Microbiology (2006), **152**, 113–121 **DOI** 10.1099/mic.0.28317-0

Lipoprotein computational prediction in spirochaetal genomes

João C. Setubal,[1] Marcelo Reis,[2] James Matsunaga[3,4] and David A. Haake[3,4]

Correspondence
David A. Haake
dhaake@ucla.edu

[1]Virginia Bioinformatics Institute, Virginia Tech, Bioinformatics 1, Box 0477, Blacksburg, VA 24060-0477, USA

[2]Laboratório de Bioinformática, Instituto de Computação, Universidade Estadual de Campinas, Caixa Postal 6076, Campinas, SP 13084-071, Brazil

[3]Division of Infectious Diseases, 111F, Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, CA 90073 USA

[4]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095 USA

Lipoproteins are of great interest in understanding the molecular pathogenesis of spirochaetes. Because spirochaete lipobox sequences exhibit more plasticity than those of other bacteria, application of existing prediction algorithms to emerging sequence data has been problematic. In this paper a novel lipoprotein prediction algorithm is described, designated SpLip, constructed as a hybrid of a lipobox weight matrix approach supplemented by a set of lipoprotein signal peptide rules allowing for conservative amino acid substitutions. Both the weight matrix and the rules are based on a training set of 28 experimentally verified spirochaetal lipoproteins. The performance of the SpLip algorithm was compared to that of the hidden Markov model-based LipoP program and the rules-based algorithm Psort for all predicted protein-coding genes of *Leptospira interrogans* sv. Copenhageni, *L. interrogans* sv. Lai, *Borrelia burgdorferi*, *Borrelia garinii*, *Treponema pallidum* and *Treponema denticola*. Psort sensitivity (13–35 %) was considerably less than that of SpLip (93–100 %) or LipoP (50–84 %) due in part to the requirement of Psort for Ala or Gly at the −1 position, a rule based on *E. coli* lipoproteins. The percentage of false-positive lipoprotein predictions by the LipoP algorithm (8–30 %) was greater than that of SpLip (0–1 %) or Psort (4–27 %), due in part to the lack of rules in LipoP excluding unprecedented amino acids such as Lys and Arg in the −1 position. This analysis revealed a higher number of predicted spirochaetal lipoproteins than was previously known. The improved performance of the SpLip algorithm provides a more accurate prediction of the complete lipoprotein repertoire of spirochaetes. The hybrid approach of supplementing weight matrix scoring with rules based on knowledge of protein secretion biochemistry may be a general strategy for development of improved prediction algorithms.

Received 2 July 2005
Revised 20 September 2005
Accepted 18 October 2005

INTRODUCTION

Lipoproteins are universal components of eubacterial membranes. Anchoring of lipoproteins to lipid bilayers occurs via fatty acids that covalently modify the amino-terminal Cys of the mature protein (Braun & Wolff, 1970). The ability of lipoproteins to decorate bacterial membranes provides for a wide variety of essential structural and functional roles. Murein lipoprotein anchors the inner face of the Gram-negative outer membrane to the peptidoglycan cell wall. Other lipoproteins function as adhesins, enzymes, transporters, binding proteins, toxins and in a variety of other capacities essential for virulence (Madan Babu & Sankaran, 2002). Given their broad distribution among bacteria and their unique structural features, it is not surprising that at least one of the toll-like receptors (TLR2) is designed to detect lipoproteins as an innate response to the presence of bacteria (Aliprantis *et al.*, 1999; Brightbill *et al.*, 1999).

Abbreviations: PPCG, predicted protein-coding gene; TS, training set; WM, weight matrix.

Supplementary data tables with the complete listing of all lipoprotein predictions described in this paper are available with the online version of this paper. The SpLip program was encoded in perl, runs on any operating system that supports the perl interpreter and is available upon request by contacting the authors.

In spirochaetes, lipoproteins are the most prominent proteins in the total membrane protein profile. Examples of highly abundant spirochaetal proteins are OspA of *Borrelia burgdorferi*, the causative agent of Lyme disease, Tpp47 of the syphilis spirochaete *Treponema pallidum* and LipL32, the major outer-membrane protein of pathogenic *Leptospira* species (Chamberlain *et al.*, 1988; Haake *et al.*, 2000; Howe *et al.*, 1985). Research on spirochaetal lipoproteins has been essential to the understanding of spirochaetal physiology and pathogenesis. In *Borrelia* and *Leptospira* species, differential expression of lipoproteins is a hallmark of the transition to life inside the mammalian host (Barnett *et al.*, 1999; Schwan *et al.*, 1995). Several lipoproteins have been shown to be involved in the interaction of pathogenic spirochaetes with host molecules. Adherence of *B. burgdorferi* to extracellular matrix proteins is mediated by DbpA (Guo *et al.*, 1998) and Bbk32 (Probert & Johnson, 1998). *B. burgdorferi* has been shown to evade activation of the complement cascade by binding of Factor H to lipoprotein OspE and related proteins (Hellwage *et al.*, 2001). A number of spirochaetal lipoproteins have been shown to be targets of a protective immune response (Haake, 2000), confirming the importance of lipoproteins in the pathogenesis of spirochaetal diseases.

Since spirochaetes form a deep branch of the phylogenetic tree, it is not surprising that spirochaetal lipoproteins frequently have no homologues in the sequence databases. The divergence of spirochaetal lipoprotein sequences from those of other bacteria includes the signal peptide 'lipobox' region recognized by lipid modification enzymes. Lipid modification has been demonstrated experimentally for a relatively large number of spirochaetal lipoproteins (Haake, 2000). From these sequences it is possible to conclude that spirochaetal lipoproteins are secreted across the cytoplasmic membrane via a signal peptide, but that the lipobox at the carboxy terminus of the signal peptide differs significantly from that of other bacteria. Differences in the lipobox sequence presumably result from differences in the active site substrate specificities of the glyceryl transferase and type II signal peptidase that transfer a diacylglyceryl group to Cys and remove the signal peptide, respectively (Paetzel *et al.*, 2002).

The von Heijne consensus lipobox pattern is based on lipoprotein sequences of *Escherichia coli* and similar Gram-negative bacteria (von Heijne, 1989). The Psort program (Nakai & Horton, 1999), which is based on the von Heijne consensus lipobox pattern, fails to recognize 43 % of experimentally verified spirochaetal lipoprotein sequences. In general, the inaccuracy of Psort reflects the increased plasticity of the spirochaetal lipobox. Recent application of the hidden Markov model approach in the LipoP program (Juncker *et al.*, 2003) showed improved accuracy for lipoprotein recognition in general for bacteria other than *E. coli*, but includes many non-spirochaetal lipoproteins in its training set (TS). The emergence of data from spirochaetal genome sequencing efforts has resulted in the need for tools to accurately and efficiently identify lipoprotein genes (Fraser *et al.*, 1997, 1998; Glockner *et al.*, 2004; Nascimento *et al.*, 2004; Ren *et al.*, 2003; Seshadri *et al.*, 2004). We set out to design an algorithm specifically tailored to identify spirochaetal lipoproteins; our program is designated SpLip. Application of the SpLip program to the six available spirochaetal genomes shows improved accuracy over existing generic lipoprotein prediction algorithms.

METHODS

SpLip. SpLip uses a position-specific scoring matrix, also known as a weight matrix (WM) (Durbin *et al.*, 1998; Mount, 2001). The SpLip WM uses a TS consisting of 28 spirochaetal proteins with experimental evidence of lipidation (Table 1). The TS includes 26 sequences described by Haake (2000) plus LipL21 (Cullen *et al.*, 2003b) and LigB (Matsunaga *et al.*, 2003) from *Leptospira interrogans*. Analysis of the TS was used to define the three regions of the spirochaetal lipoprotein signal peptide shown in Fig. 1: the carboxy-terminal region (C-region or lipobox), the hydrophobic (H-) region and amino-terminal (N-) region. The SpLip WM is focused on the lipobox because this is the most conserved region among lipoproteins. The lipobox is an ungapped motif with 4 or 5 positions, as will be seen below. Additionally, the program determines the H-region length and hydrophobicity, and net charge of the N-terminal region.

Characterization of the spirochaetal lipobox. The C-terminal region, or lipobox, of the spirochaetal lipoprotein signal peptide is defined in principle as the four positions (−1, −2, −3 and −4) upstream of the cleavage site (position −5 is also considered, see below). A Cys residue is always found in position +1. This is an absolute requirement as there is no precedent that we are aware of for amino-terminal lipid modification of membrane proteins at amino acids other than Cys.

The TS was used to characterize the spirochaetal lipobox in three steps:

(1) Analysis of the TS yielded a set of lipobox rules which are a refinement of the spirochaetal lipobox described by Haake (2000).

(a) Position −1 – only Ala, Gly, Ser, Asn or Cys are allowed;

(b) Positions −3 or −4 – at least one of these positions should contain at least one of Leu, Ile, Val or Phe;

(c) The charged amino acids Lys, Arg, Asp, Glu and His are forbidden anywhere in the lipobox.

A predicted protein-coding gene (PPCG) that has a lipobox conforming to these rules (and to other constraints not pertaining to the lipobox itself, to be described below) is considered a probable lipoprotein.

(2) The WM was built following standard procedures (Durbin *et al.*, 1998; Mount, 2001) (see below)

(3) Lipoboxes in all PPCGs of *Leptospira interrogans* sv. Copenhageni were scored according to the WM. An analysis of the high-scoring PPCGs and TS members together with the multiple alignments of significantly similar pairs of PPCGs (see below), resulted in modification of the lipobox rules, as follows.

(a) Position −1 – in addition to Ala, Gly, Ser, Asn and Cys, the related amino acids Gln and Thr are also allowed;

Table 1. A set of 28 spirochaetal proteins with experimental evidence for lipidation used as the TS for the SpLip program

Gene/product	N-region sequence	H-region sequence	Lipobox (C-region)
Borrelia afzelii			
NlpH	MK	IINILFCLFLI	MLSGC
Borrelia burgdorferi			
OppA-1	MKKENPMKYIK	IALMLIIF	SLIAC
OppA-2	MKLQR	SLFLIIFFL	TFLCC
OppA-3	MSFNKTKKIGKKIK	IVTLLMLAV	SLIAC
LA7	MYKNGFFK	NYLSLFLIF	LVIAC
OspA	MKK	YLLGIGLIL	ALIAC
OspB	MR	LLIGFALAL	ALIGC
DbpA	MIKCNNKTFNNLLK	LTILVN	LLISC
OspC	MKK	NTLSAILMTLF	LFISC
OspD	MKKLIK	ILLLSLFLL	LSISC
OspE	MNKKMK	MFIICAVFI	LIGAC
OspF	MNKKIK	MFIICAIFM	LISSC
Borrelia hermsii			
Vmp7	MRKRISAIINK	LNISIIIMTVV	LMIGC
Vmp33	MKK	NTLSAILMTLF	LFISC
Brachyspira hyodysenteriae			
SmpA	MNKK	IFTLFLVVAASAI	FAVSC
Leptospira interrogans			
LigB	MKK	IFCISIFLSM	FFQSC
LipL21	MINR	LIALSLATM	IFAAC
LipL32	MKK	LSILAISVALFA	SITAC
LipL36	MRRNIMK	IAAVAALTV	ALTAC
LipL41	MRK	LSSLISVLVLLM	FLGNC
Treponema pallidum			
GlpQ	MR	GTYCVTLWGGVFAA	LVAGC
MglB-2	MKENSCTACSRR	LALFVGAAV	LVVGC
TmpA	MNAH	TLVYSGVALACAA	MLGSC
TmpC	MREKWVR	AFAGVFCAM	LLIGC
Tp47	MKVK	YALLSAGALQL	LVVGC
Tpd	MKR	VSLLGSAAIFAL	VFSAC
Tpp15	MVKR	GGAFALCLAV	LLGAC
Tpp17	MKGSVR	ALCAFLGVGALGSA	LCVSC

(b) Position –5 is also considered to be part of the lipobox; rule (1b) above is extended to this position;

(c) In addition to Leu, Ile, Val and Phe, the hydrophobic amino acids Tyr and Met are also included as possible amino acids required in positions −3, −4 or −5.

A PPCG that has a putative lipobox conforming to these modified rules (and to other constraints not pertaining to the lipobox itself, to be described below), and not to rules in step (1), is considered a possible lipoprotein.

Characterization of the hydrophobic (H-) region. Based on analysis of the TS and the requirement for a hydrophobic signal peptide, charged residues Lys, Arg, Asp, Glu and His were forbidden in the H-region. The H-region should be at least 7 aa long for probable lipoproteins and 6 aa long for possible lipoproteins.

Characterization of the amino-terminal (N-) region. In a lipoprotein signal peptide the N-terminal region should be positively charged. The N-terminal region is considered to extend from the first residue to the last charged residue (i.e. Lys, Arg, Asp, Glu or His). The residue following the last charged residue defines the start of the H-region.

Length of the mature lipoprotein. Both probable and possible lipoproteins have to follow this additional rule: the PPCG should have at least 50 residues downstream of the +1 position.

WM construction. The standard procedure (Durbin *et al.*, 1998; Mount, 2001) for building a WM is as follows. Background frequencies for the residues are determined for each organism separately, thus obtaining a different matrix for each organism. The background frequencies are given by the first 50 residues in all PPCGs. Then residue frequencies in the TS are determined. Entry (i,j) in the matrix (where i is a residue and j is a position in the sequence) is given by $\log_2(F_{i,j}/G_{i,j})$, where $F_{i,j}$ is the observed frequency of residue i in position j in the TS, and $G_{i,j}$ is the observed background frequency of residue i in position j.

The standard WM procedure was adapted to accommodate the rules described above. The lipobox for each sequence in the TS was known.

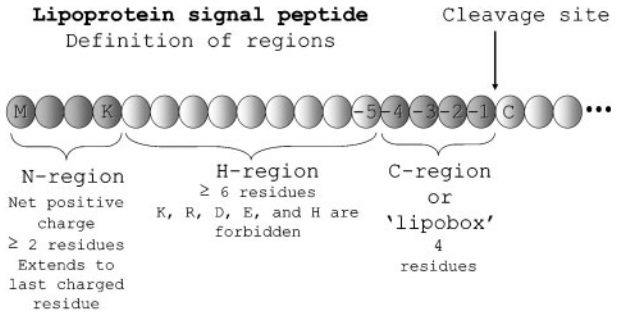

Fig. 1. Definition of spirochaetal lipoprotein signal peptide regions. Analysis of the TS and paralogous sequences of probable lipoprotein sequences was used to define signal peptide regions for use in the SpLip algorithm. The amino-terminal N-region extends from the start methionine to the last charged residue. The hydrophobic H-region extends from the last charged residue to the −5 position and must be at least 6 aa in length. The carboxy-terminal C-region, or lipobox, is 4 aa in length. At least one Leu, Phe, Val, Ile or Tyr must occur in positions −3 or −4. Position −1 must contain Ser, Asn, Ala, Gly, Cys, Thr or Gln. Charged residues (i.e. Lys, Arg, Asp, Glu and His) are forbidden in the H- and C-regions.

Matrix entries for which $F_{i,j}=0$ (the amino acid i was not observed in position j in the TS, but is not forbidden either) were changed to −2 [this is the largest integer less than the lowest value for $\log_2(F_{i,j}/G_{i,j})$ observed], except for Gln and Thr at position −1 or Met and Tyr at positions −3, −4 and −5; these residues received a score of zero at those positions. Matrix entries for which the corresponding amino acid is forbidden were set to −100.

C-region scoring. Given the WM, scoring was done in the standard way as follows. For each PPCG, all Cys residues in the first 50 residues are located, and for each Cys the four positions −1 to −4 are evaluated according to the WM. That is, if residue j is found at position i, it gets the score given by entry (i,j) in the WM. The sum of the scores for the four positions −1 to −4 is the C-region score. The putative lipobox is taken as the one with highest score (in case there is more than one Cys in the first 50 residues). If no score is positive, position −5 is also included in the analysis. If the highest score is still negative or zero, the PPCG is rejected.

H-region scoring. The H-region is defined as the region from position −5 (or −6, depending on the start position of the C-region) upstream of the putative lipobox to the position after the first charged residue defined as Lys, Arg, Asp, Glu or His. The Kyte–Doolittle hydrophobicity matrix (Kyte & Doolittle, 1982) was used to score the H-region. PPCGs with a negative H-region score or a length less than 6 residues are rejected.

N-region scoring. The SpLip algorithm calculates the net charge of the N-region according to the formula #Lys + #Arg + #His > #Asp + #Glu. PPCGs without a net positive charge in the N-region are rejected.

The final predictions are those that achieve positive scores for regions C, H and N. The pseudocode summarizing the SpLip algorithm is shown in Fig. 2.

Methodology of prediction evaluation. We compared the performance of the SpLip, Psort (Nakai & Horton, 1999) and LipoP (Juncker *et al.*, 2003) algorithms. The primary input for all three prediction algorithms was the complete set of PPCGs of *L. interrogans* sv. Copenhageni (Nascimento *et al.*, 2004), *L. interrogans* sv. Lai (Ren *et al.*, 2003), *B. burgdorferi* (Fraser *et al.*, 1997), *Borrelia garinii* (Glockner *et al.*, 2004), *T. pallidum* (Fraser *et al.*, 1998) and *Treponema denticola* (Seshadri *et al.*, 2004). Positive predictions in the case of Psort were those for which the output contained the words 'may be a lipoprotein'. Positive predictions in the case of LipoP were those with a positive score. In the case of SpLip, positive predictions included those with a positive score using rules for both probable and possible lipoproteins. The 'base set' comprised all PPCGs that were predicted to be a lipoprotein by at least one program. The base set of results was manually curated as 'true-positive' or 'false-positive' lipoproteins. We refer to the complete set of true-positive lipoproteins, including members of the TS, as the 'Liposet' for that organism. It should be noted that except for the TS, Liposet members identified in this paper have not been confirmed experimentally. True-positive lipoproteins in the base set that were not predicted to be lipoproteins by a given program were considered 'false-negative' sequences for that program. False-negative rates for each program and each genome were computed with the formula $\#fn/(\#fn+\#tp)$, where $\#fn$ is the number of false-negatives and $\#tp$ is the number of true-positives found by the program. The denominator $\#fn+\#tp$ corresponds to the size of the Liposet for that genome and it is the same number for all programs. Sensitivity is given by 1−false-negative rate. False-positive rates for each program and each genome were computed with the formula $\#fp/(\#fp+\#tp)$, where $\#fp$ corresponds to the number of false-positives for the given program and $\#tp$ is the same as above. Note that the denominator $\#fp+\#tp$ corresponds to all positive predictions of a given program and may be different for different programs. As an additional test of false-positive rates, all three algorithms were run on a set of 298 transmembrane cytoplasmic proteins of spirochaetes (known to be non-lipoproteins) available from SWISS-PROT (http://us.expasy.org) (Bairoch & Apweiler, 2000).

1. Find all cysteines (C-regions) in the first 50 residues
2. If no cysteines are found then **reject**
3. Discard C-regions with less than 50 residues downstream of +1
4. Score all remaining C-regions using WM
5. If no C-region has positive score then:
   a. include position −5
   b. re-score
   c. if no region has positive score then **reject**
6. Keep C-region with highest positive score
7. Score the H-region
8. If H-region has negative charge score then **reject**
9. Score the N-region
10. If N-region does not have net positive charge then **reject**
11. Output C-, H-, and N-region scores

Fig. 2. Pseudocode illustrating the SpLip algorithm. This pseudocode is executed for every PPCG in each of the six genomes.

RESULTS

The total number of true-positive lipoproteins for each genome is given in Table 2. Note that for *B. burgdorferi* and *T. pallidum* the number of lipoproteins listed is higher than what has been originally reported. For *B. burgdorferi*, 105

Table 2. Prediction results of all three programs

	L. interrogans sv. Copenhageni	*L. interrogans* sv. Lai	*B. burgdorferi*	*B. garinii*	*T. pallidum*	*T. denticola*
PPCGs	3660	4727*	1637	832	1031	2767
Total predictions (all 3 algorithms)	217	206	140	36	58	181
Liposet†	164	157	127	30	46	166
SpLip						
Probable	134	125	112	24	36	160
Possible	23	21	8	6	9	2
Total predictions	157	146	120	30	45	162
False-positive rate (%)‡	2 (1·3)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)
False-negative rate (%)‡	9 (5·5)	11 (7·0)	7 (5·5)	0 (0·0)	1 (2·2)	4 (2·4)
LipoP						
Predictions	163	149	116	25	31	151
False-positive rate (%)‡	47 (28·8)	44 (29·5)	13 (11·2)	5 (20·0)	8 (25·8)	12 (7·9)
False-negative rate (%)‡	48 (29·2)	52 (33·1)	24 (18·9)	10 (33·3)	23 (50·0)	27 (26·7)
Psort						
Predictions	47	46	26	5	22	48
False-positive rate (%)‡	10 (21·3)	12 (26·1)	1 (3·8)	1 (20·0)	6 (27·3)	3 (6·3)
False-negative rate (%)‡	127 (77·4)	123 (78·3)	102 (80·3)	26 (86·7)	30 (65·2)	121 (72·9)

*Errors in the *L. interrogans* sv. Lai genome annotation led to a major overestimation in the number of PPCGs.

†The Liposet is defined as the total number of true lipoproteins for each genome as determined by expert review of all lipoprotein predictions by the SpLip, LipoP, and Psort algorithms. Lipoproteins not predicted by any of the three algorithms would not have been captured using this method.

‡See Methods for definitions. The SpLip algorithm had the lowest false-positive and false-negative rates for all six genomes.

lipoproteins were originally reported (Fraser *et al.*, 1997); 127 were identified in the present study. For *T. pallidum*, 22 lipoproteins were previously reported (Fraser *et al.*, 1998); 46 were identified in the present study. To a large extent, these higher numbers reflect the greater sensitivity of the SpLip algorithm relative to previously available methods.

The *T. denticola* genome was found to encode 166 predicted lipoproteins, the largest number of any of the six spirochaete genomes. However, as a fraction of the total number of PPCGs, the *B. burgdorferi* sequence contains the highest percentage (7·8 %) of lipoproteins (Fig. 3). The relatively low percentage of lipoproteins in the *B. garinii* genome is due in large measure to the incomplete sequencing of the seven highly redundant lipoprotein-rich cp32 plasmids (Glockner *et al.*, 2004).

Leptospiral lipobox

The lipoboxes of the *L. interrogans* sv. Copenhageni Liposet were compared to those of experimentally confirmed *E. coli* lipoproteins (Gonnet *et al.*, 2004). Considerable differences were noted. As shown in Table 3, there is increased amino acid sequence variability in the leptospiral lipobox. For example, while Leu is the most common amino acid in both the −3 and −4 positions of the *E. coli* lipobox, Leu and Phe occurred with similar frequency at these positions in leptospiral lipoproteins. There was also a difference in the preferred amino acids, particularly at the −1 and −2 positions before Cys. Seventy-eight of 81 (96 %) aa at position −1 in *E. coli* lipoproteins are Ala or Gly. In contrast, the most common amino acids at position −1 of leptospiral lipoproteins are Ser and Asn. While the most common amino acid at position −2 in *E. coli* lipoproteins is Ala (29/81 = 36 %), seven other amino acids occur more frequently than Ala at this position in leptospiral lipoproteins.

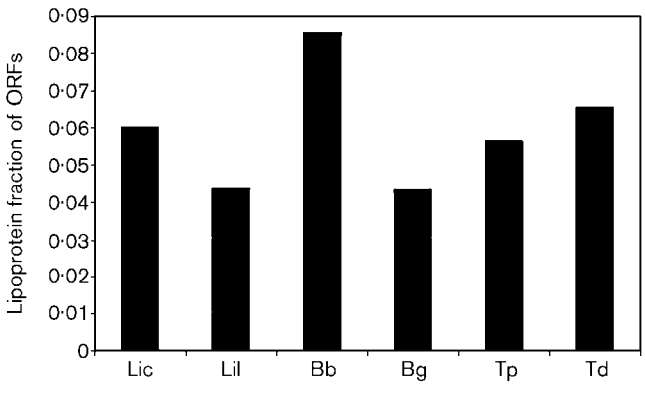

Fig. 3. Percentage of lipoprotein sequences in spirochaetal genomes. The total number of predicted lipoprotein sequences as a percentage of the total number of PPCGs was calculated for *L. interrogans* sv. Copenhageni (Lic), *L. interrogans* sv. Lai (Lil), *B. burgdorferi* (Bb), *B. garinii* (Bg), *T. pallidum* (Tp) and *T. denticola* (Td).

Table 3. Amino acid counts and frequencies (%) for the lipoboxes (C-regions) of the 164-member *L. interrogans* sv. Copenhageni Liposet identified in the current study and 81 experimentally confirmed *E. coli* lipoproteins (Gonnet *et al*., 2004)

Amino acids His, Arg, Asp, and Glu are not listed because they were not found to occur in any *L. interrogans* or *E. coli* lipoboxes.

Amino acid	164 *L. interrogans* sv. Copenhageni lipoproteins				81 *E. coli* lipoproteins			
	−4	−3	−2	−1	−4	−3	−2	−1
Phe	44 (26·8)	55 (33·5)	14 (8·5)	0 (0·0)	5 (6·2)	1 (1·2)	0 (0·0)	0 (0·0)
Ser	16 (9·8)	8 (4·9)	22 (13·4)	44 (26·8)	5 (6·2)	0 (0·0)	18 (22·2)	3 (3·7)
Tyr	0 (0·0)	1 (0·6)	5 (3·1)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)
Cys	3 (1·8)	0 (0·0)	0 (0·0)	2 (1·2)	0 (0·0)	1 (1·2)	2 (2·5)	0 (0·0)
Trp	2 (1·2)	0 (0·0)	4 (2·4)	0 (0·0)	1 (1·2)	0 (0·0)	0 (0·0)	0 (0·0)
Leu	46 (28·1)	53 (32·3)	22 (13·4)	0 (0·0)	38 (46·9)	67 (82·7)	0 (0·0)	0 (0·0)
Pro	1 (0·6)	1 (0·6)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)
Gln	0 (0·0)	2 (1·2)	9 (5·5)	10 (6·1)	0 (0·0)	0 (0·0)	2 (2·5)	0 (0·0)
Ile	12 (7·3)	16 (9·8)	23 (14·0)	0 (0·0)	2 (2·5)	5 (6·2)	2 (2·5)	0 (0·0)
Met	0 (0·0)	3 (1·8)	2 (1·2)	0 (0·0)	9 (11·1)	2 (2·5)	1 (1·2)	0 (0·0)
Thr	8 (4·9)	3 (1·8)	15 (9·2)	5 (3·1)	6 (7·4)	1 (1·2)	18 (22·2)	0 (0·0)
Asn	1 (0·6)	1 (0·6)	5 (3·1)	41 (25·0)	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)
Lys	0 (0·0)	0 (0·0)	0 (0·0)	0 (0·0)	1 (1·2)	0 (0·0)	0 (0·0)	0 (0·0)
Val	15 (9·2)	15 (9·2)	21 (12·8)	0 (0·0)	3 (3·7)	4 (4·9)	9 (11·1)	0 (0·0)
Ala	7 (4·3)	4 (2·4)	10 (6·1)	24 (14·6)	9 (11·1)	0 (0·0)	29 (35·8)	22 (27·2)
Gly	9 (5·5)	2 (1·2)	12 (7·3)	38 (22·6)	2 (2·5)	0 (0·0)	0 (0·0)	56 (69·1)

Sensitivity

As shown in Fig. 4, the SpLip and LipoP algorithms had much better sensitivity than Psort for identification of spirochaetal lipoprotein sequences. The primary reason for the poor sensitivity of Psort is the requirement of Ala or Gly in the −1 position, based on sequences of *E. coli* lipoproteins (von Heijne, 1989). Psort sensitivity ranged from 35 % in the case of *T. pallidum* to as low as 20 and 13 % for the *B. burgdorferi* and *B. garinii* genomes, respectively. These results indicate a higher diversity of lipobox amino acids in spirochaetal genomes. It should be noted that Psort failed to predict 12 out of 28 experimentally confirmed lipoprotein sequences in the SpLip TS (a false-negative rate of 43 % for members of the SpLip TS) and LipoP failed to predict one (false-negative rate of 3·6 % for members of the SpLip TS).

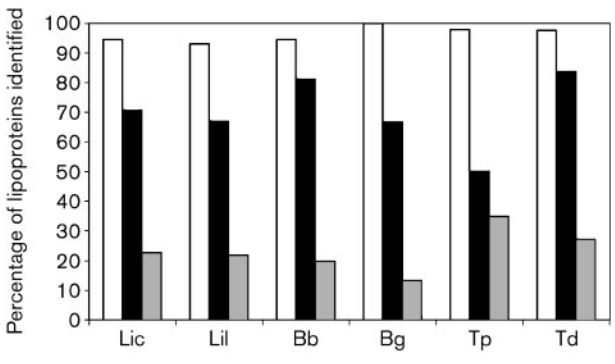

Fig. 4. Sensitivity of algorithms for spirochaetal lipoprotein sequences. Sensitivity (see Methods for definition) was calculated for SpLip (white bars), LipoP (black bars) and Psort (grey bars) for *L. interrogans* sv. Copenhageni (Lic), *L. interrogans* sv. Lai (Lil), *B. burgdorferi* (Bb), *B. garinii* (Bg), *T. pallidum* (Tp) and *T. denticola* (Td). For all three spirochaetal genomes, algorithm sensitivity was SpLip > LipoP > Psort.

False-positive predictions

The SpLip algorithm produced no false-positive lipoprotein predictions for any spirochaetal genome except for *L. interrogans* sv. Copenhageni (LIC). Two LIC sequences identified by SpLip as possible lipoproteins, DsbD and FlaB5, were judged to be false-positives because of their close homology with proteins that are known not to be lipoproteins. In *E. coli*, DsbD is a disulfide interchange protein located in the cytoplasmic membrane with nine transmembrane loops (Stewart *et al*., 1999). Rather than being conjugated with fatty acids, the Cys residue at position 30 is involved in the transfer of disulfide reducing potential from the cytoplasm to the periplasm. FlaB5 is one of five leptospiral FlaB flagellin proteins. All other known bacterial flagellin proteins are secreted without a signal peptide by the flagellar type III secretion system (Minamino & Namba, 2004). FlaB5 is unique in that it has a signal peptide, indicating sec-dependent secretion. For FlaB to be incorporated into the flagellar filament, FlaB should be secreted into the periplasm. Since lipoproteins generally remain membrane bound and are not secreted, we judged that FlaB5 is likely to be a false-positive lipoprotein prediction.

LipoP had a higher mean rate of false-positive lipoprotein sequences than either Psort or SpLip (Table 2). The percentage of false-positive LipoP predictions was highest for

leptospiral sequences. Psort false-positive predictions were slightly lower than those for LipoP. The percentage of false-positive Psort predictions was lowest for *B. burgdorferi* sequences. The LipoP and Psort algorithms tended to have different patterns of false-positive errors (see supplementary data available with the online version of this paper). The most frequent cause of LipoP errors was due to allowing unprecedented amino acids in the −1 position, frequently including charged amino acids Asp, Glu, Arg and Lys. Psort does not allow charged amino acids in the −1 position, but did allow them in other positions in the carboxy-terminal and hydrophobic regions of the lipoprotein signal peptide. Psort had a higher frequency than LipoP of unacceptably short hydrophobic regions and amino-terminal regions without a net positive charge.

As an additional control for false-positive predictions, we also tested all three algorithms using a set of 298 SWISS-PROT (Bairoch & Apweiler, 2000) spirochaetal proteins identified as having a cytoplasmic subcellular localization and that typically lack a signal peptide. The following results were obtained: SpLip had zero false-positive predictions, Psort had two and LipoP had seven. These results show that all three algorithms had low false-positive lipoprotein predictions when queried with cytoplasmic protein sequences.

Brachyspira lipoproteins

Brachyspira is a spirochaetal genus for which there are no complete genomes. However, there are reports of *Brachyspira* proteins with experimental evidence of lipidation. One of these, SmpA, is included in the SpLip TS (Table 1). In addition, published reports indicate that two additional proteins, MglB (Zhang *et al.*, 2000) of *Brachyspira pilosicoli* and BlpA (Cullen *et al.*, 2003a) of *Brachyspira hyodysenteriae* are also lipidated. Because *Leptospira* is the closest phylogenetic relative of *Brachyspira*, we tested the SpLip algorithm on the MglB and BlpA sequences using the *L. interrogans* sv. Copenhageni WM. SpLip correctly predicted that both MglB and BlpA are 'probable' lipoproteins with lipobox scores of 2·82 and 2·57, respectively. These results provide additional confirmatory evidence for the SpLip algorithm's accuracy.

DISCUSSION

We have described a novel algorithm, designated SpLip, for the prediction of spirochaetal lipoproteins. SpLip uses a hybrid approach incorporating both lipoprotein signal peptide rules and a statistical WM based on a TS consisting of experimentally verified spirochaetal lipoproteins. The SpLip algorithm appeared to be more accurate than either the rules-based Psort algorithm or the more general hidden Markov model approach represented by the LipoP algorithm. The rules governing the Psort algorithm do not take into account the higher variability of the spirochaetal lipobox. LipoP has improved sensitivity for detection of spirochaetal lipoproteins compared to Psort, but allows charged amino acids in the lipobox, which is not consistent with a lipoprotein sequence. Analysis of all available spirochaetal genome sequences using the SpLip algorithm made it possible to produce an accurate and virtually complete set of lipoproteins for six spirochaetes, including *L. interrogans* sv. Copenhageni, *L. interrogans* sv. Lai, *B. burgdorferi*, *B. garinii*, *T. pallidum* and *T. denticola*. *T. denticola* had 166 predicted lipoproteins, the largest number of lipoproteins for any of the six organisms. When the sequences of its plasmids are included, 7·8% of *B. burgdorferi* PPCGs encoded predicted lipoproteins, the highest fraction of PPCGs of any organism that we are aware of. Although *Brachyspira* genome sequences are not yet available and strict application of the SpLip methodology requires the complete PPCG set for each genome to be tested, we found that the SpLip algorithm correctly classified the sequences of two experimentally confirmed *Brachyspira* lipoprotein sequences, MglB and BlpA.

The SpLip algorithm is supplemented with rules based both on precedent and on an understanding of the biochemistry of lipoprotein signal peptides. Aside from Cys in the +1 position and the start Met, the most constrained position in the lipoprotein signal peptide is the −1 position. The structural constraints on the −1 position are largely imposed by the substrate specificity of the diacylglyceryl transferase, which transfers the initial lipid to Cys and to a lesser extent by the lipoprotein signal peptidase, which removes the signal peptide from the preprotein (Paetzel *et al.*, 2002). In many bacteria, the −1 position consists exclusively of the small non-polar amino acids Ala or Gly, as reflected in the von Heijne consensus pattern (von Heijne, 1989). In spirochaetes, a high percentage of lipoprotein signal peptides contain Ser at the −1 position, which largely accounts for the failure of Psort to correctly predict 12 of the 28 sequences in the SpLip TS of experimentally verified lipoproteins. The lipid modification of spirochaetal lipoproteins with Ser at −1 has been well documented experimentally in 9/28 sequences in the SpLip TS, including *B. burgdorferi* proteins OspC, OspD and the decorin-binding protein, DbpA. Lipoproteins with Ser at −1 are uncommon in non-spirochaetes, but examples do exist, including MltC and YddW of *E. coli* (Gonnet *et al.*, 2004), β-lactamase III of *Bacillus cereus* and VirB7 of *Agrobacterium tumefaciens*.

In contrast to Ser, the occurrence of Asn and Cys in the −1 position of lipoproteins may be unique to spirochaetes. Asn occurs in the −1 position in the leptospiral lipoprotein LipL41 (Shang *et al.*, 1996). The experimental evidence for lipidation of LipL41 includes ^{3}H-palmitate intrinsic labelling studies and inhibition of labelling with globomycin. Cys occurs in the −1 position in the *B. burgdorferi* oligopeptide-binding protein OppA-2, which has been shown to be a lipoprotein by intrinsic labelling with ^{3}H-palmitate (Kornacki & Oliver, 1998). The SpLip algorithm was initially run allowing only Ala, Gly, Ser, Asn and Cys in the −1 position. When the algorithm was re-run allowing Thr and Gln (conservative amino substitutions for Ser and Asn, respectively) in the −1 position, a number of

additional lipoproteins were identified that would have otherwise received an unfavourable score. Because there is no experimental evidence available at this time to confirm that spirochaetal proteins with Thr or Gln in the −1 position would be lipidated, we refer to predicted lipoproteins with Thr or Gln in the −1 position as 'possible' rather than 'probable' lipoproteins.

In contrast to the conservative amino acid substitutions allowed by the SpLip algorithm, the lipoproteins predicted by the LipoP algorithm contain a wide variety of amino acids in the −1 position. We observed that the LipoP algorithm predicts as lipoprotein sequences that are lipoprotein-like at other positions but have unprecedented amino acids at the −1 position. When we analysed spirochaetal genomes with the LipoP algorithm, a large number of lipoproteins were predicted with charged amino acids (Asp, Glu, Lys, Arg or His) in the −1 position and elsewhere in the hydrophobic or carboxy-terminal regions of the signal peptide of spirochaetal lipoproteins (see supplementary data available with the online version of this paper). We interpret such sequences as false-positive lipoprotein predictions because there is no precedent or justification that we are aware of for charged amino acids to occur in regions other than the amino-terminal portion of the lipoprotein signal peptide. Another measure of the inaccuracy of the LipoP algorithm is the finding that 7/298 spirochaetal sequences designated cytoplasmic membrane proteins by the SWISS-PROT database are predicted by LipoP to be lipoproteins. In contrast, SpLip correctly rejected all 298 of the SWISS-PROT cytoplasmic membrane protein sequences.

There were important differences between the TSs used to construct the models used for scoring in the LipoP and SpLip algorithms. Only 17/63 lipoproteins from the LipoP TS are spirochaetal in origin. Because there is evidence of substantial differences in the preferred amino acids in the lipobox of spirochaetes relative to other bacteria, the non-spirochaetal lipoproteins in the LipoP algorithm TS would reduce the accuracy of the LipoP algorithm for correctly scoring spirochaetal lipoproteins. LipoP incorrectly predicts that one of the sequences in the SpLip TS, the 17-kDa TpN17 *T. pallidum* lipoprotein, has a signal peptidase 1 cleavage site rather than a lipoprotein signal peptidase II cleavage site. TpN17 is included in the SpLip TS because of intrinsic labelling studies in *T. pallidum* have demonstrated lipidation (Akins *et al.*, 1993). Another problem is that the LipoP TS was obtained by searching the SWISS-PROT database for 'probable' or 'potential' lipoproteins. Consequently, many of the 17 spirochaetal lipoproteins in the LipoP TS, such as the *Borrelia* Bmp proteins, lack experimental evidence of lipidation. In contrast, the SpLip algorithm relied exclusively on lipoprotein sequences for which there was experimental evidence of lipidation. Homologous lipoprotein sequences were excluded from the LipoP TS but retained by the SpLip TS. For example, the *B. burgdorferi* oligopeptide-binding proteins OppA-1 and OppA-3 share the same lipobox sequence, SLIAC. One justification for retaining both sequences in the TS is that the signal peptide sequences of these homologous proteins are otherwise dissimilar. Another justification is that because the SpLip TS is large enough to be a representative sampling of all spirochaetal lipoproteins, sequences that occur more frequently should be given more weight in the WM.

In summary, we have developed a novel lipoprotein prediction algorithm that is a hybrid approach using a combination of rules based on a biochemical knowledge of lipoprotein signal peptides and a statistical WM based on a lipoprotein sequence TS. Application of the SpLip algorithm to six spirochaetal genome sequences resulted in a more accurate set of predicted lipoproteins, with significantly improved sensitivity and specificity than either of the previously existing programs. The lipoprotein databases provided by this study will be useful in the search for spirochaetal virulence factors and vaccine candidates. In addition, the SpLip program will be useful for analysis of emerging spirochaetal genome sequence data. More importantly, we believe that our hybrid approach, by taking advantage of the strengths of rules grounded in biochemistry and statistical empiricism, can be generalized not only to lipoprotein identification in non-spirochaetal bacteria but also to other types of sequence prediction algorithms.

SpLip program

The SpLip program was encoded in perl, runs on any operating system that supports the perl interpreter and is available upon request by contacting the authors.

ACKNOWLEDGEMENTS

This work was supported in part by VA Medical Research Funds (to J. M. and D. A. H.); Public Health Service grant AI-34431 (to D. A. H.) from the National Institute of Allergy and Infectious Diseases; FAPESP (Brazil) *Leptospira* grant (to J. C. S.) and fellowship (to M. S. R.); and CNPq (Brazil) fellowship (to J. C. S.).

REFERENCES

Akins, D. R., Purcell, B. K., Mitra, M. M., Norgard, M. V. & Radolf, J. D. (1993). Lipid modification of the 17-kilodalton membrane immunogen of *Treponema pallidum* determines macrophage activation as well as amphiphilicity. *Infect Immun* **61**, 1202–1210.

Aliprantis, A. O., Yang, R. B., Mark, M. R., Suggett, S., Devaux, B., Radolf, J. D., Klimpel, G. R., Godowski, P. & Zychlinsky, A. (1999). Cell activation and apoptosis by bacterial lipoproteins through toll-like receptor-2. *Science* **285**, 736–739.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48.

Barnett, J. K., Barnett, D., Bolin, C. A., Summers, T. A., Wagar, E. A., Cheville, N. F., Hartskeerl, R. A. & Haake, D. A. (1999). Expression and distribution of leptospiral outer membrane components during renal infection of hamsters. *Infect Immun* **67**, 853–861.

Braun, V. & Wolff, H. (1970). The murein-lipoprotein linkage in the cell wall of *Escherichia coli*. *Eur J Biochem* **14**, 387–391.

Brightbill, H. D., Libraty, D. H., Krutzik, S. R. & 11 other authors (1999). Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. *Science* **285**, 732–736.

Chamberlain, N. R., Radolf, J. D., Hsu, P. L., Sell, S. & Norgard, M. V. (1988). Genetic and physicochemical characterization of the recombinant DNA-derived 47-kilodalton surface immunogen of *Treponema pallidum* subsp. *pallidum*. *Infect Immun* **56**, 71–78.

Cullen, P. A., Coutts, S. A., Cordwell, S. J., Bulach, D. M. & Adler, B. (2003a). Characterization of a locus encoding four paralogous outer membrane lipoproteins of *Brachyspira hyodysenteriae*. *Microbes Infect* **5**, 275–283.

Cullen, P. A., Haake, D. A., Bulach, D. M., Zuerner, R. L. & Adler, B. (2003b). LipL21 is a novel surface-exposed lipoprotein of pathogenic *Leptospira* species. *Infect Immun* **71**, 2414–2421.

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

Fraser, C. M., Casjens, S., Huang, W. M. & 35 other authors (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586.

Fraser, C. M., Norris, S. J., Weinstock, G. M. & 29 other authors (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388.

Glockner, G., Lehmann, R., Romualdi, A., Pradella, S., Schulte-Spechtel, U., Schilhabel, M., Wilske, B., Suhnel, J. & Platzer, M. (2004). Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res* **32**, 6038–6046.

Gonnet, P., Rudd, K. E. & Lisacek, F. (2004). Fine-tuning the prediction of sequences cleaved by signal peptidase II: a curated set of proven and predicted lipoproteins of *Escherichia coli* K-12. *Proteomics* **4**, 1597–1613.

Guo, B. P., Brown, E. L., Dorward, D. W., Rosenberg, L. C. & Hook, M. (1998). Decorin-binding adhesins from *Borrelia burgdorferi*. *Mol Microbiol* **30**, 711–723.

Haake, D. A. (2000). Spirochaetal lipoproteins and pathogenesis. *Microbiology* **146**, 1491–1504.

Haake, D. A., Chao, G., Zuerner, R. L., Barnett, J. K., Barnett, D., Mazel, M., Matsunaga, J., Levett, P. N. & Bolin, C. A. (2000). The leptospiral major outer membrane protein LipL32 is a lipoprotein expressed during mammalian infection. *Infect Immun* **68**, 2276–2285.

Hellwage, J., Meri, T., Heikkila, T., Alitalo, A., Panelius, J., Lahdenne, P., Seppala, I. J. & Meri, S. (2001). The complement regulator factor H binds to the surface protein OspE of *Borrelia burgdorferi*. *J Biol Chem* **276**, 8427–8435.

Howe, T. R., Mayer, L. W. & Barbour, A. G. (1985). A single recombinant plasmid expressing two major outer surface proteins of the Lyme disease spirochete. *Science* **227**, 645–646.

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 1652–1662.

Kornacki, J. A. & Oliver, D. B. (1998). Lyme-disease-causing *Borrelia* species encode multiple lipoproteins homologous to peptide-binding proteins of ABC-type transporters. *Infect Immun* **66**, 4115–4122.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132.

Madan Babu, M. & Sankaran, K. (2002). DOLOP – database of bacterial lipoproteins. *Bioinformatics* **18**, 641–643.

Matsunaga, J., Barocchi, M. A., Croda, J. & 8 other authors (2003). Pathogenic *Leptospira* species express surface-exposed proteins belonging to the bacterial immunoglobulin superfamily. *Mol Microbiol* **49**, 929–945.

Minamino, T. & Namba, K. (2004). Self-assembly and type III protein export of the bacterial flagellum. *J Mol Microbiol Biotechnol* **7**, 5–17.

Mount, D. W. (2001). *Bioinformatics: Sequence and Genome Analysis*: Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

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

Nascimento, A. L., Ko, A. I., Martins, E. A. & 43 other authors (2004). Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol* **186**, 2164–2172.

Paetzel, M., Karla, A., Strynadka, N. C. & Dalbey, R. E. (2002). Signal peptidases. *Chem Rev* **102**, 4549–4580.

Probert, W. S. & Johnson, B. J. (1998). Identification of a 47 kDa fibronectin-binding protein expressed by *Borrelia burgdorferi* isolate B31. *Mol Microbiol* **30**, 1003–1015.

Ren, S. X., Fu, G., Jiang, X. G. & 36 other authors (2003). Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature* **422**, 888–893.

Schwan, T. G., Piesman, J., Golde, W. T., Dolan, M. C. & Rosa, P. A. (1995). Induction of an outer surface protein on *Borrelia burgdorferi* during tick feeding. *Proc Natl Acad Sci U S A* **92**, 2909–2913.

Seshadri, R., Myers, G. S., Tettelin, H. & 36 other authors (2004). Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc Natl Acad Sci U S A* **101**, 5646–5651.

Shang, E. S., Summers, T. A. & Haake, D. A. (1996). Molecular cloning and sequence analysis of the gene encoding LipL41, a surface-exposed lipoprotein of pathogenic *Leptospira* species. *Infect Immun* **64**, 2322–2330.

Stewart, E. J., Katzen, F. & Beckwith, J. (1999). Six conserved cysteines of the membrane protein DsbD are required for the transfer of electrons from the cytoplasm to the periplasm of *Escherichia coli*. *EMBO J* **18**, 5963–5971.

von Heijne, G. (1989). The structure of signal peptides from bacterial lipoproteins. *Protein Eng* **2**, 531–534.

Zhang, P., Cheng, X. & Duhamel, G. E. (2000). Cloning and DNA sequence analysis of an immunogenic glucose-galactose MglB lipoprotein homologue from *Brachyspira pilosicoli*, the agent of colonic spirochetosis. *Infect Immun* **68**, 4559–4565.