Support vector-based fuzzy system for the prediction of mouse class I MHC peptide binding affinity

Volkan Uslan and Huseyin Seker*

Abstract— **The performance of predictive models is crucial in order to accurately determine peptide binding affinity for major histocompatibility complex (MHC) alleles. Data sets extracted to model the relationship between the peptides and their binding affinities are often high-dimensional, complex and non-linear, which require highly sophisticated computational predictive models. Support Vector Machine (SVM)-based predictive methods have been used for such problems and have been shown to deal with such high dimensional data, however failed to take into account of uncertainty that naturally exists in this type of data. In order to address to the uncertainty issue, Fuzzy System (FS) has generally been utilised in various applications. Therefore, a hybrid method that combines FS and SVM is proposed in this study for the prediction of binding affinity of peptides in mouse class I MHC alleles. The hybrid system is successfully applied to two benchmark data sets of class I MHC peptides, each of which contains over 5000 peptide features. The assessments yield as much as 17% improvement over the previous studies that also include SVM-based experiments. The results also suggest positive impact of the concept of fuzziness on SVM-based predictive methods when combined and that the hybrid model can be generalised for similar non-linear system modelling problems.**

I. INTRODUCTION

The T-cell receptor is a molecule, present at the T-cell surface, and signicantly required to activate the T-cell by recognising antigenic peptides bound to major histocompatibility complex (MHC) molecules translocated on to the surface of the infected cells [1]. The peptide epitopes that are bound to MHC class I molecules can be recognised by the T-cells and can induce the cellular immune response. As a consequence revealing the association of peptides with the MHC molecules can be crucial for drug development. A common assessment to elicit these associations is to find peptide binding affinity.

Predicting binding affinity using computational methods is of particular interest in bioinformatics. As there are large number of peptides, laboratory-based experimental evaluation of binding affinity between proteins and peptides are costly, expensive and requires labour and resource [2]. Therefore, as though empirical methods may become inefficient and unfeasible, there is a need to develop a computational predictive model that is capable of determining the tendency and strength of the bindings in order to save time as well as experimental efforts. Wide range of applications in the prediction of peptide binding affinity reported in the literature including partial least squares [3] and artificial neural networks [4]. Nevertheless, recent research efforts have been focused on quantifying the binding predictions [5].

SVM is one of the computational methods that has been shown to effectively deal with large number of dimensions [6]. When the quantitative modelling is the case, SVMs can be extended to SVR with the aid of e-sensitive loss function [7]. SVR has been proven to lead better generalization ability and performance in a wide range of applications [5], [8]. Fuzzy systems is another non-linear method that is good at modelling uncertainty and yielding a set of interpretable if-then rules [9]. On the contrary, fuzzy systems can suffer from the curse of dimensionality in high-dimensional systems. A hybrid learning system is therefore proposed in this paper to train the parameters of the fuzzy system. The main idea behind this method is that the linear parameters of the consequent part of the fuzzy system were obtained by using SVR whereas the parameters of antecedent part of the fuzzy system that characterise each fuzzy set were obtained using the clustering method [10], [11].

SVR-based fuzzy system is generally applied in a data set with a small number of features due to the nature of fuzzy system [10]. Our recent attempt has proven its applicability in relatively large number of features where a peptide binding affinity problem can be successfully studied. Therefore, in order to further assess efficiency of the proposed hybrid method, two mouse class I MHC allele data sets are used to model their MHC-peptide binding affinity. The predictive performances are then compared with the recently published studies. The rest of the paper starts by presenting the characteristics of the MHC class I binding data sets and follows by describing the support vector-based fuzzy system and the performance measurements of the prediction models (Section II). Experimental results are then provided in Section III, and finally the conclusions are drawn in Section VI.

II. MATERIALS AND METHODS

A. *MHC Class I Binding Data Sets*

Proteins degraded into peptides by the proteaosome, and the generated peptides are translocated to the endoplasmic reticulum [12]. These translocated peptides are bind to MHC class I molecules. When the peptides are immunologically active regions and able to induce cellular immune responses, they are called T-cell epitopes. T-cell epitopes were translocated on to the surface of the infected cells so that they can be recognized by a T-cell receptor present at the T-cell surface. The Cytotoxic T-cells are the immune cells that can recognise

[1] The authors are with the Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK, email: vuslan@dmu.ac.uk, hseker@dmu.ac.uk

* Corresponding author (hseker@dmu.ac.uk)

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

TABLE I

GENERAL CHARACTERISTICS OF THE DATA SETS USED FOR THE PREDICTION OF PEPTIDE BINDING AFFINITY FOR MOUSE CLASS I MHC ALLELES

Data sets	Number of Peptide Sequences	Number of Amino Acids	Number of Peptide Sequence Descriptors
H2-D^b	65	9	5787
H2-K^b	62	8	5144

antigenic peptides bound to MHC molecules translocated on to the surface of the infected cells originated from pathogenic organisms like bacteria, fungi, parasites or viruses [13].

As the number of peptides are very large, experimental measurement of their binding affinity is difficult, therefore prediction methods have become increasingly important in the post-genome era. Such predictions help determine the accurate binding affinities of these peptides. Peptide data sets are often high-dimensional ($\sim$5000) and complex and consist of limited number of peptides ($\sim$100).

Mouse class I MHC alleles (H2-Db and H2-Kb) were used in this paper to model their MHC-peptide binding affinities. The peptides in each allele contain experimentally measured binding affinities, numerically as pIC50. Each peptide in the data sets was represented by assigning values of physico-chemical and bio-chemical descriptors to each amino acid. There are 643 descriptors (real values) for each amino acid extracted from the Comparative Evaluation of Prediction Algorithms (CoEPrA) modelling competition [14]. As shown in Table I H2-Db consists of nona-peptides that have a total of 5787 (643×9 = 5787) descriptors and H2-Kb has octa-peptides that have a total of 5144 (643×8 = 5144) descriptors.

B. Support Vector Based Fuzzy System

The rule-base of the Takagi, Sugeno and Kang (TSK) fuzzy system with r rules can be expressed as [15]:

$$\begin{array}{l} R_i : \text{IF } x_1 \text{ is } A_{1i} \text{ AND } x_2 \text{ is } A_{2i} \text{ ... AND } x_n \text{ is } A_{ni} \\ \text{THEN } y_i \ = \ c_{0i} + c_{1i}x_1 + ... + c_{ni}x_n \end{array} \tag{1}$$

where n are the input variables (x_1, x_2, ..., x_n); and A_{ni} is a fuzzy set for the variable n and rule r, generally represented by a membership function; and y_i is a linear function in the consequent part; and c_0, c_1, c_2, ..., c_n are the coefficients of input parameters. The overall output is obtained by weighted average and can be calculated as:

$$y = \sum_{i=1}^{r} \overline{f_i} y_i \tag{2}$$

$$\overline{f_i} = f_i / \sum_{k=1}^{r} f_k \tag{3}$$

where f_i is the firing strength determined by using a t-norm operator and $\overline{f_i}$ is the normalized firing strength. The t-norm operation can be defined as:

$$f_i = \prod_{j=1}^{n} \mu(x_j) \tag{4}$$

where $\mu(x_j)$ is the degree of membership for input variable x_j.

SVM is a statistical learning algorithm that searches for a hyperplane which separates the given training data set optimally [6]. Based on the structural risk minimization, this separating hyperplane will maximize the margin between two classes. SVMs can be extended to regression using the ϵ-insensitive loss function. SVR approximates a linear function $h(x)$:

$$h(x) = w^T x + b. \tag{5}$$

where the coefficients w and b represent the weight vector of the linear expression. The linear function is constrained to:

$$\min \frac{1}{2} \|w\|^2 + C \sum (\xi_+ + \xi_-). \tag{6}$$

where ξ^+, ξ^- are the two types of slack variables. The parameters w and C are regularisation and optimisation factors, respectively. This expression tolerates up to a value which deviates greater than ϵ. The minimisation function defined in Eq.6 is subject to:

$$\begin{array}{l} y' - (w^T x + b) \leq \epsilon + \xi_+ \\ (w^T x + b) - y' \leq \epsilon + \xi_- \\ (\xi_+, \xi_-) \geq 0 \end{array} \tag{7}$$

Support vectors are chosen from the certain training instances and the regression is defined by the weighted sum of the support vectors to adequately model data. TSK is generally benefited from the least squares estimation [16] to design the consequent parameters. Instead, the proposed approach uses the support vector regression concept with a linear kernel. The inputs ($\overline{f_i}$, $\overline{f_i}x_{i1}$, $\overline{f_i}x_{i2}$, ..., $\overline{f_i}x_{in}$) for each data item in the training data set along with its actual output y are given to SVR to compute the TSK consequent parameters from the coefficients w and b which represent the weight vector of the SVR linear expression. The support vector based Takagi-Sugeno-Kang fuzzy system (TSK-SVR) can be formulated as:

$$y'_i = w_{0r} + \sum_{i=1}^{n} (w_{ir} x_i) \tag{8}$$

$$y' = \sum_{i=1}^{r} (\overline{f_i} y'_i + \frac{b}{r}) \tag{9}$$

where y' is the new output formulation that now represents TSK-SVR. In order to implement SVR part of the hybrid method, LIBSVM is utilized [17].

TABLE II

ENTIRE DATA SET PREDICTION RESULTS OF THE MOUSE ALLELES

		Allele		Allele	
		H2-D^b	H2-K^b	H2-D^b	H2-K^b
Methods		q^2	q^2	AR	AR
Additive	[18]	0.602	0.370	0.403	0.443
SVRMHC	[5]	0.749	0.568	0.170	0.382
RVMMHC-1	[8]	0.840	0.664	0.297	0.527
RVMMHC-2	[8]	0.845	0.691	0.316	0.489
TSK-SVR		**0.912**	**0.792**	**0.140**	**0.340**
Improvement		7.93%	14.62%	17.65%	10.99%

TABLE III

LEAVE-ONE-OUT CROSS VALIDATED CORRELATION COEFFICIENT PREDICTION RESULTS OF THE MOUSE ALLELES

		Allele	
		H2-D^b	H2-K^b
Methods		q^2	q^2
Additive	[18]	0.401	0.454
SVRMHC	[5]	0.456	0.486
TSK-SVR		**0.462**	**0.490**
Improvement		1.32%	0.82%

C. Performance Measurements of the Prediction Models

Two performance measurements, namely the coefficient of determination (q^2) and average residual (AR) were used for the predictive models in order to keep the assessments consistent over the published results. q^2, a statistical model based upon the proportion of variability in a data set, is defined as [19]:

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(pIC50_i - pIC50_i^*)^2}{\sum_{i=1}^{n}(pIC50_i - \overline{pIC50_i})^2} \tag{10}$$

where $pIC50_i$ and $pIC50_i^*$ are the expected and predicted values of the peptide, respectively. $\overline{pIC50_i}$ denotes the average of expected values in the prediction data set. As q^2 gets closer to one, the model is better constructed. On the contrary, if it yields negative values, this indicates that the model has poorly approximated the expected values. The other measure, AR is expressed as:

$$\text{AR} = \frac{\sum_{i=1}^{n}|pIC50_i - pIC50_i^*|}{n} \tag{11}$$

where n is the number of peptides in the allele. A successful prediction can be achieved with lower values of AR whereas its higher values show poorer predictions.

III. RESULTS AND DISCUSSION

The analysis of the model requires an efficient preprocessing of features for the allele data sets. In general, physico-chemical attributes of the data sets are represented by real numbers. The features were normalized to [0, 1] so that every feature fall within the same range of values. In order to obtain attributes that have more representative capability of the underlying information, priorly, a feature selection method, namely Multi-Cluster Feature Selection method was used [20]. The normalization and feature selection were necessary to keep the low-dimensionality and support the robustness of the TSK fuzzy system. TSK fuzzy system was constructed using only two rules with the reduced features. These rules are suffice for the proposed model to build a robust and interpretable fuzzy system for the high-dimensional data set by using relatively small number of data samples. The structure of the TSK fuzzy system is constituted by automating the parameters of the antecedent and consequent parts. Values of the parameters of Gaussian membership functions that characterise each fuzzy set of the premise part were obtained by using Fuzzy c-Means clustering method [21]. The coefficients of linear functions of each rule for the consequent part were then identified using SVR.

The performance of the proposed approach (TSK-SVR) that models the relationship between the peptides and their binding affinities was evaluated using MHC alleles. In Table II, it can be seen that two different measures were used to observe their influence on the prediction error. The prediction results are comparatively better than those of the studies presented in [18], [5] and [8] for MHC alleles H2-Db and H2-Kb. The predictive performances have been improved by 7.9% (q^2) and 17.6% (AR) for the H2-Db allele; and 14.6% (q^2) and 10.9% (AR) for the H2-Kb allele. It should be noted that our literature search appears to indicate that these two data sets have been understudied due to their complexity, therefore not many papers other than the cited ones seem to have appeared in the literature.

In order to further explain the results, the construction of correlation diagram for each allele data set is used to illustrate the relationship between the experimentally measured and predicted pIC50 values. When the performance is perfect, the correlation diagram shows a straight line along the 45° diagonal. A good quality of prediction performance can be obtained when the data samples are mainly distributed along the 45° diagonal. The divergence in the line is caused

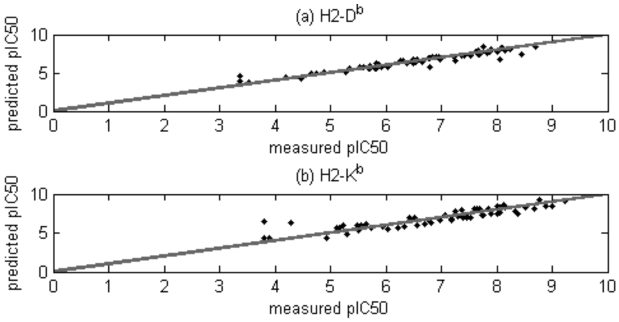

Fig. 1. Correlation diagrams for the prediction performance of mouse alleles. a) H2-Db b) H2-Kb

by the prediction error between the measured and the predicted pIC50 values.

In order to avoid the problem of overfitting the parameters need to be selected properly. Grid-search is one of the simple and reliable methods for this purpose and allows parallel computations to speed up the calculations. The optimal parameters for the MHC alleles using the q^2 measure are found to be $C = 75.0$, $\epsilon = 0.20$ for H2-Db, and $C = 25.0$, $\epsilon = 0.50$ for H2-Kb. The models contained 30 and 25 features for each MHC allele, respectively. The average residual (AR) measure values of the proposed model are: $C = 9.10$, $\epsilon = 0.05$ for allele H2-Db; and $C = 8.70$, $\epsilon = 0.05$ for allele H2-Kb. The final and refined models contained 39 and 24 features, respectively.

In addition, each model was evaluated by using leave-one-out cross validation (LOO-CV) using the cross-validated correlation coefficient. This will allow an independent predictive assessment as compared to the assessment carried out using the entire data set. As the compared methods presented in the literature did not report an LOO-CV AR measure, this assessment was excluded from the calculations. The additive method recognized 6 outliers for H2-Db and 7 outliers for H2-Kb and they have been removed before the LOO-CV calculations. These outliers were also excluded from the proposed models during the LOO-CV calculations in order to perform a consistent comparison. The optimal parameters for the MHC alleles using the q^2 measure are found to be $C = 0.45$, $\epsilon = 0.05$ for H2-Db, and $C = 3.10$, $\epsilon = 0.05$ for H2-Kb. The models contained 34 and 21 features for each MHC allele, respectively. As shown in Table III the proposed models yielded LOO-CV q^2 values of 0.462 and 0.490 which are higher predictive accuracy than the additive and SVRMHC methods. The cross-validated results suggest that a better descriptive power has been achieved over the unseen data indicating better generalisation ability of the proposed hybrid method. In addition, the incorporation of fuzzy system with SVR has enabled to improve SVR and consequently resulting in a better modeling of uncertainty even the model can only use small sample size being the nature of peptide data. As stated above, the fuzzy if-then rule set produced has been found to be useful and, due to the space limitation in the paper, will be discussed during the presentation.

IV. CONCLUSIONS

In this paper, TSK-SVR hybrid system was developed and successfully applied in the prediction of binding affinity of the mouse class I MHC alleles peptides, which is regarded as one of the complex modeling problems in the post-genome era.

The consequent and antecedent parameters of the Takagi-Kang-Sugeno Fuzzy System were designed using the SVR and Fuzzy C-Means clustering method, respectively. The results show that as much as 17% improvement was achieved over the previous studies that also include the results obtained from the SVR-based experiments. The SVR-based learning for TSK was shown to lead a better generalization and this achievement clearly highlights that the SVR is benefited from the inclusion of the fuzziness concept.

The promising results obtained show the potential application of TSK-SVR to other quantitative and complex modeling problems of bioinformatics. As a future work, TSK-SVR will be extended to a type-2 fuzzy system along with other similar bioinformatics applications in order to see whether predictive performance can further be improved.

REFERENCES

[1] P. A. Moss, W. M. Rosenberg, and J. I. Bell, "The human t cell receptor in health and disease," *Annual Review of Immunology*, vol. 10, no. 1, pp. 71–96, 1992.

[2] W. W. P. Liao and J. W. Arthur, "Predicting peptide binding to major histocompatibility complex molecules," *Autoimmunity Reviews*, vol. 10, no. 8, pp. 469–473, 2011.

[3] R. Bremel and E. J. Homan, "An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches," *Immunome Research*, vol. 6, no. 1, p. 7, 2010.

[4] S. Buus, S. L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak, "Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach," *Tissue antigens*, vol. 62, no. 5, pp. 378–384, 2003.

[5] W. Liu, X. Meng, Q. Xu, D. Flower, and T. Li, "Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models," *BMC Bioinformatics*, vol. 7, no. 1, p. 182, 2006.

[6] V. N. Vapnik, "An overview of statistical learning theory," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 988–999, 1999.

[7] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, *Support Vector Regression Machines*, ser. Advances in Neural Information Processing Systems. MIT Press, 1996, vol. 9.

[8] D. Li and W. Hu, "A relevance vector machine based quantitative prediction method for mouse class I MHC peptide binding affinity," in *International Conference on Biomedical and Pharmaceutical Engineering. ICBPE*, 2006, pp. 349–353.

[9] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[10] C.-F. Juang and C.-D. Hsieh, "A fuzzy system constructed by rule generation and iterative linear SVR for antecedent and consequent parameter optimization," *Fuzzy Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 372–384, 2012.

[11] V. Uslan and H. Seker, "Support vector-based Takagi-Sugeno fuzzy system for the prediction of binding affinity of peptides," in *the 35th Annual Inter. Conf. of IEEE EMBS*, 2013.

[12] J. Banchereau and R. Steinman, "Dendritic cells and the control of immunity," *Nature*, vol. 392, no. 6673, pp. 245–252, 03 1998.

[13] M. Bhasin and G. P. S. Raghava, "Analysis and prediction of affinity of tap binding peptides using cascade SVM," *Protein Science*, vol. 13, no. 3, pp. 596–607, 2004.

[14] O. Ivanciuc, "Comparative Evaluation of Prediction Algorithms (CoEPrA)," 2006. [Online]. Available: http://www.coepra.org/

[15] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. SMC*, vol. 15(1), pp. 116–132, 1985.

[16] J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. SMC*, vol. 23(3), pp. 665–685, 1993.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intel. Sys. and Tech.*, vol. 2(3), pp. 1–27, 2011.

[18] C. K. Hattotuwagama, P. Guan, I. A. Doytchinova, and D. R. Flower, "New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity," *Org.Biomol.Chem.*, vol. 2, no. 22, pp. 3274–3283, 2004.

[19] R. G. D. Steel, J. H. Torrie, and D. A. Dickey, *Principles and Procedures of Statistics*. McGraw-Hill, 1997.

[20] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *16th ACM SIGKDD international conference on knowledge discovery and data mining*, 2010, pp. 333–342.

[21] J. C. Bezdek, "FCM: The fuzzy c-means clustering algorithm," *Computers Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.