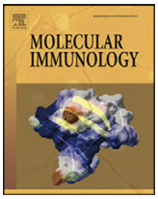
Molecular Immunology 46 (2009) 840–847

Contents lists available at ScienceDirect

Molecular Immunology

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



A machine-learning approach for predicting B-cell epitopes

Nimrod D. Rubinstein, Itay Mayrose, Tal Pupko*

Department of Cell Research and Immunology, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 25 May 2008
Received in revised form 23 July 2008
Accepted 7 September 2008
Available online 22 October 2008

Keywords:
Antigen
Antibody
Epitope
Immunogenic regions
Prediction
Properties
Surface

ABSTRACT

The immune activity of an antibody is directed against a specific region on its target antigen known as the epitope. Numerous immunodetection and immunotheraputics applications are based on the ability of antibodies to recognize epitopes. The detection of immunogenic regions is often an essential step in these applications. The experimental approaches used for detecting immunogenic regions are often laborious and resource-intensive. Thus, computational methods for the prediction of immunogenic regions alleviate this drawback by guiding the experimental procedures. In this work we developed a computational method for the prediction of immunogenic regions from either the protein three-dimensional structure or sequence when the structure is unavailable. The method implements a machine-learning algorithm that was trained to recognize immunogenic patterns based on a large benchmark dataset of validated epitopes derived from antigen structures and sequences. We compare our method to other available tools that perform the same task and show that it outperforms them.

1. Introduction

The ability of an antibody to specifically bind an antigen is used in various biomedical applications ranging from immunodetection to immunotheraputics (Irving et al., 2001; Westwood and Hay, 2001). In many such applications it is required to computationally predict protein regions with the highest potential to elicit antibodies that will strongly bind the intact protein. This task is also important for epitope-mapping (Westwood and Hay, 2001). In epitope-mapping, a phage-display library is initially scanned with the antibody; following that, the affinity selected peptides need to be mapped onto the antigen structure in order to infer the exact location of the epitope (Castrignano et al., 2007; Enshell-Seijffers et al., 2003; Halperin et al., 2003; Mayrose et al., 2007; Moreau et al., 2006; Schreiber et al., 2005). Predicting immunogenic regions can focus the mapping of the affinity-selected peptides to relevant regions on the antigen, and thus to increase the accuracy of this approach.

Several computational methods were developed for the task of predicting the most immunogenic regions of a given antigen (Emini et al., 1985; Haste Andersen et al., 2006; Hopp and Woods, 1981; Karplus and Schulz, 1985; Kolaskar and Tongaonkar, 1990; Kulkarni-Kale et al., 2005; Parker et al., 1986; Pellequer et al., 1991). Most of these methods are sequence-based: a score drawn from a propensity scale is assigned to each amino-acid. The antigen sequence is then scanned for high scoring segments, which are inferred as the candidate epitopes. Different propensity scales were suggested for this task, each reflecting a certain amino-acid physico-chemical property, e.g., hydrophilicity or backbone-flexibility. These scales were selected based on the premise that they are correlated with antigenicity. Although this approach is commonly used and has been reported to be successful to some extent, it was criticized since the correlations of the propensity scales with peaks of epitope locations are limited, and thus the predictions are, on average, only marginally better than random (reviewed in Blythe and Flower, 2005).

When the 3D structure of the antigen is available or can be reliably predicted, this information can be used to increase the accuracy of predicting immunogenic regions. For example, it is clear that immunogenic regions reside on the solvent accessible surface of the antigen. This property was used by Novotny et al. (1986), and by Kulkarni-Kale et al. (2005) who developed the Conformational Epitope Prediction (CEP) server, which searches for regions that are highly accessible. Haste Andersen et al. (2006) developed DiscoTope, which in addition to solvent accessibility uses in its prediction algorithm a propensity scale that reflects the observation that the distribution of amino-acids in epitopes varies from that of the remaining antigen. While these structural and physico-chemical properties are clearly correlated with immunogenic regions, it is

Abbreviations: ASA, accessible surface area; AUC, area under the curve; CDR, complementarity determining region; CEP, Conformational Epitope Prediction; PDB, protein data bank; ROC, receiver operating characteristic; 3D, 3-dimensional.

* Corresponding author. Tel.: +972 3 640 7693; fax: +972 3 642 2046.
E-mail address: talp@post.tau.ac.il (T. Pupko).

0161-5890/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.molimm.2008.09.009

now established that additional attributes characterize epitopes (Jones and Thornton, 1997; Rubinstein et al., 2008). Accounting for such attributes can thus boost the accuracy of algorithms for prediction of immunogenic regions. Ponomarenko and Bourne (2007) assessed the success of several 3D structure-based protein–protein binding site prediction methods (including CEP and DiscoTope), at predicting immunogenic regions. The performance of all methods was found to be mediocre, and it was hence concluded that utilizing additional features that characterize epitopes is the key for improvement.

We have recently performed a detailed computational analysis of all non-redundant antibody–antigen complexes available in the protein data bank (PDB, Berman et al., 2000), in order to reveal the specific characteristics of epitopes (Rubinstein et al., 2008). This study delineated a range of physico-chemical, structural, and geometrical properties that significantly distinguish epitopes from the remaining antigen surface. Epitopes were found to have a unique amino-acid composition, enriched with tyrosine and tryptophan residues. A strong preference for unorganized secondary structures in epitopes was also observed. Moreover, epitopes were found to display a distinct geometrical shape, with a rugged surface that resides on bulgy regions of the antigen. Interestingly, epitopes were found to be less evolutionary conserved relative to the remaining antigen surface.

Determining the major characteristics of antigenicity is the first and critical step towards predicting epitopes from antigen structures. The challenge in the next step is to utilize these characteristics in an optimal way to produce accurate predictions of immunogenic regions. In this work we have applied a machine-learning approach for predicting such regions that are candidate epitopes. We first constructed a large dataset composed of antigen structures, for which validated epitopes are available. We next trained a classifier for the prediction task, and tested its performance using the same data applying a cross-validation procedure for avoiding over-fitting the algorithm to the data. Often, the antigen 3D structure is unavailable. To predict immunogenic regions from sequence alone we repeated the above process of constructing a sequence benchmark dataset, selecting immunogenic properties relevant to sequences, training the classifier, and testing its performance. We show that our novel algorithms accurately predict immunogenic regions. Moreover, we show that they outperform other available structure and sequence-based tools for the same task.

2. Methods

2.1. Algorithm outline

The underlying assumption in this work is that epitope and non-epitope parts of an antigen surface are distinct with respect to their physico-chemical and structural–geometrical properties. We thus trained two Naïve Bayes classifiers, one for structures and one for sequences, to recognize immunogenic regions based on a large set of physico-chemical and structural–geometrical properties. A trained classifier computes for each region of a given input antigen structure or sequence a score that reflects its immunogenic potential. Specifically, the input antigen is divided into overlapping surface patches (for a 3D structure) or stretches (for a sequence), with the size of a typical epitope. Then for each patch or stretch, the trained classifier computes the probability that it is drawn from a population of epitopes, given its physico-chemical and structural–geometrical properties. The score of each patch (or stretch) is assigned to its central residue (the middle residue in a circle-shaped patch, or the middle residue in a linear stretch), which enables the inference of the immunogenic potential at the single amino-acid site resolution. Fig. 1 illustrates the flow of the algorithm for an input antigen structure.

In the sections below we first provide the formal definition of a patch, we then explain how the properties were chosen, and proceed with a description on how these properties are combined to

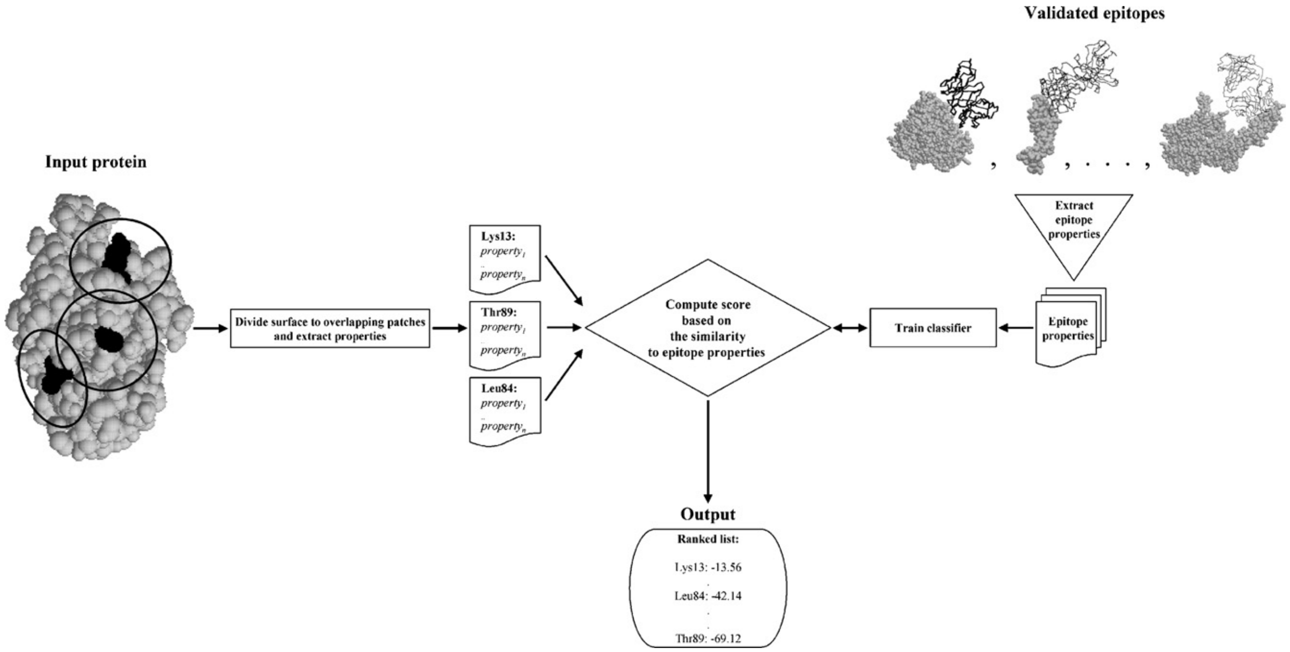


Fig. 1. Illustration of the flow of the prediction algorithm. A dataset of antibody–antigen co-crystal structures is used to derive epitopes and extract their physico-chemical and structural–geometrical properties. This collection of epitope properties is then used to train the classifier. Given an input protein structure for which immunogenic regions are sought, its surface is divided to overlapping circular patches the size of an average epitope, centered on each of the surface residues. The physico-chemical and structural–geometrical properties are extracted for each such patch. The trained classifier then computes for each patch a score that reflects its immunogenic potential based on the similarity of its properties to pre-characterized epitope properties. This score is expressed in log-likelihood terms and is thus negative. Finally, these scores are assigned to the residues on which each patch was centered, and the output is a list of residues and their scores sorted in descending order.

achieve maximal predictive power. We first describe this process when a protein 3D structure is available.

2.2. Patch definition

Formally, a surface patch was defined as the group of $n-1$ surface residues with the shortest distance to a central surface residue, and the central surface residue itself. We used $n = 20$, representing the average size of an epitope (Rubinstein et al., 2008). The distance between two residues was defined as the minimal Euclidean distance between the centers of any of their solvent-exposed non-hydrogen atoms. Any residue was defined as a surface residue if its accessible surface area (ASA), computed using the Surface Racer program (Tsodikov et al., 2002) with a probe radius of 1.4 Å, exceeded 5% of its maximal (theoretical) ASA (i.e., relative ASA > 0.05). The maximal ASA value of a residue was calculated in an extended GXG theoretical tripeptide, where G denotes glycine and X denotes the residue in question (Miller et al., 1987).

2.3. Physico-chemical and structural–geometrical properties

In order to train a classifier one must select properties and combine them to achieve optimal predictive power. The physico-chemical and structural–geometrical signature that distinguishes epitope from non-epitope surfaces (Jones and Thornton, 1997; Rubinstein et al., 2008) is not necessarily the set of properties that is optimal for the prediction task. We thus selected a wider set, comprised of 44 physico-chemical and structural–geometrical properties for the prediction task (Table 1). Of these, 28 properties were previously tested for their ability to distinguish epitope from non-epitope surfaces (Rubinstein et al., 2008): the ratio between the frequency of each of the 20 amino-acids in the patch and the remaining surface (properties 1–20); the ratio between the frequency of each of the main secondary-structure elements (helix, beta-strand, and loop, obtained according to the dictionary of secondary structure of proteins (Kabsch and Sander, 1983)) (properties 21–23); a patch's average relative solvent accessibility and average accessibility to a large probe (with radius = 9 Å, approximating a CDR), which were computed using the Surface Racer program (properties 24 and 25); the average curvature of the patch atoms (atom level geometrical shape), which was also computed using the Surface Racer program (property 26); the fraction of patch atoms that are within a distance of 4 Å from the convex hull of the protein structure (patch level geometrical shape), which was constructed using the computational geometry algorithms library (http://www.cgal.org) (property 27); and the average evolutionary conservation of the patch, which was computed using our Bayesian estimation method (Mayrose et al., 2004) (property 28). The remaining 16 properties are the average score of a patch residues according to amino-acid propensity scales, most of which have been previously used to predict epitopes: solvent accessibility (Emini et al., 1985) (property 29); exposed residues (Janin and Wodak, 1978) (property 30); composition (Grantham, 1974) (property 31); polarity (Grantham, 1974) (property 32); molecular volume (Grantham, 1974) (property 33); hydrophilicity (Hopp and Woods, 1981) (property 34); flexibility (Karplus and Schulz, 1985) (property 35); antigenicity (Kolaskar and Tongaonkar, 1990) (property 36); hydrophilicity (Parker et al., 1986) (property 37); beta-turns (Pellequer et al., 1993) (property 38); polarity (Ponnuswamy et al., 1980) (property 39); and a set of five scales (factor1–5) that were found to summarize ~500 different propensity scales using a multivariate statistical analysis (Atchley et al., 2005) (properties 40–44).

2.4. Scoring a patch based on its immunogenic properties

The model for computing the immunogenic score of a patch is the Naïve Bayes classifier, which follows the Bayes theorem:

$$P(\text{patch} = E|ip_1, \ldots, ip_n) = \frac{P(\text{patch})P(ip_1, \ldots, ip_n|E)}{P(ip_1, \ldots, ip_n)}$$

E denotes epitope and ip_i denotes immunogenic property i. Since $P(ip_1, \ldots, ip_n)_,$ is constant and all patches have the same prior probability P(patch), only the term: $P(ip_{1,} \ldots, ip_n|E)$ determines the immunogenic score. The Naïve Bayes classifier assumes conditional independence between the properties. Thus, the score computed

Table 1
Immunogenic properties for structure-based prediction

Property number	Immunogenic property	Included in the optimal set of properties
1–20	Ratio between the frequency of each of the 20 amino-acids: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, in the patch and the remaining surface	A, R, C, H, L, F, S, T, V
21–23	Ratio between the frequency of the secondary structure elements: helices, beta-strands, and loops, in the patch and the remaining surface	Helices
24	The average relative-accessibility of the patch to the solvent (probe radius = 1.4 Å)	+
25	The average accessibility of the patch to a large probe (radius = 9 Å) approximating an antibody CDR	+
26	The average curvature of the patch atoms	+
27	The proportion of patch atoms that reside within 4 Å from convex hull of the antigen	+
28	Average evolutionary rate of the patch	
29	Solvent accessibility scale	
30	Exposed residues scale	
31	Amino-acid composition scale	
32	Amino-acid polarity scale	
33	Amino-acid molecular volume scale	
34	Amino-acid hydrophilicity scale	
35	Amino-acid flexibility scale	
36	Amino-acid antigenicity scale	
37	Amino-acid hydrophilicity scale	
38	Beta-turns scale	
39	Amino-acid polarity scale	
40	Factor1 scale	
41	Factor2 scale	
42	Factor3 scale	
43	Factor4 scale	
44	Factor5 scale	

for the examined patch is:

$$P(ip_1, \ldots, ip_n|E) = \prod_{i=1}^{n} P(ip_i|E).$$

$P(ip_i|E)$ is an abbreviated expression for the probability of observing a specific value of immunogenic property i given that the patch is an epitope. These expressions are computed based on the training stage of the algorithm (described below).

2.5. Training the classifier

The value of each immunogenic property was measured for all epitopes in the train dataset (e.g., for each structure the frequency of tyrosine residues in each epitope divided by their frequency in the corresponding non-epitope surface). Then, for each property these values were binned in order to construct a discrete histogram. The histogram was initially subjected to the Grubbs test for removing outliers (Grubbs, 1969). Following that, the number of equally sized bins in each histogram was determined. To this end, we require that a histogram will be unimodal. This requirement stems from our belief that multiple local maxima are an artifact that is created because the train data do not contain enough observations to reliably populate all bins. Thus, the binning process begins by selecting an initial high number of bins, which is iteratively reduced as long as the histogram contains multiple maxima. Each such histogram is thus used to compute the probability that a certain patch is drawn from a population of epitopes given its corresponding immunogenic property. For example, if a certain patch is evaluated for its immunogenic potential according to the ratio of the frequency of tyrosines in it versus the corresponding remaining surface. The observed value of this property for that patch is 1.25 and it corresponds to a histogram bin with probability = 0.5. Then, $P(ip_{\text{tyrosine}} = 1.25|E)$=0.5 for that patch.

2.6. Benchmark dataset

We compiled a benchmark dataset of antigens for which the epitope is reliably determined. This dataset was used both for training and testing the algorithm. Each dataset member is a protein antigen, for which the epitope (or epitopes) is reliably determined. Training and testing an algorithm on the same data can significantly bias the results, due to over-fitting of the algorithm to the data. To avoid this, we used a cross-validation procedure in which one part of the data is used for training and the remaining part—for testing. This procedure is repeated several times, each time with a different part of the data serving as the train set and the remaining part as the test set. The overall performance is eventually computed as the average performance over all data partitions to train and test sets. In this work, we implemented the leave-one-out cross-validation approach, which is suitable for small datasets (Cawley, 2006). Thus, in each of the n cross-validation iterations (n is the size of the dataset) the algorithm is trained on all but one of the dataset members, and tested on this left-out dataset member.

For the structure dataset, all available antibody–antigen co-crystal structures were retrieved from the SPIN server of protein–protein complexes (http://trantor.bioc.columbia.edu/cgi-bin/SPIN/). This dataset was then subjected to a filtering process as described in Rubinstein et al. (2008). This produced a dataset of 49 co-crystal structures. Dataset members of antigens that were co-crystallized with different antibodies were united. Consequently, 47 non-redundant co-crystal structures of antibody–antigen complexes were retained. During completion of this work Ponomarenko and Bourne (2007) published a benchmark dataset of epitopes also inferred from 3D structures of antibody–antigen complexes. As both datasets share approximately 90% of the structures we decided to use only the dataset constructed by us. Validated epitope residues were defined as those for which at least one exposed atom was found to be in contact with the antibody. This was determined from the antibody–antigen complex, using the Contacts of Structural Units program (Sobolev et al., 1999).

A non-redundant set of validated linear epitopes was obtained from the Bcipep database (Saha et al., 2005). Only epitopes for which the antigen sequence is available in the NCBI protein database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=Protein) were retained from this list. In case a certain antigen sequence had more than one epitope, these epitopes were united. This resulted in 194 antigen sequences. The datasets can be found at: http://www.tau.ac.il/∼talp/EpitopePrediction.

2.7. Performance evaluation

In the algorithm training stage, each antigen residue in both the structure and the sequence datasets is regarded as either a validated "epitope residue" or a "non-epitope residue". As described above, the algorithm computes a score for each residue, which reflects its immunogenic potential. Intuitively, in a successful prediction, "epitope residues" should be scored higher than the average over all residues. We thus considered a prediction to be successful if the average score of "epitope residues" exceeds the average score over all residues. Hence, the prediction for each dataset member is regarded as either successful or not. This approach was applied to evaluate the performance of our prediction method, and for comparison to other prediction methods that also assign a score to each residue.

We also report the area under the receiver operating characteristic (ROC) curve (AUC) (Fawcett, 2006). Although the AUC is traditionally used for diagnosing the performance of classification models (e.g., Ponomarenko and Bourne, 2007 used the AUC measure for evaluating several extant epitope prediction methods), we note that it is somewhat inadequate when it comes to assessing the performance of epitope prediction methods. This is because the ROC analysis considers predictions that are not part of any validated epitope as false predictions. However, since the antigens in our data possibly contain a far larger number of epitopes than are currently known, the AUC underestimates the actual predictive power of the classifier.

2.8. Feature selection

Among the 44 immunogenic properties (Table 1), not all necessarily contribute to the prediction of epitopes. Moreover, while many properties can independently contribute to the prediction task, some may be highly correlated and such redundancy may actually prevent reaching optimal predictive power. Hence, a subset of these properties may produce optimal predictive power. To this end, an exhaustive search over all possible combinations of immunogenic properties should be performed. However, due to the large number of properties, this option is computationally unfeasible. Therefore, we applied a top-down heuristic search for this task. Starting with all 44 immunogenic properties, this set was iteratively reduced until only a single property remained (to prevent premature convergence to a suboptimal set of properties). The immunogenic property deleted at each iteration was the one for which its deletion had the least effect on the number of successful predictions (as defined above), computed for the entire test dataset, using the leave-one-out cross-validation procedure. Finally, the set of properties that resulted with the highest number of successful predictions was selected as the optimal set.

2.9. Applying the algorithm to antigen sequences

When only the sequence of the antigen is available, a few modifications for the prediction algorithm are applied. Whereas for an antigen structure each residue is scored based on a patch surrounding it, for an antigen sequence each residue is scored based on a segment of $n - 1$ consecutive residues (stretch) flanking it, and the center residue itself. We used $n = 7$, which was found to be the average length of an epitope segment in the sequence dataset. The three residues at the N and C terminals of the antigen sequence do not define any stretch of 7 residues and hence are given the same score as the first and last stretches in the sequence, respectively.

Some of the immunogenic properties derived from antigen structures are irrelevant for antigen sequences. These include for example, structural–geometrical properties that cannot be derived from the sequence. Nevertheless, a few structural features can be reliability predicted from the protein sequence (Rost et al., 2004). We thus defined 41 immunogenic properties (Table 2): the ratio between the frequency of each of the 20 amino-acids in the stretch and the corresponding remaining sequence (properties 1–20); the ratio between the frequency of each of the main secondary-structure elements (helix, beta-strand, and loop) in the stretch, and the corresponding remaining sequence, which were obtained using the PredictProtein program (Rost et al., 2004) (properties 21–23); the predicted average relative solvent accessibility (Rost et al., 2004) (property 24); the average evolutionary conservation of the stretch (Mayrose et al., 2004) (property 25); and the same 16 propensity scales used for structure-based prediction (properties 26–41). Using the same performance assessment method as for the structure-based algorithm, we also applied the top-down heuristic search to select the optimal set of immunogenic properties for sequence-based prediction.

2.10. Program availability

The algorithms described in this work are implemented in C++. The obligatory inputs for the structure-based algorithm are the PDB file of the antigen and the corresponding required chains. For the sequence-based algorithm, the obligatory input is a Fasta format file of the antigen sequence. The executable and accompanying scripts are available at: http://www.tau.ac.il/~talp/EpitopePrediction.

3. Results

3.1. Prediction performance

For evaluating the performance of our algorithm we applied the leave-one-out cross-validation procedure both for the structure and sequence predictions and computed the resulting success rate. When the entire set of 44 properties (Table 1) was used for prediction of immunogenic regions of the structure dataset, the number of successful predictions amounted to 33 out of 47 dataset members (70.3%). For prediction of immunogenic regions of the sequence dataset, the number of successful predictions, using the entire set of 41 properties (Table 2), amounted to 137 out of 194 dataset members (70.6%). Naturally, one would expect a higher success rate for the structure-based prediction given that the corresponding classifier is provided with a richer set of properties (structural–geometrical properties that are unavailable for sequences). Nevertheless, as we show below, this is probably the result of dilution of the structural signal with many sequence related properties, which constitute the majority among the set of 44 properties used to obtain the above results.

3.2. Determining the optimal set of immunogenic properties

Using an exceedingly large set of properties may result with suboptimal predictive power since the inclusion of several properties contributes more noise than signal. Thus, a feature selection procedure was applied to obtain the set of properties that maximizes the algorithm's predictive power in terms of success rate. For the structure-based prediction, the feature selection procedure reduced the number of properties from 44 to 14 (Table 1), and significantly improved the success rate from 33 to 43, out of 47 dataset members (from 70.3% to 91.4%, $P = 0.04$; G-test). Although the feature selection was not devised for optimizing the AUC measure, we note that this procedure also significantly increased the average AUC from 0.6 to 0.65 ($P < 10^{-20}$; paired t-test). Interestingly, the optimal set of properties included both structural–geometrical properties (e.g., the average curvature of the patch atoms) as well as physico-chemical properties (e.g., the frequency ratios between patch and remaining antigen surface of valine residues). In contrast, none of the propensity scales (properties 29–44,

Table 2
Immunogenic properties for sequence-based prediction

Property number	Immunogenic property	Included in the optimal set of properties
1–20	Ratio between the frequency of each of the 20 amino-acids: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, in the stretch and the remaining sequence	N, Q, L, K, M, S, T, V
21–23	Ratio between the frequency of the secondary structure elements: helices, beta-strands, and loops, in the stretch and the remaining sequence	
24	The average relative-accessibility of the stretch to the solvent	
25	Average evolutionary rate of the stretch	
26	Solvent accessibility scale	
27	Exposed residues scale	
28	Amino-acid composition scale	
29	Amino-acid polarity scale	+
30	Amino-acid molecular volume scale	
31	Amino-acid hydrophilicity scale	
32	Amino-acid flexibility scale	+
33	Amino-acid antigenicity scale	+
34	Amino-acid hydrophilicity scale	+
35	Beta-turns scale	
36	Amino-acid polarity scale	+
37	Factor1 scale	+
38	Factor2 scale	
39	Factor3 scale	
40	Factor4 scale	
41	Factor5 scale	

Table 1) were retained in the optimal set of immunogenic properties.

For the sequence-based prediction, the feature selection procedure reduced the number of properties from 41 to 14, and increased the success rate from 137 to 156, out of 194 dataset members (from 70.6% to 80.4%, $P = 0.1$; *G*-test). The corresponding improvement in average AUC was from 0.55 to 0.59 ($P < 10^{-20}$; paired *t*-test). In contrast to the structure-based prediction, none of the predicted structural properties were retained in the optimal set of properties (Table 2), however five propensity scales did remain in this set. We hypothesize that this stems from the fact that at least part of the epitopes in the sequence data were defined as such since they manifest peak scores using different propensity scales.

A comparison between the structure and sequence based predictions shows that the structure-based prediction has greater predictive power (91.4% versus 80.4%, respectively, $P < 10^{-4}$; *G*-test). Since these results are approximately optimal, it can now be reasoned that this difference stems from the fact that the classifier for the structure-based prediction is provided with a richer set of properties.

Two predictions of the structure-based algorithm are presented in Fig. 2. Fig. 2A presents a typical prediction of the algorithm applied to the interferon-gamma receptor alpha chain (PDB identifier 1jrh, Sogabe et al., 1997), co-crystallized with its binding antibody. The largest predicted highly immunogenic region on the surface of this antigen considerably overlaps the validated epitope. Several other highly immunogenic regions are predicted, yet they are much smaller and thus do not seem to be good epitope candidates. Fig. 2B presents a very successful prediction applied to the vascular endothelial growth factor (PDB identifier 1bj1, Muller et al., 1998), also co-crystallized with its binding antibody. In this case, 16 out of the 17 validated epitope residues are within the most immunogenic region predicted by our algorithm. As can be seen in the figure, the least immunogenic region on the surface of this antigen is very bulgy, which is a favorable immunogenic property. Nevertheless, its amino-acid composition is immunogenically unfavorable. Thus, when all properties are considered this region receives a low immunogenic score.

3.3. Comparison with other methods

Propensity scales have been traditionally used for predicting immunogenic regions in antigens. In this methodology, each antigen residue is assigned its corresponding score from the propensity scale and the antigen is then scanned for high scoring regions. Since both our method and the propensity-scale-based method assign a score for each residue in the antigen reflecting its immunogenic tendency, we were able to compare the two. For the structural dataset, out of all 16-propensity scales, which were evaluated independently, those that produced the highest success rate, amounting to 29 out of 47 dataset members (61.7% and a corresponding average AUC of 0.57), are factor2 and factor5 (properties 41 and 44, respectively, Table 1). Each of these propensity scales summarizes an independent structural-functional aspect of nearly 500 different propensity scales (Atchley et al., 2005), where factor2 corresponds to secondary structure propensity and factor5 corresponds to electrostatic charge propensity. To the best of our knowledge, these propensity scales have not been used for epitope prediction. Among the propensity scales that have been traditionally used for epitope prediction, the solvent accessibility scale (Emini et al., 1985) (property 29, Table 1) was the most successful, producing a success rate of 28 out of 47 dataset members (59.5%) and a corresponding average AUC of 0.55. Since our method obtained a significantly higher success rate (91.4% with AUC of 0.65, $P = 0.004$ and 0.0007; *G*-test, and paired *t*-test, respectively), it can be concluded that our methodology along with the suggested optimal set of immunogenic properties are much more appropriate for predicting immunogenic regions in protein structures.

We repeated the above comparison, this time with our sequence-based prediction algorithm applied to the sequence dataset. The two propensity scales that produced the highest success rate, amounting to 132 out of 194 dataset members (68% with corresponding average AUCs of 0.57), are two different polarity scales (Grantham, 1974; Ponnuswamy et al., 1980) (properties 29 and 36, respectively, Table 2). These two propensity scales are among the five propensity scales that were retained in the feature-selection procedure applied to the sequence-based algorithm (Table 2). Notably, the predictive power of the propensity-scale-

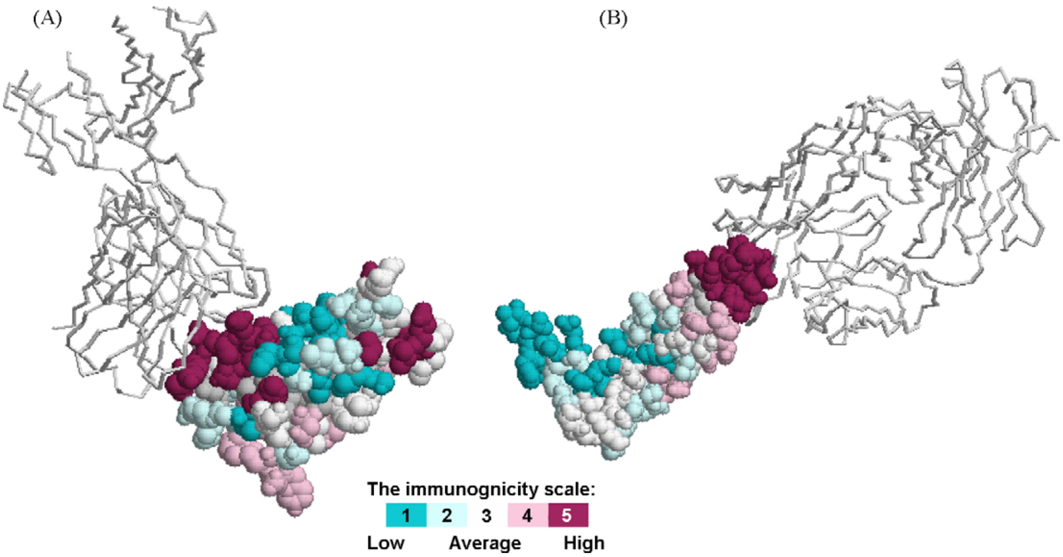


Fig. 2. Visualization of two predictions of the structure-based algorithm. The antigens and antibodies are represented in a space-fill and backbone models, respectively. Each of the antigen surface residues is color-coded according to its immunogenic level, defined in the color legend. Non-surface residues are colored gray. (A) Visualization of the prediction of immunogenic regions for interferon-gamma receptor alpha chain, co-crystallized with its binding antibody (PDB identifier: 1jrh). (B) Visualization of the prediction of immunogenic regions for vascular endothelial growth factor, co-crystallized with its binding antibody (PDB identifier: 1bj1).

based methodology is higher when applied to the sequence dataset than to the structural dataset (68% versus 61.7%, $P = 0.06$; *G*-test). This strengthens the notion that the epitopes of the sequence data were defined as such based on their peak scores computed using different propensity scales. In conclusion, this comparison suggests that relying on a set of properties rather than a single propensity scale results with higher predictive power.

We further compared our method to several additional methods that perform the same task, which do not rely on a single propensity scale. For the structure-based comparison we chose CEP (Kulkarni-Kale et al., 2005) and DiscoTope (Haste Andersen et al., 2006). In brief, CEP locates linear stretches of residues on the antigen structure that are highly accessible to the solvent, and groups them into patches if the spatial distance between them is below a certain threshold. These patches are then regarded as candidate immunogenic regions. DiscoTope on the other hand, operates in a mode more similar to the method described here. That is, it computes a score for each residue of the antigen, albeit, using fewer properties and without applying a machine-learning methodology. As the candidate immunogenic regions predicted by CEP are not computed a score, our performance evaluation method is inapplicable. We thus decided to compare our method with CEP using the AUC measure. We followed the methodology described in Ponomarenko and Bourne (2007) for computing the AUC values for CEP predictions. The average AUC for 45 out of the 47 dataset members, for which a CEP prediction could be obtained, was 0.52. For the same 45 dataset members, the average AUC obtained by our method was 0.65 ($P < 10^{-8}$; paired *t*-test). The success rate, defined in this work, of the DiscoTope method was found to be 40 out of 47 dataset members (85.1%), with a corresponding average AUC of 0.56. For the same dataset our method succeeded in 43 out of 47 dataset members (91.4%) with a corresponding average AUC of 0.65. We note that although the difference in average AUCs is statistically significant ($P < 10^{-8}$; paired *t*-test), the difference in success rates is not ($P = 0.47$; *G*-test). Still, it should be reminded that the leave-one-out cross-validation procedure was only applied to evaluate the performance of our method; hence, this comparison may be biased in favor of DiscoTope.

For comparison to our sequence-based algorithm we chose ABCpred (Saha and Raghava, 2006). ABCpred uses a recurrent neural network scheme for predicting continuous epitopes for protein sequences. In their work, the authors show that ABCpred outperforms other extant sequence-based methods. In addition, we found ABCpred most appropriate for comparison since it was trained and evaluated on essentially the same data used in this work. This comparison revealed that our method significantly outperforms ABCpred both in terms of success rate: 156 successful predictions versus 123 successful predictions, out of 194 dataset members, respectively (80.4% versus 63.4%, respectively, $P < 10^{-06}$; *G*-test), and in average AUCs: 0.59 versus 0.42, respectively ($P < 10^{-32}$; paired *t*-test).

4. Discussion

The problem of predicting immunogenic regions is one of the oldest (e.g., Arnon and Sela, 1969) and most challenging in immunoformatics (Ponomarenko and Bourne, 2007). In early methods, the extent of available data upon which they were developed, was extremely limited and thus their premises and performance could not be thoroughly assessed. Although the availability of epitope data experienced a sharp incline in the recent decade, current methods do not fully utilize the large number of features that characterize epitopes within a robust inference framework. The observation, which motivated this work was that the problem of epitope prediction is in fact a classical classification problem, and should be tackled as such with the rich methodology of machine-learning already in hand.

In their evaluation of epitope prediction methods, Ponomarenko and Bourne (2007) concluded that current methodologies perform poorly. This conclusion stemmed from the mediocre AUC values evaluated by the authors. Indeed, the use of the AUC as a performance measure is fully justified when each prediction can be accurately ascribed as true or false. However, such a reliable classification cannot be achieved for the current tested data, and it is reasonable to claim that these data contain a far larger number of epitopes than are currently known. For this reason, the resulting AUC values are an underestimate of the actual predictive power of epitope detection methods. The performance evaluation method defined in this work is free from this limitation. This method simply examines whether validated epitope residues are distinct from the corresponding non-epitope surface. Thus, in contrast to the AUC measure, which accounts both for sensitivity and specificity (i.e., true and false positive and negative predictions), in our approach these terms need not be directly defined. Moreover, a successful prediction according to our approach not only indicates that validated epitope residues obtained higher scores than average, but also that residues that are not part of a validated epitope and are unlikely to be immunogenic obtained lower scores than average. This approach is thus a more reliable measure for assessing how well a prediction algorithm succeeds in detecting its targets.

In Rubinstein et al. (2008), we detected a set of physico-chemical and structural–geometrical properties, which significantly distinguish epitopes from the remaining antigen surface. Out of such 11 properties, only six were ultimately included in the set of 14 properties, which yielded optimal epitope predictions. This partial overlap shows that not all properties that significantly distinguish epitope from non-epitope surfaces in a robust statistical comparison can be used together to produce optimal predictive power. The difference between the two sets of properties presumably stems from the inherent differences between the two computational approaches (characterizing what distinguishes epitopes versus predicting them). Perhaps the set of properties which is optimal for the prediction task is comprised of the set of properties that best distinguish epitope from non-epitope surfaces under the limitation that they are non-redundant.

The challenge in this work was to rationally combine information on epitope characteristics to develop a sophisticated tool for predicting immunogenic regions. Such a challenge was also noted by Ponomarenko and Bourne (2007) when suggesting future directions for improving epitope prediction methods. As our algorithm outperforms other methods for detection of immunogenic regions, we believe that a significant progress towards solution of this problem has been accomplished. The superiority of our method also indicates that the machine-learning approach is the natural paradigm to address the problem at hand. Currently, our structural train data consist of merely 47 antibody–antigen co-crystals. The continuous accumulation of additional data would certainly boost the training stage, and is expected to continuously improve the method's predictive power. This increase in data should also enhance the biological insights gained from this work as to which properties are most informative of immunogenicity.

Acknowledgments

This work was supported by the Wolfson Foundation, an ISF grant no. 1208/04, and a grant from the Israeli Ministry of Science (to T.P.); and, by the Edmond J. Safra Program in Bioinformatics at Tel Aviv University (to I.M. and N.D.R.).

References

Arnon, R., Sela, M., 1969. Antibodies to a unique region in lysozyme provoked by a synthetic antigen conjugate. Proc. Natl. Acad. Sci. U.S.A. 62, 163–170.

Atchley, W.R., Zhao, J., Fernandes, A.D., Druke, T., 2005. Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. U.S.A. 102, 6395–6400.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Blythe, M.J., Flower, D.R., 2005. Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci. 14, 246–248.

Castrignano, T., De Meo, P.D., Carrabino, D., Orsini, M., Floris, M., Tramontano, A., 2007. The MEPS server for identifying protein conformational epitopes. BMC Bioinformatics 8 (Suppl. 1), S6.

Cawley G.C., 2006. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. Neural Networks. IJCNN'06. International Joint Conference on Neural Networks.

Emini, E.A., Hughes, J.V., Perlow, D.S., Boger, J., 1985. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J. Virol. 55, 836–839.

Enshell-Seijffers, D., Denisov, D., Groisman, B., Smelyanski, L., Meyuhas, R., Gross, G., Denisova, G., Gershoni, J.M., 2003. The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1. J. Mol. Biol. 334, 87–101.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics 11, 1–21.

Halperin, I., Wolfson, H., Nussinov, R., 2003. SiteLight: binding-site prediction using phage display libraries. Protein Sci. 12, 1344–1359.

Haste Andersen, P., Nielsen, M., Lund, O., 2006. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 15, 2558–2567.

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78, 3824–3828.

Irving, M.B., Pan, O., Scott, J.K., 2001. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. Curr. Opin. Chem. Biol. 5, 314–324.

Janin, J., Wodak, S., 1978. Conformation of amino acid side-chains in proteins. J. Mol. Biol. 125, 357–386.

Jones, S., Thornton, J.M., 1997. Analysis of protein–protein interaction sites using surface patches. J. Mol. Biol. 272, 121–132.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Karplus, P.A., Schulz, G.E., 1985. Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. Naturwissenschaften 72, 212–213.

Kolaskar, A.S., Tongaonkar, P.C., 1990. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett. 276, 172–174.

Kulkarni-Kale, U., Bhosle, S., Kolaskar, A.S., 2005. CEP: a conformational epitope prediction server. Nucleic Acids Res. 33, W168–W171.

Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. 21, 1781–1791.

Mayrose, I., Shlomi, T., Rubinstein, N.D., Gershoni, J.M., Ruppin, E., Sharan, R., Pupko, T., 2007. Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. Nucleic Acids Res. 35, 69–78.

Miller, S., Janin, J., Lesk, A.M., Chothia, C., 1987. Interior and surface of monomeric proteins. J. Mol. Biol. 196, 641–656.

Moreau, V., Granier, C., Villard, S., Laune, D., Molina, F., 2006. Discontinuous epitope prediction based on mimotope analysis. Bioinformatics 22, 1088–1095.

Muller, Y.A., Chen, Y., Christinger, H.W., Li, B., Cunningham, B.C., Lowman, H.B., de Vos, A.M., 1998. VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2 4 A resolution and mutational analysis of the interface. Structure 6, 1153–1167.

Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R.E., Carlson, W.B., Fanning, D.W., Smith, J.A., Rose, G.D., 1986. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc. Natl. Acad. Sci. U.S.A. 83, 226–230.

Parker, J.M., Guo, D., Hodges, R.S., 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 25, 5425–5432.

Pellequer, J.L., Westhof, E., Van Regenmortel, M.H., 1991. Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol. 203, 176–201.

Pellequer, J.L., Westhof, E., Van Regenmortel, M.H.V., 1993. Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol. Lett. 36, 83–99.

Ponnuswamy, P.K., Prabhakaran, M., Manavalan, P., 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta 623, 301–326.

Ponomarenko, J.V., Bourne, P.E., 2007. Antibody–protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct. Biol. 7, 64.

Rost, B., Yachdav, G., Liu, J., 2004. The PredictProtein server. Nucleic Acids Res. 32, W321–W326.

Rubinstein, N.D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J.M., Pupko, T., 2008. Computational characterization of B-cell epitopes. Mol. Immunol. 45 (12), 3477–3489.

Saha, S., Raghava, G.P., 2006. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65, 40–48.

Saha, S., Bhasin, M., Raghava, G.P., 2005. Bcipep: a database of B-cell epitopes. BMC Genomics 6, 79.

Schreiber, A., Humbert, M., Benz, A., Dietrich, U., 2005. 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. J. Comput. Chem. 26, 879–887.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M., 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15, 327–332.

Sogabe, S., Stuart, F., Henke, C., Bridges, A., Williams, G., Birch, A., Winkler, F.K., Robinson, J.A., 1997. Neutralizing epitopes on the extracellular interferon gamma receptor (IFNgammaR) alpha-chain characterized by homolog scanning mutagenesis and X-ray crystal structure of the A6 fab-IFNgammaR1-108 complex. J. Mol. Biol. 273, 882–897.

Tsodikov, O.V., Record Jr., M.T., Sergeev, Y.V., 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J. Comput. Chem. 23, 600–609.

Westwood, O.M.R., Hay, F.C., 2001. Epitope Mapping: A Practical Approach. Oxford University Press, Oxford, UK.