Bioinformation by Biomedical Informatics Publishing Group open access

www.bioinformation.net Prediction Model

Multi-class subcellular location prediction for bacterial proteins

Paul D. Taylor [1, 2], 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 10, 2006; accepted November 22, 2006; published online November 24, 2006

Abstract:

Two algorithms, based on Bayesian Networks (BNs), for bacterial subcellular location prediction, are explored in this paper: one predicts all locations for Gram+ bacteria and the other all locations for Gram- bacteria. Methods were evaluated using different numbers of residues (from the N-terminal 10 residues to the whole sequence) and residue representation (amino acid-composition, percentage amino acid-composition or normalised amino acid-composition). The accuracy of the best resulting BN was compared to PSORTB. The accuracy of this multi-location BN was roughly comparable to PSORTB; the difference in predictions is low, often less than 2%. The BN method thus represents both an important new avenue of methodological development for subcellular location prediction and a potentially value new tool of true utilitarian value for candidate subunit vaccine selection.

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

Background:

Only proteins liable to surveillance by the immune system are likely candidate subunit vaccines. Thus, for bacteria, subcellular location can be a prime arbiter of immunogenicity. There are five principal subcellular locations in Gram- bacteria (extracellular, outer membrane, periplasmic, inner membrane, or cytoplasmic) and three locations in Gram+ bacteria (extracellular, membrane, or cytoplasmic). Components of the proteome contain signals which can direct proteins to one or more of these locations. Such signals are legion. They can, for example, be explicit sequence motifs recognised by a membrane transporter. They can also be coincidental physical properties that render certain proteins compatible with their environment and were derived through an evolutionary process. An organism can read such signals well enough *in vivo*, and there is thus much interest in effectively reproducing this *in silico*. Bioinformatician's have, therefore, attempted to identify both sequence motifs and overall physical properties of proteins indicative of protein subcellular location.

Many methods have attempted to predict subcellular location. There are two basic types of prediction method: manual construction of rules derived from factors thought to determine subcellular location and the application of data-driven machine learning methods that automatically identify factors that determine cellular localisation, using proteins of known location as training data. The degrees of accuracy differ markedly between methods and compartments, reflecting either a lack of data for a specific compartment or the complexity of factors controlling the location of certain proteins.

However, there have been few, if any, real attempts to create prediction methods for all such compartments, since most methods predict only a subset of the 'most interesting' locations. An exception to this is PSORTB, which is a sub cellular location-prediction expert system developed specifically for bacteria. **[1]** PSORTB is a modular system based on 6 prediction algorithms. A query protein undergoes analysis by each of the modules and the results are then combined. The modules that form PSORTB are: SCL-BLAST, which uses sequence similarity to known proteins to identify location; PROSITE, which detects motifs indicative of subcellular location **[2]**; HMMTOP, a method for the prediction of TM domains, to identify membrane proteins **[3]**; outer membrane protein motifs are identified using sequences occurring only in TM beta barrel proteins; SubLocC, a support vector machine based method, which assigns a cytoplasmic or non-cytoplasmic location based on amino acid-composition; and a hidden Markov model trained to identify signal peptide cleavage sites. Prediction of a query sequence location is reported as the likelihood that a query protein belongs to a particular compartment. PSORTB has a precision of 96.5% and a recall of 74.8%.

In the context of bacterial subcellular location prediction, methods based on Bayesian Networks (BNs) are explored in this paper. Two algorithms which predict all locations for Gram+ and Gram- bacteria were created. A range of variant methods was evaluated, with differences including the number of residues considered (from the N-terminal 10 residues to the whole sequence) and residue representation (amino acid-composition, percentage amino acid-composition or normalised amino acid-composition). The accuracy of the best resulting BN was then compared to PSORTB.

Methodology:

Dataset

An algorithm was used to mine the bacterial subset of SWISS-PROT release 40. **[4]** 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. **[5]** 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.

Combined bacterial subcellular location predictor method

When training the method, a variety of sequence representations were examined. Six different sequence lengths were used: residues 1-10 of the N-terminus, residues 1-20, residues 1-30, residues 1-40, residues 1-50, and the whole protein sequence. For each sub-sequence, amino acids were represented in three ways: as the residues themselves, as the amino acid-composition (for each residue, the total number of each amino acid in the sub-sequence); and as the normalised amino acid-composition (for each amino acid, the residue composition divided by the total number of amino acids in the sub-sequence).

Each representation was tested with each sub-sequence length, creating 18 Näive-Bayes networks. The amino acid-composition and normalised composition sequence representations used BNs comprising 20 input nodes and 1 output node. During training, a sub-sequence is extracted from the original protein sequence and its composition calculated. To train the BN for an individual sub-sequence, each of the 20 input nodes is assigned a different composition value: the first contains that of alanine, the second that of arginine, etc. This procedure is repeated until all sub-sequences have been used to train the network. The output node is given the value of the subcellular location of the training protein, which are different for Gram+ (5 locations) and Gram- (3 locations).

The directed acyclic graph (DAG) required when the residue representation was the actual amino acid sequence, varied when different sequence lengths were used. A length of 10 residues required a BN with 10 input nodes, for example. When the whole protein sequence was used, the DAG required as many input nodes as the protein had amino acids. Since the same network is used for all the proteins of the data-set, the longest protein determined the total number of input nodes used. For the Gram- predictor 2248 input nodes were used and for the Gram positive predictor 1852 input nodes were used. The amino acids were converted to integers, 1 to 20 according to the alphabetical order of their single letter representations i.e. alanine (A) had the value 1, cysteine (C) was 2, etc. When training the network, the first input node takes as its value the first residue, the second the second, and so on until the end of the sequence. Input nodes that do not have a corresponding amino acid, due to the training sequence being shorter than the maximum length, were assigned the value 0. The output node is given the value of the subcellular location of the training protein.

Testing of the network was performed using the training set under five-fold cross-validation. For all networks, the negative set chosen was the equivalent data-set of the opposite Gram-type. To assess the predictivity of the Bayesian approach, the same data-sets were submitted to the PSORTB predictor.

Results and Discussion:

For both Gram+ and Gram- predictors the same combination of residue representation and sub-sequence length produced the most accurate results: amino acid-composition and a sub-sequence length of 50 residues. See Tables 1 and 2. The accuracy of both predictors increased with increasing sub-sequence length, up to 50 residues. Generally, both predictors were more accurate when using amino acid-composition. The worst performing representations is the one based on residues, which tries to capture residue position specific information. Apparent inadequacies of the representation may arise from the structure of the BN DAG requiring it: the number of input nodes equalled the length of the longest sequence; all other sequences therefore had many nodes assigned a value of 0 during training. For each compartment, the longest sequence was many times larger than the average sequence, thus many input nodes for most sequences had little predictive benefit.

The sub-sequence length affected accuracy more obviously. Unsurprisingly, the accuracies of all locations are highest when the first 50 residues are considered, since this will encompass the entire length of the vast majority of signal sequences. Shorter lengths may neglect important regions within such signal peptides. Charge, length, and composition, among other properties, will vary between different signal sequences and can therefore be used to distinguish accurately between different signal peptides.

A surprising feature of the results was that in most cases the accuracy of amino acid-composition for the whole protein was close to the accuracy of just the first 50 residues. However, for the extra-cellular compartment Gram+ predictor, the whole protein composition had a higher accuracy. This was unexpected as the un-normalised composition varied significantly with sequence length. A possible explanation is that Gram+ extracellular sequences have a very different length distribution to sequences from other compartments. The average length of sequences from each compartment was calculated. For the Gram+ proteins the average sequence length of the extracellular set was 397, compared to 491 (membranous proteins) and 442 (cytoplasmic). Further support comes from the Gram- sequence lengths, which were found to be 549 (extracellular), 568 (outer membrane), 322 (periplasmic), 400 (inner membrane), and 448 (cytoplasmic). If the BN based on composition draws its predictivity from the atypical Gram+ sequence length distribution, then the accuracy for the negative set should be low, since sequence length is nearer that of Gram+ extracellular sequences.

Comparing the best performing multi-location BNs to PSORTB indicates that their accuracy is roughly equivalent; the discrepancy between predictions is typically low, often less than 2%. See table 3. Exceptions include the extracellular compartment (both Gram- and Gram+) and membrane prediction. The prediction of extracellular location is more accurate for both Gram- (8.57% higher than PSORTB) and Gram+ (7.86 higher). For membranous prediction, PSORTB has an accuracy which is 20.54% higher than that of the Gram+ multi-location predictor. This may be because PSORTB is specifically trained to identify TM spanning regions.

Sequence representation	**Sub-sequence length**	**Cytoplasmic accuracy (%)**	**Membrane accuracy (%)**	**Extracellular accuracy (%)**	**Negative set accuracy (%)**
Amino acid-composition	10	98.84	3.73	8.20	32.25%
	20	94.15	19.20	45.08	29.43%
	30	93.94	41.07	52.87	51.11%
	40	93.94	52.53	52.05	77.62%
	50	**94.89**	**70.93**	**78.28**	**92.25%**
	All sequence	94.49	59.20	48.77	94.71%
Amino acids	10	60.11	0.00	0.37	21.14%
	20	79.71	0.80	4.41	23.51%
	30	85.64	3.73	8.82	31.23%
	40	87.81	23.73	13.97	42.24%
	50	90.06	24.27	25.00	43.25**%**
	All sequence	99.12	26.93	36.40	47.14%
Normalised amino acid-composition	10	100	30.67	56.15	25.26%
	20	100	31.73	61.48	29.43%
	30	100	46.93	56.56	42.15%
	40	100	59.47	57.38	56.45%
	50	100	65.00	75.00	78.26%
	All sequence	100	65.00	55.33	76.15%

Table 1: Results of the Gram+ all compartments predictor. The best performing network is highlighted in bold

Sequence represent-ion	**Sub-sequence length**	**Cytoplasmic accuracy (%)**	**Inner Membrane accuracy (%)**	**Periplasmic accuracy (%)**	**Outer Membrane accuracy (%)**	**Extracelluar accuracy (%)**	**Negative set accuracy (%)**
Amino acid-composition	10	78.14	13.31	61.14	1.13	55.14	80.14
	20	81.61	49.14	63.73	7.14	59.62	84.72
	30	85.25	52.51	69.09	34.15	76.15	83.62
	40	86.14	74.62	72.24	59.25	71.17	89.25
	50	**84.74**	**91.16**	**79.96**	**70.75**	**86.12**	**92.42**
	All sequence	89.14	84.57	71.15	55.25	80.24	87.59
Amino acids	10	32.55	22.21	2.14	0.24	8.36	27.73
	20	35.25	27.52	7.36	2.51	14.82	32.15
	30	29.83	32.15	12.93	7.37	17.93	34.85
	40	40.24	37.86	18.74	8.52	24.99	41.84
	50	41.84	32.41	34.14	12.41	28.25	36.83
	All sequence	40.34	44.83	27.41	15.84	22.51	40.14
Normalised amino acid-composition	10	63.26	22.15	52.36	0.42	40.25	57.12
	20	64.37	53.87	58.25	2.35	43.95	58.51
	30	69.73	68.46	65.36	12.94	51.64	68.20
	40	72.63	81.36	73.36	33.73	52.72	71.14
	50	74.89	92.97	71.l4	60.26	66.67	74.28
	All sequence	71.46	91.73	70.73	52.24	63.26	71.75

Table 2: Results of the Gram- all compartments predictor. The best performing network is highlighted in bold

Gram-type	**Subcellular location**	**PSORTB accuracy (%)**	**Multi-location predictor accuracy (%)**
Gram+	Cytoplasmic	96.38	94.89
	Membranous	91.47	70.93
	Extra-cellular	70.42	78.28
	Negative set	93.86	92.25
Gram-	Cytoplasmic	91.37	84.74
	Inner membrane	94.68	91.96
	Periplasmic	84.69	79.96
	Outer membrane	83.70	70.75
	Extra-cellular	77.55	86.12
	Negative set	90.02	92.42

Table 3: Accuracy of PSORTB bacterial subcellular location predictor in comparison to the most accurate methods produced

Conclusion:

Good levels of accuracy were achieved, yet PSORTB outperformed the BN method. Since our approach attempts to utilise a single method and sequence representation to capture all information relevant to bacterial subcellular location, the performance of the BN method reported here is most encouraging. When comparing our method to PSORTB, we see a single methodology competing against an expert system, which is specifically designed to capitalise on best-in-class methods. Constructing a successful multi-outcome predictive method is difficult. Prediction is made between input variables that are very difficult to separate using any method. The generally lower degree of prediction accuracy of the BN approach is most likely due to PSORTB applying many algorithms, each specifically trained to address the individual requirements of each particular location. Clearly, this strategy is more likely to produce a significantly greater level of accuracy. However, the BN method described here is nonetheless very competitive, notwithstanding such arguments. Thus, we can aver that BNs represent an important new avenue in subcellular location prediction and that our implementation is in itself a potential powerful new tool for candidate subunit vaccine selection with real utilitarian value.

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

[02] E de Castro, *et al., Nucleic Acids Res.,* 34:W362 (2006) [PMID: 16845026]

[03] G. E. Tusnady & I. Simon, *Bioinformatics,* 17:849 (2001) [PMID: 11590105]

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

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

Edited by P. Kangueane
Citation: Taylor *et al.,* Bioinformation 1(7): 260-264 (2006)