BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 22 2010, pages 2811–2817
doi:10.1093/bioinformatics/btq530

Sequence analysis Advance Access publication September 16, 2010

Combined prediction of Tat and Sec signal peptides with hidden Markov models

Pantelis G. Bagos*, Elisanthi P. Nikolaou, Theodore D. Liakopoulos and Konstantinos D. Tsirigos

Department of Computer Science and Biomedical Informatics, University of Central Greece, Papasiopoulou 2–4, Lamia 35100, Greece

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Computational prediction of signal peptides is of great importance in computational biology. In addition to the general secretory pathway (Sec), Bacteria, Archaea and chloroplasts possess another major pathway that utilizes the Twin-Arginine translocase (Tat), which recognizes longer and less hydrophobic signal peptides carrying a distinctive pattern of two consecutive Arginines (RR) in the *n*-region. A major functional differentiation between the Sec and Tat export pathways lies in the fact that the former translocates secreted proteins unfolded through a protein-conducting channel, whereas the latter translocates completely folded proteins using an unknown mechanism. The purpose of this work is to develop a novel method for predicting and discriminating Sec from Tat signal peptides at better accuracy.

Results: We report the development of a novel method, PRED-TAT, which is capable of discriminating Sec from Tat signal peptides and predicting their cleavage sites. The method is based on Hidden Markov Models and possesses a modular architecture suitable for both Sec and Tat signal peptides. On an independent test set of experimentally verified Tat signal peptides, PRED-TAT clearly outperforms the previously proposed methods TatP and TATFIND, whereas, when evaluated as a Sec signal peptide predictor compares favorably to top-scoring predictors such as SignalP and Phobius. The method is freely available for academic users at http://www.compgen.org/tools/PRED-TAT/.

Contact: pbagos@ucg.gr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on 2 August 2010; revised on 4 September 2010; accepted on 10 September 2010

1 INTRODUCTION

In all domains of life (Bacteria, Eukarya and Archaea), the majority of secreted proteins are synthesized as precursors that carry a cleavable N-terminal signal sequence. The signal peptide possesses a modular architecture with a positively charged region at the n-terminus (*n*-region), a hydrophobic region (h-region) that spans the membrane and a c-region of mostly small and uncharged residues ending at the characteristic cleavage site (von Heijne, 1990). The signal peptide is necessary for targeting the protein to the membrane-embedded export machinery in Bacteria (Driessen and Nouwen, 2007), Eukaryotes (Rapoport *et al.*, 1999) and Archaea (Pohlschroder *et al.*, 2005). Upon translocation across the membrane, the signal peptide is cleaved from the precursor via a membrane-bound signal peptidase (Tuteja, 2005; van Roosmalen *et al.*, 2004). In Eukaryotes, proteins targeted to the organelles of bacterial origin (mitochondria and chloroplasts) also contain cleavable N-terminal targeting sequences, although they are in general different from those found in the eukaryotic or bacterial secreted proteins (Habib *et al.*, 2007; von Heijne *et al.*, 1989).

In addition to the general export pathway (Sec), Bacteria, Archaea and chloroplasts possess another major pathway that utilizes the Twin-Arginine translocase (Tat). Tat recognizes longer and less hydrophobic signal peptides carrying a distinctive pattern of two consecutive Arginines (RR) in the *n*-region (Berks *et al.*, 2005; Lee *et al.*, 2006; Teter and Klionsky, 1999). A major functional differentiation between Sec and Tat export pathways lies in the fact that the former translocates secreted proteins unfolded through a protein-conducting channel, whereas the latter translocates completely folded proteins using an unknown mechanism (Teter and Klionsky, 1999). Interestingly, in halophilic Archaea, the components of the Tat pathway are essential for viability (Dilks *et al.*, 2005; Thomas and Bolhuis, 2006) and there is evidence that Tat-dependent translocation is widely used as part of a mechanism for adaptation to extreme saline environments (Rose *et al.*, 2002).

Computational prediction of signal peptides was performed initially using weight matrices (von Heijne, 1986). However, Neural Networks (NNs) (Nielsen *et al.*, 1997, 1999) as well as Hidden Markov Models (HMMs; Nielsen and Krogh, 1998) introduced by the SignalP method have been proven to be the most successful methods currently available (Menne *et al.*, 2000). Recently, SignalP was retrained and, mainly due to better annotation and selection of the training set, yielded an even better accuracy (Bendtsen *et al.*, 2004). The Phobius (Kall *et al.*, 2004, 2007) and Philius (Reynolds *et al.*, 2008) methods followed a different approach in which a graphical model (HMM and Bayesian network, respectively) was used to predict at the same time the presence of a secretory signal peptide and transmembrane (TM) topology of a given protein minimizing thus the number of signal peptides predicted as TM segments and vice versa. Other methods such as LipoP (Juncker *et al.*, 2003) and PRED-LIPO (Bagos *et al.*, 2008) were developed during the last years for predicting lipoprotein signal peptides, which possess a distinctive cleavage site with an indispensable cysteine responsible for anchoring to the membrane (Sankaran and Wu, 1994;

*To whom correspondence should be addressed.

Sankaran *et al.*, 1995) and discriminate them from secretory signal peptides.

Although most of the methods mentioned so far are capable, up to a certain degree, of predicting the Tat signal peptides, only a few attempts were made toward predicting specifically this class of proteins. TATFIND was presented initially combining regular expression patterns and hydrophobicity analysis (Rose *et al.*, 2002), whereas few years later, TatP was presented using a combination of regular expression patterns and NNs (Bendtsen *et al.*, 2005). TatP has been shown to be more reliable, whereas TATFIND is not capable of predicting the cleavage site but only recognizes the existence of the n- and h-region. Another problem is that these methods are not trained to discriminate at the same time Tat from Sec signal peptides and thus, there is a need for combining these predictors with a generic signal peptide predictor such as SignalP. In this work, we present a novel method, PRED-TAT, based on HMMs, which is capable of predicting and discriminating Tat and Sec signal peptides. We show that this new method is more accurate than the previously developed methods and additionally, compares favorably to the top-scoring methods for the prediction of Sec signal peptides. The prediction method is available for non-commercial users at http://www.compgen.org/tools/PRED-TAT/ and we expect it to be useful to both experimentalists and bioinformaticians.

2 METHODS

2.1 Datasets

The dataset that we used for training contained 150 Tat signal peptides (119 from Gram-negative bacteria and 31 from Gram-positive ones), 328 secreted proteins containing a Sec signal peptide (216 from Gram-negative bacteria and 112 from Gram-positive ones), 288 cytoplasmic proteins (183 from Gram-negative bacteria and 105 from Gram-positive ones) and 140 sequences (segments of TM proteins) with a TM segment that have their N-terminus located to the cytoplasmic side of the membrane (90 from Gram-negative bacteria and 50 from Gram-positive ones). The Tat signal peptides were collected from the Uniprot database (Wu *et al.*, 2006). Since there were only few proteins with an experimentally verified Tat signal peptide (i.e. with the 'Tat-type signal peptide' identifier in the FT field), we decided to also include proteins with putative or potential Tat signal peptides. The initial dataset was submitted to redundancy reduction following the procedures used in SignalP papers (Nielsen *et al.*, 1997, 1999) using the algorithm of Hobohm (Hobohm *et al.*, 1992). We used a similarity threshold determined at 20 identical residues within the signal sequence after a BLAST alignment (Altschul *et al.*, 1997). After careful checking, we removed from the set some proteins that could potentially contain a lipoprotein Tat signal peptide. The Sec signal peptides, the cytoplasmic and the membrane proteins were collected as described previously in the development of PRED-LIPO (Bagos *et al.*, 2008) and CW-PRED (Litou *et al.*, 2008) with two major differences. First, since PRED-LIPO and CW-PRED were trained on Gram-positive bacteria, we repeated the same procedure in order to include additional proteins from Gram-negative bacteria. Secondly, we removed from the Sec signal peptide dataset proteins that were exported by the Tat pathway (as judged by the annotation in Uniprot). In short, Sec signal peptides were extracted from the training set of SignalPv2 (Nielsen *et al.*, 1997, 1999), cytoplasmic proteins from the set of Menne (Menne *et al.*, 2000) and TM proteins from various previously presented well-annotated datasets (Chen and Rost, 2002; Ikeda *et al*., 2003; Jayasinghe *et al*., 2001; Moller *et al*., 2000). All proteins were of bacterial origin (Gram positive or Gram negative) and we deliberately excluded archaeal sequences since the number of proteins with experimentally verified signal peptide cleavage site is very small (Bagos *et al.*, 2009).

In order to have an independent test set for evaluating the method and compare it against the other available ones, we performed an extensive search in the recent literature. We identified experimentally verified Tat signal peptides originating from Gram-negative, Gram-positive bacteria and Archaea. Proteins that were already present in the training set (i.e. those that already had an identifier indicating that were Tat substrates) were subsequently removed from the test set. By this way, for the majority of Tat signal peptides in the test set, the precise location of the cleavage site was not known. Some potential lipoproteins as well as several anchored proteins (Aldridge *et al.*, 2008; Bachmann *et al.*, 2006; Hatzixanthis *et al.*, 2003) were removed. Moreover, some wrongly annotated translation initiation sites that were discovered during the literature search, such as Q3L8N0 (Yikmis *et al.*, 2008), were reported to the Uniprot database. From UniProt, following the procedure of Menne and coworkers (Menne *et al.*, 2000) we also collected proteins having an experimentally verified Sec signal peptide from bacteria. We then removed proteins that were already present in the set of SignalPv2 (which we used for training). Proteins carrying a Sec as well as a Tat signal peptide were submitted once again to redundancy reduction, ensuring also that no similarity with proteins of the training set was present. Finally, in the test set of Tat signal peptides remained 75 sequences, including 18 sequences from Gram-negative bacteria, 45 from Gram-positive bacteria and 12 from Archaea (Supplementary Table 1). In the test set of Sec signal peptides remained 273 sequences (193 from Gram-negative bacteria and 80 from Gram-positive ones). We also retrieved bacterial 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 601 proteins (407 from Gram-negative bacteria and 194 from Gram-positive ones). Finally, in order to test our method on TM proteins, we used the experimentally verified cytoplasmic membrane proteins from bacteria used for the development of the PSORTB method (Gardy *et al.*, 2005). 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 192 TM segments (136 from Gram-negative bacteria and 56 from Gram-positive ones).

We also evaluated the HMMs, against the training and test sets used by the developers of TatP. Finally, by searching the literature we identified some additional proteins with signal peptides possessing the RR motif in the *n*-region, but which have been experimentally verified as not being Tat substrates (Palmer *et al.*, 2005; Widdick *et al.*, 2006, 2008). These proteins were used as an additional carefully selected negative test set for assessing the specificity of the methods developed here. The datasets compiled in this work are available at http://www.compgen.org/tools/PRED-TAT/supplement/datasets/.

2.2 HMMs

The HMM that we used has a modular architecture quite similar to the ones previously used by PRED-LIPO, CW-PRED and PRED-SIGNAL. It consists of four different sub-models, the Tat sub-model, corresponding to the signal peptides translocated by Tat, the Sec signal peptide sub-model corresponding to the Sec signal peptides, the N-terminal TM sub-model corresponding to the N-terminal TM domain and a globular sub-model used to model the globular N-terminal domains of cytoplasmic or membrane proteins. The novel feature of our method is the Tat sub-model (Fig. 1) that was especially designed to capture the sequence features of Tat signal peptides. It contains states modeling the N-terminal *n*-region, the RR region, the hydrophobic h-region and the cleavage site (c-region). In order to avoid overfitting, we used the same emission probabilities for the *n*-region and, for all but the last two residues of the h-region. The allowed transition probabilities were set in order to model as closely as possible the sequence features of the Tat substrates. The Sec signal peptide model is identical to the one used

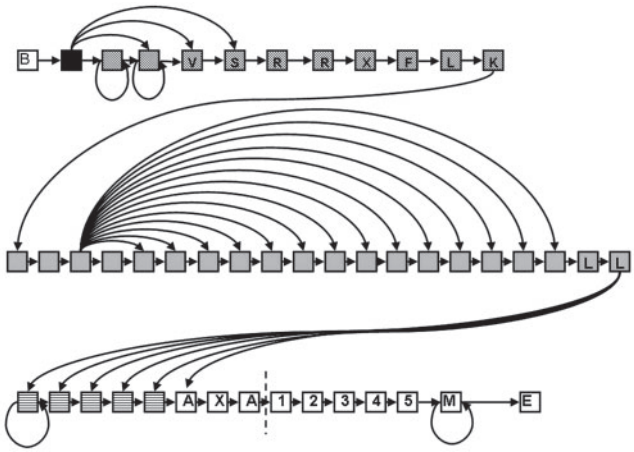

Fig. 1. The sub-model corresponding to the Tat signal peptides. States that share the same emission probabilities are depicted with the same shading and symbol (The letter denotes the dominant amino acid but only the R states within the Tat motif are invariant). The cleavage site is presented with a dashed vertical line between states A and 1. Allowed transitions are depicted with arrows.

before (Bagos *et al.*, 2008, 2009) and in many respects resembles the SignalP models. Moreover, the emission probabilities for the cleavage site were tied to the ones used for the cleavage site of Tat substrates. The TM sub-model, is identical to the one used by the HMM-TM predictor for alpha-helical membrane proteins (Bagos *et al.*, 2006), whereas the globular sub-model consists simply of a self-transitioning state. The total number of the model's states is 142 (including begin and end states) with 240 freely estimated transitions. On the other hand, the total number of freely estimated emission probabilities is 551 (29×19), yielding a total number of 791 freely estimated parameters.

The model was trained using the Baum–Welch algorithm for labeled sequences (Krogh, 1994) and the decoding was performed using the standard Viterbi algorithm (Durbin *et al.*, 1998). In addition to the Viterbi decoding that produces the optimal path of states through the model, and hence predicts simultaneously the type of the sequence as well as the cleavage sites (if any), we report also the S1 reliability index (Melen *et al.*, 2003) which takes values within the range of 0–1 and it is a measure of the reliability of the prediction, useful in a lot of situations.

The reported results correspond to a 30-fold cross-validation procedure, where each set contains an equally balanced number of Tat signal peptides, Sec signal peptides, TM and globular proteins. The training procedure consists of removing 1 of the 30 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 (Tat substrates versus non-Tat substrates, signal peptides versus non-signal peptides), we used the percentage of correctly classified positive examples (sensitivity), the percentage of correctly classified negative examples (specificity) and the Mathews Correlation coefficient (MCC) that summarizes in a single measure true postitives (TP), false positives (FP), true negatives (TN) and false negatives (FN) (Baldi *et al*., 2000).

We also created two profile Hidden Markov Models (pHMMs) using the HMMER 2.3.2 package (Eddy, 1998). The pHMM is a special case of HMM 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 classical 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, since it has been shown that under certain circumstances, it can also be used to model the sequence features of signal peptides (Zhang and Henzel, 2004; Zhang and Wood, 2003). Initially, we created the multiple alignments as advised previously for eukaryotic proteins (Zhang and Henzel, 2004; Zhang and Wood, 2003) with minor manual interventions to handle the longer *n*-regions of Tat signal peptides. We then built the pHMMs using the hmmbuild command of the HMMER package and we performed searches using the hmmpfam command of the same package. The multiple alignments and the HMMs are available at http://www.compgen.org/tools/PRED-TAT/.

2.3 Comparison to other prediction methods

For comparison, we used mainly the TatP (Bendtsen *et al*., 2005) and TATFIND (Rose *et al.*, 2002) methods, which are until now the only available predictors concerning Tat signal peptides. TatP is based on a combination of regular expression patterns and NNs and it is available at http://www.cbs.dtu.dk/services/TatP/. TATFIND, on the other hand, is based on a combination of regular expression patterns and hydrophobicity analysis (Rose *et al.*, 2002) and it is available online at http://signalfind.org/tatfind.html. We also evaluated two profile HMMs especially designed to recognize Tat signal peptides, which are deposited in public databases: the PF10518 profile of the PFAM database (Finn *et al.*, 2006) and the TIGR01409 profile of the TIGR database (Haft *et al.*, 2003). In all cases (as well as when running the methods developed in this work), we used the sequences truncated to their first 100 residues. The profile HMMs were run using the trusted cutoffs reported in the respective databases.

For analyses concerning the accuracy in predicting Sec signal peptides, we used SignalPv3 (Bendtsen *et al*., 2004), Phobius (Kall *et al*., 2004), Philius (Reynolds *et al.*, 2008), RPSP (Plewczynski *et al.*, 2008) and PrediSi (Hiller *et al*., 2004). Concerning SignalPv3, we used both the NN and the HMM modules. For methods that use different parameters for Gram-positive and Gram-negative bacteria (SignalP, PrediSi) we used the appropriate predictor for each test sequence. In all cases we used the default parameters, with the submitted sequences truncated to their first 70 residues.

3 RESULTS

The detailed results obtained from the HMM method (PRED-TAT) on the 30-fold cross-validation procedure are presented in the form of a confusion matrix in Supplementary Table 1A, whereas those concerning the independent test set in Supplementary Table 1B.

In the cross-validation, the model performs very well correctly classifying the 148 out of the 150 Tat signal peptides (98.67%) and predicts falsely only 9 Tat substrates (1.25%, all are actually Sec signal peptides), yielding this way an MCC equal to 0.96 (Table 1). PRED-TAT in the cross-validation clearly outperforms the other available methods (TatP, TATFIND and the profile HMMs), being equally specific but far more sensitive even though the results obtained for these methods are not cross-validated (i.e. some of the proteins tested were previously used for training these methods). TATFIND seems to perform better compared to TatP, a fact mainly caused by TatP's lower specificity in excluding Sec signal peptides. Interestingly, the profile HMM that we created using HMMER (PRED-TAT$_{\text{HMMER}}$) outperforms both the other profile HMMs as well as TATFIND and TatP, being almost equally good to the custom HMM (PRED-TAT). Concerning the Sec signal peptides, PRED-TAT also produces satisfactory results predicting correctly the presence of a signal peptide in 315 out of the 328 proteins (96.04%) and excludes non-signal peptides with a rate of 92.29%, giving an overall MCC of 0.88 (Table 3). These results compare favorably to those obtained by the other methods trained specifically

Table 1. Results obtained from the Tat predictors in the training set

Method	Tat SPs	Sec SPs	Cyto	TMs	MCC
PRED-TAT	148/150 (98.67%)	319/328 (97.26%)	288/288 (100.00%)	140/140 (100.00%)	0.96
PRED-TAT $_{HMMER}$	148/150 (98.67%)	312/328 (95.12%)	288/288 (100%)	139/140 (99.3%)	0.93
TATFIND	134/150 (89.33%)	326/328 (99.39%)	287/288 (99.65%)	140/140 (100.00%)	0.92
TatP	130/150 (86.67%)	284/328 (86.59%)	283/288 (98.26%)	133/140 (95.00%)	0.73
PF10518	15/150 (10.00%)	328/328 (100.00%)	288/288 (100.00%)	140/140 (100.00%)	0.29
TIGR01409	105/150 (70.00%)	327/328 (99.70%)	288/288 (100.00%)	140/140 (100.00%)	0.81

The results concerning PRED-TAT were obtained from the 30-fold cross-validation procedure as described in the Section 2. The MCC is computed by comparing Tat signal peptides versus non-Tat sequences (Sec signal peptides, cytoplasmic and TM sequences).

Table 2. Results obtained from the Tat predictors in the independent test set

Method	Tat SPs	Sec SPs	Cyto	TMs	MCC
PRED-TAT	71/75 (94.67%)	265/273 (97.07%)	598/601 (99.50%)	190/192 (98.96%)	0.89
PRED-TAT$_{HMMER}$	72/75 (96.00%)	259/273 (94.87%)	601/601 (100.00%)	190/192 (98.96%)	0.88
TATFIND	60/75 (80.00%)	270/273 (98.90%)	599/601 (99.67%)	192/192 (100.00%)	0.85
TatP	62/75 (82.67%)	231/273 (84.62%)	594/601 (98.84%)	177/192 (92.19%)	0.61
PF10518	9/75 (12.00%)	273/273 (100.00%)	601/601 (100.00%)	192/192 (100.00%)	0.34
TIGR01409	47/75 (62.67%)	272/273 (99.63%)	601/601 (100.00%)	192/192 (100.00%)	0.77

The test set contains no similar sequences to those included in the training set (see Section 2). The MCC is computed by comparing Tat signal peptides versus non-Tat sequences (Sec signal peptides, cytoplasmic and TM sequences).

for the prediction of Sec signal peptides. In particular, PRED-TAT slightly outperforms RPSP and PrediSi, whereas is surpassed by SignalPv3, Philius and Phobius. Nevertheless, as it is apparent from Table 2, all these predictors perform comparably yielding MCCs in the range of 0.87–0.93. Moreover, we should emphasize that the results of the other predictors are not cross-validated.

On the independent test set (Table 3), PRED-TAT continues to perform better than TatP and TATFIND, classifying correctly 71 out of the 75 Tat sequences (with a sensitivity equal to 94.67%). Concerning the specificity of the method in detecting Tat signal peptides, PRED-TAT produces approximately equal number of false positive findings (compared with TatP and TATFIND) among the secreted, TM and cytoplasmic proteins yielding thus an MCC that is equal to 0.89. PRED-TAT$_{HMMER}$ once again performs comparable to PRED-TAT. TATFIND performs better compared with TatP, since the latter fails to discriminate a larger number of Sec signal peptides. Similar results are obtained when we test all the methods in the sets used for training TatP (Supplementary Tables). As expected, the overall accuracy of TatP is increased

Table 3. Results obtained from the Sec predictors in the training set

Method	Sec SPs	Cyto	TMs	MCC
PRED-TAT	315/328 (96.04%)	265/288 (92.01%)	130/140 (92.86%)	0.88
PRED-TAT$_{HMMER}$	285/328 (86.89%)	285/288 (98.96%)	130/140 (92.86%)	0.88
RPSP	303/328 (92.38%)	287/288 (99.65%)	116/140 (82.86%)	0.87
PrediSi	317/328 (96.65%)	280/288 (97.22%)	108/140 (77.14%)	0.87
SignalPv3 (NN)	323/328 (98.48%)	280/288 (97.22%)	117/140 (83.57%)	0.91
SignalPv3 (HMM)	325/328 (99.09%)	283/288 (98.26%)	114/140 (81.43%)	0.91
Phobius	318/328 (96.95%)	281/288 (97.57%)	129/140 (92.14%)	0.93
Philius	318/328 (96.95%)	274/288 (95.14%)	132/140 (94.29%)	0.91

The results concerning PRED-TAT were obtained from the 30-fold cross-validation procedure as described in Section 2. The training set includes a significant portion of sequences used to train the remaining predictors. The set does not include Tat signal peptides and in such case, the specificity of the Sec signal peptide predictors would be lower. The MCC is computed by comparing Sec signal peptides versus non-Sec sequences (cytoplasmic and TM sequences).

(MCC = 0.89) in this set, but PRED-TAT is still more accurate yielding an MCC of 0.94 (PRED-TAT$_{HMMER}$ performs slightly better). Lastly, the methods for predicting Tat signal peptides were evaluated against the dataset of proteins carrying an RR motif but which have been experimentally verified not to be Tat substrates. In this set, the profile HMMs and TATFIND are found to be more specific, but nevertheless, PRED-TAT correctly excludes 86.64% of the sequences and PRED-TAT$_{HMMER}$ 84.1%, whereas TatP excludes only 50% (Supplementary Table 6). Two similar datasets, even though not experimentally verified, were compiled by the authors of TatP and on these, PRED-TAT excludes 1291 out of the 1336 proteins (96.63%), PRED-TAT$_{HMMER}$ excludes 1295 (96.93%), whereas TATFIND excludes 1299 (97.23%) and TatP 1289 (96.48%).

Concerning the classification of proteins bearing a Sec signal peptide, in the independent test set (Table 4), the method correctly classifies 252 out of the 273 proteins (92.31%). The specificity of the method in detecting secretory signal peptides is very satisfactory since it wrongly assigns a signal peptide in 31 out of the 601 cytoplasmic proteins and in 25 out of the 192 TM proteins. These results correspond to a specificity of 92.94%, with an MCC of 0.82. PRED-TAT$_{HMMER}$ is more specific but less sensitive; nevertheless, achieves the second higher MCC among the methods tested. We should note that the overall MCCs reported here concerning SignalP correspond clearly to lower values compared with those reported in the original publications (Bendtsen *et al.*, 2004; Nielsen *et al.*, 1997) where the authors reported larger MCCs for bacterial predictors mainly caused by SignalP's lower specificity when tested on TM proteins. Since we have reasons to believe that many of the proteins in the test set were also present in the datasets used for training the other methods overestimating thus the results, we compiled another test set of proteins with potential signal peptides deposited in Uniprot from 2009 onwards. After removing sequences similar

Table 4. Results obtained from the Sec predictors in the independent test set

Method	Sec SPs	Cyto	TMs	MCC
PRED-TAT	252/273 (92.31%)	570/601 (94.84%)	167/192 (86.98%)	0.82
PRED-TAT$_{HMMER}$	238/273 (87.18%)	597/601 (99.33%)	174/192 (90.62%)	0.86
RPSP	249/273 (91.21%)	601/601 (100.00%)	146/192 (76.04%)	0.83
PrediSi	260/273 (95.24%)	579/601 (96.34%)	114/192 (59.38%)	0.76
SignalPv3 (NN)	252/273 (92.31%)	599/601 (99.67%)	150/192 (78.12%)	0.85
SignalPv3 (HMM)	264/273 (96.70%)	593/601 (98.67%)	134/192 (69.79%)	0.83
Phobius	249/273 (91.21%)	594/601 (98.84%)	154/192 (80.21%)	0.84
Philius	253/273 (92.67%)	582/601 (96.84%)	181/192 (94.27%)	0.88

The test set contains no similar sequences to those included in the training set (see Section 2). The set does not include Tat signal peptides and in such case, the specificity of the Sec signal peptide predictors would be lower. The MCC is computed by comparing Sec signal peptides versus non-Sec sequences (cytoplasmic and TM sequences).

to our training set we came up with 95 bacterial proteins. In this set, PRED-TAT and SignalPv3-HMM correctly classify 92.63%, Phobius 93.68% and Philius 91.58%. All the other methods perform worse (70–90%, data not shown).

Concerning the precise location of the cleavage site of Tat signal peptides, PRED-TAT outperforms TatP in the cross-validation on the training set (70% versus 64.67%) and performs equally well when tested on the TatP training set (both methods yield 75.24%). When we consider as correct a prediction within ±2 residues of the actual cleavage site, we obtain similar results (i.e. 78.67% versus 72.67% in the former set and 81.90% versus 80.95% in the latter). As we already mentioned, TATFIND and the profile HMMs of PFAM and TIGR are not capable of predicting the location of the cleavage site, and thus cannot be tested. Moreover, in the independent test set the precise location of the cleavage sites was not known. PRED-TAT$_{HMMER}$ performs similarly in both sets. Concerning the accuracy in predicting the cleavage sites of Sec signal peptides, PRED-TAT predicts correctly the location of the cleavage site in 78.66% of the proteins in the training set (89.94% when counting predictions within ±2 residues of the actual cleavage site) and 78.02% of the proteins in the independent test set (83.88% when counting predictions within ±2 residues of the actual cleavage site). These results are approximately equally good compared with RPSP, PrediSi, Phobius and Philius and slightly worse compared with both versions of SignalPv3 (data not shown). PRED-TAT$_{HMMER}$ performs slightly worse in this dataset.

The most important feature of the method is the fact that it is capable of simultaneously predicting and discriminating Sec from Tat signal peptides. If for example, one performed the analysis concerning Sec signal peptides and included in the dataset a number of Tat signal sequences, the majority of general signal peptide predictors would have predicted a significant number of them to be Sec signal peptides. If on the other hand, one tries to combine two different predictors (i.e. TatP and SignalP), then he/she would have to devise an algorithm for prioritization. For instance, we could give

Fig. 2. Sequence logo (Schneider and Stephens, 1990) of the RR region of the Tat signal peptides included in the training set (upper) and in the test set (lower). Even though the sets were submitted to redundancy reduction, the conserved regions are quite similar, as one would expect. The logos were created using WebLogo (Crooks *et al.*, 2004).

priority to TatP, or alternatively, the user may choose the predictor with the highest D-score. This strategy was also applied for the profile HMMs (PRED-TAT$_{HMMER}$) that we developed, which also can be considered as a combination of two independent predictors and the results are presented in Supplementary Tables 2–4, where the superiority of the combined predictor is highlighted.

4 CONCLUSIONS

In this work, we proposed an efficient method for the combined prediction of Tat and Sec signal peptides. The HMM method (PRED-TAT) that we developed was tested on various carefully selected datasets and in all cases was shown to outperform TatP and TATFIND in almost all measures of accuracy (sensitivity, specificity and MCC, as well as the correct prediction of the cleavage site). TATFIND seems to be close in overall accuracy, but its main disadvantage is that it cannot predict the location of the cleavage site resulting this way in a reduced functionality. The profile HMMs from PFAM and TIGR seem to be very specific, but suffer in terms of sensitivity, at least when the suggested trusted cutoff is used; additionally, they also are incapable of predicting the cleavage site. The profile HMMs generated by us (PRED-TAT$_{HMMER}$) seem to be a very easily applicable and useful alternative. PRED-TAT was also proved to be very accurate in the correct prediction of Sec signal peptides, comparing favorably even to the top-scoring predictors. Nevertheless, we have clearly shown that a combined predictor such as PRED-TAT performs better predicting simultaneously Tat and Sec signal peptides.

The HMM, as a machine learning method, is capable of detecting the general preferences of Tat signal peptides that discriminate them from Sec signal peptides, such as the longest *n*-region, the RR motif and the less hydrophobic h-region. However, variations do exist between organisms due to differences in the Tat machinery itself. For instance, in some rare cases, one of the two arginines in the RR motif can be substituted by lysine without affecting the Tat-dependent export (Hinsley *et al.*, 2001), whereas there are also variants with an intervening Asparagine (RNR motif) (Ignatova *et al.*, 2002). These rare variants are not tolerated by PRED-TAT, which was trained using only 'canonical' Tat sequences (Fig. 2). TatP allows the

user to change the regular expression pattern and thus to perform predictions by relaxing the RR motif assumption. This approach, however, has limited usefulness in large-scale analyses since it is not statistically validated (i.e. the number of false positives will increase drastically). Mutagenesis studies in various model organisms have also shown different properties concerning the adjacency to the RR amino acids (DeLisa *et al*., 2002; Kreutzenbeck *et al*., 2007; Kwan and Bolhuis, 2010; Li *et al*., 2006). Moreover, the situation is more complicated, since in many cases some signal peptides possess intermediate properties. For instance, in some cases, signal peptides are rerouted to the Sec machinery only after Tat deactivation and in others, the signal sequences possess promiscuity under physiological conditions (Tullman-Ercek *et al*., 2007). Thus, it would be non-realistic to expect having a single predictor that could be predicting all these real-life situations. Nevertheless, in a dataset of RR-containing signal peptides that were experimentally verified not to be TAT substrates, PRED-TAT performs satisfactory, a fact which indicates that indeed the algorithm extracted the general preferences of Tat-substrates succesfully.

The method presented here was trained on a combined set of Gram-positive and Gram-negative bacterial sequences. Even though subtle differences exist between the signal peptides' cleavage sites between these classes, the limited size of the training set dictated this choice and the results are satisfactory. Perhaps in the near future, where larger datasets may become available, a different predictor may be feasible to be constructed. The test set also contained several archaeal Tat signal sequences and these were predicted correctly as well. A limitation of PRED-TAT (as well as the previously proposed methods) is that it cannot discriminate Tat lipoproteins. Lipoproteins exported using the Tat machinery have been observed lately, but there are a limited number of well-characterized examples (Gimenez *et al.*, 2007; Shruthi *et al.*, 2010a). Thus, only approaches based on regular expression patterns could be developed for the prediction of such proteins (Shruthi *et al.*, 2010b). However, in the future, where more examples will become available, the flexible architecture of the HMM allows easily the extension to such cases. Another possible extension would be to discriminate between signal anchored proteins exported with the Tat pathway. Such proteins were excluded from the training and test sets, but a future extension of the method could incorporate them in the final prediction.

PRED-TAT is publicly available for non-commercial users at http://www.compgen.org/tools/PRED-TAT/ where the user may submit a single sequence in order to receive a detailed prediction or upload a file in order to perform batch predictions (i.e. in a whole genome). The profile HMMs (PRED-TAT$_{HMMER}$) are also freely available for download as an easy to use alternative for large-scale analyses.

ACKNOWLEDGEMENTS

The authors thank the reviewers for the valuable comments.

Conflict of Interest: none declared.

REFERENCES

Aldridge,C. *et al.* (2008) Tat-dependent targeting of Rieske iron-sulphur proteins to both the plasma and thylakoid membranes in the cyanobacterium Synechocystis PCC6803. *Mol. Microbiol.*, **70**, 140–150.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.

Bachmann,J. *et al.* (2006) The Rieske protein from Paracoccus denitrificans is inserted into the cytoplasmic membrane by the twin-arginine translocase. *FEBS J.*, **273**, 4817–4830.

Bagos,P.G. *et al.* (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics*, **7**, 189.

Bagos,P.G. *et al.* (2008) Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model, *J. Proteome Res.*, **7**, 5082–5093.

Bagos,P.G. *et al.* (2009) Prediction of signal peptides in archaea, *Protein Eng. Des. Sel.*, **22**, 27–35.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Bendtsen,J.D. *et al.* (2005) Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, **6**, 167.

Berks,B.C. *et al.* (2005) Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr. Opin. Microbiol.*, **8**, 174–181.

Chen,C.P. and Rost,B. (2002) Long membrane helices and short loops predicted less accurately. *Protein Sci.*, **11**, 2766–2773.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

DeLisa,M.P. *et al.* (2002) Genetic analysis of the twin arginine translocator secretion pathway in bacteria. *J. Biol. Chem.*, **277**, 29825–29831.

Dilks,K. *et al.* (2005) Genetic and biochemical analysis of the twin-arginine translocation pathway in halophilic archaea. *J. Bacteriol*, **187**, 8104–8113.

Driessen,A.J. and Nouwen,N. (2008) Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.*, **77**, 643–667.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis, Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Gardy,J.L. *et al.* (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.

Gimenez,M.I. *et al.* (2007) Haloferax volcanii twin-arginine translocation substates include secreted soluble, C-terminally anchored and lipoproteins. *Mol. Microbiol.*, **66**, 1597–1606.

Habib,S.J. *et al.* (2007) Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol.*, **80**, 761–781.

Haft,D.H. *et al.* (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.

Hatzixanthis,K. *et al.* (2003) A subset of bacterial inner membrane proteins integrated by the twin-arginine translocase. *Mol. Microbiol.*, **49**, 1377–1390.

Hiller,K. *et al.* (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.*, **32**, W375–W379.

Hinsley,A.P. *et al.* (2001) A naturally occurring bacterial Tat signal peptide lacking one of the 'invariant' arginine residues of the consensus targeting motif. *FEBS Lett.*, **497**, 45–49.

Hobohm,U. *et al.* (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.

Ignatova,Z. *et al.* (2002) Unusual signal peptide directs penicillin amidase from Escherichia coli to the Tat translocation machinery. *Biochem. Biophys. Res. Commun.*, **291**, 146–149.

Ikeda,M. *et al.* (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.*, **31**, 406–409.

Jayasinghe,S. *et al.* (2001) MPtopo: a database of membrane protein topology. *Protein Sci.*, **10**, 455–458.

Juncker,A.S. *et al.* (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.*, **12**, 1652–1662.

Kall,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Kall,L. *et al.* (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.

Kreutzenbeck,P. *et al.* (2007) Escherichia coli twin arginine (Tat) mutant translocases possessing relaxed signal peptide recognition specificities. *J. Biol. Chem.*, **282**, 7903–7911.

Krogh,A. (1994) Hidden Markov models for labelled sequences. In *Proceedings of the12th IAPR International Conference on Pattern Recognition (Jerusalem)*, IEEE, pp. 140–144.

Kwan,D. and Bolhuis,A. (2010) Analysis of the twin-arginine motif of a haloarchaeal Tat substrate. *FEMS Microbiol. Lett.*, **308**, 138–143.

Lee,P.A. *et al.* (2006) The bacterial twin-arginine translocation pathway. *Annu. Rev. Microbiol.*, **60**, 373–395.

Li,H. *et al.* (2006) Impact of amino acid changes in the signal peptide on the secretion of the Tat-dependent xylanase C from Streptomyces lividans. *FEMS Microbiol. Lett.*, **255**, 268–274.

Litou,Z.I. *et al.* (2008) 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.

Melen,K. *et al.* (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.

Menne,K.M. *et al.* (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.

Moller,S. *et al.* (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Nielsen,H. and Krogh,A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 122–130.

Nielsen,H. *et al.* (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Nielsen,H. *et al.* (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.

Palmer,T. *et al.* (2005) Export of complex cofactor-containing proteins by the bacterial Tat pathway. *Trends Microbiol.*, **13**, 175–180.

Plewczynski,D. *et al.* (2008) Prediction of signal peptides in protein sequences by neural networks. *Acta Biochim. Pol.*, **55**, 261–267.

Pohlschroder,M. *et al.* (2005) Protein transport in Archaea: Sec and twin arginine translocation pathways. *Curr. Opin. Microbiol.*, **8**, 713–719.

Rapoport,T.A. *et al.* (1999) Posttranslational protein translocation across the membrane of the endoplasmic reticulum. *Biol. Chem.*, **380**, 1143–1150.

Reynolds,S.M. *et al.* (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.*, **4**, e1000213.

Rose,R.W. *et al.* (2002) Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Mol. Microbiol.*, **45**, 943–950.

Sankaran,K. and Wu,H.C. (1994) Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol. *J. Biol. Chem.*, **269**, 19701–19706.

Sankaran,K. *et al.* (1995) Modification of bacterial lipoproteins. *Methods Enzymol.*, **250**, 683–697.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Shruthi,H. *et al.* (2010a) Twin arginine translocase pathway and fast-folding lipoprotein biosynthesis in E. coli: interesting implications and applications. *Mol. Biosyst.*, **6**, 999–1007.

Shruthi,H. *et al.* (2010b) TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes. *J. Mol. Evol.*, **70**, 359–370.

Teter,S.A. and Klionsky,D.J. (1999) How to get a folded protein across a membrane. *Trends Cell Biol.*, **9**, 428–431.

Thomas,J.R. and Bolhuis,A. (2006) The tatC gene cluster is essential for viability in halophilic archaea. *FEMS Microbiol. Lett.*, **256**, 44–49.

Tullman-Ercek,D. *et al.* (2007) Export pathway selectivity of Escherichia coli twin arginine translocation signal peptides. *J. Biol. Chem.*, **282**, 8309–8316.

Tuteja,R. (2005) Type I signal peptidase: an overview. *Arch. Biochem. Biophys.*, **441**, 107–111.

von Heijne,G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.*, **14**, 4683–4690.

von Heijne,G. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.

von Heijne,G. *et al.* (1989) Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.*, **180**, 535–545.

van Roosmalen,M.L. *et al.* (2004) Type I signal peptidases of Gram-positive bacteria. *Biochim. Biophys. Acta*, **1694**, 279–297.

Widdick,D.A. *et al.* (2006) The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. *Proc. Natl Acad. Sci. USA*, **103**, 17927–17932.

Widdick,D.A. *et al.* (2008) A facile reporter system for the experimental identification of twin-arginine translocation (Tat) signal peptides from all kingdoms of life. *J. Mol. Biol.*, **375**, 595–603.

Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.

Yikmis,M. *et al.* (2008) Secretion and transcriptional regulation of the latex-clearing protein, Lcp, by the rubber-degrading bacterium Streptomyces sp. strain K30. *Appl. Environ. Microbiol.*, **74**, 5373–5382.

Zhang,Z. and Henzel,W.J. (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci.*, **13**, 2819–2824.

Zhang,Z. and Wood,W.I. (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, **19**, 307–308.