Predicting Protective Bacterial Antigens Using Random Forest Classifiers

Yasser EL-Manzalawy[2,3] Drena Dobbs[1] Vasant Honavar[2]
[1]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA
[2]Department of Computer Science, Iowa State University, Ames, IA 50011, Ames, IA, USA
[3]Department of Systems and Computer Engineering, Al-Azhar University, Cairo, Egypt
{yasser, ddobbs, honavar}@iastate.edu

ABSTRACT

Identifying protective antigens from bacterial pathogens is important for developing vaccines. Most computational methods for predicting protein antigenicity rely on sequence similarity between a query protein sequence and at least one known antigen. Such methods limit our ability to predict novel antigens (i.e., antigens that are not homologous to any known antigen). Therefore, there is an urgent need for alignment-free computational methods for reliable prediction of protective antigens.

We evaluated the discriminative power of four different amino acid composition derived feature representations using three classification methods (Logistic Regression, Support Vector Machine, and Random Forest) on a cross validation data set of 193 protective bacterial antigens and 193 non-antigenic bacterial proteins. Our results show that, with all four data representations, Random Forest classifiers consistently outperform other classifiers. We compared HRF50, one of the best performing Random Forest classifiers with VaxiJen and SignalP on independent test sets derived from the *Chlamydia trachomatis* and *Bartonella* proteomes. Our results show that our HRF50 predictor outperforms VaxiJen and is competitive with SignalP and ANTIGENpro in predicting protective antigens. We further showed that when we combine SignalP with HRF50, the resulting method, which we call BacGen, yields performance that is comparable to or better than that of ANTIGENpro in predicting antigens in bacterial sequences. We conclude that amino acid sequence composition derived features can be effectively used to design alignment-free methods for predicting protein antigenicity using Random Forest classifiers. BacGen is available as an online server at:http://ailab.cs.iastate.edu/bacgen/.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics

General Terms

Algorithms

Keywords

Computational immunology, protein antigenicity prediction

1. INTRODUCTION

Vaccination is one of the most cost-effective tools for preventing infectious diseases and minimizing their impact on the human populations [1]. Conventional vaccines have been developed by isolating or purifying antigenic components from target pathogens. Conventional approaches to vaccine design use live but weakened (i.e., attenuated) pathogens, whole killed pathogens, or a part of the target pathogen. Unfortunately, not all pathogens can be cultured, and insufficient manipulation of killed or attenuated pathogens can result in the contamination of the vaccine by virulent organisms.

Advances in whole-genome sequencing, high-throughput protein characterization, and bioinformatics have led to a new approach for vaccine development called Reverse Vaccinology (RV) [2, 3, 4, 5, 6]. RV starts with the entire proteome of a target pathogen and screens all the proteins to identify potential protective antigens to be tested in vivo and in vitro for their immunogenicity. The major advantages of the RV approach is its applicability to a broad range of pathogens and its efficiency in quickly finding vaccine targets [4]. Advances in RV rely almost entirely on the development of sufficiently accurate bioinformatics tools for predicting protective antigens [6].

Existing computational approaches for predicting protective antigens include [7]: i) subcellular location predictors, which attempt to identify potential protective antigens by predicting proteins exposed on the cell surface and hence accessible to neutralizing antibodies. An important limitation of this approach is the lack of validated, high-quality data sets for developing reliable predictors [7] ; ii) homology-based predictors that identify proteins in the target pathogen that share a high degree of sequence similarity with known protective antigens. An important limitation of this approach is the difficulty of predicting novel antigens that lack high sequence similarity with previously identified antigens [8]; iii) machine learning and statistical methods that train a classifier to discriminate between antigenic and non-antigenic proteins using sequence-derived features. Such methods are alignment-free and facilitate the identification of novel antigens.

VaxiJen [8] was the first method for predicting protective antigens using machine learning to train a classifier on amino acid sequence derived features. The method includes three separate predictors for predicting protective antigenic sequences in bacteria, viruses, and tumors. The performance of each predictor was estimated using a cross-validation data set of 100 antigen and 100 non-antigen sequences from the target category. To address the lack of experimental data for training VaxiJen classifiers, Magnan et al [9] proposed ANTIGENpro, a machine learning based method for predicting protein antigenicity trained using data curated from literature and high-throughput protein microarray data. Reported comparisons of ANTIGENpro with VaxiJen and SignalP [10], a program for predicting secreted proteins, on an independent test set of 1463 proteins from *Bartonella*, including 73 antigenic proteins and 1390 non-antigens, demonstrated superior performance of ANTIGENpro over VaxiJen, while the performance of SignalP was competitive with that of ANTIGENpro. An important difference between ANTIGENpro and VaxiJen is that ANTIGENpro is trained to discriminate between antigens and non-antigens while VaxiJen predictors are trained to discriminate between protective antigens and non-antigens. In fact, only a fraction of potential antigens are protective in the sense that they are specifically targeted by the acquired immune response of the host and are able to induce protection in the host against infectious and non-infectious diseases [11].

Against this background, we constructed a non-redundant data set of 193 experimentally verified protective bacterial antigens and 193 bacterial non-antigens (almost double the size of the VaxiJen bacterial data set) and used it to evaluate the performance of 32 distinct classifiers generated using all possible combinations of four classification methods, four protein sequence data representations, and a wavelet filter. One of the best performing classifiers, based on a Random Forest classifier combined with a wavelet filter for preprocessing the input data in the form of amino acid moment descriptors (AAMD) [12], was further evaluated for predicting antigenic proteins in *Chlamydia trachomatis* and *Bartonella henselae* proteomes. Our results confirm the findings of Magnan et al [9] that the SignalP program is competitive in performance with machine learning based methods for predicting protein antigenicity. We also show that an improvement in performance can be obtained by combining SignalP predictions with predictions from our Random Forest based predictor. Based on these results, we propose a method for predicting protective antigens in bacteria using Random Forest classifiers and a hybrid method for predicting antigens in bacterial sequences using consensus predictions of SignalP and Random Forest based predictor. Implementations of both methods are freely accessible at http://ailab.cs.iastate.edu/bacgen/.

2. METHODS

2.1 Data sets

Cross-validation data set. For our cross-validation evaluation tests, we used a data set of 193 protective bacterial antigens and 193 non-antigens constructed using the following procedure. First, 257 protective bacterial antigen sequences were downloaded from Protegen database [11] (as of September 2011). The sequence redundancy filtering step using BLASTCLUST [13] and 30% cutoff produced a final non-redundant set of 193 protective bacterial antigens. A set of 193 non-antigens were randomly selected from a pool of 144090 pathogenic bacterial proteins downloaded from UniRef such that no pair of sequences in the pool has >50% mutual sequence identity. A randomly selected protein was added to our non-antigenic set if it does not share more than 30% identity with any sequence in the antigenic set or with any sequence in the incrementally growing list of non-antigens.

Validation data sets. For comparison with existing methods for predicting antigenic proteins, we constructed two test sets from a set of 895 proteins corresponding to the complete *Chlamydia trachomatis* DUW-3CX proteome. The first data set, called *Chlamydia trachomatis* data set, consists of a set of 83 *Chlamydia trachomatis* antigens compiled by Finco et al [14] from several recent high-throughput studies [15, 16, 17, 18, 19] and served as positive data and the remaining 812 proteins were considered as negative data. The second data set, called balanced *Chlamydia trachomatis* data set, is a balanced version of the first data set in which the negative instances were reduced to only 83 proteins selected at random from the set of 812 non-antigens.

We also used an independent test set, *Bartonella* data set, which has been used previously for evaluating ANTIGENpro [9]. The data set consists of 1463 proteins from *Bartonella henselae* pathogen. Out of these 1463, 73 proteins are antigenic and the remaining 1390 proteins are considered as non-antigens. More information about this data set can be found in [9, 20].

2.2 Feature representation

Predicting antigenic proteins from amino acid sequence can be seen as a binary classification task in which a query protein sequence is to be classified into one of two categories: antigenic or non-antigenic. Such classifiers accept as input, a (typically fixed length) representation of the protein sequence and produce as output, a class label. *Amino acid composition* (AAC), the frequency of each amino acid type in a given protein sequence, is a widely used feature representation of amino acid sequences. AAC has been successfully used in several protein sequence classification tasks including protein subcellular localization prediction [21], and protein function [22] prediction. An inherent limitation of amino acid composition feature representation is that sequence order information is lost. To overcome this limitation, several studies (e.g., [23, 24, 25, 26]) have proposed the inclusion of additional features that capture sequence order information. In this study, we compared AAC with three amino acid sequence representations that utilize some sequence order information: *Dipeptide Composition* (DC); *Composition-Transition-Distribution* (CTD) [23]; and *Amino Acid Moment Descriptors* (AAMD) [12].

2.3 Wavelet transform

Wavelet transform is a widely used technique in signal processing and multi-resolution analysis [27]. A wavelet is a waveform of limited duration that has an average value of zero. Wavelet transform decomposes a signal into shifted and scaled versions of the original wavelet. Wavelet transform can be categorized into continuous wavelet transform (CWT) and discrete wavelet transform (DWT). CWT operates over every possible scale and shift parameter, whereas DWT uses a discretized scale and shift parameters [28].

Wavelet transform has found many applications [29]. including data and image compression [30, 31], pattern recognition [32], transient detection [33], texture analysis [34], and noise reduction [35], and more recently, in bioinformatics [36] e.g., genome sequence analysis [37], analysis of microarray data [38], and retrieval of protein structures from databases [39].

Haar wavelet transform (HWT) [40] is the simplest form of discrete wavelet transform. In this transform, given an input signal represented by a list of 2^n numbers, HWT simply pairs up input values, storing the difference and passing the sum. This process is repeated recursively, pairing up the sums to provide the next scale, finally resulting in $2n - 1$ differences and one final sum. HWT has been successfully used in several applications including digital image processing [41], feature extraction [42], solving non-linear integral and differential equations [43], and image and signal de-noising [44].

2.4 Classification methods

We experimented with three classification algorithms implemented in WEKA machine learning toolkit [45]: Logistic Regression (LR) [46]; Support Vector Machine (SVM) [47] and ; and Random Forest (RF) [48]. For the three classification methods, WEKA default parameters settings have been used unless stated otherwise.

Support vector machine (SVM) classifiers [47] have proven successful in several protein classification tasks (e.g., [23, 24, 25, 21]). Moreover, they have been successfully applied to the problem of predicting protein antigenicity [8, 9]. Optimizing the performance of SVM classifiers often requires tuning some parameters. To avoid over-optimistic performance estimates, the test data should not be used to guide the choice of the optimal parameters. Due to the relatively small size of our cross-validation data set, we found that using a subset of the training data as a validation set for determining the optimal parameters produces a final SVM model whose performance is lower than an SVM predictor trained using the default parameters. Therefore, we decided not to tune any of the SVM parameters (except setting ON the parameter that allows the classifier to return predicted probabilities instead of binary predictions).

Random forest (RF) classifiers [48] have been shown to outperform SVM classifiers on several tasks, including DNA-binding site prediction [49], conformational B-cell epitope prediction [50], gene expression profile classification [51], and prediction of RNA-protein interaction partners [52]. We used the WEKA default settings of the RF classifier, except for the number of trees which was set to 50 instead of 10.

2.5 Performance evaluation

We used five-fold cross-validation test to evaluate different classifiers developed in this study. The predictive performance of each classifier was assessed using prediction Accuracy (ACC), Sensitivity (S_n), Specificity (S_p), and Mathew Correlation coefficient (MCC) metrics defined as [53]:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$S_n = \frac{TP}{TP + FN} \; and \; S_p = \frac{TN}{TN + FP} \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \tag{3}$$

where TP, FP, TN, FN are the numbers of true positives, false positives, true negatives, and false negatives respectively.

In addition to these commonly used threshold-dependent metrics, we report area under the receiver operating characteristic curve (AUC). The receiver operating characteristic (ROC) curve is obtained by plotting the true positive rate as a function of the false positive rate as the discrimination threshold of the binary classifier is varied. An ideal classifier will have an AUC = 1, while a classifier assigning labels at random will have an AUC = 0.5, and any classifier performing better than random will have an AUC value that lies between these two extremes.

3. RESULTS AND DISCUSSION

3.1 Random Forest classifiers outperform SVM classifiers

Table 1 compares the performance of a Random Forest classifier with 50 iterations (RF50) with Logistic regression (LR) and Support Vector Machine classifiers using linear (SVML) or RBF (SVMRBF) kernels on four different representations of the cross-validation data set. Several interesting conclusions can be drawn from these results. First, based on AUC, LR performs best using AAC representation and the inclusion of amino acid sequence order information results in a drop in the LR classifier performance. Second, SVM classifiers are sensitive to the representation of the data (e.g., observed AUC for SVM classifiers range from 0.79 to 0.85) while RF50 classifiers have a consistent performance independent of the data representation (e.g., observed AUC values for RF50 are between 0.88 and 0.89). Third, for SVM classifier the best kernel is representation-dependent. For example, using DC representation, SVMRBF performs better than SVML while for the remaining three representations, a better performance is observed using linear kernel. Fourth, for all the classifiers except LR, the inclusion of extra features for encoding sequence order information yields only a slight improvement in classifier performance (if any). Finally, the best performing classifier (in terms of ACC, MCC, and AUC) is RF50 using AAMD representation.

The observation that RF50 classifiers outperform SVM classifiers might be attributed to the presence of some irrelevant features in the data. Another possibility is that a better classification can be obtained by assigning more weights to some features, i.e., a better classification can be obtained by assigning higher weights to certain types of amino acids (e.g., glycosylated amino acids). While SVM classifiers treat all features equally, tree classifiers like RF indirectly treat them unequally by discarding irrelevant attributes and using informative/discriminative features more frequently in the learned tree model.

3.2 Effect of filtering the data with Haar wavelet transform

We tested whether the application of a wavelet filter to the data can eliminate noisy signals and improve the predictive performance of the four classification algorithms considered in this study. Table 2 summarizes the performance of different classifiers on four data representations of amino acid

Table 1: Performance comparison of Logistic Regression (LR), Support Vector Machine with linear kernel (SVML), Support Vector Machine with RBF kernel (SVMRBF), and Random Forest with 50 trees (RF50) on the cross-validation data set using four different representations (See Methods for more details)

Representation	Classifier	ACC	Sn	Sp	MCC	AUC
AAC	LR	0.75	0.77	0.74	0.51	0.83
	SVML	0.74	0.74	0.73	0.48	0.83
	SVMRBF	0.73	0.74	0.71	0.45	0.81
	RF50	0.79	0.79	0.80	0.59	0.89
DC	LR	0.67	0.76	0.58	0.35	0.66
	SVML	0.74	0.79	0.70	0.49	0.79
	SVMRBF	0.75	0.76	0.74	0.49	0.84
	RF50	0.80	0.75	0.85	0.60	0.89
CTD	LR	0.72	0.76	0.68	0.44	0.78
	SVML	0.77	0.76	0.78	0.54	0.85
	SVMRBF	0.73	0.72	0.75	0.47	0.82
	RF50	0.79	0.75	0.83	0.58	0.88
AAMD	LR	0.74	0.77	0.71	0.48	0.80
	SVML	0.75	0.75	0.76	0.51	0.84
	SVMRBF	0.76	0.78	0.74	0.52	0.81
	RF50	0.80	0.80	0.81	0.61	0.89

sequences filtered with Haar wavelet transform. For almost all the classifiers no change in AUC values is observed. However, for some classifiers (e.g, RF50 using AAMD representation) a slight improvement in ACC or MCC was observed. Results in Table 1 and Table 2 suggest that several RF50 classifiers could effectively serve as our final model for predicting potential protective bacterial antigens from amino acid sequence. We decided to use RF50 using Haar wavelet transformed AAMD representation of the data as our final machine learning based predictor. RF50 consistently has the highest AUC valuse using the filtered and unfiltered representations of the data. The highest observed MCC of 0.61 was obtained using both filtered and unfiltered AAMD representation of the data and the highest ACC of 0.81 was noted for RF50 classifier using the filtered AAMD data representation.

3.3 Comparison with existing servers

To the best of our knowledge, there are only two machine learning based methods for predicting protective antigens from bacterial pathogens based on protein sequence: VaxiJen [8] and ANTIGENpro [9].ANTIGENpro, which was originally developed to predict protein antigenicity has shown surprisingly good performance on the related but different task of predicting protective antigens [9]. Implementations of the two methods as Web servers are freely available online. Unfortunately, ANTIGENpro server restricts submissions to only one protein sequence per submission, making the application of ANTIGENpro on a genome-wide scale impractical. Therefore, we were unable to directly compare our server with ANTIGENpro on *Chlamydia trachomatis* test sets. Hence, we provide an indirect comparison with ANTIGENpro server here via a comparison with SignalP [10], which has been reported to yield performance that is competitive with that of ANTIGENpro [9]. In addition, we provide a direct comparison of our method with ANTIGENpro on an independent test set, *Bartonella* data set [9], that has previously been used for comparing ANTIGENpro with VaxiJen and SignalP servers [9].

Table 3 compares the performance of VaxiJen v2 server, SignalP, HRF50, and BacGen (a consensus prediction of SignalP and HRF50) on the balanced *Chlamydia trachomatis* data set. The performance of the HRF50 is competitive with that of SignalP and both HRF50 and SignalP outperform VaxiJen. This is consistent with the results in [9] which show that the performance of SignalP is competitive with that of ANTIEGNPro, and that both methods outperform VaxiJen on the *Bartonella* data set. Our result confirm that tools for identifying secreted proteins (e.g., SignalP) can be used for predicting antigens in pathogenic bacterial genomes. Our results also demonstrate that SignalP can complement machine learning methods for predicting antigens as shown by the superior performance of our proposed method, BacGen, which combines predictions of SignalP with that of HRF50.

We noticed that the performance of HRF50 classifier on a balanced version of the *Chlamydia trachomatis* data set is worse than its estimated performance on the cross-validation data set. Similar observation holds in the case of VaxiJen [8]. To understand this discrepancy, it is important to note that both HRF50 and VaxiJen classifiers are trained to discriminate between protective antigens and non-antigens, whereas proteins in the balanced *Chlamydia trachomatis* data set are either antigens (an unknown fraction of which is expected to be protective) or non-antigens. Although it might appear to be unfair to test VaxiJen and HRF50 on predicting protein antigenicity when they are in fact designed to predict a subset of antigens (protective antigens), this comparison yields two interesting findings: First, in predicting antigenicity, HRF50 outperforms VaxiJen iand has performance that is competitive with that of SignalP, which has been previously shown to be competitive with the ANTIGENpro, a server designed to predict protein antigenicity. Second, combining SignalP with HRF50 predictions results in improved performance in predicting protein antigenicity.

The performance of different antigen prediction tools reported in Table 3 is likely to be over-optimistic due to the evaluation on balanced data. To obtain a more realistic performance estimate, we evaluated the four methods on the entire *Chlamydia trachomatis* test set (See Table 4). In this data set, only 9% of the proteins are antigens and the re-

Table 2: Performance comparison of Logistic Regression (LR), Support Vector Machine with linear kernel (SVML), Support Vector Machine with RBF kernel (SVMRBF), and Random Forest with 50 trees (RF50) on the cross-validation data set using four different representations and Haar wavelet filter as a pre-processing step.

Representation	Classifier	ACC	Sn	Sp	MCC	AUC
AAC	LR	0.75	0.77	0.74	0.51	0.83
	SVML	0.74	0.74	0.73	0.47	0.83
	SVMRBF	0.71	0.70	0.72	0.41	0.78
	RF50	0.79	0.77	0.81	0.58	0.88
DC	LR	0.67	0.75	0.59	0.35	0.66
	SVML	0.75	0.74	0.75	0.49	0.80
	SVMRBF	0.74	0.75	0.73	0.48	0.84
	RF50	0.79	0.74	0.84	0.59	0.88
CTD	LR	0.72	0.76	0.68	0.44	0.78
	SVML	0.77	0.76	0.78	0.55	0.84
	SVMRBF	0.73	0.71	0.75	0.46	0.82
	RF50	0.80	0.76	0.84	0.60	0.89
AAMD	LR	0.74	0.77	0.71	0.48	0.80
	SVML	0.75	0.76	0.74	0.51	0.84
	SVMRBF	0.75	0.75	0.74	0.50	0.81
	RF50	0.81	0.80	0.82	0.61	0.88

maining proteins are considered as non-antigens. For such unbalanced data, the use of traditional performance measures can be problematic. For instance, a classifier that predicts every query protein as non-antigen will have 91% accuracy on the unbalanced *Chlamydia trachomatis* test set. Therefore, we decided to follow the approach used in [9] for evaluating ANTIGENpro on unbalanced test set derived from sequences of *Bartonella* genome. First, all query protein sequences were submitted to the predictor to be assigned a predicted score (e.g., probability that the query protein is an antigen). Second, all proteins were ranked in a descending order by their predicted scores. Third, the enrichment of antigens among top k% proteins (k = 2, 5, 10, 15, 20, 25) was computed as (% of antigens among top ranked subset)/(% of antigens in the entire data set). The expected enrichment of a random ranking is 1 and higher values correspond to better performing classifiers [9]. In addition, we compared the predictors performance using the percentage of true positives (%TP) in the top ranked subset calculated as (number of positives among top ranked subset)/(number of positives among the entire data set). Table 4 shows the percentage of true positives and the enrichment estimates for VaxiJen, SignalP, HRF50, and BacGen from the ranked top 2%, 5%, 10%, 15%, 20%, and 25% subsets of *Chlamydia trachomatis* entire proteome sequences. In general, SignalP has higher %TP and enrichment estimates than VaxiJen and HRF50, and the combination of HRF50 and SignalP, BacGen method, leads to improvements in performance.

Table 5 compares our server with VaxiJen, SignalP, and ANTIGENpro servers using *Bartonella* data set. The results show that combining HRF50 predictions with SignalP predictions, BacGen method, provides a computational method for predicting protein antigenicity from amino acid sequence that is highly competitive with ANTIGENpro. An important advantage of BacGen over ANTIGENpro is that the latter relies on some sequence extracted information and predictions obtained by applying four programs to the query protein: i) SSpro [54] for predicting protein secondary structure; ii) DOMpro [55] for predicting the number of domains in a protein, ; iii) ACCPro [54] for predicting relative solvent accessibility; iv) TMHMM [56] for predicting the number of transmembrane helices (TMH) and the expected number of residues in TMHs. On the other hand, BacGen relies only on sequence compositional features and SignalP predictions. Therefore, BacGen is more easily applicable for large scale data sets. Also, BacGen server allows users to submit multiple proteins while ANTIGENpro limits submissions to a single protein at a time.

Table 3: Performance comparison of different predictors on predicting antigenic proteins in the balanced *Chlamydia trachomatis* test set

Method	ACC	Sn	Sp	MCC	AUC
VaxiJen	0.54	0.39	0.68	0.08	0.56
SignalP	0.63	0.35	0.91	0.32	0.72
HRF50	0.63	0.77	0.49	0.27	0.72
BacGen	0.66	0.38	0.94	0.38	0.78

3.4 BacGen Server

Implementations of HRF50 classifier and SignalP program [10] are provided as an online Web server which is freely accessible at http://ailab.cs.iastate.edu/bacgen. The server accepts as input one or more query protein sequences in FASTA format and returns predictions using either HRF50, SignalP, or BacGen (consensus predictions of HRF50 and SignalP). If SignalP predictions are requested, then the user will be prompted to specify whether the query proteins belong to gram+ or gram- bacteria. SignalP and BacGen predict antigenic proteins in bacterial pathogens while HRF50 predicts protective antigens in bacterial proteins. The output page associates each query protein with the predicted probability that the protein is an antigen or protective antigen.

Table 4: Performance comparison of different predictors on predicting antigenic proteins in the unbalanced *Chlamydia trachomatis* test set

Method	Top(%)	TP(%)	Enrichment
VaxiJen	2	2	1.2
SignalP	2	9	4.3
HRF50	2	11	5.5
BacGen	2	11	5.5
VaxiJen	5	10	2.0
SignalP	5	17	3.4
HRF50	5	16	3.2
BacGen	5	20	3.9
VaxiJen	10	17	1.7
SignalP	10	30	3.0
HRF50	10	21	2.1
BacGen	10	28	2.8
VaxiJen	15	26	1.7
SignalP	15	37	2.4
HRF50	15	28	1.9
BacGen	15	37	2.4
VaxiJen	20	37	1.8
SignalP	20	43	2.1
HRF50	20	35	1.8
BacGen	20	44	2.2
VaxiJen	25	40	1.6
SignalP	25	50	2.0
HRF50	25	44	1.8
BacGen	25	45	1.8

Table 5: Performance comparison of different predictors on predicting antigenic proteins in the unbalanced *Bartonella* test set

Method	Top(%)	Enrichment
VaxiJen	2	2.1
ANTIGENpro	2	5.5
HRF50	2	0.7
SignlP	2	4.8
BacGen	2	5.5
VaxiJen	5	1.6
ANTIGENpro	5	4.4
HRF50	5	1.4
SignlP	5	2.7
BacGen	5	5.2
VaxiJen	10	1.9
ANTIGENpro	10	3.4
HRF50	10	1.1
SignlP	10	2.1
BacGen	10	3.4
VaxiJen	25	1.6
ANTIGENpro	25	2.1
HRF50	25	0.9
SignlP	25	2.0
BacGen	25	2.0

4. CONCLUSION

Predicting antigens or protective antigens from a pool of protein sequences (i.e., entire set of proteins encoded by a pathogen genome) is a challenging problem [7]. Computational methods for reliably predicting antigenic proteins from amino acid sequence can dramatically expedite the identification of vaccine candidates and can contribute to development of diagnostic tests. Amino acid composition features have been shown to be effective for various protein classification tasks. In this work, we systematically evaluated amino acid composition features and three approaches for capturing sequence-order information for developing protein antigenicity classifiers based on Logistic Regression, Support Vector Machine, and Random Forest classification algorithms. Our results showed that good performance (AUC = 0.89 and ACC = 79% on a non-redundant data set of 193 protective antigens and 193 non-antigens using 5-fold cross-validation test) can be reached using only amino acid composition features and Random Forest classification. Slight improvements were noted using extra features for modeling sequence order information and filtering the input features using Haar wavelet transformation. Comparisons of one of the best performing classifiers considered in this study, HRF50, with VaxiJen and SignalP on independent test sets derived from the *Chlamydia trachomatis* and *Bartonella henselae* proteomes, respectively, showed that SignalP is highly competitive with machine learning based methods for predicting antigens, but a better performance is observed when SignalP and HRF50 predictions are combined. Based on these findings, we propose alignment-free methods for two important classification tasks: i) Predicting protective antigens in bacterial sequences using HRF50, a Random Forest classifier trained using Haar wavelet transformed features of amino acid compositions and amino acid moment descriptors [12]; ii) Predicting antigens in bacterial sequences using BacGen, a method combining HRF50 and SignalP predictions. Implementations of our methods are freely available as an online Web server at http://ailab.cs.iastate.edu/bacgen.

Despite the acceptable performance (AUC close to 0.9) of many classifiers trained using machine learning on the cross-validation data set of protective antigens used in this study, the performance of HRF50, VaxiJen, SignalP, and BacGen in identifying antigens in *Chlamydia trachomatis* proteome is far from satisfactory: If we treat the sequences ranked among top 15% with respect to the score assigned by the classifier in the case of the *Chlamydia trachomatis* proteome as predicted antigens, only 37% of the antigens reported in high-throughput studies of *Chlamydia trachomatis* antigenicity profiles are predicted to be antigens. This suggests that discriminating protective antigens from non-antigens may be much easier than discriminating antigens from non-antigens. A similar observation has been reported in [57], where we showed that classifiers trained to predict protective linear B-cell epitopes have better predictive performance than classifiers trained to predict linear B-cell epitopes. We conjecture that protective antigens can be discriminated from antigens on the basis of some sequence features. Our work in progress is aimed at identifying useful features for discriminating protective antigens from antigens. Such an analysis might help improve our understanding of what makes an antigen protective and lead to the development of improved methods for identifying protective antigens.

5. ACKNOWLEDGMENTS

This work was funded in part by the National Institutes of Health grant GM066387 to Vasant Honavar and Drena Dobbs and in part by a research assistantship funded by the Center for Computational Intelligence, Learning, and Discovery. The authors sincerely thank Pierre Baldi for inspiring discussions and for sharing Bartonella data set. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] S. Bambini, R. Rappuoli, The use of genomics in microbial vaccine development, Drug Discovery Today 14 (5-6) (2009) 252–260.

[2] R. Rappuoli, Reverse vaccinology, Current Opinion in Microbiology 3 (5) (2000) 445–450.

[3] J. Adu-Bobie, B. Capecchi, D. Serruto, R. Rappuoli, M. Pizza, Two years into reverse vaccinology, Vaccine 21 (7-8) (2003) 605–610.

[4] R. Rappuoli, A. Aderem, A 2020 vision for vaccines against HIV, tuberculosis and malaria, Nature 473 (7348) (2011) 463–469.

[5] A. Sette, R. Rappuoli, Reverse vaccinology: developing vaccines in the era of genomics, Immunity 33 (4) (2010) 530–541.

[6] D. Jones, Reverse vaccinology on the cusp, Nature Reviews Drug Discovery 11 (3) (2012) 175–176.

[7] D. Flower, I. Macdonald, K. Ramakrishnan, M. Davies, I. Doytchinova, Computer aided selection of candidate vaccine antigens, Immunome Research 6 (2010) 1–16.

[8] I. Doytchinova, D. Flower, VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines, BMC Bioinformatics 8 (1) (2007) 4.

[9] C. Magnan, M. Zeller, M. Kayala, A. Vigil, A. Randall, P. Felgner, P. Baldi, High-throughput prediction of protein antigenicity using protein microarray data, Bioinformatics 26 (23) (2010) 2936–2943.

[10] J. Dyrløv Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: Signalp 3.0, Journal of Molecular Biology 340 (4) (2004) 783–795.

[11] B. Yang, S. Sayers, Z. Xiang, Y. He, Protegen: a web-based protective antigen database and analysis system, Nucleic Acids Research 39 (suppl 1) (2011) D1073–D1078.

[12] J. Shi, S. Zhang, Y. Liang, Q. Pan, Prediction of protein subcellular localizations using moment descriptors and support vector machine, Pattern Recognition in Bioinformatics (2006) 105–114.

[13] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389–3402.

[14] O. Finco, E. Frigimelica, F. Buricchi, R. Petracca, G. Galli, E. Faenzi, E. Meoni, A. Bonci, M. Agnusdei, F. Nardelli, et al., Approach to discover t-and b-cell antigens of intracellular pathogens applied to the design of chlamydia trachomatis vaccines, Proceedings of the National Academy of Sciences 108 (24) (2011) 9969–9974.

[15] F. Follmann, A. Olsen, K. Jensen, P. Hansen, P. Andersen, M. Theisen, Antigenic profiling of a chlamydia trachomatis gene-expression library, Journal of Infectious Diseases 197 (6) (2008) 897–905.

[16] R. Coler, A. Bhatia, J. Maisonneuve, P. Probst, B. Barth, P. Ovendale, H. Fang, M. Alderson, Y. Lobet, J. Cohen, et al., Identification and characterization of novel recombinant vaccine antigens for immunization against genital chlamydia trachomatis, FEMS Immunology & Medical Microbiology 55 (2) (2009) 258–270.

[17] D. Molina, S. Pal, M. Kayala, A. Teng, P. Kim, P. Baldi, P. Felgner, X. Liang, L. De la Maza, Identification of immunodominant antigens of chlamydia trachomatis using proteome microarrays, Vaccine 28 (17) (2010) 3014–3024.

[18] J. Wang, L. Chen, F. Chen, X. Zhang, Y. Zhang, J. Baseman, S. Perdue, I. Yeh, R. Shain, M. Holland, et al., A chlamydial type iii-secreted effector protein (tarp) is predominantly recognized by antibodies from humans infected with chlamydia trachomatis and induces protective immunity against upper genital tract pathologies in mice, Vaccine 27 (22) (2009) 2967–2980.

[19] J. Sharma, Y. Zhong, F. Dong, J. Piper, G. Wang, G. Zhong, Profiling of human antibody responses to chlamydia trachomatis urogenital tract infection using microplates arrayed with 156 chlamydial fusion proteins, Infection and Immunity 74 (3) (2006) 1490–1499.

[20] A. Vigil, R. Ortega, A. Jain, R. Nakajima-Sasaki, X. Tan, B. Chomel, R. Kasten, J. Koehler, P. Felgner, Identification of the feline humoral immune response to Bartonella henselae infection by protein microarray, PloS One 5 (7) (2010) e11447.

[21] K. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, Bioinformatics 19 (13) (2003) 1656–1663.

[22] C. Cai, W. Wang, L. Sun, Y. Chen, Protein function classification via support vector machine approach, Mathematical Biosciences 185 (2) (2003) 111–122.

[23] C. Cai, L. Han, Z. Ji, X. Chen, Y. Chen, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Research 31 (13) (2003) 3692–3697.

[24] K. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins: Structure, Function, and Bioinformatics 43 (3) (2001) 246–255.

[25] Z. Feng, C. Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, Journal of Protein Chemistry 19 (4) (2000) 269–275.

[26] R. Sokal, B. Thomson, Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population, American Journal of Physical Anthropology 129 (1) (2006) 121–131.

[27] C. Chui, An introduction to wavelets, Vol. 1, Academic Pr, 1992.

[28] D. Lee, A. Yamamoto, Wavelet analysis: Theory and applications, Hewlett-Packard Journal (1994) 44–52.

[29] J. Goswami, A. Chan, Fundamentals of wavelets: theory, algorithms, and applications, Vol. 219, Wiley, 2011.

[30] E. Hamid, Z. Kawasaki, Wavelet-based data compression of power system disturbances using the minimum description length criterion, IEEE Transactions on Power Delivery 17 (2) (2002) 460–466.

[31] R. DeVore, B. Jawerth, B. Lucier, Image compression through wavelet transform coding, IEEE Transactions on Information Theory 38 (2) (1992) 719–746.

[32] Y. Tang, Wavelet theory and its application to pattern recognition, Vol. 36, World Scientific Pub Co Inc, 2000.

[33] M. Riera-Guasp, J. Antonino-Daviu, M. Pineda-Sanchez, R. Puche-Panadero, J. Perez-Cruz, A general approach for the transient detection of slip-dependent fault components based on the discrete wavelet transform, IEEE Transactions on Industrial Electronics 55 (12) (2008) 4167–4180.

[34] T. Chang, C. Kuo, Texture analysis and classification with tree-structured wavelet transform, IEEE Transactions on Image Processing 2 (4) (1993) 429–441.

[35] M. Lang, H. Guo, J. Odegard, C. Burrus, R. Wells Jr, Noise reduction using an undecimated discrete wavelet transform, IEEE Signal Processing Letters 3 (1) (1996) 10–12.

[36] P. Lio, Wavelets in bioinformatics and computational biology: state of art and perspectives, Bioinformatics 19 (1) (2003) 2–9.

[37] A. Elloumi Oueslati, Z. Lachiri, N. Ellouze, Detecting particular features in c. elegans genomes using synchronous analysis based on wavelet transform, International Journal of Bioinformatics Research and Applications 7 (2) (2011) 183–201.

[38] G. Bidaut, F. Manion, C. Garcia, M. Ochs, WaveRead: automatic measurement of relative gene expression levels from microarrays using wavelet analysis, Journal of Biomedical Informatics 39 (4) (2006) 379–388.

[39] Z. Aung, K. Tan, Rapid retrieval of protein structures from databases, Drug Discovery Today 12 (17-18) (2007) 732–739.

[40] A. Haar, Zur theorie der orthogonalen funktionensysteme, Mathematische Annalen 69 (3) (1910) 331–371.

[41] P. Porwik, A. Lisowska, The Haar-wavelet transform in digital image processing: its status and achievements, Machine Graphics and Vision 13 (2004) 79–98.

[42] C. Papageorgiou, T. Poggio, A trainable system for object detection, International Journal of Computer Vision 38 (1) (2000) 15–33.

[43] Ü. Lepik, Application of the haar wavelet transform to solving integral and differential equations, Proceedings of the Estonian Academy of Sciences. Physics, Mathmatics 56 (1) (2007) 28–46.

[44] F. Luisier, C. Vonesch, T. Blu, M. Unser, Fast Haar-wavelet denoising of multidimensional fluorescence microscopy data, in: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging: From Nano to Macro, IEEE Press, 2009, pp. 310–313.

[45] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. Witten, L. Trigg, Weka-a machine learning workbench for data mining, Data Mining and Knowledge Discovery Handbook (2010) 1269–1277.

[46] S. Le Cessie, J. Van Houwelingen, Ridge estimators in logistic regression, Applied Statistics (1992) 191–201.

[47] V. Vapnik, The nature of statistical learning theory, Springer-Verlag New York Inc, 2000.

[48] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[49] J. Wu, H. Liu, X. Duan, Y. Ding, H. Wu, Y. Bai, X. Sun, Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature, Bioinformatics 25 (1) (2009) 30–35.

[50] W. Zhang, Y. Xiong, M. Zhao, H. Zou, X. Ye, J. Liu, Prediction of conformational b-cell epitopes from 3d structures by random forests with a distance-based feature, BMC bioinformatics 12 (1) (2011) 341.

[51] K. Moorthy, M. Mohamad, Random forest for gene selection and microarray data classification, Bioinformation 7 (3) (2011) 142.

[52] U. Muppirala, V. Honavar, D. Dobbs, Predicting rna-protein interactions using only sequence information, BMC Bioinformatics 12 (1) (2011) 489.

[53] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[54] J. Cheng, A. Randall, M. Sweredoski, P. Baldi, SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Research 33 (suppl 2) (2005) W72–W76.

[55] J. Cheng, M. Sweredoski, P. Baldi, DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks, Data Mining and Knowledge Discovery 13 (1) (2006) 1–10.

[56] A. Krogh, B. Larsson, G. Von Heijne, E. Sonnhammer, Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1, Journal of Molecular Biology 305 (3) (2001) 567–580.

[57] Y. EL-Manzalawy, D. Dobbs, V. Honavar, Predicting protective linear B-cell epitopes using evolutionary information, in: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, 2008, pp. 289–292.