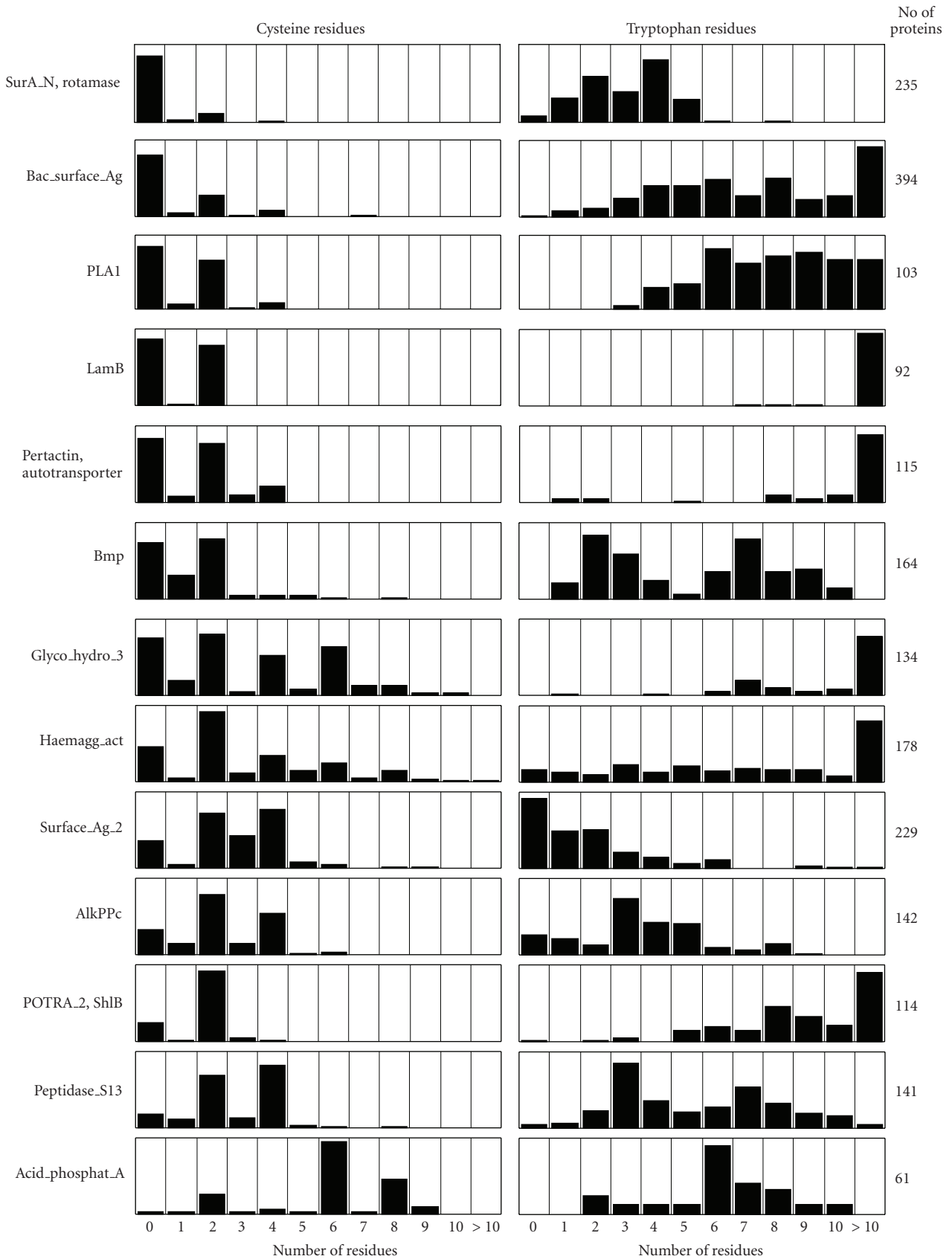
(a)
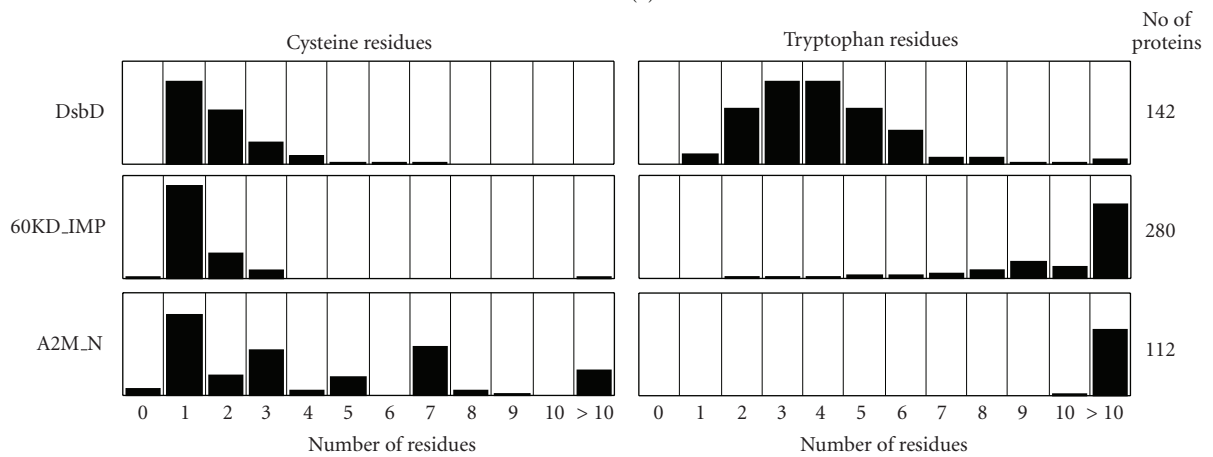
(b)

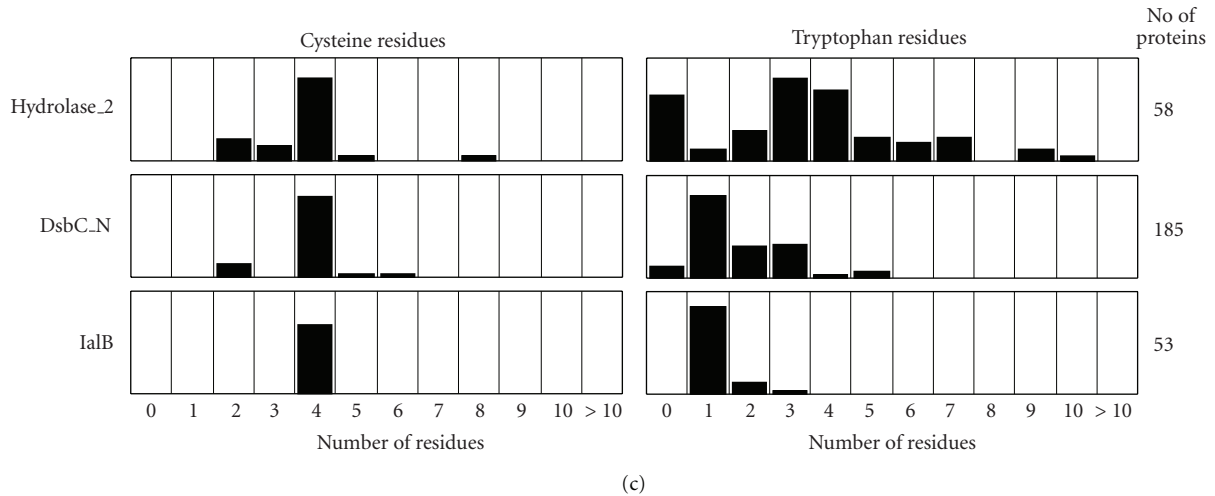Figure 3: Continued.

Figure 3: Continued.

(c)

FIGURE 3: Abundance and distribution analysis of cysteine (left panel) and tryptophan residues (right panel) in secreted proteins belonging to 28 different families and secreted by many species of pathogenic and non/pathogenic Proteobacteria. The protein families were arranged by cysteine abundance and distribution to optimize the visualization of patterns.

TABLE 1: Abundance of proteins devoid of cysteine residues (CFP) or tryptophan residues (WFP), included those predicted as secreted (sCFP and sWFP), in 12 species of arthropod-borne microorganisms.

Species	Proteins*	CFP (%)	WFP (%)	sCFP	sWFP	sCFP/sWFP
Alphaproteobacteria						
A. phagocytophilum HZ	1264	12 (0.94)	77 (6.09)	7	20	0.35
O. tsutsugamushi str. Boryong	1182	15 (1.26)	71 (6.00)	3	8	0.37
E. chaffeensis str. Arkansas	1105	20 (1.80)	133 (12.03)	12	34	0.35
R. rickettsii str. "Sheila Smith"	1345	40 (2.97)	99 (7.36)	14	21	0.66
W. endosymbiont of *D. melanogaster*	1195	36 (3.01)	90 (7.53)	13	14	0.92
B. bacilliformis KC583	1282	86 (6.70)	74 (5.76)	43	12	3.58
Gammaproteobacteria						
W. glossinidia endosymbiont of *G. brevipalpis*	611	26 (4.25)	71 (11.62)	6	6	1.00
F. tularensis subsp. *tularensis* FSC198	1605	78 (4.85)	133 (8.28)	54	19	2.84
C. burnetii RSA 493	2016	114 (5.65)	78 (3.86)	63	12	5.25
Y. pestis Nepal 516	3981	339 (8.51)	157 (3.94)	216	32	6.75
Spirochaetes						
B. burgdorferi B31	851	140 (16.45)	138 (16.21)	89	36	2.47
Apicomplexa						
P. falciparum 3D7	5267	64 (1.21)	567 (10.76)	26	115	0.22

*Chromosomal proteins only.

different functional categories to conduct the cysteine- and tryptophan-abundance and distribution analysis. The domains identified in the proteins selected for this analysis allowed for access (using a SMART window to proteins of similar composition) to the sequence of orthologs present in other species of Proteobacteria. The number of cysteine and tryptophan residues present in the mature proteins selected was recorded and a figure indicating their relative abundance was visually inspected for emergence of patterns (Figure 3). When considering the abundance and distribution of cysteine residues in these

protein families, three different patterns were detected. Pattern-I, which was found in proteins carrying 9 different domains (LolB, MCE, TolB_N, VacJ, Flip, Lipoprotein_9, Lipoprotein_18, CsgG, and GagX), is characterized by a predominance of CFPs over proteins expressing one or more cysteine residues. Pattern-II, which was found in proteins carrying the DsdB, 60KD_IMP, or A2M_N domains, shows a predominance of proteins expressing one cysteine residue and rarely expressing CFPs. Finally, Pattern-III was found in proteins expressing 16 different sets of domains (Sur_A_N/Rotamase, Bac_surface_Ag, PLA1,

Table 2: Domain and functional groups associated with each cysteine-abundance and distribution-pattern detected in pathogenic and nonpathogenic Proteobacteria.

Patterns	Functional category	Domains
Pattern-I	Lipoproteins	CsgG, Lipoprotein_9, Lipoprotein_18, LolB, VacJ
	Part of protein-secretion systems	CagX, FliP, MCE, TolB_N
Pattern-II	Endopeptidase inhibitory activity	A2M_N
	Generation of reduced environment in periplasm	DsdB
	Protein insertion into membranes	60KD_IMP
Pattern-III	Adhesin	Haemagg_act, Pertactin, Autotransporter
	Invasin	IalB
	Hemolysin	POTRA_2, ShlB
	Enzyme	PLA1, Acid_phosphatase_A, AlkPPc, Glyco_hydro_3, Hydrolase_2, Peptidase_S13
	Surface antigen	Bac_surface_Ag, BMP, Surface_Ag_2
	Membrane transporter	LamB

LamB, Pertactin/Autotransporter, Bmp, Glyco_hydro_3, Haemagg_act, Surface_Ag_2, AlkPPc, POTRA_2/ShlB, Peptidase_S13, Acid_phosphatases_A, Hydrolase_2, DsbC_N, and IalB) and was characterized by a wider variety of cysteine abundance, but with a clear preference for even numbers of residues. It is possible that this segregation into 3 patterns is a reflection of a functional spectrum defined by structural requirements of reduced sulfhydryl groups (as in proteins of Pattern-II), correct folding and rigidity conferred by a defined number of disulfide bonds (as in proteins of Pattern-III), and high flexibility and promiscuous interactivity conferred by the absence of cysteine residues (as in proteins of Pattern-I). In support of this interpretation, a correlation was found between functionality and cysteine abundance and distribution patterns for proteins secreted by Proteobacteria (Table 2). In contrast to the clear definition of cysteine-expression patterns in the proteins analyzed, no clear patterns emerged when analyzing the abundance and distribution of tryptophan residues. With the exception of proteins carrying the Surface_Ag_2 domain (which have a dominant WFPs profile) and those carrying the DsbC_N and IalB domains (which preferentially express a single tryptophan residue), no particular pattern of tryptophan residues was observed, with most proteins expressing a variable number of tryptophan residues, frequently more than 10 per protein.

3.3. Secreted CFPs in Pathogenic Arthropod-Borne Proteobacteria.

Many sCFPs were found in pathogenic arthropod-borne Proteobacteria, including 79 in Alphaproteobacteria and 333 in Gammaproteobacteria (Table 1). In order to identify those that could represent virulence and pathogenicity determinants or potential vaccine targets, we eliminated from consideration those with domains detected in sCFPs of any of the arthropod endosymbionts included in the study. Still, a large list of sCFP-associated domains remained with 42 present in Alphaproteobacteria (Table 3) and 137 in Gammaproteobacteria (Table 4). In addition to these proteins, a group of sCFPs, characterized by the absence of any of the domains

Table 3: Domains associated with secreted CFPs in four species of arthropod-borne Alphaproteobacteria.

Species	Domains
A. phagocytophilum	OEP, Peptidase_M16, Peptidase_M17, Peptidase_S49, TPR_2
B. bacilliformis	ABC_sub_bind, Ami_3, **Autotransporter**, **Bac_surface_Ag**, BPD_transp_1, DPBB_1, DUF165, DUF374, DUF541, FlgH, FlgI, **FliP**, HAMP, HATPase_C, HisKA, HlyD, **Lipoprotein_9**, LolA, MotA_ExbB, OEP, Oxidored_q3, PDZ, Peptidase_M23, Peripla_BP_2, **Pertactin**, Porin_2, SBP_bac_9, **Surf_Ag_VNR**, **TolB_N**, **PD40**, Trypsin, TSPc
E. chaffeensis	**Bac_surface_Ag**, BON, OEP, PDZ, Peptidase_S49, **Surf_Ag_VNR**, TerC, Trypsin
R. rickettsii	AAA, **Bac_surface_Ag**, DUF6, FTR1, OmpW, PBPb, PDZ, Peptidase_M23, **Surf_Ag_VNR**, Trypsin, TSPc

Bold font domains selected for definition of cysteine-abundance and distribution patterns (Figure 3).

registered in the SMART and Pfam databases and conformed by segments of low complexity, intrinsic disorder, coiled-coil structure, transmembrane regions, and/or internal repeats, was found among the CFPs secreted by these pathogens. These sequences were found to be particularly abundant in Gammaproteobacteria, with 14 present in *C. burnetii,* 22 in *F. tularensis,* and 28 in *Y. pestis.* Only one was found in both *A. phagocytophilum* and *O. tsutsugamushi,* three in *R. rickettsii,* four in *E. chaffeensis,* and seven in *B. bacilliformis.* One strategy to identify proteins of interest in such a large list of candidates is to focus on those with the longest sequence of amino acid residues. After all, the significance of absence of a particular residue in a protein is highly dependent on protein length. When only the precursors of at least 500 amino acid residues were considered, few domains were identified (Table 5). Of these, the Pertactin and Autotransporter domains are of interest to understand

TABLE 4: Domains associated with secreted CFPs in arthropod-borne Gammaproteobacteria.

Species	Domains
C. burnetii	AAL_decarboxy, Acid_phosphat_B, **Bac_surface_Ag**, Bac_Ubq_Cox, BPD_transp_1, Cation_efflux, DUF2066, DUF502, FKBP_N, FtsQ, FtsX, HlyD, **Lipoprotein_9**, **LolB**, Lys M, **MCE**, MotA_ExbB, MreC, MVIN, OmpA_membrane, OpuAC, Oxidored_q1, PDZ, Peptidase_M16, Peptidase_M22, Peptidase_M23, Peptidase_M50, Peptidase_S24, PNTB, Polysacc_deac_1, POTRA_1, Sec_GG, SecD_SecF, SecY, SLT, SPOR, **Surf_Ag_VNR**, **TolB_N**, **PD40**, Toluene_X, Trypsin, TSPc
F. tularensis	**Bac_surface_Ag**, DUF1239, DUF165, G_glu_transpept, GDPD, HlyD, **LolB**, Lys_M, **MCE**, Neur_chan_LBD, OEP, **PD40**, Peptidase_M50, Peroxidase, PQQ, RmuC, SBP_bac_9, Sec_GG, SecD_SecF, Secretin, SOUL, Sugar_tr, **SurA_N**, **Surf_Ag_VNR**, **TolB_N**, **VacJ**, VWA, WZZ
Y. pestis	Abhydrolase_BP_2, ADC_lyase, Ami_2, ArsB, AsmA, **Autotransporter**, **Bac_surface_Ag**, BON, BPD_transp_1, Cache_1, **CsgG**, DPBB_1, DUF1284, DUF1454, DUF1471, DUF187, DUF218, DUF481, DUF541, DUF6, DUF799, DUF945, Ell-Sor, FKBP_N, Flgl, **FliP**, FMN_bind, GDPD, **Glyco_hydro_3**, GntP_permease, **Haemagg_act**, HAMP, HATPase_C, HIM, HisKA, HlyD, KdgM, KdpC, Kelch_2, Lactamase_B, LamB, Lipase_GDSL, **Lipoprotein_18**, **Lipoprotein_9**, **LolB**, Lyase_8, Lyase_catalyt, Lyase_N, Lys M, MA, MalM, MgtC, MotA_ExbB, MS_channel, OEP, Omptin, OmpW, Oxidored_q1, PAS, Patatin, PbH1, PBP5_C, PBPb, **PD40**, PDZ, Pectate_lyase_2, Peptidase_M16, Peptidase_M23, Peptidase_M48, Peptidase_S11, Peripla_BP_1, Peripla_BP_2, **Pertactin**, PG_binding_1, Pili_assembly_N, **PLA1**, Plug, Polysacc_deac_1, Porin_1, PQQ, **Rotamase**, RpiB, SBP_bac_1, SBP_bac_5, SBP_bac_7, SBP_bac_9, Secretin, Slp, SLT, SPOR, STAS, Sulfatase, **SurA_N**, **Surf_Ag_VNR**, TM_helix, Tol_tol_Ttg2, **TolB_N**, Toluene_X, TonB, TonB_dep_Rec, TPR_2, Transgly, Trypsin, TSPc, UPF0118, **VacJ**, VWA, YscJ_FliF, ZipA_C

Bold font domains selected for definition of cysteine-abundance and distribution patterns (Figure 3).

the role of sCFPs in the immunobiology of ABDs because they have been identified as virulence and pathogenicity determinants [17, 18] and because both have been found to be coexpressed in very large sCFPs, including a *Y. pestis* protein of 3710 amino acids (GenBank accession number gi: 45443160). Another domain of interest, Bac_surface_Ag, represents a potential target for multipathogen vaccines directed to sCFPs because it is the only domain found associated with sCFPs of six species of arthropod-borne Proteobacteria (Table 5) and in a 821-amino acid residue outer membrane CFP secreted by *B. burgdorferi* Spirochetes (GenBank accession number gi: 15595140).

4. Discussion

Several technical hurdles have complicated the development of effective vaccines to control the transmission of ABDs, including (1) identification of protective epitopes expressed on the pathogen and/or its arthropod vector, (2) induction of safe and long-lasting immunity that overcomes immune evasion mechanisms used by the pathogens and amplified by immuno-modulators present in arthropod saliva [19, 20], and (3) induction of an immunity expected to operate effectively in human populations of tremendous diversity. Some of these problems derive from the assumption that the induction of species-specific immunity is the most effective approach to develop vaccines to prevent infectious diseases. Research into an alternative approach, the induction of immunity of broad specificity, can benefit from the great diversity of pathogens and vectors involved in the etiopathogenesis of ABDs in order to identity the cross-reactive epitopes required for the induction of cross-protective immunity. Just as adaptation of arthropods to hematophagy is an example of convergent evolution [21], the adaptation of microbial pathogens to the complex biological processes that guarantees their survival and multiplication in two very different kinds of hosts, vertebrate and invertebrate, can be seen as an example of convergent evolution as well. Under such consideration, the identification of common structural determinants in molecules secreted by pathogens into the host-parasite interface might shed light into vulnerabilities of the host defense systems, providing opportunities for the design of alternatives to control ABDs. The availability of complete genomes in many of the etiologic agents of ABDs offers the possibility to identify, via the comparative analysis of their secretomes described herein, such structural determinants.

4.1. Cysteine Residues in Protein Function and Antigenicity/Immunogenicity. One of the amino acid residues with a prominent effect on the structure of proteins, cysteine, can be expected to play critical roles in protein function, antigenicity and immunogenicity. The role played by cysteine residues in protein function has been demonstrated in numerous experiments using chemical modification techniques and cysteine scanning mutagenesis [22, 23], but its role in antigenicity and immunogenicity of proteins is only beginning to be characterized using the same techniques [24]. The flexibility that characterizes a group of proteins with a clear cysteine use bias, Intrinsically Unstructured Proteins [10–13], can play a prominent role in host-parasite relationships by favoring protein-protein interactions of low affinity and specificity. From an interactomics perspective, this kind of protein-protein interactivity might favor the interaction of pathogen-secreted proteins with nondefense proteins present in host tissues, thus interfering with pathogen recognition by defense proteins [25]. As a result, immunization with these cysteine-depleted proteins

TABLE 5: Domains associated with sCFPs of at least 500 residues in length and present in more than one species of the indicated species of arthropod-borne Proteobacteria: *A. phagocytophilum* (A.p), *B. bacilliformis* (B.b), *E. chaffeensis* (E.c), *R. rickettsii* (R.r), *C. burnetii* (C.b), *F. tularensis* (F.t), and *Y. pestis* (Y.p).

Domains	A.p	B.b	E.c	R.r	C.b	F.t	Y.p
PBP_dimer/Transpeptidase	—	—	—	—	1	1	—
Peptidase_M16	1	—	—	—	—	—	1
Pertactin/Autotransporter	—	1	—	—	—	—	3
Sec_GG/SecD_SecF	—	—	—	—	1	1	—
Secretin	—	—	—	—	—	1	1
Bac_surface_Ag	—	2	1	1	1	1	2

might shift the balance toward protective immunity against pathogens that secreted them [26].

4.2. Evolutionary Considerations on the Secretion of CFPs by Proteobacteria.

Judged by the contrast between cysteine- and tryptophan-abundance and distribution expression patterns identified in 28 families of proteins secreted by Proteobacteria, a restrictive evolutionary process must be in operation to keep some of these proteins in a cysteine-free state (Figure 3). This is particularly evident in all proteins classified in Pattern-I which includes nine families of surface lipoproteins and components of protein-transport systems. This expression pattern is better explained by the idea that these families of proteins were derived from ancestors that were themselves free of cysteine residues. During the radiation of species a variable number of mutations might have occurred that introduced additional tryptophan residues in the proteins without restriction. While a similar process might have led to the diversification in the number of cysteine residues, a restrictive force must have kept it to a minimum. Given that this occurred in all species of Proteobacteria, including those known to be pathogenic and nonpathogenic to vertebrates and arthropods, it is unlikely that the immune systems of these hosts were responsible for the restriction. Most likely, the function of the proteins belonging to this group of proteins is seriously compromised by the introduction of cysteine residues. A somewhat similar mechanism might have operated during the evolution of the three families of proteins classified in Pattern-II. An ancestor with a single cysteine residue may have evolved under conditions that allowed the acquisition of cysteine residues but restricted its loss. The case of proteins carrying the A2M_N domain is peculiar because it is the only family of proteins where a variable, and odd number, of cysteine residues predominates. This suggests that while the accumulation of cysteine residues through the evolution of this family was fully permitted, those individuals with an odd number of residues had a survival advantage, probably because the function of the A2M_N domain, just as it is known to occur in vertebrate proteins carrying it, depends on the availability of a free sulfhydryl group [27]. One scenario that could explain the more complex cysteine-abundance and distribution analysis of the families of proteins classified in Pattern-III is that the ancestor of these protein families,

starting with an even number of cysteine residues, evolved under conditions that allowed either the acquisition or loss of cysteine residues. Once again the survival advantage of having, in this case, an even number of cysteine residues explains the overall pattern. Under this hypothetical scenario it becomes necessary to explain a peculiar jump in protein evolution. In the case of protein families carrying the Bmp, Glyco_hydro_3, Haemagg_act, Surface_Ag_2, AlkPPc, POTRA_2/ShlB, and Peptidase_S13 domains (which would be predicted to derive from ancestors with two cysteine residues and thus the potential to have a disulfide bond in the protein structure, Figure 3) could be transformed into a protein where the formation of disulfide bonds is no longer possible because of the loss of two cysteine residues. How much protein functionality is affected by the disappearance of the only disulfide bond present in the structure of these proteins? Similarly one could ask: How much is the function of a protein devoid of cysteine residues affected when accumulation of two cysteine residues allows the emergence of disulfide bonds in the protein structure? This complementary scenario could occur with proteins carrying the Sur_A/Rotamase, Bac_surface_Ag, PLA1, LamB, and Pertactin/Autotransporter domains.

5. Conclusions

It is apparent that cysteine residues play a dramatic role in protein function and that their removal, or introduction, into protein sequences represents a critical event of great significance in pathogen evolution and host-parasite relationships. The use of the a simple and novel bioinformatics tool, amino acid-abundance and distribution analysis, can prove useful in the clarification on some of these ideas and address others that were not included in the discussion, such as the role of sWFPs in host-parasite relationship, in particular in the context of two arthropod-borne diseases that are caused by pathogens that secrete them abundantly, *B. burgdorferi* and *P. falciparum*. The use of cysteine- and/or tryptophan-directed mutagenesis will prove instrumental in clarifying the role that these amino acid depleted proteins may have in the immunobiology of ABDs. It will also open to experimentation the usefulness of vaccines based on a new kind of epitope, one defined not by what it contains in its sequence, but rather by what it lacks.

Acknowledgments

References

[1] N. G. Gratz, "Emerging and resurging vector-borne diseases," *Annual Review of Entomology*, vol. 44, pp. 51–75, 1999.

[2] R. W. Sutherst, "Global change and human vulnerability to vector-borne diseases," *Clinical Microbiology Reviews*, vol. 17, no. 1, pp. 136–173, 2004.

[3] T. V. Inglesby, D. T. Dennis, D. A. Henderson, et al., "Plague as a biological weapon: medical and public health management. Working Group on Civilian Biodefense," *The Journal of the American Medical Association*, vol. 283, no. 17, pp. 2281–2290, 2000.

[4] P. C. F. Oyston, A. Sjostedt, and R. W. Titball, "Tularaemia: bioterrorism defence renews interest in *Francisella tularensis*," *Nature Reviews Microbiology*, vol. 2, no. 12, pp. 967–978, 2004.

[5] A. F. Azad, "Pathogenic rickettsiae as bioterrorism agents," *Clinical Infectious Diseases*, vol. 45, supplement 1, pp. S52–S55, 2007.

[6] S. C. Arya, "Yellow fever vaccine safety: a reality or a myth?" *Vaccine*, vol. 20, no. 31-32, pp. 3627–3628, 2002.

[7] A. Nalca, P. F. Fellows, and C. A. Whitehouse, "Vaccines and animal models for arboviral encephalitides," *Antiviral Research*, vol. 60, no. 3, pp. 153–174, 2003.

[8] C. Pinilla, R. Martin, B. Gran, et al., "Exploring immunological specificity using synthetic peptide combinatorial libraries," *Current Opinion in Immunology*, vol. 11, no. 2, pp. 193–202, 1999.

[9] K. W. Wucherpfennig, P. M. Allen, F. Celada, et al., "Polyspecificity of T cell and B cell receptor recognition," *Seminars in Immunology*, vol. 19, no. 4, pp. 216–224, 2007.

[10] P. Tompa, "Intrinsically unstructured proteins," *Trends in Biochemical Sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[11] P. Tompa and M. Fuxreiter, "Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions," *Trends in Biochemical Sciences*, vol. 33, no. 1, pp. 2–8, 2008.

[12] P. Romero, Z. Obradovic, and A. K. Dunker, "Natively disordered proteins: functions and predictions," *Applied Bioinformatics*, vol. 3, no. 2-3, pp. 105–113, 2004.

[13] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein in intrinsic disorder," *BMC Bioinformatics*, vol. 7, article 208, 2006.

[14] J. A. Greenbaum, P. H. Andersen, M. Blythe, et al., "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 75–82, 2007.

[15] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork, "SMART 5: domains in the context of genomes and networks," *Nucleic Acids Research*, vol. 34, database issue, pp. D257–D260, 2006.

[16] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: signalP 3.0," *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, 2004.

[17] P. Emsley, I. G. Charles, N. F. Fairweather, and N. W. Isaacs, "Structure of *Bordetella pertussis* virulence factor P.69 pertactin," *Nature*, vol. 381, no. 6577, pp. 90–92, 1996.

[18] I. R. Henderson, F. Navarro-Garcia, and J. P. Nataro, "The great escape: structure and function of the autotransporter proteins," *Trends in Microbiology*, vol. 6, no. 9, pp. 370–378, 1998.

[19] R. G. Titus, J. V. Bishop, and J. S. Mejia, "The immunomodulatory factors of arthropod saliva and the potential for these factors to serve as vaccine targets to prevent pathogen transmission," *Parasite Immunology*, vol. 28, no. 4, pp. 131–141, 2006.

[20] J. S. Mejia, J. V. Bishop, and R. G. Titus, "Is it possible to develop pan-arthropod vaccines?" *Trends in Parasitology*, vol. 22, no. 8, pp. 367–370, 2006.

[21] W. C. Black IV and B. C. Kondratieff, "Evolution of arthropod disease vector," in *Biology of Disease Vectors*, W. C. Marquardt, et al., Ed., pp. 9–23, Elsevier, London, UK, 2nd edition, 2005.

[22] R. B. Bass, S. L. Butler, S. A. Chervitz, S. L. Gloor, and J. J. Falke, "Use of site-directed cysteine and disulfide chemistry to probe protein structure and dynamics: applications to soluble and transmembrane receptors of bacterial chemotaxis," *Methods in Enzymology*, vol. 423, pp. 25–51, 2007.

[23] M. Palmer, P. Saweljew, I. Vulicevic, A. Valeva, M. Kehoe, and S. Bhakdi, "Membrane-penetrating domain of streptolysin O identified by cysteine scanning mutagenesis," *Journal of Biological Chemistry*, vol. 271, no. 43, pp. 26664–26667, 1996.

[24] A. Rimon, C. Hunte, H. Michel, and E. Padan, "Epitope mapping of conformational monoclonal antibodies specific to NhaA Na^+/H^+ antiporter: structural and functional implications," *Journal of Molecular Biology*, vol. 379, no. 3, pp. 471–481, 2008.

[25] J. S. Mejia, F. Moreno, C. Muskus, I. D. Véez, and R. G. Titus, "The surface-mosaic model in host-parasite relationships," *Trends in Parasitology*, vol. 20, no. 11, pp. 508–511, 2004.

[26] A. Mohan, W. J. Sullivan Jr., P. Radivojac, A. K. Dunker, and V. N. Uversky, "Intrinsec disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes," *Molecular BioSystems*, vol. 4, no. 4, pp. 328–340, 2008.

[27] L. Sottrup-Jensen, "α-macroglobulins: structure, shape, and mechanism of proteinase complex formation," *Journal of Biological Chemistry*, vol. 264, no. 20, pp. 11539–11542, 1989.