Quantitative prediction of peptide binding affinity by using hybrid fuzzy support vector regression

Volkan Uslan[a,b], Huseyin Seker[c,*]

[a]*Bio-Health Informatics Research Group, Centre for Computational Intelligence, De Montfort University, Leicester, LE1 9BH, The United Kingdom*
[b]*Faculty of Engineering, Mevlana (Rumi) University, Konya, Turkiye*
[c]*Faculty of Engineering and Environment, The University of Northumbria at Newcastle, Newcastle-upon-Tyne, NE1 8ST, The United Kingdom*

Abstract

Support Vector Machines has a wide use for the prediction problems in life sciences. It has been shown to offer more generalisation ability in input-output mapping. However, the performance of predictive models is often negatively influenced due to the complex, high-dimensional, and non-linear nature of the post-genome data. Soft computing methods can be used to model such non-linear systems. Fuzzy systems are one of the widely used methods of soft computing that model uncertainties. It is formed of interpretable rules aiding one to gain insight into applied model. This study is therefore concerned to provide more interpretable and efficient biological model with the development of a hybrid method that integrates the fuzzy system and support vector regression. In order to demonstrate the robustness of this new hybrid method, it is applied to the prediction of peptide binding affinity being one of the most challenging problems in the post-genomic era due to diversity in peptide families and complexity and high-dimensionality in the characteristic features of the peptides. Having used four different case studies, this hybrid predictive model has yielded the highest predictive power in all the four cases and achieved an improvement of as much as 34% compared to the results presented in the literature. (The algorithm's user friendly MATLAB code will be made available at our research group's website).

Keywords: Fuzzy Systems, Support Vector Regression, Peptide Binding Affinity

1. Introduction

Peptide binding plays vital roles in the molecular biology of the cell. The process of the peptide binding can activate the cytotoxic T-cells in the immune system [1]. One of the most challenging and complex aspect of the peptide binding is the prediction of protein-peptide binding affinity. These bindings are very crucial in that they induce cellular immune

*Corresponding author
Email addresses: vuslan@mevlana.edu.tr (Volkan Uslan), huseyin.seker@northumbria.ac.uk (Huseyin Seker)

responses [2]. On the other hand, due to diversity of peptide families, there are quite large number of peptides available and still being discovered (e.g., potentially over 512 billion peptides for each MHC molecule [3]).

Biological experiments for the measurement of the binding affinity between proteins and peptides are costly and time-consuming. In this regard, computational methods are of particular interest in bioinformatics for finding feasible approaches to this problem [4], [5]. Predictive models in the identification of peptide binding affinity are often used to find out whether a binding exists between peptide and MHC molecule [6]. The qualitative models further improved and focused on modeling to classify binders as strong and weak binders [7], [8], [9]. Recent research efforts have been focused on quantifying the binding predictions. Additive model is one of the earliest quantitative approaches that is proposed to model MHC-peptide for finding precise binding affinities [10]. After that, studies are focused on non-linear approaches and they achieved a better performance compared to linear models such as the additive method. Non-linear modelling approach has been taken by a number of later methods such as regularization methods [11], partial least squares [12] and random forests [13] to reveal the real-value of the binding affinity. However, complexity and nonlinearity that exist in such data sets have led the necessity of more robust and sophisticated methods.

Fuzzy systems are able to model uncertain and imprecise knowledge in complex and non-linear data sets, and form a structure for representing human reasoning. Among various fuzzy systems, Takagi-Sugeno-Kang (TSK) is commonly used for modelling complex systems [14], [15]. TSK fuzzy systems (TSK-FS) can be combined with other methods, particularly learning methods, and enhanced with learning and adaptation capabilities [16]. In TSK models, rule antecedent is in the form of membership functions and the rule consequent is a linear function of inputs. Although there are many methods proposed to model TSK-FS, general approach is to keep the premise parameters constant whereas values of the consequent parameters are computed. This computation is done by least square estimation (LSE) which is a statistical modeling that assumes a linear relationship that exists between input and output variables. LSE is based on the minimising the empirical risk and constitutes an essential part of the TSK fuzzy systems [17], [18]. One drawback of least squares learning algorithm is that even though the training error is minimised, the model can badly suffer from the overfitting [19]. However, there are methods that have been explored for addressing the problems in the least square estimation (e.g., neuro-fuzzy systems [17], genetic-fuzzy systems [20]).

Support vector regression (SVR) [21], [22] is an efficient and robust method and provides high generalizability and performance. Applications of SVR have demonstrated considerably better modelling of various non-linear systems and minimising the structural risk than least squares approach. It is considered that, this concept can be incorporated with TSK-FS in order to better train its consequent part [23]. However, there are not many methods reported in the literature for the utilisation of support vector based methods at the consequent part of the fuzzy system [24], [25], [26].

In this paper, a support-vector based Takagi-Sugeno-Kang fuzzy system (TSK-SVR) is proposed and applied to the quantitative prediction of the binding affinities between major

histocompatibility complex proteins (MHCs) and peptides which is an important problem in biology and medicine with applications for drug design. This paper extends the initial work [27] and improves initial results by yielding as much as 34% improvement in prediction accuracy than what has been presented in the recently published papers. The rest of the paper is organised as follows. Section II introduces the peptide binding affinity problem. In Section III background methodology is explained. Section IV presents the SVR-based TSK type-1 fuzzy prediction model. Section V presents the results and discussion. Finally, conclusions are drawn in Section VI.

2. Peptide Binding Affinity

This section presents the problem statement and data sets to be used.

2.1. Problem Statement

A peptide presented by MHC class I molecules is a short number of amino acid sequence that generally contain eight to eleven amino acids [28]. Peptides bind to protein molecules in order to induce cellular immune responses. Affinity indicates the tendency or strength of the binding. As there is a quite larger number of peptides (potentially over 512 billion binding peptides for each MHC molecule [3]), there is a need for prediction methods to help determine binding affinities of these peptides. In addition, in order to avoid this time consuming task, a computational predictive model should be developed. The difficulty of the peptide prediction problems when building a prediction model is the number of features being very large (in this study ≥ 5000) whereas the number of peptides in the training data set is relatively small (in this study ≤ 150).

2.2. The Data Sets

The high-dimensional peptide data sets provided at the Comparative Evaluation of Prediction Algorithms (CoEPrA) modeling competition [29] were used in this study in order to further improve predictivity of the affinity of peptides and, in particular, to test predictive capability of the proposed TSK-SVR model for the given data sets. As shown in Table 1 each task contains calibration (training) and prediction (test) data sets and physico-chemical descriptors have been provided for each small peptide (for both calibration and prediction data sets).

In addition to two different amino acid data sets used in the literature that consists of physico-chemical and bio-chemical properties of amino acids (e.g., AAindex database [30] and CISAPS [31]), to be consistent with the CoEPrA, each amino acid in a peptide is described by 643 descriptors. It should be noted that, these descriptors were picked mostly from AAindex database. Task 2 consists of octa-peptides that have a total of 5144 ($643\times8 = 5144$) descriptors whereas all other tasks have nona-peptides that have a total of 5787 ($643\times9 = 5787$) descriptors (Table 1). The statistics (range, mean and standard deviation) of the binding affinities of the peptides of each task are given in Table 2.

Table 1: General Characteristics of the data sets used for the prediction of peptide binding affinity.

Data Sets	Number of Peptide Sequences		Number of Peptide Sequence Descriptors
	Training	Testing	
Task 1	89	88	5787
Task 2	76	76	5144
Task 3	133	133	5787
Task 4	133	47	5787

Table 2: The statistical characteristics of the values of peptide binding affinities.

Data Sets	Training				Testing			
	Min	Max	Mean	Std	Min	Max	Mean	Std
Task 1	2.94	8.65	5.41	1.01	3.13	8.17	5.41	0.95
Task 2	5.01	8.34	7.55	0.77	5.01	8.40	7.58	0.74
Task 3	4.30	8.77	7.08	0.82	5.08	8.96	7.10	0.80
Task 4	4.30	8.77	7.08	0.82	13	121	60.96	33.94

3. Background Methodology

The proposed approach consists of a number of components to be implemented for the prediction of peptide binding affinity. This section will provide background information related to these components.

3.1. Takagi-Sugeno-Kang Fuzzy System

The Takagi-Sugeno-Kang (TSK) fuzzy system rules are defined as conditional statements that are presented by using a linear function in the consequent part. A fuzzy rule-base with n input variables (x_1, x_2, ..., x_n), r rules can be written as:

$$\begin{aligned} &R_r : \text{IF } x_1 \text{ is } A_{1r} \text{ AND } x_2 \text{ is } A_{2r} \text{ ... AND } x_n \text{ is } A_{nr} \\ &\text{THEN } y_r \ = \ f(x_1, x_2, ..., x_n) \end{aligned} \tag{1}$$

where A_{nr} is a fuzzy set for the input variable n and rule r, generally represented by a membership function, and y_r is a linear function in the consequent part and can be defined as:

$$y_r = f(x_1, x_2, ..., x_n) = m_0 + \sum_{i=1}^{n} (m_i x_i) \tag{2}$$

where m_0, m_1, m_2, ..., m_n are the coefficients of input parameters (x_1, x_2, ..., x_n). In the TSK model each rule generates a crisp output and then the final output is obtained by aggregating all the rule outputs. This process is called defuzzification, and the weighted average defuzzification value y can be defined as:

$$y = \sum_{i=1}^{r} \overline{f_i} y_i \tag{3}$$

$$\overline{f_i} = f_i \,/\, \sum_{k=1}^{r} f_k \tag{4}$$

where f_i and $\overline{f_i}$ are the firing strength and normalized firing strength of the fuzzy rule, respectively, and f_i is determined by using a t-norm operator that can be defined as:

$$f_i = \prod_{j=1}^{n} \mu(x_j) \tag{5}$$

where $\mu(x_j)$ is the membership degree of input variable x_j. The fuzzy sets (e.g., A_{ij}) can be described by any form of membership functions. In this study, Gaussian membership function is used and can be defined as:

$$\mu(x_j) = e^{-\frac{(x_j - c_{ij})^2}{2(\sigma_{ij})^2}} \tag{6}$$

where c and σ are the centre and standard deviation, respectively.

3.2. Support Vector Regression

Support Vector Machine (SVM) is a statistical learning architecture based on the structural risk minimization [32]. SVM learning algorithm finds the optimal separating hyperplane by training a classifier for a given training data. The optimal separating hyperplane is the one that maximizes the margin between two classes. SVMs can be generalized to perform regression using its linear model. Other than the traditional square error loss function, the ϵ-insensitive loss function is used in SVR [33]. This chosen error function tolerates errors up to ϵ. One other advantage of using this error function is its tolerance against noise.

SVR searches for a linear function $h(x)$

$$h(x) = w^T x + b. \tag{7}$$

where w and b represent the coefficients of the weight vector of the linear expression. This linear function is constrained to the following mathematical expressions:

$$\min \frac{1}{2}\|w\|^2 + C \sum (\xi_+ + \xi_-). \tag{8}$$

subject to

$$\begin{aligned} & y' - (w^T x + b) \leq \epsilon + \xi_+ \\ & (w^T x + b) - y' \leq \epsilon + \xi_- \\ & (\xi_+, \xi_-) \geq 0 \end{aligned} \tag{9}$$

where two types of slack variables ξ_+ and ξ_- measure the deviations of training samples out of the ϵ-region [22]. The values of these variables are computed during the training of SVR as in (9). The parameter C is a pre-specified value and works as a regularization factor between minimizing w and up to the value which deviations greater than ϵ can be tolerated. Certain training instances are chosen to be support vectors. Then, the weighted sum of the support vectors are used to define the regression and adequately model data.

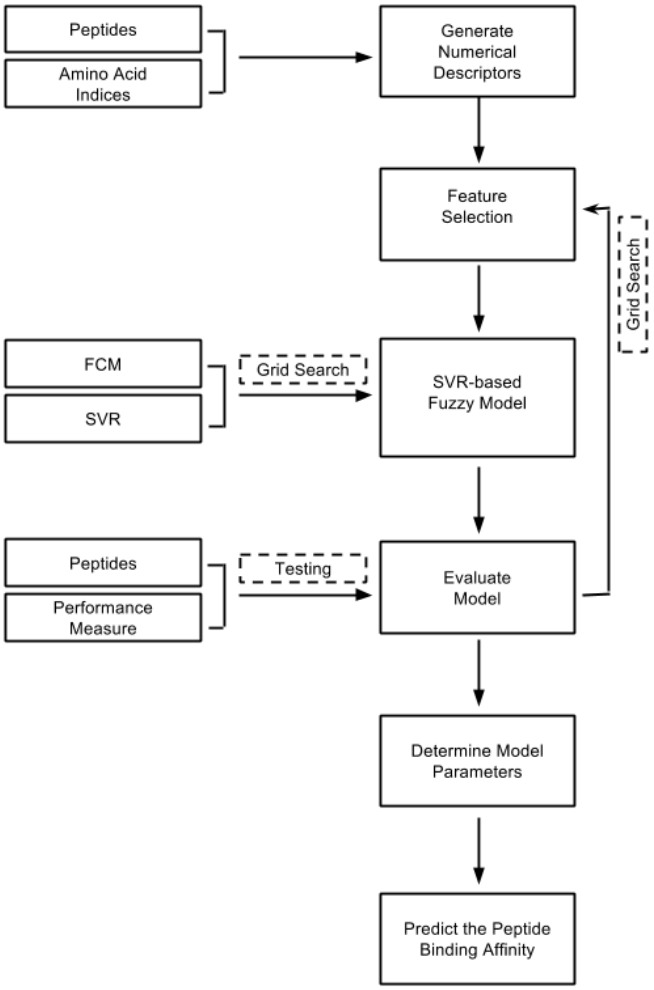

Figure 1: Flowchart of the TSK-SVR model for the prediction of peptide binding affinity.

4. SVR-based TSK Type-1 Fuzzy Prediction Model

This section presents the implementation of the SVR-based TSK type-1 fuzzy prediction model. The flowchart of the proposed model is shown in Fig. 1.

4.1. Preprocessing

The amino acids of the peptides that form the data set turned into numerical descriptors using amino acid indices. Then the analysis started with normalising the data set in order for every feature to fall within the same range of values. The descriptors are normalised to a scale in the interval [0, 1] as expressed in (10)

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{10}$$

4.2. Reducing the High Dimensionality

Feature selection is a process to reduce dimensionality by choosing a subset of relevant features leading to a better performance of the system or the model. In this regard, feature selection algorithms are widely used in bioinformatics aiming at finding the least number of features that improve the accuracy and performance of the models [34]. There are several feature selection methods available. In this study, the problem of feature selection is addressed by utilizing the multi-cluster feature selection (MCFS) [35] for the proposed model as its superiority has recently been shown over different application domains [36], [37], [38]. MCFS is an unsupervised feature selection method and uses information contained in eigenvectors by solving the generalised eigen-problem to preserve the multi-cluster structure of the data. In this study, the number of used eigenvectors parameter of MCFS is set to the the number of features to be selected.

4.3. Identifying Antecedent Parameters

Fuzzy c-Means (FCM) method partitions data set into a number of clusters in a way that each data object is assigned a degree of membership for each cluster [39]. The FCM model aims to minimise an optimisation function. The clustering process iteratively calculates cluster centres and degrees of memberships of each data point until the optimisation function is satisfied or the number of iterations reaches a preset value.

For construction of rule-base and membership functions to automate the rule-based fuzzy system, clustering based methods have been commonly used, in particular, for type-1 fuzzy systems [40], [41]. The fuzzy sets involved in the rules are fully characterised by their membership functions. As explained in Section 3.1., the Gaussian membership function was utilised to develop the fuzzy rule base. The centroids of the clusters and their corresponding standard deviations obtained from FCM are used to determine the values of the parameters of the Gaussian membership functions. In this study, the degree of fuzzification is chosen to be two for FCM and number of clusters have been used to determine the number of rules.

4.4. Identifying Consequent Parameters

The least square estimation is a common method used to compute values of the consequent parameters of TSK-FS [18]. Given the support vector regression concept with a linear kernel, this can be potentially utilized to compute values of the consequent parameters of TSK-FS. The variables ($\overline{f_i}$, $\overline{f_i}x_{i1}$, $\overline{f_i}x_{i2}$, ..., $\overline{f_i}x_{in}$) defined using the normalized firing strength in (4) form inputs to SVR to derive w parameters that correspond to the consequent parameters in TSK-FS. Finally, SVR-based TSK-FS can be formulated by combining (11) and (12).

$$y' = \sum_{i=1}^{r} (\overline{f_i} y'_i + \frac{b}{r}) \tag{11}$$

where

$$y'_r = f(x_1, x_2, ..., x_n)' = w_{0r} + \sum_{i=1}^{n} (w_{ir} x_i) \tag{12}$$

‘ y' now represents the formulation of the SVR-based TSK-FS. For the sake of simplicity, in order to implement support vector regression part, LIBSVM library was used [42].

4.5. Searching the Optimal Parameters

There are three important parameters that are likely to affect the performance of the models. They are C and ϵ used to optimise the SVR linear kernel part, and the number of rules (i.e., clusters) for the TSK-FS. Due to the fact no generally accepted methods exist to determine these parameters optimally, the grid-search method has been decided to be employed as a parameter selection method in order to find the optimal parameter set. The grid-search method is simple and reliable and allows to implement parallel computations. The parameter range is searched with a step size of 0.05 for finding the optimal SVR kernel linear parameters. For the features, the search range was decided to be between 1 and 250. It is hoped that these ranges broadly cover all the possibilities that may contain optimal measure. Therefore, these parameters as well as different combinations of the features are assessed and their results were presented. Fig. 2 depicts how the grid-search conducted on SVR kernel parameters (C and ϵ) for their given ranges.

4.6. Performance Measurements of the Prediction Models

There are different measurements used to assess capability of the predictive models. However, in order to maintain consistency over the published results and perform consistent comparison, the following measures; coefficient of determination (q^2) and spearman rank correlation coefficient (ρ) are used that can be expressed as:

$$q^2 = 1 - \frac{\sum^{n}(y_{exp} - y_{prd})^2}{\sum^{n}(y_{exp} - \overline{y}_{exp})^2} \tag{13}$$

$$\rho = 1 - \frac{6\sum(y_{exp} - y_{prd})^2}{n(n^2 - 1)} \tag{14}$$

where y_{exp} and y_{prd} are the expected and predicted values of the peptide binding affinity, respectively, n is the number of peptides and $\overline{y}_{exp}$ is the mean of all expected values in the data set.

The measure q^2 is a statistical model based upon the proportion of variability in a data set [43]. When q^2 is close to 1 it suggests a model that has been successfully constructed. Negative q^2 values indicate that model poorly approximates the expected values. The spearman rank correlation coefficient (ρ) [44] is used to measure the statistical dependence between two variables. The value of ρ ranges between +1 and -1 showing perfect correlation at each end. The measures are calculated for each task (both training and testing). The metric q^2 is used to assess performance of the predictable models for the first three tasks whereas the fourth task was assessed by ρ in the competition.

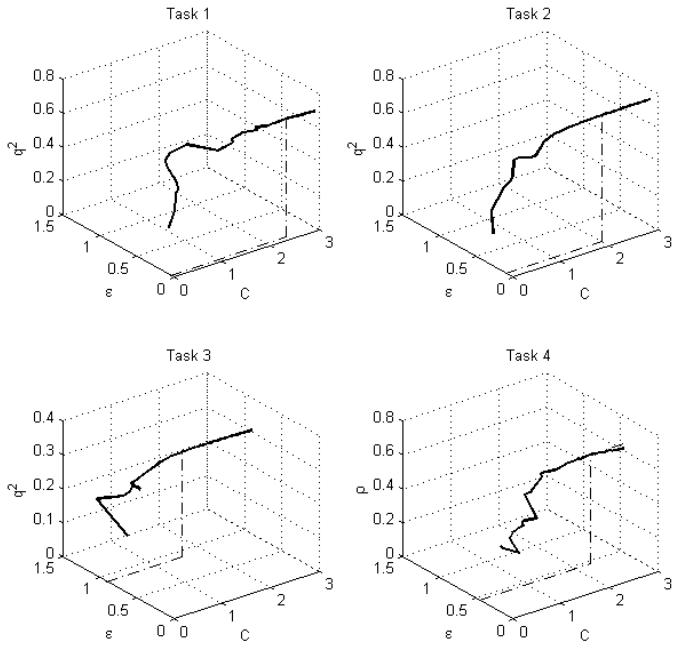

Figure 2: Illustration of grid-search for TSK-SVR to find the optimum values of parameters (C and ϵ) and prediction performance q^2 or ρ based on the selected features for the peptide binding affinity Tasks 1-4. For simplicity, the grid-search iterates C parameter with a step size 0.05 in the range 0.05-3.00 while remaining epsilon fixed. a) Task 1 with seven rules and 161 features yielded a q^2 value of 0.696 with SVR parameters C = 2.4 and $\epsilon = 0.05$. b) Task 2 with three rules and 247 features yielded a q^2 value of 0.743 with SVR parameters C = 1.9 and $\epsilon = 0.10$. c) Task 3 with three rules and 172 features yielded a q^2 value of 0.310 with SVR parameters C = 1.45 and $\epsilon = 0.90$. d) Task 4 with two rules and 141 features yielded a ρ value of 0.643 with SVR parameters C = 2.3 and $\epsilon = 0.45$.

5. Experimental Results and Discussion

The results of the experiments carried out will be discussed in three sub-sections. In the first part, the robustness of the proposed hybrid TSK-SVR method will be demonstrated over the four data sets. In the second part, SVR and TSK-SVR are compared in order to demonstrate the performance with and without the fuzzy concept. The latter part will present the outcome of the feature selection methods showing amino acid locations and amino acid scales.

5.1. TSK-SVR Results

In this section, the results of the proposed model (support vector based fuzzy system) are presented. To test the performance of the proposed model, four peptide data sets obtained from the (CoEPrA) competition are used. The proposed approach takes into account of predictive problem with very large number of attributes rather than simulated or practical data sets which only have very small number of features and consist of noise-free samples.

The data sets that have been used almost contain over 5000 descriptors for each peptide. One difficulty for the analysis of post-genome data is the curse of dimensionality. The curse of dimensionality is a term usually related to significant challenges that may occur when working with high-dimensional data sets [45]. Small sample size is another important characteristics of the peptide data sets. As a consequence, the high-dimensional nature of the data negatively effects the performance of the prediction methods and the proposed approach also has no exceptions. Since thousands of features are available for peptides, a feature selection process is integrated to the proposed model as an initial step to obtain low dimensional feature space. MCFS was able to deal with large number of attributes of the peptide data sets efficiently, and the reduced feature subset was used as input variables of the rule-based fuzzy system.

FCM is used in this study to construct the fuzzy rule-base. The centroids of the clusters and their corresponding standard deviations obtained from FCM are used to design Gaussian membership functions of the fuzzy models. The rule-base for the fuzzy systems (in this study, TSK fuzzy system) can be driven by using clustering methods where each cluster generally represents a fuzzy rule. Therefore, the number of clusters is equivalent to the number of rules in the fuzzy system. Determining the optimum number of clusters (consequently, number of rules) in the clustering methods can be generally achieved by considering the outcome of experimental studies where different number of clusters is explored and the cluster structure that yields the best outcome (e.g., minimum error) could be regarded as the best set of clusters. Following this concept, we studied the number of clusters from 2 to 7 and their results were presented in Table 4 and Figs. 3-8. It should be noted that the cluster centers and the membership matrix is randomly initialized in the fuzzy clustering stage. Thereby, random initialization in FCM may have some effect on the performance.

Along with the number of rules (clusters), further experiments were carried out to find optimum values of the parameters of TSK-SVR model for each rule structure as demonstrated in Figs. 3-8. In addition, one common problem in the support vector based approach is that it is not easy to determine which kernel function can be used [46]. In this study, SVR is trained with a linear kernel to learn the parameters of the consequent part of the fuzzy model. Therefore, the parameters C and ϵ are required to be optimised. The optimisation of SVR parameters was achieved by the grid-search where several thousands of the values of the parameters were tested over each rule base in order to find the best set of the values of the parameters for each rule base that yields the highest q^2 (first three tasks) and ρ (last task). The grid-search is repeated for each of the feature selection step and then, at the end of the process, the best model is selected.

For Task 1, graphs show fluctuations and reach local maximums particularly in the first 100 features. They rose gradually then and reach the global maximum at 161 features. After reaching the global maximum they become steady. For Task 2, graphs increase gradually as the number of features selected grew. They reach local maximums in the first 75 features and reach the global maximum at 247 features (at 248 features in Fig. 3). For Task 3, slight fluctuations are observed throughout the graphs, reaching local maximums in the first 150 features and then reaching global maximum at 165 features (at 172 features in Fig. 4). For Task 4, substantial fluctuations are observed throughout the graphs, reaching local

Table 3: Prediction results of TSK-SVR for each rule-base. For each rule, two results are presented. The former shows the best results obtained with the lowest possible feature set as compared to literature. The latter shows the best result and its number of features.

Number of Rules	Task 1		Task 2		Task 3		Task 4	
	q^2	f	q^2	f	q^2	f	ρ	f
	0.692	161	0.671	172	0.236	31	0.598	101
2	0.692	161	0.739	246	0.299	165	**0.643**	141
	0.693	161	0.669	176	0.236	31	0.594	101
3	0.693	161	**0.743**	247	**0.310**	172	0.638	141
	0.693	161	0.671	172	0.236	31	0.587	101
4	0.693	161	0.743	247	0.299	165	0.643	141
	0.694	161	0.670	172	0.236	31	0.573	67
5	0.694	161	0.743	247	0.299	165	0.639	141
	0.695	161	0.668	172	0.236	31	0.582	67
6	0.695	161	0.740	247	0.299	165	0.628	141
	0.696	161	0.664	172	0.236	31	0.577	67
7	**0.696**	161	0.736	247	0.299	165	0.626	121

f: Number of Features.

maximums after 50 features until reaching global maximum at 141 features (at 121 features in Fig. 8). As an example, illustration of the fuzzy rules for Task 4 (with two rules only) is provided in Fig. 3. As this model has 141 features and is not possible to fit in the paper, only three features were presented.

For each rule-base the proposed method is able to build a robust and interpretable fuzzy system for a high-dimensional data set with a relatively small number of data samples. It is observed that an optimum predictive model for each task was obtained by using different sets of rules as presented in Table 3. While the number of rules needed was smaller for Tasks 2, 3 and 4, Task 1 seems to require more rules to obtain the best possible outcome. Only two rules for Task 4 and three rules for Tasks 2 and 3 were enough to yield the best performance whereas the fuzzy-rule base with six or seven rules seems a requirement for the optimum modelling of Task 1.

The outcomes of the experiments clearly highlighted the strengths of TSK-SVR. The fuzziness has positively contributed towards the modelling of the tasks. To illustrate the performance of the proposed hybrid method, it is compared to the recently published results. In the (CoEPrA) competition Task 1 and 2 contained more than ten participants. Task 3 and 4 contained more than five participants. As shown in Table 4, the results outperform the competition results in which each participant competed with their best model (e.g., SVR, RF, PLS) [29]. In addition, for each task the results obtained are comparatively better than the recent studies presented in [11], [29], [12] and [13]. As compared to the best model presented in the literature, the predictive performance for Tasks 1, 2, 3 and 4 have been improved by 0.7%, 11.2%, 33.6% and 9.7%, respectively. The overall improvement gain for all tasks is found to be 13.8%.

Table 4: Prediction results of TSK-SVR compared to the results found in the literature. The performance of the method along with its selected number of features (f) are presented. Those methods that do not report the number of features for their models remained not available.

Methods		Task 1		Task 2		Task 3		Task 4	
		q^2	f	q^2	f	q^2	f	ρ	f
SVR	[29]	0.677	All	0.401	N/A	0.154	N/A	0.565	N/A
Partial Least Squares (PLS)	[29]	0.602	All	0.735	34	0.201	148	0.593	148
Random Forest (RF)	[29]	0.626	N/A	N/A		0.236	115	0.472	15
K-Nearest Neighbours	[29]	-0.322	N/A	0.612	N/A	N/A		N/A	
Gaussian Process	[29]	0.615	1864	-0.324	1289	0.065	2044	0.467	2044
Lasso	[11]	0.667	50	0.642	43	0.205	56	0.548	41
Ridge w/ Lasso	[11]	0.691	50	0.668	43	0.131	56	0.586	41
Partial Least Squares (PLS)	[12]	0.691	584	0.590	147	0.219	180	N/A	
SVR	[13]	0.682	200	0.639	100	0.232	100	N/A	
Random Forest (RF)	[13]	0.661	200	0.607	200	0.208	100	N/A	
TSK-SVR		**0.696**	161	**0.743**	247	**0.310**	172	**0.643**	141
I%		0.7%		11.2%		33.6%		9.7%	

I%: Percent Improvement of TSK-SVR with respect to the best methods in the literature.
f: Number of Features. N/A: Not Available.

5.2. Comparison of SVR and TSK-SVR

There have been a number of studies that present the prediction of peptide biding affinity by using SVR-based analysis. As TSK-SVR is a hybrid method that combines SVR with a fuzzy-rule base, namely TSK in this study, it will be important to compare the performances of SVR with and without the fuzzy concept. As detailed in Table 5, there is clear evidence over all the tasks that, based on the recent literature where SVR has been used for the prediction of the same data sets with the same training and test cases, the proposed TSK-SVR algorithm outperforms its solo version and yields an improvement of 2.1%, 16.3%, 33.6% and 13.8% for each of the tasks, respectively. This outcome demonstrates superiority of the proposed hybrid approach in mapping the input on the output over this challenging high-dimensional regression problem. The optimal parameters of TSK-SVR for the peptide binding affinity tasks are found to be $C = 2.40$, $\epsilon = 0.05$, and rule size of seven for Task 1; $C = 1.90$, $\epsilon = 0.10$, and rule size of three for Task 2; $C = 1.45$, $\epsilon = 0.90$, and rule size of three for Task 3; and $C = 2.30$, $\epsilon = 0.45$, and rule size of two for Task 4. The TSK-SVR models contained 161, 247, 172, 141 features for each peptide binding affinity task, respectively. It is worth noting that, our approach (TSK-SVR) not only benefited from SVR-based training but also handled the uncertainties in the peptide binding data set using the fuzzy modelling.

Table 5: The Parameters and Correlation Coefficient Results of SVR and TSK-SVR.

Methods		Task 1		Task 2		Task 3		Task 4	
		q^2	f	q^2	f	q^2	f	ρ	f
SVR	[29]	0.677	All	0.401	N/A	0.154	N/A	0.565	N/A
SVR	[13]	0.682	200	0.639	100	**0.232**	100	N/A	
TSK-SVR		**0.696**	161	**0.743**	247	**0.310**	172	**0.643**	141
I%		2.1%		16.3%		33.6%		13.8%	

I%: Percent Improvement of TSK-SVR with respect to SVR.
f: Number of Features. N/A: Not Available.

5.3. Analysis of Selected Descriptors

The SVR-based experiments were carried out for four different peptide affinity data sets. For each rule-base (rules that range between two and seven), feature selection (between 1 and 250 features) was conducted to reduce the number of features. The amino acid features that contributed most to the efficiency of the proposed models are given in Table 6 - Table 9.

For Task 1, eight amino acid features contributed to the output in more than four separate locations. The amino acid feature numbered with 481 (Hydrophobicity coefficient in reversed phase high performance liquid chromatography) contributed highest as it is represented in seven separate locations on each of the nona-peptide within the data set.

For Task 2, eleven amino acid features contributed to the output in more than five separate locations. The amino acid feature numbered with 364 (Zimm-Bragg parameter sigma x 1.0E4) contributed highest as it is represented in seven separate locations on each of the octa-peptide within the data set.

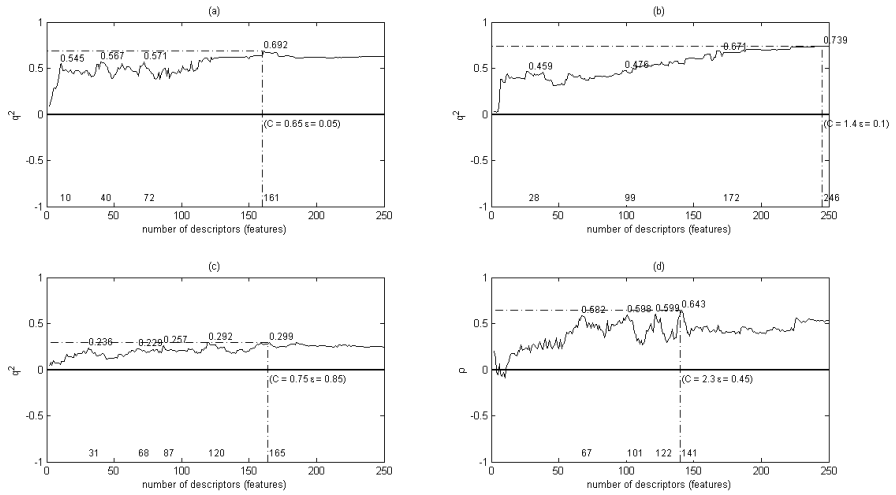

Figure 3: The performance of 2-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 0.65 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 246 with the SVR parameters (C = 1.4 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 165 with the SVR parameters (C = 0.75 and $\epsilon = 0.85$). d) Task 4: Graph reaches highest peak at 141 with the SVR parameters (C = 2.3 and $\epsilon = 0.45$).

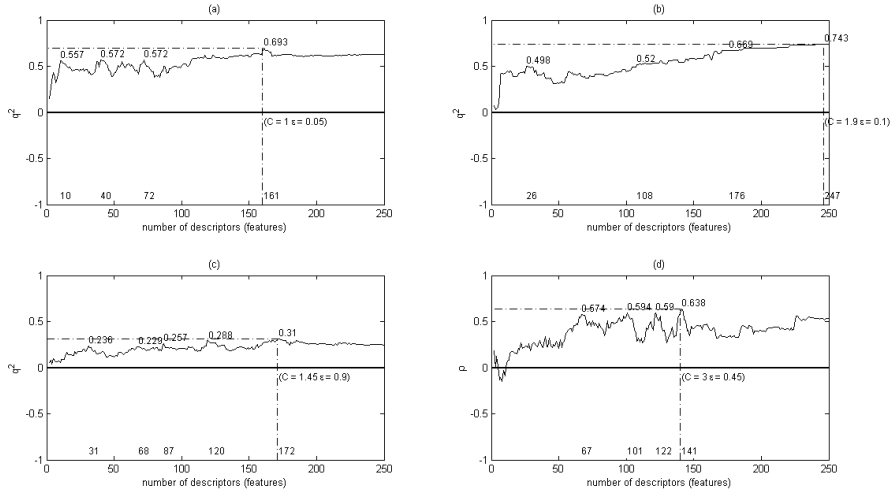

Figure 4: The performance of 3-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 1.0 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 247 with the SVR parameters (C = 1.9 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 172 with the SVR parameters (C = 1.45 and $\epsilon = 0.9$). d) Task 4: Graph reaches highest peak at 141 with the SVR parameters (C = 3.0 and $\epsilon = 0.45$).

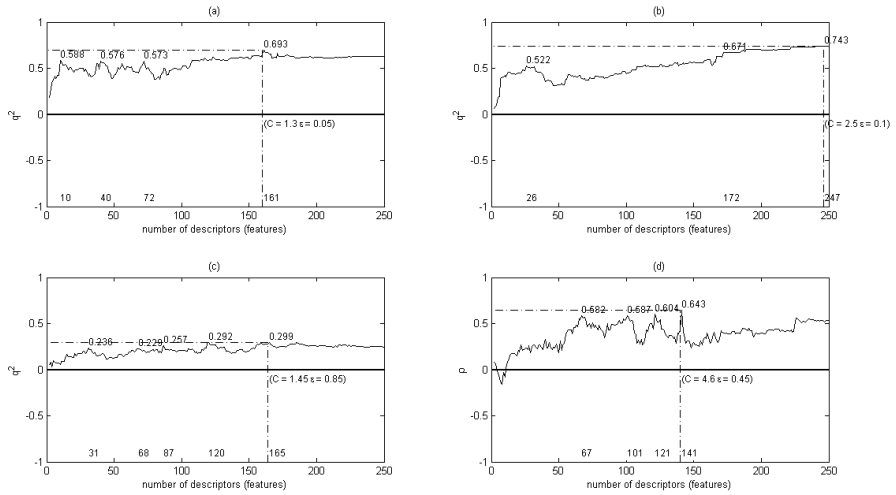

Figure 5: The performance of 4-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 1.3 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 247 with the SVR parameters (C = 2.5 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 165 with the SVR parameters (C = 1.45 and $\epsilon = 0.85$). d) Task 4: Graph reaches highest peak at 141 with the SVR parameters (C = 4.6 and $\epsilon = 0.45$).

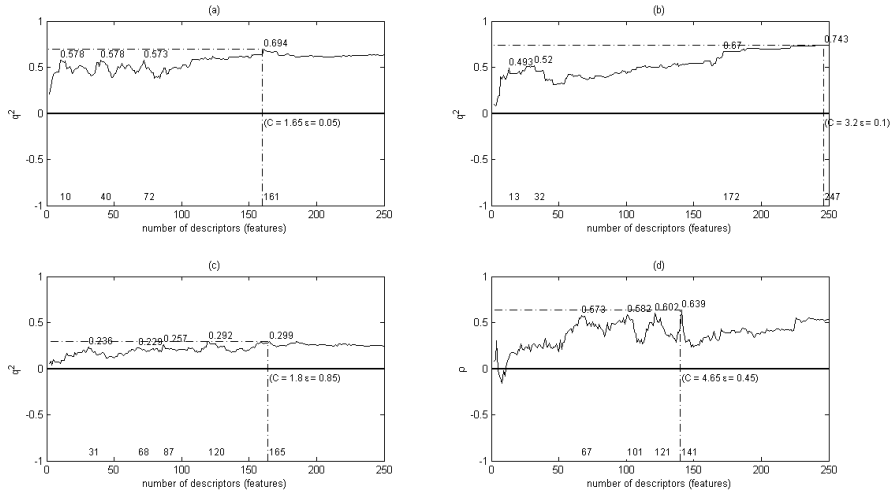

Figure 6: The performance of 5-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 1.65 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 247 with the SVR parameters (C = 3.2 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 165 with the SVR parameters (C = 1.8 and $\epsilon = 0.85$). d) Task 4: Graph reaches highest peak at 141 with the SVR parameters (C = 4.65 and $\epsilon = 0.45$).

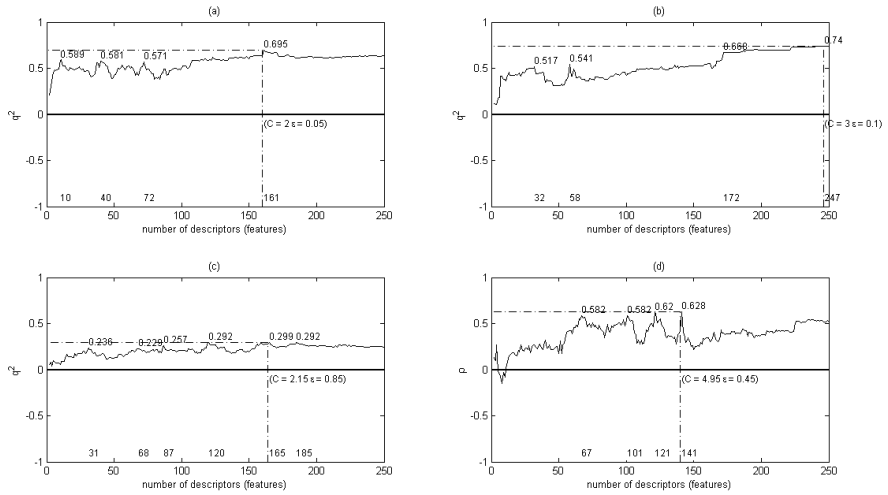

Figure 7: The performance of 6-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 2.0 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 247 with the SVR parameters (C = 3.0 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 165 with the SVR parameters (C = 2.15 and $\epsilon = 0.85$). d) Task 4: Graph reaches highest peak at 141 with the SVR parameters (C = 4.95 and $\epsilon = 0.45$).

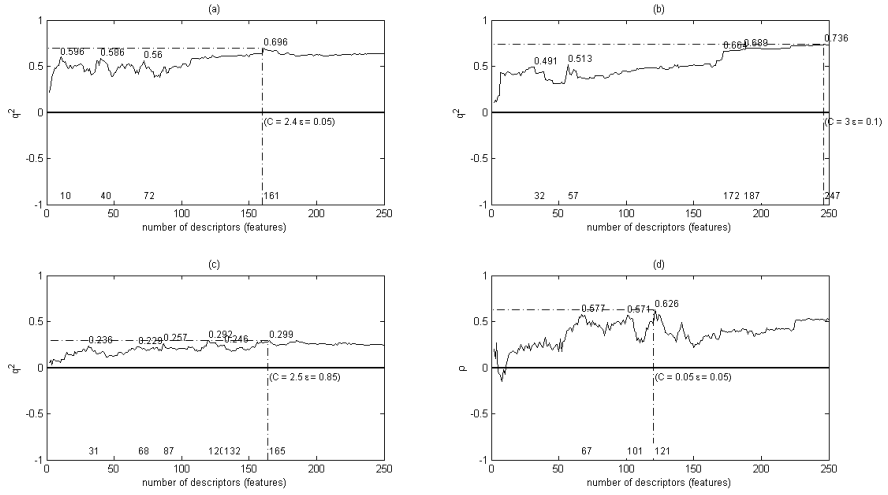

Figure 8: The performance of 7-rule fuzzy model based on the number of descriptors. a) Task 1: Graph reaches highest peak at 161 with the SVR parameters (C = 2.4 and $\epsilon = 0.05$). b) Task 2: Graph reaches highest peak at 247 with the SVR parameters (C = 3.0 and $\epsilon = 0.1$). c) Task 3: Graph reaches highest peak at 165 with the SVR parameters (C = 2.5 and $\epsilon = 0.85$). d) Task 4: Graph reaches highest peak at 121 with the SVR parameters (C = 0.05 and $\epsilon = 0.05$).

Table 6: Top most frequent amino acid features for the optimal model of Task 1.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	481	7	1	1	1	0	1	0	1	1	1
2	302	6	0	1	1	0	1	1	1	0	1
3	367	6	1	1	0	0	1	1	0	1	1
4	31	5	0	0	1	1	0	1	1	1	0
5	613	5	1	1	0	0	0	1	1	0	1
6	259	4	0	1	0	1	0	1	0	1	0
7	359	4	0	0	1	1	0	0	1	1	0
8	400	4	0	1	0	1	0	0	0	1	1

Table 7: Top most frequent amino acid features for the optimal model of Task 2.

No	Amino Acid Index	Number of Occurrences	Location							
			1	2	3	4	5	6	7	8
1	364	7	1	1	0	1	1	1	1	1
2	31	6	1	1	1	1	1	0	0	1
3	379	6	1	0	0	1	1	1	1	1
4	400	6	1	1	0	1	0	1	1	1
5	476	6	1	0	0	1	1	1	1	1
6	30	5	1	0	1	1	0	0	1	1
7	235	5	0	1	1	1	1	0	1	0
8	302	5	0	1	1	1	0	0	1	1
9	380	5	1	0	0	0	1	1	1	1
10	386	5	0	1	1	1	1	0	1	0
11	609	5	1	1	0	1	1	1	0	0

Table 8: Top most frequent amino acid features for the optimal model of Task 3.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	110	4	0	1	0	1	0	1	0	0	1
2	338	4	0	0	0	1	0	1	1	1	0
3	376	4	0	0	0	1	0	1	1	1	0
4	405	4	1	1	1	0	0	0	1	0	0
5	25	3	0	0	1	1	0	0	0	1	0
6	88	3	0	0	1	1	0	1	0	0	0
7	220	3	0	0	0	1	0	0	1	1	0
8	221	3	1	0	0	0	0	1	0	1	0
9	232	3	0	1	0	1	0	0	0	1	0
10	296	3	1	0	0	1	0	0	0	1	0
11	299	3	0	0	0	0	1	1	0	1	0
12	345	3	0	0	0	0	0	1	1	1	0
13	349	3	0	0	1	0	1	0	0	1	0
14	367	3	1	0	0	0	0	0	1	1	0
15	373	3	1	0	0	0	0	1	0	1	0
16	400	3	1	0	0	0	0	0	1	1	0
17	452	3	1	0	0	1	1	0	0	0	0
18	455	3	0	0	1	1	0	0	0	1	0
19	481	3	0	0	0	0	1	0	1	1	0

Table 9: Top most frequent amino acid features for the optimal model of Task 4.

No	Amino Acid Index	Number of Occurrences	Location								
			1	2	3	4	5	6	7	8	9
1	306	4	0	0	0	1	0	1	1	1	0
2	338	4	0	0	0	1	0	1	1	1	0
3	110	3	0	1	0	0	0	1	0	0	1
4	125	3	0	0	0	0	1	1	0	1	0
5	221	3	1	0	0	0	0	1	0	1	0
6	232	3	0	1	0	1	0	0	0	1	0
7	251	3	0	0	0	1	0	0	1	0	1
8	373	3	1	0	0	0	0	1	0	1	0
9	405	3	1	1	1	0	0	0	0	0	0
10	420	3	1	0	0	0	0	1	1	0	0

For Task 3, nineteen amino acid features contributed to the output in more than three separate locations. The amino acid features numbered with 110 (Composition), 338 (Relative preference value at C"), 376 (Relative population of conformational state A), 405 (Normalized positional residue frequency at helix termini N") contributed highest as they are represented in four separate locations on each of the nona-peptide within the data set.

For Task 4, ten amino acid features contributed to the output in more than three separate locations. The amino acid features numbered with 306 (Average relative fractional occurrence in A0(i-1)), 338 (Relative preference value at C"), 110 (Composition), 125 (Normalized relative frequency of double bend) contributed highest as they are represented in seven separate locations on each of the nona-peptide within the data set. The amino acid feature numbered with 400 (Polarity) appeared in Task 1, Task 2 and Task 3 as a common feature with location occurrences of 4, 6 and 3, respectively. Therefore, the polarity of an amino acid is considered as one of the highly discriminating feature in these data sets.

The results also appear to suggest that different sets of amino acid descriptors effect the result, and that exploration of the feature selection methods may further help accelerate the predictive power of the proposed hybrid method.

6. Conclusions

In this paper, a hybrid system (TSK-SVR) that has helped improve the predictive ability of TSK-FS significantly with the aid of support-based vector method was developed and demonstrated with the successful applications in the prediction of peptide binding affinity being regarded as one of the difficult modelling problems in bioinformatics. As far as an algorithmic approach is concerned, two important conclusions can be driven:

- SVR is enhanced by adding the fuzziness concept.
- TSK-FS is benefited from SVR-based training.

Predictive performances have been improved as much as 34% when compared to the best performance presented in the literature. The overall improvement gain for all tasks is found to be 13.8%. Apart from improving the prediction accuracy, this research study has also identified amino acid features "Polarity," "Hydrophobicity coefficient," "Zimm-Bragg parameter," and "Composition" being the highly discriminating features in the peptide binding affinity data sets. Therefore, the amino acid features may be potentially considered for better design of peptides with appropriate binding affinity.

The developed hybrid framework used for non-linear system modelling is based on TSK fuzzy model, consequent part of which is formed by a set of linear equations. As the support vectors in SVR were used to help form the consequent part of the model, it can be extended to type-2 fuzzy system with a closed-form type reduction and defuzzification method where Biglarbegian-Melek-Mendel (BMM) based type-2 fuzzy system could be explored as an example [47], [48]. Similarly, the concept could be further generalised to explore type-n fuzzy system for which the defuzzification phase could be performed using such approach. Further research is being carried out in this direction.

Acknowledgements

V. Uslan was a full PhD tuition fee bursary student funded by the De Montfort University Leicester UK. The authors thank to Dr Ovidiu Ivanciuc for organizing the CoEPrA contest that provided the peptide binding affinity data sets. The authors also thank to Dr Ozgur-Demir Kavuk for his assistance in providing the binding affinities of the test data sets.

References

[1] C. Yanover, T. Hertz, Predicting protein-peptide binding affinity by learning peptide-peptide distance functions 3500 (2005) 456–471.

[2] J. A. Bristol, J. Schlom, S. I. Abrams, Development of a murine mutant Ras CD8+ CTL peptide epitope variant that possesses enhanced MHC class I binding and immunogenic properties, The Journal of Immunology 160 (5) (1998) 2433–2441.

[3] W. W. P. Liao, J. W. Arthur, Predicting peptide binding to major histocompatibility complex molecules, Autoimmunity Reviews 10 (8) (2011) 469–473.

[4] P. Donnes, A. Elofsson, Prediction of MHC class I binding peptides, using SVMHC, BMC Bioinformatics 3 (1) (2002) 25.

[5] M. Bhasin, G. P. S. Raghava, Analysis and prediction of affinity of TAP binding peptides using cascade SVM, Protein Science 13 (3) (2004) 596–607.

[6] A. Sette, S. Buus, E. Appella, J. A. Smith, R. Chesnut, C. Miles, S. M. Colon, H. M. Grey, Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis, Proc Natl Acad Sci U S A 86 (9) (1989) 3296–3300.

[7] J. D. Stone, A. S. Chervin, D. M. Kranz, T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity., Immunology 126 (2) (2009) 165–176.

[8] K. Roomp, I. Antes, T. Lengauer, Predicting MHC class I epitopes in large datasets, BMC Bioinformatics 11 (1) (2010) 90.

[9] R. Bremel, 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 6 (1) (2010) 7.

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

[11] O. Demir-Kavuk, M. Kamada, T. Akutsu, E.-W. Knapp, Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features, BMC Bioinformatics 12 (1) (2011) 412.

[12] C. Bergeron, T. Hepburn, C. M. Sundling, M. P. Krein, W. P. Katt, N. Sukumar, C. M. Breneman, K. P. Bennett, Prediction of peptide bonding affinity: kernel methods for nonlinear modeling.

[13] A. Srivastava, S. Ghosh, N. Anantharaman, V. K. Jayaraman, Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests, Journal of Immunological Methods 387 (1-2) (2013) 284–292.

[14] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, Systems, Man and Cybernetics, IEEE Transactions on SMC-15 (1) (1985) 116–132.

[15] M. Sugeno, G. T. Kang, Structure identification of fuzzy model, Fuzzy Sets and Systems 28 (1) (1988) 15–33.

[16] O. Cordon, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, Ten years of genetic fuzzy systems: current framework and new trends, Fuzzy Sets and Systems 141 (1) (2004) 5–31.

[17] J.-S. R. Jang, C.-T. Sun, Neuro-fuzzy modeling and control, Proceedings of the IEEE 83 (3) (1995) 378–406.

[18] J.-S. R. Jang, ANFIS: adaptive-network-based fuzzy inference system, Systems, Man and Cybernetics, IEEE Transactions on 23 (3) (1993) 665–685.

[19] S. Chen, E. Chng, K. Alkadhimi, Regularized orthogonal least squares algorithm for constructing radial basis function networks, International Journal of Control 64 (5) (1996) 829–837.

[20] O. Cordon, Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases, Vol. 19, World Scientific, 2001.

[21] V. N. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.

[22] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, V. N. Vapnik, Support Vector Regression Machines, Vol. 9 of Advances in Neural Information Processing Systems, MIT Press, 1996.

[23] J. M. Leski, TSK-fuzzy modeling based on e-insensitive learning, Fuzzy Systems, IEEE Transactions on 13 (2) (2005) 181–193, iD: 1.
[24] C.-F. Juang, S.-H. Chiu, S.-J. Shiu, Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 37 (6) (2007) 1077–1087.
[25] C.-F. Juang, C.-D. Hsieh, J.-L. Hong, Fuzzy clustering-based neural fuzzy network with support vector regression, in: Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on, 2010, pp. 576–581.
[26] C.-F. Juang, 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 20 (2) (2012) 372–384.
[27] V. Uslan, H. Seker, Support vector-based Takagi-Sugeno fuzzy system for the prediction of binding affinity of peptides, in: Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE, 2013, pp. 4062–4065.
[28] A. Stryhn, L. O. Pedersen, A. Holm, S. Buus, Longer peptide can be accommodated in the MHC class I binding site by a protrusion mechanism, European Journal of Immunology 30 (11) (2000) 3089–3099.
[29] O. Ivanciuc, CoEPrA.
[30] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: Amino acid index database, Nucleic Acids Res 27 (1) (1999) 368–369.
[31] C. Chrysostomou, H. Seker, N. Aydin, CISAPS: Complex informational spectrum for the analysis of protein sequences, Advances in Bioinformatics.
[32] V. N. Vapnik, An overview of statistical learning theory, Neural Networks, IEEE Transactions on 10 (5) (1999) 988–999.
[33] A. J. Smola, B. Scholkopf, A tutorial on support vector regression, Statistics and Computing 14 (3) (2004) 199–222.
[34] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[35] D. Cai, C. Zhang, 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.
[36] M. Anastasiadou, A. Hadjipapas, M. Christodoulakis, E. S. Papathanasiou, S. S. Papacostas, G. D. Mitsis, Detection and removal of muscle artifacts from scalp EEG recordings in patients with epilepsy, in: Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on, 2014, pp. 291–296.
[37] V. Uslan, H. Seker, The quantitative prediction of HLA-B*2705 peptide binding affinities using support vector regression to gain insights into its role for the spondyloarthropathies, in: Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE, 2015, pp. 7651–7654.
[38] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, Exploiting statistical energy test for comparison of multiple groups in morphometric and chemometric data, Chemometrics and Intelligent Laboratory Systems 146 (2015) 10–23.
[39] J. C. Bezdek, FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences 10 (2-3) (1984) 191–203.
[40] T. A. Runkler, J. C. Bezdek, Alternating cluster estimation: a new tool for clustering and function approximation, Fuzzy Systems, IEEE Transactions on 7 (4) (1999) 377–393.
[41] H. Cao, L. Jia, G. Si, Y. Zhang, A clustering-analysis-based membership functions formation method for fuzzy controller of ball mill pulverizing system, Journal of Process Control 23 (1) (2013) 34–43.
[42] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 1–27.
[43] R. G. D. Steel, J. H. Torrie, D. A. Dickey, Principles and Procedures of Statistics, McGraw-Hill, 1997.
[44] J. L. Myers, A. D. Well, Research Design & Statistical Analysis, 2nd Edition, Lawrence Erlbaum Associates, 2003.
[45] R. E. Bellman, Adaptive control processes - A guided tour, Princeton University Press, Princeton, New

Jersey, U.S.A., 1961.

[46] C. J. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery 2 (2) (1998) 121–167.

[47] M. Biglarbegian, W. W. Melek, J. M. Mendel, Parametric design of stable type-2 TSK fuzzy systems, in: Fuzzy Information Processing Society, 2008. NAFIPS 2008. Annual Meeting of the North American, IEEE, 2008, pp. 1–6.

[48] M. Biglarbegian, W. W. Melek, J. M. Mendel, On the stability of interval type-2 TSK fuzzy logic control systems, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 40 (3) (2010) 798–818.