Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Toward bacterial protein sub-cellular location prediction: single-class discrimminant models for all gram- and gram+ compartments

Paul D. Taylor [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 November 24, 2006; accepted December 02, 2006; published online December 02, 2006

Abstract:

Based on Bayesian Networks, methods were created that address protein sequence-based bacterial subcellular location prediction. Distinct predictive algorithms for the eight bacterial subcellular locations were created. Several variant methods were explored. These variations included differences in the number of residues considered within the query sequence - which ranged from the N-terminal 10 residues to the whole sequence - and residue representation - which took the form of amino acid composition, percentage amino acid composition, or normalised amino acid composition. The accuracies of the best performing networks were then compared to PSORTB. All individual location methods outperform PSORTB except for the Gram+ cytoplasmic protein predictor, for which accuracies were essentially equal, and for outer membrane protein prediction, where PSORTB outperforms the binary predictor. The method described here is an important new approach to method development for subcellular location prediction. It is also a new, potentially valuable tool for candidate subunit vaccine selection.

Key Words: Bayesian Networks; prediction method; subcellular location; membrane protein; periplasmic protein; secreted protein

Background:

Only certain microbial components - those open to surveillance by a host immune system - are likely to be potential subunit vaccines. Thus, subcellular location is, for bacteria, a principal determinant of immunogenicity. There are five subcellular locations in Gram- bacteria and three locations in Gram+ bacteria. There have been few attempts to generate prediction methods for all compartments, as most methods predict only certain locations. There are two basic types of prediction method. First, the manual construction of rules derived from our current knowledge of the diverse factors determining subcellular location, and secondly, the application of data-driven machine learning methods which automatically identify factors that determine subcellular location from features of proteins of known location.

The degrees of accuracy differ markedly between methods and compartments, reflecting either a paucity of data for a specific compartment or the lack of proper understanding of what determines protein location. Subcellular location prediction methods are often classified according to two factors: the input data required and the process of constructing prediction rules. Input data includes, the amino acid composition of the whole protein; sequence derived features of the protein (i.e. hydrophobic regions); the presence of certain motifs; and the sequence itself, whether whole or in part. It is common to combine several types of input data. Expression patterns have been used, together with sequence motifs, since expression levels and subcellular location are often correlated. Phylogenetic profiles can identify protein location through sequence similarity, based on the premise that the location of close homologues can be assumed to be equal.

The output of prediction methods can be binary or multi-category. A binary predictor indicates if a protein is located in one category or not. A multi-category predictor will attempt to sort the query sequence into one of several possible locations. Binary predictors often have a high rate of false positive prediction. Multi-category prediction methods suffer reduced accuracy for certain location due to a scarcity of data or the complexity of the signal for that compartment or both. As always, the better the data the more accurate and reliable the resulting method. Signal complexity is a more complicated issue. A very complex signal requires abundant data to ensure that all signals are recognised with a high degree of confidence. A simple signal can also be problematic, in that many proteins may posses the sorting signal by chance alone. For example, SWISS-PROT contains twice as many non-perioxisomal proteins containing the PTS1 sorting signal than actual perioxisome proteins with the PTS1 signal.

Several predictors exist. PSORT **[1]** is a knowledge-based, multi-category program for subcellular location prediction. It is often used as the gold standard for such prediction. PSORT consists of two different programs: PSORT I (predicting 17 subcellular compartments; trained on 295 proteins) and PSORT II (predicting 10 locations; trained on 1,080 yeast proteins), each using different algorithms.

PSORT I uses "if-then" rules sequentially in a tree-like manner. PSORT II scores sequence-derived features for each localisation and then predicts using *k* nearest-neighbours classification. PSORT II is the more accurate of the two methods. Using a test set of 940 plant proteins and 2,738 non-plant proteins, the accuracy of PSORT I and II was 69.8% and 83.2%. PSORT-B is specifically for the prediction of bacteria. **[2]** NNPSL **[3]** is a neural network method that predicts four locations - cytoplasmic, extracellular, mitochondrial and nuclear - using amino acid composition. 66.1% accuracy was reported for multi-category prediction. Mitochondrial localisation is predicted by MitoProt II **[4]**, which has certain similarities with bacterial prediction. When tested on 3,419 human mitochondrial proteins obtained from SWISS-PROT, an accuracy of 86.7% was reported.

In this paper, we describe the development of an array of binary predictors, one for each of the eight Gram+ and Gram- bacterial compartments. In contrast to other methods, these predictors use a single methodology based on Bayesian Networks (BNs), which makes use of the same predictive architecture, the same sequence representation, and focuses on the same region of the sequence. We have also developed a predictor which discriminates between soluble and non-soluble proteins. Together these methods form a cohesive, integrated, and standardised approach to subcellular location prediction. After validating this method using cross-validation and test sets, we compared its predictivity to that of PSORTB.

Methodology:

Dataset

An algorithm was used to mine the bacterial subset of SWISS-PROT release 40. **[5]** Initially, bacterial status was confirmed using the OC line code of the SWISS-PROT entry. Entries were split into Gram+ and Gram- at the superfamily level. The following were assigned as Gram+: actinobacteria; deinococcus; thermus; firmicutes; planctomycetes; and thermotogae, and the following assigned as Gram-: chlamydia; verrucomicrobia; cyanobacteria; chloroflexi; fusobacteria; nitrospirae; proteobacteria; spirochaetes; chlorobi; and bacteroidete. The SWISS-PROT subcellular location descriptions (lines labelled CC) were then searched to identify if the subcellular location was known. To remove proteins of uncertain location, only entries not labelled as 'potential', 'probable', 'hypothetical', 'possibly' or 'by similarity', were incorporated into the final data-set. A non-redundant data-set of proteins was obtained using CLUSTALW. **[6]** If two or more proteins were found to have sequence similarity higher than 90% then all but one were removed from the data-set. The algorithm and subsequent CLUSTALW analysis produced a Gram+ data-set of were 272 extracellular proteins, 375 membranous proteins and 1500 cytoplasmic proteins, while the final Gram+ data-set contained 185 extracellular, 159 outer membrane, 432 periplasmic, 273 inner membrane and 2480 cytoplasmic proteins.

Single protein class Gram- and Gram+ prediction

Individual methods were created for all eight bacterial locations, both Gram+ (cytoplasmic, membrane and extracellular) and Gram- (cytoplasmic, inner membrane, periplasm, outer membrane and extracellular). Three sequence representations (actual residues, amino acid composition, and normalised amino acid composition) and six sub-sequence length were used (1-10, 1-20, 1-30, 1-40, 1-50 and the whole sequence). Each variation of the method was used to train a single Näive-Bayes network for each of the eight locations, thus 24 BNs were constructed per location. For amino acid composition, BNs possessed 20 input nodes, each representing the raw number or normalised percentage residue composition. The number of input nodes in the actual residue BNs varies dependent on the location. The Gram+ cytoplasmic, membranous and extracellular BNs had 1436, 1852 and 1627 input nodes respectively. The Gram- cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular BNs have 1368, 874, 1014, 2248 and 1848 input nodes respectively. The individual location networks are binary predictors and therefore have a single output node that is either on or off for the predicted location.

A combined soluble class predictor, able to distinguish soluble (cytoplasmic and periplasmic) proteins from all other locations, was also created. This was based on considering the whole protein using pseudo-amino acid composition. **[7]** Pseudo-amino acid composition is used in preference to simple amino acid composition, as it attempts to model sequence-order effect. It is calculated using a formula that assigns 40 scores to 40 input nodes for each sequence: 20 representing normal amino acid composition and 20 representing sequence-order effects. The output node can possess the values of soluble or non-soluble.

Testing of the networks was performed using their respective training sets under five-fold cross-validation. Each location-specific network was trained using known positive data for that location, while the negative training set was all other sequences, except where a location is present in both Gram- and Gram+ bacteria i.e. cytoplasmic, membrane (in Gram+) and inner membrane (in Gram-), and extracellular. For these networks, sequences from the equivalent location in the other Gram class were excluded. For the soluble predictor, the positive data-set was all cytoplasmic and periplasmic sequences obtained from SWISS-PROT and the negative data-set was the sequences from all other locations. To assess the predictivity of the Bayesian approach, the same data-sets were submitted to PSORTB.

Results and Discussion:

Results for the eight individual compartments are shown in tables 1a, 1b and 2. The most accurate prediction was achieved using amino acid composition for the first 50 residues. Both Gram+ and Gram- predictors achieved their best accuracies under the same conditions. Prediction accuracy generally increases with increasing sub-sequence length from 10 residues to 50 residues, and then tails off slightly when the whole sequence is considered. For the range of sub-sequence lengths the amino acid composition sequence representation consistently outperformed the other two representations. The sub-sequence size used affects the accuracy in a much more obvious manner. The N-terminal compartment sorting signals are present in the sub-sequence lengths used in all networks and a method that best models the sorting signals will achieve the best accuracy.

Unsurprisingly the accuracies of all locations are highest when the first 50 residues are considered as this length will encompass the full lengths of the vast majority of signal sequences. The shorter sub-sequence lengths will only consider the n region and possibly part of the h region. The variances in n-region charge and, therefore residues, does vary from signal peptide types and therefore can be used with some degree of accuracy to distinguish between different signal peptide types. The h-region also varies in length and composition between different signal peptide types and therefore when both are considered a higher degree of accuracy is achieved.

The performance of the methods was compared to that of PSORTB (See Table 3). The individual location methods outperform PSORTB with the exception of the Gram+ cytoplasmic proteins, in which the accuracies were approximately equivalent, and the outer membrane proteins, in which PSORTB outperforms the individual method. This may be because outer membrane proteins have a notoriously variable amino acid composition due to TM strands being exposed to both membrane and pore. The amino acid composition method therefore may not be suitable for the prediction of outer membrane proteins.

Sequence representation	Sub-sequence length	Cytoplasmic accuracy (%)		Membrane accuracy (%)		Extracellular accuracy (%)	
		Spec	Sens	Spec	Sens	Spec	Sens
Amino acid composition	10	88.42	34.25	73.54	55.25	76.23	42.55
	20	89.84	42.83	72.92	58.03	73.73	38.08
	30	94.52	55.90	84.25	67.06	77.81	65.97
	40	93.6	77.68	89.03	78.94	80.51	81.60
	50	**96.78**	**94.24**	**96.30**	**89.51**	**82.53**	**93.90**
	All sequence	91.51	90.82	91.41	80.03	84.91	74.84
Actual amino acids	10	52.51	22.36	12.14	1.41	0.04	1.14
	20	63.35	26.51	15.52	6.77	3.62	2.62
	30	64.93	34.99	24.05	9.59	9.27	5.01
	40	68.34	38.27	29.15	17.97	15.73	12.63
	50	69.41	48.15	32.33	16.11	18.42	11.09
	All sequence	72.42	58.73	36.87	23.72	16.64	14.78
Normalised amino acid composition	10	89.52	29.86	69.93	52.93	77.36	40.42
	20	89.72	38.11	70.95	61.09	78.41	47.73
	30	91.42	44.19	74.01	73.60	82.25	57.80
	40	92.20	61.07	79.44	85.26	81.09	71.03
	50	93.13	79.14	83.10	97.88	83.98	84.76
	All sequence	90.96	73.77	84.76	93.61	80.15	84.08

Table 1a: Prediction accuracies of the Gram+ individual location predictors. The results of highest accuracy are shown in bold. Specificity refers to the accuracy of prediction from the positive test set while sensitivity refers to accuracy of prediction for the negative test set

Sequence representation	Sub-sequence length	Cytoplasmic accuracy (%)	Inner Membrane accuracy (%)	Periplasmic accuracy (%)	Outer Membrane accuracy (%)	Extra-celluar accuracy (%)
Amino acid composition	10	98.35	78.42	84.35	48.85	71.14
	20	92.52	81.09	88.89	56.50	76.91
	30	94.99	91.75	90.14	63.88	81.06
	40	96.34	89.33	93.98	69.25	82.37
	50	**97.48**	**96.83**	**94.57**	**77.90**	**87.97**
	All sequence	91.41	94.79	94.02	73.21	81.96
Actual amino acids	10	68.53	64.36	24.79	13.16	52.62
	20	74.52	60.23	33.05	14.93	58.35
	30	77.90	61.85	41.21	17.09	52.70
	40	74.08	66.33	45.82	24.51	55.08
	50	79.76	64.67	53.68	22.74	61.98
	All sequence	73.13	63.16	53.68	25.88	59.22
Normalised amino acid composition	10	94.32	77.35	80.41	51.51	71.01
	20	93.45	84.24	83.85	53.86	75.25
	30	93.78	86.94	87.02	57.12	72.09
	40	96.26	90.24	91.97	63.26	73.63
	50	94.78	93.12	93.52	61.03	77.60
	All sequence	93.21	91.51	92.87	67.73	74.28

Table 1b: Prediction accuracies of the Gram- individual location predictors. The results of highest accuracy are shown in bold

Sequence representation	Sub-sequence length	Cytoplasmic accuracy (%)	Inner Membrane accuracy (%)	Periplasmic accuracy (%)	Outer Membrane accuracy (%)	Extra-cellular accuracy (%)
Amino acid composition	10	51.03	83.52	44.02	37.85	73.31
	20	53.64	81.09	58.23	51.67	76.90
	30	64.07	84.24	62.68	64.69	85.02
	40	81.75	88.31	79.43	77.11	86.42
	50	**90.13**	**94.76**	**92.01**	**86.36**	**92.85**
	All sequence	88.24	93.41	84.42	86.02	88.22
Actual amino acids	10	40.03	53.59	20.52	12.05	23.24
	20	41.49	58.21	23.84	16.73	22.86
	30	48.79	68.84	55.08	22.41	26.72
	40	55.32	61.33	42.21	25.62	30.55
	50	58.62	63.71	49.33	32.41	29.86
	All sequence	64.28	59.35	43.57	34.79	30.08
Normalised amino acid composition	10	44.63	84.04	41.93	44.32	74.59
	20	48.32	88.68	56.26	46.72	77.35
	30	61.04	87.14	63.17	51.48	81.68
	40	71.38	93.73	71.87	56.37	84.24
	50	77.20	96.26	78.88	68.53	85.32
	All sequence	83.56	92.47	73.29	59.25	84.99

Table 2: Prediction accuracies of the Gram- individual location predictors for the negative test sets. The results of highest accuracy are shown in bold

Gram-type	Subcellular location	PSORTB accuracy (%)	Individual location predictors accuracy (%)
Gram+	Cytoplasmic	96.38	96.78
	Membranous	91.47	96.30
	Extra-cellular	70.42	82.53
Gram-	Cytoplasmic	91.37	97.48
	Inner membrane	94.68	96.83
	Periplasmic	84.69	94.57
	Outer membrane	83.70	77.90
	Extra-cellular	77.55	87.97

Table 3: Results of the individual method predictions compared to the PSORTB algorithm

Conclusion:

In the search for viable subunit vaccines, highly accurate methods for sub-cellular location prediction, such as the set of binary predictors we describe here, can aid reverse vaccinology directly. In particular, the soluble predictor will significantly reduce the number of sequences that must be tested as vaccine targets. This is a breakthrough method since cytoplasmic and periplasmic proteins are usually predicted as the residue left after positive predictions for all other locations. Binary predictors discriminate between two classes; they are often more accurate than multi-outcome predictors, such as PSORTB, which is itself built from several binary predictors. We judged the accuracy of our approach and that of PSORTB and found that out set of location-specific methods, which is built on a single underlying methodology, compared favourably. However, when several binary predictors are used to create a multi-outcome prediction method, an efficient and effective way of combining the disparate outputs is still required. Nonetheless, our new BN approach is an important, competitive advance in the development of subcellular location prediction methods. It should, it is own right, prove a powerful tool for candidate subunit vaccine selection.

Acknowledgement:

PDT wishes to thank the MRC for a priority area studentship. The Jenner Institute, (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] K. Nakai & P. Horton, *Trends Biochem Sci.*, 24:34 (1999) [PMID: 10087920]

[02] J. L. Gardy, *et al.*, *Bioinformatics*, 21:617 (2005) [PMID: 15501914]

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

[04] M. G. Claros, *Comput Appl Biosci.*, 11:441 (1995) [PMID: 8521054]

[05] M. Schneider, *et al.*, *Plant Physiol Biochem.*, 42:1013 (2004) [PMID: 15707838]

[06] R. Chenna, *et al.*, *Nucleic Acids Res.*, 31:3497 (2003) [PMID: 12824352]

[07] K. C. Chou, *Proteins*, 43:246 (2001) [PMID: 11288174]

Edited by P. Kangueane

Citation: Taylor *et al.*, Bioinformation 1(8): 276-280 (2006)