Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

TATPred: a Bayesian method for the identification of twin arginine translocation pathway signal sequences.

Paul D. Taylor[1], Christopher P. Toseland[1], Teresa K. Attwood[2] and Darren R. Flower[1*]

[1*]The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20 7NN, UK; [2]Faculty of Life Sciences & School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PT, UK;

Darren R. Flower* - E-mail: darren.flower@jenner.ac.uk; Phone: +44 1635 577954; Fax: +44 1635 577908;

* Corresponding author

received July 12, 2006; revised July 24, 2006; accepted July 24, 2006; published online July 25, 2006

Abstract:

The twin arginine translocation (TAT) system ferries folded proteins across the bacterial membrane. Proteins are directed into this system by the TAT signal peptide present at the amino terminus of the precursor protein, which contains the twin arginine residues that give the system its name. There are currently only two computational methods for the prediction of TAT translocated proteins from sequence. Both methods have limitations that make the creation of a new algorithm for TAT-translocated protein prediction desirable. We have developed TATPred, a new sequence-model method, based on a Naïve-Bayesian network, for the prediction of TAT signal peptides. In this approach, a comprehensive range of models was tested to identify the most reliable and robust predictor. The best model comprised 12 residues: three residues prior to the twin arginines and the seven residues that follow them. We found a prediction sensitivity of 0.979 and a specificity of 0.942.

Keywords: Twin arginine motif; Bayesian Network; TAT translocation; Signal sequence; Vaccine.

Background:

The bacterial Sec-independent protein export pathway is also known as the TAT system: since precursor proteins are targeted to it by a signal peptide containing a characteristic sequence motif comprising two consecutive arginine residues. The most remarkable feature of the TAT pathway is the apparent transportation of fully folded proteins across the cytoplasmic membrane without rendering the membrane permeable to ions. **[1]** Most substrates secreted via the TAT pathway are proteins that bind cofactors within the cytoplasm, and must therefore be folded before export. Such proteins function predominantly in respiratory and photosynthetic electron transfer, and are a critical component of bacterial energy metabolism. Proteins without cofactors can also be transported via the TAT pathway. **[2, 3]**

TAT pathway signal peptides have a tripartite structure comprising a charged amino terminus (N-region), a hydrophobic central core (H-region), and a cleavage site containing the carboxy terminus (C-region). There are several distinct differences between TAT and Sec-dependent signal sequences, the most notable of which is the presence of a characteristic sequence motif, (S/T)-R-R-x-F-L-K, at the N/H-region boundary. The twin arginine residues of the motif are probably invariant, while the other residues occur at a frequency of greater than 50%. **[2]** The initial residue of the motif may act as an N-cap for α-helix formation by the H-region, since a potentially helix-breaking praline is often present at the end of the H-region. **[4]** The C-region is characterised by a high proportion of basic amino acids **[3]**, while the Sec-dependent pathway is biased against positively charged residues near to the signal peptidase cleavage site. **[5]** There is also a significant size difference: an extended N-region gives TAT signal peptides an average length that is 14 amino acids longer than that of Sec signal peptides. In addition, a greater frequency of glycine and threonine, and a lower abundance of leucine residues, is observed in the H-region, giving rise to a significant decrease in its hydrophobicity. **[4]**

Secreted or surface-expressed proteins are open to direct surveillance by the immune system, and are thus of potential interest as vaccine candidates. Identification of such proteins is a vital component in the rational discovery of vaccines. In order to predict, *in silico*, the subcellular location of bacterial proteins, one must consider the export of proteins by all possible pathways, including the TAT secretion system. Currently, the only available Web-server for the prediction of TAT signal sequences is TATP. **[6]** The method is not generic and requires specific tailoring for each genome. A simpler method for identifying TAT signal sequences is the well known TAT motif **[1, 2]**: this consists of the conserved residues described above, followed by a hydrophobic stretch of residues. Such methods have limitations as they allow only a small variation within the TAT signal sequence. In response to this limitation, we have developed a rapid, yet accurate alternative: TATPred. TATPred uses a Bayesian methodology to characterise and predict TAT signal peptides.

Methodology:

To train the method, we used data-sets of TAT translocated proteins and proteins not translocated by TAT yet which still contain a twin arginine motif. Based on annotations, proteins with an N-terminal RR motif were extracted from Swiss-Prot release 42: 117 TAT - translocated (positive set) and 1178 non-TAT translocated (negative set). Only proteins which were designated as "TAT translocated" were accepted for the positive set. Non-TAT translocated proteins were of

© 2006 Biomedical Informatics Publishing Group

cytoplasmic or membraneous subcellular location or were translocated via different export pathways. In all cases, only confirmed annotations were accepted. All entries with relevant annotations containing "hypothetical" or "putative" or "possible" or "putative" were excluded.

It is common practice to use non-redundant test and training data-sets by removing similar sequences. However, one could not remove redundancy within the TAT translocated set as it was too small. Instead, a separate set of proteins, not used in testing, was used to assess the method. Of the 117 putative TAT translocated proteins, 12 had experimental evidence confirming that they were transported via the TAT system. To form a test-set of confirmed TAT-translocated proteins, these 12 sequences were removed from the training set. To increase the size of the test-set, 23 other TAT translocated proteins from an annotated *E. coli* set (www.jic.bbsrc.ac.uk/staff/tracy-plamer/signals.htm) were added. This test-set was also used to assess the predictive ability of TatP. The non-redundant non-TAT translocated protein set comprised 714 proteins. The second negative test-set, which was only used in testing, comprised Sec-dependent/Type II signal peptides. This was taken from the SignalP v2 data-set **[7]**, and contained 250 bacterial proteins. This set is referred to as the signal set hereafter.

A sequence model was developed which used the invariant twin arginines as a reference point. Models were selected with lengths of 3 to 22 residues. A comprehensive range of models was tested, with 1 to 10 residues on either side of the twin arginines. Each model was used to train a Naïve-Bayesian Network (BN) with a structure such that one input node was used for each residue, while the one output node could take the values of TAT or non-TAT. This approach required the creation of 110 BNs. To train the BNs on the negative data-set, a random residue was selected from within the first 50 residues of each protein in the negative training set.

On completion of network training, query sequences were processed as follows: the first 50 residues of the query sequence were scanned to identify the presence of consecutive arginines. When two such arginines were found, appropriate sequences for the specific network were extracted. The residues were entered into the input nodes of the BN, with the first residue of the extracted sub-sequence being entered into node one, the second into node two, *etc*. The BN assesses whether or not the query sub-sequence is an instance of a TAT signal sequence and also returns an associated probability (threshold for a positive score is 80%). Testing was conducted on both the negative and positive data-sets using five-fold cross-validation; overall accuracy was obtained by averaging the five values. The signal set was then tested using the same methodology.

Start Position	End Position	Model Length	Sensitivity	Specificity
-10	10	22	0.809	0.968
-9	9	20	0.830	0.960
-8	8	18	0.872	0.964
-7	7	16	0.894	0.953
-6	6	14	0.915	0.931
-5	5	12	0.957	0.939
-4	4	10	0.957	0.939
-3	3	8	0.957	0.939
-2	2	6	1.000	0.903
-1	1	4	1.000	0.885
0	0	2	1.000	0.833
-1	3	5	**0.943**	**0.942**

Table 1: Representative sample of results for the 110 networks used to develop TATPred. The best performing network is shown in bold

Method	Sensitivity	Specificity	Signal Sequence Accuracy (%)
TATP	0.914	0.981	96.4
TAT Motif	0.743	1.000	100.0
TATPred	0.943	0.942	100.0

Table 2: Representative sample of results for the 110 networks used to develop TATPred when tested using the signal set

Results and Discussion:

Performance of different sequence lengths

When tested using the test-set and the non-TAT translocated set, the 110 sequence models investigated

© 2006 Biomedical Informatics Publishing Group

produced significantly different results, of which an informative sample is shown in Table 1. The best performing network, which comprises one residue before the twin arginines and the two residues after them, has a high degree of sensitivity and specificity (0.943 and 0.942 respectively). Only two from the test set were not predicted as TAT-translocated: YEDY_ECOLI and YCBK_ECOLI.

The best performing model considers only 5 residues. While remarkably small, it reaffirms previous studies indicating residue conservation close to the twin arginines. Differences in prediction accuracy reflect the ability of Bayesian methods to account for sequence variability. The BNs provide a probabilistic classification of which residues at which positions are likely to be consistent with a TAT signal sequence.

There is a correlation between the number of residues in a sequence model prior to the twin arginine motif and the sensitivity of prediction: the greater the number, the lower the concomitant sensitivity. This confirms previous studies showing residues prior to the twin arginine motif (with the exception of those immediately before) are not conserved. Sequence models with a high proportion of residues after the twin arginine motif show a higher specificity of prediction, though this is not as marked as the low specificity described above. This trend is clearly apparent in Table 1. The -10 +2 model, for example, has a prediction sensitivity of 0.254, while the -2 +10 mode has a sensitivity of 0.867.

Performance of different sequence lengths in the signal test-set

Assessment of the ability of the 110 sequence models to discriminate between TAT-translocated proteins and Sec signal sequences were performed using the Signal dataset. The -1 +3 sequence model, which performed best previously (see Table 1), also had the highest sensitivity of prediction; it correctly identified all the dataset as non-TAT translocated. Networks with significant proportions of their corresponding sequence model located after the twin arginine motif could not distinguish between TAT-translocated proteins and those possessing a Sec signal sequence. Higher specificity may be due to the sequence models also considering the hydrophobic region in the TAT signal sequence. Prediction accuracy for the signal set for these networks tends to be lower, as these networks are not able to distinguish the hydrophobic region on the TAT-signal peptide from the hydrophobic region of a Sec signal peptide.

Comparison of performance with other TAT-signal sequence prediction methods

Compared to other methods for the detection of TAT-translocated proteins, increased accuracy is seen over both simple motif methods and the TATP algorithm. The results of the comparison are shown in Table 2. As Table 2 shows, TATPred achieves a higher sensitivity of prediction than the two publicly available methods. Specificity of prediction is marginally lower than the other two methods. Strikingly, TATPred correctly identifies all the Signal dataset as being non-TAT translocated, despite 36 of the proteins possessing a twin arginine motif.

Differences in accuracy between the three prediction methods can be explained by differences in methodology. As with all regular-expression-based methods, the motifs cannot account for variation within TAT signal sequences, and thus only distinguish the most typical TAT-translocated proteins. TATP is a more sophisticated and accurate predictor, yet it is partially dependent on a motif to determine TAT signal peptides. The rigidity of a sequence motif may explain why it correctly identifies less of the test set. By contrast, TATPred exhibits high sensitivity for a range of organisms. Its Bayesian-based sequence model methodology offers a more flexible probabilistic approach that allows it to identify more variant TAT signal peptide sequences.

The specificity of TATPred is marginally lower than the other methods. The flexibility of TAT signal-sequence recognition, which confers high sensitivity, causes a slightly higher rate of false-positive predictions. A high sensitivity ensures that more candidate proteins are identified; false positives can then be screened out using other prediction methods. This is of particular importance in the area of target discovery, where it is vitally important to identify all possible targets. Algorithms, such as TATPred, could thus form a valuable part of multi-algorithmic approaches to the accurate prediction of subcellular location.

Conclusion:

TATPred deals with large, genomic-scale data-sets with high sensitivity and specificity of prediction. The method allows rapid analysis and identification of TAT signal sequences. Only proteins that are exposed on the surface of the pathogen will be accessible to surveillance by the immune system and hence produce a protective response. The computational identification of subcellular location can thus be used to screen pathogen genomes for surface-exposed proteins and greatly reduce the number of candidates that require *in vitro* testing. **[8]** An *in silico* analysis of the pathogen genome may mean that more vaccine candidates could be identified, allowing the development of vaccines for pathogens that have previously proved difficult to develop using conventional vaccinology techniques. Screening of pathogen genomes to identify such proteins can increase the efficiency of vaccine or drug discovery, reducing the development time and cost of new therapies. There are many freely available algorithms that facilitate reverse vaccinology **[9, 10]**; TATPred is a potentially useful new approach which can complement such approaches.

Acknowledgement:

PDT wishes to thank the MRC for a priority area studentship. We should like to thank Andrew Worth for his invaluable technical assistance. The Jenner Institute

© 2006 Biomedical Informatics Publishing Group

(Formally, The Edward Jenner Institute for Vaccine Research) wishes to thank its sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the UK Department of Health.

References:

[01] B. C Berks, *et al.*, *Mol. Microbiol.*, 35:260 (2000) [PMID: 10652088]

[02] B. C. Berks, *Mol. Microbiol.*, 22:393 (1996) [PMID: 8939424]

[03] T. Bruser, *et al.*, *FEMS Microbiol. Lett.*, 164:329 (1998) [PMID: 9682482]

[04] S. Cristobal, *et al.*, *Embo J.*, 18:2982 (1999) [PMID: 10357811]

[05] G. von Heijne, *J. Mol. Biol.*, 192:287 (1986) [PMID: 3560218]

[06] J. D. Bendtsen, *et al.*, *BMC Bioinformatics*, 6: 167 (2005) [PMID: 15992409]

[07] H. Nielsen, *et al.*, *Int. J. Neural. Syst.*, 8:581 (1997) [PMID: 10065837]

[08] M. Pizza, *et al.*, *Science*, 287:1816 (2000) [PMID: 10710308]

[09] A. S. Juncker, *et al.*, *Protein Sci.*, 12:1652 (2003) [PMID: 12876315]

[10] P. D. Taylor, *et al.*, *Nucleic Acids Res.*, 31:3698 (2003) [PMID: 12824397]

Edited by P. Kangueane

Citation: Taylor *et al.,* Bioinformation 1(5): 184-187 (2006)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

© 2006 Biomedical Informatics Publishing Group