research articles Journal of proteome research

Prediction of Lipoprotein Signal Peptides in Gram-Positive Bacteria with a Hidden Markov Model

Pantelis G. Bagos,*[,†,‡] Konstantinos D. Tsirigos,[†] Theodore D. Liakopoulos,[†] and Stavros J. Hamodrakas[†]

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 15701, Greece, and Department of Informatics with Applications in Biomedicine, School of Applied Sciences, Lamia 35100, Greece

Received March 1, 2008

We present a Hidden Markov Model method for the prediction of lipoprotein signal peptides of Gram-positive bacteria, trained on a set of 67 experimentally verified lipoproteins. The method outperforms LipoP and the methods based on regular expression patterns, in various data sets containing experimentally characterized lipoproteins, secretory proteins, proteins with an N-terminal TM segment and cytoplasmic proteins. The method is also very sensitive and specific in the detection of secretory signal peptides and in terms of overall accuracy outperforms even SignalP, which is the top-scoring method for the prediction of signal peptides. PRED-LIPO is freely available at http://bioinformatics.biol.uoa.gr/PRED-LIPO/, and we anticipate that it will be a valuable tool for the experimentalists studying secreted proteins and lipoproteins from Gram-positive bacteria.

Keywords: lipoproteins • signal peptide • hidden markov model • prediction • bacteria

Introduction

Signal peptides[1] in Bacteria are mainly divided into the secretory signal peptides that are cleaved by Signal Peptidase I (SPase I),[2,3] and those cleaved by Signal Peptidase II (SPase II or Lsp),[4] which characterize the membrane-bound lipoproteins. The secretory signal peptides have been extensively studied for years, revealing a structure comprised of a short, positively charged N-region, a hydrophobic H-region that spans the membrane, a C-region of mostly small and uncharged residues and a cleavage site (known as the A-X-A motif, in which A stands for alanine and X for any amino acid), that is recognized by the peptidase that cleaves the peptide and releases the mature protein.[5−7] The signal peptide of bacterial lipoproteins possesses a similar structure,[8] with main differences being the comparatively shorter length and the unique pattern in the C-region (which is commonly denoted by [LVI]-[AST]-[GA]-C and termed as "lipobox") that is recognized for cleavage by SPase II.[4] The cysteine in the last position of the particular pattern is indispensable in both Gram-positive and Gram-negative bacteria, and is necessary for membrane anchoring. The post-translational lipid modification involves three enzymes that act sequentially: the prolipoprotein diacylglyceryl transferase (Lgt), that transfers a diacylglyceride to the cysteine sulfydryl group, the signal peptidase II (SPase II or Lsp) that cleaves the signal peptide at the residue before the cysteine forming an apolipoprotein, and the apolipoprotein *N*-acyltransferase (Lnt) which acylates the α-amino group of the apolipoprotein N-terminal cysteine forming the mature lipoprotein.[9,10] The proteins carrying a secretory signal peptide can be directed to the membrane through the action of the Sec translocase,[11,12] although another major pathway has been discovered, utilizing the Twin-Arginine (TAT) translocase which recognizes (longer in general) signal peptides that are carrying a distinctive pattern of two consecutive arginines (R-R) in the N-region.[13−15] Translocation of lipoproteins through the TAT pathway has been postulated based on sequence analysis,[16] but only recently has been proven for Bacteria (*Desulfovibrio vulgaris*[17]) and Archaea (*Haloferax volcanii*[18]).

The discovery of globomycin, a specific inhibitor of SPase II, represented a major breakthrough in the biochemical studies of lipoprotein maturation.[19,20] Bacteria treated with globomycin, as well as SPase II deficient strains, show accumulation of lipid-modified prolipoproteins.[21] Nevertheless, extensive studies in SPase II deficient strains showed that absence of SPase II results in rather pleiotropic effects on the composition of the extracellular proteome, since some prolipoproteins were released in the medium, whereas the synthesis of others was strongly reduced.[22,23] Conversely, only in the case of Lgt deficient strains, significantly more lipoproteins are observed in the growth medium.[22,24] The most excellent, however, proof that a protein is a lipoprotein would be labeling with [^{3}H] or [^{14}C] palmitate in the presence/absence of globomycin (or in wild-type and SPase II or Lgt deficient strains), combined with immunoblotting, immunoprecipitation, protein fractionation and protease accessibility assays to investigate its extracellular localization.[25,26]

Computational prediction of secretory signal peptides was performed initially using weight matrices.[27] However, the

* To whom correspondence should be addressed. E-mail: pbagos@ucg.gr, pbagos@biol.uoa.gr.
† University of Athens.
‡ University of Central Greece.

10.1021/pr800162c CCC: $40.75
Published on Web 11/01/2008

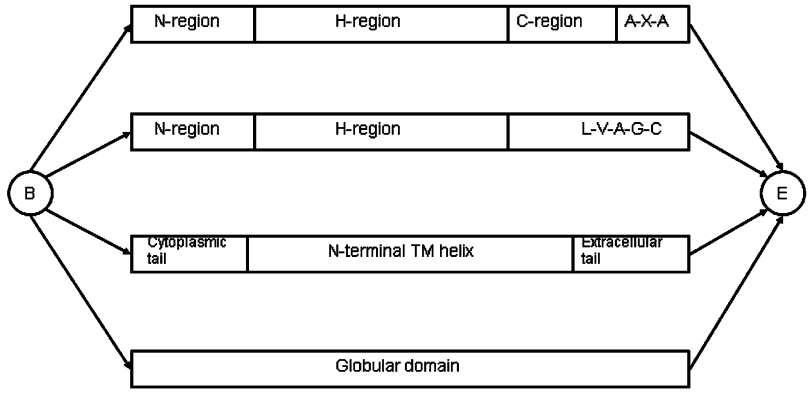

Figure 1. The topology of the full model with the four branches (submodels) corresponding to the secreted signal peptides, lipoprotein signal peptides, N-terminal TM segments and the cytoplasmic domain.

Neural Networks introduced by the SignalP method,[28,29] as well as Hidden Markov Models (HMM),[30] have been proven to be the most successful methods currently available.[31] Recently, the SignalP method was upgraded, mainly due to better annotation and selection of the training set, yielding an ever better accuracy,[32] whereas TatP has been presented offering the most accurate classification of TAT signal peptides.[33] A different approach has been followed in the Phobius method,[34,35] where a HMM was used to predict simultaneously the presence of a secretory signal peptide and the TM topology of a given protein. Following this approach, the authors showed that they can minimize the number of signal peptides predicted as transmembrane (TM) segments and vice versa. Concerning lipoproteins, for years, regular expression patterns were used based on the von Heijne rule,[8] with various modifications.[26,36−38] Recently, a method called LipoP was developed, which is based on HMMs and was trained exclusively on Gram-negative bacteria lipoproteins.[39] LipoP performs not only lipoprotein signal peptide prediction but also discrimination from secretory signal peptides, N-terminal TM helices and cytoplasmic proteins. LipoP has been reported to accurately classify ∼97% of Gram-negative bacteria lipoproteins with an error rate (on nonlipoproteins) of approximately 0.3%. When used, however, on lipoproteins from Gram-positive bacteria, the sensitivity of the prediction dropped to 90−92%.[39] In this work, we present a HMM-based method for performing the same task that is trained exclusively on experimentally verified lipoproteins from Gram-positive bacteria. We performed an extensive literature search in order to overcome the problem of limited experimentally verification and annotation of Gram-positive bacteria lipoproteins found in public databases, and in this way, we collected a data set of 67 such lipoproteins, the largest such set compiled so far. By analyzing these sequences, we show that they possess slightly different characteristics compared to their Gram-negative bacteria counterparts, providing, thus, a justification of our approach for constructing a different predictor. The method discriminates also very accurately secretory signal peptides (and predicts their cleavage site) as well as N-terminal TM anchored proteins. We show that the method developed here (PRED-LIPO) outperforms LipoP when applied to lipoproteins from Gram-positive bacteria, and we validate it on a number of different data sets. We also show that the module that predicts secretory signal peptides is also very accurate and compares favorably even to the currently top-scoring method, SignalP. Thus, the method developed here (http://bioinformatics.biol.uoa.gr/PRED-LIPO/) can be used also as a general predictor for signal peptides of Gram-positive bacteria, and we anticipate that it will be useful in proteomics applications and in large-scale genome analyses.

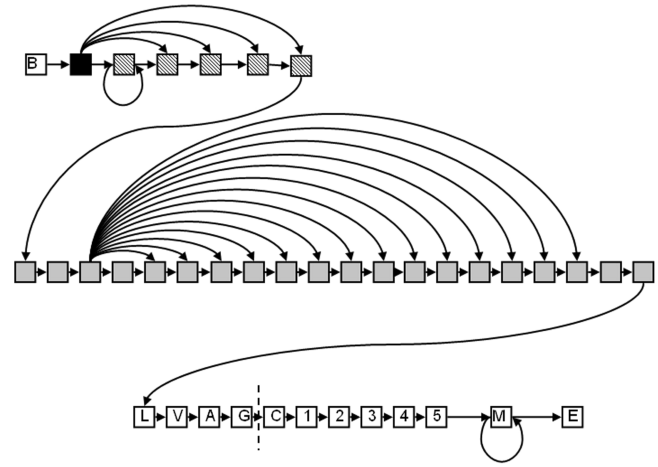

Figure 2. The model corresponding to the lipoprotein signal peptides. States in the n- and h-region that share the same emission probabilities are depicted with the same color. The cleavage site is presented with a dashed vertical line between G and C. Allowed transitions are depicted with arrows.

Materials and Methods

The Hidden Markov Model. The Hidden Markov Model (HMM) that we used is quite similar to the one proposed by LipoP. It consists of four different submodels (Figure 1), the Lipoprotein submodel, corresponding to the signal peptides cleaved by SPase II, the Signal Peptide submodel corresponding to the secretory signal peptides cleaved by SPase I, the N-terminal TM submodel corresponding to the N-terminal TM domain, and a globular submodel used to model the globular N-terminal domains of cytoplasmic or membrane proteins. The Lipoprotein submodel (Figure 2) was especially designed to capture the sequence features of Gram-positive bacteria lipoproteins. It contains states modeling the N-terminal n-region, the hydrophobic h-region and the lipobox (lipoprotein c-region). We used the same emission probabilities for the states in each region (with the exception of the lipobox) in order to avoid overfitting and the allowed transition probabilities were set in order to model as closely as possible the sequence features of the known lipoproteins. The secretory signal peptide model (Figure 3) is very similar to the lipoproteins' model, with the exception of the precaution for the existence of longer

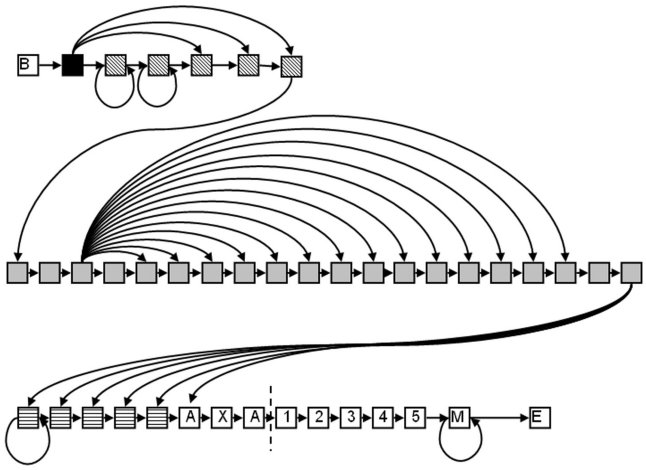

Figure 3. The model corresponding to the secretory signal peptides. States in the n- and h-region that share the same emission probabilities are depicted with the same color. The cleavage site is presented with a dashed vertical line between A and 1 (first amino acid of the mature protein). Allowed transitions are depicted with arrows.

n-regions, the variable length of c-regions and the different patterns of amino acids at the cleavage site. The TM submodel is identical to the one used by the HMM-TM predictor for α-helical membrane proteins,[40] whereas the globular submodel consists simply of a self-transitioning state. The total number of the model's states is 134 (including start and end states) with 227 freely estimated transitions. On the other hand, the total number of freely estimated emission probabilities is 589 (31 × 19), yielding a total number of freely estimated parameters equal to 816.

The model was trained using the Baum-Welch algorithm for labeled sequences[41] and the decoding was performed using the standard Viterbi algorithm,[42] although more advanced techniques such as the Posterior-Viterbi decoding[43] and the Optimal Accuracy Posterior Decoder[44] yield nearly identical results. With respect to this, the model introduced here differs from LipoP which uses the Forward decoding algorithm for choosing between the various submodels. In addition to the Viterbi decoding which produces the optimal path of states through the model, and hence predicts simultaneously the type of the sequence as well as the cleavages site (if any), we also report the S1 reliability index,[45] which takes values in the range 0−1 and it is a measure of the reliability of the prediction, useful in many situations.

The reported results correspond to a 33-fold cross-validation procedure, where each set consists of 11 proteins with an equally balanced number of SPase I cleaved signal peptides (3 or 4), SPase II cleaved signal peptides (2 or 3), TM (1 or 2) and globular proteins (3 or 4). The training procedure consists of removing 1 of the 33 subsets from the training set, training the model with the remaining proteins and performing the test on the proteins of the set that was removed. This process is tandemly repeated for all subsets in the training set, and the final prediction accuracy summarizes the outcome of all independent tests. For measures of accuracy in each binary classification problem (lipoproteins vs nonlipoproteins, signal peptides vs nonsignal peptides), we used the percentage of correctly classified positive examples (sensitivity), the percentage of correctly classified negative examples (specificity) and the Mathews Correlation coefficient that summarizes in a single measure True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).[46] For estimating the rate of correct predictions in the genome analysis, of particular importance is also the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) defined, respectively, as the percentage of true positives among the positive predictions (TP/TP + FP) and the percentage of true negatives among the negative predictions (TN/TN + FN). The method is available online at http://bioinformatics.biol.uoa.gr/PRED-LIPO/.

For comparison purposes, we created also two profile Hidden Markov Models (pHMMs) using the HMMER 2.3.2 package.[47] The pHMM is a special case of Hidden Markov Model and can be seen also as an extension of sequence profiles. It uses a HMM to model in a statistical manner a multiple alignment of related sequences. The pHMM, in contrast to the simple HMM described above, uses position specific parameters (emission and transition probabilities) and in general has a larger number of freely estimable parameters. It is well-suited for modeling protein families, but we used it here for comparison, since it has been shown that, under certain circumstances, it can also be used to model the sequence features of signal peptides.[48,49] We created the multiple alignments as advised previously,[48,49] we built the pHMMs using the hmmbuild command of the HMMER package, and we performed searches using the hmmpfam command of the same package.

Data Sets. The Data set that we used for training contained 67 experimentally verified lipoproteins from Gram-positive bacteria, 127 secreted proteins containing a signal peptide cleaved by SPase I from Gram-positive bacteria, 111 cytoplasmic proteins from Gram-positive bacteria and 58 Gram-positive bacterial sequences with an N-terminal TM segment that have their N-terminus located to the cytoplasmic side of the membrane. The 67 experimentally verified lipoproteins (Table 1) contain the 33 verified lipoproteins previously reported[26] that were used already for the construction of the G + LPP regular expression pattern. One of these sequences (MBL of *Streptococcus equi*) could not be retrieved from UniProt and we identified it using a BLAST[50] search against the genome sequence at http://www.sanger.ac.uk/Projects/S_equi/. In addition to these and given the low quality of the annotation in public databases regarding the experimental verification of Gram-positive bacteria lipoproteins, we performed an extensive literature search to identify additional such proteins. In total, we identified additionally 34 such proteins from various species of Gram-positive bacteria (*Mycoplasma*, *Mycobacterium*, *Corynebacterium*, *Spiroplasma*, *Streptomyces*, *Streptococcus*, *Staphylococcus*, *Bacillus*), which are listed in Table 1 along with the original references. Interestingly, in one of these proteins (CseA), the start codon reported in UniProt from previous publications was misassigned,[51] and the error was reported to UniProt database.[52] The identified papers provided results of varying degrees of reliability. The majority of the identified papers used chemical labeling of cysteine coupled with verification of the extracellular localization by subcellular fractionization and/or immunoblotting.[53−59] Others used site-directed mutagenesis in the lipobox region,[51,60] others relied only in the results obtained by treatment with globomycin coupled with subcellular localization techniques,[61,62] and one proteomic study used Lgt deficient strains.[24] Finally, several studies were included based only on indirect evidence[63−66] in order to obtain an as large as possible training set.

Table 1. The training set of 67 experimentally verified lipoproteins used in this study. We list the UniProt AC, the organism and the reference

the 34 newly identified lipoproteins		original set of 33 lipoproteins from ref 26	
UniProt AC[52]	organism	UniProt AC[52]	organism
Q8KVR9	*Mycoplasma mycoides*[54]	MBL (SEQ1660)	*Streptococcus equi*
O05121	*Mycoplasma gallisepticum*[58]	Q9RHZ6	*Alicyclobacillus acidocaldarius*
Q50327	*Mycoplasma pneumoniae*[57]	P06548	*Bacillus cereus*
P55801	*Mycoplasma mycoides*[66]	P00808	*Bacillus licheniformis*
P29230	*Mycoplasma hyorhinis*[59]	Q56247	*Bacillus PS3*
Q9X775	*Mycoplasma agalactiae*[64]	P34957	*Bacillus subtilis*
P0A671	*Mycobacterium bovis*[60]	P24327	*Bacillus subtilis*
P21625	*Spiroplasma melliferum*[55,56]	P46922	*Bacillus subtilis*
Q46023	*Corynebacterium diphtheriae*[61]	Q08429	*Bacillus subtilis*
O05471	*Streptococcus equisimilis*[62]	P24011	*Bacillus subtilis*
Q70UQ6	*Streptococcus uberis*[65]	P46338	*Bacillus subtilis*
Q9ZEP5	*Streptomyces coelicolor*[51]	Q93HZ4	*Corynebacterium glutamicum*
Q99VY4	*Staphylococcus aureus*[90]	O69087	*Heliobacterium gestii*
Q8VQS9	*Staphylococcus aureus*[90]	P14308	*Lactococcus lactis*
Q5HET4	*Staphylococcus aureus*[53]	Q03490	*Mycobacterium intracellulare*
Q2FI86	*Staphylococcus aureus*[53]	Q10790	*Mycobacterium tuberculosis*
Q7A603	*Staphylococcus aureus*[53]	P11572	*Mycobacterium tuberculosis*
Q99U04	*Staphylococcus aureus*[53]	P15712	*Mycobacterium tuberculosis*
Q600S6	*Mycoplasma hyopneumoniae*[63]	P96278	*Mycobacterium tuberculosis*
Q5ZZQ6	*Mycoplasma hyopneumoniae*[63]	P00807	*Staphylococcus aureus*
P40409	*Bacillus subtilis*[24]	Q9ZIN7	*Staphylococcus carnosus*
P37580	*Bacillus subtilis*[24]	Q7CCL6	*Staphylococcus epidermidis (strain ATCC 12228)*
O34385	*Bacillus subtilis*[24]	Q9Z692	*Streptococcus equi*
O34335	*Bacillus subtilis*[24]	O05471	*Streptococcus equisimilis*
P24141	*Bacillus subtilis*[24]	P31306	*Streptococcus gordonii Challis*
P36949	*Bacillus subtilis*[24]	Q00749	*Streptococcus mutans*
O34966	*Bacillus subtilis*[24]	P18791	*Streptococcus pneumoniae*
O05497	*Bacillus subtilis*[24]	P97008	*Streptococcus pneumoniae*
O31567	*Bacillus subtilis*[24]	Q51933	*Streptococcus pneumoniae*
O34348	*Bacillus subtilis*[24]	Q99Y38	*Streptococcus pyogenes*
P54535	*Bacillus subtilis*[24]	Q53919	*Streptomyces chrysomallus*
O05410	*Bacillus subtilis*[24]	Q9X9R7	*Streptomyces reticuli*
O32167	*Bacillus subtilis*[24]	O68456	*Thermoanaerobacter ethanolicus*
P54941	*Bacillus subtilis*[24]		

The 127 secreted proteins containing a SPase I cleaved signal peptide were retrieved from the set for training the SignalP method,[30] taking into consideration the corrections made later concerning wrongly annotated cleavage sites, initial mthionines and false annotations.[32] We did not try to eliminate proteins translocated through the Twin-Arginine Translocation (TAT) machinery,[13,14] in either the secretory or the lipoprotein set. The 111 proteins not containing either a signal peptide or an α-helical TM segment within the first 70 amino acids were collected from the well-curated data set of Menne et al.,[31] that was used to check the accuracy of signal peptide predictors. Putative TM proteins in this set were removed using TMHMM.[67] Finally, in order to model the N-terminal transmembrane (TM) domains, we scrutinized various well-annotated data sets[68−71] in order to compile a nonredundant set of transmembrane proteins from Gram-positive bacteria with experimentally verified topology. The final set consisted of 22 such transmembrane proteins, and from these, we extracted the TM segments with orientation from the cytoplasm to the extracellular space (In → Out), in a procedure similar to the one followed in the development of LipoP[39] and CW-PRED.[72] Thus, if a particular TM segment was localized in a 60-residue long window not overlapping with another TM segment, it was included in the set. In case of closely packed TM segments from multispanning TM proteins, we included only the upstream and downstream regions corresponding to the half of the proximal loop (extracellular or cytoplasmic).

To have an independent test set to further evaluate the method and compare it against the other available ones, we searched once again the recent literature. Experimentally verified lipoproteins from Gram-positive bacteria are, as we discussed earlier, very difficult to find. We have found, however, several proteomic analyses,[22,23,73,74] in which dozens of proteins were identified as potential lipoproteins. By this way, and also by searching the 278 experimentally verified bacterial lipoproteins from DOLOP,[36,37] a total number of 117 Lipoproteins have been collected. From UniProt, following Menne and coworkers,[31] we collected proteins having an experimentally verified signal peptide from Gram-positive bacteria, and after removing proteins that are already present in the set of SignalP (that we used for training), we came up with 89 proteins. The proteins carrying a secretory as well as a lipoprotein signal peptide were submitted to redundancy reduction following the procedures used in SignalP papers[28,29] (18 identical residues in the first 40 residues of the sequence). This reduction was extended further to the proteins of the training set in order to have a truly objective evaluation of the accuracy of the method. Finally, in the set of lipoproteins remained 110 sequences and in the set of secretory signal peptides remained 80 sequences.

We also retrieved cytoplasmic proteins from UniProt by searching the "Subcellular Localization" field and excluding entries marked as "Potential", "Putative" and "By Similarity". Given that the number of sequences was large, these were submitted to redundancy reduction using full sequences (30%

identities in an alignment of at least 80 residues), and once again, we removed proteins present (or having a homologue) in the training set, leaving us with 198 proteins. Finally, to test our method on TM proteins, we used the 106 experimentally verified cytoplasmic membrane proteins from Gram-positive bacteria used for the development of the PSORTB method.[75] From this set, we removed proteins present in the training set, proteins with a putative signal peptide (based on the annotation) and we performed redundancy reduction at 30% identical residues in an alignment of at least 80 residues, leaving finally 66 TM proteins.

We also validated the method on 109 secreted proteins of *Bacillus subtilis* whose signal peptides are sufficient for protein secretion in a novel expression system,[76] and 713 cytoplasmic proteins from the same organism identified by proteomic analyses.[73] From the work of Brockmeier et al.,[76] we also retrieved 25 proteins with predicted signal peptide that were not expressed in the heterologous expression system for various reasons, and 35 proteins that even though expressed they were not detected in the medium by any of the two reporters used (Cutinase or Esterase). These data sets were used to further assess the sensitivity and the specificity of the methods developed here, since these proteins were predicted to possess a signal peptide by SignalP. Once again, the proteins in the independent sets were compared against the sequences used for training in order to avoid redundancy, and cross-checked between the sets to avoid mistakes arising from the low-resolution proteomics experiments. Thus, from the 713 initially identified cytoplasmic proteins, 11 were also found in the other two subsets (lipoproteins and secreted), and thus, they were removed. All the protein sequences were retrieved from SubtiList.[77] Finally, we used the complete sequenced genomes of Gram-positive bacteria from the NCBI repository in order to perform predictions and compare the results.

Comparison to Other Prediction Methods. For comparison, we used mainly the LipoP[39] method, which is based on HMMs and was trained on Gram-negative bacteria lipoproteins, since it is the only available machine learning method for the same task. LipoP[39] possesses a model architecture similar to the one we used (discrimination between lipoprotein signal peptides, secretory signal peptides, N-terminal TM helices and cytoplasmic proteins). Besides the major difference that the method was trained on Gram-negative bacteria lipoproteins, another difference is the fact that LipoP is based on the forward decoding, whereas the method proposed here is using the Viterbi decoding algorithm.

Traditionally, the identification of bacterial lipoproteins was based on regular expression patterns. The first such pattern was the one proposed by von Heijne, back in 1989[8] which is [LVI]-[ASTG]-[GA]-C, requiring only one match to the first 2 positions (for all the patterns here we use the notation of Prosite[78]). The most widely used pattern is the PS00013 pattern of the Prosite database,[79] {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C, with the additional rule that the cysteine (C) must be between position 15 and 35, and at least one lysine (K) or arginine (R) must be in one of the first seven positions of the signal peptide. Recently, this pattern has been replaced by a Position Specific Scoring Matrix (PSSM) with Accession number: PS51257; this PSSM was also used in the analysis using ScanProsite,[80] but we chose to keep PS00013 in the analysis for historical reasons. Later, a pattern especially designed for Gram-positive bacteria lipoproteins (G + LPP) was developed based on observations on 33 experimentally characterized lipoproteins.[26] This pattern is: <[MV]-X(0,13)-[RK]-{DERKQ}(6,20)-[LIVMFESTAG]-[LVIAM]-[IVMSTAFG]-[AG]-C. Lastly, another pattern was developed by the creators of the DOLOP database,[36,37] which is a slightly modified version of the von Heijne pattern: [LVI]-[ASTVI]-[ASG]-C, requiring the additional rules that the cysteine (C) must be placed within the first 40 amino acids, at least one lysine (K) or arginine (R) must be in one of the first seven positions of the signal peptide, and that this positively charged residue should be 7−22 amino acids far from the cysteine. The regular expression patterns used here were implemented locally using PERL scripts.

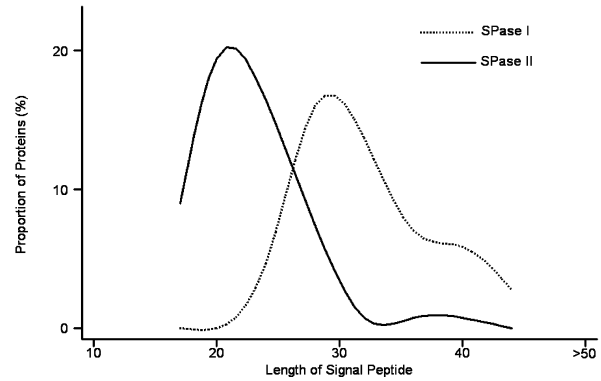

Figure 4. Smoothened histogram of the length distribution of lipoprotein and secretory signal peptides of Gram-positive bacteria. The latter are significantly longer with a mean length of approximately 31 amino acid (compared to 22). Notice the second mode in both distributions, accounting for signal peptides of length approximately 40 amino acids long. These could be instances of false annotations concerning the initial methionine or TAT signal peptides (see Results and Discussion).

Two other advanced methods have been proposed for the prediction of bacterial lipoproteins. The first one (SPEPlip) uses a combination of a Neural Network predictor to detect the presence of the signal peptide, and afterward filters out nonlipoproteins by using the PS00013 pattern[81] This eliminates much of the false positive predictions made by the regular expression pattern alone. However, the method was not trained in experimentally verified but rather on putative lipoproteins, and moreover, the web-server does not allow massive submissions; thus, we did not use it on our evaluation. Furthermore, since the method uses the PS00013 pattern, we expect that its sensitivity would be equal, although it may be more specific. The second method is based on the concept of probabilistic alignments of sequences with patterns (motifs),[82] and the authors used as an illustrative example for the development of the method, the case of Gram-positive bacteria lipoproteins. Later, they also applied the same method in *Escherichia coli* lipoproteins.[83] However, the original method was trained on all *B. subtilis* putative lipoproteins, and furthermore, no prediction tool is available to run the tests.

Finally, for analyses concerning accuracy in predicting secretory signal peptides, we also used SignalPv2,[28] SignalPv3,[32] Phobius[34] and PrediSi[84] in order to predict the putative signal-sequences of the proteins tested. We used both the Neural Network (NN) and the HMM modules of SignalP, trained on Gram-positive bacteria, using the default parameters, with the submitted sequences truncated to their first 70 residues.

Results and Discussion

In Figure 4, one can see the different length distributions of the signal peptides cleaved by SPase I (secreted proteins) and

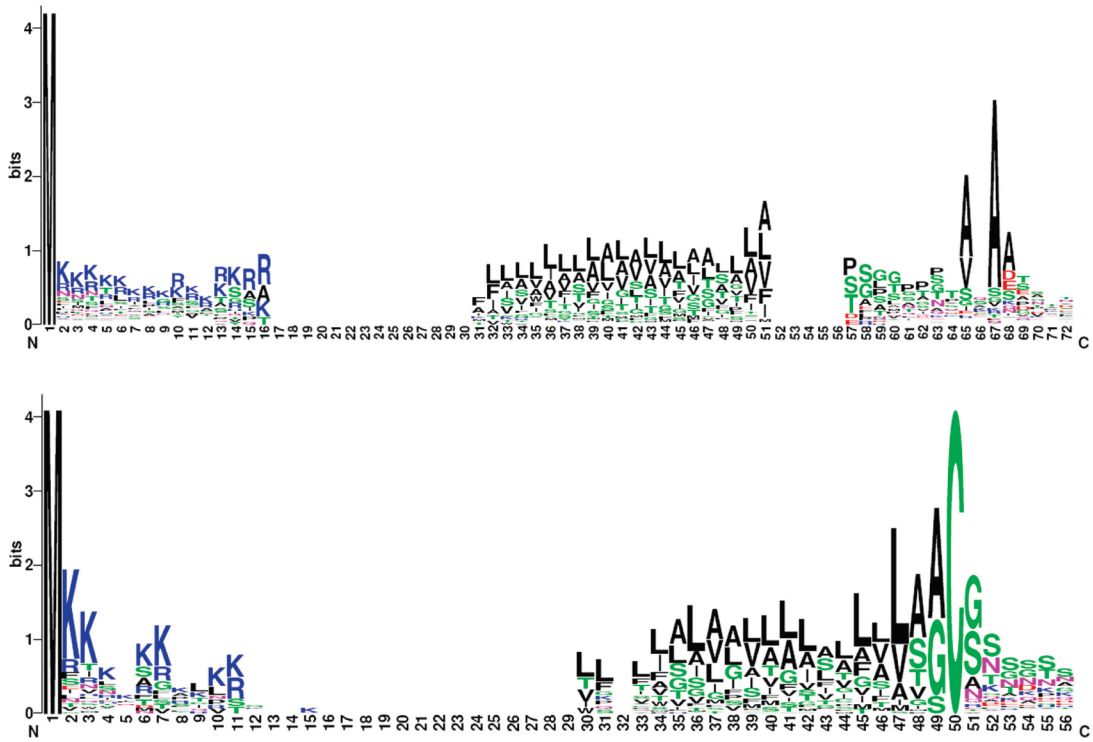

Figure 5. Sequence logo of secretory signal peptides (upper) and lipoprotein signal peptides (lower) of Gram-positive bacteria. The three regions (n-, h-, and c-region) were aligned as described previously,[48,49] and the logo was constructed using WebLogo.[86]

Table 2. The Results Obtained Using the HMM Method Developed Here in the Form of a Confusion Matrix[a]

		predicted				
		lipoprotein	membrane	cytoplasmic	signal peptide	total
			A			
Observed	Cytoplasmic	0 (0.00%)	2 (1.80%)	108 (97.30%)	1 (0.90%)	111 (100%)
	Membrane	0 (0.00%)	52 (89.66%)	5 (8.62%)	1 (1.72%)	58 (100%)
	Lipoprotein	65 (97.01%)	1 (1.49%)	1 (1.49%)	0 (0.00%)	67 (100%)
	Signal Peptide	0 (0.00%)	3 (2.36%)	4 (3.15%)	120 (94.49%)	127 (100%)
			B			
	Lipoprotein	107 (91.45%)	1 (1.92%)	0 (0%)	5 (9.62%)	117 (100%)
	Membrane	0 (0.00%)	26 (83.87%)	0 (0%)	5 (16.13%)	31 (100%)
	Cytoplasmic	0 (0.00%)	1 (0.51%)	195 (98.48%)	2 (1.01%)	198 (100%)
	Signal Peptide	0 (0.00%)	6 (6.74%)	2 (2.25%)	81 (91.01%)	89 (100%)

[a] (A) The results on the training set using a 33-fold cross-validation procedure. (B) The results on the independent test set.

those cleaved by SPase II (lipoproteins). Although Gram-positive bacteria lipoproteins have a comparable mean length (22 amino acids) to their Gram-negative bacteria counterparts, the secreted proteins' signal peptides are significantly longer (with a mean of 31 amino acids). The amino acid preferences are highlighted in the Sequence Logo[85] constructed using WebLogo[86] in Figure 5, where one can observe some slight differences from the respective patterns of Gram-negative bacteria. For instance, alanine (A) is still the dominant amino acid in position −2, but serine (S) is also very common. In position −1 (just before the cleavage site), alanine (A) is slightly more frequent than glycine (as opposed to what is the case in Gram-negative bacteria). Furthermore, glycine (G) is the most frequent amino acid in the first position immediately after the cleavage site (instead of serine). In general, in the first six positions of the mature protein following cysteine, the most common amino acids are serine, glycine, threonine and asparagine. These observations justify our approach to use a HMM to model the distinct characteristics of the Gram-positive bacteria lipoproteins.

The results obtained from the HMM method (PRED-LIPO) on the 33-fold cross-validation procedure are presented in Tables 2 and 3. The model performs very well correctly classifying the 65 out of the 67 lipoproteins (97.01%). Concerning the SPase-I cleaved signal peptides, the method also produces satisfactory results, correctly predicting the presence of a signal peptide in 120 out of the 127 proteins (94.49%) and excludes nonsignal peptides with a rate of 95.12%, giving an overall MCC of 0.95. These results are comparable with those reported by SignalP for Gram-positive bacteria (tested on similar data sets),[28,32] which was designed specifically to predict the secretory signal peptides (and hence not the lipoproteins' ones). The method developed here is also very specific, as it can be seen by the rate of false positive findings (Table 3). In total, it correctly classifies all of the 296 nonlipoproteins as nonlipoproteins (100% specificity). The method is also very specific concerning the prediction of secretory signal peptides, since it wrongly assigns such a signal peptide only in two cases out of the 236 proteins (0.85% false positives). On the same data set, LipoP which was developed for Gram-negative

Table 3. The Comparison of the Method Developed Here (PRED-LIPO) Concerning the Correct Classification of Lipoproteins (vs Nonlipoproteins) and the Comparison against the Other Available Methods[a]

A	on the training set (363 sequences)					
		correctly identified nonlipoproteins				
	correctly identified lipoproteins	signal peptide	cytoplasmic	TM proteins	total	MCC
PRED-LIPO	**65/67(97.01%)**	**127/127(100%)**	**111/111(100%)**	**58/58(100%)**	**296/296(100%)**	**0.97**
LipoP	58/67(86.57%)	126/127(99.21%)	111/111(100%)	58/58(100%)	295/296(99.66%)	0.91
PS00013	50/67(74.63%)	123/127(96.85%)	111/111(100%)	58/58(100%)	292/296(98.65%)	0.80
PS51257	62/67(92.54%)	123/127(96.85%)	111/111(100%)	58/58(100%)	292/296(98.65%)	0.92
G+LPP	55/67(82.09%)	126/127(99.21%)	111/111(100%)	58/58 (100%)	295/296(99.66%)	0.88
Von Heijne pattern	51/67(67.12%)	126/127(99.21%)	110/111(99.09%)	58/58(100%)	294/296(99.32%)	0.83
DOLOP pattern	50/67(74.63%)	126/127(99.21%)	111/111(100%)	58/58(100%)	295/296(99.66%)	0.83
B	on the independent test set (454 sequences)					
		correctly identified nonlipoproteins				
	correctly identified lipoproteins	signal peptide	cytoplasmic	TM proteins	total	MCC
PRED-LIPO	**99/110(90%)**	**80/80(100%)**	**198/198(100%)**	**66/66(100%)**	**344/344(100%)**	**0.93**
LipoP	91/110(82.72%)	80/80(100%)	198/198(100%)	66/66(100%)	344/344(100%)	0.88
PS00013	95/110(86.37%)	78/80(97.50%)	198/198(100%)	66/66(100%)	342/344(99.42%)	0.90
PS51257	98/110(89.09%)	77/80(96.25%)	198/198(100%)	66/66(100%)	341/344(99.13%)	0.91
G+LPP	68/110(61.81%)	79/80(98.75%)	198/198(100%)	66/66(100%)	343/344(99.71%)	0.73
Von Heijne pattern	69/110(62.73%)	79/80(98.75%)	197/198(99.49%)	66/66 (100%)	343/344(99.71%)	0.74
DOLOP pattern	77/110(70%)	80/80(100%)	198/198(100%)	66/66(100%)	344/344(100%)	0.80

[a] (A) The results on the training set using a 33-fold cross-validation procedure. (B) The results on the independent test set.

bacteria performs satisfactory, but it is, as expected, significantly worse than PRED-LIPO. LipoP correctly identifies 58 out of the 67 lipoproteins (86.57%), and correctly excludes 295 out of the 296 nonlipoproteins (99.66%). The methods that are based on regular expression patterns and sequence profiles (the von Heijne pattern, the G + LPP pattern, the PS00013, the DOLOP pattern and the PS51257 PSSM) perform well filtering out the nonlipoproteins (98−99%) but fail to correctly classify large numbers of lipoproteins (67−75% correct classification rates). The only exception here is the PS51257 PSSM that is "trained" on a large number of putative lipoproteins and is (as a PSSM), in nature, very closely related to the HMM models like PRED-LIPO, LipoP and the profile HMMs developed here. When looking at the MCC, which takes into account not only the TP and TN but also the FP and FN, PRED-LIPO (0.97) is clearly superior to LipoP (0.90) and PS51257 (0.91), whereas the pattern-based methods follow with correlations ranging from 0.80−0.88. We have to notice though, that even specificities larger than 98−99% seem well, when it comes to genome annotation concerning a rare protein class (such as lipoproteins) would produce a significant number of false positive findings (see below).

On the independent test set (Table 3), PRED-LIPO performs better than LipoP and the other available methods, classifying correctly 99 out of the 110 lipoproteins (with sensitivity equal to 90%). Concerning the specificity of the method in detecting lipoproteins, PRED-LIPO does not produce once again even a single false positive finding among the 344 secreted, TM and cytoplasmic proteins giving an MCC that is equal to 0.93. On the same data set, the performance and the rating of the pattern-based methods are similar to the ones in the set of 363 proteins mentioned above. They all are correctly excluding nonlipoproteins (99−100%) but fail to accurately classify lipoproteins with sensitivity better than 90%, yielding MCCs in the range (0.73−0.91, Table 3B). We should mention here that the most refined among these patterns (G + LPP), even though developed based on the 33 experimentally verified lipoproteins that we also used for training, fails to correctly classify most of the newly identified cases. This is something known from years and somehow expected, since it has to do with the generalization ability of regular expression patterns. On the other hand, PSSMs and machine learning methods like HMMs and Neural Networks have a superior performance since they can (if trained properly) tolerate unusual examples and classify them correctly.

Concerning the classification of secreted proteins bearing a signal peptide cleaved by SPase I (Table 4), the method correctly classifies 73 out of the 80 proteins (91.25%). The specificity of the method in detecting secretory signal peptides is very satisfactory since it wrongly assigns a signal peptide in 2 out of the 198 cytoplasmic proteins and in 3 out of the 66 TM proteins. These results correspond to a specificity of 98.11% in total with a MCC of 0.90, the best among the available predictors (including all versions of SignalP). Even though the sensitivity of PRED-LIPO is worse compared to the other predictors (except Phobius), its specificity is significantly superior due to its better ability to correctly exclude proteins with an N-terminal TM segment (all methods correctly identify approximately 99% of cytoplasmic proteins). Since in a real-world application (i.e., when searching a complete genome), there is a considerable chance that a candidate protein possesses indeed an N-terminal TM segment, the capability of a predictor to correctly exclude it is of significant importance. This was also the rationale behind the development of Phobius, but as it is apparent from Table 4, PRED-LIPO clearly outperforms even this specialized predictor. The overall MCCs reported here concerning SignalP correspond clearly to lower values compared to those reported in the original publications[28,32] where the authors reported MCCs for the Gram-positive predictor that was in the range 0.95−0.96. In the evaluation performed by Menne and co-workers (that included only SignalPv2),[31] the authors did not supply an estimate or MCC but clearly, the overall false positive rates (13−18% for SignalPv2) are in accordance to our estimates (13−17% for

Table 4. The Comparison of the Method Developed Here (PRED-LIPO) Concerning the Correct Classification of Secretory Signal Peptides (vs Nonsignal Peptides) and the Comparison against the Other Available Methods

	on an independent test set (344 sequences)				
		correctly identified nonsignal peptides			
	correctly identified signal peptides	**Cytoplasmic**	**TM proteins**	**Total**	MCC
PRED-LIPO	**73/80 (91.25%)**	**196/198 (98.99%)**	**63/66 (95.45%)**	**259/264 (98.11%)**	**0.90**
SignalPv2-NN	77/80 (96.25%)	197/198 (99.49%)	24/66 (36.36%)	221/264 (83.71%)	0.71
SignalPv2-HMM	77/80 (96.25%)	191/198 (96.46%)	41/66 (62.12%)	232/264 (87.88%)	0.77
SignalPv3-NN	73/80 (91.25%)	197/198 (99.49%)	50/66 (75.75%)	247/264 (93.56%)	0.81
SignalPv3-HMM	76/80 (95%)	196/198 (98.99%)	49/66 (74.24%)	244/264 (92.42%)	0.82
PrediSi	73/80 (91.25%)	195/198 (98.48%)	49/66 (74.24%)	243/264 (92.05%)	0.80
Phobius	71/80 (88.75%)	196/198 (98.99%)	57/66 (86.36%)	253/264 (95.83%)	0.85

Table 5. The Comparison of the Method Developed Here (PRED-LIPO) Concerning the Accuracy in Determining the Correct Cleavage Site in Secretory Signal Peptides and the Comparison against the Other Available Methods[a]

	correctly predicted cleavage site (percent of correctly classified proteins)	correctly predicted cleavage site within ±2 residues (percent of correctly classified proteins)
PRED-LIPO	52 (65%)	59 (73.75%)
SignalPv2-NN	58 (72.50%)	64 (80%)
SignalPv2-HMM	62 (77.50%)	69 (86.25%)
SignalPv3-NN	59 (73.75%)	61 (76.25%)
SignalPv3-HMM	63 (78.75%)	68 (85%)
PrediSi	58 (72.50%)	64 (80%)
Phobius	50 (62.5%)	52 (65%)

[a] The comparison has been performed on the independent test set of 80 secretory signal peptides.

SignalPv2 and 7−8% for SignalPv3). The obvious discrepancies are easily explained by the simple fact that our independent test set, as well as the one used by Menne and co-workers, includes proteins with an N-terminal TM segment. Thus, from the above, it is clear that PRED-LIPO, even though its primary objective was to predict lipoproteins, outperforms any other available secretory signal peptide predictor (concerning Gram-positive bacteria). PRED-LIPO is less specific compared to other methods; however, it is more sensitive, which is important when it comes to genome-wide analyses where the majority of the submitted proteins do not contain a signal peptide. Concerning the precise location of a cleavage site (Table 5), all methods perform satisfactory (perhaps with the exception of Phobius) predicting correctly the cleavage site in 62−79% of the cases and within ±2 residues in 65−86% of the cases. Interestingly, SignalPv3-HMM seems to perform uniformly better, outperforming even SignalPv3-NN.

The observation that SignalP is optimized toward sensitivity (few false negatives), whereas PRED-LIPO toward specificity (few false positives), is highlighted also in the analysis of the additional data set that includes 702 cytoplasmic and 169 proteins with putative (i.e., predicted by SignalP) secretory signal peptides of *B. subtilis* (of which 109 were detected in the medium, 35 were not detected and 25 were not expressed at all). On the secreted proteins, the method correctly classifies 94 out of the 109 proteins (86.24%), whereas among the cytoplasmic proteins, the method wrongly assigns a signal peptide in 8 out of the 702 proteins (1.14% false positives). However, we should emphasize here that this set consists of proteins that were identified and classified with methods of varying reliability and the results should be interpreted cautiously. Thus, the 109 secreted proteins in this data set were selected based on the predictions of SignalP, and their potential signal sequence was fused to a reporter protein which was detected in the extracellular medium in a heterologous expression system. Even in such a case, one cannot ignore the possibility that the putative signal sequence is in fact an N-terminal TM segment that directs the protein to the membrane where it is retained.[23] On the other hand, cytoplasmic proteins were detected after separation and analysis of the cytoplasmic fraction in proteomics analyses. In both cases, spurious annotations are expected in such a large data set arising from low resolution experiments and having the above in mind, and we may re-evaluate now the results of the independent test set. Thus, among the 8 presumed cytoplasmic proteins predicted as secreted or TM ones by our method ("false positives"), there are a lot of hypothetical proteins and enzymes and noticeably, ComEA[87] which is clearly a TM protein, raising thus the specificity to 99.01%. Among the proteins that presumably are containing a secretory signal peptide and were misclassified by our method, noticeable examples are CccA (cytochrome c550) which is TM-anchored, Mdr (Multidrug efflux transporter) which is multispanning membrane protein, rpmGB (50S ribosomal protein L33 2) which is cytoplasmic, and MotB which is a flagellar protein implicated in flagellar motor rotation. Interestingly, some of the cytoplasmic or secreted proteins misclassified by our predictor (AtpF, CccA, SpoIIQ) were recently reannotated as TM-anchored ones based on a combination of low-resolution proteomics analyses and computational predictions.[23] Thus, at least 5 out of the 15 "false negatives" of this set are more likely to be true negatives, raising the sensitivity to 90.83%, closer to the estimate of the independent test set mentioned earlier.

On the other hand, among the 25 proteins with a putative signal peptide (predicted by SignalP) that were not tolerated in the heterologous expression system proposed in the work of Brockmeier and co-workers,[76] 14 (DltD, FliL, LytR, MreC, PbpB, PhrG, SpoIIP, YocA, YrrL, YrrR, YrrS, YunB, YveB, YyaB) were classified by our method as transmembrane ones and one as cytoplasmic (YunA). Of the 35 proteins that were not detected in the medium, 3 were predicted as cytoplasmic (YwcI, YpcP, YwgB), 8 as transmembrane ones (YlbL, YoqH, YwdK, YwqC, YwqO, YwtC, YycP, tatAc) and one (YusW) as a lipoprotein. The simple fact that the majority of these proteins are predicted as TM ones correlates well with the observation that SignalP is more prone to false positive prediction in this particular class of proteins. A closer examination of these proteins reveals further that FliL is a flagelar protein, whereas YpcP is a DNA polymerase homologue and tatAC is the Twin-Arginine translocase of the TAT system, making thus unlikely

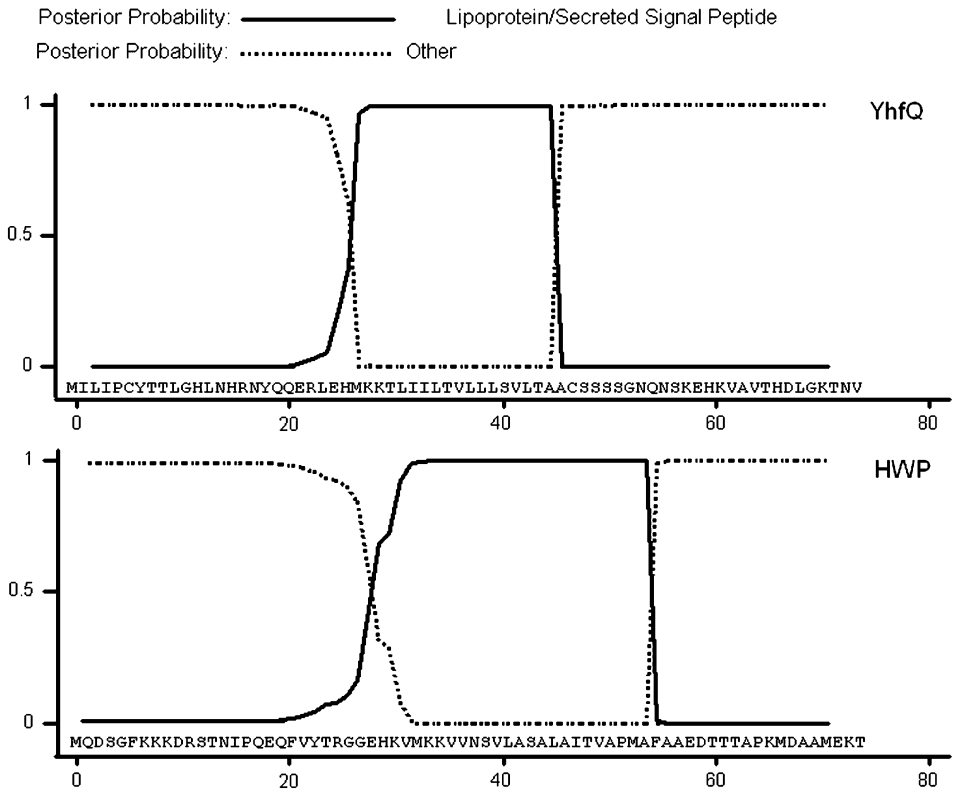

Figure 6. The posterior probabilities for YhfQ (upper) and HWP (lower) reveal some unusual signal peptides that potentially account for wrongly annotated initial methionine.

to possess a signal-peptide. The fact that among these 60 proteins, 27 (45%) were classified as not having a signal peptide should be an indication that at least some of these proteins indeed do not possess a SPase I cleaved signal peptide. This prediction contradicts the predictions made by SignalP, and is in accordance with the experimental data of Brockmeier and co-workers.[76] Considering also that the false negative prediction rate of our method concerning signal peptides is in the range 4−9%, it is highly unlikely that in this small set rises up to 45%. Thus, these results should be evaluated as results providing evidence for the specificity of our method, taken into account the experimental information. Furthermore, since we have shown (as also as Menne and co-workers) that SignalP produces false positives with a rate ∼7−13%, we expect that, among the 173 sequences initially used based on these predictions, there would be approximately 12−23 false positives and indeed, our method excluded 27. Of course, some of these could indeed possess a signal peptide that for a number of reasons discussed by Brockmeier and co-workers[76] could not be expressed in the expression system or could not be sufficient for exporting the particular reporters used to the extracellular medium. However, these results strongly suggest that the HMM method proposed in this work is highly specific in detecting both SPase I and II signal sequences, and even more specific than SignalP. This is of particular importance, since both classes of proteins (and especially lipoproteins) are rare ones in a completely sequenced genome, and when we are making predictions, we often want these predictions to be reliable. When a majority of the sequences submitted to such a predictor do not belong to the particular classes, we then try to avoid the false positive predictions that would affect the positive predictive value of the method.

The profile HMMs generated by HMMER were also used on the independent test sets. However, even though the profile for lipoprotein signal peptides performed very well (similar to PRED-LIPO on lipoproteins, and only one false positive on nonlipoproteins), the model used for secretory signal peptide prediction performs worse. In particular, although it is equally sensitive compared to PRED-LIPO, it fails at discriminating N-terminal TM segments (data not shown). Therefore, we do not recommend its use. Lastly, we have to mention that, since we used the training set of SignalPv2, all proteins belonging to the test set were checked against that set. Thus, some of the proteins contained in the test were used for training by SignalPv3, which was trained on a larger data set (153 secreted proteins instead of 111). Therefore, the results obtained using SignalPv3 (as well as Phobius and PrediSi) may be biased and slightly overestimated.

The analysis of the completely sequenced genomes revealed once again that PRED-LIPO is more suitable (compared to LipoP and the pattern-based methods) at correctly detecting lipoproteins in whole genome searches. PRED-LIPO predicts constantly a slightly smaller number of lipoproteins in the analyzed genomes (10 490 in total, accounting for a 1.98%), and given that PRED-LIPO is (as judged based on the training and test sets) more specific and more sensitive compared to the other methods, we are justified to conclude that the majority of them are truly lipoproteins. LipoP predicts 11 000 proteins in total (2.07%), DOLOP and G + LPP predict even smaller numbers (9044 and 8801 proteins, respectively) and the von Heijne, the PS51257 and PS00013 patterns predict 11 902, 12 634 and 12 097 lipoproteins, respectively, clearly producing overestimates. The detailed results are available at http://biophysics.biol.uoa.gr/PRED-LIPO-results/. For instance, in the

well-annotated proteome *B. subtilis*, PRED-LIPO predicts 88 lipoproteins, whereas LipoP 103 and G + LPP 71. Interestingly, in the same proteome, PRED-LIPO predicts 233 secretory signal peptides (accounting for 5.676% of the proteome), whereas SignalP-NN predicts 565 (13.764%) which is clearly an overestimate, since all available studies suggest that the "secretome" of the particular organism consists of approximately 200 proteins.[22,23] Thus, it is clear once again that SignalP is optimized toward increased sensitivity, and relying on its predictions without for instance taking into account other factors (i.e., TM predictions) underlies the risk of many false positives.

Finally, the HMM method developed here is also very helpful in identifying spurious cleavage sites and wrongly annotated translation initiation sites. For instance, one of the lipoproteins contained in the test set (YhfQ) possess an unusual long signal peptide of 44 amino acids (that among other things, makes the particular prediction impossible for some of the regular expression patterns). The prediction made by PRED-LIPO (Figure 6, upper panel) indicates that residues 1−25 are probably misannotated since the predicted start of the signal peptide is the methionine at position 26. The reliability of this prediction is very high (0.989) making this explanation highly probable. XlnC is another example of a spurious finding. The predicted signal peptide ends at position 49 and is in agreement with the experimentally verified site. However, the predicted starting methionine is at position 21. Interestingly, the particular protein is proven to be a TAT substrate, providing an explanation for the unusual length of the c-region of this signal peptide.[88] Similar results hold also for a number of other proven TAT signal peptides (data not shown). Finally, the Hexagonal Wall Protein (HWP, UniProt AC: P38538) from *Brevibacillus choshinensis* shows a very similar plot with a high reliability index (0.963). The predicted cleavage site is identical to the annotated one (AFA-A, position 54); however, the model strongly suggests that the initial methionine is the one located at position 31 of the precursor. Indeed, in the original publication, it is clearly stated that although the N-terminal sequence of the mature protein had been determined (AEDTTTA...), the authors were inconclusive concerning the precise translation initiation site, since M31, which is in-frame with M1, possesses all the necessary characteristics to be the starting methionine.[89]

Conclusions

We have presented a HMM method (PRED-LIPO) for predicting the secretory and lipoprotein signal peptides in Gram-positive bacteria and discriminate them from cytoplasmic and N-terminal TM proteins. The model resembles the well-known LipoP method which is trained on Gram-negative bacteria, and is the first such method especially designed for Gram-positive bacteria. PRED-LIPO is, as expected, superior to LipoP when tested on lipoproteins of Gram-positive bacteria, and should be used exclusively for predicting such sequences. One of the main advantages of the prediction method is the high specificity, since it predicts very few (<0.3%) false positives. Similar results hold also for the module that predicts the presence of the secretory signal peptides and this makes PRED-LIPO comparable even to the top-scoring method SignalP that is considered to be the most accurate predictor for signal peptides. The secretory signal peptide module of PRED-LIPO was validated further in various experimentally verified data sets, and the results strongly suggest that the method is more specific compared to SignalP and outperforms it in terms of overall accuracy. The method is freely available online at http://bioinformatics.biol.uoa.gr/PRED-LIPO/, and we anticipate that will be a valuable tool for the experimentalists studying secreted proteins and lipoproteins from Gram-positive bacteria. The method can be run in two modes, the single sequence mode, where the user submits one sequence and receives a detailed output (graphs, reliability, etc.), and in multiple sequence mode, where the user may submit up to 500 sequences at a time and receives the summary results (type of signal peptide, cleavage site) in an easily readable format. The later option would facilitate large-scale analyses of bacterial genomes and high-throuput proteomics applications. For larger submissions, interested users may contact the authors.

Acknowledgment. P.G.B. was supported by a postdoctoral research fellowship from the National Scholarships' Foundation (NSF) of Greece for the program "Machine Learning Algorithms in Bioinformatics" at the University of Athens. The authors would like to thank the numerous experimentalists cited in this work for making their data available on the public databases, the curators of these databases and the three anonymous reviewers that provided very useful comments that helped in improving the quality of the manuscript.

References

(1) von Heijne, G. The signal peptide. *J. Membr. Biol.* **1990**, *115* (3), 195–201.
(2) van Roosmalen, M. L.; Geukens, N.; Jongbloed, J. D.; Tjalsma, H.; Dubois, J. Y.; Bron, S.; van Dijl, J. M.; Anne, J. Type I signal peptidases of Gram-positive bacteria. *Biochim. Biophys. Acta* **2004**, *1694* (1−3), 279–97.
(3) Tuteja, R. Type I signal peptidase: an overview. *Arch. Biochem. Biophys.* **2005**, *441* (2), 107–11.
(4) Sankaran, K.; Wu, H. C. Bacterial prolipoprotein signal peptidase. *Methods Enzymol.* **1995**, *248*, 169–80.
(5) von Heijne, G. Signal sequences. The limits of variation. *J. Mol. Biol.* **1985**, *184* (1), 99–105.
(6) von Heijne, G. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. *EMBO J.* **1984**, *3* (10), 2315–8.
(7) von Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* **1983**, *133* (1), 17–21.
(8) von Heijne, G. The structure of signal peptides from bacterial lipoproteins. *Protein Eng.* **1989**, *2* (7), 531–4.
(9) Sankaran, K.; Gupta, S. D.; Wu, H. C. Modification of bacterial lipoproteins. *Methods Enzymol.* **1995**, *250*, 683–97.
(10) Sankaran, K.; Wu, H. C. Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol. *J. Biol. Chem.* **1994**, *269* (31), 19701–6.
(11) Driessen, A. J.; Nouwen, N. Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.* **2008**, *77*, 643–67.
(12) Gold, V. A.; Duong, F.; Collinson, I. Structure and function of the bacterial Sec translocon. *Mol. Membr. Biol.* **2007**, *24* (5−6), 387–94.
(13) Lee, P. A.; Tullman-Ercek, D.; Georgiou, G. The bacterial twin-arginine translocation pathway. *Annu. Rev. Microbiol.* **2006**, *60*, 373–95.
(14) Berks, B. C.; Palmer, T.; Sargent, F. Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr. Opin. Microbiol.* **2005**, *8* (2), 174–81.
(15) Teter, S. A.; Klionsky, D. J. How to get a folded protein across a membrane. *Trends Cell Biol.* **1999**, *9* (11), 428–31.
(16) Widdick, D. A.; Dilks, K.; Chandra, G.; Bottrill, A.; Naldrett, M.; Pohlschroder, M.; Palmer, T. The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (47), 17927–32.
(17) Valente, F. M.; Pereira, P. M.; Venceslau, S. S.; Regalla, M.; Coelho, A. V.; Pereira, I. A. The [NiFeSe] hydrogenase from Desulfovibrio vulgaris Hildenborough is a bacterial lipoprotein lacking a typical lipoprotein signal peptide. *FEBS Lett.* **2007**, *581* (18), 3341–4.

(18) Gimenez, M. I.; Dilks, K.; Pohlschroder, M. Haloferax volcanii twin-arginine translocation substates include secreted soluble, C-terminally anchored and lipoproteins. *Mol. Microbiol.* **2007**, *66* (6), 1597–1606.
(19) Ichihara, S.; Hussain, M.; Mizushima, S. Characterization of new membrane lipoproteins and their precursors of Escherichia coli. *J. Biol. Chem.* **1981**, *256* (6), 3125–9.
(20) Hussain, M.; Ichihara, S.; Mizushima, S. Mechanism of signal peptide cleavage in the biosynthesis of the major lipoprotein of the Escherichia coli outer membrane. *J. Biol. Chem.* **1982**, *257* (9), 5177–82.
(21) Hayashi, S.; Wu, H. C. Lipoproteins in bacteria. *J Bioenerg Biomembr* **1990**, *22* (3), 451–71.
(22) Tjalsma, H.; Antelmann, H.; Jongbloed, J. D.; Braun, P. G.; Darmon, E.; Dorenbos, R.; Dubois, J. Y.; Westers, H.; Zanen, G.; Quax, W. J.; Kuipers, O. P.; Bron, S.; Hecker, M.; van Dijl, J. M. Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. *Microbiol. Mol. Biol. Rev.* **2004**, *68* (2), 207–33.
(23) Tjalsma, H.; van Dijl, J. M. Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* **2005**, *5* (17), 4472–82.
(24) Antelmann, H.; Tjalsma, H.; Voigt, B.; Ohlmeier, S.; Bron, S.; van Dijl, J. M.; Hecker, M. A proteomic view on genome-based signal peptide predictions. *Genome Res.* **2001**, *11* (9), 1484–502.
(25) Sutcliffe, I. C.; Russell, R. R. Lipoproteins of gram-positive bacteria. *J. Bacteriol.* **1995**, *177* (5), 1123–8.
(26) Sutcliffe, I. C.; Harrington, D. J. Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology* **2002**, *148* (Pt 7), 2065–77.
(27) von Heijne, G. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* **1986**, *14* (11), 4683–90.
(28) Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **1997**, *10* (1), 1–6.
(29) Nielsen, H.; Brunak, S.; von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **1999**, *12* (1), 3–9.
(30) Nielsen, H.; Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998**, *6*, 122–30.
(31) Menne, K. M.; Hermjakob, H.; Apweiler, R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **2000**, *16* (8), 741–2.
(32) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340* (4), 783–95.
(33) Bendtsen, J. D.; Nielsen, H.; Widdick, D.; Palmer, T.; Brunak, S. Prediction of twin-arginine signal peptides. *BMC Bioinf.* **2005**, *6*, 167.
(34) Kall, L.; Krogh, A.; Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004**, *338* (5), 1027–36.
(35) Kall, L.; Krogh, A.; Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W429–32.
(36) Madan Babu, M.; Sankaran, K. DOLOP--database of bacterial lipoproteins. *Bioinformatics* **2002**, *18* (4), 641–3.
(37) Madan Babu, M.; Priya, M. L.; Selvan, A. T.; Madera, M.; Gough, J.; Aravind, L.; Sankaran, K. A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins. *J. Bacteriol.* **2006**, *188* (8), 2761–73.
(38) Setubal, J. C.; Reis, M.; Matsunaga, J.; Haake, D. A. Lipoprotein computational prediction in spirochaetal genomes. *Microbiology* **2006**, *152* (Pt 1), 113–21.
(39) Juncker, A. S.; Willenbrock, H.; Von Heijne, G.; Brunak, S.; Nielsen, H.; Krogh, A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **2003**, *12* (8), 1652–62.
(40) Bagos, P. G.; Liakopoulos, T. D.; Hamodrakas, S. J. Algorithms for incorporating prior topological information in HMMs: Application to transmembrane proteins. *BMC Bioinf.* **2006**, *7* (1), 189.
(41) Krogh, A. Hidden Markov models for labelled sequences. *Proceedings of the12th IAPR International Conference on Pattern Recognition*, October 9−13, 1994, Jerusalem, Israel, pp 140144.
(42) Durbin, R.; Eddy, S. R.; Krogh, A.; Mithison, G. *Biological Sequence Analysis, Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, U.K., 1998.
(43) Fariselli, P.; Martelli, P. L.; Casadio, R. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinf.* **2005**, *6* (Suppl 4), S12.
(44) Kall, L.; Krogh, A.; Sonnhammer, E. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **2005**, *21 Suppl 1*, i251–7.
(45) Melen, K.; Krogh, A.; von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **2003**, *327* (3), 735–44.
(46) Baldi, P.; Brunak, S.; Chauvin, Y.; ersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16* (5), 412–24.
(47) Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **1998**, *14* (9), 755–63.
(48) Zhang, Z.; Henzel, W. J. Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci.* **2004**, *13* (10), 2819–24.
(49) Zhang, Z.; Wood, W. I. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **2003**, *19* (2), 307–8.
(50) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–402.
(51) Hutchings, M. I.; Hong, H. J.; Leibovitz, E.; Sutcliffe, I. C.; Buttner, M. J. The sigma(E) cell envelope stress response of Streptomyces coelicolor is influenced by a novel lipoprotein, CseA. *J. Bacteriol.* **2006**, *188* (20), 7222–9.
(52) Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O'Donovan, C.; Redaschi, N.; Suzek, B. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **2006**, *34* (Database issue), D187–91.
(53) Stoll, H.; Dengjel, J.; Nerz, C.; Gotz, F. Staphylococcus aureus deficient in lipidation of prelipoproteins is attenuated in growth and immune activation. *Infect. Immun.* **2005**, *73* (4), 2411–23.
(54) Persson, A.; Jacobsson, K.; Frykberg, L.; Johansson, K. E.; Poumarat, F. Variable surface protein Vmm of Mycoplasma mycoides subsp. mycoides small colony type. *J. Bacteriol.* **2002**, *184* (13), 3712–22.
(55) Le Henaff, M.; Fontenelle, C. Chemical analysis of processing of spiralin, the major lipoprotein of Spiroplasma melliferum. *Arch. Microbiol.* **2000**, *173* (5−6), 339–45.
(56) Le Henaff, M.; Cremet, J. Y.; Fontenelle, C. Purification and characterization of the major lipoprotein (P28) of Spiroplasma apis. *Protein Expression Purif.* **2002**, *24* (3), 489–96.
(57) Pyrowolakis, G.; Hofmann, D.; Herrmann, R. The subunit b of the F0F1-type ATPase of the bacterium Mycoplasma pneumoniae is a lipoprotein. *J. Biol. Chem.* **1998**, *273* (38), 24792–6.
(58) Jan, G.; Le Henaff, M.; Fontenelle, C.; Wroblewski, H. Biochemical and antigenic characterisation of Mycoplasma gallisepticum membrane proteins P52 and P67 (pMGA). *Arch. Microbiol.* **2001**, *177* (1), 81–90.
(59) Cleavinger, C. M.; Kim, M. F.; Wise, K. S. Processing and surface presentation of the Mycoplasma hyorhinis variant lipoprotein VlpC. *J. Bacteriol.* **1994**, *176* (8), 2463–7.
(60) Vosloo, W.; Tippoo, P.; Hughes, J. E.; Harriman, N.; Emms, M.; Beatty, D. W.; Zappe, H.; Steyn, L. M. Characterisation of a lipoprotein in Mycobacterium bovis (BCG) with sequence similarity to the secreted protein MPB70. *Gene* **1997**, *188* (1), 123–8.
(61) Schmitt, M. P.; Talley, B. G.; Holmes, R. K. Characterization of lipoprotein IRP1 from Corynebacterium diphtheriae, which is regulated by the diphtheria toxin repressor (DtxR) and iron. *Infect. Immun.* **1997**, *65* (12), 5364–7.
(62) Gase, K.; Liu, G.; Bruckmann, A.; Steiner, K.; Ozegowski, J.; Malke, H. The lppC gene of Streptococcus equisimilis encodes a lipoprotein that is homologous to the e (P4) outer membrane protein from Haemophilus influenzae. *Med. Microbiol. Immunol.* **1997**, *186* (1), 63–73.
(63) Meens, J.; Selke, M.; Gerlach, G. F. Identification and immunological characterization of conserved Mycoplasma hyopneumoniae lipoproteins Mhp378 and Mhp651. *Vet. Microbiol.* **2006**, *116* (1−3), 85–95.
(64) Rosati, S.; Pozzi, S.; Robino, P.; Montinaro, B.; Conti, A.; Fadda, M.; Pittau, M. P48 major surface antigen of Mycoplasma agalactiae is homologous to a malp product of Mycoplasma fermentans and belongs to a selected family of bacterial lipoproteins. *Infect. Immun.* **1999**, *67* (11), 6213–6.
(65) Smith, A. J.; Ward, P. N.; Field, T. R.; Jones, C. L.; Lincoln, R. A.; Leigh, J. A. MtuA, a lipoprotein receptor antigen from Streptococcus uberis, is responsible for acquisition of manganese during growth in milk and is essential for infection of the lactating bovine mammary gland. *Infect. Immun.* **2003**, *71* (9), 4842–9.

(66) Cheng, X.; Nicolet, J.; Miserez, R.; Kuhnert, P.; Krampe, M.; Pilloud, T.; Abdo, E. M.; Griot, C.; Frey, J. Characterization of the gene for an immunodominant 72 kDa lipoprotein of Mycoplasma mycoides subsp. mycoides small colony type. *Microbiology* **1996**, *142* (Pt 12), 3515–24.
(67) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001**, *305* (3), 567–80.
(68) Moller, S.; Kriventseva, E. V.; Apweiler, R. A collection of well characterised integral membrane proteins. *Bioinformatics* **2000**, *16* (12), 1159–60.
(69) Ikeda, M.; Arai, M.; Okuno, T.; Shimizu, T. TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.* **2003**, *31* (1), 406–9.
(70) Chen, C. P.; Rost, B. Long membrane helices and short loops predicted less accurately. *Protein Sci.* **2002**, *11* (12), 2766–73.
(71) Jayasinghe, S.; Hristova, K.; White, S. H. MPtopo: A database of membrane protein topology. *Protein Sci.* **2001**, *10* (2), 455–8.
(72) Litou, Z. I.; Bagos, P. G.; Tsirigos, K. D.; Liakopoulos, T. D.; Hamodrakas, S. J. Prediction of Cell Wall sorting signals in Gram-positive bacteria with a Hidden Markov Model: application to complete genomes. *J. Bioinform. Comput. Biol.* **2008**, *6* (2), 387–401.
(73) Eymann, C.; Dreisbach, A.; Albrecht, D.; Bernhardt, J.; Becher, D.; Gentner, S.; Tam le, T.; Buttner, K.; Buurman, G.; Scharf, C.; Venz, S.; Volker, U.; Hecker, M. A comprehensive proteome map of growing Bacillus subtilis cells. *Proteomics* **2004**, *4* (10), 2849–76.
(74) Bunai, K.; Ariga, M.; Inoue, T.; Nozaki, M.; Ogane, S.; Kakeshita, H.; Nemoto, T.; Nakanishi, H.; Yamane, K. Profiling and comprehensive expression analysis of ABC transporter solute-binding proteins of Bacillus subtilis membrane based on a proteomic approach. *Electrophoresis* **2004**, *25* (1), 141–55.
(75) Gardy, J. L.; Laird, M. R.; Chen, F.; Rey, S.; Walsh, C. J.; Ester, M.; Brinkman, F. S. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **2005**, *21* (5), 617–23.
(76) Brockmeier, U.; Caspers, M.; Freudl, R.; Jockwer, A.; Noll, T.; Eggert, T. Systematic screening of all signal peptides from Bacillus subtilis: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. *J. Mol. Biol.* **2006**, *362* (3), 393–402.
(77) Moszer, I.; Jones, L. M.; Moreira, S.; Fabry, C.; Danchin, A. SubtiList: the reference database for the Bacillus subtilis genome. *Nucleic Acids Res.* **2002**, *30* (1), 62–5.
(78) Bucher, P.; Bairoch, A. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 53–61.
(79) Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; De Castro, E.; Langendijk-Genevaux, P. S.; Pagni, M.; Sigrist, C. J. The PROSITE database. *Nucleic Acids Res.* **2006**, *34* (Database issue), D227–30.
(80) de Castro, E.; Sigrist, C. J.; Gattiker, A.; Bulliard, V.; Langendijk-Genevaux, P. S.; Gasteiger, E.; Bairoch, A.; Hulo, N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W362–5.
(81) Fariselli, P.; Finocchiaro, G.; Casadio, R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* **2003**, *19* (18), 2498–9.
(82) Gonnet, P.; Lisacek, F. Probabilistic alignment of motifs with sequences. *Bioinformatics* **2002**, *18* (8), 1091–101.
(83) Gonnet, P.; Rudd, K. E.; Lisacek, F. 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* **2004**, *4* (6), 1597–613.
(84) Hiller, K.; Grote, A.; Scheer, M.; Munch, R.; Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W375–9.
(85) Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18* (20), 6097–100.
(86) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14* (6), 1188–90.
(87) Inamine, G. S.; Dubnau, D. ComEA, a Bacillus subtilis integral membrane protein required for genetic transformation, is needed for both DNA binding and transport. *J. Bacteriol.* **1995**, *177* (11), 3045–51.
(88) Faury, D.; Saidane, S.; Li, H.; Morosoli, R. Secretion of active xylanase C from Streptomyces lividans is exclusively mediated by the Tat protein export system. *Biochim. Biophys. Acta* **2004**, *1699* (1−2), 155–62.
(89) Ebisu, S.; Tsuboi, A.; Takagi, H.; Naruse, Y.; Yamagata, H.; Tsukagoshi, N.; Udaka, S. Conserved structures of cell wall protein genes among protein-producing Bacillus brevis strains. *J. Bacteriol.* **1990**, *172* (3), 1312–20.
(90) Brady, R. A.; Leid, J. G.; Camper, A. K.; Costerton, J. W.; Shirtliff, M. E. Identification of Staphylococcus aureus proteins recognized by the antibody-mediated immune response to a Biofilm infection. *Infect. Immun.* **2006**, *74* (6), 3415–26.

PR800162C