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 $\leq$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 communication). Set one comprised 30 peptides of length up to 20 amino acids. Set two consisted of 343 peptides of length up to 25 amino acids. None of the nonamer motifs in the external test sets appeared in the training sets. MHC binding affinity for external test set compounds was expressed as pI_{50} values. For this study the external test set compounds were classified into activity classes in a similar way to the training set data. External test set sequences are available from the authors on request.

2.4. Descriptors

The peptide sequences were converted into molecular descriptors. Each amino acid in the peptide was converted into a representation. This was done in two ways.

2.4.1. Binary (Bin20) descriptors

These were indicator variables describing the identity of the amino acid at each position in the peptide sequence. At each amino acid position one of 20 indicator variables is set to 1 to denote the presence of that amino acid. As the peptide motifs used in building the models were nonamers, and there are 20 possible amino acids at each position, there were 9 × 20 = 180 bin20 descriptors.

2.4.2. Property descriptors

These descriptors were designed to be a more compact representation of amino acid properties. They were based on physical properties of the 20 naturally-occurring amino acids. For each position in the peptide sequence there were seven descriptors, resulting in 9 × 7 = 63 property descriptors for each nonamer. The property descriptors for the twenty amino acids are summarized in Table 1.

2.5. Protein-peptide QSAR modelling

The models were derived using a Bayesian regularized neural network written in MATLAB code running on a personal computer. We used a three layer neural network architecture containing a single hidden layer with one or two neurodes to determine how well the Bayesian net could train and generalize. A single output neurode was used and the output mapped to the four biological response categories denoted above. Sigmoidal transfer functions were used in the hidden layer and linear transfer functions in the output layer. This architecture is substantially less complex than that employed by Gulukota et al. [6] who used 50 neurodes in the hidden layer. We trained the networks until we attained a maximum in the evidence. Details of Bayesian regularization applied to backpropagation neural networks may be found in previous publications [13,14] so only a brief summary is given here.

Regression is an ill-posed problem in statistics and regularization is used to improve the modeling performance. This is achieved by adding a term to the cost function (in non-regularized regressions the cost function is simply the sum of the square errors) that is minimized during the regression. In neural network-based regression, this additional regularization term is proportional to the square of the weights. Consequently, overly complex models with large weights are penalized by such cost functions. Bayesian statistics are used to find the correct balance between bias (where the model is too simple to describe the underlying relationship in the data) and variance (where the model is too complex and fits the noise as well as the underlying relationship). This is done by finding the optimum coefficients (α and β) for the two terms in the cost function

Table 1
Property descriptors for the amino acids

No.	Amino acid	Code	N_{accept}	N_{donor}	N_{hetero}	Mol. volume	log P	rot. bonds	Charge
1	Alanine	A	0	0	0	45	1.8	0	0
2	Cysteine	C	0	1	1	62	1.2	0	0
3	Aspartate	D	2	0	2	69	−1.5	1	−1
4	Glutamate	E	2	0	2	87	1.2	2	−1
5	Phenylalanine	F	0	0	0	115	3.3	1	0
6	Glycine	G	0	0	0	29	1.1	0	0
7	Histidine	H	2	2	2	97	1.0	1	0
8	Isoleucine	I	0	0	0	96	3.3	1	0
9	Lysine	K	0	1	1	108	−1.2	3	1
10	Leucine	L	0	0	0	96	3.3	1	0
11	Methionine	M	0	0	1	97	2.0	2	0
12	Asparagine	N	1	1	2	76	−1.5	1	0
13	Proline	P	0	0	0	76	2.9	0	0
14	Glutamine	Q	1	1	2	93	−1.0	2	0
15	Arginine	R	0	3	3	128	3.3	4	1
16	Serine	S	1	1	1	54	−0.2	0	0
17	Threonine	T	1	1	1	70	0.1	0	0
18	Valine	V	0	0	0	77	2.7	0	0
19	Tryptophan	W	0	1	1	145	2.6	1	0
20	Tyrosine	Y	1	1	1	124	2.7	1	0

N_{accept}, N_{donor}, N_{hetero} are the numbers of hydrogen bond donor, acceptors and heteroatoms, respectively. Mol. volume, log P, rot. bonds, and charge represent the molar volume, log of the octanol–water partition coefficient, number of rotatable bonds, and formal charge at physiological pH, respectively.

relating to the sum of the squared errors and sum of the squared weights.

In addition, backpropagation neural net training methods are usually variations of maximum likelihood algorithms that aim to find a single set of network weights that maximize the fit to training data. Applying a Bayesian framework to the neural net results in a probability distribution of weights not a single set of weights.

2.6. Statistics

As this is a classification exercise, contingency tables and A_{ROC} (area under the receiver operating characteristic (ROC) curve) were used as the primary yardsticks of performance [16]. An A_{ROC} value of 1 indicates a perfect model and a value of 0.5 denotes a model no better than chance. The A_{ROC} measure removes biases due to differing numbers of binding and non-binding peptides, and biases due to arbitrary defined decision thresholds [17]. This measure is also not overly affected by the presence of a small number of outliers, some of which may result from classification ambiguities near the decision boundaries.

3. Results

3.1. Modelling of the training data

We derived four models for each training data set (101 and 301). The models varied the type of peptide descriptor used (bin20, or property descriptors), and the number of neurodes in the hidden layer (one, or two).

Tables 2 and 3 summarize the training statistics for two neural net architectures, two training sets, and two types of descriptors. In the tables the descriptors column indicates the number of descriptors used to represent the nonamers, and $\mathrm{Par}_{\mathrm{eff}}$ is the number of effective parameters used by the neural net (approximately the number of non-trivial weights in the trained neural network). Examples of the quality of the training set and internal test set prediction are shown in Fig. 1.

Table 2
Results of Bayesian neural net training and validation studies

Model	Descriptors	A_{ROC} train	A_{ROC} test	$\mathrm{Par}_{\mathrm{eff}}$
101/1/bin20	180	1.00	0.91	152
101/2/bin20	180	0.99	0.84	312
101/1/prop	63	0.78	0.74	63
101/2/prop	63	0.77	0.71	114
301/1/bin20	174	0.96	0.88	154
301/2/bin20	174	0.96	0.88	268
301/1/prop	63	0.76	0.75	55
301/2/prop	63	0.85	0.79	108

The first column represents: data set/number of hidden layer neurodes/descriptor type. $\mathrm{Par}_{\mathrm{eff}}$ represents the number of effective parameters in the model.

Table 3
Statistics of Bayesian neural net model training using all data in model (no internal test set)

Model	Descriptors	A_{ROC}	$\mathrm{Par}_{\mathrm{eff}}$
101/1/bin20	180	0.99	160
101/2//bin20	180	0.99	319
101/1/prop	63	0.78	64
101/2/prop	63	0.74	115
301/1/bin20	174	0.94	154
301/2/bin20	174	0.98	297
301/1/prop	63	0.82	57
301/2/prop	63	0.82	111

The model column is data set/number of hidden layer neurodes/descriptor type.

3.1.1. 101 data set

All models were quite robust, with repeated training from random initial weights resulting in essentially identical training and test set statistics. The bin20 descriptors clearly resulted in better training statistics than the property-based descriptors as illustrated by the A_{ROC} values. The models derived from neural nets with two hidden layer neurodes had very similar training statistics to models in which a single hidden layer neurode was used. The property-based descriptors produced reproducible models and only required seven property descriptors per amino acid, compared with twenty for the bin20 descriptors.

Both neural net architectures produced models with good generalization abilities as judged by the A_{ROC} values for the internal test set. The results suggested that the property-based descriptors generalize less well than the bin20 descriptors. The models using two hidden layer neurodes were also slightly worse at generalizing than the simpler neural network architecture.

When trained on the full data set (Table 3), the bin20 descriptors again produced superior statistics to the property-based descriptors, and there was little difference between the two neural net architectures.

3.1.2. 301 data set

As with the 101 data set, the property-based descriptors gave inferior training and internal test statistics compared with the bin20 descriptors but they still gave models that generalized well. There was very little difference in performance between the two neural network architectures

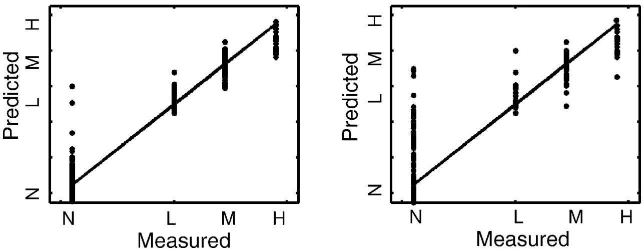

Fig. 1. Measured vs. predicted classes for training set (70% of peptides, left graph) and internal test set (30% of peptides, right graph) for the 301 data set, two neurodes in hidden layer, bin20 descriptors.

for training and prediction for the bin20 descriptors. The two hidden layer neurode models for the property-based descriptors gave better training and generalization statistics than the single neurode model. This suggests that, for this data set, the more parsimonious descriptors needed a more complex neural network to map the structure-activity relationship.

Training on the full data set containing all nonamer peptides (Table 3) produced models with better training and generalization statistics for the bin20 descriptors compared with the property-based descriptors and little difference between the two neural network architectures.

3.2. Performance of models on external test sets

The eight models derived from each of the full training data sets (101 and 301) (Table 3) were used to predict the activity classes of the external test sets (set1-101, set1-301, set2-101 and set2-301). None of the nonamers in the training sets appeared in the external test sets.

To predict the activity of each peptide in the external test sets, we generated all possible contiguous nonamers in the peptide and scored them using the relevant model. The highest-ranking nonamer was used as the activity of the peptide from which it was derived. The predicted activity classes were compared to the experimentally determined binding classes. Tables 4 and 5 summarize the classification performance of each of the eight models on the relevant external test sets. Table 4 shows the results of the most rigorous test of performance of the models on the independent, external test set – a four level classification (N, L, M, H). The results show that the models were able to correctly classify up to 60% of peptides. Allowing one mismatch in classification for decision boundary ambiguity (e.g. H classified as M), up to 100% (30 peptide test set) and 83% (343 peptide test set) of peptides to be correctly classified. The 101 and 301 external test data sets were predicted equally well by the models. The two hidden layer neurode models were, on average, equally successful at correctly classifying the test sets compared with the single hidden layer neurode models. The bin20 and property-based descriptors were also equally good at classifying peptides in the external test sets.

Table 4
Percentage of external test set peptides classified into correct class (column 0), or misclassified by ±1 for models generated using alleles, one and two neurodes in hidden layer, and two combinations of descriptors (four-class classification)

Model	N_{hidden}	Test set 1 (30 peptides)		Test set 2 (343 peptides)	
		0	±1	0	±1
101 test sets					
Bin20	1	60	73	53	78
Property	1	60	80	55	84
Bin20	2	60	70	48	72
Property	2	53	66	49	79
301 test sets					
Bin20	1	50	100	55	76
Property	1	57	90	44	83
Bin20	2	53	100	49	74
Property	2	47	93	40	83

Table 5
Percentage of each external test set activity class correctly predicted (two class model – binder/non binders (see text)) by the models

Model	N_{hidden}	Test set 1 (30 peptides)				Test set 2 (343 peptides)			
		N	L	M	H	N	L	M	H
101 test set									
Bin20	1	100	33	0	25	93	20	31	48
Property	1	100	33	0	100	87	19	43	56
Bin20	2	94	33	0	25	79	29	44	63
Property	2	78	33	0	75	75	33	52	56
301 test set									
Bin20	1	83	55	100	100	74	46	56	56
Property	1	92	55	67	100	60	36	62	31
Bin20	2	92	36	100	100	70	50	54	56
Property	2	100	18	33	100	51	42	68	50

Explanations of terms as in Table 4.

Table 5 shows the results of a less rigorous, but more 'real world' test of the performance of the models – a two-class classification (binders and non-binders). The Table summarizes the percentage of each class (N, L, M, H) correctly predicted by the model. The procedure of Brusic et al. [15] was adopted where a binary activity classification was adopted to score whether a molecule was corrected predicted. This procedure assumes that for peptides experimentally determined to be non-binders only those predicted as non-binders are counted. For peptides experimentally measured as low binders (L), any peptides predicted as active (L, M, or H) count as a correct prediction. Similar two-category methods were used for medium (M) and high (H) binders. This is equivalent to allowing up to a two class misclassification within the active predictions. Such predictions would be useful in a real screening situation as any peptides predicted to have activity would be further tested in an in vitro MHC binding assay.

Table 6 shows results of a truth table analysis of the traditional two class results (binders versus non binders). Results were converted into a 2 × 2 truth table consisting of non-binders correctly predicted (true negatives, TN), binders correctly predicted (true positives, TP), binders incorrectly predicted as non-binders (false negatives, FN), and non-binders incorrectly predicted as binders (false positives, FP). The numbers of peptides in each part of the truth table were converted into figures of merit [6] for comparison of the models. Sensitivity is the proportion of all binders correctly predicted, Se = TP/(TP + FN). Specificity is the proportion of non-binders correctly predicted, Sp = TN/(TN + FP). Positive predictive value is the probability that a predicted binder is in fact a binder, PPV = TP/(TP + FP). Negative predictive value is the probability that a predicted non-binder is in fact a non-binder, NPV = TN/(TN + FN).

Table 6
Truth table statistics for external test sets for two class prediction – binders/non binders) for various models

Model	N_{hidden}	Test set 1 (30 peptides)					Test set 2 (343 peptides)				
		Se	Sp	PPV	NPV	Acc	Se	Sp	PPV	NPV	Acc
101 test sets											
Bin20	1	16	100	100	64	67	28	93	78	60	63
Property	1	42	100	100	67	77	33	87	68	60	62
Bin20	2	16	94	67	63	63	39	79	62	60	61
Property	2	33	78	50	64	60	42	75	59	60	60
301 test sets											
Bin20	1	72	83	87	67	77	51	74	57	69	65
Property	1	67	92	92	65	77	45	60	43	62	54
Bin20	2	61	92	92	61	63	52	70	54	68	63
Property	2	30	100	100	52	63	52	51	42	61	52

Se, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value; Acc, accuracy (see text).

Accuracy is the percentage of all predictions that are correct, Acc = (TP + TN)/total (Table 6).

4. Discussion

4.1. Training

This peptide classification problem is of substantially different character to the type of modelling to which Bayesian neural nets have been applied previously. However the results show that the method is capable to producing good, predictive models of MHC-peptide binding, due to its ability to deal with nonlinear response surfaces and interactions between components.

The quality of all models produced, as assessed by the training and internal test sets statistics, is high. Some general conclusions can be drawn on the effects of neural net architecture, classification method, and amino acid descriptors on the model training.

Models with A_{ROC} values of 0.88–0.91 for the internal test set, and 0.99–1.00 for the training set (A_{ROC} calculated from the predicted activities of the training set) are achievable. We did note several large outliers in some models, prompting a recheck of these points for typographical or measurement errors.

Table 7
Examples of variation in motif binding activity

Test data set	Peptide (motif in bold)	IC_{50} (nM)	Class
A 30 peptides	TISSYF**VGKMYFNLI**DTK	10000	L
	YF**VGKMYFNLI**DTKCYKL	300	H
A 343 peptides	GGGQI**VGGVYLLPR**R	100000	N
	VGGVYLLPRRGPRLG	14000	L
	HPELI**FDITKLLIA**I	100000	N
	FDITKLLIAILGPLM	1500	M
	SRCWV**ALTPLAARN**	35	H
	ALTPLAARNVTIPT	90000	L

The peptide classification problem is not a good example for assessing the influence of neural net architecture on performance. Unlike Gulukota et al. [6] we found one neurode in the hidden layer was usually sufficient to classify the peptides, with the use of two hidden layer neurodes producing models of similar or occasionally lower quality. The response surface is probably quite complex, as shown by the tendency of peptides in the 'H' category to be slightly under predicted by networks with a single hidden layer neurode (see graphs in training section). Two neurodes in the hidden layer removed this under prediction, but did not substantially improve the model overall. An additional factor is the use of indicator variables as descriptors. The fact that the descriptors can only adopt values of 0 or 1 can reduce the influence of higher order terms. The flexibility of neural nets as general classifiers in problems with complex response surfaces is well illustrated here.

The property descriptors tended to produce less statistically significant training models and internal test set statistics. The observation that models using these descriptors sometimes performed better with more complex neural networks architectures suggests they require a more flexible modelling method to take into account a larger contribution from cross terms or nonlinearity than with the bin20 descriptors. As these descriptors were property-based, they may generalize better in 'peptide space' than other models that learn the associations between motifs and activity. As illustrated in the external test set discussion, this superior generalization can compensate for a slightly worse model from the training data.

4.2. Prediction of external test sets

There are $20^9 \sim 10^{11}$ possible nonamer peptides if all 20 naturally occurring amino acids are allowed in the binding motifs. Given that the training set size is 10^3, the ability to make reliable predictions in this much larger 'nonamer' space is a stringent test of the predictive power of any peptide QSAR model. Our models were able to usefully predict the classifications of peptides not used in the training and internal test set procedures. In addition, as the nonamer motifs used in training do not appear in any of the external test sets, this is an excellent assessment of the 'blind' predictive ability of the models. It must be stressed that prediction of such external test sets is a much more rigorous measure of the predictivity of the model than predicting activities of peptides used in commonly-used procedures such as cross validation [18]. This needs to be considered when comparing our statistics on the accuracy of external test set predictions with statistics from a study where cross-validation was used.

All four-class models gave good predictions (see Table 4). The best models were able to correctly classify 60% of peptides in the small 101 external test set, 55% in the large 101 external test set, 53% of peptides in the small 301 external test set, and 55% of peptides in the large 301 external test set. When a single category mismatch was allowed (e.g. M classified as H or L), these figures increase to 80, 84, 100 and 83%, respectively.

The performance of the models in the less rigorous two-category classifications (Table 5) was also very good given the small size of the training set relative to the prediction space. The models were able to correctly predict 100% of non-binding peptides in the small 101 test set, 93% of non-binding peptides in the large 101 test set, 100% of non-binding peptides in the small 301 test set, and 76% of non-binding peptides in the large 301 test set. Table 5 shows that the models can correctly predict 100% of 'H' binding peptides in the small 101 test set, 63% of 'H' binding peptides in the large 101 test set, 100% of 'H' binding peptides in the small 301 test set, and 63% of 'H' binding peptides in the large 301 test set. Considering the goodness of prediction measures summarized in Table 6, it appears that two hidden layer neurode models have better sensitivity than the single hidden layer neurode models for the larger external test sets (343 peptides) for which conclusions are more valid. This suggests that the more complex neural network architecture is better at eliminating false negatives. However, the reverse if true for specificity, which is a measure of how well false positives are eliminated, but the specificity of all models is still good. All models have similar abilities to predict non-binders correctly (NPV) but their abilities to predict binders (PPV) varies markedly, from a low of 42% for property-based descriptors in the 301 test set, to a high of 78% for bin20 descriptors in the 101 external test set. The prediction accuracies for all models are similar for the 101 external test set and the bin20 descriptors in the 301 external test set, with the property-based descriptors for the 301 external test set showing lower accuracy. For the two smaller external test sets (30 peptides), the prediction accuracies of all models were similar.

4.3. Comparison with other MHC peptide QSAR models

Gulukota et al. compared sequence motifs with a polynomial method and standard backpropagation neural network for modeling a set of 463 nonamers binding to the MHC class I HLA-A2.1 allele [6]. The neural network used by these authors employed a single hidden layer with 50 neurodes. As in our work, they also used one hundred and eighty bin20 descriptors to encode the properties of the nonamer peptides and they used activity cutoffs (IC_{50} of 500 nM) to define two classes – binders/non binders. Their modeling is problematic in that their network used over nine thousand weights ($181 \times 50 + 51$) assuming biases were used in the network. Given the size of their training set (approx. 200–300), this is a very large number of adjustable weights in the model. Curiously, these authors identified this overfitting problem in their discussion of possible methods of relaxing the independent binding approximation, but have not recognized the potential for the same problem with neural networks. Our data sets are larger than those modelled by Gulukota et al. and we have restricted the number of hidden layer neurodes to one or two to ensure that the number of weights is considerably less than the number of training set peptides. These authors found the polynomial and neural network methods superior to structural motifs for prediction, and polynomial methods complementary to neural networks in terms of sensitivity and specificity. For the neural network they typically obtained sensitivities of 45% PPV of 60% and accuracies of 70% for binding to this class I allele. Results for the polynomial method depended on the threshold chosen but gave sensitivities of 50–80%, PPV of 15–35% and accuracies of 60–80%. In comparison, our best Bayesian neural net models had sensitivities of 52%, PPV of 78% and accuracies of 65% for a larger data set and the more complex class II binding regime. Although the results for our test set predictions are similar or superior to (depending on the criteria used) those of Gulukota et al., closer comparisons of the methods are problematic because of concerns about the predictivity of their models due to overfitting, and the fact that Gulukota et al. used a different MHC class and HLA allele to us.

The recent paper by Doytchinova, Blythe and Flower is relevant to our work because they have used the same bin20 peptide representation we employed [3]. In addition, like Gulukota et al., these authors have attempted to relax the independent binding approximation by including interaction terms between amino acids at a given position and nearest, or next nearest neighbors. Again this work is based on an MHC class I allele, not class II as in our work. The ability of neural networks to function as universal approximators and account for nonlinear and interaction terms is their major strength in building predictive models of MHC peptide binding. In Doytchinova, Blythe and Flower's work, they attempted to account for interactions by explicitly including the required interaction terms in a linear model. As the number of possible terms and cross terms was very large (6180) they employed partial least squares methods to build a linear model of the peptide binding so as to avoid overfitting. They obtained reasonable quantitative predictions of peptide binding IC_{50} values with a mean residual of 0.51 log units. However, their method still required subjective decisions as to which interactions were important to the model, and the model was still linear. Our neural network models can account for interaction and nonlinear behaviour inherent in the data in a non-subjective way as we explained above. Comparisons between our work and that of Doytchinova, Blythe and Flower is difficult because they used continuous rather than categorical data, a different MHC class and HLA allele, and they did not use an independent, external test set which allowed comparison. However, their leave-one-out cross validation statistics are relatively poor ($q^2 \sim 0.3$), suggesting that the model would have poor predictivity for a test set.

Brusic et al. reported a neural net study of MHC class II binding which most directly comparable to our study [17]. They trained a standard backpropagation neural network using the same type of peptide representation we used. They used an independent test set of 63 peptides to quantify the prediction accuracy of their best models. They obtained an accuracy of 52% when all peptides were considered, and 33% when predicting only the categories of the peptides exhibiting some degree of binding. This compares with accuracies of 65% for our best models using a larger external test set of lengths up to 25 peptides.

An interesting recent study was reported by Bhasin and Raghava, who employed support vector machines to classify peptides binding to the MHC class II allele, HLA-DRB1*0401 [8]. SVMs represent an excellent nonlinear method that projects the classification problem into a higher dimensional space in which a classification boundary can be found. These authors claimed an 86% accuracy of classification for peptides binding to this allele using a five-fold cross validation method which is not as rigorous as the internal and external test sets used in our study. Comparing our external test set prediction results shown in Table 6 with Bhasin and Raghava's SVM cross validation results shows that our method has similar or superior accuracy, NPV, PPV, sensitivity and specificity when predicting the small (30 peptide) external test sets. Prediction of the large external test sets yields superior specificity and comparable PPV and NPV to those published by Bhasin and Raghava, but lower values of accuracy and sensitivity. However, their cross-validated results are likely to overestimate the performance on independent test sets as Golbraikh and Tropsha have shown [18]. Our results are also equal to or superior to those of other workers on this allele as summarized in Bhasin and Raghava's paper.

4.4. Factors causing misclassification

4.4.1. Fuzziness of class membership

It is clear that the boundaries between classes are not sharp, and that class membership is a fuzzy set problem. The use of crisp sets to classify what is essentially continuous data results in ambiguities in class membership. This applies to both the training and test sets, and results in some misclassification. Scanning of the predicted class memberships showed a number of examples where peptide motifs had values just under the class boundary and were classified as non-binders, when experimentally they were low affinity binders.

4.4.2. Invariance of nonamer motif

As the activity of peptides in the external test sets was predicted by scoring all possible overlapping nonamers in the peptide and choosing the best, it became apparent that the same active nonamer motifs were occurring in different peptides with different activities. This is illustrated in Table 7 by several examples. It is clear that in some cases the same motif can exist as the highest ranked nonamer in several larger peptides, where those peptides can exhibit activity differing by at least two classes (two to three orders of magnitude in IC_{50}).

Consequently, the assumption that the activity of a nonamer motif was invariant with respect to the rest of the peptide in which it is embedded is not always true, although in the majority of cases it is approximately true. This would explain some of the misclassification.

4.4.3. Diversity and coverage of the training set

As we discussed above, any realistic training set constitutes a very small selection of 'peptide space' (1 in 10^8 in our study). Consequently, the diversity of the training set can have a major influence on the performance of predictive models derived from it. The few examples of highly active peptides in the external test set that are predicted to have no binding attest to these limitations in training set diversity.

4.4.4. Quality of representation and ability to generalize

Clearly, different representations of peptides will produce models with varying abilities to generalize well. Our results with the property-based descriptors show that it is possible to obtain models that generalize well using reduced descriptor sets chosen to capture peptide properties rather than simply indicate the presence or absence of a given amino acid at a certain position. Limitations of descriptors in capturing all of the relevant properties (some of which will clearly depend on three-dimension structures adopted by the nonamer motifs) will cause misclassifications.

4.4.5. Assumption that peptide side chains bind independently

Gulukota et al. [6] have discussed this assumption at some length. They conjecture that the degree to which this assumption is true depends on the level of detail at which prediction is attempted. For binary classification (binding versus non-binding) they found that the assumption appeared to be justified and they proposed ways of refining predictions to relax this independent binding of sub chains assumption. As the number of classes increases it us likely this assumption will be less valid. One advantage of neural networks is that they can accommodate cross terms between descriptors, achieving some degree of relaxation of the strict independent binding assumption.

4.4.6. Relevance of descriptors used

It is clear that the types of descriptors we used are fairly simple. The binary descriptors do no more than identify each amino acid. The property-based descriptors attempt to incorporate molecular properties into the description. There are many ways amino acids could be described and some of these may produce better models. However, our work suggests that using a robust, model-free, nonlinear method to build models relating descriptors to activity can be surprisingly successful, even with relatively simple descriptors. Development of more efficient descriptors is an active area of research.

5. Conclusions

We have applied our Bayesian neural net methods to the modelling of complex MHC class-II binding peptide activity. The quality of the models was assessed via internal test sets randomly partitioned from the data, and external, independent test sets. From the training data sets we have derived robust SAR models. We found that Bayesian neural networks are able to build good predictive models for MHC class II peptide binding. These models are able to make useful predictions of binding activity in a relatively large region of 'peptide space', even when the sequences being predicted have not appeared in the training sets for the models. These models are able to accurately predict the activities of a large number of peptides in the four independent, external test sets. Our methods allow all of the available data to be used in the training, and they greatly minimize overtraining. Models can be developed quickly, robustly and without the need to optimize the neural net architecture. Models derived by these methods would be very useful in rapidly developing T-cell epitopes without the need to screen large libraries of peptides.

Acknowledgement

We thank Dr. Vladimir Brusic (A*Star Laboratories, Singapore) for supplying the data used in this analysis, for his permission to report the results of the modelling, and for his very helpful comments on our MHC peptide binding QSAR modelling studies.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2005.03.001.

References

[1] S. Buus, Description and prediction of peptide-MHC binding: The 'Human MHC Project', Curr. Opin. Immunol. 11 (1999) 209–213.

[2] I.A. Doytchinova, D.R. Flower, Towards the quantitative prediction of T-Cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the Class I MHC molecule HLA-A*0201, J. Med. Chem. 44 (2001) 3572–3581.

[3] I.A. Doytchinova, M.J. Blythe, D.R. Flower, Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class 1 molecule HLA-A*0201, J. Proteome Res. 1 (2002) 263–272.

[4] A. Logean, A. Sette, D. Rognen, Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions, Bioorg. Med. Chem. Lett. 11 (2000) 675–679.

[5] V. Brusic, K. Bucci, C. Schönbach, N. Petrovsky, J. Zelezvikow, J.K. Kazura, Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding, J. Mol. Graph. Modell. 19 (2001) 405–411.

[6] K. Gulukota, J. Sidney, A. Sette, C. DeLisi, Two complementary methods for predicting peptides binding major histocompatibility complex molecules, J. Mol. Biol. 267 (1997) 1258–1267.

[7] E.C. De Hann, M.H.M. Wauben, M.C. Grosfeld-Stulemeyer, J.A.W. Kruijtzer, R.M.J. Liskamp, E.E. Moret, Major histocompatibility complex class II binding characteristics of peptoid-peptide hybrids, Biorg. Med. Chem. 10 (2002) 1939–1945.

[8] M. Bhasin, G.P.S. Raghava, SVM-based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence, Bioinformation 20 (2004) 421–423.

[9] M.J. Polley, D.A. Winkler, F.R. Burden, Broad-based QSAR of farnesyltransferase inhibitors using a Bayesian regularized neural network, J. Med. Chem. 47 (2004) 6230–6238.

[10] D.A. Winkler, F.R. Burden, Modelling blood brain barrier partitioning using Bayesian neural nets, J. Mol. Graph. Modell. 22 (2004) 499–508.

[11] F.R. Burden, D.A. Winkler, A QSAR model for the acute toxicity of substituted benzenes towards *Tetrahymena pyriformis* using Bayesian regularized neural networks, Chem. Res. Toxicol. 13 (2000) 436–440.

[12] M.J. Sorich, R.A. McKinnon, D.A. Winkler, F.R. Burden, J.O. Miners, P.A. Smith, Comparison of linear and nonlinear classification algorithms: prediction of drug metabolism by UDP-glucuronosyltransferase isoforms, J. Chem. Inf. Comput. Sci. 43 (2003) 2019–2024.

[13] D.A. Winkler, F.R. Burden, Robust QSAR models from novel descriptors and Bayesian regularized neural networks, Mol. Simul. 24 (2000) 243–258.

[14] F.R. Burden, D.A. Winkler, Robust QSAR models using Bayesian regularized artificial neural networks, J. Med. Chem. 42 (1999) 3183–3187.

[15] V. Brusic, G. Rudy, L.C. Harrison, MHCPEP, a database of MHC-binding peptides: update, Nucleic Acids Res. 26 (1998) 368–371.

[16] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.

[17] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, L. Harrison, Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, Bioinformation 14 (1998) 121–130.

[18] A. Golbraikh, A. Tropsha, Beware of q2, J. Mol. Graphics Modell. 20 (2002) 269–276.