ELSEVIER

Journal of Immunological Methods 320 (2007) 143 – 154

JIM Journal of Immunological Methods

www.elsevier.com/locate/jim

Research paper

Prediction of supertype-specific HLA class I binding peptides using support vector machines

Guang Lan Zhang [a,b], Ivana Bozic [c], Chee Keong Kwoh [b], J. Thomas August [d], Vladimir Brusic [e,*]

[a] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore*
[b] *School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, Singapore 639798, Singapore*
[c] *Faculty of Mathematics, University of Belgrade, Belgrade, Serbia and Montenegro*
[d] *Department of Pharmacology and Molecular Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA*
[e] *Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, MA 02115, USA*

Received 27 November 2006; accepted 20 December 2006
Available online 25 January 2007

Abstract

Experimental approaches for identifying T-cell epitopes are time-consuming, costly and not applicable to the large scale screening. Computer modeling methods can help to minimize the number of experiments required, enable a systematic scanning for candidate major histocompatibility complex (MHC) binding peptides and thus speed up vaccine development. We developed a prediction system based on a novel data representation of peptide/MHC interaction and support vector machines (SVM) for prediction of peptides that promiscuously bind to multiple Human Leukocyte Antigen (HLA, human MHC) alleles belonging to a HLA supertype. Ten-fold cross-validation results showed that the overall performance of SVM models is improved in comparison to our previously published methods based on hidden Markov models (HMM) and artificial neural networks (ANN), also confirmed by blind testing. At specificity 0.90, sensitivity values of SVM models were 0.90 and 0.92 for HLA-A2 and -A3 dataset respectively. Average area under the receiver operating curve (A_{ROC}) of SVM models in blind testing are 0.89 and 0.92 for HLA-A2 and -A3 datasets. A_{ROC} of HLA-A2 and -A3 SVM models were 0.94 and 0.95, validated using a full overlapping study of 9-mer peptides from human papillomavirus type 16 E6 and E7 proteins. In addition, a large-scale experimental dataset has been used to validate HLA-A2 and -A3 SVM models. The SVM prediction models were integrated into a web-based computational system MULTIPRED1, accessible at antigen.i2r.a-star.edu.sg/multipred1/.

Keywords: T-cell epitope; Human Leukocyte Antigen supertype; Promiscuous binding peptide; Support vector machines

1. Introduction

Cellular immunity in vertebrates is mediated by T cells of the immune system which generate highly specific and lasting immune responses to pathogens (Fabbri et al., 2003). T-cell-based immune responses are mediated by antigenic peptides presented by major histocompatibility complex (MHC) molecules (Pamer and Cresswell, 1998; Yewdell and Bennink, 2001). Antigenic peptides bind MHC molecules and form peptide/MHC complexes. Peptide/MHC complexes shown to be recognized by T cells are called T-cell

[*] Corresponding author. Tel.: +1 617 632 3824; fax: +1 617 632 3351.
E-mail address: Vladimir_Brusic@DFCI.HARVARD.EDU (V. Brusic).

0022-1759/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jim.2006.12.011

epitopes. Identifying promiscuous peptides that bind multiple Human Leukocyte Antigen (HLA, human MHC) alleles is a basis for T cell epitope mapping and epitope-based vaccine development (Berzofsky et al., 2001; Srinivasan et al., 2004a; De Groot, 2006). HLA genes are the most polymorphic human genes known (Williams, 2001), with more than 2400 allelic variants identified in the human population as of July 2006 (www.anthonynolan.org.uk/HIG/). Because of the high HLA polymorphism, identifying promiscuous peptides that bind more than one HLA allele is essential for the development of vaccines with a broad and unbiased coverage of the human population. HLA alleles that share sequence similarity and that bind largely overlapping sets of peptides define HLA supertypes (Sette and Sidney, 1999; Doytchinova et al., 2004; Lund et al., 2004). Promiscuous peptides have been reported in the context of HLA supertypes (Threlked et al., 1997; Wilson et al., 2003; Srinivasan et al., 2004b). Epitope-based vaccines show great potential in fighting infectious diseases (Sette et al., 2000; Ada, 2003; Wilson et al., 2003), and they are also investigated for control of cancers, allergy, autoimmunity, and even dementia (Alexander et al., 2002; Durrant and Ramage, 2005; Quintana and Cohen, 2005; Verhagen et al., 2005; Wisniewski and Frangione, 2005; De Groot, 2006).

Experimental validation of peptide binding to HLA molecules is time-consuming and costly, and thus not applicable to large scale screening across multiple HLA alleles. Computational methods are instrumental for systematic large-scale identification of MHC-binding peptides (Schirle et al., 2001; Brusic et al., 2004). One type of methods is structure-based approach that relies on structural conservation observed in 3D structure of peptide–MHC complexes (Schueler-Furman et al., 2000; Bui et al., 2006; Tong et al., 2006). These methods are computationally intensive, and have mainly been applied to MHC molecules with known crystal structures. Data-driven approaches include statistical methods based on experimental peptide binding measurements. These methods include binding motifs (Rammensee et al., 1993), quantitative matrices (Parker et al., 1994; Singh and Raghava, 2003; Reche and Reinherz, 2005; Peters and Sette, 2005), artificial neural networks (ANN) (Honeyman et al., 1998; Christensen et al., 2003), hidden Markov models (HMM) (Mamitsuka, 1998; Brusic et al., 2002), decision trees (Savoie et al., 1999; Segal et al., 2001), discriminant analysis (Mallios, 2001), multivariate regression (Lin et al., 2004), ensemble classifier (Xiao and Segal, 2005), support vector machines (SVM) (Donnes and Elofsson, 2002; Zhao et al., 2003; Bhasin and Raghava, 2004; Riedesel et al., 2004; Bozic et al., 2005; Liu et al., 2006; Cui et al., 2007), and biosupport vector machine which is modified from a conventional support vector machine by introducing a biobasis function so that the non-numerical attributes of amino acids can be recognized without a feature extraction process (Yang and Johnson, 2005). Recently a structure- and sequence-based method was reported, in which residue-based energy terms from the molecular dynamics simulations are used as features to train SVM prediction models for peptide/MHC class I binding (Antes et al., 2006).

SVM-based models showed higher accuracy than other prediction methods in studies of peptide binding to a single HLA molecule. We have employed SVM models with a novel data representation, which captures information of the interaction between a peptide and an HLA molecule and allows the use of a single model for prediction of peptide binding to a multiplicity of alleles that belong to a particular HLA supertype. Earlier we reported the application of HMM (Brusic et al., 2002) and ANN (Zhang et al., 2005b) for prediction of peptide binding to the HLA-A2 supertype. A web-based prediction system, MULTIPRED (Zhang et al., 2005a), was developed using HMM and ANN models. In this study we extended MULTIPRED by applying SVM models. The SVM-MULTIPRED was applied to prediction of HLA class I supertype-specific promiscuous binding peptides in the context of HLA-A2 and -A3. Extensive testing, including blind testing and 10-fold cross-validation, were performed to assess the performance of the prediction models. Validation of the models was conducted using experimental data from human papillomavirus (HPV) type 16 E6 and E7 proteins and a large-scale experimental dataset made available recently by Peters et al. (2006). The performance of the SVM models were compared with that of HMM and ANN models. MULTIPRED1 is the updated version of MULTIPRED (Zhang et al., 2005a). MULTIPRED1 is accessible at antigen.i2r.a-star.edu.sg/multipred1/.

2. Materials and methods

2.1. Data and data representation

Nine-mer peptide data were extracted from the MHCPEP database (Brusic et al., 1994), published articles, and a set of HLA non-binding peptides (Brusic, V. unpublished data). The HLA-A2 supertype dataset, named as Dataset1, has 3050 peptides (664 binders and 2386 non-binders) related to 15 alleles (Table 1) of