Computational Biology and Chemistry 65 (2016) 117–127

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research article

SVM and SVR-based MHC-binding prediction using a mathematical presentation of peptide sequences

Davorka R. Jandrlić

University of Belgrade, Faculty of Mechanical Engineering, Kraljice Marije 16, Belgrade, Serbia

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 16 September 2016
Accepted 24 October 2016
Available online 27 October 2016

Keywords:
T-cell epitope
MHC I binding prediction
Data mining
Support vector machine
Encoding scheme

ABSTRACT

At present, there are a number of methods for the prediction of T-cell epitopes and major histocompatibility complex (MHC)-binding peptides. Despite numerous methods for predicting T-cell epitopes, there still exist limitations that affect the reliability of prevailing methods. For this reason, the development of models with high accuracy are crucial. An accurate prediction of the peptides that bind to specific major histocompatibility complex class I and II (MHC-I and MHC-II) molecules is important for an understanding of the functioning of the immune system and the development of peptide-based vaccines. Peptide binding is the most selective step in identifying T-cell epitopes. In this paper, we present a new approach to predicting MHC-binding ligands that takes into account new weighting schemes for position-based amino acid frequencies, BLOSUM and VOGG substitution of amino acids, and the physicochemical and molecular properties of amino acids. We have made models for quantitatively and qualitatively predicting MHC-binding ligands. Our models are based on two machine learning methods support vector machine (SVM) and support vector regression (SVR), where our models have used for feature selection, several different encoding and weighting schemes for peptides. The resulting models showed comparable, and in some cases better, performance than the best existing predictors. The obtained results indicate that the physicochemical and molecular properties of amino acids (AA) contribute significantly to the peptide-binding affinity.

1. Introduction

The binding of peptides, derived by the intracellular processing of protein antigen(s) (Ag(s)) to MHC proteins, is the most selective step in defining T-cell epitopes. Computational methods, based on reverse immunology, are essential steps in the identification of T-cell epitope candidates, and complement epitope screening by predicting the best binding peptides. Computational epitope-prediction programs are trained on the known peptide-binding affinities to a particular MHC molecule (or a defined set of MHC molecules) and fall into two categories (Brusic et al., 2004; Yang and Yu, 2009): sequence-based and structure-based. The first category is focused on the primary structure of the analyzed protein Ags and the identification of binding peptides, while the second makes use of the 3D structure of MHC molecules or their binding sites. Sequence-based methods for the prediction of MHC-binding peptides include binding motifs, quantitative matrices, artificial neural networks, hidden Markov models (HMMs) and molecular modeling. Structure-based methods, developed in structural biology for the prediction of potentially good MHC-binders, involve docking of peptides, threading algorithms, binding energy and molecular dynamics to discriminate between binding and non-binding peptides (Patronov and Doytchinova, 2013). In this study, we have only considered sequence-based methods. Of the current sequence-based methods, the most prevalent are those based on machine learning, because they are better at balancing the cost/performance ratio. Experimental methods are very expensive and time consuming, the application of prediction methods that reduce the number of experiments, and thereby cost, can significantly lower the time and money needed for experiments (Huang et al., 2013). Prediction methods are also used when the time needed for the identification of epitopes is crucial for rapid immunization.

The first developed methods for predicting epitopes included only the analysis of sequences and epitope motif alignment (Sette and Fikes, 2003). These methods were later upgraded to position-specific scoring matrix (PSSM) approaches (Yu et al., 2002). A disadvantage of these approaches is the negation of the link between an amino acid (AA) and neighboring molecules, i.e. the

E-mail address: djandrlic@mas.bg.ac.rs (D.R. Jandrlić).

http://dx.doi.org/10.1016/j.compbiolchem.2016.10.011
1476-9271/

assumption that an AA independently appears at an appropriate position and contributes alone to the binding affinity. Another flaw in these methods is the poorer accuracy in predicting T-cell epitopes, which was one of the motives for using the advanced machine learning algorithms mentioned above (Yu et al., 2002).

Novel and more advanced models for predicting T-cell epitopes with a high accuracies frequently appear. However, when we tried to include all these predictors in our previous research (Mitić et al., 2014; Pavlović et al., 2014; Jandrlić et al., 2016; Jandrlić, 2016), we were confronted with numerous restrictions. Predictions were made for single allele or a small number of alleles, models were trained on at most 200–300 peptides. In the lack of experimental data, the data produced through predictions were combined, or taken from different sources, irrespective of the methods used to define the epitopes (Peters et al., 2006). The development of a good model for the prediction of T-cell epitopes is a difficult task because of this lack of well-documented experimental data. It is common to utilize data from published articles or from specialized databases providing information related to T-cell epitopes, such as Syfpeithi (Rammensee et al., 1999), MHCBN (Bhasin et al., 2003), AntiJen (Toseland et al., 2005), HLA Ligand (Sathiamurthy et al., 2003), FIMM (Schönbach et al., 2000). However, algorithm developers are not always aware of the implications of mixing data from different experimental approaches, such as T-cell response, MHC ligand elution and MHC binding data. Even within a single assay category, such as MHC-binding experiments, mixing data from different sources without further standardization can be problematic. The data often had conflicting classifications into both binding and nonbinding peptides (Peters et al., 2006).

1.1. Existing methods and models

This section describes the current most reliable predictors of MHC-binding peptides and their methodology. Predictors from the CBS group (http://www.cbs.dtu.dk/) have proved to be the most reliable, accurate and provide support for a large number of different HLA alleles, even in the absence of experimental data (pan-specific). These predictors belong to two families: NetMHC and NetMHCpan, both are based on artificial neural networks (ANNs). These predictors combine several ANN models, which are based on sparse sequence encoding and BLOSUM50 encoding. ANNs can be utilized to make both qualitative and quantitative predictions. These predictors are still being developed and improved, and are included in a number of tools, including the Immune Epitope Database–IEDB (http://tools.iedb.org/main/tcell/) and CBS (Luo et al., 2015). The next group of predictors, which is also regularly updated and accessible for large-scale analysis, is that from the IEDB (http://tools.immuneepitope.org/mhci/). Most of the predictors available at the IEDB are ANN-based or matrix-based. ANN predictors (Nielsen et al., 2003) use a combination of sparse encoding, BLOSUM encoding and input derived from hidden Markov models as a sequence presentation for different neural networks; these models are then combined. The ARB (average relative binding) predictor (Bui et al., 2005) is a matrix-based method that directly predicts IC50 values; SMM and SMMPMBEC are also matrix-based methods that predict peptide binding to MHC molecules, peptide transport by the transporter associated with antigen processing (TAP) and proteasomal cleavage of proteins (Peters and Sette, 2005). The PickPocket is based on receptor pocket similarities (Kim et al., 2009). There are also predictors that are not mentioned here; a detailed list of predictors with their prediction precision is described (Luo et al., 2015). Most of the predictors are based on sparse or BLOSUM50 encoding of sequences, while different methods of machine learning are used: ANN, HMM, Decision Tree and SVM. The reason why these predictors are not described in detail here is because they are not available as stand-alone applications, and because they cannot be easily tested for larger proteins or for more proteins. To add to this, even when only one protein is considered, the result is obtained only after some time, or the end result is an error message. Of all the predictors presented in the above mentioned work, only POPI (Tung and Ho, 2007) uses physicochemical (*PC*) properties as input features. However, this predictor only gives a qualitative evaluation of prediction (it is an epitope – it isn't an epitope), and with very low reported accuracy. Beside mentioned predictors, there are proposed methods which are based on orthonormal encoding strategies and binary encoded *PC* properties which suggest that certain combination of *PC* properties can significantly improve classification performance (Gok and Ozcerit, 2012a,b). However, proposed methods are only developed for qualitative classification of peptides.

In our previous work (Mitić et al., 2014; Pavlović et al., 2014; Jandrlić et al., 2016), we used predictors from the CBS group; however, our intention to include other predictors that do not belong to CBS group or IEDB tool to compare results with other proteins characteristics, was met by all of the abovementioned problems. This was the reason for creating the support vector machine (SVM) models used for binary classification and regression, based on different sequence encoding strategies. The obtained models predict MHC-binding ligands with great accuracy. Of particular interest was the finding that the combination of *PC* properties greatly influences the binding affinity of peptides. In order to avoid the problems of inconsistent data to obtain reliable models, such as differing measures of binding affinity, etc., we chose to use only data from the IEDB (http://www.iedb.org/), which is regularly updated, as the most reliable source of MHC-binding ligands.

2. Materials and methods

2.1. Datasets

The data source was the Immune Epitope Database (IEDB), June 2015 version. All experimentally proven MHC-binding ligands for all available alleles were downloaded. We limited our research to peptides 9 amino acids (AAs) in length because nonamers are the most common MHC-I epitopes, and because for MHC-I there exist enough experimental data for the construction of good models for a number of alleles. We discarded the data:

- where there was insufficient information for the construction of a good model, for individual alleles.
- those ligands for which there were no qualitative and quantitative measures (affinity binding and verification of whether a ligand is positive or negative).
- ligands labeled as both positive and negative, and as being found in the same protein at the same position.
- peptides containing a non-proteinogenic amino acid, where AA not in $\alpha = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.

Of the remaining alleles with a sufficient number of peptides required to produce reliable model, herein we present the results for the first 15 alleles according to the amount of data available for each allele. Table 1 presents the alleles for which results are provided, and for each allele the numbers of positive and negative epitopes involved in developing and testing the model are shown.

We made separate models for each of the presented alleles. All data used to build and test the models are available at http://147.91.26.24/mrepo/CBIC-MHC-I-binding-prediction-DJandrlic.zip. Also attached are the obtained models, and the programs used to process the peptides and convert them to the appropriate vector.