Verma and Melcher *BMC Bioinformatics* 2012, **13**(Suppl 15):S9
http://www.biomedcentral.com/1471-2105/13/S15/S9

BMC Bioinformatics

PROCEEDINGS Open Access

A Support Vector Machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins

Ruchi Verma, Ulrich Melcher*

From Proceedings of the Ninth Annual MCBIOS Conference. Dealing with the Omics Data Deluge Oxford, MS, USA. 17-18 February 2012

Abstract

Background: Members of the phylum Proteobacteria are most prominent among bacteria causing plant diseases that result in a diminution of the quantity and quality of food produced by agriculture. To ameliorate these losses, there is a need to identify infections in early stages. Recent developments in next generation nucleic acid sequencing and mass spectrometry open the door to screening plants by the sequences of their macromolecules. Such an approach requires the ability to recognize the organismal origin of unknown DNA or peptide fragments. There are many ways to approach this problem but none have emerged as the best protocol. Here we attempt a systematic way to determine organismal origins of peptides by using a machine learning algorithm. The algorithm that we implement is a Support Vector Machine (SVM).

Result: The amino acid compositions of proteobacterial proteins were found to be different from those of plant proteins. We developed an SVM model based on amino acid and dipeptide compositions to distinguish between a proteobacterial protein and a plant protein. The amino acid composition (AAC) based SVM model had an accuracy of 92.44% with 0.85 Matthews correlation coefficient (MCC) while the dipeptide composition (DC) based SVM model had a maximum accuracy of 94.67% and 0.89 MCC. We also developed SVM models based on a hybrid approach (AAC and DC), which gave a maximum accuracy 94.86% and a 0.90 MCC. The models were tested on unseen or untrained datasets to assess their validity.

Conclusion: The results indicate that the SVM based on the AAC and DC hybrid approach can be used to distinguish proteobacterial from plant protein sequences.

Background

Bacterial plant pathogens are a major threat to global food security [1]. Half of the bacterial species causing major food losses in the world belong to the major phylum Proteobacteria (Figure 1). They are found predominantly in the class Gammaproteobacteria (*Xanthomonas, Pseudomonas* and *Erwinia*) and also in the class Betaproteobacteria (*Ralstonia*). Gammaproteobacteria include in addition a wide variety of several medically, ecologically and scientifically important groups such as Enterobacteriaceae (*Escherichia coli*), Vibrionaceae and Pseudomonadaceae. Also, beneficial bacteria, such as nitrogen fixing, ammonia oxidizing and iron fixing bacteria are members of this phylum. Betaproteobacteria also include ammonia oxidizing and arsenic resistant bacteria with Burkholderiales as one of the major classes. Alphaproteobacteria is dominated mostly by nitrogen-fixing bacteria and agrobacteria. Deltaproteobacteria and Epsilonproteobacteria have aerobic genera and curved to spirilloid *Wolinella* spp., respectively. Zetaproteobacteria is composed of a sole member: *Mariprofundus ferrooxydans* which oxidizes ferrous to ferric iron [2].

Several methods are being developed to detect phytopathogens involving macromolecular sequencing,

* Correspondence: ulrich.melcher@okstate.edu
Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078 USA

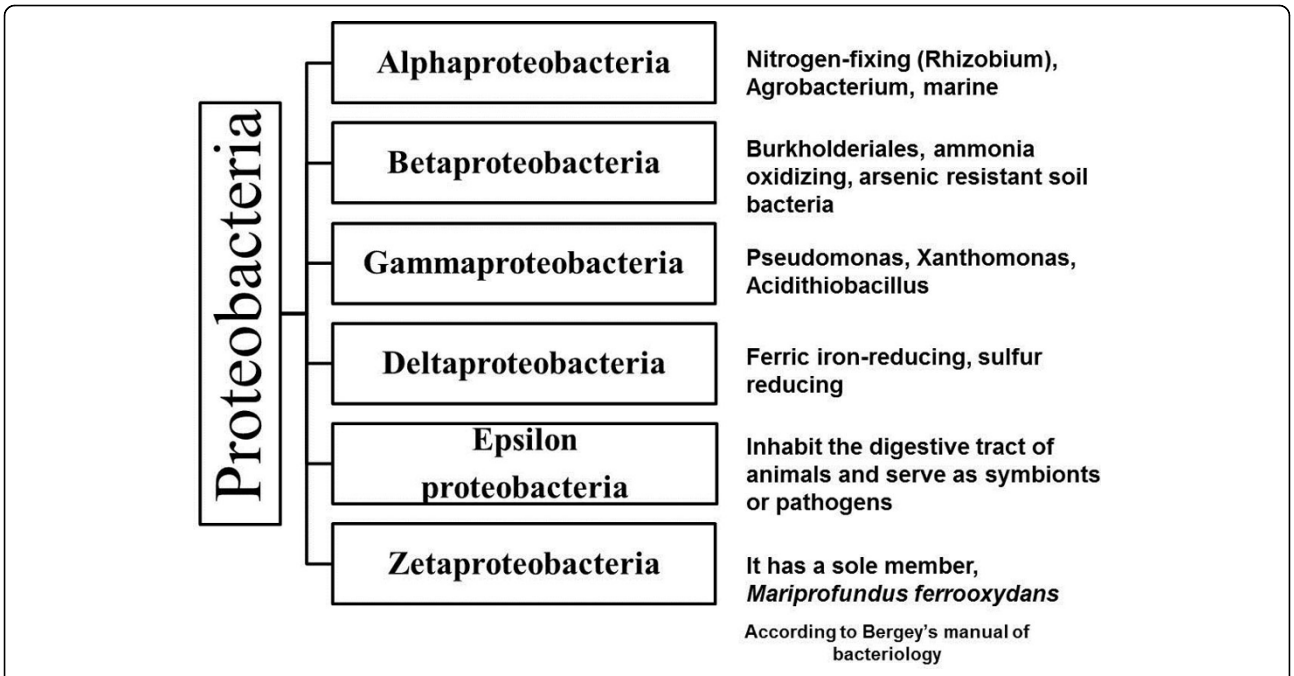


Figure 1 The subgroups of proteobacteria and the main members of each subgroup.

especially nucleotide sequencing [3,4]. With the advent of next generation sequencing, testing of diseased or quarantined plants for the presence of proteobacteria will rely increasingly on massive DNA sequencing. Peptide mass spectroscopy also shows promise in such screening. The analysis of nucleotide sequences typically involves assembly of sequence reads into contigs followed by analysis using Blast [5] search to identify pathogen-derived contigs. This approach is limited in that it only identifies potential pathogens whose nucleotide sequences are included in the searched database. Thus, there is a strong need for methods to find the organismal origin of unknown DNA or peptide fragments to identify potential pathogen sequences. Machine learning techniques, such as support vector machines (SVMs) and neural networks have been used successfully to develop classifiers for a number of different biolgocial problems including predicting different categories of proteins [6-14]. As a first step towards detecting pathogenic bacteria spp., we evaluated whether a machine learning algorithm, SVM, could distinguish between proteobacterial (potential pathogen) and plant (host) proteins. Thus, we assembled datasets of proteobacterial and plant host proteins for this study. We focused on amino acid, rather than nucleotide residues, because of the greater variety of residues that can be present at any one position, allowing subtle evolutionary forces to play a role in shaping the protein sequence and its properties.

Methods

Training datasets

Amino acid sequences of proteobacteria and plants were downloaded from the Uniprot website [UniProt release 2012_01-Jan 25, 2012] http://www.uniprot.org/. Only reviewed protein sequences were taken into consideration. A total of 3508 proteins (mean length, 322 ± 202) from nine species of proteobacteria (of which, three are phytopathogens) and 3206 proteins (mean length, 376 ± 308) from ten plant species were used initially for training. We used Blastclust [15] to remove redundant proteins, defined as those having greater than a specified % identity (a % redundancy value) from the data. Redundancy filtering was performed both before and after combining proteins from different species. Datasets were constructed at 90%, 50% and 30% redundancy values. Thus, with the 90% redundancy set we obtained 3408 proteobacterial and 2631 plant host proteins. For the 50% and 30% redundancy sets we obtained 3230 proteobacterial proteins, 2284 plant host proteins and 3203 proteobacterial proteins, 2277 plant host proteins, respectively. As the goal of this study was to identify bacterial proteins, the proteobacterial protein set was taken as the positive class and the plant protein set as the negative class (Tables 1 and 2). Test and training sets were designed from a five-fold cross-validation to create a model for the classification of new sequences (Figure 2). Thus each dataset was in both training and testing sets. To further validate the performance of our

Table 1 Total number of pathogen proteins taken from Uniprot and number of proteins remaining after redundancy filtering at 3 different percentages.

Positive dataset (Pathogen)	Total number of proteins (reviewed)	90% redundancy	50% redundancy	30% redundancy
Agrobacterium tumefaciens (Rhizobium radiobacter)	104	103	103	103
Burkholderia phymatum	333	333	333	333
Pseudomonas aeruginosa (ATCC)	1217	1216	1211	1211
Xanthomonas oryzae pv. Oryzae	411	410	410	410
Ralstonia solanacearum	601	601	599	599
Rhizobium etli (ATCC)	424	421	421	421
Rhizobium meliloti	48	47	47	47
Methylobacterium nodulans	213	213	213	213
Desulfobacterales autotrophicum (ATCC)	157	157	157	157
Total	3508	3501	3494	3494
Total after blastclust on cumulative data	-	3408	3230	3203

best-trained models, we tested the models on unseen/blind or untrained data not used for training the SVM. From Uniprot we downloaded non-redundant proteins for three species of proteobacteria (*Serratia marcesens, Acidovorax citrulli, Rhizobium fredii*) and three plant species (*Solanum lycopersicum, Phaseolus vulgaris, Cucurbita pepo*).

Feature Vectors used

Amino Acid Composition (AAC): Each protein was represented as a vector of 20 features, each corresponding to the fractional composition of an amino acid. This set of feature vectors was presented as input to SVM. Separate amino acid frequencies were calculated for both sets of proteins (proteobacteria and plants). The AAC was calculated by the following equation:

$$\text{Fraction of amino acid x} = \frac{\textit{Total number of amino acid x}}{\textit{Total number of amino acids in protein}}$$

where x can be any amino acid residue.

Dipeptide Composition (DC): Each protein was represented as a vector of 400 features for the 20 × 20 possible combinations of amino acids. The DC was calculated by the following equation:

$$\text{Fraction of dipeptide (xy)} = \frac{\textit{Total number of dipep xy}}{\textit{Total number of all possible dipeptides}}$$

where dipeptide (xy) is one of 400 possible dipeptides.

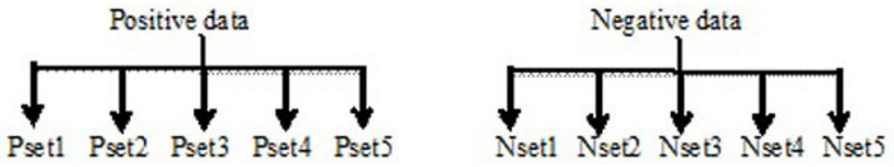
Hybrid (AAC+DC): The AAC and DC feature vectors were merged to yield feature vectors of 420 features (20 +400).

Support Vector Machine

An SVM is a kernel-based margin classifier, which uses both statistics and optimization. It draws an optimal hyper-plane in a high dimensional feature space that defines a boundary that maximizes the margin between data samples in two classes, therefore giving a better generalization property (Figure 3). Specifically, SVM^{light}, which is an implementation (in C language)

Table 2 Total number of plant proteins taken from Uniprot and number of proteins remaining after redundancy filtering at 3 different percentages.

Negative dataset (Plant host)	Total number of proteins (reviewed)	90% redundancy	50% redundancy	30% redundancy
Triticum aestivum	357	315	292	291
Oryza sativa	87	86	86	86
Solanum tuberosum	390	314	308	308
Arabidopsis thaliana	1000	968	857	852
Cucurbita maxima	26	25	25	25
Citrus sinensis	93	91	91	91
Vitis vinefera	161	154	152	152
Hordeum vulgare	348	323	307	307
Pisum sativum	371	347	335	334
Glycine max	373	346	324	324
Total	3206	2969	2777	2770
Total after blastclust on cumulative data	-	2631	2284	2277



Training set		Testing set	
Set1	Pset1 Nset1 Pset2 Nset2 Pset3 Nset3 Pset4 Nset4	Pset5 Nset5	Set1
Set2	Pset1 Nset1 Pset2 Nset2 Pset3 Nset3 Pset5 Nset5	Pset4 Nset4	Set2
Set3	Pset1 Nset1 Pset2 Nset2 Pset4 Nset4 Pset5 Nset5	Pset3 Nset3	Set3
Set4	Pset1 Nset1 Pset3 Nset3 Pset4 Nset4 Pset5 Nset5	Pset2 Nset2	Set4
Set5	Pset2 Nset2 Pset3 Nset3 Pset4 Nset4 Pset5 Nset5	Pset1 Nset1	Set5

Figure 2 Construction of datasets using five-fold cross validation. Pset is for positive dataset (proteobacteria) and Nset is for negative dataset (plants).

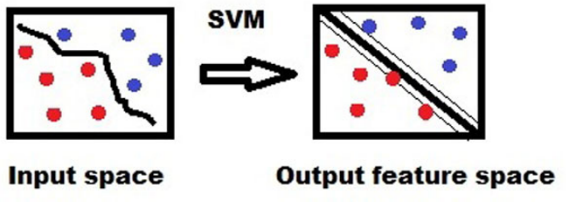


Figure 3 The concept of Support Vector Machine (SVM) in feature differentiation.

of SVM, has been used in this study. The SVM^{light} package can be downloaded from http://www.joachims.org for non-commercial or academic use [16]. In this study we used the SVM concept for the classification of proteobacteria and plant (host) proteins. Learning was carried out by using three kinds of kernels: the linear (t = 0), the polynomial (t = 1) and the Radial Basis Function (RBF) (t = 2). We obtained the best performance from the RBF.

Evaluation

Evaluation of the performance of the three models is threshold dependent. The performance of our method was computed by using the following standard parameters [17,18].

(a) Sensitivity or coverage of positive examples: percent of proteobacterial proteins correctly predicted

$$\text{Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

(b) Specificity or coverage of negative examples: percent of plant proteins correctly predicted as plant protein

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100$$

(c) Accuracy: percent of correctly predicted proteins (bacterial and plant proteins).

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{FN}} \times 100$$

(d) Matthews correlation coefficient (MCC) is considered to be the most robust parameter of any class prediction method [19]. MCC equal to 1 is regarded as perfect prediction while 0 suggests completely random prediction.

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP represents truly predicted proteobacterial proteins, and TN represents truly predicted plant proteins. FP and FN are falsely predicted proteobacterial and plant proteins, respectively.

Results and discussion

To test whether the AAC of proteobacterial and plant proteins differ significantly, we calculated AAC for both the proteobacterial (Table 1) and plant (Table 2) proteins datasets (Figure 4). We observe differences of AAC between proteobacteria and plants with respect to alanine, cysteine, glycine, lysine, arginine and serine. We also calculated the DC for these two datasets (figure not shown). We input the following vector sets for the SVM: AAC, DC and a hybrid of AAC and DC [20] models. We trained all three kernels (linear (Table 3), polynomial (Table 4) and RBF (Table 5) to identify the best-trained kernel. Comparison of the accuracies and MCCs obtained by all three kernels revealed that the RBF kernel performed best with all three redundancy percentages (Table 5). At 90% redundancy the SVM achieved a maximum accuracy of 92.44% and a 0.85 MCC for the AAC model [RBF parameters: g = 0.04, c = 4, j = 1], a maximum accuracy of 94.67% and 0.89 MCC for the DC model [g = 0.02, c = 6, j = 2] and for the hybrid model a maximum accuracy of 94.86% and a

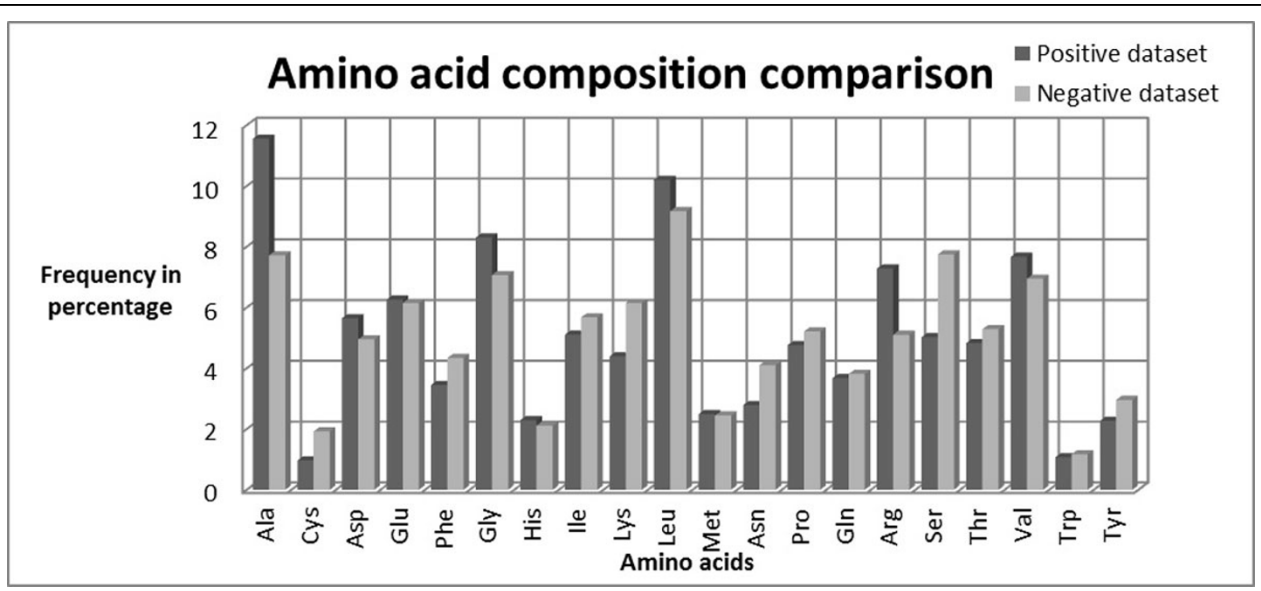


Figure 4 Comparative amino acid compositions of positive dataset (proteobacteria) and negative dataset (plants).

Table 3 Results of SVM models based on AAC, DC and hybrid (AAC+DC) features at three different redundancy percentages using the linear kernel (t = 0).

Redundancy(percentage)	Amino acid composition (AAC)		Dipeptide composition (DC)		Hybrid (AAC+DC)	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
30	87.87	0.75	90.74	0.81	89.10	0.78
50	87.45	0.74	90.87	0.81	91.81	0.83
90	87.45	0.74	90.87	0.81	89.33	0.78

0.90 MCC [g = 0.01, c = 8, j = 1]. At 50% redundancy, maximum accuracies for the AAC, DC and hybrid models were 91.62% (MCC 0.83) [g = 0.04, c = 2, j = 1], 94.12% (MCC 0.88) [g = 0.02, c = 4, j = 1] and 94.49% (MCC 0.89) [g = 0.01, c = 8, j = 2] respectively. At 30% redundancy, the maximum accuracies of the AAC, DC and hybrid models were 92.30% (MCC 0.84) [g = 0.05, c = 1, j = 1], 93.72% (MCC 0.87) [g = 0.03, c = 2, j = 2] and 93.84% (MCC 0.88) [g = 0.01, c = 4, j = 2]. As shown in Table 5 we achieved maximum accuracy with the hybrid model at 90% redundancy.

The result of the validation datatsets on six species (on all three models) are shown in Table 6. The hybrid model trained at 90% redundancy had the best accuracy only with exception of *Rhizobium fredii* for which the 50% redundant model was better. As can be seen from Table 5, the hybrid model at 90% redundancy performed best overall for most species. It is possible that the decrease in performance obtained by removing more proteins based on their similarities is not due to the identity value, but due to a resulting imbalance in the training datasets since the redundancy criteria affected proteobacterial protein numbers more strongly than they did the plant protein numbers. Because these estimates are sensitive to the threshold for distinguishing positives from negatives, we constructed a ROC curve to examine the model's accuracy. ROC has been used to show the accuracy of constructed models [21-29]. The ROC curve is a graphical representation of sensitivity (true positive rate) vs. one minus specificity (false positive rate or true negative rate) for any binary classifier system [30]. It is a threshold independent evaluation parameter and gives a value known as Area Under Curve (AUC) (Figure 5) which shows the performance of a classifier in a two class problem [31]. The higher the AUC, the more accurate the model. In the present study the AUC for hybrid model was 0.985 and therefore demonstrated the accuracy of the model.

This SVM model can be used to assign a query sequence as to whether it originated from a plant or proteobacterium, thus enabling timely detection of the infection. It may also be used to identify food contamination with bacteria by screening samples by sequencing. SVM models can be used to work in the area of animal proteins. As we have developed a model for plant and proteobacteria, another model can be designed for animal protein and pathogenic proteobacterial proteins. Thus, SVMs can be used in a variety of fields of study.

Conclusion

The SVM models based on the hybrid approach using both amino acid and dipeptide features exhibited the maximum accuracy on both threshold dependent and threshold independent parameters. Best results were obtained with an RBF kernel and considering protein sets that did not contain any proteins that are more than 90% identical to another protein in the dataset. SVMs have great potential to handle large datasets and thus can be used for sorting proteobacterial sequences from a mixed background, like those found in metagenomic sequence data. As such, an SVM classifier would be a step forward in surveillance techniques for bacteria that lack previously characterized relatives. It may be useful for determining protein sequences obtained from non-sequenced genomes not yet present in Genbank. Other features like domains specific to nitrogen oxidising or fixing bacteria can also be used even to distinguish a pathogenic proteobacterium from a non-pathogenic proteobacterium. This may be used to

Table 4 Results of SVM models based on AAC, DC and hybrid (AAC+DC) features dataset at three different redundancy percentages using polynomial kernel (t = 1)[d is another parameter used in this kernel and its value is given in parentheses].

Redundancy(percentage)	Amino acid composition (AAC)		Dipeptide composition (DC)		Hybrid (AAC+DC)	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
30	(d5) 89.63	0.78	(d2) 92.06	0.84	(d6) 91.29	0.82
50	(d4) 88.88	0.77	(d2) 91.98	0.83	(d4) 90.68	0.81
90	(d5) 89.46	0.78	(d2) 92.54	0.85	(d4) 91.26	0.82

Table 5 Results of SVM models based on AAC, DC and hybrid (AAC+DC) features at three different redundancy percentages using RBF kernel (t = 2)

Redundancy(percentage)	Amino acid composition (AAC)		Dipeptide composition (DC)		Hybrid (AAC+DC)	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
30	92.30	0.84	93.72	0.87	93.84	0.88
50	91.62	0.83	94.12	0.88	94.49	0.89
90	92.44	0.85	94.67	0.89	94.86	0.90

Table 6 Validation (accuracy) percentage by SVM models trained on AAC, DC and hybrid features.

AAC	Serratia marcescens (127)	Acidovorax citrulli (314)	Rhizobium fredii (16)	Solanum lycopersicum (413)	Phaseolus vulgaris (159)	Cucurbita pepo (15)
30	78.74	98.09	75	93.46	94.97	100
50	70.87	96.18	75	94.19	96.23	100
90	70.87	97.77	75	93.46	96.86	100
DC						
30	73.23	96.82	75	94.19	97.48	100
50	75.59	97.45	68.75	94.92	96.84	100
90	74.8	97.77	62.5	95.16	98.72	100
Hybrid (AAC +DC)						
30	74.8	96.5	75	95.4	98.74	100
50	79.53	99.04	81.25	93.46	97.48	100
90	81.1	99.04	79.53	93.95	98.11	100

The accuracy is calculated by dividing the number of correct predictions by total number of protein inputs. The numbers of reviewed proteins are shown in parentheses.

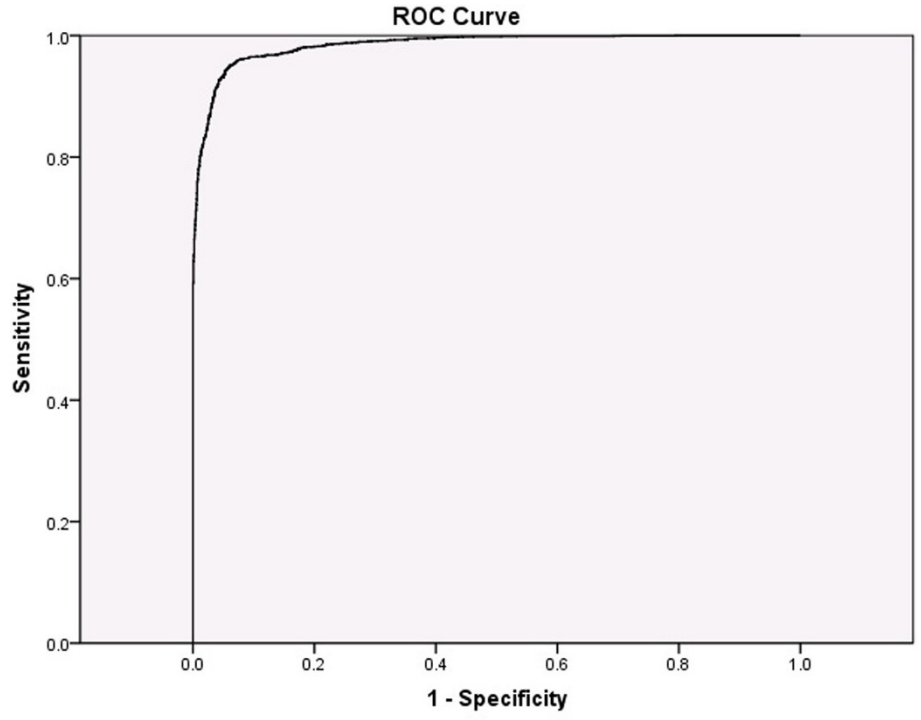


Figure 5 The ROC curve (Relative Operating Characteristic) and the area under curve for the best hybrid model at 90% redundancy.

determine the kinds of bacterial pathogens present in food samples thus improving food security. Human pathogens that are proteobacterial in nature also exist. Specific SVM models can be trained or designed to distinguish them. Thus SVMs hold greater potential for solving a variety of problems in biology.

List of abbreviations used

SVM: Support Vector Machine; AAC: Amino Acid Composition; DC: Dipeptide Composition; FAO: Food and Agricultural Organization; ROC: Receiver Operating Characteristic; AUC: Area Under The Curve; TP: True Positive; TN: True Negative; FN: False Negative; FP: False Positive.

Acknowledgements

The authors acknowledge the support from USDA-NIFA grant number 2010-85605-20542, and the Oklahoma Agricultural Experiment Station whose Director has approved the manuscript for publication. The authors thank William Schneider and Jacqueline Fletcher for helpful discussion. We also thank James Borrone and Rakesh Kaundal for reading of a draft manuscript. This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 15, 2012: Proceedings of the Ninth Annual MCBIOS Conference. Dealing with the Omics Data Deluge. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S15

Authors' contributions

RV designed the study and performed the machine learning analysis and drafted the manuscript. UM conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 11 September 2012

References

1. Strange RN, Scott PR: **Plant disease: a threat to global food security.** *Annual review of phytopathology* 2005, **43**:83-116.
2. Emerson D, Rentz JA, Lilburn TG, Davis RE, Aldrich H, Chan C, Moyer CL: **A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities.** *PloS one* 2007, **2(7)**:e667.
3. Melcher U, Grover V: **Genomic approaches to discovery of viral species diversity of non-cultivated plants.** In *Recent Advances in Plant Virology.* Norfolk UK: Caister Academic Press;Caranta C, Aranda MA, Tepfer M, López-Moya JJ 2011:321-342.
4. Fletcher J, Bender C, Budowle B, Cobb WT, Gold SE, Ishimaru CA, Luster D, Melcher U, Murch R, Scherm H, *et al*: **Plant Pathogen Forensics: Capabilities, Needs and Recommendations.** *MMBR* 2006, **70(2)**:450-471.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215(3)**:403-410.
6. Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GP: **Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles.** *BMC bioinformatics* 2008, **9**:201.
7. Kaundal R, Raghava GP: **RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information.** *Proteomics* 2009, **9(9)**:2324-2342.
8. Hu X, Wong KK, Young GS, Guo L, Wong ST: **Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma.** *Journal of magnetic resonance imaging: JMRI* 2011, **33(2)**:296-305.
9. Choi S, Jiang Z: **Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique.** *Computers in biology and medicine* 2010, **40(1)**:8-20.
10. Magnin B, Mesrob L, Kinkingnehun S, Pelegrini-Issac M, Colliot O, Sarazin M, Dubois B, Lehericy S, Benali H: **Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI.** *Neuroradiology* 2009, **51(2)**:73-83.
11. Vert JP: **Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings.** *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2002, 649-660.
12. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16(10)**:906-914.
13. Dharmasaroja P, Dharmasaroja PA: **Prediction of intracerebral hemorrhage following thrombolytic therapy for acute ischemic stroke using multiple artificial neural networks.** *Neurological research* 2012, **34(2)**:120-128.
14. Naguib IA, Darwish HW: **Support vector regression and artificial neural network models for stability indicating analysis of mebeverine hydrochloride and sulpiride mixtures in pharmaceutical preparation: a comparative study.** *Spectrochimica acta Part A, Molecular and biomolecular spectroscopy* 2012, **86**:515-526.
15. Dondoshansky IWY: *BLASTCLUST - BLAST score-based single-linkage clustering* 2000.
16. Joachims T: **Learning to classify text using support vector machines.** Boston: Kluwer Academic Publishers; 2002.
17. O'Dwyer L, Lamberton F, Bokde AL, Ewers M, Faluyi YO, Tanner C, Mazoyer B, O'Neill D, Bartley M, Collins DR, *et al*: **Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment.** *PloS one* 2012, **7(2)**:e32441.
18. Ansari HR, Raghava GP: **Identification of conformational B-cell Epitopes in an antigen from its primary sequence.** *Immunome research* 2010, **6**:6.
19. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16(5)**:412-424.
20. Verma R, Varshney GC, Raghava GP: **Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile.** *Amino acids* 2010, **39(1)**:101-110.
21. Lu Q, Cui Y, Ye C, Wei C, Elston RC: **Bagging optimal ROC curve method for predictive genetic tests, with an application for rheumatoid arthritis.** *Journal of biopharmaceutical statistics* 2010, **20(2)**:401-414.
22. He X, Frey E: **ROC, LROC, FROC, AFROC: an alphabet soup.** *Journal of the American College of Radiology: JACR* 2009, **6(9)**:652-655.
23. Chappell FM, Raab GM, Wardlaw JM: **When are summary ROC curves appropriate for diagnostic meta-analyses?** *Statistics in medicine* 2009, **28(21)**:2653-2668.
24. Algarabel S, Pitarque A: **ROC parameters in item and context recognition.** *Psicothema* 2007, **19(1)**:163-170.
25. Higashida Y, Ideguchi T, Muranaka T, Tabata N, Miyajima R, Akazawa F, Ikeda H, Morimoto K, Ohki M, Toyofuku F, *et al*: **[ROC analysis of detection of interval changes in interstitial lung diseases on digital chest radiographs using the temporal subtraction technique].** *Nihon Igaku Hoshasen Gakkai zasshi Nippon acta radiologica* 2004, **64(1)**:35-40.
26. Wiebringhaus R, John V, Muller RD, Hirche H, Voss M, Callies R: **[ROC analysis of image quality in digital luminescence radiography in comparison with current film-screen systems in mammography].** *Aktuelle Radiologie* 1995, **5(4)**:263-267.
27. Daures JP: **[Use of ROC curves in medical imaging].** *Journal de radiologie* 1991, **72(8-9)**:445-461.
28. Hannequin P, Liehn JC, Delisle MJ, Deltour G, Valeyre J: **ROC analysis in radioimmunoassay: an application to the interpretation of thyroglobulin measurement in the follow-up of thyroid carcinoma.** *European journal of nuclear medicine* 1987, **13(4)**:203-206.
29. Creelman CD, Donaldson W: **ROC curves for discrimination of linear extent.** *Journal of experimental psychology* 1968, **77(3)**:514-516.
30. Balakrishnan N: **Handbook of the logistic distribution.** New York: Dekker; 1992.
31. Zahr N, Arnaud L, Marquet P, Haroche J, Costedoat-Chalumeau N, Hulot JS, Funck-Brentano C, Piette JC, Amoura Z: **Mycophenolic acid area under the curve correlates with disease activity in lupus patients treated with mycophenolate mofetil.** *Arthritis and rheumatism* 2010, **62(7)**:2047-2054.

doi:10.1186/1471-2105-13-S15-S9
Cite this article as: Verma and Melcher: A Support Vector Machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins. *BMC Bioinformatics* 2012 **13**(Suppl 15):S9.