Vaccine 29 (2011) 8156–8164

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Improving reverse vaccinology with a machine learning approach

Brett N. Bowman[a], Paul R. McAdam[b,c,1], Sandro Vivona[d], Jin X. Zhang[b], Tiffany Luong[b], Richard K. Belew[e], Harpal Sahota[b,c], Donald Guiney[f], Faramarz Valafar[a], Joshua Fierer[b,f,g], Christopher H. Woelk[b,f,*]

[a] Bioinformatics and Medical Informatics, San Diego State University, San Diego, CA 92182, USA
[b] Veterans Affairs San Diego Healthcare System, San Diego, CA 92161, USA
[c] Department of Biology, University of York, York YO10 5YW, UK
[d] Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA 94305, USA
[e] Department of Cognitive Science, University of California San Diego, La Jolla, CA 92093, USA
[f] Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[g] Department of Pathology, University of California San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 9 April 2011
Received in revised form 19 July 2011
Accepted 28 July 2011
Available online 22 August 2011

Keywords:
Reverse vaccinology
Bacterial pathogens
Protective antigen
Support vector machines

ABSTRACT

Reverse vaccinology aims to accelerate subunit vaccine design by rapidly predicting which proteins in a pathogenic bacterial proteome are putative protective antigens. Support vector machine classification is a machine learning approach that has been applied to solve numerous classification problems in biological sciences but has not previously been incorporated into a reverse vaccinology approach. A training data set of 136 bacterial protective antigens paired with 136 non-antigens was constructed and bioinformatic tools were used to annotate this data for predicted protein features, many of which are associated with antigenicity (*i.e.* extracellular localization, signal peptides and B-cell epitopes). Annotation was used to train support vector machine classifiers that exhibited a maximum accuracy of 92% for discriminating protective antigens from non-antigens as assessed by a leave-tenth-out cross-validation approach. These accuracies were superior to those achieved when annotating training data with auto and cross covariance transformations of *z*-descriptors for hydrophobicity, molecular size and polarity, or when classification was performed using regression methods. To further validate support vector machine classifiers, they were used to rank all the proteins in six bacterial proteomes for their antigenicity. Protective antigens from the training data were significantly recalled (enriched) in the top 75 ranked proteins for all six proteomes as assessed by a Fisher's exact test ($p < 0.05$). This paper describes a superior workflow for performing reverse vaccinology studies and provides a benchmark training data set that can be used to evaluate future methodological improvements.

Published by Elsevier Ltd.

1. Introduction

Reverse vaccinology (RV) has emerged over the past decade as a method that uses bioinformatic algorithms to identify putative protective antigens in bacterial proteomes that may represent potential vaccine candidates [1,2]. This emergence has been driven by the rapid accumulation of whole genome sequencing data from over 4000 bacteria [3], which are used to identify the protein coding genes that constitute the raw material for RV approaches. The main advantage of RV over conventional vaccinology is the speed and reduced cost with which potential vaccine candidates can be identified since there is no requirement for culturing bacteria [4]. In addition, RV can identify all the putative protective protein antigens for a bacterial species and not just the most abundant antigens isolated from bacterial cultures via a conventional approach.

Vaccination is one of the most important and cost-effective methods of preventing infectious diseases and no other method has had a similar impact in reducing morbidity and mortality [5]. The use of recombinant proteins as immunogens has been described as the most promising approach to meet the demands of vaccinology in the future [6]. RV has been used to limit the number of vaccine candidates that are validated using immunogenicity assays in animal models prior to vaccine formulation and clinical trials [7,8]. The advantages of subunit vaccines include increased safety, reduced cost, less antigenic competition, and direct targeting to the site of infection [9], when compared to live vaccines, which suffer from problems associated with attenuation and reversion

* Corresponding author at: University of California San Diego, Department of Medicine, Stein Clinical Research Building, Rm. 326, 9500 Gilman Drive, #0679, La Jolla, CA 92093-0679, USA. Tel.: +1 858 552 8585x7193; fax: +1 858 552 7445.
E-mail address: cwoelk@ucsd.edu (C.H. Woelk).
[1] Current address: Laboratory for Bacterial Evolution and Pathogenesis, Centre for Infectious Diseases and The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, Scotland, UK.

0264-410X/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.vaccine.2011.07.142

to virulence [10]. Subunit vaccine delivery strategies compatible with antigens predicted by RV include: non-pathogenic vectors, non-living antigen delivery systems, and DNA vaccines.

Previous RV studies fall into the two main categories of filtering [7,11–20] or classifying [21]. Pizza et al. [7] completed the first published RV study in order to develop a vaccine against serogroup B *Neisseria meningitidis*. Bioinformatic tools predictive of subcellular localization (*e.g.* PSORT) and the annotation derived from homology searches were used in a filtering approach to identify 570 open reading frames (ORFs) that coded for potential surface associated proteins. Proteins were successfully cloned and expressed from 350 ORFs and used to generate mouse immune sera. The immune sera were used to evaluate protein expression on the bacterial surface and bactericidal activity. Five of the most promising antigens that conferred immunity across a broad range of *Neisseria* strains and serogroups were combined to form the rMenB vaccine that has recently completed a phase II clinical trial [8].

The New Enhanced Reverse Vaccinology Environment (NERVE) provided the first user-friendly software package for performing RV studies [20]. The filters used by NERVE to identify potential antigens include: (1) the probability the protein is extracellular or an adhesin using PSORTb and SPAAN, respectively [22,23], (2) the ease with which the protein can be cloned, expressed and purified for experimental validation (*i.e.* $\leq$2 transmembrane domains predicted by HMMTOP) [24], (3) the likelihood of inducing an autoimmune response, (4) the ability to confer protection across a number of bacterial strains (*i.e.* conservation), and (5) homology with known antigens. NERVE could potentially be used to rank the proteins in a bacterial proteome for their antigenicity. However, at default settings NERVE discards proteins (*i.e.* filtering) and only ranks proteins based on their conservation across bacterial species. NERVE was used to examine the recall of known antigens in ranked bacterial proteomes but only for those antigens that were non-cytoplasmic with $\leq$2 transmembrane domains and no statistical test was applied to evaluate the significance of recall.

Doytchinova and Flower [21] developed a bacterial antigen classifier based on regression methods. Training data consisted of 100 known antigens from the literature record and the same number of randomly selected non-antigens from matched bacterial proteomes. This training data was annotated using auto and cross covariance (ACC) transformations of *z*-descriptors for hydrophobicity, molecular size and polarity [25]. The *z*-descriptors are amino acid descriptors obtained by applying principal component analysis to groups of physiochemical variables [25]. These descriptors have been used to model the relationship between peptide function and amino acid composition [26,27], and for predicting the activity of β-lactam antibiotics [28]. The annotation derived from ACC transformations of *z*-descriptors was used to train an antigen classifier to discriminate antigens from non-antigens based on two-class discriminant analysis using partial least squares (DA-PLS). A leave-one-out cross-validation (LOOCV) procedure demonstrated that this classifier could distinguish antigens from non-antigens with 82% accuracy and this method can be implemented using the "VaxiJen" server (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html).

Support vector machine (SVM) methods have been applied to a wide range of classification problems in biology including the discrimination of catalytic residues from noncatalytic residues [29], epitope prediction [30], the identification of translation initiation sites [31], and the diagnosis of disease states using microarray gene expression data [32]. SVMs are capable of non-linear modeling and are less prone to over-fitting (or over-training) compared to DA-PLS [33]. Furthermore, annotation derived from the protein annotation tools, such as those employed by NERVE, may provide a more biologically relevant data source than ACC transformations of *z*-descriptors. Therefore, we hypothesized that increased accuracies would be attained from antigen classifiers constructed using SVMs rather than DA-PLS, when trained on annotation derived from an expanded set of protein annotation tools instead of ACC transformations of *z*-descriptors. This was indeed the case and SVM classifiers were further validated by assessing the significance with which they recalled known antigens from the training data when in the background of a whole bacterial proteome.

2. Materials and methods

2.1. Training data

Two sources of positive training data (*i.e.* antigens) were used in order to train classifiers. The VaxiJen data set was previously published by Doytchinova and Flower [21] and contains 100 antigens for which details are available as supplementary information to their publication. The VaxiJen data set contains 74 bacterial protective antigens (BPAs) that led to significant protection ($p < 0.05$) in an animal model (*i.e.* bacterial load reduction or survival assay) following immunization and subsequent challenge with the bacterial pathogen. These 74 BPAs were combined with a further 62 BPAs curated from the literature record to derive a second training data set referred to as the BPA data set. Protein sequence data for 36 of the 74 BPAs from the VaxiJen data set were substituted in the BPA data set in order to better reflect the bacterial strain used for cloning and expression, and thus immunization in animal models. Further descriptive annotation for the 136 BPAs is available as supplemental information (Supplemental Table 1) as well as their protein sequence data in FASTA format (Supplemental File 1).

For both the VaxiJen and BPA data sets, non-antigen (negative training) data was generated to match the antigen (positive training) data using enhancements to the method described by Doytchinova and Flower [21]. Briefly, for each antigen, a non-antigen was randomly selected from the proteome of the same bacterial strain. If a selected non-antigen had homology (*i.e.* blastp *E*-value $\leq$10E−3) with a non-antigen already in the negative training data then it was discarded and resampled in order to increase diversity. In contrast to Doytchinova and Flower [21], non-antigens that matched with high identity (>98%) to antigens in the positive training data were also discarded and resampled in order to prevent the inclusion of known antigens in the negative training data. This resulted in balanced training data sets of antigens and non-antigens. Five sets of non-antigen training data were created in this manner for both the VaxiJen and BPA data sets in order to assess the effects of randomly selecting negative training data. Protein sequence data in FASTA format is provided in the supplemental information for the 5 non-antigen data sets (Supplemental Files 2 to 6).

2.2. Data annotation

Proteins in the training data sets (VaxiJen and BPA) and bacterial proteomes were annotated using either protein annotation tools (Supplementary Table 2) or in an identical manner to Doytchinova and Flower [21], using ACC transformations of *z*-descriptors. In more detail, proteins were annotated using output from 19 different bioinformatics tools that were available through web interfaces or standalone packages. The output of these 19 protein annotation tools was parsed to derive a total of 122 annotation features. In general, when multiple sites in a protein were given a score for a particular annotation feature then the following variables were calculated: number of predicted sites (Count), number of predicted sites normalized for protein length (CorrCount), maximum scoring site (MaxScore), and average score over all predicted sites (AvgScore). Hydrophobicity, molecular size and polarity are

reflected in *z*-descriptors [25], which were used to annotate every amino acid in a protein sequence. Each protein annotated in this manner was then subjected to ACC transformations resulting in a uniform vector of 45 annotation features per protein [34]. Briefly, ACC scores were calculated between pairs of amino acids separated by between 1 and 5 positions and then averaged over the entire protein length for all possible combinations of *z*-descriptors ($3^2 = 9$) to give a total of 45 annotation features.

2.3. Machine learning and regression classifiers

Classifiers were constructed to distinguish antigens from non-antigens in training data sets using SVM, DA-PLS and simpler linear regression methods. SVM methods were implemented using the LIBSVM toolkit [35], whereas DA-PLS and linear regression models were constructed using the Weka data-mining software package [36]. The annotation derived from training data sets was scaled between −1 and 1 using the normalization function in Weka prior to classifier construction. An in depth description of SVM methodology is beyond the scope of the methods presented here but a full description is provided in the work of Cristianini and Shawe-Taylor [37]. To summarize, SVM classification uses a kernel function to project the input vectors (representing annotation features) into an abstract, high dimensional feature space so that the two classes of these vectors (representing antigens versus non-antigens) can be clearly separated by a hyperplane. As with all inductive classification techniques, a key issue is to find a generalizable pattern across the training data instead of "overfitting" to its particulars. SVMs accomplish this by effectively searching through potential data projections so as to maximize the "margin" separating the antigen versus non-antigen classes. This margin is defined as the distance between the hyperplane and the nearest points of each class in feature space. A simple radial basis function (RBF) was selected for the kernel function because it can discriminate non-linear relationships between input vectors and class labels using only a couple of parameters. These parameters include γ and *C*, which control the width of the RBF kernel and the trade off between expanded classification margins and misclassification error, respectively. A grid search was used to examine the effect of combinations of different values for γ and *C* parameters on cross-validated accuracy and the optimal parameters were selected as $\gamma = 4.88\text{E}{-4}$ and $C = 3.28\text{E}{+4}$.

Classification accuracy was assessed using a leave-tenth-out cross-validation (LTOCV) procedure. Briefly, a random tenth of the training data was removed and the ability of the SVM classifier to correctly predict the class (*i.e.* antigen versus non-antigen) of proteins in this excluded tenth was assessed. These predictions were used to populate a 2×2 contingency table for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and the whole process repeated until every tenth had been excluded at least once. The 2×2 contingency table was used to calculate classification accuracy (TP + TN/TP + TN + FP + FN). For all classifiers, the minimum number of annotation features that led to maximum classification accuracy was identified by feature selection using a greedy backward elimination algorithm. Greedy backward elimination aims to remove the least informative feature from the classifier but this process often identifies features with tied information content. In order to evaluate the effect of randomly breaking such ties during feature selection the greedy backward elimination algorithm was iterated 10 times.

2.4. Known antigen recall

The ability of antigen classifiers to identify known antigens from the positive training data when in a background of a whole bacterial proteome was assessed. Classifiers were used to rank the proteins in a bacterial proteome based on their antigenicity. For SVM classifiers this ranking was based on the distance of each protein from the hyperplane and for DA-PLS classifiers this ranking was derived from regression statistics. Matches of known antigens from training data sets to proteins in bacterial proteomes were identified using a blastp *E*-value cut-off of 1E−40. A one-tailed Fisher's exact test was used to determine whether known antigens were enriched above a number of different cut offs (i.e. 25, 50, 75 and 100) in the ranked proteome. This test determines whether the number of known antigens in the top 100 ranked proteins, for example, was significantly enriched with respect to all of the known antigens in a background of the entire bacterial proteome. A one-tailed Fisher's exact test is based on a hypergeometric probability function and was implemented in the R statistical package [38] using the *dhyper* test.

2.5. Gene ontology analysis

The top 100 proteins ranked for antigenicity by the SVM classifier trained on the BPA data set in the proteome of *Streptococcus pneumoniae* (strain TIGR4) were subjected to gene ontology (GO) analysis. Initially, Blast2GO [39] was used to annotate the *S. pneumoniae* proteome for GO terms and those terms that were enriched for proteins in the top 100 with a false discovery rate (FDR) corrected *p*-value <0.05 were identified using BiNGO [40]. GO annotation was available for 74 of the 100 top ranked proteins of which 59 could be mapped to GO terms related to biological processes.

3. Results

3.1. Improvements to antigen training data

The VaxiJen data set was previously published by Doytchinova and Flower [21] and consists of 100 antigens (positive training data) paired with non-antigens (negative training data) from the same species. However, this data set contains protein sequence data from organisms that are not bacterial in origin (Supplemental Fig. 1) and antigens that were not significantly protective in animal models. Therefore, a BPA data set was constructed consisting solely of antigens ($N = 136$) from bacterial species. A BPA is defined as an antigen that when used for immunization studies in animal models led to significant protection following challenge with the bacterial pathogen from which it was derived. The BPA data set contains a larger number of protective antigens from a greater diversity of bacterial species compared to the VaxiJen data set (Supplemental Fig. 1). In the following text, for simplicity, the term "antigen" is used to describe proteins in the positive training data, however, for the BPA data set these antigens are strictly BPAs. Non-antigen negative training data was generated for both the BPA and VaxiJen data sets using a similar approach to Doytchinova and Flower's [21] by randomly selecting a protein from the same proteome for each antigen but with a modification to ensure that no antigens themselves were selected.

3.2. Annotation of training data with bioinformatic protein annotation tools

Nineteen different bioinformatic tools for annotating protein sequence data for biological features were identified and their output parsed to derive 122 annotation features (Supplemental Table 2). The first task was to determine if an antigen classifier could more accurately discriminate antigens from non-antigens when trained on annotation from these protein annotation tools or when trained on the 45 annotation features generated from ACC transformations of *z*-descriptors [21]. To this end, a DA-PLS classifier was trained on the VaxiJen data set annotated using both protein annotation tools and ACC transformations of *z*-descriptors. Plots of classification accuracies derived from DA-PLS classifiers

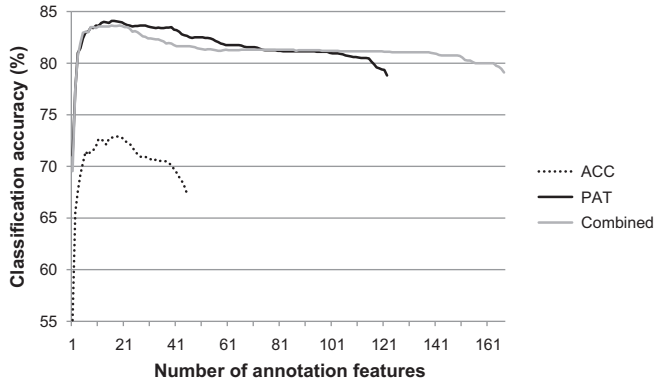

Fig. 1. The ability of a DA-PLS classifier to discriminate 100 antigens from 100 non-antigens in the VaxiJen data set using annotation generated from ACC transformations of *z*-descriptors ($N = 45$), protein annotation tools (PAT, $N = 122$), or a combination of both annotations (Combined, $N = 167$). Classification accuracy was assessed for DA-PLS classifiers consisting of different numbers of annotation features selected by greedy backward elimination using a LTOCV approach.

consisting of different numbers of annotation features selected by greedy backward elimination indicated that higher accuracies were attained using annotation derived from the protein annotation tools, and combining this annotation with ACC transformations of *z*-descriptors showed no improvement to classifier accuracy (Fig. 1). Therefore, only annotation derived from protein annotation tools was selected when assessing the ability of different types of classifier (SVM, DA-PLS and linear regression) to discriminate antigens from non-antigens.

3.3. Antigens are best distinguished from non-antigens using support vector machine classifiers

The ability of three different classification methods (SVM, DA-PLS and linear regression) to discriminate antigens from non-antigens in both the VaxiJen and BPA data sets when annotated with protein annotation tools was assessed. The main aim was to determine if greater classification accuracies were achieved with more sophisticated SVM methods and with the well-characterized BPA data set containing sequence data solely from bacterial pathogens. Fig. 2 clearly indicates that SVM classifiers achieved higher classification accuracies than other classification methods. A negligible difference was observed in classification accuracies when the SVM classifier was trained on either the BPA (SVM-BPA) or the VaxiJen (SVM-VaxiJen) data set. Therefore, SVM classifiers trained on the BPA training data annotated with protein annotation tools were further investigated.

Each line plotted in Fig. 2 represents the average of five individual analyses derived from pairing antigen data with different non-antigen data sets ($N = 5$). Therefore, it was decided to examine the variation in classification accuracies associated with different non-antigen data sets for the SVM-BPA classifier. Minimal differences in classification accuracies were observed when using different non-antigen data sets and all 5 classifiers resulted in peak accuracies greater than 86% (Fig. 3A). Furthermore, 10 iterations were performed for each of the 5 pairings of antigen and non-antigen data in order to assess the impact of randomly breaking ties during feature selection by greedy backward elimination. The best performing pairing in Fig. 3A was "Data set 3" and was selected in order to visualize the variation in classification accuracies associated with the 10 iterations. Again, minimal differences in classification accuracies were observed with different iterations, all of which resulted in classifiers with peak accuracies greater than 87% (Fig. 3B).

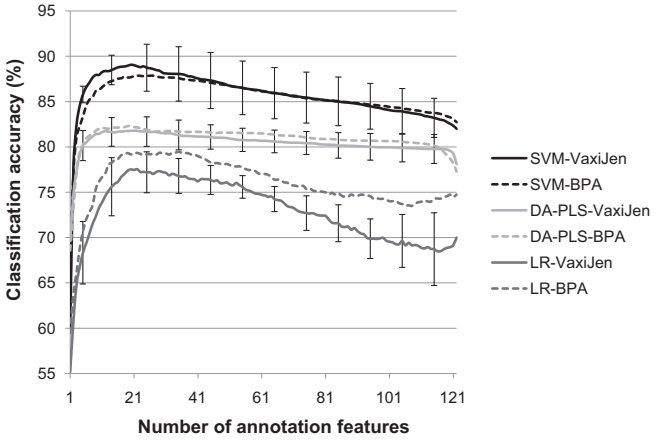

Fig. 2. The ability of SVM, DA-PLS and linear regression (LR) classifiers to discriminate antigens from non-antigens in the VaxiJen and BPA data sets annotated using protein annotation tools. Classification accuracy was assessed for classifiers comprised of different numbers of annotation features selected by greedy backward elimination using a LTOCV approach. Each line represents an averaging of classification accuracies from 5 different analyses resulting from the pairing of antigen data with 5 different non-antigen data sets. Error bars were calculated as 95% confidence intervals derived from the standard error of the mean and displayed at intervals of 10 features for the VaxiJen data set.

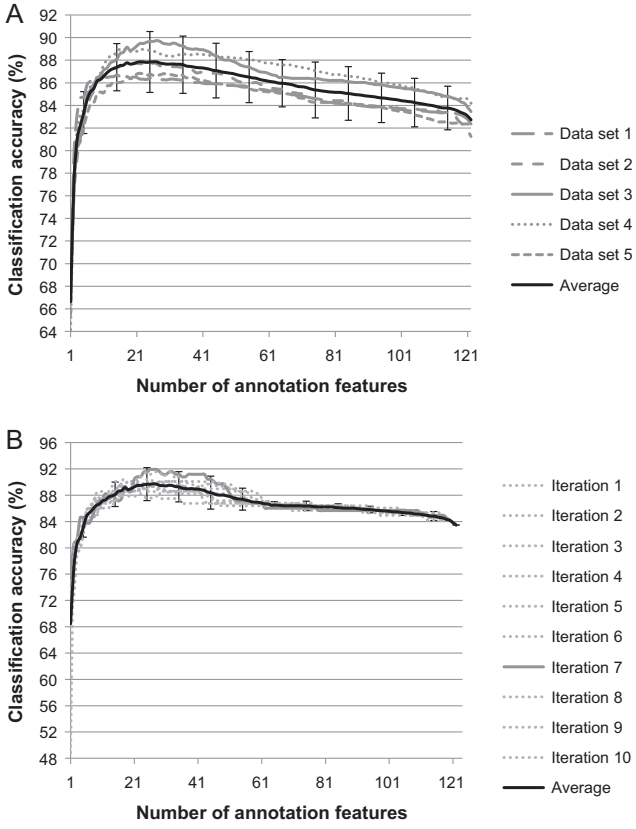

Fig. 3. Variation in SVM classification accuracy in the BPA data set for each pairing of positive training data (136 BPAs) with 5 different non-antigen data sets (A), and when one of these pairings (Data set 3) was selected to examine the variation associated with 10 iterations used to randomly break ties resulting from feature selection by greedy backward elimination (B). Classification accuracy was assessed for classifiers consisting of different numbers of annotation features selected by greedy backward elimination using a LTOCV approach. The average line (black) represents an averaging of the classification accuracies from the 5 pairings of antigen with different non-antigen data (A), or the 10 iterations performed to randomly break ties associated with greedy backward elimination (B). Error bars were calculated as 95% confidence intervals derived from the standard error of the mean or the standard deviation where appropriate, and displayed at intervals of 10 features along the average line.

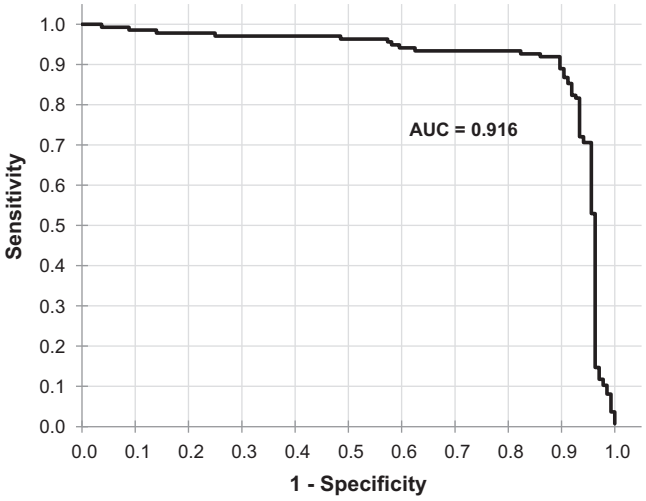

Fig. 4. An ROC curve was generated to assess the ability of the SVM-BPA classifier to discriminate antigens from non-antigens by plotting the true positive rate (Sensitivity) against the false positive rate (1 – Specificity) for different classification thresholds. The area under the curve (AUC) for this ROC curve (0.916) indicated that the SVM-BPA classifier was performing near optimal (AUC = 1).

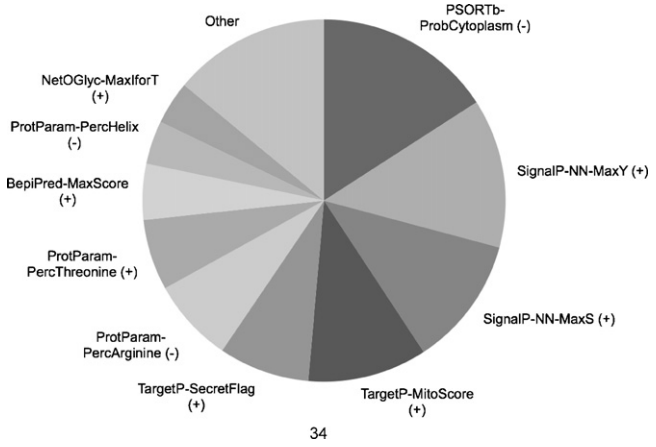

Fig. 5. Pie chart depicting the top 10 annotation features derived from protein annotation tools used by the SVM-BPA classifier to discriminate antigens from non-antigens. The syntax for the annotation features is fully explained in Supplemental Table 2. Feature selection using greedy backward elimination identified a total of 26 annotation features required by the SVM-BPA classifier in order to achieve maximum classification accuracy (*i.e.* 92%). An *F*-score was calculated for each annotation feature and the top 10 selected for the pie chart where the segment size is proportional to *F*-score. The positive or negative suffix indicates whether an annotation feature was positively or negatively correlated with antigenicity.

The final SVM-BPA classifier selected to assess recall of known antigens from the positive training data in a background of the entire bacterial proteome was taken as the best performing iteration in Fig. 3B (*i.e.* Iteration 7). From this iteration, 26 annotation features were selected for the final SVM-BPA classifier, which represent the minimum number of features required for a maximum classification accuracy of 92% as assessed by LTOCV. A receiver operating characteristic (ROC) curve was constructed in order to further assess the ability of the SVM-BPA classifier to separate antigens from non-antigens (Fig. 4). An ROC curve is a plot of the true positive rate (Sensitivity) versus the false positive rate (1 – Specificity) and describes the performance of the SVM-classifier across the entire range of classification thresholds. The area under the curve (AUC) was calculated from the ROC curve and represents an estimate of the probability that when an antigen and a non-antigen were selected at random, then the SVM-BPA classifier would assign a higher score to the antigen versus the non-antigen. An AUC of 0.916 was estimated for the SVM-BPA classifier, which is close to 1, and indicates that the SVM-BPA classifier was near optimal for discriminating antigens from non-antigens (Fig. 4).

3.4. Annotation features related to subcellular localization are important for antigen classification

An *F*-score was calculated for each annotation feature in the SVM-BPA classifier, whereby the larger the *F*-score, the greater the potential contribution of the feature for discriminating antigens from non-antigens. The 10 annotation features with the largest *F*-scores in the SVM-BPA classifier indicated that annotation related to the prediction of subcellular localization is critical for classifying proteins as antigens or non-antigens (Fig. 5). Specifically, antigens exhibited negative correlations with cytoplasmic localization and positive correlations with secreted and extracellular proteins as predicted by PSORTb, SignalP and TargetP. Fundamental protein properties (percent helix, arginine and threonine), humoral immunity (prediction of linear B-cell epitopes with BepiPred), and sites of O-glycosylation (NetOGlyc) may also be important for discriminating antigens from non-antigens. Although the prediction of subcellular localization appears central to discriminating antigens from non-antigens, it is probable that other annotation features may also be informative.

3.5. Classifiers significantly recall antigens when in the background of a whole bacterial proteome

In Fig. 2 the ability of antigen classifiers to discriminate antigens from non-antigens in the training data was assessed using cross-validation. A more practical measure of classifier utility is to assess the ability of classifiers to identify (*i.e.* recall) antigens from the positive training data when in a background of their native bacterial proteome. The SVM-BPA classifier was compared to the original VaxiJen method (DA-PLS-VaxiJen) for its ability to recall antigens from the proteomes of six different bacterial species: *Borrelia burgdorferi*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Treponema pallidum* (Fig. 6). The original VaxiJen method was implemented using the VaxiJen Server v2.0 (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) and utilizes a DA-PLS classifier trained on the VaxiJen data set annotated with ACC transformations of *z*-descriptors. Each classifier was used to rank annotated proteins in each bacterial proteome and then the position of known antigens from the training data in these ranked proteomes was assessed. The SVM-BPA classifier recalled all of the previously known antigens in all six bacterial proteomes with a higher rank than the DA-PLS-VaxiJen classifier (Fig. 6). The exception was recall of the 5th antigen, citrate synthase, for *H. pylori* where the DA-PLS-VaxiJen classifier outperforms the SVM-BPA classifier (Fig. 6B).

The BPA data set consisted of 136 antigens and the VaxiJen data set consisted of 100 antigens with an overlap of 74 antigens (*i.e.* BPAs) between the two data sets. When assessing recall between the two classifiers in Fig. 6, care was taken to select only those antigens that were shared between the BPA and VaxiJen data sets. In this way, recall could be compared between classifiers in an unbiased manner using a common set of antigens for each bacterial species. This avoided situations whereby the SVM-BPA classifier might appear to be recalling antigens with greater significance than the DA-PLS-VaxiJen classifier because certain antigens had never been used to train the DA-PLS-VaxiJen classifier, and vice versa. The six bacterial proteomes selected for recall analysis maximized the overlap between the BPA and VaxiJen data sets.

The ability of the SVM-BPA classifier to recall all of the antigens in the BPA data set for which a match could be found in the

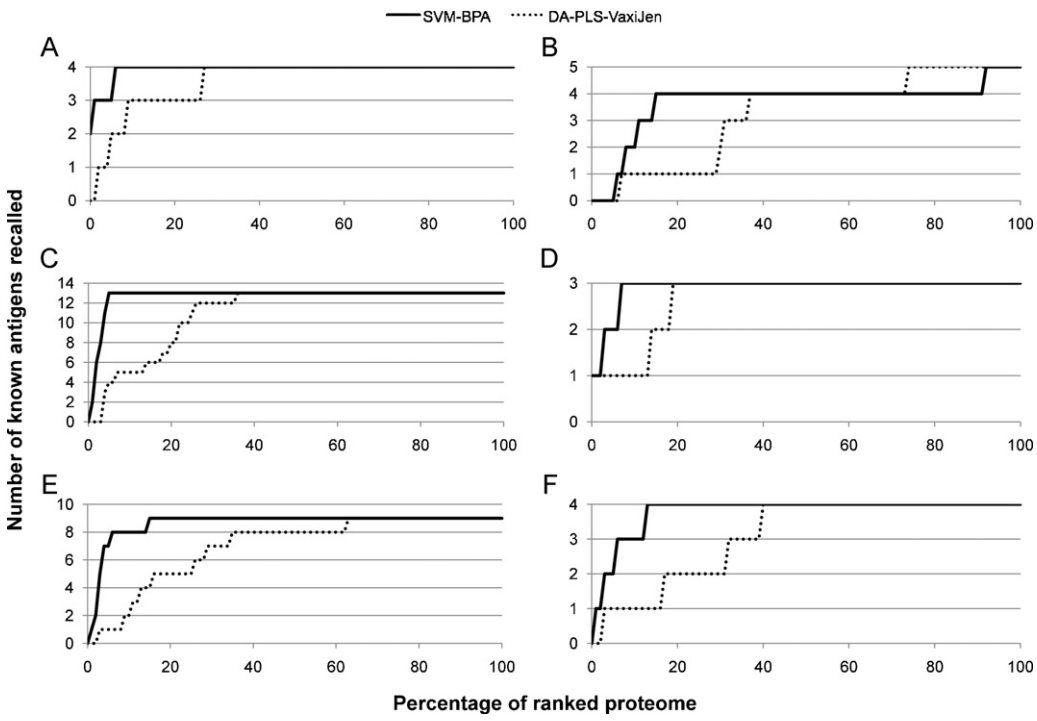

Fig. 6. Plots depicting the ability of the SVM-BPA and the DA-PLS-VaxiJen classifiers for recalling known antigens in the training data from the following bacterial proteomes: (A) *Borrelia burgdorferi*, (B) *Helicobacter pylori*, (C) *Mycobacterium tuberculosis*, (D) *Staphylococcus aureus*, (E) *Streptococcus pneumoniae*, and (F) *Treponema pallidum*. Strain, Gram staining and sequence (*i.e.* NCBI genome accession number) information for these bacterial species is available in Table 1. The original VaxiJen method (DA-PLS-VaxiJen) was implemented using the VaxiJen Server v2.0. Each classifier was used to rank the proteins in each bacterial proteome for their antigenicity and recall assessed as the position of the known antigens from the training data in this ranked list. The position of a known antigen in the ranked proteome is expressed as a percentage of the entire proteome.

Table 1
The significance of known antigen recall within the top 25, 50, 75, or 100 proteins when bacterial proteomes were ranked for antigenicity using the SVM-BPA classifier.[a]

Species	Strain	Gram	KA	TP	Acc. No.	25	*P*	50	*P*	75	*P*	100	*P*
Borrelia burgdorferi	80a	−ve	6	1989	NZ_ABJU00000000	4	**<1.00E−05**	5	**<1.00E−05**	5	**<1.00E−05**	5	**<1.00E−05**
Helicobacter pylori	26695	−ve	7	1570	NC_000915	1	1.02E−01	1	1.84E−01	2	**3.73E−02**	3	**6.81E−03**
Mycobacterium tuberculosis	H37Rv	NA	16	3987	NC_000962	1	9.16E−02	2	**1.56E−02**	3	**2.83E−03**	6	**<1.00E−05**
Staphylococcus aureus	N315	+ve	3	2537	NC_002745	1	**2.90E−02**	2	**1.12E−03**	2	**2.51E−03**	2	**4.44E−03**
Streptococcus pneumoniae	TIGR4	+ve	10	2125	AE005672	1	1.06E−01	2	**2.03E−02**	6	**<1.00E−05**	8	**<1.00E−05**
Treponema pallidum	Nichols	−ve	4	1030	NC_000919	2	**3.25E−03**	2	**1.26E−02**	3	**1.38E−03**	3	**3.23E−03**

[a] *Gram*, refers to the Gram stain for each bacteria and is listed as not applicable (NA) for *M. tuberculosis*; *KA*, is the total number of known bacterial protective antigens from the positive training data that had a matching homologous protein in the bacterial proteome of the strain selected for recall analysis with a blastp *E*-value <1E−40 (*Note*: a match for one antigen for *S. aureus* was not found below this *E*-value cut off, which is why the numbers of known antigens for this bacterial species does not correspond to the number presented in Supplemental Fig. 1); *TP*, the total number of protein coding genes in the proteome of each bacterial species; *P*, *p*-value associated with the significance of enrichment of known antigens for each cut off (25, 50, 75, and 100) as determined by a one-tailed Fisher's exact test and is in bold when <0.05. It should be noted that the proteomes of *H. pylori*, *M. tuberculosis*, *S. aureus* and *T. pallidum* are 3, 1, 46 and 6 proteins smaller, respectively, then their accession numbers indicate due to the removal of proteins with undetermined amino acids (*e.g.* X) which could not be annotated using certain bioinformatic tools (*e.g.* ProtParam).

proteome of each bacterial species, and not just the antigens that overlapped with the VaxiJen data set, was also assessed. In this case, the significance of antigen recall could be determined using strict cut-offs (*i.e.* 25, 50, 75 and 100), which represent different depths from the top of the list of proteins ranked by the SVM-BPA classifier for each bacterial proteome. Significant recall of known antigens was achieved for all 6 bacteria (Table 1) within the top 75 ranked proteins ($p < 0.05$, Fisher's exact test) and with even greater significance within the top 100 ranked proteins ($p < 0.01$). The top 100 antigens predicted for each bacterial species by the SVM-BPA classifier are included as supplemental data (Supplemental Table 3). GO analysis was performed to determine which biological processes were enriched for proteins in the top 100 ranked proteins from the proteome of *S. pneumoniae* (Supplemental Fig. 2). GO terms associated with pathogenesis, adhesion, cell wall biogenesis and transport across the cell membrane, were among the many that exhibited significant enrichment (FDR corrected *p*-value <0.05, hypergeometric test).

4. Discussion

RV approaches have the ability to rapidly accelerate the selection of antigen candidates for immunogenicity testing and subunit vaccine formulation [1,2]. For the first time, a powerful machine learning method (*i.e.* SVMs) has been incorporated into an RV approach in order to identify protective antigens in bacterial proteomes. Our approach to RV is summarized in Fig. 7 and represents a significant enhancement and extension to the methods originally proposed by Doytchinova and Flower [21]. Enhancements for discriminating antigens from non-antigens were demonstrated when training data was annotated with more biologically relevant information derived from bioinformatic tools

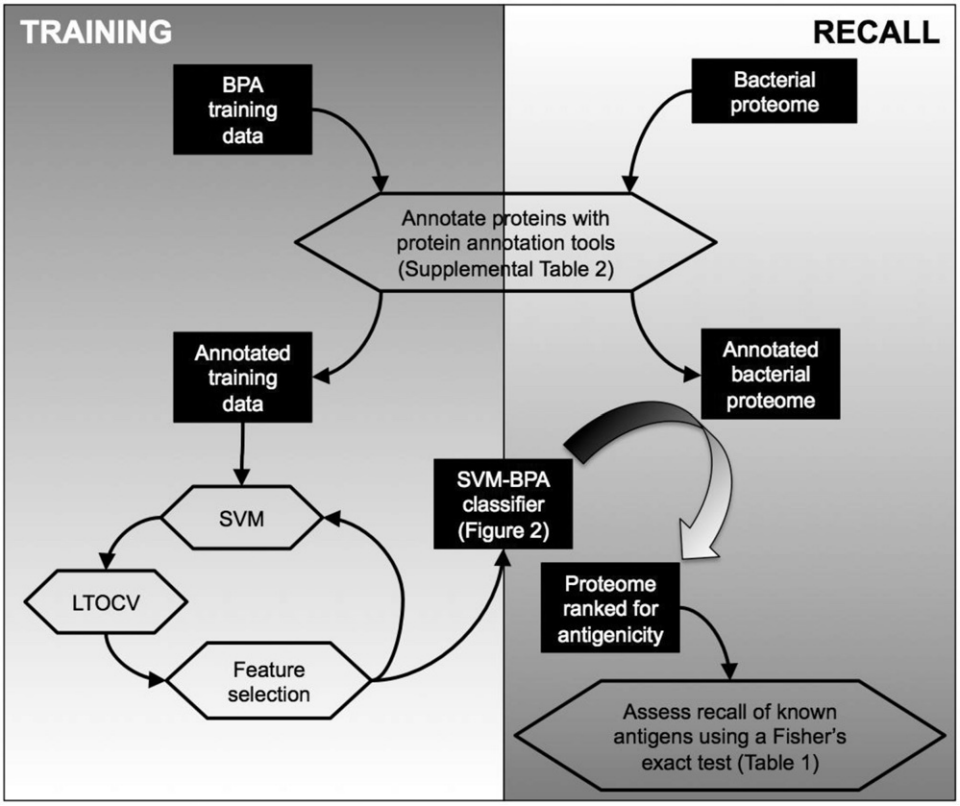

Fig. 7. Flowchart depicting the RV process used to train and assess the recall capacity of the SVM-BPA classifier. Bioinformatic tools for protein annotation were used to annotate proteins in the BPA training data set and in bacterial proteomes for 122 different features. This annotated data was used to train an SVM classifier whose accuracy was assessed by LTOCV. Feature selection by greedy backward elimination identified the minimum number of features ($N = 26$) that gave the maximum classification accuracy (*i.e.* 92%). The resulting SVM-BPA classifier was used to rank the annotated proteins in six bacterial proteomes for their antigenicity. A Fisher's exact test confirmed whether known antigens from the training data were significantly enriched at the top of the ranked bacterial proteome.

for protein annotation (Supplemental Table 2) instead of ACC transformations of *z*-descriptors (Fig. 1). In addition, SVM classifiers outperformed DA-PLS classifiers when separating antigens from non-antigens in the training data (Fig. 2) and when recalling known antigens in bacterial proteomes (Fig. 6). The optimal SVM classifier achieved an accuracy of 92% for discriminating antigens from non-antigens in the training data. In an informative extension to previous RV methods it was demonstrated that SVM classifiers significantly recall ($p < 0.05$, Fisher's exact test) known antigens from the training data within the top 75 ranked proteins for each of the six bacterial proteomes analyzed (Table 1).

The quality of the training data did not appear to have a large impact on SVM classification accuracy. The SVM-BPA and SVM-VaxiJen classifiers performed comparably well with respect to distinguishing antigens from non-antigens in the training data (Fig. 2). This was surprising since the BPA training data set was larger, not contaminated with sequence data from eukaryotic pathogens, and contained only antigens that were shown to be protective in animal models, when compared to the VaxiJen data set. However for future RV studies, the BPA training data set is recommended over VaxiJen as a benchmark data set because of its size and composition (Supplemental Fig. 1).

SVM classifiers represent state-of-the-art techniques that generally provide accurate models capable of capturing non-linearities in the data [41]. One of the drawbacks of SVM classifiers is that they typically utilize incomprehensible black-box models such that the evaluation of the exact criteria used to distinguish antigens from non-antigens is unclear. In an attempt to better understand the annotation features that were important for classification, a greedy backward elimination method of feature selection was used to identify the minimum number of features ($N = 26$) that resulted in maximum classification accuracy (*i.e.* 92%) and then *F*-scores were calculated for these features. The *F*-score takes into account both precision and recall and a larger *F*-score suggests that the annotation feature is more discriminative [42]. The 10 annotation features with the largest *F*-scores and their correlation with antigenicity are depicted in Fig. 5. Future research will focus on utilizing machine learning methods (*e.g.* random forests and decisions trees), which may be less accurate than SVMs, but more transparent with respect to the exact criteria used for discriminating antigens from non-antigens.

In agreement with previous RV studies [7,20], it appears that subcellular localization is important for discriminating antigens from non-antigens. The largest *F*-scores were associated with annotation features resulting from predictions of subcellular localization (Fig. 5). Extracellular localizations predicted by SignalP and TargetP were positively correlated with antigenicity and intracellular localizations predicted by PSORTb (*i.e.* probability of cytoplasmic localization) were negatively correlated. In addition, the vast majority of GO terms enriched for proteins in the top 100 potential antigens predicted by the SVM-BPA classifier in the proteome of *S. pneumoniae* were related to biological processes associated with the bacterial cell surface (Supplemental Fig. 2). These results conform to the general tenet of an RV approach, which states that potential vaccine candidates are bacterial proteins that are surface exposed or exported and thus accessible to antibody binding during infection [7,43,44]. This tenet is further confirmed since the maximum scoring linear B-cell epitope in each protein, an annotation feature predicted by BepiPred, was within the top 10 *F*-scores and exhibited a positive correlation with antigenicity (Fig. 5). In

contrast to previous RV studies that simply filtered out proteins lacking extracellular annotation [7,11–20], the SVM-BPA classifier uses both extracellular and intracellular annotation to classify bacterial proteins as potential protective antigens. Therefore, it is not surprising that the SVM-BPA classifier failed to recall the antigen citrate synthase [45] at a high rank in the *H. pylori* proteome (Fig. 6B) since PSORTb (v3.0) predicted with high probability that this particular antigen is localized to the cytoplasm (data not shown). The remaining annotation features with a top 10 *F*-score included the percentage of threonine and the highest scoring signal for O-glycosylation at a threonine residue, which were positively correlated with antigenicity, and the percentage of arginine or α-helices, which were negatively correlated with antigenicity (Fig. 5). However, the relationship between these features and antigenicity is less clear. For example arginine rich sequences have been used as adjuvants in vaccination regimens [46,47], associated with porin outer membrane proteins [48], and found in proteins secreted from the cytoplasm by the twin-arginine translocation (Tat) pathway [49]. Therefore arginine content might be expected to be positively correlated with antigenicity instead of negatively correlated. At this time the relationship between arginine content and antigenicity remains unclear and the importance of arginine content for classification of antigens in larger training data sets needs to be confirmed.

The existence of a causal link between antigenicity and features associated with subcellular localization and epitope prediction is quite probable. However, it is difficult to determine if a causal or simply a correlative role can be associated with the other types of annotation feature. It is quite possible that these annotation features simply correlate with some underlying and more informative causal feature. However, since the main aim of this study was to develop a classifier, it is not important whether annotation features have a correlative or causal relationship with antigenicity as long as these features can be used consistently to discriminate antigens from non-antigens as demonstrated here.

The SVM-BPA classifier was trained on a data set that contains a large diversity of bacterial species and different types of antigens from mycobacteria, and Gram positive and Gram negative bacteria. Therefore, this classifier should be able to identify putative protective antigens in bacterial species not represented in the training data or in newly sequenced bacterial species for which antigens have not been experimentally confirmed. In order to assess the generalizability of the SVM-BPA classifier, the proteomes of *Klebsiella pneumoniae* (strain 78578) and *Vibrio cholerae* (serovar O1, biovar El Tor, strain N16961) were selected, since they were not represented in the training data. For the *K. pneumoniae* proteome, which contains 4774 proteins, the outer membrane protein A (OmpA), fimbrial adhesion protein (MrkD), and OmpC (also known as OmpK36) antigens were significantly protective in animal models [50,51] and recalled by the SVM-BPA classifier at ranks of 84, 222, and 268, respectively. For the *V. cholerae* proteome ($N = 3820$), the putative porin, toxin co-regulated pilin A (TcpA), OmpW, and OmpU antigens were significantly protective in a rabbit ileal loop [52] or an infant mouse challenge model [53] and recalled by the SVM-BPA classifier at ranks of 34, 90, 257 and 764, respectively. When assessing the significance of antigen recall within the top 100 ranked proteins using a Fisher's exact test, known antigens were significantly recalled by the SVM-BPA classifier in the proteome of *V. cholerae* ($p = 3.87E{-}03$) and approached significance for *K. pneumoniae* ($p = 0.06$). It should be noted that the proteome of *K. pneumoniae* is by far the largest proteome analyzed for recall and a cut off of 100 therefore represents a much more conservative cut off for this bacteria. In summary, it is clear that the SVM-BPA classifier is capable of recalling known antigens for bacterial species that were not included in the training data.

What are the practical implications of the RV methodology presented in Fig. 7 for vaccine design? The bacterial immunologist will want to know how many predicted antigens need to be evaluated in animal models for their protective capacity in order to identify novel vaccine candidates. In this respect, we have shown that there was significant enrichment of known antigens in the top 75 ranked proteins for all of the bacterial proteomes assessed for recall (Table 1). The cloning and expression of 75 antigens for subsequent immunogenicity testing in animal models may be a tall order for a single academic laboratory but should be within reach of a small number of collaborating laboratories. This number ($N = 75$) is certainly an improvement with respect to previous RV studies [7] that used filtering approaches ($N = 570$). Such immunogenicity testing in animal models is required in order to assess the ability of this classifier to identify truly novel protective antigens. Prior to immunogenicity testing, predicted antigens for a given bacterial species should be screened for those that contain more than two transmembrane domains and thus represent proteins that may be difficult to clone and express [2,7]. This criteria would not limit the study of the top 75 potential antigens predicted by the SVM-BPA classifier for each of the 6 bacterial species analyzed in this study since *S. pneumoniae* was the bacterial species whose top 75 contained the most proteins with more than two transmembrane domains of which there were only 6 (Supplemental Table 3). In addition, it would be beneficial to ensure that the bacteria under study express predicted antigens during the pathogenic phase of infection. If bacteria do not express a particular protein, then this protein is unlikely to trigger an immune response. In fact, as gene expression data from bacterial microarrays becomes more common [54], it can be used to annotate training data in order to train classifiers. Testing the protective capacity in animal models of the novel antigens predicted by SVM classifiers is beyond the scope of the methods presented here but will be performed in the future to further validate and refine our RV approach.

In summary, RV approaches based on SVM classifiers appear to be the most robust developed to date as assessed by their ability to distinguish antigens from non-antigens in training data sets, and their power for recalling known antigens from the proteomes of pathogenic bacteria. The 136 BPAs and 5 different sets of matching non-antigen data curated through the course of this study are available in Supplemental Files 1 to 6. RV researchers are encouraged to use this data set as benchmark in order to evaluate future methodological improvements.

Acknowledgements

This work was performed with the support of the Genomics Core at the UCSD Center for AIDS Research (AI36214), the Biomedical Informatics Research Center (http://informatics.sdsu.edu/) at San Diego State University and its High Performance Computing facility, and a United Health scholarship to Paul R. McAdam from the San Diego Veterans Medical Research Foundation. This material is based upon work supported in part by the Department of Veterans Affairs (VA), Veterans Health Administration, Office of Research and Development. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government. Gratitude is extended to Drs. Benjamin M. Good, Art Poon, Sergei Kosakovsky Pond, Theo N. Kirkland, Sharon Reed, Roberto Badaro and Francesco Filippini, who engaged in invaluable discussions throughout this study, and the reviewers of the original manuscript whose critiques resulted in a much more coherent body of work. This paper is dedicated to the memory of the senior author's aunt, Lee Ann Cowhig, who passed away in 2010 due to a preventable infectious disease.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.vaccine.2011.07.142.

References

[1] Mora M, Veggi D, Santini L, Pizza M, Rappuoli R. Reverse vaccinology. Drug Discov Today 2003;8:459–64.
[2] Vivona S, Gardy JL, Ramachandran S, Brinkman FS, Raghava GP, Flower DR, et al. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. Trends Biotechnol 2008;26:190–200.
[3] Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2010;38:D346–54.
[4] Adu-Bobie J, Capecchi B, Serruto D, Rappuoli R, Pizza M. Two years into reverse vaccinology. Vaccine 2003;21:605–10.
[5] Rogan D, Babiuk LA. Novel vaccines from biotechnology. Rev Sci Tech 2005;24:159–74.
[6] Meinke A, Henics T, Nagy E. Bacterial genomes pave the way to novel vaccines. Curr Opin Microbiol 2004;7:314–20.
[7] Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000;287:1816–20.
[8] Tan LK, Carlone GM, Borrow R. Advances in the development of vaccines against *Neisseria meningitidis*. N Engl J Med 2010;362:1511–20.
[9] Babiuk LA. Broadening the approaches to developing more effective vaccines. Vaccine 1999;17:1587–95.
[10] Minor PD, John A, Ferguson M, Icenogle JP. Antigenic and molecular evolution of the vaccine strain of type 3 poliovirus during the period of excretion by a primary vaccinee. J Gen Virol 1986;67(Pt 4):693–706.
[11] Montigiani S, Falugi F, Scarselli M, Finco O, Petracca R, Galli G, et al. Genomic approach for analysis of surface proteins in Chlamydia pneumoniae. Infect Immun 2002;70:368–79.
[12] Finco O, Bonci A, Agnusdei M, Scarselli M, Petracca R, Norais N, et al. Identification of new potential vaccine candidates against Chlamydia pneumoniae by multiple screenings. Vaccine 2005;23:1178–88.
[13] Gamberini M, Gomez RM, Atzingen MV, Martins EA, Vasconcellos SA, Romero EC, et al. Whole-genome analysis of Leptospira interrogans to identify potential vaccine candidates against leptospirosis. FEMS Microbiol Lett 2005;244:305–13.
[14] Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, et al. Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science 2005;309:148–50.
[15] Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infect Immun 2001;69:1593–8.
[16] Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, et al. Identification of vaccine candidate antigens from a genomic analysis of Porphyromonas gingivalis. Vaccine 2001;19:4135–42.
[17] Ariel N, Zvi A, Grosfeld H, Gat O, Inbar Y, Velan B, et al. Search for potential vaccine candidate open reading frames in the Bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening. Infect Immun 2002;70:6817–27.
[18] Gat O, Grosfeld H, Ariel N, Inbar I, Zaide G, Broder Y, et al. Search for Bacillus anthracis potential vaccine candidates by a functional genomic-serologic screen. Infect Immun 2006;74:3987–4001.
[19] Lei B, Liu M, Chesney GL, Musser JM. Identification of new candidate vaccine antigens made by Streptococcus pyogenes: purification and characterization of 16 putative extracellular lipoproteins. J Infect Dis 2004;189:79–89.
[20] Vivona S, Bernante F, Filippini F. NERVE new enhanced reverse vaccinology environment. BMC Biotechnol 2006;6:35.
[21] Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 2007;8:4.
[22] Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, et al. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 2005;21:617–23.
[23] Sachdeva G, Kumar K, Jain P, Ramachandran S. SPAAN a software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 2005;21:483–91.
[24] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–50.
[25] Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 1987;30:1126–35.
[26] Guan P, Doytchinova IA, Walshe VA, Borrow P, Flower DR. Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex HLA-A0201. J Med Chem 2005;48:7418–25.
[27] Siebert KJ. Quantitative structure-activity relationship modeling of peptide and protein behavior as a function of amino acid composition. J Agric Food Chem 2001;49:851–8.
[28] Wanchana S, Yamashita F, Hara H, Fujiwara S, Akamatsu M, Hashida M. Two- and three-dimensional QSAR of carrier-mediated transport of beta-lactam antibiotics in Caco-2 cells. J Pharm Sci 2004;93:3057–65.
[29] Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 2007;16:216–26.
[30] Bhasin M, Raghava GP. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine 2004;22:3195–204.
[31] Gao T, Yang Z, Wang Y, Jing L. Identifying translation initiation sites in prokaryotes using support vector machine. J Theor Biol 2010;262:644–9.
[32] Woelk CH, Beliakova-Bethell N, Goicoechea M, Zhao Y, Du P, Rought SE, et al. Gene expression before HAART initiation predicts HIV-infected individuals at risk of poor CD4+ T-cell recovery. AIDS 2010;24:217–22.
[33] O'Boyle NM, Palmer DS, Nigsch F, Mitchell JB. Simultaneous feature selection and parameter optimisation using an artificial ant colony: case study of melting point prediction. Chem Cent J 2008;2:21.
[34] Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 1993;277:239–53.
[35] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm; 2001.
[36] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann; 1999.
[37] Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
[38] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
[39] Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674–6.
[40] Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21:3448–9.
[41] Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: an overview of issues and application in credit scoring. In: Diederich J, editor. Rule extraction from support vector machines: studies in computational intelligence. Berlin, Heidelberg: Springer-Verlag; 2008. p. 33–63.
[42] Chang YW, Lin CJ., Feature ranking using linear svm. In JMLR Workshop and Conference Proceedings, volume 3: WCCI 2008 causality challenge, Hong Kong, June 3-4 2008, p 53-64.
[43] Barocchi MA, Censini S, Rappuoli R. Vaccines in the era of genomics: the pneumococcal challenge. Vaccine 2007;25:2963–73.
[44] De Groot AS, Rappuoli R. Genome-derived vaccines. Expert Rev Vaccines 2004;3:59–76.
[45] Dunkley ML, Harris SJ, McCoy RJ, Musicka MJ, Eyers FM, Beagley LG, et al. Protection against Helicobacter pylori infection by intestinal immunisation with a 50/52-kDa subunit protein. FEMS Immunol Med Microbiol 1999;24:221–5.
[46] Mattner F, Fleitmann JK, Lingnau K, Schmidt W, Egyed A, Fritz J, et al. Vaccination with poly-L-arginine as immunostimulant for peptide vaccines: induction of potent and long-lasting T-cell responses against cancer antigens. Cancer Res 2002;62:1477–80.
[47] Riedl P, Stober D, Oehninger C, Melber K, Reimann J, Schirmbeck R. Priming Th1 immunity to viral core particles is facilitated by trace amounts of RNA bound to its arginine-rich domain. J Immunol 2002;168:4951–9.
[48] Schulz GE. Bacterial porins: structure and function. Curr Opin Cell Biol 1993;5:701–7.
[49] DeLisa MP, Tullman D, Georgiou G. Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. Proc Natl Acad Sci USA 2003;100:6115–20.
[50] Kurupati P, Ramachandran NP, Poh CL. Protective efficacy of DNA vaccines encoding outer membrane protein A and OmpK36 of Klebsiella pneumoniae in mice. Clin Vaccine Immunol 2011;18:82–8.
[51] Li Y, Han WY, Li ZJ, Lei LC. Klebsiella pneumoniae MrkD adhesin-mediated immunity to respiratory infection and mapping the antigenic epitope by phage display library. Microb Pathog 2009;46:144–9.
[52] Das M, Chopra AK, Cantu JM, Peterson JW. Antisera to selected outer membrane proteins of Vibrio cholerae protect against challenge with homologous and heterologous strains of V. cholerae. FEMS Immunol Med Microbiol 1998;22:303–8.
[53] Rollenhagen JE, Kalsy A, Cerda F, John M, Harris JB, Larocque RC, et al. Transcutaneous immunization with toxin-coregulated pilin A induces protective immunity against Vibrio cholerae O1 El Tor challenge in mice. Infect Immun 2006;74:5834–9.
[54] Eriksson S, Lucchini S, Thompson A, Rhen M, Hinton JC. Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. Mol Microbiol 2003;47:103–18.