Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Prediction of binding peptides to class I Major Histocompatibility Complex using modified scoring matrices and data splitting strategies

Dina A. Salem [a,*], *Rania A. Abul Seoud* [b], *Yasser M. Kadah* [c]

[a] *Computer Department, Faculty of Engineering, Misr University for Science and Technology, Giza, Egypt*
[b] *Department of Electronics and Communication, Faculty of Engineering, Fayoum University, Fayoum, Egypt*
[c] *Electrical and Computer Engineering Department, King Abdulaziz University, Jeddah, Saudi Arabia*

ARTICLE INFO

Article history:
Received 10 October 2015
Received in revised form
5 March 2016
Accepted 11 April 2016
Available online 12 May 2016

Keywords:
Data splitting
Epitope prediction
Major Histocompatibility Complex
Position specific scoring matrix
Vaccine design

ABSTRACT

Predicting peptides that can bind to MHC class I molecules is an important step in the vaccine design process. Computational approaches have potential to provide good predictive models that save both time and cost of the process. Position Specific Scoring Matrix (PSSM) is a reliable approach when dealing with amino acid sequences. PSSM formation involves carefully selecting its constructing data and parameters. In this work, we apply three different data splitting strategies and propose alternative values for the embedded PSSM parameters. The basic principle of data splitting is to choose train data that is able to represent the whole data. We propose using the Kennard–Stone algorithm to highlight the importance of choosing the data constituting the PSSM. Furthermore, this work proposes modifications to PSSM parameters and studies the model behavior in response to each change. The model is applied to experimental data for the Major Histocompatibility Complex of class I. Performance of modified parameters show either comparable or better results to conventional parameters. Moreover, Kennard–Stone data splitting algorithm contributed to significant model performance enhancement.

1. Introduction

Vaccination process aims to aid our bodies to develop a defense mechanism against infectious diseases by artificial induction of immunity [1]. Immune system is characterized by having memory cells fulfilling main role in response enhancement to subsequent encounters with that same pathogen. Vaccines are the tools of this process of acquired immunity. Vaccines can be classified into three main forms; namely, live attenuated, inactivated/killed or subunit/conjugate vaccines. Subunit vaccines offer the advantage of having lower chances of adverse reaction because they contain only antigens with best immune response stimulation [2]. Therefore, to reach an effective subunit vaccine, the process relies on selecting the proper antigens [3]. Antigenic peptide in immunology

* *Corresponding author at*: Computer Department, Faculty of Engineering, Misr University for Science and Technology, Giza, Egypt.
E-mail address: dena.salem@mail.com (D.A. Salem).

http://dx.doi.org/10.1016/j.bbe.2016.04.003

is defined as any structural peptide that is able through a specific pathway to elicit the receptors of the adaptive immune system. T-cell receptors cannot recognize antigens unless processed into small fragments that can bind Major Histocompatibility Complex (MHC). MHC class I molecules (called Human Leukocyte Antigen or HLA in humans) are responsible for presenting intracellular proteins to the cytotoxic T Lymphocytes (CTLs) after binding its extracellular part with the antigenic peptide. The corresponding CTLs to the formed – peptide complex in turn detect and kill the presenting cell in case of non-self-epitopes [4].

The MHC class I peptide complex must be immunogenic in the sense that it is able to trigger the immune response. The degree of immunogenicity of the complex depends on the strength of binding between allele and the antigenic peptide. According to [5], the value of binding affinity for an immunogenic complex must be more than 500 nM. Binding affinity can be expressed in either qualitative or a quantitative forms. A qualitative property indicates whether a peptide is a binder or a non-binder to MHC class I molecule. Alternatively, a quantitative score is a description of how strong the binding is. For some epitope databases such as the Immune Epitope Database (IEDB) [6], the half-maximal inhibitory concentration (IC50) gives an indication to the binding affinity of a peptide to a specific allele. Using both the quantitative scores along with the qualitative properties, computational analysis of vaccines provide a major step forward in enhancing the prediction of antigenic peptides that are likely able to bind MHC class I molecules.

Computer-aided vaccine design contributed to the influential antigenic peptide prediction step through two main tools: scoring matrices and machine learning algorithms. Position Specific Scoring Matrices (PSSM) are the most commonly used scoring matrices in the area of amino acid sequence alignment showing comparable results to machine learning tools [7]. They have the characteristics of easy construction from scratch and adaptability of its parameters for the employed prediction goal. They result in a consensus score for each peptide that decides according to a specific threshold if this peptide is MHC class I binder or not [8].

Discovering T-cell epitopes is still an open area in the vaccine design field. So far, relying on computational techniques only does not seem to be sufficient. Available models need more optimization to enhance the accuracy of the predictive goal. For example, PSSM present many effective parameters that need optimization altogether. The model must thoroughly choose and filter its constructing data so as not to result in misleading outcomes. Statistical measures of the evaluation phase must be comprehensive and reflective of the real performance. In this work, we investigate how to build PSSM taking into consideration all its constituting parameters. This will facilitate the process of modifying all the effective steps including different proposed estimates to parameters that seem to play a role in improving the prediction. This investigation studies the role of each of the PSSM parameters and provide an insight into which of them would be more fruitful to pursue for further optimization. Moreover, this study suggests a new approach that emphasizes the importance of carefully selecting the model constructing data based on data splitting strategies. Three different data splitting algorithms are utilized. Finally, the evaluation metrics of the outcomes incorporate the sensitivity and specificity statistics in addition to the commonly used prediction accuracy statistic to reveal more information about the performance and comparison of different methods used.

2. Previous work

Reverse immunology is the use of previously experimental detected epitopes to build models to decide easily if new peptides are to be considered as epitopes or not. The reverse immunology systems employing *in-silico* techniques were extensively used for epitope prediction with high confidence during the past two decades [9]. For this reason, computational techniques designed for prediction of peptides binding to MHC molecules were at the focus of many research articles with the goal of enhancing prediction accuracy. The main blocks that influence the ability to reach an acceptable degree of model confidence are data aggregation and model building methodology.

One of the common online servers used for prediction of MHC class I binding is the ‘‘RANKPEP: prediction of binding peptides to class I and class II MHC molecules’’ [8,10]. It gets a protein sequence as an input to rank all its possible peptides according to PSSM coefficients. The data used to build its model are a collection of MHC class I binding peptides downloaded from MHCPEP database [11]. In this database, all peptides are binders with unknown, low, moderate or high binding affinities. Data collected are filtered by discarding low binders and then applying a similarity reduction where no two peptides in the same dataset have at least four residues in common. Data are then divided according to the peptide length into five specific-length sets of 8-mer, 9-mer, 10-mer, 11-mer and 12^{+}-mer. The model constructs a profile for each dataset with at least five sequences. Implementation of profiles used the sequence weighted frequency model available by PROFILEWEIGHT and BLK2PSSM included in the BLIMPS package [12,13]. The profiles adopted its background probabilities from the amino acid frequencies in the SWISSPROT database.

Moving along the same direction, a group of researchers focused mainly on collecting immune epitopes from different sources to offer the presence of benchmark datasets on one online resource called the Immune Epitope Database (IEDB) [6]. This same resource evaluates three prediction methods using their collected data. One of these is based on Artificial Neural Networks (ANN), which is machine learning based method. The other two are based on Average Relative Binding (ARB) and Stabilized Matrix Method (SMM), which are matrix based methods. Their results were very promising which indicates the credibility of their data [14]. Another study used MHC binding peptides collected from the EPIMHC database [15]. It selected 9-mer binder peptides with high binding affinity to MHC allele HLA-A0201 only as a filtration step. HLA-A0201 refers to the Human Leukocyte Antigen of super-type A that is one of three different major types (HLA-A, HLA-B and HLA-C). The next two digits (02) defines the serotype and the last two digits (02) is for the HLA protein produced [16]. This study relied on a collection of software packages for deriving profiles