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

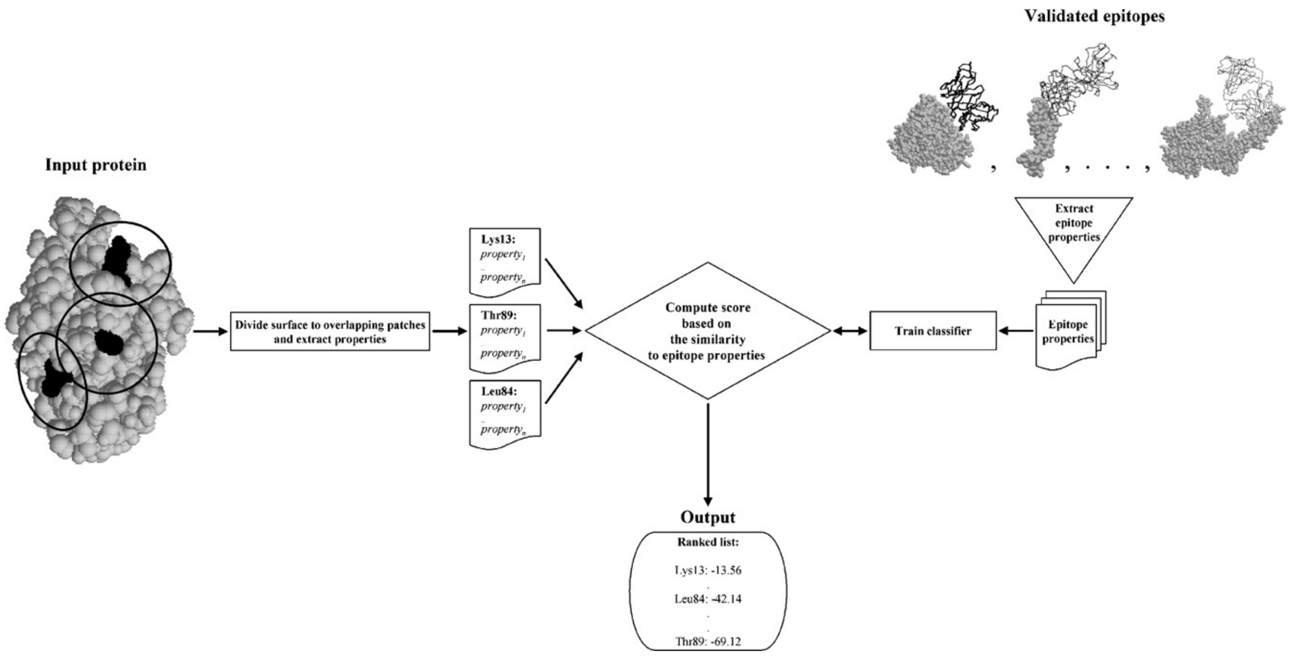

Fig. 1. Illustration of the flow of the prediction algorithm. A dataset of antibody–antigen co-crystal structures is used to derive epitopes and extract their physico-chemical and structural–geometrical properties. This collection of epitope properties is then used to train the classifier. Given an input 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.