CW-PRED: A HMM-Based Method for the Classification of Cell Wall-Anchored Proteins of Gram-Positive Bacteria

Danai K. Fimereli[1], Konstantinos D. Tsirigos[1], Zoi I. Litou[1],
Theodore D. Liakopoulos[2], Pantelis G. Bagos[2], and Stavros J. Hamodrakas[1]

[1] Department of Cell Biology and Biophysics, Faculty of Biology,
University of Athens, Athens 157 01, Greece
fdanai@gmail.com, {ktsirig,zlitou,shamodr}@biol.uoa.gr
[2] Department of Computer Science and Biomedical Informatics,
University of Central Greece, Papasiopoulou 2-4 Lamia 35100, Greece
{liakop,pbagos}@ucg.gr

Abstract. Gram-positive bacteria have surface proteins that are often implicated in virulence. A group of extracellular proteins attached to the cell wall contains an LPXTG-like motif that is target for cleavage and covalent coupling to peptidoglycan by sortase enzymes. A Hidden Markov Model (HMM) was developed for predicting the LPXTG and LPXTG-like cell-wall proteins of Gram-positive bacteria. The model is the first capable of predicting alternative (i.e. other than LPXTG-containing) substrates. Our analysis of 177 completely sequenced genomes identified 1456 cell-wall proteins, a number larger compared to the previously available methods. Among these, apart from the previously identified 1283 proteins carrying the LPXTG motif, we identified 39 newly identified proteins carrying NPXTG, 53 carrying LPXTA and 81 carrying the LAXTG motif. The tool is freely available for academic use at http://bioinformatics.biol.uoa.gr/CW-PRED/.

Keywords. Gram-positive bacteria, cell-wall proteins, sortase substrates, LPXTG-like motifs, Hidden Markov Models, proteome analysis.

1 Introduction

Surface proteins of pathogenic bacteria carry out many important functions including invasion of host cells, evasion of the immune response and adhesion to the site of infection and may be used as drugs or vaccine targets [1]. Most of the cell-wall attached proteins have a conserved C-terminal region containing an LPXTG motif, which is required for linking to the cell wall envelope[2, 3]. The C-terminal signal, required for the sorting of the protein to the cell wall, consists of the LPXTG sequence motif (where X denotes any amino acid), followed by a hydrophobic domain and a short positively charged tail[4-6]. Membrane-associated transpeptidases, called sortases, are responsible for the covalent attachment of the LPXTG-like proteins to the Gram-positive bacterial cell wall. Sortases cleave their protein substrate between the threonine (Thr) and glycine (Gly) residues of the LPXTG motif [7] and an amide

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 285–290, 2012.
© Springer-Verlag Berlin Heidelberg 2012

bond is formed between the C-terminal of the threonine and the amino group of the pentaglycine cross-bridge of peptidoglycan. The hydrophobic region of the sorting signal then passes through the plasma membrane and, together with the charged tail, both act as a stop transfer signal [2, 3, 8, 9]. The surface protein linked to peptidoglycan is then displayed on the microbial surface [10].

Apart from SrtA which cleaves LPXTG substrates, it has been shown that other sortases can process proteins that do not fit to the canonical pattern[11-13]. For example, SrtB from *Staphylococcus aureus* recognizes the NPQTN motif [14], whereas SrtC recognizes the LPXTA motif in proteins of *Bacillus anthracis* [10]. Traditionally, LPXTG-like proteins were predicted using regular expression patterns and Hidden Markov Models. Among these methods, CW-PRED has been shown to be the most successful both in terms of sensitivity and specificity [15]. However, the limited number of experimentally verified non-canonical substrates limits also the applicability of such methods in detecting other sortase substrates. This work presents a HMM model, that extends the previous model developed by Litou and coworkers [15] for predicting LPXTG-like cell-wall proteins of Gram-positive bacteria.

2 Materials and Methods

For training the initial version of the model [15], 55 experimentally verified proteins were used, none of which had a sorting signal that differed from the canonical LPXTG motif (SrtA substrates) [16]. In order to extend the model, we performed an extensive literature search in order to find experimentally verified surface proteins. We scrutinized more than 100 published articles published up to October 2010 and we requested experimental evidence for the localization of the protein to the cell-surface. The sequences of these cell-wall anchored proteins were subsequently retrieved from the UNIPROT database, version 14 [17].

We also considered previously described datasets [3] from which we extracted 65 additional experimentally verified cell-wall anchoring proteins of Gram-positive bacteria. After redundancy reduction we finally came up with a total of 132 proteins (the largest set ever compiled), of which 122 had the canonical LPXTG motif, 5 had the NPXTG motif, 3 had the LAXTG motif and 2 had the LPXTA motif. Even though LAXTG-containing proteins are most likely cleaved by SrtA, we considered them as a separate category for computational convenience. In the HMM sub-model that corresponds to the cleavage site pattern, we created four additional branches consisting of states that model the LPXTG, NPXTG, LAXTG and LPXTA variants, while the rest of the states remained the same. The old HMM model was parsimonious in terms of the number of freely estimated parameters, and it has proved to be very sensitive and specific. Thus, we updated only the transition end emission parameters of the HMM that correspond to these states, whereas the other model parameters remained unchanged. The model architecture is shown in Figure 1.

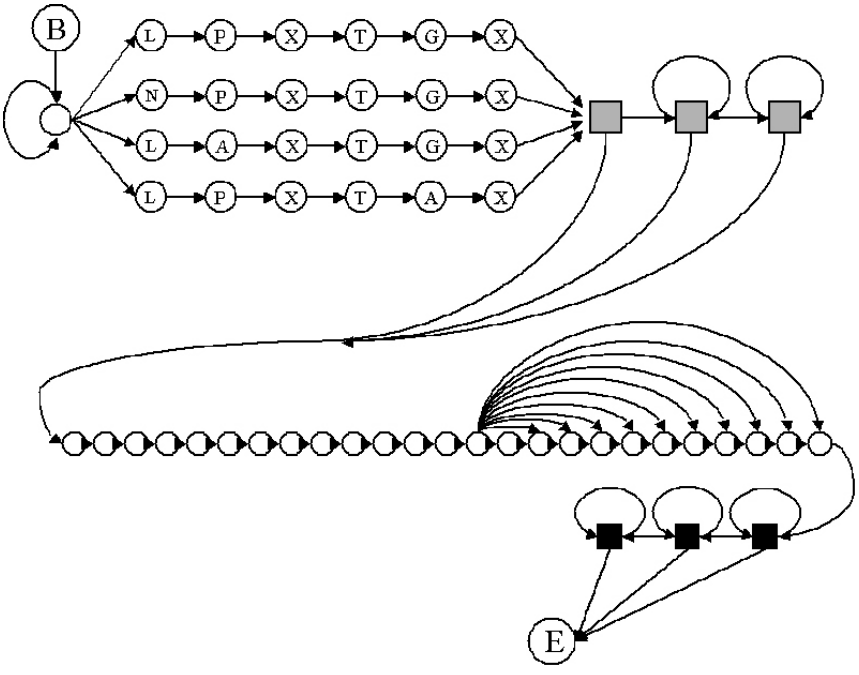

Fig. 1. A graphical representation of the LPXTG anchor submodel of the HMM. Different states are used sequentially to model the LPXTG cleavage site, the variable-length gap region, the hydrophobic helix, and the positively charged tail. Arrows represent allowed transitions among states, and states having 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. Different states are used sequentially to model the LPXTG cleavage site (upper left section), the variable-length gap region (upper right section; gray squared states), the hydrophobic helix (middle section; small pointed circle states), and the positively charged tail (bottom dark squared states) [15].

3 Results

The original version of the predictor performs already very well in predicting experimentally verified cell-wall anchored proteins (100% specificity and sensitivity in a 25-fold cross-validation procedure) and this is also the case for the updated version. Due to the lack of an independent test set, we could not measure performance quantitatively. Therefore, as an alternative, we analyzed 177 completely sequenced Gram-positive bacterial genomes retrieved from the NCBI, in order to test the method's prediction accuracy on large, 'unknown' datasets. The detailed results are deposited in a database available as supplementary material in http://bioinformatics.biol.uoa.gr/CW-PRED-results/. The updated HMM predictor identified a total of 1456 proteins, a number larger compared to previously available methods [15]. Apart from the 1283 proteins carrying the LPXTG motif that could be

identified by the old model too, by using the new model we identified and classified 39 additional proteins carrying NPXTG, 53 carrying LPXTA and 81 carrying the LAXTG motif.

Table 1. Novel proteins experimentally verified as sortase substrates that are correctly predicted by the CW-PRED algorithm

Protein from Davies, et al. 2009 [18] (Uniprot AC)	Prediction by CW-PRED	Protein from Egan, et al., 2010 [19] (Uniprot AC)	Prediction by CW-PRED
SGO_0854 (A8AWJ3)	Sortase A	sub0135 (B9DT15)	Sortase A
SGO_1148 (A8AXC5)	Sortase A	sub0145 (B9DT25)	Sortase A
SGO_0707 (A8AW49)	Sortase A	sub0207 (B9DT84)	Sortase A
SGO_0210 (A8AUS0)	Sortase A	sub0826 (B9DS05)	Sortase A
SGO_0211 (A8AUS1)	Sortase A	sub0888 (B9DUA9)	Sortase A
SGO_1487 (A8AYA6)	Sortase A	sub1095 (B9DSF3)	Sortase A
SGO_1247 (A8AXM1)	Sortase A	sub1154 (B9DSH4)	Sortase A
SGO_0890 (A8AWM6)	Sortase A	sub1370 (B9DV27)	Sortase A
SGO_2005 (A8AZP4)	Sortase A	sub1730 (B9DW17)	Sortase A
SGO_0966 (A8AWU7)	Sortase A	sub0135 (B9DT15)	Sortase A
SGO_0208 (A8AUR8)	Sortase C	sub0145 (B9DT25)	Sortase A
SGO_0317 (A8AV26)	Sortase A	sub0207 (B9DT84)	Sortase A
SGO_0316 (A8AV25)	Sortase A	sub0826 (B9DS05)	Sortase A
SGO_0388 (A8AV94)	Sortase A	sub0888 (B9DUA9)	Sortase A
SGO_1415 (A8AY35)	Sortase A	sub1095 (B9DSF3)	Sortase A
SGO_0107 (A8AUG9)	Sortase A	sub1154 (B9DSH4)	Sortase A
SGO_0430 (A8AVD6)	Sortase A	AAM99322 (Q8E1E1)	Sortase A
SGO_2004 (A8AZP3)	Sortase A	AAN00204 (Q8DYY9)	Sortase A
SGO_1651 (A8AYR8)	Sortase A	CAW99349 (C0MFU4)	Sortase A
SGO_1650 (A8AYR7)	Sortase A	CAW94597 (C0MAN5)	Sortase A
SGO_1182 (A8AXF9)	Sortase A	CAW92812 (C0M9K8)	Sortase A
		CAW92309 (C0MAH2)	Sortase C
		AAL00574 (Q8DNF3)	Sortase A
		AAT87853 (Q5X9R0)	Sortase A
		AAL97965 (Q8P0G8)	Sortase A

Of the 1456 proteins identified, 778 (53.4%) had an annotation suggesting a definite localization to the bacterial cell wall (cell-wall bound, anchored, LPXTG-bound etc), or belong to families of proteins known to be LPXTG-bound (C5A peptidase, dextranase, sialidase, M protein etc). Furthermore, 171 proteins (11.7%) belong to the same category, having however annotations such as putative, probable or possible. We also identified 40 (mostly extracellular) enzymes (2.8%) without however having any indication as to whether these proteins are cell wall-bound or not, and 30 other proteins (2.1%) of various annotations that may also be LPXTG-bound proteins, but they may also constitute false positive findings. We also identified 419 hypothetical

proteins (28.8%), having absolutely no annotation concerning their function or localization. Finally, only 18 proteins (1.2%) possessed an annotation suggesting that they were putative or known transmembrane proteins; this figure should be considered as an estimate for the false positive prediction rate of the method. The majority of proteins with non-canonical motifs are newly identified cell-wall anchored proteins that are presumably cleaved by sortases with different substrate specificities [18]. The annotation was carried out based on the respective Uniprot entries of the proteins.

4 Discussion

We presented a HMM model for predicting the LPXTG-like cell-wall proteins of Gram-positive bacteria. In contrast to the previous HMM model [15], the new model predicts more proteins that contain all possible motifs in their carboxy-terminal, namely the NPXTG, LAXTG and LPXTA motifs, while these proteins are regarded to be cleaved by different sortases. When evaluated at the 94 sequenced genomes previously analyzed [15], the new updated nethod is better at detecting proteins containing non-canonical sortase substrates. Additionally, 83 newly sequenced genomes were also analyzed (a total of 177 sequenced genomes) and proteins having all possible motifs were detected (http://bioinformatics.biol.uoa.gr/CW-PRED-results/). Most importantly, proteins not included in the training set that were identified recently in different organisms using experimental methods [19, 20] were also predicted as sortase substrates by our model (Table 1).

Taken together, these findings suggest that CW-PRED is a reliable tool for predicting cell-wall proteins of Gram-positive bacteria that contain all possible motifs in their C-terminal. The user may submit either a single sequence and receive detailed results or multiple sequences (up to 1000 per submission) and receive summary prediction in an easily readable format. The prediction method (along with the training set used) and the results from the analysis are freely available for academic users at http://bioinformatics.biol.uoa.gr/CW-PRED/. As far as computational requirements is concerned, a benchmark test on an Intel Xeon CPU server machine with 4GB of RAM memory showed that it takes approximately 15 minutes for a submission of 500 protein sequence entries. This shows that CW-PRED can be used quite efficiently for large genome analysis projects as well.

References

1. Lee, S.G., Pancholi, V., Fischetti, V.A.: 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, 46912–46922 (2002)
2. Cabanes, D., Dehoux, P., Dussurget, O., Frangeul, L., Cossart, P.: Surface proteins and the pathogenic potential of Listeria monocytogenes. Trends Microbiol. 10, 238–245 (2002)
3. Navarre, W.W., Schneewind, O.: Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol. Mol. Biol. Rev. 63, 174–229 (1999)

4. Fischetti, V.A., Pancholi, V., Schneewind, O.: Conservation of a hexapeptide sequence in the anchor region of surface proteins from gram-positive cocci. Mol. Microbiol. 4, 1603–1605 (1990)
5. Marraffini, L.A., Dedent, A.C., Schneewind, O.: Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria. Microbiol. Mol. Biol. Rev. 70, 192–221 (2006)
6. Roche, F.M., Massey, R., Peacock, S.J., Day, N.P., Visai, L., Speziale, P., Lam, A., Pallen, M., Foster, T.J.: Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences. Microbiology 149, 643–654 (2003)
7. Guttilla, I.K., Gaspar, A.H., Swierczynski, A., Swaminathan, A., Dwivedi, P., Das, A., Ton-That, H.: Acyl enzyme intermediates in sortase-catalyzed pilus morphogenesis in gram-positive bacteria. J. Bacteriol. 191, 5603–5612 (2009)
8. Mazmanian, S.K., Ton-That, H., Su, K., Schneewind, O.: An iron-regulated sortase anchors a class of surface protein during Staphylococcus aureus pathogenesis. Proc. Natl. Acad. Sci. U S A 99, 2293–2298 (2002)
9. Ton-That, H., Mazmanian, S.K., Faull, K.F., 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, 9876–9881 (2000)
10. Marraffini, L.A., Schneewind, O.: Targeting proteins to the cell wall of sporulating Bacillus anthracis. Mol. Microbiol. 62, 1402–1417 (2006)
11. Maresso, A.W., Schneewind, O.: Sortase as a target of anti-infective therapy. Pharmacol. Rev. 60, 128–141 (2008)
12. Ton-That, H., Marraffini, L.A., Schneewind, O.: Protein sorting to the cell wall envelope of Gram-positive bacteria. Biochim. Biophys. Acta 1694, 269–278 (2004)
13. Zhou, M., Boekhorst, J., Francke, C., Siezen, R.J.: LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics 9, 173 (2008)
14. Boekhorst, J., de Been, M.W., Kleerebezem, M., Siezen, R.J.: Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. J. Bacteriol. 187, 4928–4934 (2005)
15. 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. 6, 387–401 (2008)
16. Hobohm, U., Scharf, M., Schneider, R., Sander, C.: Selection of representative protein data sets. Protein Science: A Publication of the Protein Society 1, 409–417 (1992)
17. 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. 34, D187–D191 (2006)
18. Comfort, D., Clubb, R.T.: A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. Infect Immun. 72, 2710–2722 (2004)
19. Davies, J.R., Svensater, G., Herzberg, M.C.: Identification of novel LPXTG-linked surface proteins from Streptococcus gordonii. Microbiology 155, 1977–1988 (2009)
20. Egan, S.A., Kurian, D., Ward, P.N., Hunt, L., Leigh, J.A.: Identification of sortase A (SrtA) substrates in Streptococcus uberis: evidence for an additional hexapeptide (LPXXXD) sorting motif. J. Proteome Res. 9, 1088–1095 (2010)