UNIVERSITY THOUGHT
Publication in Natural Sciences, Vol. 6, N^{o} 1, 2016, pp. 60-66.
doi:10.5937/univtho6-10768
Original Scientific Paper

THE RULE BASED CLASSIFICATION MODELS FOR MHC BINDING PREDICTION AND IDENTIFICATION OF THE MOST RELEVANT PHYSICOCHEMICAL PROPERTIES FOR THE INDIVIDUAL ALLELE

Davorka Jandrlić[1*]

[1]Faculty of Mechanical Engineering, University of Belgrade, Belgrade, Serbia.

ABSTRACT

Binding of proteolyzed fragments of proteins to MHC molecules is essential and the most selective step that determines T-cell epitopes. Therefore, the prediction of MHC-peptide binding is principal for anticipating potential T cell epitopes and is of immense relevance in vaccine design. Despite numerous methods for predicting MHC binding ligands, there still exist limitations that affect the reliability of a prevailing number of methods. Certain important methods based on physicochemical properties have very low reported accuracy. The aim of this paper is to present a new approach of extracting the most important physicochemical properties that influence the classification of MHC-binding ligands. In this study, we have developed rule based classification models which take into account the physicochemical properties of amino acids and their frequencies. The models use k-means clustering technique for extracting the relevant physicochemical properties. The results of the study indicate that the physicochemical properties of amino acids contribute significantly to the peptide-binding and that the different alleles are characterized by a different set of the physicochemical properties.

Key words:K – mean clustering, The rule based classification, MHC - peptide binding.

1. INTRODUCTION

Binding of peptides, derived by the intracellular processing of protein antigen(s) (Ag(s)) to MHC proteins, is the most selective step in defining T-cell epitopes. Computational methods, based on reverse immunology, are essential steps in the identification of T-cell epitope candidates, and complement epitope screening by predicting the best binding peptides.

Computational epitope-prediction programs are trained on the known peptide-binding affinities to a particular MHC molecule (or a defined set of MHC molecules) and fall into two categories (Brusić et al., 2004), (Yang et al.,2009): sequence-based and structure-based. The focus of this study is on sequence-based methods. A detailed list of the majority of currently available predictors with their prediction precision is described in (Luo et al., 2014).

Most of the predictors are based on sparse or BLOSUM50 encoding of peptide sequences, and different methods of machine learning are used: ANN, HMM, Decision Tree and SVM. Of all the predictors presented above, only POPI (Tung et al., 2007) uses physicochemical (PC) properties as input features. However, this predictor gives a evaluation qualitative of prediction with very low reported accuracy (~ 60%) (Luo et al., 2015). This was the reason for the investigation of the influence of different PC properties on peptide-binding.

The models obtained here predict MHC-binding ligands with very high accuracy. However, the main purpose of the developed models was not the prediction of MHC binding ligands, but rather identification of those PC characteristics that have the greatest impact on the classification of peptides into binders and non-binders. Fromour previous research (Mitić et al., 2014), (Pavlović et al., 2014), (Jandrlić et al, 2016) there was evidence that there were some characteristics that influence the binders appearance, for example, hydrophobicity, hydrophilicity, appertaining to ordered protein structure, etc., and there is certainly a great need for reliable methods based on high-quality experimental data for classification and prediction of MHC binding ligands and epitopes as a complement to existing ones.

2. EXPERIMENTAL

2.1. Materials and methods

In order to avoid the problems of inconsistent data to obtain reliable models, such as differing measures of binding affinity, etc., we chose to use only the data from the Immune Epitope Database (IEDB) (http://www.iedb.org/), June 2015 version, which is regularly updated, as the most reliable source of MHC-binding ligands. All experimentally proven MHC-binding ligands for all available alleles were downloaded. The research has been limited to peptides of 9 amino acids (AAs) in length because nonamers are the most common MHC-I epitopes, and to peptides of 15 amino acids (AAs) in length for MHC-II class, because there are only enough experimental data to construct good models for that length of the peptide. The data for those ligands for which there were no qualitative and quantitative measures (binding affinity and verification of whether a ligand is positive or negative), ligands labeled as both positive and negative, and as being found in the same protein at the same position and peptides containing AA not in α = {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}, were discarded. For all the alleles listed in Table 1 we created separate models and provided the results. For each of the allele the numbers of positive and negative epitopes involved in developing and testing the model are shown.

Table 1. Overview of alleles for which prediction models were made, as well as the number of positive and negative ligands (peptides) per studied allele.

Allele	No. of ligands	Trainset (Positive/Negative)	Test set (Positive/Negative)
HLA-A*02:01	5915	2010/2130	416/789
HLA-A*03:01	4014	970/1839	416/789
HLA-A*11:01	3306	972/1341	417/576
HLA-A*02:03	2304	793/819	340/352
HLA-B*15:01	3160	1010/1201	433/516
HLA-B*07:02	2606	97/1127	299/483
HLA-A*01:01	2577	347/1456	149/625
HLA-A*02:06	1695	733/452	315/195
HLA-A*24:02	1568	609/488	261/210
HLA-A*02:02	1339	618/319	265/137
HLA-B*08:01	1570	388/710	167/305
HLA-A*26:01	2534	244/1528	106/656
HLA-A*A31:01	2970	644/1434	277/615
HLA-DRB1*04:01	1321	657/267	282/115
HLA-DRB1*01:01	4987	2977/513	1276/221

2.1.1. Model building

The most important step in designing a reliable classification model is the choice of methodology for representing the data and feature selection. The most convenient way to represent the input data is in the form of a vector. That is why each peptide has to be represented by a vector whose components are obtained by the application of some weighting function. The representation of peptides in the form of vectors of its PC properties is also not uncommon, but as mentioned above, these methods are not shown very well in term of accuracy. The possible problem could be that each allele is generalized and all are characterized by the same group of PC properties (by applying principal component analyses or factor analyses). In this study, the peptide is represented using its combination unigrams and bigrams frequencies and specific PC properties, as described below.

2.1.2. Calculating the frequency of AAs

Calculating the frequency of AAs at appropriate positions in a peptide is aimed at extracting the features of occurrence of AAs in peptide binders and non-binders, which would enable easier classification. Instead of the standard calculation of AA frequency by position in a peptide, we used a modified calculation of frequency which was successfully implemented in document classification. Δ-TFIDF technique was first introduced in (Martineau et al., 2009) for the document classification problems, where this technique was applied to terms within a text, providing better results in classification of documents, than those obtained by a simple calculation of the frequency of terms or binary encoding. The task of classifying documents can easily be turned into the task of classifying peptides into binding and non-binding ligands. In a similar way, we calculated the Δ-TFIDF for individual AAs and bigrams in a peptide.

In our case, the term t, element of this method, is an AA (unigram) or bigram from the set of all AAs that occur in a peptide. For instance, in a peptide p = LVIKALLEV, t could be from the set {L, V, I, K, A, L, L, E, V, LV, VI, IK, KA, AL, LL, LE, EV}. Table 2 shows the basic definition and equations used for calculation of AA frequencies, applying Δ-TFIDF technique.

Table 2. Frequency measures

Equation	Definition
$df(t,S)$	Represents the frequency of the term *t* in a set of peptides *S*.
$idf(t,S) = log_2 \frac{\|S\|}{df(t,S)}$	Represents the inverse frequency of the term *t* in the set of peptides *S*, where $\|S\|$ is the cardinality of set S.
$tf(t, Peptide)$	Represents the number of occurrences of the term *t* in the peptide *Peptide* itself.

Taking into account equations from the table 2, the Δ -TFIDF measure, is introduced as:

$$\Delta tfidf(t_i, Peptide, S^+, S^-) = tf(t_i, Peptide) * \log_2 \frac{|S^+|}{df(t_i, S^+)} * \frac{df(t_i, S^-)}{|S^-|} \quad (1)$$

where t_i is AA or bigram in peptide Peptide from set S, at position i. S^+ and S^- are the subsets of S, of positive ligands (binders) and negative ligands (non binders), respectively. $|S^+|$ and $|S^-|$ are the cardinalities of the positive and negative sets. This way of transforming the frequency of AAs in a peptide gives greater importance to AAs that are not equally represented in the positive and negative sets, and less importance to those AAs that are more or less equally represented (the same holds for bigrams). Using the Δ -TFIDF measure, each peptide is represented as a vector of weights of its AAs and bigrams. If peptides are 9 AAs in length, an assigned vector is 9 + 8 = 17 in length. An obvious problem arises with peptides that have AAs or bigrams that do not occur in both classes, i.e. where $df(t, S^{\pm}) = 0$. Another problem is the non-linearity of AA frequency. These issues are resolved with the introduction of smoothing factors (Joachims, 1997). Here we engaged the Δ-BM25 smoothing measure (Joachims, 2005), a corrected frequency calculation, so as to avoid division with 0 in those cases where an AA occurs in only one of the sets. The Δ-BM25-IDF measure was chosen because it was found to be the best solution in the classification of texts (Joachims, 2005). With the inclusion of the Δ-BM25 smoothing factor, frequency is calculated through the equation:

$$\Delta BM25\ idf(t_i) = log \frac{(|S^+| - \Delta idf(t_i, S^+) + 0.5) * \Delta idf(t_i, S^-) + 0.5}{(|S^-| - \Delta idf(t_i, S^-) + 0.5) * \Delta idf(t_i, S^+) + 0.5} \quad (2)$$

This technique is simple and easily applied and understood.

2.1.3. Encoding peptide using physicochemical properties

The PC properties of the AAs are the information that could point to the similarity between AAs or bigrams within the peptide. The 119 PC properties (23 kinds of electronic properties, 37 kinds of steric properties, 54 kinds of hydrophobic properties and 5 kinds of hydrogen bond) were taken from the paper (Tian et al., 2009). Instead of using all PC properties for peptide representation, the common practice of reducing the number of properties which uses principal component analysis (PCA) or factor analyses (FA) and then selects some of the principal components or factors. This way, all the alleles are generalized, irrespective of whether certain properties are characteristic for a single allele. The aim of this research was to investigate specific allele characteristics and to evaluate influence of each individual PC property on classification peptides into MHC binders or non-binders. The peptide is firstly encoded with single PC property, in this way peptide is represented with vector of length 17 (9 + 8) with the numerical value obtained by applying PC property on appropriate consecutive AAs and bigrams from that peptide. This procedure is carried out for every single PC.

2.1.4.Classification rule based model building and selection of the most important physicochemical properties

To evaluate the importance of single PC property, for each PC property *fk* (k = 1,..., 119) and every single allele, a new rule based classification model was constructed. The set of all peptides associated with single allele is divided into a training and test subsets (70:30%). The peptides in both sets were

encoded with Δ-BM25 technique applied on unigrams and bigrams (see 2.1.2) and the component values were multiplied by the appropriate PC property value (see 2.1.3). For example, for the first PC property the following vector was obtained:

$$\begin{aligned}[w_{f_1}] &= w \cdot f_1 \\ &= [w_{11} \cdot f_1(AA_1), \dots, w_{19} \cdot f_1(AA_9), \\ &\quad w_{21} \cdot \frac{f_1(AA_1) + f_1(AA_2)}{2}, \dots, w_{28} \\ &\quad \cdot \frac{f_1(AA_8) + f_1(AA_9)}{2}]\end{aligned} \tag{3}$$

Then the k-means clustering technique (Hartigan, 1975) was applied to the positive (binders) and negative (non-binders) subsets of the training set, individually, with k = 3, three centroids per set are found that define the clusters. As a measure of the distance of the vector from the centroids, Euclidean distance was used. The number of clusters was determined empirically. Increasing the number of clusters increases the precision of the methods, but decreases the recall. Silhouette method was used for validation of consistency of the data within clusters (Rousseeuw, 1986).

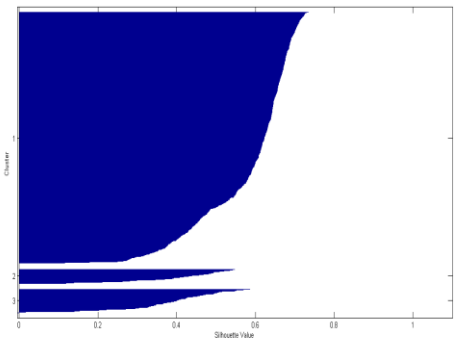

Fig. 1 Visualization of the clusters from thepositive training set for the HLA-A*1101 allele, where the peptides (binders) are encoded with the best PC property.

Fig. 1. represents clusters visualization of silhouette measure forallele HLA-A*1101. The silhouette value for each point *i* is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value ranges from -1 to +1. A high silhouette value indicates that *i* is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. The main purpose of clustering is to find the centroids, cluster representatives, that define all peptides from positive set and negative set. The centroids could serve as leading points for separating binders from non-binders.

The rule based binary classification models were made and identification of the best PC properties was done through the following steps:

1. Each peptideis encoded in the abovementioned manner; it is understood that if a peptide is closer to one of the positive class centroids, then it is a binder, but if it is closer to one of the negative class centroids, then it can be considered a non-binder.
2. For individual PC property the model is constructed (step 1.), the model was applied on the test set and the Kappa statistic, Accuracy, Precision and Recall are calculated (as described in the chapter 2.2);
3. The procedure (step 1. and 2.) is carried out for each of the PC properties.
4. Ten models with the highest accuracy are selected. The PC property involved in peptide encoding for the input into these models have been chosen as the best.
5. Finally, a consensus model is constructed from these ten best models. The consensus model takes into account results from *k* number of models (*k* is fewer or equal to 10) and if *k* number of these models are in agreement that peptide is binder then the observed peptide is considered a binder. Otherwise, the peptide is considered as non-binder. The choice of threshold for value *k* is treated as a maximization problem in terms of accuracy of the final model, on the training set. The minimal number of individual models is chosen to achieve the best accuracy.
6. The consensus model is tested on a set of peptides which were not presented during the training stage (Test Set, see Table 1).

The 119 models obtained in step 2. (for single allele) served for the selection of the 10 “best” PC properties (for unigrams and bigrams); and their construction was to compare the contributions of individual PC properties in the separation of binders from non-binders for a particular allele. The results of the consensus model constructed from the 10 best models are shown in Table 3, in the chapter Results. The list of 10 best PC properties (identified from these models) for each individual allele is presented in Supplement.

2.2. Performance evaluation

To evaluate the performance of our methods comprehensively, we report standard performance

measures, including accuracy, precision, recall and Kappa statistic as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FP + TN + FN}$$

$$Precision = \frac{\mathrm{TP}}{TP+FP} \quad (4)$$

$$Recall = \frac{\mathrm{TP}}{TP + FN}$$

where *TP*, *TN*, *FP* and *FN* respectively denote the number of true positives (correctly predicted binders), true negatives (correctly predicted non-binders), false positives (falsely predicted binders) and false negatives (falsely predicted non-binders).

Calculation of Cohen's kappa is performed according to the following formula:

$$Kappa = \frac{\Pr\ (a)\ - \Pr\ (e)}{1 - \Pr\ (e)}$$

$$\Pr(a) = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$\Pr(e) = \frac{(TP + FP) * (TP + FN) + (TN + FP) * (TN + FN)}{(TP + FN + FP + TN)^2}$$

Pr(a) represents the actual observed agreement, and *Pr(e)*represents chance agreement (Roy et al., 2015). Cohen's kappa analysis returns values between −1 (no agreement) and 1 (complete agreement).

3. RESULTSAND DISCUSSION

The results and performance measures of developed models for each individual allele are presented in the table 3. The measures of the training set indicate how well binders and non-binders are separated by calculated centroids of positive and negative clusters respectively. The measures of accuracy for all models are over 94% indicating that the centroids represent the binders and non binders well, i.e. binders indeed cluster around the 3 positive cluster centroids selected in this way. Similarly, non-binders cluster around negative3 centroids. Chosen PC properties are very good for characterizing binders related to individual allele.

Furthermore, clustering peptide into two groups - binders and non binders - is not a feasible task, and cannot be done just by applying *k*-means clustering algorithm on the entire set of peptides. In this research, it was established that a selection of three appropriate centroids for positive ligands (binder), based on extracted PC properties, describes very well the entire bindersset for single alleles (the same conclusion holds for negative ligands, as well). The results on test sets indicate how well the developed rule-based classification model generalizes on the blind set. Accuracy for almost all models on test sets is close to or over 80%, which confirms that certain combinations of physicochemical properties of AAs in peptide separates binders from non-binders with high accuracy. All models are tested on another MHC class I and II dataset collected from the MHCBN repository (http://www.imtech.res.in/raghava/mhcbn/mhcbinder_download.html). There is no overlapping between these datasets and IEDB datasets. The results on this dataset for all models also show high accuracy (see Supplement where comparative results for all datasets are shown).

Table 3. Comparative performance evaluation in terms of precision/recall/accuracy for: developed rule based classification models on training and test sets, the best currently existing predictor NetMHCpan (for MHC I)/NetMHCIIpan (for MHC class II) and MHCPred predictor.

Allele	Training	Test	NetMHCPan2.8	MHCPred
HLA-A0201	0.93 / 0.95 / 0.94	0.87 / 0.82 / 0.85	0.97 / 0.83 / 0.91	0.73 / 0.76 / 0.74
HLA-A0301	0.88 / 0.98 / 0.95	0.76 / 0.71 / 0.82	0.90 / 0.86 / 0.92	0.43 / 0.86 / 0.56
HLA-A1101	0.93 / 0.97 / 0.95	0.86 / 0.72 / 0.83	0.97 / 0.86 / 0.93	0.42 / 0.99 / 0.43
HLA-A0203	0.95 / 0.98 / 0.97	0.81 / 0.75 / 0.79	0.96 / 0.88 / 0.92	0.54 / 0.88 / 0.57
HLA-B1501	0.95 / 0.98 / 0.97	0.90 / 0.79 / 0.86	0.99 / 0.80 / 0.90	N/A
HLA-B0702	0.95 / 0.99 / 0.98	0.84 / 0.65 / 0.82	0.92 / 0.92 / 0.94	N/A
HLA-A0101	0.93 / 0.99 / 0.97	0.78 / 0.51 / 0.88	0.72 / 0.97 / 0.92	0.55 / 0.71 / 0.83
HLA-A0206	0.98 / 0.99 / 0.98	0.79 / 0.84 / 0.76	0.94 / 0.89 / 0.90	0.63 / 0.83 / 0.59
HLA-A2402	0.98 / 0.98 / 0.98	0.72 / 0.71 / 0.70	0.88 / 0.85 / 0.85	N/A
HLA-A3101	0.89 / 0.97 / 0.95	0.72 / 0.61 / 0.81	0.86 / 0.82 / 0.90	0.41 / 0.41 / 0.63
HLA-A0202	0.99 / 0.99 / 0.99	0.81 / 0.91 / 0.79	0.96 / 0.85 / 0.88	0.77 / 0.78 / 0.70
HLA-B0801	0.98 / 0.99 / 0.98	0.84 / 0.74 / 0.86	0.97 / 0.87 / 0.95	N/A
HLA-A2601	0.87 / 0.99 / 0.98	0.73 / 0.51 / 0.91	0.78 / 0.90 / 0.95	N/A
Allele	**Training**	**Test**	**NetMHCIIPan20**	**MHCPred**
HLA-DRB10101	0.99 / 0.94 / 0.95	0.88 / 0.92 / 0.82	0.97 / 0.92 / 0.91	0.85 / 0.99 / 0.85
HLA-DRB10401	0.98 / 0.98 / 0.98	0.78 / 0.93 / 0.76	0.95 / 0.80 / 0.64	0.72 / 0.99 / 0.72

The consensus threshold value represents the number of models, based on individual PC property, that are enough to include in the consensus model to achieve the best accuracy. The values of threshold k smaller than ten indicate that for a good classification model it is not necessary to use all ten PC properties and that a smaller number of k properties have enough influence to separate binders from non-binders for that allele. An interesting finding is that for MHC II class alleles (HLA-DRB1*01:01 and HLA-DRB1*04:01) there are only three PC properties that have enough influence on decision when the peptide is a potential binder. The list of PC properties associated for all involved alleles is presented in the Supplement. On the basis of alleles that are associated with the same physicochemical properties we can make a conclusion about similarity of alleles themselves. It is expected that alleles classified into the same supertype (Sidney et al., 2001), share data about most important PC

properties. It also can be seen that some of PC information is shared between different supertypes. Fig. 2 presents six of the most important *PC* properties, in the sense of being common for the largest number of alleles, their relationship is depicted with arrows from allele to PC property. It could be seen that an allele HLA-A*68:02 is characterized with PC properties: PC21, PC90, PC58, PC1 (pK-C, Partition coefficient, Flexibility parameter for no rigid neighbors, alpha-NH chemical shifts) and share importan PC properties as allele HLA-A*11:01 and HLA-A*31:01 which is to be expected as they belong to the same supertype A3. However, for allele HLA-A*02:03 and allele HLA-B*08:01 the common PC properties are crucial for separating binders from non-binders even if these two alleles do not belong to the same supertype. This confirm the claim by Heckerman (Heckerman et al., 2007) that there are some characteristics that are common for binders independent off allele and supertypes specific characteristics. All this information could be used in consideration for making better MHC binding prediction models. We have analyzed this information and used it as input features in different machine learning models, which produce even better results (this part of research is submitted for publication elsewhere).

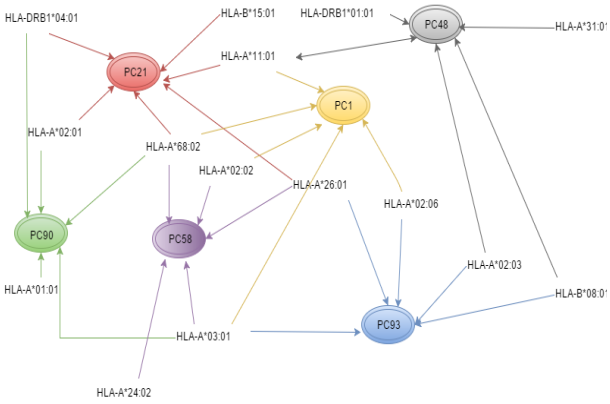

Fig. 2. The first six PC properties, according to number of alleles that are assigned to, that best separate binders from non binders: PC1 - alpha-NH chemical shifts; PC48 – Bulkiness; PC58 - Flexibility parameter for no rigid neighbors; PC90 - Partition coefficient; PC21- pK-C; PC93 Average gain ratio in surrounding hydrophobicity (all of them are taken from (Tian et al., 2009)and HLA – alleles related to them.

4. Conclusion

In this study, the new position-dependent method for classifying peptides into MHC binders and non-binders, is presented. Binary rule based classification models were made for 13 different alleles of MHC class I and 2 alleles for MHC class II. Vector components used as input features into models that represent peptides were obtained based on the physicochemical properties of the consecutive amino acids and bigrams contained in the peptides and application of the weighting technique for frequency calculation well established in the problems of the text categorization and opinion mining problems, but had never before been used for this type of problem. The models obtained have high accuracy. The goal of developed models was not to beat currently existing predictors, but to identify the most relevant PC properties for classification peptides into MHC binding ligands and non-binding ligands. However, in order to demonstrate the usefulness and reliability of the models developed, we compared them with two predictors: NetMHC(II)pan (Nielsen et al., 2007), (Nielsen et al, 2008) which is refered as the best curruntly existing predictors and MHCPred (Pingping et al, 2003). Summary results are shown in Table 3. and graphical presentation of comparative analyses is illustrated in Fig 3. The developed models had a significanty higher predictive performance than MHCPred predictor, but lower that NetMHCpan predictor.It should be noted that the test set for evaluation of these three predictors was blind test only for models developed here, but probably involved in training for NetMHC(II)pan models.

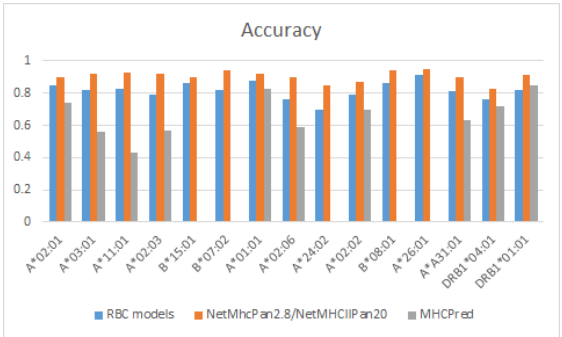

Fig. 3. The comparative analysis of the accuracy of the rule-based classification (RBC) models, NetMHC(II) Pan and MHC Pred predictor.

It is demonstrated that the certain PC properties have huge impact on the separation binders from non-binders and that their combination with a frequency measure can serve as input features in the new models for predicting MHC binding ligands. One would

expect that machine learning models with the these input parameters provide even better performance.

ACKNOWLEDGEMENT

The work presented has been financially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Projects No. 174002.

REFERENCES

Brusic, V., Bajic, V.B., & Petrovsky, N. 2004. Computational methods for prediction of T-cell epitopes: A framework for modelling, testing, and applications. Methods, 34(4), pp. 436-43, pmid:15542369.

Hartigan, J.A. 1975. Clustering Algorithms.New York, NY: USA John Wiley & Sons..

Heckerman, D., Kadie, C., & Listgarten, J. 2007. Leveraging information across HLA alleles/supertypes improves epitope prediction. J Comput Biol, 14(6), pp. 736-746. doi:10.1089/cmb.2007.R013.

Jandrlić, R.D., Lazić, M.G., Mitić, S.N., & Pavlović, D.M. 2016. Software tools for simultaneous data visualization and T cell epitopes and disorder prediction in proteins. Journal of Biomedical Informatics, . doi:10.1016/j.jbi.2016.01.016.

Joachims, T. 1997. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. . In: InInternational Conference on Machine Learning, ICML.

Joachims, T. 2005. Text categorization with Support Vector Machines: Learning with many relevant features. Sprnger.Lecture Notes in Computer Science, 1398, pp. 137-142.

Luo, H., Ye, H., Ng, H.W., Shi, L., Tong, W., Mendrick, D.L., & Hong, H. 2015. Machine Learning Methods for Predicting HLA-Peptide Binding Activity. Bioinform Biol Insights, 9(3), pp. 21-29, doi:10.4137/BBI.S29466.

Mitić, S.N., Pavlović, D.M., & Jandrlić, R.D. 2014. Epitope distribution in ordered and disordered protein regions: Part A. T-cell epitope frequency, affinity and hydropathy. J Immunol Methods, 406, pp. 83-103, doi:10.1016/j.jim.2014.02.012.

Martineau Justin, , & Finin Tim, 2009. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. . In: Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media. San Jose, CA: AAAI Press. May.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., & et al., 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence.PLoS One, 2(8). doi:10.1371/journal.pone.0000796.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., & et al., 2008. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol, 4(7), p. 1000107, doi:10.1371/journal.pcbi.1000107.

Pavlović, D.M., Jandrlić, R.D., & Mitić, S.N. 2014. Epitope distribution in ordered and disordered protein regions. Part B: Ordered regions and disordered binding sites are targets of T- and B-cell immunity. J Immunol Methods, 407, pp. 90-107, doi:10.1016/j.jim.2014.03.027.

Pingping Guan, I.A.D., Christianna Zygouri, , & Flower, D.R. 2003. MHCPred: A server for quantitative prediction of peptide-MHC binding., pp. 3621-3624.

Rousseeuw, P.J. 1986. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, . doi:10.1016/0377-0427(87)90125-7.

Roy, K., Kar, S., & Das, R.N. 2015. A Primer on QSAR/QSPR Modeling. Retrieved from http://www.springer.com/978-3-319-17280-4

Sidney, J., Southwood, S., Mann, D.L., Fernandez-Vina, M.A., Neuman, M.J., & Sette, A. 2001. Majority of peptides binding HLA-A*0201 with high affinity crossreact with other A2-supertype molecules. Hum Immunol, 62, pp. 1200-1216.

Tian, F., Yang, L., Lv, F., Yang, Q., & Zhou, P. 2009. In silico quantitative prediction of peptides binding affinity to human MHC molecule: An intuitive quantitative structure-activity relationship approach. Amino Acids, 36(3), pp. 535-554, doi:10.1007/s00726-008-0116-8.

Tung, C.W., & Ho, S.Y. 2007. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics, 23(8), pp. 942-949, doi:10.1093/bioinformatics/btm061.

Yang, X., & Yu, X. 2009. An introduction to epitope prediction methods and software. Rev Med Virol, 19(2), pp. 77-96, doi:10.1002/rmv.602

.

* E-mail: djandrlic@mas.bg.ac.rs