Sequence analysis

High-throughput prediction of protein antigenicity using protein microarray data

Christophe N. Magnan1, Michael Zeller1, Matthew A. Kayala1, Adam Vigil2, Arlo Randall1, Philip L. Felgner1,2 and Pierre Baldi1,∗

1Institute for Genomics and Bioinformatics, School of Information and Computer Sciences and 2Department of Medicine, Division of Infectious Diseases, University of California, Irvine, CA 92697, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Discovery of novel protective antigens is fundamental to the development of vaccines for existing and emerging pathogens. Most computational methods for predicting protein antigenicity rely directly on homology with previously characterized protective antigens; however, homology-based methods will fail to discover truly novel protective antigens. Thus, there is a significant need for homology-free methods capable of screening entire proteomes for the antigens most likely to generate a protective humoral immune response.

Results: Here we begin by curating two types of positive data: (i) antigens that elicit a strong antibody response in protected individuals but not in unprotected individuals, using human immunoglobulin reactivity data obtained from protein microarray analyses; and (ii) known protective antigens from the literature. The resulting datasets are used to train a sequence-based prediction model, ANTIGENpro, to predict the likelihood that a protein is a protective antigen. ANTIGENpro correctly classifies 82% of the known protective antigens when trained using only the protein microarray datasets. The accuracy on the combined dataset is estimated at 76% by cross-validation experiments. Finally, ANTIGENpro performs well when evaluated on an external pathogen proteome for which protein microarray data were obtained after the initial development of ANTIGENpro.


Contact: pf baldi@ics.uci.edu

Received on August 11, 2010; revised on September 8, 2010; accepted on September 23, 2010

1 INTRODUCTION

Identification of antigen proteins capable of triggering a significant humoral immune system response is important for addressing fundamental questions in immunology, virology and bacteriology. It is also important for practical purposes ranging from diagnostic applications to vaccine design. The goal of this article is to develop and test a predictor of protein antigenicity that can be used on a high-throughput scale on existing or new proteomes to identify key antigenic proteins that may have protective qualities and may be used in vaccine design applications. The predictor is developed by applying machine-learning methods to training data resulting from a unique high-throughput proteomic chip technology originally developed in the Felgner Laboratory (Davies et al., 2005), as well as data extracted from the literature and public databases.

From the outset, it must be recognized that the notion of protein antigen is similar to many other notions in biology (e.g. gene, consciousness) that are very useful but not tightly defined. Given the complexity, variability and flexibility of the immune system, at one extreme, one could take the position that every protein has the potential of being antigenic and of triggering a humoral immune response. On the other hand, it is well known that immune systems respond differentially to the various proteins of a pathogen and that there are commonalities among the humoral immune responses of different individuals exposed to the same pathogen. Within the proteins of a pathogen considered to be antigenic, one can further distinguish several overlapping subclasses with fuzzy boundaries such as protective antigens, serodiagnostic antigens and cross-reactive antigens. To a first degree of approximation: protective antigens are important for conferring protection, serodiagnostic antigens are associated with a differential humoral antibody response between naive and exposed individuals and are important for diagnostics purposes and cross-reactive antigens are associated with a strong humoral antibody response in both naive and exposed individuals. The primary focus of this work is on protective antigen prediction.

Protecting populations against infectious pathogens is an important priority and vaccination is widely recognized as one of the most reliable preventive approaches. By simulating the presence of a given pathogen, a vaccine elicits a specific protective immune response. Although most proteins produced by a given pathogen can be considered to be antigens, only some, denoted protective antigens, induce an effective immune response against the whole pathogen (Rappuoli, 2001). These protective antigens are usually surface-exposed or exported proteins accessible to the immune system (Rodriguez-Ortega et al., 2006). For safety purposes, the historical trend has been toward creating subunit vaccines or epitope vaccines (Schmidt, 1989) containing only full or partial protective antigens, as opposed to early vaccines based on attenuated whole pathogens. Identification of protective antigenic proteins or determinants is therefore a top priority in current vaccine development projects (Dooyenova and Flower, 2007a). Thanks to advances in genomic technologies, pathogen genomes can now be rapidly obtained (Rappuoli and Covacci, 2003), opening the door for in silico screening of a pathogen’s entire proteome...
for the antigens most likely to elicit a protective immune response. Prediction of subunit antigenic determinants, B-cell epitopes, by computational methods has been an active area of research for a long time (Hopp and Woods, 1981; Kolaskar and Tongaonkar, 1990; Thornton et al., 1986; Welling et al., 1985) and is still an accepted approach for modern vaccinology (Andersen et al., 2006; Larsen et al., 2006; Odorico and Pellequer, 2003; Rubinstein et al., 2009; Saha and Raghava, 2006; Sollner and Mayer, 2006; Sweredoski and Baldi, 2008, 2009). However, because of the non-linear nature of epitopes in active folded proteins and the immune system’s ability for adaptive response to antigens, the relevance of such predictors remains unclear and is debated often (Blythe and Flower, 2005; Greenbaum et al., 2007; Ponomarenko and Bourne, 2007).

Prediction of protective antigens is usually referred to as reverse vaccinology (Rappuoli, 2001; Rappuoli and Covacci, 2003). This approach has obtained several notable successes (Rappuoli and Covacci, 2003) since the seminal work of Pizza et al. (2000). In previous work, the best vaccine candidates have been selected using both similarity to known protective antigens and predicted characteristics, such as the protein localization. For instance, Pizza et al. (2000) used BLAST (Altschul et al., 1990), GCG (Accelrys Software Inc, 2005), FASTA (Pearson, 1990) and PSORT (Nakai and Horton, 1999) to predict features typical of surface-associated proteins. These methods rely on sequence similarity with known protective antigens to predict relevant vaccine candidates in the proteome of a new pathogen. While homology to known protective antigens is a reasonable criteria for identifying some new protective antigens, it is by no means exhaustive and, by definition, will miss novel non-homologous protective antigens.

Doytchinova and Flower (2007a, b) have proposed a sequence-based method that does not rely on homology as the methods described above do. Three distinct datasets were prepared to predict protective bacterial antigens, protective viral antigens and tumor antigens, respectively. Each dataset consists of 100 antigens collected from the literature as positive examples and 100 proteins randomly selected from the same set of species as negative examples. The corresponding predictive models were derived by applying a two-class discriminant analysis using partial least squares applied to a uniform representation of the protein sequences, obtained by the auto cross-covariance method described in Wold et al. (1993). The estimated prediction accuracy of the resulting predictor, Vaxijen, ranges from 70% to 89% depending on the evaluation method and training set. The same method has been applied to 33 protective fungal antigens and 117 protective parasite antigens in Doytchinova and Flower (2008), the corresponding models are also included in Vaxijen. To the best of our knowledge, the Vaxijen predictor is the first and only alignment-free bioinformatics tool available for protective antigen prediction.

Considering the small size of the datasets used for training and evaluation and the protocol for selecting negative examples, it is reasonable to suspect that the published accuracy estimates for Vaxijen are somewhat biased and that improvements ought to be possible with larger annotated datasets. However, collecting data on protective and non-protective antigens remains a particularly difficult task in spite of the numerous B-cell epitope databases available like Antilens (Toseland et al., 2005), IEDB (Peters et al., 2005), Bic geopolitical (Saha et al., 2005) or AntigenDB (Ansari et al., 2010). Since information about the neutralization of the parent protein or pathogen is rarely available. Reviewing the vaccinology literature to find relevant proteins to build a large training set is laborious and yields relatively small datasets. Thus, new high-throughput approaches must be developed to obtain larger data sets.

Here, we take advantage of a new high-throughput technology (Barbour et al., 2008; Davies et al., 2005; Sundaresh et al., 2006) to study the humoral immune response to pathogen infection using protein microarrays. The technology uses a proprietary in vitro expression system to express the proteins encoded in the genome of a pathogen and print them on nitrocellulose arrays where they can be probed with sera from different individuals (e.g. naive versus exposed versus vaccinated). Secondary antibody coupled with fluorescent methods are used to visualize the entire response profile as with DNA microarrays. The proteomes of several pathogens such as Francisella tularensis (Eyles et al., 2007), Burkholderia pseudomallei (Felgner et al., 2009) and Plasmodium falciparum (Crompton et al., 2008, 2010) have been partially or entirely printed and probed against the sera of individuals with either positive or negative clinical tests for the corresponding pathogen. The resulting reactivity data provide a reliable estimate of the humoral immune response to each pathogen protein for each sera sample. Thus, these protein microarray data can be used to curate relevant datasets to train sequence-based machine-learning methods for predicting the degree of humoral immune response to novel proteins. Although the protein microarray data does not directly provide information about whether or not a particular antigen is protective, our working hypothesis is that the actual protective antigens are significantly overrepresented among the set of antigens for which the protected individuals elicit a significant antibody response, and the unprotected individuals do not.

Here we begin by curating a large and non-redundant set of antigens with known immunogenicity using this high-throughput technology and combine it with data extracted from the literature and the existing databases. Only pathogen proteins are analyzed in this work, thus, self-antigens from tumors and non-peptide antigens are not utilized. From this protein set, we extract several sequence-based features and develop a two-stage machine-learning architecture to predict protective antigenicity from the protein primary sequence. The predictive abilities of the resulting system are assessed in four different ways: (i) by direct comparison with the method proposed in Doytchinova and Flower (2007b); (ii) by standard 10-fold cross-validation; (iii) by repeated cross-validations using each pathogen protein subset as a fold; and (iv) by external validation on a pathogen proteome for which protein microarray data were obtained after the initial development of ANTIGENpro.

2 DATASETS AND METHODS

In this section, we describe the methodological steps of our study: (i) preparation of a rigorous set of antigenic and non-antigenic proteins; (ii) extraction of several sequence-based feature sets for each protein; and (iii) derivation of ANTIGENpro to predict protein antigenicity. For convenience, protein sets appear in bold and feature sets appear in square brackets.

2.1 Protein datasets for protective antigen prediction

We curate seven independent sets of proteins for the purpose of studying the prediction of protein antigenicity from primary sequence. Five of the datasets are curated using protein microarray data analysis for each of the following pathogens: Candida albicans, Plasmodium falciparum, Brucella melitensis, Burkholderia pseudomallei and Mycobacterium tuberculosis.
Table 1. Description of the five protein microarray experiments used in this study for the preparation of the microarray datasets (Section 2.1.1).

<table>
<thead>
<tr>
<th>Pathogen (Strain)</th>
<th>Microarray proteins</th>
<th>Exposed samples</th>
<th>Naive samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candida albicans (SC5314)</td>
<td>109</td>
<td>31</td>
<td>62</td>
</tr>
<tr>
<td>Burkholderia pseudomallei (K92624)</td>
<td>132</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>Plasmodium falciparum (3D7)</td>
<td>2279</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>Brucella melitensis (16M)</td>
<td>1404</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis (H37Rv)</td>
<td>3883</td>
<td>13</td>
<td>69</td>
</tr>
</tbody>
</table>

Details and protocols followed to prepare these sets are given in Section 2.1.1. In addition, a dataset containing known protective antigens found in the literature and public databases, P\text{Antigen}, was curated and the details are given in Section 2.1.2. After the development of ANTIGENpro, an additional protein microarray dataset was obtained for the pathogen Bartonella henselae, and this dataset is used for external validation of the method. The details of this external dataset are provided in Section 2.1.3.

2.1.1 Protein microarray data analysis

Technology developed in the Felgner Laboratory, described in detail in Davies et al. (2005), was used to obtain reactivity data for most of the proteins of the five pathogens listed in Table 1. In short, for a given pathogen, the process begins by PCR amplification of primers for all deduced open reading frames (ORFs) in the sequenced genome. The amplified DNA is cloned with an in vivo recombination method into a plasmid expression vector. Then the vector is run through an in vitro transcription/translation process to produce polypeptides. These peptides are then printed on an enzyme-linked immunosorbent assay (ELISA) (Ingvald and Perffman, 1971) array for screens against patient sera. ELISA uses fluorescence of a tagged dye probe to indirectly measure binding between antibodies in sera and the proteins printed on the array. A single microarray experiment consists of probing an array with patient sera, known as the primary, then probing with a secondary that binds with antibodies present in the primary, and finally a tagged dye conjugated with a molecule known to bind to the secondary is applied as a tertiary probe. The arrays are then scanned via laser, and fluorescence intensities are measured. The printing, probing and scanning aspects of these experiments are similar to the processes used with cDNA microarrays.

Normalization and differential analysis protocols developed for DNA microarrays have been extended to these new protein arrays (Sundaresh et al., 2006, 2007). Each dataset is prepared for analysis in the following manner: the VSN transformation (Huber et al., 2002) is applied to the dataset of class, typically using a plasmid expression vector with no inserted cloned DNA as a control. Differential analysis is done using the Cyber-T software described in Baldi and Long (2001) and available online at http://cybert.ucsc.edu. Cyber-T calculates a Bayesian regularized estimate of the variance of the signal intensity levels. Then these variance estimates are used to compare the groups with a t-test. The Benjamin–Hochberg multiple test correction is used to scale the P-values, which are used as a measure of differential reactivity between naive and exposed groups.

For each pathogen used in this study the pathogen name, specific strain, number of proteins with microarray data, number of exposed sera samples, and number of naive sera samples are displayed in Table 1. The ORFs and primers used for each dataset are available at http://portal.proteomics.scs.uci.edu/virus/portal.php. Each pathogen has at least 40 associated sera samples. The sera samples are divided into naive and exposed groups as described below.

- The 
  C.albicans dataset (Candida) exposed group sera comes from patients being treated for candidiasis (recovery of Candida from blood cultures) at Shands Teaching Hospital at the University of Florida (STI-UF). For the candidiasis patients, most sera were collected within 7 days from the first date that blood cultures were positive. The naive control group sera comes from hospitalized patients with negative tests at the Infectious Diseases Consultation Service of STI-UF and from volunteers at the University of California, Irvine (Mochon et al., 2010).
- The 
  P.falciparum dataset (Malaria) consists of patients from a 2006 longitudinal study conducted in malaria endemic rural Md. Sera samples were taken in May, shortly before the beginning of malaria season. To avoid age-related factors, only subjects aged 8–10 years were analyzed. The subjects were monitored throughout the following malaria season. At the end of the season if a patient had at least one malaria episode, their May sera sample was categorized as naive, and if they had no episodes their May sera was categorized as exposed. The dataset is described in detail in Crompton et al. (2008).
- The 
  B.mellitensis dataset (Brucella) exposed group sera comes from patients in Lima, Peru confirmed to have acute brucellosis by both Rose Bengal screening test and positive blood culture. All exposed patients had their first known episode of brucellosis, and sera was drawn within 1–3 weeks of onset of symptoms. The naive group sera come from ambulatory healthy control individuals from Lima, Peru. The dataset is described in detail in Liang et al. (2010).
- The 
  M.tuberculosis dataset (Tuberculosis) consists of sera samples collected from Colombia where the exposed group tests positive using both sputum smear and bacterial culture tests for M.tuberculosis and the naive control group is negative for both tests. Additionally, all sera are HIV negative in order to minimize the confounding relationship between HIV positive sera and the antibody response to the M.tuberculosis antigens.

As discussed in the introduction, the notions of ‘antigen’ and ‘protective antigen’ are fuzzy. Here we describe the general approach and specific steps taken to curate sets of positive and negative examples, based on protein microarray analysis, that are relevant for the task of predicting the likelihood that a particular protein is a protective antigen. Since no direct information about protection is available, we identify the sets of antigens which elicit: (i) the strongest response among protected individuals; and (ii) the most significant differential response between protected and unprotected individuals. The working hypothesis is that these antigens are significantly more likely to contribute to a protective immune response than the other proteins in the proteome. This key assumption will be tested using a set of known protective antigens from the literature, P\text{Antigen}, described in Section 2.1.2.

Specifically, the following protocol is applied to each pathogen dataset to classify proteins as antigenic or non-antigenic. First, the proteins are ranked according to the mean antibody reactivity calculated from the exposed sera group. The most reactive 10% are reordered by the P-value calculated by the t-test between the exposed and naive sera groups and the 50% most differentially reactive of this subset (5% of total), with the lowest P-values, are labeled as antigenic. The non-antigenic proteins are selected in the opposite fashion. Starting from the complete list of proteins ranked by the mean reactivity, the least reactive 20% are reordered by the P-value and the 50% least differentially reactive of this subset (10% of total), with the highest P-values, are labeled as non-antigenic.

Finally, the antigen and non-antigen sets from the five pathogens are merged and redundancy reduced using BLASTCLUST (Altschul et al., 1997) with a 30% similarity threshold, as for P\text{Antigen} (Section 2.1.2). In addition, proteins in the merged pathogen set with more than 30% similarity with any protein in P\text{Antigen} are also removed. The size and composition of the six final datasets are reported in Table 2.

2.1.2 Known protective antigens: P\text{Antigen}

As discussed in the introduction, collecting relevant data for predicting protective antigens is
difficult. To the best of our knowledge, there are no public databases dedicated specifically to this purpose. B-cell epitope databases such as IEDB (Peter et al., 2005), AntigenDB (Jans et al., 2010) or AntigenDB (Jans et al., 2010) represent a post source of data for this type of study. Most of the reported B-cell epitopes are either part of the same set of extensively studied and epitope-mapped antigens or are peptides for which the neutralization of the parent protein, let alone the pathogen, is unknown and cannot be assumed from the neutralization of the epitope (Greenbaum et al., 2007). The immunology literature remains the principal source of relevant data for this prediction problem.

Doytchinova and Flower (2007b) collected 100 bacterial and 100 viral protective antigens by reviewing the literature. We merged these two protein sets together with antigens reported in Kolaskar and Tongaonkar (1990), in Rodriguez-Ortega et al. (2006) and 12 immunogenic antigens found in the Bcipep database resulting in a set of 246 known protective antigens. Sequence redundancy was then reduced with a 30% similarity cutoff using BLASTCLUST (Altschul et al., 1997). The final PAntigens set contains 213 non-redundant protective antigens. UniProt (The UniProt Consortium, 2007) identifiers of the 213 protective antigens are reported in Table 1. Note that there are several important differences between PAntigens and the Vaxijen (Doytchinova and Flower, 2007b) datasets. First, only pathogen proteins are considered in this study, therefore tumor antigens are not included. In addition, different groups of pathogens (e.g. bacteria, viruses or yeasts) are not separated to train distinct prediction models. Finally, the proteins classified as non-antigenic in this study are curated by selecting proteins with low seroreactivity according to the protein microarray experiments, whereas in other studies proteins classified as non-antigenic were selected randomly.

2.1.3 External proteome: Bartonella

The pathogen Bartonella henselae was recently analyzed by the Folger Laboratory using the methods described in Section 2.1.1. The corresponding dataset, Bartonella, is used in this work for external validation of ANTIGENpro (Section 2.3). As highlighted in Bloeker et al. (2003), such validation provides a bias-reduced evaluation of a predictor. The dataset B.henselae consists of 1463 proteins. The procedure to select the most antigenic proteins based on the microarray data described in Section 2.1.1, was applied to B.henselae resulting in a set of 73 antigenic proteins. The remaining 1390 were then classified as non-antigenic for the purpose of evaluating the ability of ANTIGENpro to recover the most antigenic proteins from an entire proteome. Some basic information on the source of the sera samples for B.henselae is provided below.

- The B.henselae dataset (Bartonella) consists of sera collected from cats admitted to animal shelters in the San Francisco area. The dataset consists of 62 sera samples exposed to Bartonella as confirmed by the gold standard IFA titers test. The naive set is composed of 67 samples confirmed as not exposed via IFA titers in addition to sera from eight specific pathogen free (SPF) cats. The dataset is described in detail in Vigil et al. (2010).

Table 2. Size and composition of the six protein sets used to train ANTIGENpro

<table>
<thead>
<tr>
<th>Protein set</th>
<th>Size</th>
<th>Antigenic</th>
<th>Non-antigenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAntigens</td>
<td>213</td>
<td>213</td>
<td>0</td>
</tr>
<tr>
<td>Brucella</td>
<td>206</td>
<td>70</td>
<td>136</td>
</tr>
<tr>
<td>Burkholderia</td>
<td>17</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Candida</td>
<td>13</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Malaria</td>
<td>333</td>
<td>114</td>
<td>219</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>542</td>
<td>171</td>
<td>371</td>
</tr>
<tr>
<td>Total</td>
<td>1324</td>
<td>576</td>
<td>748</td>
</tr>
</tbody>
</table>

Table 3. Size of the initial feature sets

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Size</th>
<th>Feature set</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Hydropo-5:MI]</td>
<td>5</td>
<td>[Computed]</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 4. Amino acid alphabets used to compute frequencies of monopeptides and dipeptides

<table>
<thead>
<tr>
<th>Name</th>
<th>Amino acid groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydropho-5</td>
<td>CPFTPMW, HSTY, DEER, AG, HP</td>
</tr>
</tbody>
</table>

Amino acids groups are separated by commas.

2.2 Sequence-based features

From the immunology literature detailed in Section 1, only a small set of primary sequence characteristics are usually considered indicative of protein antigenicity. Here to train ANTIGENpro we take a broader approach by using a large number of features and feature sets. We follow a protocol similar to the one proposed to predict protein solubility on overexpression in Magnum et al. (2009). Among the 23 distinct feature sets proposed in this previous study, only eight feature sets showed any correlation with protein antigenicity in preliminary experiments. The details of these sets are provided below. For consistency and convenience, the names assigned to each feature set in Magnum et al. (2009) are also used here.

Six of the eight feature sets are frequencies of amino acid monomers and dimers using three different amino acid alphabets described in Table 4. The six sets are denoted by [Name−5:XI] where Name−5 is the name given to the alphabet in Table 4, 5 is the size of the corresponding alphabet and X takes the value M or D associated with the frequencies of monomers and dimers over the corresponding alphabet (e.g. Hydropho-5:MI). The two remaining feature sets, [Computed] and [Predicted] are described below. Note that the set of predicted features [Predicted] was modified to include TMJMM (Kneif et al., 2001) predictions. Feature set sizes are reported in Table 3. Each feature is normalized to [−1,1] in the following experiments (Section 2.3).

- [Computed] Features directly computed from the sequence.
  (1) Sequence length $n$.
  (2) Turn-forming residues fraction: $\frac{|X|=|\psi|}{n}$, where, for instance, $N$ is the number of serine or threonine residues in the sequence.
  (3) Absolute charge per residue: $\frac{|R|=|\psi|}{n} - 0.33$.
  (4) Molecular weight.
  (5) GRAVY Index defined as the averaged hydrophathy value (Kyte and Doolittle, 1982) of the amino acids in the primary sequence.
  (6) Aliphatic index: $A = 2.9V + 3.9D + 3.9L/\pi(DK)$.
- [Predicted] Features predicted from the sequence.
  (1) Beta residues fraction, as predicted by SSpro (Cheng et al., 2005).
  (2) Alpha residues fraction, as predicted by SSpro.
Table 5. Evaluation of ANTIGENpro by repeated 10-fold cross-validations (results in bold)

<table>
<thead>
<tr>
<th>Method</th>
<th>Vaxijen*</th>
<th>ANTIGENpro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>59.48 ± 0.140</td>
<td>75.51 ± 0.992</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>89.69 ± 0.000</td>
<td>75.88 ± 1.937</td>
</tr>
<tr>
<td>Specificity</td>
<td>25.85 ± 0.742</td>
<td>75.14 ± 1.480</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.20 ± 0.008</td>
<td>0.51 ± 0.020</td>
</tr>
<tr>
<td></td>
<td>0.67 ± 0.006</td>
<td>0.81 ± 0.012</td>
</tr>
</tbody>
</table>

Vaxijen is also evaluated using the bacterial and viral antigens of the same datasets. Standard deviations appear in italic.

*Method proposed in Doytchinova and Flower (2007b), evaluated using the web server available at http://www.darellfloweronline/Vaxijen

Table 6. Evaluation of ANTIGENpro by repeated cross-validations over the six distinct protein sets

<table>
<thead>
<tr>
<th>Test set</th>
<th>Test set size</th>
<th>Training set size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAntigens</td>
<td>213*</td>
<td>726</td>
<td>81.60</td>
</tr>
<tr>
<td>Brucella</td>
<td>140</td>
<td>1012</td>
<td>78.00</td>
</tr>
<tr>
<td>Burkholderia</td>
<td>10</td>
<td>1142</td>
<td>66.00</td>
</tr>
<tr>
<td>Candida</td>
<td>6</td>
<td>1146</td>
<td>66.67</td>
</tr>
<tr>
<td>Malaria</td>
<td>228</td>
<td>924</td>
<td>59.96</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>342</td>
<td>754</td>
<td>68.30</td>
</tr>
</tbody>
</table>

* This set only contains positive examples of protective antigens. All other protein sets are balanced (50% positive examples, 50% negative examples).
3 RESULTS AND DISCUSSION

Here we present a three-pronged approach to assess the ANTIGENpro methodology. First, an internal validation is performed to determine if ANTIGENpro can discriminate the most antigenic from the least antigenic proteins in our curated data. For this purpose, ANTIGENpro is evaluated on the six protein sets described in Section 2.1 by repeated cross-validations using standard 10-fold cross-validation and dataset-fold cross-validation following the protocol described in Section 2.3.2. Results of the standard 10-fold cross-validations are reported in the last column of Table 5 and in Figure 1. Results for the dataset-fold cross-validation runs are reported in Table 6. These results clearly show that ANTIGENpro performs well at the internal discrimination task.

Next, a validation experiment is performed using the set PAntigens to assess ANTIGENpro’s ability to recognize confirmed protective antigens. For this experiment, ANTIGENpro is trained using only the microarray data and the accuracy on PAntigens is calculated. The results of this experiment indicate that a predictor trained solely on protein microarray data can predict truly protective antigens with accuracy that is significantly better than random.

Finally, armed with a classifier that we know can discriminate the most antigenic from the least antigenic proteins and can recognize confirmed protective antigens, we perform an external validation on an entire proteome to assess ANTIGENpro’s ability to recover likely protective antigens in a real-world setting. The Bartonella dataset, described in Section 2.1.3, is used for this purpose and the results of the experiment are reported in Tables 7 and 8. These results indicate that ANTIGENpro can be used effectively to identify likely protective antigens from an entire proteome.

3.1 Internal validation

The overall accuracy of ANTIGENpro, evaluated by repeated 10-fold cross-validations, is 75.51% with a prediction threshold of 0.5. The predictor correctly classifies 75.88% of the antigenic proteins and 75.14% of the non-antigenic proteins. The Matthews correlation coefficient (MCC in Table 5) is 0.51 and the area under the ROC curve (AUC) is 0.81. These results indicate that the prediction model has learned relevant characteristics of the most antigenic versus least antigenic proteins. Finally, the small SDs over the five runs attest to the stability of the method.

The evaluation of the Vaxijen predictor (Doytchinova and Flower, 2007b) on the datasets used for the 10-fold cross-validation experiments discussed above (results in Table 5) shows that this predictor is outperformed by ANTIGENpro. In fact, Vaxijen correctly classified 59.48% of the bacterial and viral proteins in these datasets. The other antigens (around 25% of the antigens in our dataset) cannot be tested since no prediction model is available for these pathogens. This is a clear shortcoming of Vaxijen. In addition, the Vaxijen sensitivity (89.69%) and specificity (25.85%) show an important bias toward positive predictions. Using higher prediction thresholds for this predictor drops the sensitivity and does not improve the overall accuracy. The other evaluation criteria and the ROC curve (Fig. 1) also tend to indicate that ANTIGENpro is more suitable than Vaxijen for both two-class prediction and ranking.

<table>
<thead>
<tr>
<th>Method</th>
<th>ANTIGENpro</th>
<th>SignalP</th>
<th>Vaxijen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top ranked 2%</td>
<td>5.5</td>
<td>1.7</td>
<td>2.1</td>
</tr>
<tr>
<td>Top ranked 5%</td>
<td>4.4</td>
<td>2.7</td>
<td>1.6</td>
</tr>
<tr>
<td>Top ranked 10%</td>
<td>3.4</td>
<td>2.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Top ranked 25%</td>
<td>2.1</td>
<td>2.2</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Note that the dataset is unbalanced. Bartonella contains 73 antigenic proteins and 1390 non-antigenic proteins. The expected enrichment of a random ranking is 1.0.

3.2 Known protective antigens: PAntigens

During the standard 10-fold cross-validation experiments, we observe that nearly all of the protective antigens in PAntigens are correctly classified (around 90%). In addition, when ANTIGENpro is trained using the five microarray datasets only, the predictor correctly classifies 81.6% of the protective antigens in PAntigens using an unbiased decision threshold of 0.5 (first line of Table 6). The high accuracy of ANTIGENpro on PAntigens supports the working hypothesis that protein microarray data can be used to predict the likelihood that a protein is a protective antigen. This result is significant because new alignment-independent methods are required for the discovery of truly novel antigens, and our data preparation protocol ensures that none of the sequences in the microarray datasets is homologous with any sequence in PAntigens.

3.3 External proteome: Bartonella

Here we assess ANTIGENpro’s ability to recover the most antigenic proteins from an entire proteome using the external dataset Bartonella. For this test every protein in the dataset must be classified as either antigenic or non-antigenic, resulting in an
unbalanced set consisting of only 5% positive examples. This situation approximates how the method can be applied, and how it may perform on a new proteome in a real-world situation.

Specifically, the scores produced by computational methods can be used to screen proteomes for subsets of proteins for further testing. For this type of application, the enrichment among the top ranked proteins is a useful metric. The enrichment is calculated as: (\% of positives among top ranked subset)/(% of positives among the entire proteome), thus the expected enrichment of a random ranking would be 1.0. For instance, among the proteins ranked in the top 10% by ANTIGENpro on *Bartonella*, 17.0% of the antigenic proteins are recovered while only 5% of the entire proteome is classified as antigenic, this corresponds to a 3.4-fold enrichment.

The probability estimates produced by ANTIGENpro are used to rank all the proteins in the proteome. Similarly, the ranking process is repeated using the scores produced by Vaxijen and SignalP. SignalP predicts the likelihood a protein contains a signal sequence, which is one criteria that is frequently used to screen for potential antigens. The enrichment results for ANTIGENpro, Vaxijen and SignalP are presented in Table 7 using the top 2, 5, 10 and 25% ranked proteins. Using the 2, 5 and 10% thresholds, the ANTIGENpro enrichment values are higher than for both SignalP and Vaxijen. At the 25% threshold, the SignalP result of 2.2 is slightly higher than the ANTIGENpro result of 2.1. Overall, the enrichment results indicate that ANTIGENpro can be used effectively to screen for likely protective antigens for further analysis or experiments.

The two-class prediction accuracies on the external dataset are reported in Table 8. These results show that using a threshold of 0.5 to decide antigenicity yields a low accuracy of 55.64%. However, using higher prediction thresholds results in a significant increase in the accuracy of the predictor while the decrease in sensitivity is less significant. This is clearly a desirable characteristic of the predictor. In fact, when the threshold is set to 0.65, the accuracy is 70.54% significant. This is clearly a desirable characteristic of the predictor. The results obtained during the evaluation experiments provide several interesting conclusions. First, despite the inherent noise in protein microarray data, it can be used to effectively categorize both antigen and non-antigenic proteins for training purposes.

The results demonstrate that this source of data can be used effectively for this problem: 81.6% of the known protective antigens are correctly classified when the model is trained only on balanced protein microarray datasets and an unbiased decision threshold of 0.5 is used. In addition, the cross-validation results indicate that ANTIGENpro can predict protein antigenicity from sequence alone with accuracy that is significantly better than random (75.51%). These cross-validation results also show that it significantly outperforms Vaxijen, the only previously reported method that does not directly rely on homology to known protective antigens.

The results on the external validation dataset demonstrate that ANTIGENpro performs well when ranking entire proteomes according to likely antigenicity, when compared with Vaxijen and SignalP. In addition, the method presented here can rapidly take advantage of the massive amount of data that will be available thanks to the high-throughput protein microarray technology developed for studying the humoral immune response to pathogen infections. ANTIGENpro is available online at http://scratch.protomics.ics.uci.edu.

Funding: Work supported by NIH Biomedical Informatics Training Grant (LM-07443-01), NSF MRI grant (EIA-0321390), NSF grant 0513376, and a Microsoft Faculty Research Award (to P.B.). Original collection of the data was in part supported by NIH grants U01AI078213 and U54AI063539 (to PF.).

Conflict of Interest: none declared.

REFERENCES

Acelryx Software Inc, formerly Genetics Computer Group Inc (GCG). 10188 Telesis Court, Suite 100; San Diego, CA 92121, USA.


Prediction of protein antigenicity


