Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Server**

LIPPRED: A web server for accurate prediction of lipoprotein signal sequences and cleavage sites

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* - Email: darren.flower@jenner.ac.uk; Phone: +44 1635 577954, Fax: +44 1635 577908;

* Corresponding author

received July 11, 2006; revised July 18, 2006; accepted July 18, 2006; published online July 19, 2006

Abstract:

Bacterial lipoproteins have many important functions and represent a class of possible vaccine candidates. The prediction of lipoproteins from sequence is thus an important task for computational vaccinology. Naïve-Bayesian networks were trained to identify SpaseII cleavage sites and their preceding signal sequences using a set of 199 distinct lipoprotein sequences. A comprehensive range of sequence models was used to identify the best model for lipoprotein signal sequences. The best performing sequence model was found to be 10-residues in length, including the conserved cysteine lipid attachment site and the nine residues prior to it. The sensitivity of prediction for LipPred was 0.979, while the specificity was 0.742. Here, we describe LipPred, a web server for lipoprotein prediction; available at the URL: http://www.jenner.ac.uk/LipPred. LipPred is the most accurate method available for the detection of SpaseII-cleaved lipoprotein signal sequences and the prediction of their cleavage sites.

Keywords: lipoprotein signal sequences; Naïve-Bayesian networks; reverse vaccinology; prediction; server

Background:

Bacterial lipoproteins in Gram-positive bacteria perform a variety of important roles: promote antibiotic resistance, cell signalling and substrate binding in ABC transport systems, protein export, sporulation, germination, bacterial conjugation, and many others. **[1]** Biosynthesis of bacterial lipoproteins is conducted via a pathway that appears to be highly conserved and unique to prokaryotes. **[2]** Following signal peptide-directed export of the prolipoprotein, processing occurs by the enzyme prolipoprotein diacylglycerol transferase (Lgt). Lgt uses phospholipid substrates and catalyses the addition of a diacylglycerol unit onto the thiol of a crucial conserved cysteine, which is located within the 'lipobox' motif at the cleavage region of the prolipoprotein signal peptide. **[3]** The lipid-modified prolipoprotein is processed by SpaseII, a lipoprotein-specific signal peptidase. SpaseII cleaves the signal peptide from the lipoprotein at the conserved cysteine present at the C-terminal end of the lipobox. This cysteine therefore forms the N-terminus of the mature lipoprotein. **[3]** It has been experimentally confirmed that these two steps are necessary and sufficient for protein lipidation in Gram-positive bacteria. **[4]** In certain organisms, the N-terminus of the lipoprotein is further modified by the addition of an amide-linked fatty acid. The additional processing step only occurs in some organisms as the enzymes responsible, lipoprotein aminoacyl transferases, are not found in low G+C Gram-positive bacteria.

Lipoprotein processing is controlled by two factors: signal peptide structure, which directs protein export, and an appropriately placed lipobox, which is essential for prolipoprotein recognition and modification by the appropriate enzymes. Lipoprotein signal peptide features are typical of signal sequences and consist of a positively charged N-region (owing to the presence of lysine and/or arginine), a central hydrophobic region and a cleavage C-region. **[5]** Differences in structure and composition between a lipoprotein signal peptide and a typical signal peptide have been described. Both Gram-negative and Gram-positive lipoprotein signal peptides are usually shorter in length than the typical signal peptide primarily due to shorter C-regions, which also possess apolar amino acids. The decreased length and apolar composition of the C-region effectively makes it a continuation of the H-region, which is primarily distinguished by sequence conservation that precedes the invariant lipid-modified cysteine. **[5]** This conserved sequence is referred to as the lipobox and is present at positions –3 to +1, typically taking the form of leucine, alanine/serine, glycine/alanine and cysteine.

The lipobox lipidation motif is represented in PROSITE by the regular expression {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C (PS00013). The permitted lipobox amino acids preceding the invariant cysteine at positions –1 to –4 are characterised by a lack of charged residues (no D, E, R, K) within the H-region. The PROSITE expression is also constrained by additional features of lipoprotein signal sequences: the cysteine at the cleavage site must be present between residues 15-35 of the sequence; and one lysine or arginine must be present in the first seven positions of the sequence. Taxon-specific variations in the lipobox have been discovered. Hakke showed that the

lipobox consensus of an experimentally determined set of spirochaetal lipoproteins varied in composition compared to that of *E. coli*. **[6]** A subset of Gram-positive lipobox sequences were also shown to be more restrictive than the PROSITE expression at positions –3 and –2. **[2]**

Lipoproteins are important vaccine targets in pathogenic bacteria. **[7]** Current vaccine design endorses the utilisation of *in silico* bacterial genome analysis to discover putative candidate vaccine. An algorithm for identification of lipoproteins would be essential to implementing a viable approach to 'reverse vaccinology' (the *in silico* and *in vitro* screening of whole genomes to identify antigens and hence candidate vaccines). In this context, we present LipPred, a web server for the identification of lipoproteins in bacteria, whose characteristics are tuned to the needs of reverse vaccinology.

Methodology:

Datasets

The training set for LipPred comprises 199 GRAM +VE and GRAM –VE bacterial lipoproteins obtained from DOLOP. **[8]** The use of unverified experimental data in model building raises obvious concerns over the ultimate quality of prediction. However, the number of experimentally verified lipoproteins is low, and using unverified data combined with probable lipoproteins increases accuracy.

Three distinct test sets were used. Firstly, data from Reinhardt and Hubbard **[9]**, which comprised 2031 non-membranous eukaryotic sequences, 268 prokaryotic inner membrane α-helical sequences and 997 non-membranous prokaryotic sequences. Both eukaryotic and prokaryotic sequences were used in the negative set to test the method's ability to distinguish bacterial lipoproteins from other sequences. Secondly, to assess accurately the ability of the method to predict experimentally confirmed lipoproteins, the 81 lipoproteins described by Gonnet *et al*. were used. A third test set was used to test the ability of the method to distinguish between lipoproteins and proteins with Type I signal peptides: this consisted of 307 GRAM +VE and GRAM -VE proteins extracted from a SignalP **[11]** non-redundant secreted data-set.

Algorithm

A probabilistic sequence model was used to allow for lipobox sequence variations. A range of sequence models were tested from 3-21 residues in length. As all lipoproteins possess a conserved cysteine at the C-terminus this was used as the central residue of the sequence model. A Naive-Bayesian network was trained using this data as described elsewhere (unpublished data). The network structure used one input node for each residue of the sequence model (hence 21 for the first model tried) and one output node. The output node can take the value of lipoprotein or non-lipoprotein. To train using the negative data-set, a cysteine was searched for between residues 10 and 50, and a sequence model was built centred on that cysteine. A lack of cysteine reduced the final negative data-set of 695 proteins.

To test a query protein, residues 10-50 were scanned to identify the presence of a cysteine, and a probability score calculated for whether the protein is a lipoprotein. The threshold for a positive score was taken to be 70%. Scanning for a cysteine was repeated up to the 50th residue, allowing the best-scoring lipobox to be found. Testing was conducted on all lipoproteins of the test-set using five-fold cross-validation, overall accuracy being obtained by averaging the five test-set results. The secreted protein set was also used to test the method. To be able to benchmark LipPred against other lipoprotein prediction methods, the same data-sets were used to query the LipoP server **[12]** and PROSITE. The SignalP data-set was also used to query LipoP to ascertain the ability of the method to distinguish Type I signal sequences from Type II lipoprotein signal sequences.

Implementation

The method is available as a web server: www.jenner.ac.uk/LipPred. Sequences can be entered or uploaded in either FastA or plain text format. Results are returned in standard or simple formats. Standard format provides a comprehensive description of lipoboxes and cleavage site locations, with associated probabilities. Simple format is plain text on one line, detailing the protein name (if provided), prediction of the protein class and the most likely cleavage site.

Utility and Caveats:

The best performing sequence model was found to be the -9 +0 model (data not shown); predicting with a sensitivity of 0.979 and a specificity of 0.742. To make a comparison with existing methods, the same positive and negative sequences were analysed using an algorithm that exploits the PROSITE regular expression PS00013 and the LipoP algorithm. **[12]** The results of all three methods are shown in Table 1. Comparison with other well-known methods of lipoprotein prediction indicates that LipPred achieves a significantly higher sensitivity but lower specificity of prediction. A detailed examination of the results indicated the relatively low specificity resulted from a high number of false positives. These sequences were of proteins of extra-cellular location, which is unsurprising as the lipoprotein signal sequence has a composition similar to that of the classical N-terminal signal sequence.

To assess the ability of the method to discriminate between lipoproteins and Sec-dependent signal sequences, the SignalP data-set was used. LipPred correctly identified 80% of the SignalP sequences as non-lipoproteins, compared to 95% by LipoP and 100% using regular expressions. It is good practice, when classifying proteins on the basis of sequence, to combine methods and thus obtain the best prediction. When SignalP was combined with LipPred, all of

the false positives of extra-cellular nature were identified as containing Type I signal peptides. This filter increases the overall specificity of LipPred to 0.846.

Method	Current dataset			SN
	SN	**SP**	**VLP**	**SPD**
LipPred	0.979	0.742	1.000	0.80
LipoP	0.860	0.989	0.926	0.95
PS00013 regex	0.791	0.996	0.951	1.00

Table 1: Results of LipPred in comparison with two currently used lipoprotein prediction methods. SN = sensitivity; SP = specificity**;** VLP = verified lipo-proteins; SPD = SignalP dataset

As the results show, the PS00013 regular expression also had the lowest sensitivity. This approach is a rather inflexible, simplistic method for protein function prediction. The method relies on a motif that is not always present within lipoproteins and hence cannot act as the definitive sequence marker for the protein class. The higher specificity of the LipoP method is likely a consequence of the method also identifying SpaseII-cleaved signal peptides and transmembrane proteins. The additional functionality built into LipoP gives an impressive degree of false-positive filtering, but this can be achieved by using other dedicated methods. The lower specificity of LipPred results from its flexibility in recognising highly variable lipoprotein signal sequences, but which also produces a higher rate of false-positive predictions. This flexibility makes LipPred a useful tool for lipoprotein discovery. The high sensitivity of LipPred is well illustrated by the accuracy of prediction obtained when the verified lipoproteins are used as the query set. 100% of the verified data-set was correctly identified as being lipoproteins, while LipoP identified 92.59%. This validates the use of training sets which lack conclusive experimental evidence of lipoprotein identity in order to achieve a high degree of prediction sensitivity.

LipPred is a fast, accurate tool for bacterial lipoprotein identification, whose characteristics are tuned to meet the needs of *in silico* 'reverse vaccinology'. Lipoproteins represent a class of vaccine target that has been exploited in many pathogenic bacteria. The outer-surface lipoprotein A of *Borrelia burgdorferi* was used as the basis of a vaccine for Lyme disease. **[7]** Studies using a variety of animal models have also demonstrated the ability of lipoproteins to provide protective immunity against bacterial pathogens. **[13]**

The suitability of lipoproteins as vaccine candidates results from their ability to be potent modulators of the host immune system owing to the presence of the lipolyated N-terminus. The lipoproteins of *M. gallisepticum* have been shown to be the most active immunogens during experimental chicken infections **[14]** with similar results having been obtained in cattle infections by *M. mycides* and *M. bovis.* **[15]**

Conclusion:

Current advances in vaccine development promote the use of *in silico* analysis of pathogen genomes to identify viable vaccine candidates. An algorithm to identify lipoproteins would form an essential part of any such 'reverse vaccinology' analysis, as their surface-exposed nature makes them more accessible to the receptors of the immune system. LipPred provides a highly sensitive algorithm for the identification of lipoproteins from sequence and the recognition of possible cleavage sites, together with their associated probabilities. The method is also capable of accepting genome-size data-sets. Its properties are honed to the needs of *in silico* vaccine identification: the high sensitivity of prediction provides an all-inclusive approach to protein classification, reducing the likelihood that a true lipoprotein will be missed. In this regard, LipPred compares well with alternative strategies. A false negative is, potentially, a worse misclassification than a false-positive prediction, which can be filtered out using methods specific for other protein classes or features. As lipoproteins are key vaccine targets **[7]**, this is of special relevance to computational vaccinology, where it is not desirable to miss proteins that may be vaccine candidates. LipPred is designed to address these requirements.

Acknowledgement:

PDT thanks the UK Medical Research Council for a Priority Area Studentship. We should like to thank Andrew Worth for his technical assistance. The Jenner Institute (previously the Edward Jenner Institute for Vaccine Research) wishes to thank its former sponsors: GlaxoSmithKline, the UK Medical Research Council, the UK Biotechnology and Biological Sciences Research Council, and the UK Department of Health.

References:

[01] I. C. Sutcliffe & D. J. Harrington, *Microbiology*, 148:2065 (2002) [PMID: 12101295]

[02] I. C. Sutcliffe & R. R. Russell, *J Bacteriol.*, 177:1123 (1995) [PMID: 7868582]

[03] H. Y. Qi, *et al., J Bacteriol.*, 177:6820 (1995) [PMID: 7592473]

[04] C. M. Petit, *et al., FEMS Microbiol. Lett.*, 200:229 (2001) [PMID: 11425480]

[05] G. von Heijne, *Protein Eng.*, 2:531 (1989) [PMID: 2664762]

[06] D. A. Haake, *Microbiology*, 146:1491 (2000) [PMID: 10878114]

[07] A. C. Steere, *et al., N. Engl. J. Med.*, 339:209 (1998) [PMID: 9673298]

[08] M. Madan Babu & K. Sankaran, *Bioinformatics*, 18:641 (2002) [PMID: 12016064]

[09] A. Reinhardt & T. Hubbard, *Nucleic Acids Res.*, 26: 2230 (1998) [PMID: 9547285]

[10] K. E. Gonnet, *et al., Proteomics*, 4:1597 (2004) [PMID: 15174130]

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

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

[13] D. West, *et al., Infect. Immun.*, 69:1561 (2001) [PMID: 11179327]

[14] G. Jan, *et al., Res. Microbiol.*, 146:739 (1995) [PMID: 8584796]

[15] A. Behrens, *et al., Microbiology*, 142:1863 (1996) [PMID: 8757750]

Edited by P. Kangueane

Citation: Taylor *et al.*, Bioinformation 1(5): 176-179 (2006)