Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 23 (2005) 481–489
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Predictive Bayesian neural network models of MHC class II peptide binding

Frank R. Burden [b,c], David A. Winkler [a,c,*]

[a] Centre for Complexity in Drug Discovery, CSIRO Molecular Science, Clayton, Australia
[b] SciMetrics, Harrow Enterprises Ltd., Vic., Australia
[c] Chemistry Department, Monash University, Clayton, Australia

Received 1 October 2004; accepted 18 March 2005
Available online 6 May 2005

Abstract

We used Bayesian regularized neural networks to model data on the MHC class II-binding affinity of peptides. Training data consisted of sequences and binding data for nonamer (nine amino acid) peptides. Independent test data consisted of sequences and binding data for peptides of length ≤25. We assumed that MHC class II-binding activity of peptides depends only on the highest ranked embedded nonamer and that reverse sequences of active nonamers are inactive. We also internally validated the models by using 30% of the training data in an internal test set.

We obtained robust models, with near identical statistics for multiple training runs. We determined how predictive our models were using statistical tests and area under the Receiver Operating Characteristic (ROC) graphs (A_{ROC}). Most models gave training A_{ROC} values close to 1.0 and test set A_{ROC} values >0.8.

We also used both amino acid indicator variables (bin20) and property-based descriptors to generate models for MHC class II-binding of peptides. The property-based descriptors were more parsimonious than the indicator variable descriptors, making them applicable to larger peptides, and their design makes them able to generalize to unknown peptides outside of the training space.

None of the external test data sets contained any of the nonamer sequences in the training sets. Consequently, the models attempted to predict the activity of truly unknown peptides not encountered in the training sets. Our models were well able to tackle the difficult problem of correctly predicting the MHC class II-binding activities of a majority of the test set peptides.

Exceptions to the assumption that nonamer motif activities were invariant to the peptide in which they were embedded, together with the limited coverage of the test data, and the fuzziness of the classification procedure, are likely explanations for some misclassifications.

Keywords: Bayesian neural networks; Quantitative structure-activity relationships; T-cell epitope; Major histocompatibility complex; Peptide binding

1. Introduction

Major histocompatibility complex (MHC) proteins are cell surface glycoproteins present on antigen presenting cells. When they recognize and bind peptides the complexes are identified by CD4+ T cells resulting in activation of the T-cell. Consequently, MHC-bound peptides play a crucial role in initiation, enhancement and suppression of immune responses, and in cytotoxicity. MHC molecules form two classes, depending on whether they bind peptides derived by degradation of intracellular proteins (class I), or extracellular proteins (class II). MHC class-II-binding peptides, which induce and recall T-cell responses, are called T-cell epitopes.

It is important to be able to identify T-cell epitopes for developing diseases therapies (e.g. malaria), and several groups have attempted to develop QSAR models to aid in identifying potent MHC binders. Buus described how privileged binding motifs exist in peptide binders, and how QSAR methods could be used to build predictive models of human immune reactivities [1]. Doytchinova and Flower employed the 3D QSAR methods CoMFA and CoMSIA to model the affinity of a small set of peptides for the class I MHC HLA-A*0201 molecule [2]. They found CoMSIA

* Corresponding author. Tel.: +61 3 9545 2477; fax: +61 3 9545 2446.
E-mail address: dave.winkler@csiro.au (D.A. Winkler).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.03.001

superior to CoMFA in predicting the affinities of the peptides. In a more recent paper Doytchinova, Blythe, and Flower used an ''additive'' linear regression method to predict MHC protein peptide binding [3]. They assumed that binding affinity was an additive function of the contributions of amino acids in each position of the peptide, essentially a type of Free-Wilson approach, with additional allowance for interactions between a given amino acid and its neighbors. They were able to predict the pI_{50} values of a test set of 89 compounds within 0.5 log units. Whilst not a QSAR study, Logean, Sette, and Rognen derived a customized free energy scoring function to predict the binding affinity of 26 peptides to the class I MHC HLA-B*2705 protein [4]. Their Fresno method was able to rank the affinities of the peptides, and predict numerical values for their binding energies within 3–4 kJ/mol. Brusic et al. used backpropagation neural networks to derive a QSAR model and identify potent HLA-A11 binders from a training set of nonamer (nine amino acid) peptides with known binding affinities [5]. Their cyclically refined models were able to identify peptides that bound but did not conform to a putative binding motif. Gulukota et al. published a study comparing sequence motifs to a backpropagation neural net and a polynomial method as means of predicting binding or peptides to MHC molecules [6]. More recently, De Hann et al. elucidated the relative individual contributions of side chain hydrogen bonding, and flexibility to MHC binding affinity of peptides using peptoid surrogates [7]. A novel support vector machine (SVM) method was used to classify a relatively large set of peptides binding to HLA-DRB1*0401 by Bhasin and Raghava [8]. They claimed an 86% accuracy of prediction using SVM.

MHC class II peptide recognition is a more complex process to model than class I recognition. It is clear from previous studies that the interaction of peptides with the MHC is nonlinear and complex, with interactions between amino acids being important modulators of affinity. Buus [1] reviewed a number of general approaches for MHC binding affinity prediction and advocated strongly for the application of neural networks. Buus felt they were much better suited to recognizing complicated peptide patterns than binding motifs (anchors) and other algorithmic methods. We have developed a robust structure-property mapping methodology able to model relationships between chemical structure and a wide variety of properties. Using these methods we have built predictive models of drug target activity [9], ADME properties [10], toxicity [11], and phase II metabolism [12], amongst other properties.

Our methodology employs Bayesian regularized neural networks and novel molecular descriptors to build predictive QSAR models [13]. Bayesian methods have a number of advantages over traditional backpropagation neural networks used in previous QSAR studies, including those modeling peptide binding to the MHC. Like standard backpropagation neural nets they are 'universal approximators', able to model complex, nonlinear response surfaces. The advantages of Bayesian neural are that they are robust, difficult to overtrain, minimize the risk of overfitting, are tolerant of noisy or missing data, automatically find the least complex model which explains the data, and can automatically optimize their architecture [14].

We have employed Bayesian neural network methods to build QSAR models explaining the more complex MHC class II-binding activity of peptides to two HLA protein alleles, HLA-DRB1*0101 and HLA-DRB1*0301.

2. Materials and methods

2.1. Training data sets

The peptide binding data were a superset of the data in the MHCPEP database curated by Brusic et al. [15]. We used two peptide-binding data sets to build predictive MHC binding models. These data related to binding of peptides to the HLA-DRB1*0101 (data set 101) and HLA-DRB1*0301 (data set 301) alleles, respectively. Training set 101 contained 1408 peptides and training set 301 contained 849 peptides. The two training data sets were used to derive separate models for peptide binding to the two HLA alleles. Peptides that bind to these MHCs have recognition motifs consisting of nine amino acids. The data sets consisted of the nonamer peptide sequences in single letter codes, together with an activity class of 1, nil MHC class II-binding activity (class N); 5, low MHC class II-binding activity (class L); 7, moderate MHC class II-binding activity (class M); and 9, high MHC class II-binding activity (class H). These classes correlated approximately with the $-\log IC_{50}$ (pI_{50}) of the test set values and were a logical choice.

2.2. Internal test sets

Traditionally, validation sets are required to stop neural net training to prevent overtraining and degradation of the ability of the network to generalize. In contrast, Bayesian neural networks do not require a validation set as the maximum in the evidence is used to terminate training. However, purely to illustrate the robustness of training and gave an additional (albeit less rigorous) indication of predictive ability, we have also used a internal test set. Each of the two training data sets (101 and 301) were randomly partitioned into a new training set (70% of peptides), and an internal test set (30% of peptides). Models were derived using the new training set, and assessed for predictive ability using the internal test sets. However, when building models to predict the external test sets we use all of the available training data in the models.

2.3. External test sets

We employed two independent external test sets for each of the 101 and 301 models (V. Brusic, private communica-