Journal of Bioinformatics and Computational Biology
Vol. 6, No. 2 (2008) 387–401

Imperial College Press
www.icpress.co.uk

PREDICTION OF CELL WALL SORTING SIGNALS IN GRAM-POSITIVE BACTERIA WITH A HIDDEN MARKOV MODEL: APPLICATION TO COMPLETE GENOMES

ZOI I. LITOU[*,‡], 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

[†]*Department of Biomedical Informatics*
University of Central Greece
Lamia 35100, Greece
[‡]*zlitou@biol.uoa.gr*
[§]*pbagos@biol.uoa.gr*
[¶]*ktsirig@biol.uoa.gr*
[‖]*liakop@biol.uoa.gr*
[**]*shamodr@biol.uoa.gr*

Received 25 July 2007
Revised 25 September 2007
Accepted 18 December 2007

Surface proteins in Gram-positive bacteria are frequently implicated in virulence. We have focused on a group of extracellular cell wall-attached proteins (CWPs), containing an LPXTG motif for cleavage and covalent coupling to peptidoglycan by sortase enzymes. A hidden Markov model (HMM) approach for predicting the LPXTG-anchored cell wall proteins of Gram-positive bacteria was developed and compared against existing methods. The HMM model is parsimonious in terms of the number of freely estimated parameters, and it has proved to be very sensitive and specific in a training set of 55 experimentally verified LPXTG-anchored cell wall proteins as well as in reliable data sets of globular and transmembrane proteins. In order to identify such proteins in Gram-positive bacteria, a comprehensive analysis of 94 completely sequenced genomes has been performed. We identified, in total, 860 LPXTG-anchored cell wall proteins, a number that is significantly higher compared to those obtained by other available methods. Of these proteins, 237 are hypothetical proteins according to the annotation of SwissProt, and 88 had no homologs in the SwissProt database — this might be evidence that they are members of newly identified families of CWPs. The prediction tool, the database with the proteins identified in the genomes, and supplementary material are available online at http://bioinformatics.biol.uoa.gr/CW-PRED/.

Keywords: Gram-positive bacteria; cell wall; sortase; hidden Markov model; LPXTG.

[§]Corresponding author.

1. Introduction

Pathogenic bacteria display an array of surface proteins with important functions including invasion of host cells and tissues, adhesion to the site of infection, and evasion of the immune response. These proteins are frequently required for virulence[1] and are potential drug or vaccine targets.[2] One of the major types of cell surface-displayed proteins of Gram-positive bacteria are the LPXTG-like proteins. These cell wall-associated surface proteins are the only currently recognized proteins that are covalently linked (anchored) to the cell wall peptidoglycan of Gram-positive bacteria and share common features.[3,4] They are linked to the cell wall envelope by a mechanism that requires a C-terminal sorting signal with a conserved LPXTG motif.[3,5]

The mechanism of protein attachment to the peptidoglycan and major components of the cell wall anchor structure of surface proteins are conserved in Gram-positive bacteria.[6] The carboxy-terminal sorting signal, essential for attachment, consists of an LPXTG sequence motif (where X denotes any amino acid), followed by a C-terminal hydrophobic domain comprising ∼20 amino acids and a positively charged tail.[4,7,8] LPXTG proteins are covalently attached to the Gram-positive bacterial cell wall by membrane-associated transpeptidases called sortases.[3,4,9–11] Sortases catalyze a transpeptidation reaction by first cleaving surface protein substrates at the cell wall sorting signal.[5] The polypeptide is retained in the cell membrane compartment, and the cleavage takes place between the threonine and glycine residues of the LPXTG motif. The carboxy terminal of the cleaved polypeptide chain joins the bacterial cell wall via an amide linkage to the amino group of the pentaglycine cross-bridge of the peptidoglycan, and this reaction is catalyzed by sortases as well; while the hydrophobic region next to the LPXTG motif passes through the plasma membrane and the charged tail, both possibly acting as a stop-transfer signal.[3–6,12,13] The surface protein linked to the peptidoglycan is then incorporated into the envelope and displayed on the microbial surface.[14]

The fact that there are multiple signal peptidase genes and sortase genes suggests that there is more than one surface protein transport pathway. It has recently been shown that not all proteins which have been experimentally verified to be sortase substrates contain a cell wall-sorting motif that fits the LPXTG pattern. For example, sortase B (SrtB) from *Staphylococcus aureus* recognizes the NPQTN motif.[11] In a recent work,[9] a computational strategy based on similarity searches was used to find all available sortase substrates in 72 bacterial genomes. However, these results still await experimental verification, since there are not enough experimentally verified cell wall-attached proteins (CWPs) available that do not conform to the classic LPXTG pattern.

Apart from the C-terminal sorting signal, which is essential for attachment, surface protein precursors are first initiated into the secretory pathway via N-terminal signal peptides.[3,10] The presence of an N-terminal secretory signal sequence is a feature of LPXTG proteins, and many of them contain a conserved sequence,

(Y/F)SIRK or variants of it. Although this conserved sequence has been observed, it is not exclusively found in these proteins and not all of them contain it.[8]

This work presents a novel hidden Markov model (HMM) approach for predicting the LPXTG-anchored CWPs of Gram-positive bacteria. The method is compared against the tripartite motif that has already been suggested, which takes into account three principal requirements for efficient sorting[15]; against the characteristic profile HMM deposited in the PFAM database[16]; and against a recently compiled profile HMM[11] that was developed with the intention of modeling not only the LPXTG-containing sequences, but also the substrates of other sortases. The method developed here has also been used for screening complete genomes in order to identify novel, putative CWPs.

2. Materials and Methods

2.1. *The hidden Markov model*

The hidden Markov model (HMM) that we used consists of three different submodels (Fig. 1): the LPXTG submodel, corresponding to the cell wall anchoring domain; the transmembrane (TM) submodel, corresponding to the C-terminal TM domain; and a globular submodel, used to model the globular C-terminal domains of secreted or membrane proteins. The LPXTG submodel (Fig. 2) was especially designed to capture the sequence features of the cell wall anchored proteins. It contains a region comprised of six states corresponding to the LPXTG cleavage site pattern, a three-state region with self-transitions used to model the nonfixed-length gap region, the membrane anchoring region comprised of 25 states, and lastly the C-terminal region used to model the positively charged tails. We used the same emission probabilities for the states in each region (with the exception of the cleavage site pattern) to avoid overfitting, and the allowed transition probabilities were

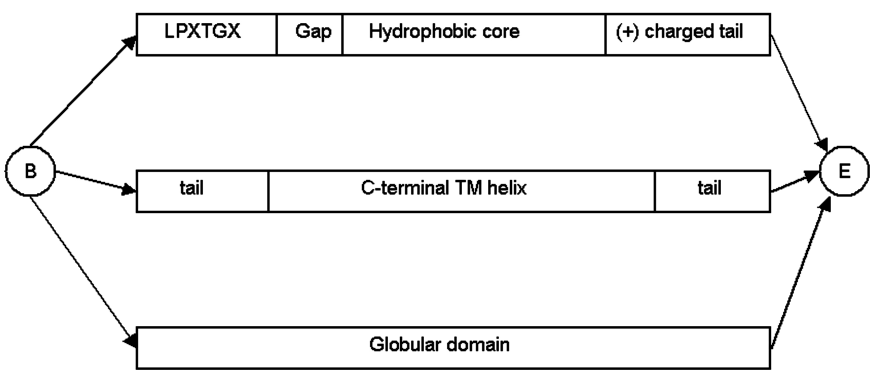

Fig. 1. Overview of the hidden Markov model (HMM). The model consists of three different submodels: the LPXTG anchor submodel, the C-terminal TM helix submodel, and the globular submodel.

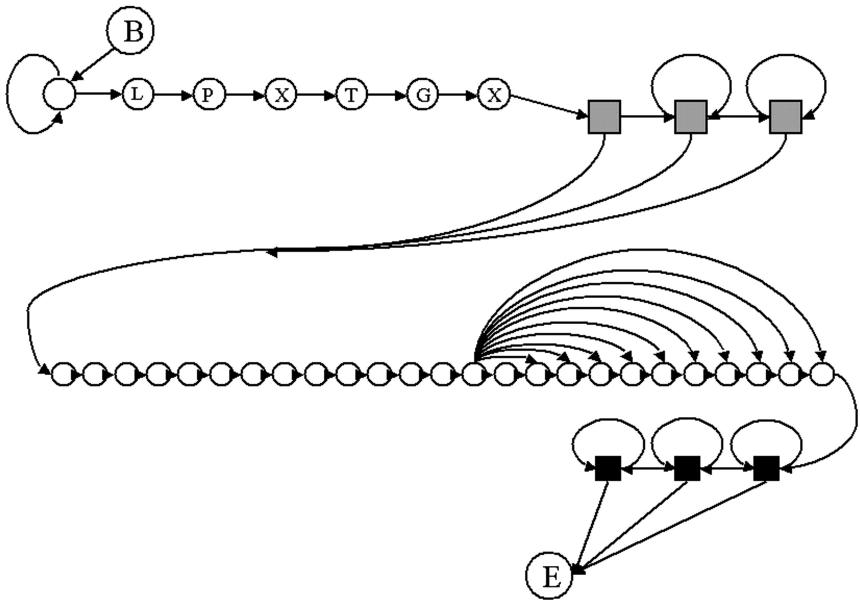

Fig. 2. A graphical representation of the LPXTG anchor submodel of the HMM. Different states are used sequentially to model the LPXTG cleavage site, the variable-length gap region, the hydrophobic helix, and the positively charged tail. Arrows represent allowed transitions among states, and states having the same emission probabilities are depicted using the same shape and color. Regions that are expected to have variable (nonfixed) length are modeled using self-transitioning states. The beginning state is denoted by B, and the end state by E.

set to model as closely as possible the sequence features of the known CWPs. We chose to model the gap region using self-transitioning states which allow arbitrary length, since it was observed that in some species the length distribution of these regions might be different from that of the majority of the known proteins, thus causing these proteins to be misclassified by the regular expression patterns previously used.[15] The TM submodel is identical to the one used by the HMM-TM predictor for alpha-helical membrane proteins,[16] whereas the globular submodel consists simply of a self-transitioning state. The total number of the model's states is 95 (including beginning and end states), with 78 freely estimated transitions; on the other hand, the total number of freely estimated emission probabilities is 285 (15×19), yielding a total number of freely estimated parameters of 363. Thus, the model is far more parsimonious than the profile HMM deposited in PFAM[17] and the profile HMM compiled by Boekhorst *et al.*,[11] which consist of position-specific parameters, while it is at the same time more flexible than the tripartite pattern based on regular expressions.[15]

The model was trained using the Baum–Welch algorithm for labeled sequences,[18] and the decoding was performed using the standard Viterbi algorithm,[19] seeing as more advanced techniques such as the posterior-Viterbi

decoding[20] yielded identical results. The reported results correspond to a 25-fold cross-validation procedure, where each set consists of an equally balanced number of CWPs, TM proteins, and globular proteins. The training procedure starts by removing 1 of the 25 subsets from the training set, training the model with the remaining proteins, and then performing the test on the proteins of each subset which 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.

2.2. *Data sets*

For training the HMM, we needed a data set comprised of LPXTG-anchored proteins, globular C-terminal domains of globular proteins, and C-terminal transmembrane helices. The data set used for training the LPXTG submodel was compiled from a list of 65 experimentally verified cell wall anchoring proteins of Gram-positive bacteria.[5] The sequences were retrieved from UniProt,[21] and their C-terminal segments including the LPXTG-anchored region (60 residues long) were used. In order to avoid overfitting arising from redundancy in the training set, we performed a redundancy reduction procedure with Algorithm 2 of Hobohm *et al.*,[22] using BLAST[23] for the pairwise alignment, considering two sequences as similar if they possessed more than 70% identical residues in a full-length (60-residue-long) alignment. With this procedure, the final set consisted of 55 sequences.

The sequences used to train the globular submodel were compiled from the well-annotated set of Menne *et al.*,[24] from which we used only secreted proteins (containing a signal peptide) from Gram-positive bacteria that had no predicted transmembrane helices according to TMHMM.[25] From these proteins, we used once again the last 60 residues and the final set included 119 such globular proteins. Finally, in order to train the C-terminal TM submodel, we scrutinized various well-annotated datasets[26–29] 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 extracellular space to the cytoplasm (Out $\rightarrow$ In), in a procedure similar to the one followed in the development of LipoP.[30] 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 the case of closely packed TM segments from multi-spanning TM proteins, we included only the upstream and downstream regions corresponding to the half of the proximal loop (extracellular or cytoplasmic). Consequently, the training set included 50 such pseudo-TM sequences. We should emphasize here that, although these pseudosequences may not be a representative example of C-terminal TM segments, this is of no concern, since our primary intention was to filter out some putative false positives while correctly discriminating LPXTG-anchored proteins, and not to construct a reliable method for predicting TM helices. Since the HMM

consists of different submodels for modeling the three distinct protein classes, the whole dataset ($55 + 119 + 50 = 224$ sequences in total) was used simultaneously in the 25-fold cross-validation procedure (see above).

To further test the rate of false-positive predictions (i.e. specificity), we used a data set of globular proteins with three-dimensional structures deposited in the Protein Data Bank (PDB).[31] This set was compiled using the PAPIA server,[32] with the sequence similarity threshold set to 25%, and excluding membrane proteins, proteins with a length shorter than 80 residues, and proteins with at least one unidentifiable residue in the sequence; finally, this set included 1,100 sequences of globular proteins of various structural classes. For the same reason, we also used the 267 alpha-helical transmembrane proteins with reliable topology (of any type), deposited in the TransMembrane Protein DataBase (TMPDB).[27] These two sets are quite large and representative, and the results on them should offer useful information concerning the specificity of the method. The complete proteomes of Gram-positive bacteria that were used were downloaded from the NCBI website at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. For an additional test of the specificity of the method, we also used two genomes of Gram-negative bacteria, *E. coli* and *S. typhi*, comprised respectively of 4,237 and 4,935 sequences.

2.3. *Comparison to other prediction methods*

In order to compare the results of the HMM that we developed, we used the profile HMM (PF00746) deposited in the PFAM database[17] and the recently compiled profile HMM of Boekhorst *et al.*[11] All of the searches were performed locally using HMMER[33] version 2.3.2, relying on the PFAM trusted cut-off (TC) for the score. In the case of the profile of Boekhorst *et al.*,[11] where no trusted cut-off was available, we used a cut-off for the *E*-value (10^{-6}). We also implemented locally, and consequently used, the tripartite regular expression pattern compiled in a previous work.[15] Finally, in the genome analysis, we used the SignalP v3 web server[34] to predict the putative signal sequences of the proteins tested. We used the neural network (NN) module trained on Gram-positive bacteria, using the default parameters, with submitted sequences truncated to their first 70 residues.

3. Results

In the 25-fold cross-validation procedure, none of the 55 LPXTG proteins was misclassified as either TM or globular (Table 1). Furthermore, no TM or globular proteins were predicted to be LPXTG-anchored ones (no false positives). The 5 out of the 50 TM proteins predicted as globular ones do not constitute a reason for concern, since, as already discussed, the primary intention of the work was to filter the false-positive findings and not to accurately predict the C-terminal TM helices. Thus, the HMM developed here for detecting LPXTG-anchoring proteins shows 100% sensitivity and 100% specificity in the cross-validation procedure. For comparison, in the same data set, the tripartite pattern failed to correctly classify

Table 1. The results obtained through the 25-fold cross-validation procedure.

	Observed		
	Cell wall	Membrane	Globular
Predicted			
Cell wall	55 (100%)	0 (0%)	0 (0%)
Membrane	0 (0%)	45 (90%)	5 (10%)
Globular	0 (0%)	0 (0%)	120 (100%)

the dextranase of *S. mutans* (DEXT_STRMU), leading to a sensitivity of 98.2%. Concerning the PF00746 profile, of the 55 proteins finally used, 13 scored below the TC of 19.1, and one (NISP_LACLA) scored even below 0 (E-value = 0.023). Thus, we do not rely on the TC proposed by the curators of the PFAM database, as we may also miss true positives when using the particular profile HMM. Furthermore, if we disregard the TC and choose to rely on the E-values, we are undertaking the risk of obtaining false positives, since the noise cut-off (NC) for the same profile is set to a score of 19.0, meaning that only hits exceeding this value are not at risk of being false positives. Similar, if not worse, is the situation concerning the profile HMM of Boekhorst *et al.*[11]: of the 55 positive examples, only 23 produced an E-value lower than 10^{-6} and only 34 lower than 10^{-5}. The last of the positive examples produced an E-value of 0.027 with a score of 9.1; and even though no negative example (globular or membrane protein) scored better, this complicates the situation regarding the decision for a discriminative threshold in the genome-wide search. An issue of great importance is the rate of false positives, that is, how often the method we propose predicts LPXTG-anchored proteins when the submitted proteins do not belong to that class. As we stated in Sec. 2.2, for this purpose, we used two large data sets containing soluble proteins with known three-dimensional structure (1,100 sequences) and experimentally verified transmembrane proteins with a varying number of transmembrane segments and topologies (279 sequences). In the two additional data sets that were analyzed, the HMM predictor developed here did not predict even a single false positive (thus giving 100% specificity). This is very encouraging since, even though we cannot rule out the possibility of a false-positive prediction in a large data set, we can be fairly sure that the probability of such a finding will be rather small.

We also analyzed 94 completely sequenced Gram-positive bacterial genomes, using our method as well as the tripartite pattern, the PF00746 profile from PFAM, and the profile of Boekhorst *et al.*[11] The results obtained from selected organisms are presented in Table 2, where we list in parentheses the number of sequences from each genome predicted by SignalP, to possess a cleavable signal sequence. The detailed results from all of the analyzed genomes are deposited in a relational database available as supplementary material in our website (http://bioinformatics.biol.uoa.gr/CW-PRED-results/). The currently developed HMM method predicted a total of 860 proteins in the analyzed genomes (657 with

Table 2. Results obtained in 30 selected completely sequenced genomes of Gram-positive bacteria. We list the results obtained with our HMM, those obtained by the tripartite pattern, the results obtained by the PF00746 profile HMM from PFAM, and the results obtained using the profile HMM of Boekhorst *et al.*[11] Numbers in parentheses correspond to the predicted proteins that additionally possess a signal sequence predicted by SignalP. The complete list of the analyzed genomes as well as detailed information about the predicted sequences can be found in our website (http://bioinformatics.biol.uoa.gr/CW-PRED-results/).

Organism (strain)	Total no. of proteins	HMM (current work)	Tripartite pattern[15]	PF00746[16]	Boekhorst *et al.*[11]
Bacillus anthracis (Ames)	5311	7 (7)	6 (6)	4 (4)	5 (5)
Bacillus cereus (ATCC 10987)	5603	11 (11)	9 (9)	5 (5)	7 (7)
Bacillus halodurans	4066	8 (7)	6 (5)	4 (4)	4 (3)
Bacillus licheniformis (DSM 13)	4196	4 (4)	2 (2)	1 (1)	2 (2)
Bacillus subtilis	4105	2 (1)	2 (1)	1 (0)	0 (0)
Bacillus thuringiensis (konkukian)	5117	9 (9)	8 (8)	6 (6)	7 (7)
Bifidobacterium longum	1727	4 (2)	1 (0)	2 (2)	5 (4)
Clostridium acetobutylicum	3672	2 (2)	2 (2)	0 (0)	2 (2)
Clostridium perfringens	2660	12 (12)	11 (10)	4 (4)	7 (7)
Corynebacterium diphtheriae	2272	8 (8)	5 (5)	1 (1)	10 (10)
Clostridium tetani(E88)	2373	3 (3)	3 (3)	1 (1)	0 (0)
Enterococcus faecalis (V583)	3113	28 (21)	21 (14)	20 (17)	27 (23)
Geobacillus kaustophilus (HTA426)	3498	2 (2)	2 (2)	0 (0)	0 (0)
Lactobacillus johnsonii (NCC 533)	1821	16 (9)	14 (8)	11 (5)	16 (9)
Lactobacillus plantarum	3009	23 (19)	15 (12)	6 (5)	15 (12)
Lactococcus lactis	2321	10 (6)	6 (3)	4 (3)	7 (4)
Listeria monocytogenes	2846	39 (38)	34 (33)	11 (11)	24 (24)
Mycobacterium leprae	1605	2 (1)	0 (0)	0 (0)	0 (0)
Mycobacterium tuberculosis (CDC1551)	4189	2 (1)	0 (0)	0 (0)	0 (0)
Nocardia farcinica (IFM10152)	5683	3 (0)	0 (0)	0 (0)	0 (0)
Staphylococcus aureus (MSSA476)	2579	18 (14)	15 (12)	13 (9)	15 (11)
Staphylococcus epidermidis (RP62A)	2494	10 (7)	9 (6)	7 (5)	5 (3)
Staphylococcus haemolyticus	2676	14 (11)	12 (9)	6 (4)	5 (4)
Staphylococcus saprophyticus	2446	2 (0)	2 (0)	0 (0)	1 (0)
Streptococcus agalactiae (2603)	2124	20 (13)	20 (12)	10 (7)	18 (13)
Streptococcus mutans	1960	6 (5)	5 (5)	4 (4)	4 (3)
Streptococcus pneumoniae (R6)	2043	15 (9)	12 (7)	9 (6)	8 (5)
Streptococcus pyogenes (MGAS6180)	1894	24 (18)	22 (16)	14 (10)	13 (8)
Streptomyces avermitilis	7577	3 (2)	6 (1)	0 (0)	8 (7)
Thermobifida fusca (YX)	3110	6 (6)	7 (6)	1 (1)	1 (1)
Total	98090	313 (248)	257 (197)	145 (115)	216 (174)

a predicted signal sequence); the tripartite pattern, 723 proteins (551 with a signal sequence); the PF00746 profile, 445 proteins (331 with a predicted signal peptide); and the profile HMM compiled by Boekhorst *et al.*,[11] 618 proteins (431 with a predicted signal peptide). From these results, it is obvious that the HMM method that we developed predicts a significantly larger number of CWPs compared to the tripartite pattern, the PF00746 profile, or the profile of Boekhorst *et al.*[11] The differences are statistically significant, according to the Wilcoxon matched-pairs signed-ranks test (p-value < 0.001 in all cases). Since we have shown that the HMM method is more sensitive, without at the same time lacking any specificity, we can be fairly sure that the majority of the overpredicted proteins are truly novel CWPs, not easily identified by any other available method. Considering the PF00746 profile, even when we did not use the TC but relied on the E-value ($< 10^{-6}$) instead, we obtained no more than 80 additional predictions (525 proteins in total). Even in such a case, the PFAM profile still predicted the smallest number of proteins, and the majority of these additional sequences are true-positive hits missed by the TC. This fact highlights the inherent problems present when using approaches based on E-values, and it is in contradiction to our approach which directly predicts the C-terminal anchor without the need to choose a particular threshold. Concerning the profile HMM of Boekhorst *et al.*,[11] when we considered sequences that score above zero (instead of using the E-value), we came out with a total of 1,237 sequences (913 with a signal peptide), a number that is significantly larger compared to all other available methods.

In order to be able to evaluate these results, we proceeded with an analysis of the annotations of the identified proteins. Of the 860 proteins identified with the use of our method, 356 (41.4%) had an annotation suggesting a definite localization to the bacterial cell wall (cell wall-bound, cell wall-anchored, etc.) or belong to families of proteins that are known to be LPXTG-bound (C5A peptidase, dextranase, sialidase, M protein, etc.). Furthermore, 110 proteins (12.8%) belong to the same category, having, however, annotations such as putative, probable, or possible. We also identified 100 (mostly extracellular) enzymes (11.6%) without, however, having any indication as to whether these proteins are cell wall-bound or not; and 47 other proteins (5.5%) of various annotations that may be LPXTG-bound proteins, but that may also constitute false-positive findings. We also identified 237 hypothetical proteins (27.5%) having absolutely no annotation concerning their function or localization. Finally, 10 proteins (1.2%) possessed an annotation suggesting that they were putative or known transmembrane proteins. If these annotations are correct, then these proteins should be considered as the sole false-positive findings of our method. In total, only 14 out of the 860 predicted proteins (1.62%) possessed more than three transmembrane helices predicted by TMHMM. Of the 1,237 proteins that scored above zero using the profile HMM of Boekhorst *et al.*,[11] only 375 had a definite annotation suggesting a localization to the outer surface. Most importantly, a number of at least 72 proteins annotated as transmembrane were predicted as anchored ones (5.8%). In total, 101 sequences (8.16%) possessed more than three

transmembrane helices predicted by TMHMM. Thus, it is clear that, despite the fact that the method of Boekhorst *et al.*[11] (using the score cut-off) predicts a larger number of proteins, a significant portion of them are clearly false-positive findings. On the other hand, the particular method is much more sensitive using the *E*-value cut-off at the cost of reduced specificity. Lastly, we have to mention that the method developed here was also tested on two genomes of Gram-negative bacteria (*E. coli* and *S. typhi*) and produced only one false-positive finding (0.01% in total), suggesting once again a high level of specificity.

When the 237 hypothetical (according to the genome annotation) proteins that we identified were used in a BLAST[23] search against the SwissProt subset of the UniProt database,[21] 6 (2.5%) were found to have a perfect match, 143 (60.3%) had a significant hit (E-value $< 10^{-6}$), but, most importantly, 88 (37.2%) were not found to have a single homolog. These proteins could not have been identified using a simple similarity search — this fact indicates the importance of our approach, which specifically identifies the C-terminal anchor region. These previously uncharacterized proteins should also be further investigated, as they may constitute novel LPXTG-bound proteins and may provide evidence for novel families of such proteins. The list of the 88 proteins is available at http://bioinformatics.biol.uoa.gr/CW-PRED-results/.

The N-terminal secretory signal sequences of the predicted LPXTG proteins were also tested in order to identify if they contain the conserved sequence (Y/F)SIRK. There were only about 12% of predicted LPXTG proteins containing this motif. This search confirmed previous observations suggesting that, although common among LPXTG protein signal peptides, this motif is neither exclusive nor universal.[8,35]

4. Discussion

We have presented a hidden Markov model approach for predicting the LPXTG-anchored cell wall proteins of Gram-positive bacteria. The model was trained in order to discriminate the C-terminal sequence of such proteins from both C-terminal TM helices and soluble globular domains. We have shown that the method is more sensitive than other currently available methods, without any lack of specificity. Compared to the tripartite regular expression pattern, the HMM is far more flexible, as it allows different length distributions of the various regions as well as different amino acid compositions. Compared to the profile HMMs (PF00746 and that of Boekhorst *et al.*[11]), our method is more parsimonious, having a significantly smaller number of free parameters, while at the same time it does not involve any arbitrary choice of cut-off for the score or the *E*-value. Thus, the HMM which we developed produces an outcome that directly discriminates LPXTG proteins from TM or globular ones, and consequently it does not need any choice of cut-off value. The prediction method that we developed is freely accessible to the scientific community at http://bioinformatics.biol.uoa.gr/CW-PRED/. We also performed a comprehensive analysis of the currently available, completely sequenced genomes of

Gram-positive bacteria, comparing our method against the already existing ones. Such an effort was performed on a large scale for the first time, and the results advocate in favor of the newly proposed methodology. We have provided evidence for dozens of previously uncharacterized CWPs in the completely sequenced bacterial genomes that may lead to the discovery of novel families of cell wall-bound proteins. Further work has to be done in the future in order to classify these proteins according to their functional properties, which are essential for Gram-positive bacteria.

The constant finding that a fraction of 75%–80% of the predicted (by all methods discussed here) LPXTG-anchored proteins shows evidence for the presence of a signal peptide warrants special attention. It is well known that such proteins should possess a signal sequence necessary for the secretion. However, the inability to identify such a sequence in all of these proteins may not be necessarily taken as evidence against their localization in the cell wall, but instead may indicate sequencing errors (such as wrong annotation of the N-terminal[34]), wrongly assigned coding sequences in the genome annotation,[36] or prediction errors (false negatives) produced by SignalP. For instance, it has been shown[37] that some LPXTG-anchored proteins possess an unusually long signal sequence (>80 residues long) that cannot be easily predicted since, by default, SignalP restricts its predictions to the first 70 residues of the submitted sequences. Using these proteins on the PrediSi prediction method[38] or the SignalP without truncating the proteins to their first 70 amino acids, we found that 78 out of the 203 (38.4%) indeed possess evidence for the presence of a signal peptide. Furthermore, 38 additional proteins (18.7%) provide evidence based on TMHMM[25] and HMMTOP[39] that indicates the presence of an N-terminal transmembrane helix, a fact that possibly indicates the existence of a signal peptide which was initially missed by the prediction methods. We should note here that only 15 out of the 203 proteins (7.4%) contain more than three predicted transmembrane segments, and the majority of them were included in the proteins with a "probable" annotation as integral membrane proteins which we discarded as "false positives" in the previous section. It has been also shown that a large portion of the open reading frames (20%–40%) that supposedly code for proteins, especially in the bacterial genomes, are in fact overannotations, mainly due to errors in the gene-finding procedure.[36] Thus, even though we cannot exclude the possibility of some false-positive findings among our results, we can be fairly sure that most of them originate from other errors and are not solely due to the particular prediction method. Lastly, we have to mention that the N-terminal (Y/F)SIRK motif cannot be used as a determinant of specificity of cell wall anchored proteins as it is found to be poorly conserved.

It has been recently shown that, apart from the typical LPXTG pattern which is responsible for cleavage by sortase, other similar patterns do exist that are responsible for cleavage by other sortases. For instance, sortase B (SrtB) from *Staphylococcus aureus* recognizes the motif NPQTN, as already mentioned.[3] However, large data sets of such experimentally verified proteins are still pending, and thus we cannot incorporate them in the prediction method. In a recent

work, Comfort and Clubb[9] devised a strategy using similarity searches in order to find all of the available sortase enzymes in 72 bacterial genomes; they subsequently built profile HMMs for each one of the five categories. Using a combination of hydropathy analysis, signal peptide prediction, and pattern searches, they totally identified 892 CWPs. Despite the fact that these figures are not too far from our estimates, these results may be biased towards hypothetical data, since there are not enough experimentally verified SrtB substrates. At this point, we should emphasize that, apart from the documented overprediction using the Boekhorst *et al.*[11] method, our method is the only one capable of predicting alternative sortase substrates. However, as we have already mentioned, these proteins have not been experimentally verified to be sortase substrates. We anticipate that available data will emerge in the near future, and the HMM could easily be expanded to include them in a prediction system and to allow for a proper discrimination of the cleavage specificity between different types of sortases. A possible extension of the proposed HMM could be formulated by constructing different submodels (within the cell wall anchor submodel) to account for different amino acid preferences in the sortase recognition signal (i.e. providing states with an alternative NPQTN cleavage site in addition to the LPXTG site; see Fig. 2). The HMMs allow such modification due to their modular architecture, although a minor problem is the fact that we would have to set the appropriate transitions by hand instead of performing training. This would be necessary, since the limited amount of data prohibits the idea of maximum likelihood training.

Acknowledgments

The authors would like to thank the two anonymous reviewers for their valuable comments and the constructive criticism that helped to improve the quality of the manuscript. P. G. Bagos would like to thank the State Scholarship Foundation (SSF) of Greece for the financial support of the project "Machine Learning Algorithms for Bioinformatics".

References

1. Foster TJ, Hook M, Surface protein adhesins of *Staphylococcus aureus*, *Trends Microbiol* **6**(12):484–488, 1998.
2. Lee SG, Pancholi V, Fischetti VA, Characterization of a unique glycosylated anchor endopeptidase that cleaves the LPXTG sequence motif of cell surface proteins of Gram-positive bacteria, *J Biol Chem* **277**(49):46912–46922, 2002.
3. Mazmanian SK, Ton-That H, Schneewind O, Sortase-catalysed anchoring of surface proteins to the cell wall of *Staphylococcus aureus*, *Mol Microbiol* **40**(5):1049–1057, 2001.
4. Cabanes D, Dehoux P, Dussurget O, Frangeul L, Cossart P, Surface proteins and the pathogenic potential of Listeria monocytogenes, *Trends Microbiol* **10**(5):238–245, 2002.

5. Navarre WW, Schneewind O, Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope, *Microbiol Mol Biol Rev* **63**(1):174–229, 1999.
6. Ton-That H, Mazmanian SK, Faull KF, Schneewind O, Anchoring of surface proteins to the cell wall of *Staphylococcus aureus*. Sortase catalyzed *in vitro* transpeptidation reaction using LPXTG peptide and NH_2-Gly_3 substrates, *J Biol Chem* **275**(13):9876–9881, 2000.
7. Fischetti VA, Pancholi V, Schneewind O, Conservation of a hexapeptide sequence in the anchor region of surface proteins from Gram-positive cocci, *Mol Microbiol* **4**(9):1603–1605, 1990.
8. Roche FM, Massey R, Peacock SJ, Day NP, Visai L, Speziale P, Lam A, Pallen M, Foster TJ, Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences, *Microbiology* **149**(Pt 3):643–654, 2003.
9. Comfort D, Clubb RT, A comparative genome analysis identifies distinct sorting pathways in Gram-positive bacteria, *Infect Immun* **72**(5):2710–2722, 2004.
10. Ton-That H, Marraffini LA, Schneewind O, Protein sorting to the cell wall envelope of Gram-positive bacteria, *Biochim Biophys Acta* **1694**(1–3):269–278, 2004.
11. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ, Genome-wide detection and analysis of cell wall-bound proteins with LPXTG-like sorting motifs, *J Bacteriol* **187**(14):4928–4934, 2005.
12. Schneewind O, Model P, Fischetti VA, Sorting of protein A to the staphylococcal cell wall, *Cell* **70**(2):267–281, 1992.
13. Schneewind O, Mihaylova-Petkov D, Model P, Cell wall sorting signals in surface proteins of Gram-positive bacteria, *EMBO J* **12**(12):4803–4811, 1993.
14. Marraffini LA, Dedent AC, Schneewind O, Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria, *Microbiol Mol Biol Rev* **70**(1):192–221, 2006.
15. Janulczyk R, Rasmussen M, Improved pattern for genome-based screening identifies novel cell wall-attached proteins in Gram-positive bacteria, *Infect Immun* **69**(6):4019–4026, 2001.
16. Bagos PG, Liakopoulos TD, Hamodrakas SJ, Algorithms for incorporating prior topological information in HMMs: Application to transmembrane proteins, *BMC Bioinformatics* **7**(1):189, 2006.
17. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A, Pfam: Clans, web tools and services, *Nucleic Acids Res* **34**(Database issue):D247–D251, 2006.
18. Krogh A, Hidden Markov models for labelled sequences, *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pp. 140–144, 1994.
19. Durbin R, Eddy SR, Krogh A, Mithison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.
20. Fariselli P, Martelli PL, Casadio R, A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins, *BMC Bioinformatics* **6**(Suppl 4):S12, 2005.
21. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS, The Universal Protein Resource (UniProt), *Nucleic Acids Res* **33**(Database issue):D154–D159, 2005.
22. Hobohm U, Scharf M, Schneider R, Sander C, Selection of representative protein data sets, *Protein Sci* **1**(3):409–417, 1992.

23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res* **25**(17):3389–3402, 1997.
24. Menne KM, Hermjakob H, Apweiler R, A comparison of signal sequence prediction methods using a test set of signal peptides, *Bioinformatics* **16**(8):741–742, 2000.
25. Krogh A, Larsson B, von Heijne G, Sonnhammer EL, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol* **305**(3):567–580, 2001.
26. Moller S, Kriventseva EV, Apweiler R, A collection of well characterised integral membrane proteins, *Bioinformatics* **16**(12):1159–1160, 2000.
27. Ikeda M, Arai M, Okuno T, Shimizu T, TMPDB: A database of experimentally-characterized transmembrane topologies, *Nucleic Acids Res* **31**(1):406–409, 2003.
28. Chen CP, Rost B, Long membrane helices and short loops predicted less accurately, *Protein Sci* **11**(12):2766–2773, 2002.
29. Jayasinghe S, Hristova K, White SH, MPtopo: A database of membrane protein topology, *Protein Sci* **10**(2):455–458, 2001.
30. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A, Prediction of lipoprotein signal peptides in Gram-negative bacteria, *Protein Sci* **12**(8):1652–1662, 2003.
31. Berman HM *et al.*, The Protein Data Bank, *Acta Crystallogr D Biol Crystallogr* **58**(Pt 6 No 1):899–907, 2002.
32. Noguchi T, Akiyama Y, PDB-REPRDB: A database of representative protein chains from the Protein Data Bank (PDB) in 2003, *Nucleic Acids Res* **31**(1):492–493, 2003.
33. Eddy SR, Profile hidden Markov models, *Bioinformatics* **14**(9):755–763, 1998.
34. Bendtsen JD, Nielsen H, von Heijne G, Brunak S, Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol* **340**(4):783–795, 2004.
35. Tettelin H *et al.*, Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*, *Science* **293**(5529):498–506, 2001.
36. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A, On the total number of genes and their length distribution in complete microbial genomes, *Trends Genet* **17**(8):425–428, 2001.
37. Bensing BA, Sullam PM, An accessory *sec* locus of *Streptococcus gordonii* is required for export of the surface protein GspB and for normal levels of binding to human platelets, *Mol Microbiol* **44**(4):1081–1094, 2002.
38. Hiller K, Grote A, Scheer M, Munch R, Jahn D, PrediSi: Prediction of signal peptides and their cleavage positions, *Nucleic Acids Res* **32**(Web Server issue):W375–W379, 2004.
39. Tusnady GE, Simon I, The HMMTOP transmembrane topology prediction server, *Bioinformatics* **17**(9):849–850, 2001.

Zoi I. Litou received her B.Sc. in Cell and Molecular Biology in 2001 from the University of Essex, Colchester, UK. She is currently a Ph.D. student in the Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens, Greece. She is currently working on computational analysis of membrane proteins, focusing on the automated recognition and classification of single-spanning membrane proteins.

Pantelis G. Bagos received his B.Sc. in Biology in 1997, his M.Sc. in Biostatistics in 2002, and his Ph.D. in Bioinformatics in 2005, all from the University of Athens, Athens, Greece. Since then, he has been a post-doctoral research fellow at the Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens; and a visiting Assistant Professor at the University of Central Greece, Lamia, Greece, where he is currently teaching Biology and Bioinformatics. His research interests include computational analysis of protein sequences, biological databases, machine learning algorithms in bioinformatics, and genetic epidemiology.

Konstantinos D. Tsirigos is a final-year undergraduate student at the Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens, Greece. He is currently preparing his Diploma Thesis in Bioinformatics. His research interests include the development of biological databases, Internet programming, and computational analysis of protein sequences.

Theodore D. Liakopoulos is an independent computer software professional. He has a B.Sc. (Biology) and an M.Sc. (Bioinformatics) from the University of Athens, Athens, Greece. His scientific interests include machine learning approaches to model biological sequences, information extraction from large data sets, and software integration.

Stavros J. Hamodrakas received his B.Sc from the Physics Department of the University of Athens, Athens, Greece, in 1970; and his Ph.D. from the Astbury Department of Biophysics of the University of Leeds, Leeds, U.K., in 1974. He is currently a full Professor at the Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens. He is also the General Director of a post-graduate program in "Bioinformatics". His research interests include structural and functional studies of insect chorion (eggshell) and insect cuticle; prediction of protein structure, function, and interactions; creation of relational and object-oriented specialized protein databases; automatic analyses of genomes; study of fibrous and globular protein structure, function, and interactions; and studies of structure and self-assembly mechanisms of amyloids.