Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Combining algorithms to predict bacterial protein sub-cellular location: Parallel versus concurrent implementations

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 05, 2006; published online December 06, 2006

Abstract:

We describe a novel and potentially important tool for candidate subunit vaccine selection through *in silico* reverse-vaccinology. A set of Bayesian networks able to make individual predictions for specific subcellular locations is implemented in three pipelines with different architectures: a parallel implementation with a confidence level-based decision engine and two serial implementations with a hierarchical decision structure, one initially rooted by prediction between membrane types and another rooted by soluble versus membrane prediction. The parallel pipeline outperformed the serial pipeline, but took twice as long to execute. The soluble-rooted serial pipeline outperformed the membrane-rooted predictor. Assessment using genomic test sets was more equivocal, as many more predictions are made by the parallel pipeline, yet the serial pipeline identifies 22 more of the 74 proteins of known location.

Keywords: beta barrel transmembrane protein**;** prokaryotic membrane proteins**;** Bayesian Networks**;** prediction method**;** subcellular location

Background:

Subcellular location is amongst the prime arbiters of host immunogenicity within Bacteria. Many algorithms have been developed which predict one or a few locations, most often when a sequence is a membrane protein. For a successful - and practical - *in silico* reverse-vaccinology analysis we need to predict multiple localisations reliably and consistent accuracy. Thus our goal is take a genome and partition its gene products between multiple subcellular compartments. A viable strategy for achieving this aim is to combine together a set of individual binary predictors, which discriminate between a positive and a negative class, and thus develop a functional reverse-vaccinology pipeline. **[1, 2]**

Pipelines are a commonly used computational framework for combining separate modules into one. There are two common implementations of pipelines for prediction of properties from data: parallel and serial. In a parallel implementation, the multiple modules that are to be combined are run simultaneously and a further new module is required to decide which module has made the correct prediction. Parallel pipelines, in the majority of cases, will produce a more accurate prediction (with the presumption that the combination method is of a good standard), as the range of methods used when conducting a prediction is greater. Serially implemented pipelines, employ a tree-like method for the execution of the individual binary modules. The modules that make the final classification are at the top of the tree with modules that narrow down the possibilities, before the final binary predictors forming the lower branches. To make a classification with a serial pipeline, the root module is used first to analyse the data and then, based on its binary result, the pipeline will progress down one of the possible branches.

With a serial implementation, not all modules are used to produce the final result and therefore the runtime of the pipeline is nearly always quicker than an equivalent parallel implementation. The major flaw in a serially implemented predictive pipeline is that if the final module produces a negative prediction then no prediction is returned. The removal of such a “no prediction” result is dependent on two factors: the production of end module binary predictors that have an accuracy as close to perfect as possible, which would never produce false negative predictions; and having root and branch prediction modules that are as efficient as possible narrowing down the possible results so that the correct end module is always used.

We implement here serial and parallel pipelines for prediction of bacterial subcellular location and compare both the execution performance and the accuracy of the results using literature data.

Methodology:

Dataset

An algorithm was used to mine the bacterial subset of SWISS-PROT release 40. **[3]** Initially, bacterial status was confirmed using the OC line code of the SWISS-PROT entry. Entries were split into Gram-positive and Gram-negative at the superfamily level. The following were assigned as Gram-positive: actinobacteria; deinococcus; thermus; firmicutes; planctomycetes; and thermotogae, and the following assigned as Gram-negative: 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. Although CD-HIT **[4, 5, 6]** would, on reflection, have been a better choice, nonetheless a useable non-redundant data-set of proteins was obtained using CLUSTALW. **[7]** 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 procedure generated a Gram-positive data-set of were 272 extracellular proteins, 375 membranous proteins and 1500 cytoplasmic proteins, while the final Gram-negative data-set contained 185 extracellular, 159 outer membrane, 432 periplasmic, 273 inner membrane and 2480 cytoplasmic proteins.

Parallel and serial pipelines for subcellular location prediction

The parallel pipeline has a pre-defined workflow structure, which was controlled by a Perl script. A user-defined FASTA format file of protein sequences is read and then each sequence is entered, successively, into nine algorithms: five for Gram-negative locations, three for Gram-positive locations, plus a lipoprotein algorithm. **[8]** Each algorithm produces a "yes or no" prediction for one individual subcellular location. Overall output from the pipeline took one of seven values: cytoplasmic, inner membrane, periplasmic, outer membrane, Gram-positive membrane, extracellular and lipoprotein. The prediction confidence score from each binary prediction was combined to generate a final predicted location using a Näive-Bayes network. The experimentally-defined location of each protein was used to learn probabilities associated with each location.

Construction of the serial pipeline relies on the appropriate positioning of different algorithms within the tree-like decision structure. Two root algorithms were evaluated for the serial pipeline: the membrane class predictor and the soluble predictor. Perl scripts again controlled Pipeline workflow. A query protein is entered into the root algorithm. Depending on its outcome, the query protein is then sent to the next algorithm in the bifurcating pipeline. This process continues until a definite outcome is reached (purple boxes in Figures 1 and 2).

Validation

Validation was undertaken in three ways. First, serial and parallel pipelines were tested using training data under five fold cross-validation. Second, possible vaccine targets (Gram-positive membrane proteins, extracellular proteins and Gram-negative outer membrane proteins) were compared to non-targets (all other sequences). Thirdly, the ability of a pipeline to predict on a genomic scale was assessed: the translated ORFs of the Gram-negative bacteria *Neisseria meningitidis* MC58 and the Gram-positive bacteria *Staphylococcus aureus* subsp. aureus MRSA252 was run through the parallel pipeline and the best performing serial pipeline.

Results and Discussion:

Under cross-validation, the parallel predictor had an all-compartments accuracy of 93.64% compared to 92.64% for the soluble-rooted pipeline and 89.65% for the membrane-rooted serial pipeline. For vaccine-targets versus non-vaccine-targets, the parallel pipeline had the highest accuracy and the highest sensitivity, which was significantly greater than the serial predictors. The parallel pipeline had a positive specificity of 91.27% and a negative specificity of 93.93%; the membrane-rooted serial pipeline had equivalent values of 84.77% and 90.05%; and soluble-rooted serial pipeline had corresponding values of 82.27% and 94.41%. The high specificity of the soluble-rooted serial method, which outperformed the parallel method, is due to the highly accurate identification of cytoplasmic proteins: the overwhelming majority of the data-set. Computational runtime for the parallel, membrane-rooted serial, and soluble-rooted serial networks were 70,131 seconds, 44,229 seconds, and 46,112 seconds, respectively.

The difference in specificity for the serial pipelines highlights the importance of choosing the appropriate root algorithm. The soluble predictor has a greater accuracy than the membrane class predictor, misclassifying fewer query sequences. To quantify this, we calculated the percentage of unclassified predictions compared to the percentage of wrong predictions. Of the 10.45% of the data-set incorrectly predicted by the membrane rooted serial predictor 6.79% was predicted as unclassified, while of the 7.36% incorrectly predicted of the soluble rooted serial predictor only 2.04% was unclassified. This equates to 249 proteins, any of which may be a possible new vaccine target.

For the 2079 ORFs of *Neisseria meningitidis* MC58 841, the parallel pipeline predicted 841 cytoplasmic proteins, 46 periplasmic proteins, 252 inner membrane proteins, 56 outer membrane protein, 34 extracellular proteins, 124 ORFs had multiple locations, and 726 were not predicted. The soluble-rooted serial pipeline predicted 773 cytoplasmic proteins, 39 periplasmic proteins, 248 inner membrane proteins, 41 outer membrane protein, 22 extracellular proteins, and 785 ORFs had no predicted location. For the 2656 ORFs of the *Staphylococcus aureus* subsp. aureus MRSA252 genome, the parallel pipeline predicted 1468 cytoplasmic, 442 membrane proteins, 87 ORFs extracellular, 139 ORFS had multiple locations, while 520 had no predicted location. The soluble rooted serial pipeline predicted 764 cytoplasmic, 376 membrane proteins, 63 extracellular, and 785 ORFs had no predicted location.

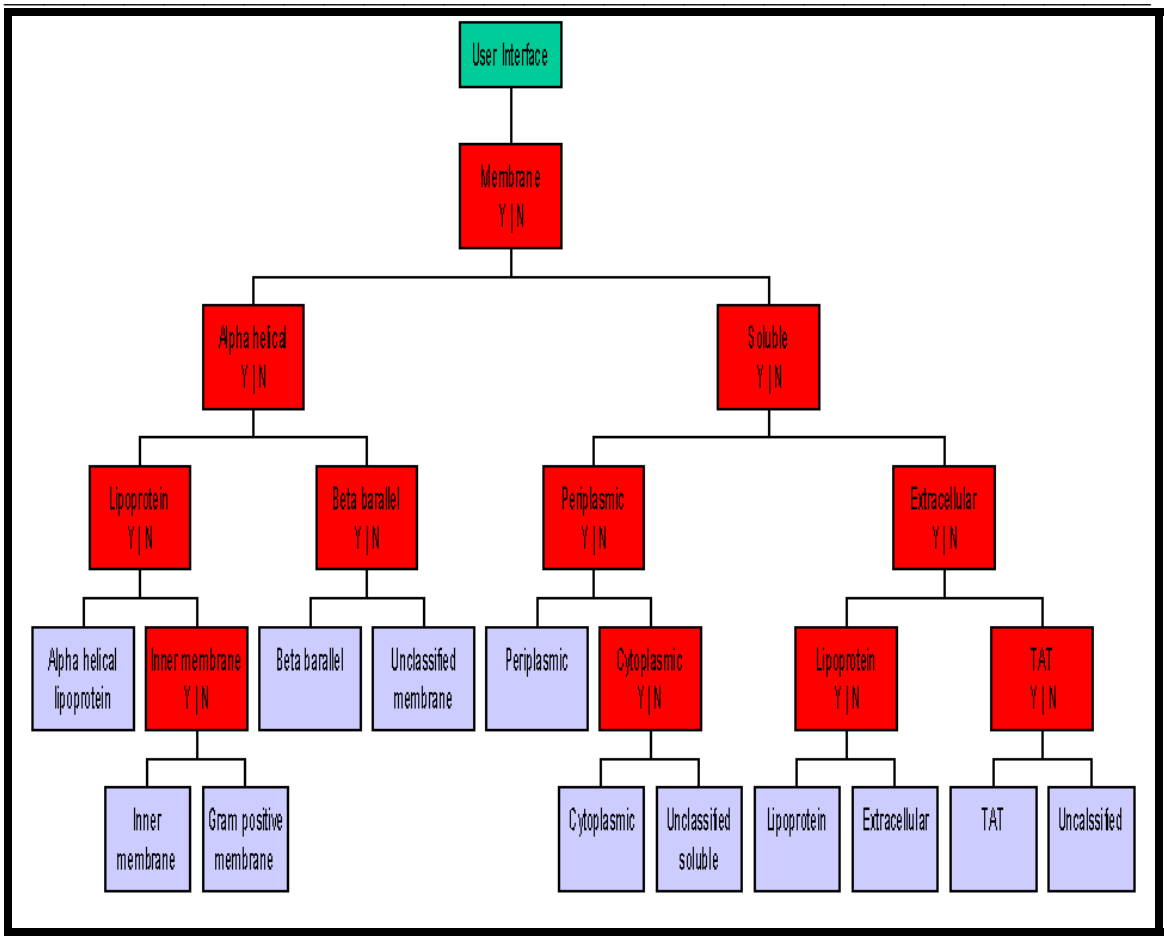

Figure 1: The membrane class predictor rooted serial pipeline. Red boxes represent algorithms while the purple boxes represent outcomes of the pipeline

The parallel pipeline predicted a higher number of protein locations than the serial pipeline for both the *N. meningitidis* and the *S. aureus* genomes. Most proteins from these genomes have no confirmed locations, thus it is impossible to asses the accuracy of all predictions. However, a partial assessment can be undertaken for those proteins with a confirmed location. 217 of the 2079 *N. meningitidis* MC58 proteins have an annotated location in SWISS-PROT. Of these, 149 were not confirmed leaving only 74 proteins with a definite location. The parallel pipeline correctly identified 36 of the 74; while of the 149 proteins with putative locations the parallel pipeline agreed with 84 SWISS-PROT annotations. The serial pipeline correctly identified 58 of the 74 proteins of confirmed location while the serial pipeline agreed with annotations of 78 of the 149 proteins of putative location. Unfortunately only two SWISS-PROT entries for *S. aureus* subsp. aureus MRSA252 protein exist and only one has a confirmed location. Both pipelines correctly identified this protein as cytoplasmic. A further assessment of the accuracies can be achieved when the prediction results are compared to the previous predictions by reverse-vaccinology for the genome, when seven proteins able to induce immunity were identified. The parallel predictor identified five, but the serial pipeline identified all seven. So for this measure at least, the serial pipeline may be more accurate.

The sensitivity of the parallel pipeline is enhanced because it considers the confidence level of each prediction. Moreover, the decision engine which drives ultimate prediction output is clearly important for the performance of the parallel pipeline. Consider the overall accuracy of the parallel method compared to that of its individual algorithms. The parallel pipeline outperforms the individual methods for three of the compartments, which are all of interest to vaccinologists: Gram-positive extracellular proteins (11.24% more accurate), Gram-negative extracellular proteins (3.58% more accurate) and outer membrane proteins (8.04% more accurate). The differences in accuracy for the other four compartments are slight, and range from 2.37% to 4.61%. Proteins from certain compartments have properties that resemble those characteristic of other compartments: for example, periplasmic and cytoplasmic proteins have similar

compositions and are easily confused by prediction methods. The combination method learns from such correlations, thus increasing the capacity of the network to determine the correct location of proteins despite their discombobulatingly similar amino acid compositions.

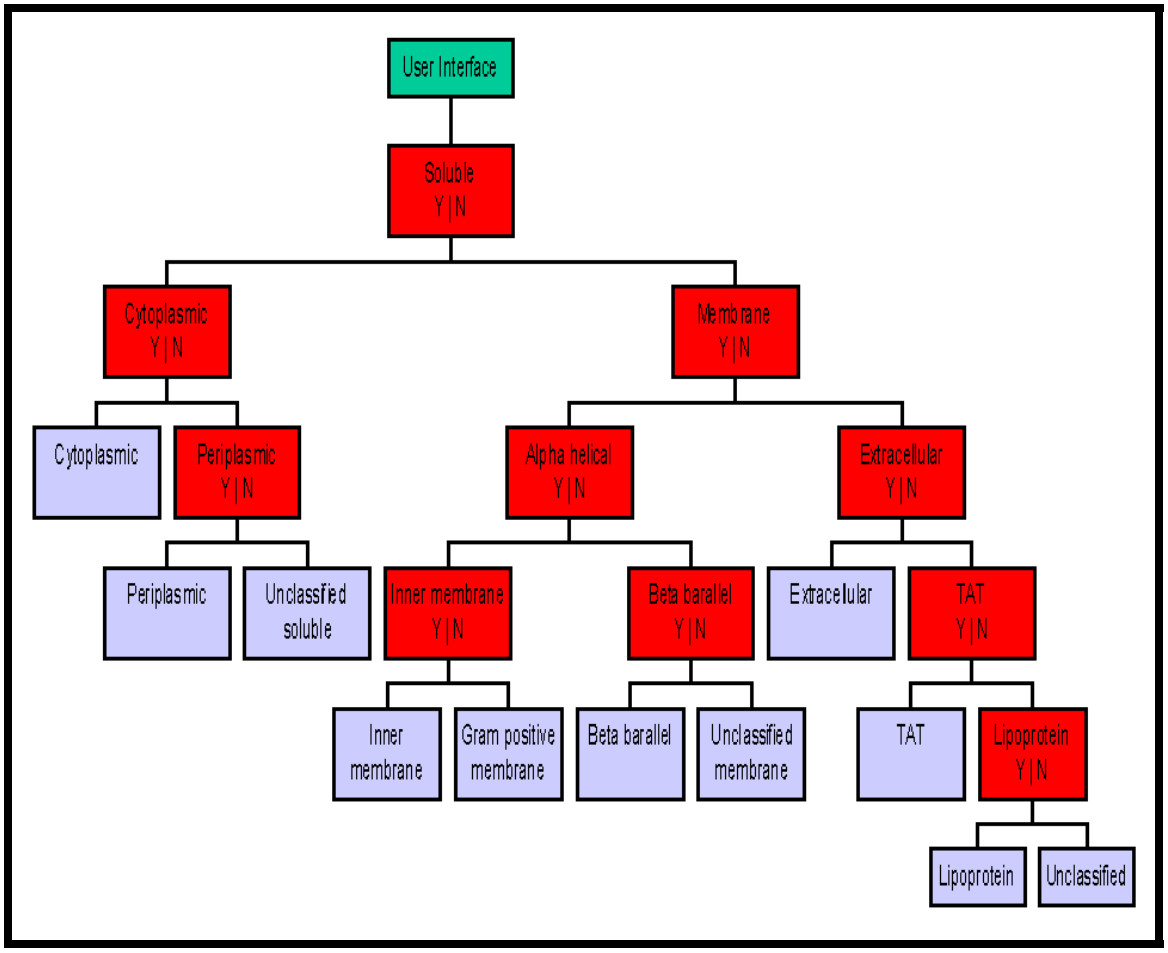

Figure 2: The soluble algorithm rooted serial pipeline. Red boxes represent algorithms while the purple boxes represent outcomes of the pipeline

Conclusion:

The principle purpose of *in silico* reverse-vaccinology is to identify potential vaccine targets (high sensitivity), but it is also important to reduce significantly the number of targets to be tested by successfully removing intracellular proteins (high specificity). The parallel pipeline outperformed the serial pipeline, but it took almost twice as long to execute. When assessed using genomic test sets, the relative performance of prediction was harder to quantify. Although many more predictions are made using the parallel pipeline the serial pipeline identifies 22 more of the 74 proteins of known location. As ever, because the test data available is not comprehensive the results we obtain, although immanently impressive, remain somewhat ambivalent and equivocal when viewed in a wider context. Thus, our study generally supports the view that a parallel implementation of subcellular location prediction is the more accurate as it utilises all available information. However, the serial implementations, which execute much more efficiency, are nonetheless potent predictors in their own right. Much work, in terms of both on-going testing and validation, is still required, yet either method should, ultimately, prove a powerful approach 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 erstwhile sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the UK Department of Health.

References:

[01] P. D. Taylor, *et al., Bioinformation,* 1:260 (2006)

[02] P. D. Taylor, *et al., Bioinformation,* 1:276 (2006)

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

[04] W. Li, *et al., Bioinformatics,* 17:282 (2001) [PMID: 11294794]

[05] W. Li, *et al., Bioinformatics,* 18:77 (2002) [PMID: 11836214]

[06] W. Li & A. Godzik, *Bioinformatics,* 22:1658 (2006) [PMID: 16731699]

[07] R. Chenna, *Nucleic Acids Res.,* 31:3497 (2005) [PMID: 12824352]

[08] P. D. Taylor, *et al., Bioinformation,* 1:176 (2006)

Edited by P. Kangueane

Citation: Taylor *et al.,* Bioinformation 1(8): 285-289 (2006)