World J Microbiol Biotechnol (2008) 24:2377–2382
DOI 10.1007/s11274-008-9795-2

REVIEW

Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria

Obaidur Rahman · Stephen P. Cummings · Dean J. Harrington · Iain C. Sutcliffe

Received: 30 April 2008 / Accepted: 15 June 2008 / Published online: 27 June 2008

Abstract Bacterial lipoproteins are a diverse and functionally important group of proteins that are amenable to bioinformatic analyses because of their unique signal peptide features. Here we have used a dataset of sequences of experimentally verified lipoproteins of Gram-positive bacteria to refine our previously described lipoprotein recognition pattern (G+LPP). Sequenced bacterial genomes can be screened for putative lipoproteins using the G+LPP pattern. The sequences identified can then be validated using online tools for lipoprotein sequence identification. We have used our protein sequence datasets to evaluate six online tools for efficacy of lipoprotein sequence identification. Our analyses demonstrate that LipoP (http://www.cbs.dtu.dk/services/LipoP/) performs best individually but that a consensus approach, incorporating outputs from predictors of general signal peptide properties, is most informative.

Keywords Lipoproteins · Signal peptides · Bioinformatics · Genomics · *Firmicutes* · Actinobacteria

Electronic supplementary material The online version of this article (doi:10.1007/s11274-008-9795-2) contains supplementary material, which is available to authorized users.

O. Rahman · S. P. Cummings · I. C. Sutcliffe
Northumbria University, Newcastle upon Tyne NE1 8ST, UK

D. J. Harrington
University of Bradford, West Yorkshire BD7 1DP, UK

I. C. Sutcliffe (✉)
Biomolecular and Biomedical Research Centre, School of Applied Science, Northumbria University, Newcastle upon Tyne NE1 8ST, UK
e-mail: iain.sutcliffe@unn.ac.uk

Introduction

Bacterial lipoproteins (Lpp) are a functionally diverse class of membrane anchored proteins that typically represent ca. 2% of the bacterial proteome (Sutcliffe and Harrington 2002; Sutcliffe and Harrington 2004; Babu et al. 2006; Sutcliffe and Hutchings), although in some taxa the proportion is even higher (Bendtsen et al. 20052007; Setubal et al. 2006). Lpp are of particular significance in Gram-positive bacteria as, in the absence of an outer membrane, various proteins must be tethered to the plasma membrane in order to be retained within the cell envelope. Thus many Lpp of Gram-positive bacteria have functions directly comparable to those of periplasmic or surface proteins in Gram-negative bacteria. For example, the substrate binding proteins which deliver substrates to the integral membrane components of ABC importer systems are typically Lpp in Gram-positive bacteria and periplasmic proteins in Gram-negative bacteria (Sutcliffe and Russell 1995). Consequently, many of the known or predicted functions of Gram-positive bacterial Lpp reflect their predicted localisation at the interface between the cell membrane and the extracytoplasmic compartment. Thus, in addition to the well defined category of substrate binding Lpp, a brief selection of Lpp functions include roles as enzymes; in sensing environmental cues; in membrane-associated redox processes; and in correct protein export and localisation (Sutcliffe and Russell 1995; Sutcliffe and Harrington 2004; Sutcliffe and Hutchings 2007). This functional versatility means that it is extremely useful to be able to identify putative Lpp in order to gain further insights into the biology of biotechnologically and medically significant organisms. Moreover, the accurate prediction of protein localisation by sequence analysis is clearly an important aspect of genome annotation and, eventually, understanding of protein function (Gardy and Brinkman 2006).

Bacterial Lpp are anchored to cellular membrane(s) as the result of their post-translational modification with, as a minimum, a diacylglyceride group which is added to an essential cysteine. This cysteine is located in the C-terminal region of a signal peptide that directs precursor-Lpp translocation across the plasma membrane prior to lipid modification (Braun and Wu 1994; Sutcliffe and Harrington 2002). The stretch of amino acids preceding the cysteine is relatively well conserved (the 'lipobox') and this means that, in combination with the recognition of other conserved signal peptide features (Fig. 1), Lpp are highly amenable to identification by bioinformatic analyses. However, there is evidence for subtle taxon-specific differences in the signal peptide features of Lpp from different bacterial taxa (Setubal et al. 2006; Sutcliffe and Harrington 2002). In order to refine the methods for the bioinformatic analysis of Lpp from Gram-positive bacteria, we have curated a true positive (TP) dataset of 90 experimentally proven Lpp and a true negative (TN) dataset of sequences not considered to be Lpp. These datasets have been used to test the performance of several online applications in accurately identifying Gram-positive bacterial Lpp.

Screening Gram-positive bacterial genomes for putative Lpp

The conserved signal peptide and lipobox features of bacterial Lpp can be expressed in regular sequence patterns. Following the work of Klein et al. (1988) and von Heijne (1989), the Prosite profile PS51257 (formerly Prosite pattern PS00013) was defined to allow bacterial Lpp sequences to be recognised. Subsequently, we refined the pattern search approach and defined a pattern, denoted G+LPP, with greater accuracy (higher specificity) for the recognition of Lpp from Gram-positive bacteria (Sutcliffe and Harrington 2002; Table 1). Both the Prosite profile

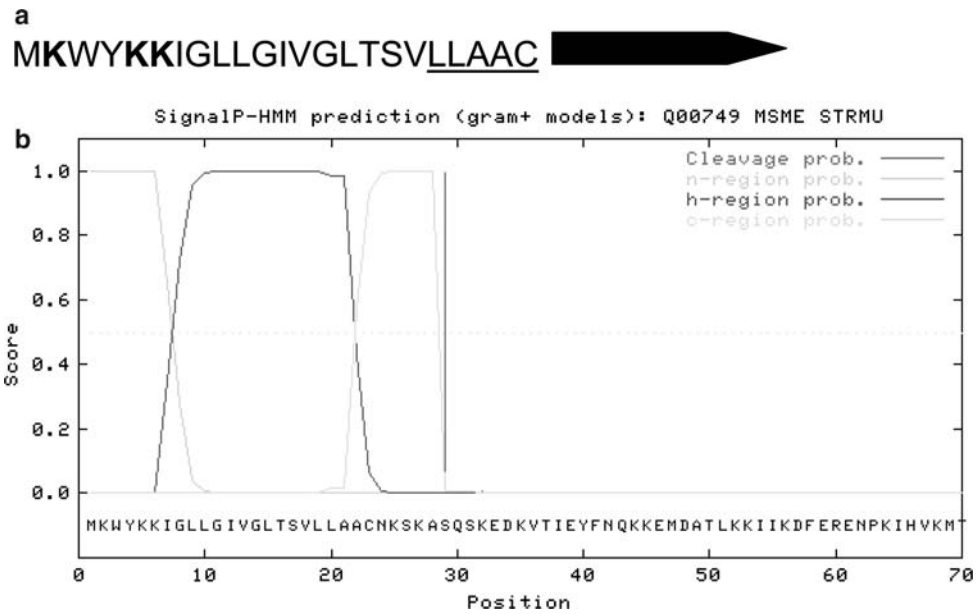

Fig. 1 Signal peptide features of a typical bacterial Lpp. (**a**) The sequence shown is that of the substrate binding protein MsmE of *Streptococcus mutans*, an experimentally verified Lpp (Sutcliffe et al. 1993). The positively charged N-region amino acids are shown in bold followed by the hydrophobic H-region. The lipoprotein specific lipobox, culminating in the crucial cysteine, is underlined. The arrow represents the mature protein sequence. (**b**) Output from SignalP-HMM for MsmE, demonstrating how this tool typically predicts the signal peptide H-region of bacterial Lpp to end in close proximity to the lipobox cysteine

Table 1 Refined G+LPP pattern expression for the identification of Gram-positive bacterial lipoproteins. The original G+LPP pattern, written in Prosite syntax, was described by analysis of the signal peptide features of 33 experimentally verified lipoproteins (Sutcliffe and Harrington 2002)

Pattern	Pattern expression
G+LPP	<[MV]-X(0,13)-[RK]-{DERKQ}(6,20)-[LIVMFESTAG]-[LVIAM]-[IVMSTAF[a]G]-[AG]-C
G+LPPv2	<[MV]-X(0,13)-[RK]-{DERK}(6,20)-[LIVMFESTAG**PC**]-[LVIAM**FTG**]-[IVMSTAG**CP**]-[AG**S**]-C
PS51257[b]	{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C

[a] F was incorrectly included as a permissible residue in the -2 position when the original pattern was described

[b] Additional rules apply i.e. that there must be a K or R in the first seven amino acids and that the cysteine must appear between amino acids 15 and 35

The extended TP dataset reported here has allowed us to refine the G+LPP pattern to create G+LPPv2. Newly recognised permissible amino acids in G+LPPv2 are emphasised in bold font. In addition to the cysteine, the amino acid positions -4 to -1 are those typically described as the 'lipobox'. The Prosite profile P51257 which is notably more relaxed in the -2 and -3 positions is also shown

Table 2 Online tools used for the identification of Gram-positive bacterial lipoproteins

Tool	Utility	Methodology	URL	Reference
DOLOP	Lpp[a]	Pattern matching	http://www.mrc-lmb.cam.ac.uk/genomes/dolop/	Babu and Sankaran (2002)
LIPO	Lpp	Pattern matching	http://services.cbu.uib.no/tools/lipo	Berven et al. (2006)
LipoP	Lpp	Hidden Markov Model	http://www.cbs.dtu.dk/services/LipoP/	Juncker et al. (2003)
LipPred	Lpp	Naive-Bayesian network	http://www.jenner.ac.uk/LipPred/	Taylor et al. (2006)
Phobius	SP[b]	Hidden Markov Model	http://phobius.sbc.su.se/	Käll et al. (2004)
PSORT	Prediction of protein localisation	Pattern matching	http://psort.ims.u-tokyo.ac.jp/form.html	Nakai and Horton (1999)
SignalP-HMM	SP	Hidden Markov Model	http://www.cbs.dtu.dk/services/SignalP/	Nielsen and Krogh (1998), Bendtsen et al. (2004)
ScanProsite	Pattern matching	Pattern matching	http://us.expasy.org/tools/scanprosite/	De Castro et al. (2006)
SPEPLip	Lpp	Neural network	http://gpcr.biocomp.unibo.it/cgi/predictors/spep/pred_spepcgi.cgi	Fariselli et al. (2003)

[a] Lpp, tool specifically optimised for lipoprotein identification

[b] SP, tool optimised for secretory protein signal peptide prediction

PS51257 and the G+LPP pattern are useful as the major protein sequence databases (UniProt, TrEMBL; Bairoch et al. 2005) can be screened for matching sequences using the ScanProsite tool (De Castro et al. 2006; Table 2). These screens can be restricted to identify matches from particular bacterial taxa at the species or higher taxonomic levels. Thus specified bacterial genomes can be rapidly mined to identify putative Lpp (for examples see Sutcliffe and Harrington 2004; Sutcliffe and Hutchings 2007). However, these datasets have to be analysed further to eliminate potential false-positives which have coincidentally matched these sequence patterns (see below).

The G+LPP pattern was based on an analysis of the signal peptide features of 33 experimentally verified Lpp (Sutcliffe and Harrington 2002). To further refine the G+LPP pattern, additional experimentaly verified Lpp were identified by extensive literature surveys. Proteins were considered experimentaly verified Lpp if they satisfied the previously described criteria (Sutcliffe and Harrington 2002) or, additionally, if aberrant protein processing has been demonstrated in mutants of the Lpp biosynthetic machinery (for good examples see Réglier-Poupet et al. 2003; Baumgärtner et al. 2006). These allowed us to define an extended TP dataset (Supplementary Table 1) containing 90 experimentally verified Lpp from Gram-positive bacteria. Analysis of the signal peptide features of the additional proteins in this dataset has allowed us to refine the G+LPP pattern to yield G+LPPv2 (Table 1). Use of the revised pattern will improve the sensitivity of future pattern searches of whole genomes. The performance metrics of the G+LPPv2 pattern are evaluated below.

Evaluation of online tools for the identification of putative Lpp.

Several feature-based online tools are available for the prediction of whether bacterial protein sequences are likely to be Lpp (Table 2). All are readily accessible and easy to use. Thus these tools can be used to evaluate the sequences derived from genome sequence screens to allow the elimination of predicted false positives (typically ca. 10% of the sequences recovered by pattern searches). However, to our knowledge, the comparative performance of these tools has not yet been evaluated. As two of the prediction tools for Lpp identification are trained to identify Lpp sequences from Gram-negative bacteria (LipoP and LIPO; Table 2), whereas others are more generic, we wished to evaluate which applications performed best in identifying putative Lpp encoded in Gram-positive bacterial genomes. To complement the TP dataset described above, a TN dataset (Supplementary Table 2) was constructed from sequences with 'coincidentally' placed N-terminal cysteine residues, such as integral membrane proteins or cytoplasmic proteins identified as false positives (FP) in previous genome analyses (for example Sutcliffe and Harrington 2002; Sutcliffe and Harrington 2004; Sutcliffe and Hutchings 2007). DOLOP, LIPO, LipoP, LipPred and SpepLIP were used with their default settings. PSORT, an integrated tool for prediction of protein localisation (Nakai and Horton 1999) that allows assessment of whether bacterial proteins sequences are putative Lpp based on the rules described by von Heijne (1989), was used in its original version (Table 2) that allows Gram-positive bacteria to be selected as the source of input sequence. Correct predictions were

Table 3 Performance evaluation metrics for the refined G+LPP pattern expression and five tools for the prediction of lipoprotein signal peptides

	True positive dataset		True negative dataset		Sensitivity[a]	Specificity[b]	Overall performance[c]
	TP	FN	TN	FP			
G+LPPv2	90	0	23[d]	11[d]	1.000	0.891	0.911
DOLOP	70	20	30	4	0.778	0.946	0.806
LIPO	85[e]	4	12	22	0.955	0.794	0.789
LipoP	86	4	32	2	0.956	0.977	0.952
LipPred	86	4	12[f]	12[f]	0.956	0.878	0.860
PSORT	81	9	30	4	0.900	0.953	0.895
SPEPLip	82	8	24	10	0.911	0.891	0.855

[a] Sensitivity (recall) is calculated as TP/(TP+FN)

[b] Specificity (precision) is calculated as TP/(TP+FP)

[c] Overall performance is calculated as (TP+TN)/(TP+FN+TN+FP)

[d] G+LPPv2 was manually matched to the signal peptide features of each sequence in the TN dataset

[e] LIPO was unable to give a prediction with one sequence with two potential lipobox cysteines

[f] LipPred generated no-prediction with 10/34 sequences

scored as either TP or TN. Incorrect predictions were scored as either false positive (FP) or false negative (FN). Specificity (precision) and sensitivity (recall) were calculated as described by Gardy and Brinkman (2006). SignalP-HMM (Nielsen and Krogh 1998; Bendtsen et al. 2004) was used with Gram-positive bacteria as the selected organism group and the graphic outputs visually inspected to determine the end of the signal peptide H-region, which is expected to be congruent with the lipobox cysteine in putative Lpp sequences (our unpublished observations; Fig. 1 and see below). Signal peptides were also predicted using Phobius (Käll et al. 2004) and the end position of the h-region noted.

Outputs from each of the prediction servers under evaluation for each dataset are summarised in Supplementary Tables 1 and 2 and the performance metrics are summarised in Table 3. When considering sensitivity, three of the online tools (LipoP, LIPO and LipPred) performed with >95% sensitivity, whereas DOLOP, PSORT and SPEPlip were less effective. DOLOP performed better when considering specificity and its sensitivity reflects a rather restricted group of permitted amino acids at the -3 position relative to the lipobox cysteine. The direct derivation of the revised G+LPPv2 pattern from the TP dataset accounts for its 100% sensitivity value (Table 3) and although it is less specific, its overall performance is still very good (0.911). Despite being trained on target sequences from Gram-negative bacteria, the Hidden Markov Model-based tool LipoP was the single best performing tool and the only online application with an overall accuracy of >90%. The sensitivity (0.955) of LipoP was improved compared to that previously reported (0.929) by the authors for a smaller dataset of Lpp from Gram-positive bacteria (Juncker et al. 2003). LipoP has been previously reported to have a sensitivity of 0.964 when tested against a dataset of 28 experimentally verified spirochaetal Lpp (Setubal et al. 2006). The specificity of LipoP is also impressive (Table 3) given that our TN dataset contains only 'confounding' sequences with an appropriately placed cysteine, rather than simply being a dataset of proteins known to be non-Lpp (such as the integral membrane protein dataset used by Setubal et al. [2006]) which may lack the crucial cysteine residue. Cumulatively these data confirm that LipoP is a highly sensitive and specific tool for the validation of Lpp from bacteria from a variety of phylogenetic lineages. An additional advantage of this tool is that it also presents a second best prediction that can be informative when the margin between the best and second best scores is low.

The present analysis confirms our previous observation (Sutcliffe and Harrington 2002) that it is also informative to relate the outputs from the Lpp specific tools to signal peptide analyses using the SignalP HMM ouput (Nielsen and Krough 1998) and Phobius (Supplementary Tables 1 and 2). Thus for 82/86 (95.3%) of the TP where signal peptides were predicted by SignalP-HMM, the hydrophobic h-region was predicted to end within 3 amino acids of the lipobox cysteine (Fig. 1b). This clearly reflects the hydrophobic nature of the typical lipobox amino acids (Table 1; Fig. 1). Moreover, the mean cleavage site probability (0.635 ± 0.203, $n = 86$; range 0.239–0.999) for the TP dataset was notably lower than that expected for Type I (secretory) signal peptides, reflecting the processing of Lpp by a lipoprotein specific (type II) signal peptidase (Tjalsma et al. 1999) rather than Type I signal peptidases. A similar, novel observation from the present study is that the end of the h-region as predicted by Phobius falls before the

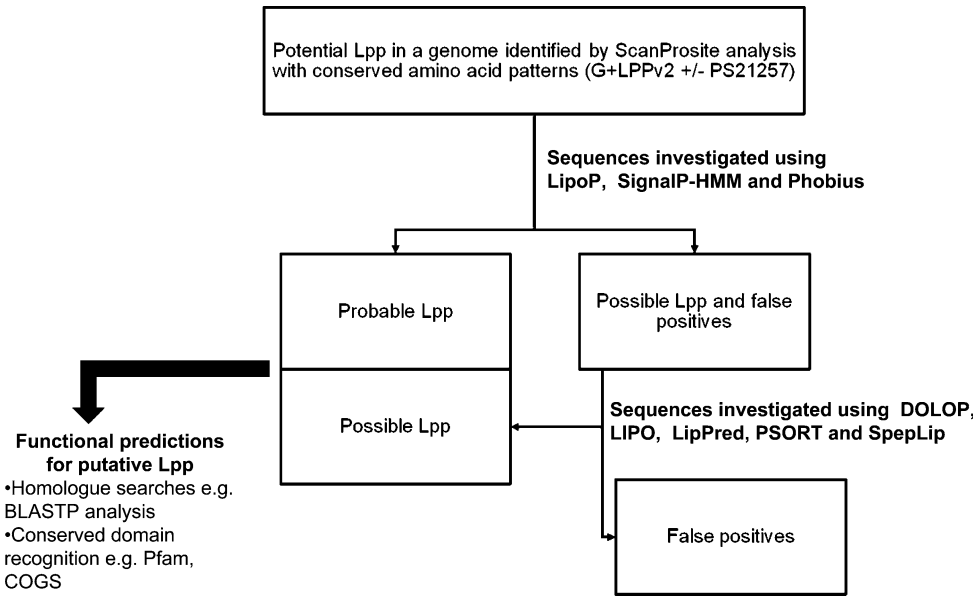

Fig. 2 Schematic protocol for the recovery and evaluation of putative Lpp sequences from Gram-positive bacterial genomes. The online tools recommended are described in the text. Individual protein sequences can also be analysed by the same strategy, including visual inspection for match to the G+LPPv2 pattern

lipobox cysteine for 89/89 (100%) predicted signal peptides in the TP dataset.

A consensus strategy to identify Gram-positive bacterial Lpp.

To our knowledge, this is the first critical evaluation of the performance of the full range of online tools for bacterial lipoprotein identification. Cumulatively these analyses suggest LipoP is the best performing single tool for predicting Lpp sequences of Gram-positive bacteria. However, we recommend an integrated approach (Fig. 2) wherein genomes are first screened using ScanProsite and the G+LPPv2 pattern. To minimise false negatives, genomes can also be rapidly scanned again using the PS51527 profile to detect any additional sequences retrieved. These two pattern search approaches are highly efficient at yielding a starter dataset for further analysis. However, the fact that the PS51527 profile is less specific than G+LPPv2 should be taken into account when validating sequences recovered with this pattern only. The datasets retrieved can then be re-assessed by analysis with a range of tools, including both those specific for the identification of Lpp signal peptides and those which can be used to identify general signal peptide features (SignalP and Phobius). As a minimum we recommend analysing sequences with LipoP, SignalP and Phobius (Fig. 2), the latter being useful to include as it also predicts transmembrane domains in integral membrane proteins. Sequences should be considered putative Lpp if they are scored 'SPII' (type II signal peptide) by LipoP and also have clearly predicted signal peptides wherein the h-region is predicted to end in close proximity to or prior to the lipobox cysteine. Sequences that are not identified as putative Lpp by LipoP should be reinvestigated by using the remaining five tools and a consensus drawn from these outputs in conjuction with the "2nd best" prediction determined by LipoP and the outputs from SignalP and Phobius.

The principle goal of Lpp identification is to link information on the predicted protein location to knowledge of the predicted protein function. The dataset of putative Lpp for a given bacterium should therefore be investigated further using bioinformatic methods (such as sequence homology searches and conserved domain analysis) as a preliminary step towards defining appropriate biochemical, genetic and structural studies. Moreover, examination of genomic context is also an important feature of this process, as many Lpp are located in operons.

Acknowledgement The authors thank Northumbria University for financial support from the 'Research into Teaching' programme.

References

Babu M, Sankaran K (2002) DOLOP—database of bacterial lipoproteins. Bioinformatics 18:641–643. doi:10.1093/bioinformatics/18.4.641

Babu MM, Priya ML, Selvan AT, Madera M, Gough J, Aravind L et al (2006) A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins. J Bacteriol 188:2761–2773. doi:10.1128/JB.188.8.2761-2773.2006

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S et al (2005) The universal protein resource (UniProt). Nucleic Acids Res 33:D154–D159. doi:10.1093/nar/gki070

Baumgärtner M, Kärst U, Gerstel B, Loessner M, Wehland J, Jänsch L (2006) Inactivation of Lgt allows systematic characterization of lipoproteins from *Listeria monocytogenes*. J Bacteriol 189:313–324. doi:10.1128/JB.00976-06

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795. doi:10.1016/j.jmb.2004.05.028

Bendtsen JD, Binnewies TT, Hallin PF, Sicheritz-Pontén T, Ussery DW (2005) Genome update: prediction of secreted proteins in 225 bacterial proteomes. Microbiology 151:1725–1727. doi:10.1099/mic.0.28029-0

Berven FS, Karlsen OA, Straume AH, Flikka K, Murrell JC, Fjellbirkeland A et al (2006) Analysing the outer membrane subproteome of *Methylococcus capsulatus* (Bath) using proteomics and novel biocomputing tools. Arch Microbiol 184:362–377. doi:10.1007/s00203-005-0055-7

Braun V, Wu HC (1994) Lipoproteins: structure function, biosynthesis and model for protein export. N Comp Biochem 27:319–341

De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362-W365. doi:10.1093/nar/gkl124

Fariselli P, Finocchiaro G, Casadio R (2003) SPEPLip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics 19:2498–2499. doi:10.1093/bioinformatics/btg360

Gardy JL, Brinkman FSL (2006) Methods for predicting bacterial protein subcellular localisation. Nat Rev Microbiol 4:741–751. doi:10.1038/nrmicro1494

Juncker AS, Willenbrock H, von Heijne G, Nielsen H, Brunak S, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12:1652–1662. doi:10.1110/ps.0303703

Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. doi:10.1016/j.jmb.2004.03.016

Klein P, Somorjai RL, Lau PCK (1988) Distinctive properties of signal sequences from bacterial lipoproteins. Protein Eng 2:15–20. doi:10.1093/protein/2.1.15

Nakai K, Horton P (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–35. doi:10.1016/S0968-0004(98)01336-X

Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. In: Proceedings of the sixth international conference on intelligent systems for molecular biology (ISMB 6), AAAI Press, Menlo Park, California, pp 122–130

Réglier-Poupet H, Frehel C, Dubail I, Beretti JL, Berche P, Charbit A, Raynaud C (2003) Maturation of lipoproteins by type II signal peptidase is required for phagosomal escape of *Listeria monocytogenes*. J Biol Chem 278:49469–49477

Setubal JC, Reis M, Matsunaga J, Haake DA (2006) Lipoprotein computational prediction in spriochaetal genomes. Microbiology 152:113–121. doi:10.1099/mic.0.28317-0

Sutcliffe IC, Harrington DJ (2002) Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. Microbiology 148:2065–2077

Sutcliffe IC, Harrington DJ (2004) Lipoproteins of *Mycobacterium tuberculosis*: an abundant and functionally diverse class of cell envelope components. FEMS Microbiol Rev 28:645–659. doi:10.1016/j.femsre.2004.06.002

Sutcliffe IC, Hutchings MI (2007) Putative lipoproteins identified by bioinformatic genome analysis of *Leifsonia xyli* subsp. *xyli*, the causative agent of sugarcane ratoon stunting disease. Mol Plant Pathol 8:121–128. doi:10.1111/j.1364-3703.2006.00377.x

Sutcliffe IC, Russell RRB (1995) Lipoproteins of Gram-positive bacteria. J Bacteriol 177:1123–1128

Sutcliffe IC, Tao L, Ferretti JJ, Russell RRB (1993) MsmE, a lipoprotein involved in sugar transport in *Streptococcus mutans*. J Bacteriol 175:1853–1855

Taylor PD, Toseland CP, Attwood CK, Flower DR (2006) LIPPRED: a web server for accurate prediction of lipoprotein signal sequences and cleavage sites. Bioinformation 1:176–179

Tjalsma H, Zanen G, Venema G, Bron S, van Dijl JM (1999) The potential active site of the lipoprotein-specific (type II) signal peptidase of *Bacillus subtilis*. J Biol Chem 274:28191–28197. doi:10.1074/jbc.274.40.28191

Von Heijne G (1989) The structure of signal peptides from bacterial lipoproteins. Protein Eng 2:531–534. doi:10.1093/protein/2.7.531