Letters in Drug Design & Discovery, 2017, *14*, 2-9

RESEARCH ARTICLE

Proteochemometrics for the Prediction of Binding to the MHC Proteins

Ventsislav Yordanov, Ivan Dimitrov and Irini Doytchinova*

Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., Sofia 1000, Bulgaria

ARTICLE HISTORY

Received: November 18, 2015
Revised: January 25, 2016
Accepted: February 22, 2016

DOI:
10.2174/1570180813666160630122341

Abstract: The recognition of a foreign protein (antigen) by host B and/or T cells and the induction of immune response have a key role for the human immune system. The peptide binding to MHC proteins is a prerequisite for B or T-cell recognition. Determining the peptide-binding preferences exhibited by the MHC extensive set of alleles requires enormous experimental work. A more rational approach is the development of bioinformatics prediction methodologies. Among others, proteochemometrics (PCM) is quite suitable for MHC binding prediction as it simultaneously models the bioactivity of multiple ligands against multiple protein targets. Handling multiple targets is the key to "ligand polypharmacology" and the development of multi-target drugs. The few applications of PCM for MHC binding prediction showed that the developed models have excellent predictive ability.

Keywords: Immunoinformatics, proteochemometrics, HLA-DRB1, HLA-DQ, peptide binding, MHC binders.

INTRODUCTION

Nowadays, it is difficult for one to imagine a therapeutic area where the immune system is not involved and where targeted modulation of the immune response could not provide substantial health benefits. Vaccines are one of the most important achievements in the field of infectious diseases, and while we have witnessed the use of whole pathogens (attenuated viruses or bacteria), or their purified single proteins, current research is focused on smaller peptides, rationally designed to contain a single or multiple epitopes, or even DNA-based vaccines, which can make cells directly produce an antigen and result in acquired immunological response [1, 2].

The recognition of a foreign protein (antigen) by host B and T cells and the induction of immune response have a key role for the human immune system. In the case of an autoimmune disease, the host B and T cells recognize the self proteins as foreign, attack them and destroy their own tissues. According to antigen origin, the antigen processing in the host cells undergoes two different pathways: intracellular and extracellular. The intracellular proteins, such as viral or self proteins, initially are degraded by the proteasome to oligopeptides, and then some of the peptides are transported to the endoplasmic reticulum (ER) by a transporter associated with antigen processing (TAP). In ER, the peptides bind to major histocompatibility complex (MHC) class I proteins and the peptide-protein complexes are presented on the cell surface where they are recognized by CD8+ T cells. The extracellular proteins, such as bacterial, parasite or fungal proteins, enter the cells by endocytosis. In the endosome, they are degraded to oligopeptides and bind to MHC class II proteins. These peptide-MHC class II protein complexes also are presented on the cell surface where they are recognized by CD4+ T cells. Not all peptides presented on the cell surface are recognized by the T cells. Those of them that are recognized are called T-cell epitopes 9 [3].

The MHC proteins are extremely polygenic (*i.e.* there are many MHC class I and class II genes) and polymorphic (*i.e.* there are many alleles of each gene). The human MHC are referred as Human Leukocyte Antigens (HLA). The IMGT/HLA database (release 3.25, July 2016) lists 11,000 class I and 3,920 class II alleles [4]. The MHC class I proteins are found in every nucleated cell of the body, whereas MHC class II proteins present in specialized cell types, called antigen-presenting cells (APC) as B cells, macrophages and dendritic cells. MHC class I proteins are encoded by three loci: HLA-A, HLA-B and HLA-C. MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ and HLA-DP. The MHC polymorphism is localized on the peptide binding site [5-10]. The MHC class I binding site is closed at both ends and only short peptides between 8 and 11 amino acids in length are able to bind, whereas the MHC class II binding side is open at both ends and the binding peptides may have length from 12 to 25 residues.

In order to predict immunogenicity of a potential vaccine candidate or an unwanted immune response towards a drug, it is crucial to determine its binding profile towards MHC class proteins. For vaccines, it is important to be able to trigger immune response in a broad number of individuals inside a population (which genetic profile should be considered while development of a new product) [11]. Contrary, keeping a "low immunogenicity profile" in other pharmaceutical products is essential for maintaining their efficacy and safety profile. This is an emerging issue given the extensive use of

*Address correspondence to this author at the Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., Sofia 1000, Bulgaria; Tel: +359 2 9236506; Fax +359 2 9879874; E-mail: idoytchinova@pharmfac.net

1875-628X/17 $58.00+.00

Table 1. Sequence-based and structure-based methods for MHC binding prediction.

Sequence-based Methods	Structure-based Methods
Motif search-based approach	Docking of peptides and screening of peptide libraries
Prediction by artificial neural network	Application of threading algorithms
Prediction by support vector machine	Binding energy calculations
Hidden Markov models	Molecular dynamics simulations
Prediction by quantitative matrices-driven methods	

monoclonal antibodies in the recent years and the limitations of those types of therapies due to unwanted immune response [12, 13]. Determining the peptide-binding preferences exhibited by this extensive set of alleles requires an enormous experimental work. A more rational approach is the development of bioinformatics prediction methodologies.

MHC BINDING PREDICTION

The immune system involves a vast number of interacting molecules, with variation both between and within individuals. Over the years, the use of high throughput techniques cumulated huge amounts of experimental data which now require the development of new computational approaches to analyze this massive information pool. A new field in immunology has emerged, called immunoinformatics, which is focused on algorithms for mapping and studying epitopes *in silico* and immunogenicity prediction. Utilizing this information is essential for the development of new vaccines, prediction of allergies, understanding the autoimmune diseases, analyzing the genome of a pathogen and identification of its antigenic proteins, and modification of the immune response in general [14].

The search for B- and/or T-cell epitopes as components of epitope-based vaccines led to the development of a great variety of algorithms and methods for MHC binding prediction. The current state of this branch of immunoinformatics has recently regularly reviewed [15-17]. Briefly, the methods for MHC binding prediction can be divided into sequence-based methods that use the amino acid sequences as starting information and structure-based methods that derive information from the protein 3D structures (Table **1**).

Historically, the first methods for T-cell epitope prediction were based on MHC-binding motifs search [18, 19]. Motif is the combination of preferred amino acids at the peptide anchor binding positions. Extended variations of the motif search are the quantitative matrices (QM) [20-24]. The QM takes into account the quantitative contribution of each amino acid at each position in the binding peptide. More complicated sequence-based methods for MHC binding prediction use Hidden Markov Models (HMM) [25], artificial neural networks (ANN) [26], support vector machines (SVM) [27, 28]. Most of the models derived by these methods are freely accessible in the web and widely used to reduce the subsequent experimental work.

The structure-based methods for MHC binding prediction rely on the crystallographic 3D structures of MHC proteins. Steric, electrostatic and hydrophobic complementarities are essential between peptide and MHC protein in order to be formed a stable complex. Knowing the structure of the binding pockets along the MHC binding site, the favorite anchor peptide residues could be identified. The structure-based methods for MHC binding prediction developed by now include docking studies [29], free energy calculations [30], threading algorithms [31], molecular dynamics simulations [32]. The structure-based methods are more accurate in predictions than the sequence-based but are more time-consuming and the models derived by them usually are inaccessible in the web.

PROTEOCHEMOMETRICS

Quantitative Structure – Activity Relationships (QSAR) use different statistical approaches to create models, which relate chemical structure to biological activity using molecular descriptors (Fig. **1**). This method is based on the compound similarity principle and is used to optimize lead compounds for target activity and other properties (*e.g.*, ADME and toxicity). Disadvantages are that its models consider only ligand properties and analyze interactions with only one target at a time. Hence, QSAR models are unable to generalize between multiple targets – they have minimal ability to extrapolate. QSAR requires sufficient data to be available on a specific target before construction of a meaningful model, which is not always the case, especially when modeling novel targets [33-35].

Proteochemometrics (PCM) was developed by Lapins *et al.* to describe interactions between proteins and their ligands – it is a method to model simultaneously the bioactivity of multiple ligands against multiple protein targets [36, 37]. It could be considered as an extension of QSAR which combines the information of ligands and targets into a single matrix, which allows extrapolation to bioactivity of new compounds to new targets [35] (Fig. **1**). The ability of PCM to connect neighboring QSAR data sets makes it quite similar to inductive learning and is the main reason to outperform conventional QSAR models. Handling multiple targets is the key to "ligand polypharmacology" and the development of multi-target drugs [34]. Important disadvantage in PCM is its dependence on the variation of ligands and targets; if a particular amino acid plays significant role in the interaction, but is conserved throughout the dataset, PCM will be unable to assess its importance [38]. In PCM typically three descrip-

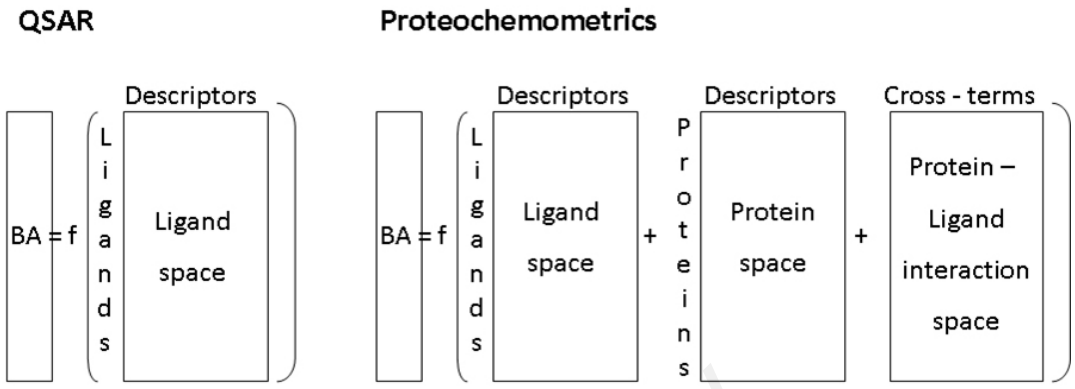

Fig. (1). QSAR and proteochemometrics.

tor blocks could be distinguished: the ligand block (L), protein block (P), and ligand-protein cross-term (LP) [37].

LIGAND SPACE

Descriptors are usually classified according to their dimensionality, ranging from one dimensional (1D) to three-dimensional (3D) properties. Simple and fast are the binary descriptors (1D), used by Lapins *et al.* [36]. Base structure for ligands was 4-piperidyl oxazole, 12 compounds were included and all combinations of substituents were coded using binary descriptors. This type of descriptors do not allow extrapolation, interpretability is low since there is need of translation of the results back to the functional group combination [34]. Physicochemical compound descriptors (such as molecular weight, van der Waals volume, electronegativity, polarizability, molar refractivity, polar surface area, log P, etc) have better interpretability, but due to the high number of possible descriptors, data preprocessing is needed in order to avoid covariance. Lapinsh *et al.* started with 61 compound descriptors and after checking for mutual correlation only 31 descriptors remained in the training set [39]. 2D topological descriptors are widely used, and incorporate both atomic and bond properties. While graph-based 2D methods are computationally slow, fingerprint-based methods (where pre-defined structural elements are encoded as bit strings) are much faster, easy to derive, handle and compare. 3-D information can also be translated into a bit string, and comparing bit strings is much easier than comparing structures, since alignment step is avoided [12]. From 3D descriptors GRINDs are example of a successful use in PCM by Lapins *et al.* [38, 40].

PROTEIN SPACE

Amino-acid sequence information is readily obtainable, however, due to variable lengths within a protein family and due to insertions/deletions (for example, sequence lengths of human GPCRs range from 290 to 6200 residues) an alignment of amino-acid sequences is required which can be tricky. Therefore, it is advisable to focus on specific protein substructures representing binding sites. Widely used are physicochemical sequential descriptors – each amino acid could be described by its volume, lipophilicity, polarizability, polarity, etc or by a principal component obtained after a PCA compression of a large set of physicochemical properties [34]. 3D descriptors use is limited by alignment problems and high number of variables [34], therefore the use of 3D "local" descriptors is more appropriate, as shown by Strömbergsson *et al.* [41].

PROTEIN-LIGAND INTERACTION SPACE

When using linear methods, such as PLS, there is a need to describe non-linear ligand-protein interaction effects. This is typically achieved by deriving cross-terms between ligand and protein descriptors [33]. Ligand-protein cross-terms are usually defined by multiplying each ligand descriptor with each receptor descriptor of particular ligand- receptor pairs [37]. A practical approach for deriving cross-terms is to use principal components of the original descriptors [33]. It is important to denote that for non-linear machine-learning methods cross-terms can show no improvement, even decrease in performance due to introduction of large number of parameters; also, since they are derived from multiplication of protein and ligand descriptors, calculation is quite time consuming. Therefore, their use should be carefully considered and preprocessing and selection of variables prior cross terms calculation is desirable [34].

Interesting approach in description of ligands and targets was presented by Pérot *et al.* [42] Readily interpretable descriptors of the ligands and the binding pockets were used, such as volume, polarity, charge, topology (roughness, planarity, narrowness) to describe the pocket-ligand pair space. Hierarchical classification of the pocket-ligand pairs was obtained, where five main clusters were outlined (*a*- and *b*-*clusters* – "small" pocket-ligand pairs with *a* and *b* differing in polarity and planarity; *d* and *e* – "large"; *c* – "average" in descriptor values). This comprehensive classification is a useful approach for a preliminary prediction models, and further development would provide a better understanding about correspondences between pockets and ligands.

DATA PREPROCESSING

In order to make descriptors compatible and to prevent one set of descriptors to mask another, scaling and mean centering should be performed. Then, different methods should be applied to remove covariance. For variable extraction, *principal component analysis* (PCA) is most commonly

applied. PCA and its regression extension p*artial least squares* (PLS) provide the ability of a reduction of a multi-dimensional variable space into a limited number of descriptor variables, which are called principal components [33, 34, 38].

For variable selection (also known as *variable importance projection*, VIP) different algorithms such as SVM, genetic algorithms, taboo search, DT and others are used to exclude variable of minor (negligible) importance [34]. VIP explains to what extent a descriptor contributes to the dependent variable Y [38]. On theory, selection of variables is not advised, as adding features should not decrease the accuracy. On practice, it is highly recommended, as it provides faster and more cost-effective predictors [43]. Two approaches are possible in order to conduct the selection – either eliminating variables from a full set (backward selection), or with addition starting from a single variable [34].

MACHINE LEARNING TECHNIQUES IN PCM

Cortes *et al.* describes the most important features of a "good algorithm": it should be easily interpretable with low training time; should be able to provide individual interval of confidence for the predictions and also able to consider the experimental uncertainty [35].

PLS is one of the most commonly used, highly interpretable, but requires cross terms. It can be considered as an extension of PCA [33]. Examples of use are seen in the works of Lapinsh *et al.* for the modeling of melanocortin receptors and HIV protease susceptibility [39, 44], dengue virus NS3 proteases by Prusis *et al.* [45]. *Rough set modeling* is non-linear rule-based machine learning method using IF-THEN rules for classification. It is very interpretable, but is unable to provide numerical values [46, 47]. *SVM* are also non-linear, interpretation of models is difficult, but can successfully extrapolate information to retrieve new active compounds. *Random forests* (RF) can be used for both classification and regression. Interpretability is better than SVM, while maintaining the accuracy; it can also measure the relative importance of descriptors and compound similarity [34]. RF could be described as multiple decision trees (DT, another method) with randomly selected variables; disadvantages are the high memory requirements and the lack of error estimate output [35]. Example of use can be seen in the work of De Bruyn *et al.* who utilized RF in modeling inhibitors for OATP1B1/3 [48]. *Gaussian Processes* (GP) are non-parametric Bayesian techniques; they are able to provide individual interval of confidence for the predictions and also able to consider the experimental uncertainty. Drawback is the requirement of long training time [35]. Other methods worth to mention include *Neural net modeling*, *Naïve Bayesian classifier*, *Decision tree* (DT).

VALIDATION METHODS

Y-scrambling, also known as Y-randomization ensures that the model is not based on chance correlations. This is achieved by reassigning activity values to different molecules and repeating the modeling. If the random models show comparable performance to the original one, then it is not a valid model [49]. Three types of internal validation (also known as cross validation) can be outlined: *leave-one-out* (LOO, with the subtypes of leave-one-target-out LOTO and leave-one-compound-out LOCO), *double loop cross validation* and *n-fold cross validation*. In double loop cross validation in the outer loop all data objects are randomly split into two subsets referred to as training and test set, the latter exclusively used for model assessment. The training set is used in the inner (internal) loop for model building and model selection; it is repeatedly split into construction and validation data sets [50]. In *n*-fold cross validation the initial set is divided to *n* subsets. Iteratively each of the subsets is excluded and the model is trained on the remaining *n*-1 subsets [34].

In *external validation* an external dataset with known activity values and completely unknown to the model is used for evaluation; such is used by Prusis *et al.* [37] *Prospective validation* is when model performance is confirmed experimentally; hence it is the best possible way to validate. Examples could be found in the works of Yabuuchi *et al.* [51], van Westen *et al.* [52], Dakshanamurthy *et al.* [53], De Bruyn *et al.* [48] and others.

APPLICATIONS

PCM is a method with a broad applicability, not only in rational drug design, but also in drug repositioning, interactions and toxicity predictions, repositioning of approved pharmaceutical products. It could be used not only for protein targets, but for complex systems, such as cell lines [35]. The reason PCM is such a universal tool is behind its simple requirements – there should be (i) consistent interaction data; (ii) numerical descriptions of relevant properties of ligand and target space; (iii) and a proper non-linear method for correlation. PCM do not require high-resolution of 3D structures and can encompass many targets into a single model [33]. Over the last 15 years many works report benefits after use of PCM. Detailed examples for PCM applications could be found in the excellent reviews by Bender [34, 35].

LIMITATIONS AND PRECAUTIONS

At its origin, PCM is a ligand-based method for drug design and the generated PCM models have an applicability domain, *i.e.* they are valid only for the chemical space defined by the tested ligands [34,35]. At the same time, PCM also is a structure-based method because it uses information about the structure of the studied targets, which means that the PCM models are valid only for the tested targets. These limitations make the extrapolated predictions risky and unreliable.

The poor predictive ability of the PCM models could come from several sources. First, the biological data (IC_{50}, pK_i, EC_{50}, LC_{50}, MIC, etc.) might contain experimental errors (systematic or random), errors due to a low resolution of the used devices, variations coming from the application of different methods and conditions. Second, the descriptors used to describe the chemical structure and properties of the studied ligands and proteins are of a different order and metrics, some are quantitative, other – binary, most of them are

collinear. These differences accumulate particularly in the cross-terms. In order to assess the real contribution (weight) of each descriptor, all descriptors used in the model should be scaled (to unit variance) or/and normalized (normally distributed). Third, the PCM matrices usually contain more variables than observations. This high dimensionality might lead to overfitting [54]. Robust validation procedures and techniques should be used to avoid chance correlations and to derive models with good predictive ability. Finally, a trade-off between interpretability and accuracy of the models exist in QSAR and particularly in PCM. Occam's razor ("The simplest explanation is usually the correct one") is helpful in most cases.

PROTEOCHEMOMETRIC MODELLING OF PEPRIDE BINDING TO HLA CLASS II PROTEINS

Since high affinity MHC binders can be potential vaccine candidates, tools for their *in silico* prediction are of a particular interest for immunologists. There are available databases such as Immune Epitope Database (IEDB) [55] and data, extracted from them has been already a subject of PCM modeling. While for class II there are existing PCM studies for HLA-DR and HLA-DQ [56, 57], for HLA-DP only classical QSAR approaches has been conducted [58] and proteochemometrics is still pending.

To obtain the models, iterative self-consistent PLS-based (ISC-PLS) algorithm has been used. In this technique, the first model is extracted from the initial training set. After cross validation and estimation of the root mean squared error of prediction (RMSEP), the optimum number of principal components is derived. The best predicted binders are used to create a second training set and derive a second model. The second model, in turn, is used to predict binding affinity of the initial training set and the best binders again are collected and used for a training set (third set). This procedure is repeated until two identical (or close to identical) consecutive training sets are obtained [56, 57]. Detailed description of ISC-PLS algorithm is described elsewhere [59]. Models have been validated using cross-validation techniques and external prediction. Based on them was developed EpiTOP – a tool for MHC class II binding prediction, which is freely accessible at http://www.ddg-pharmfac.net/epitop [60].

Proteochemometric Modelling of Peptide Binding to HLA-DRB1 Proteins

For HLA-DRB1 proteins Dimitrov *et al.* [56] used a training set of 2666 peptides, extracted from IEDB. The test set was originating from another database – AntiJen [61], and consisted of 356 binders, unknown to the models. Overlapping nonamers from the binders were extracted and encoded with *z*-descriptors (L-block, 27 descriptors). Cross terms for adjacent positions (L12), every second position (L13) and combinations of them (L123) were also included to deal with the non-linearity.

The HLA class II proteins included 12 sequences, namely: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1301 and DRB1*1501. Data was collected from the IMGT/HLA database [4]. Only the polymorphic amino acids from the binding site were considered in the analysis, which were 18 residues from the β-chain. Set of 54 (18 x 3) formed the P-block.

The whole *X* matrix consisted of five blocks: L, P, LP, L12, L13 and L123. Different combinations of them were used to derive the models via the ISC-PLS algorithm. Models had moderate goodness of fit, as expressed by r^2, ranging from 0.685 to 0.732. Their internal predictive ability was good, ranging from 0.678 to 0.719. The most predictive model was L +P + L12 model, which had the highest r^2_{pred} value. Further analysis described the pocket and anchor positions of importance.

Proteochemometric Modelling of Peptide Binding to HLA-DQ Proteins

The second PCM study was performed on HLA-DQ proteins [57], which are strongly associated with different autoimmune diseases – either protection or susceptibility. Such are type 1 Diabetes mellitus [62], celiac disease [63], multiple sclerosis [64], pemphigus vulgaris [65], rheumatoid arthritis [66]. Certain alleles (non-$Asp^{57\beta}$) are known to interact with peptides such as insulin β-chain, gliadin, glutenin; hence aspartic acid at position 57 is known to protect from type 1 diabetes [67].

The main difference between HLA-DQ and HLA-DR modeling is that in former both α- and β-chains exhibit polymorphism in the binding site. Another modeling obstacle is a deletion at position 54. Protein binding site consists of four binding pockets, which correspond to primary and secondary binding positions in ligands. Proteins selected for the study were the five most frequent alleles – DQA1*04:01/DQB1*04:02, DQA1*01:01/DQB1*05:01, DQA1*01:02/DQB1*06:02, DQA1*05:01/DQB1*03:01, DQA1*03:01/DQB1*03:02.

Again binders were extracted from IEDB, this time they were used to form both the training and the test set. For this purpose, they were sorted by their affinity value and divided into five groups. Twenty percent of each group was selected randomly to form the test set.

Descriptors were similar as those in HLA-DR, with the difference that here the P-block consisted of two separate blocks – PA for the α-chain (21 x 3 = 63 descriptors), and PB-block for the β-chain (24 x 3 = 72 descriptors). Descriptors for the deletion at position 52 took zero value. Cross-terms were introduced based on the distance between polymorphic residues in ligand and protein. The whole *X* matrix consisted of 312 variables in four blocks: L (27), PA (63), PB (72) and LP (150). ISC-PLS algorithm was used to derive the model, which was further assessed by r^2 (goodness of fit), q^2 (cross validation in 10 groups), r^2_{pred} (external validation by a test set) and r^2_{LOTO} (leave-one-target-out cross validation). Independent test set showed excellent predictive ability with r^2_{pred} = 0.808, AUC = 0.965, 92.5% accuracy at threshold of pIC_{50} = 5.3 and average sensitivity of 83% among the top 10% best predicted nonamers.

EpiTOP – a Proteochemometric Tool for MHC Binding Prediction

The results for both HLA-DR and HLA-DQ models were implemented in EpiTOP [60]. EpiTOP is easy to use, freely accessible online tool, which uses a quantitative matrix to predict binding affinities for MHC proteins (http://www.ddg-pharmfac.net/epitop). It is written in Python and HTML, and integrating the MySQL database environment. EpiTOP identifies peptides binding to different alleles within protein sequences, with options to vary HLA allele and cutoff. Performance has been assessed separately for HLA-DR and HLA-DQ predictions.

AntiJen, IEDB and Lin's datasets were used for HLA-DR proteins. EpiTOP was compared to eight other servers at five different thresholds [56]. The best results were obtained with Lin's dataset, where comparing the average AUC values EpiTOP ranked second (after NetMHCIIpan). When tested to identify peptide binding core, EpiTOP identified 8 out of 11 (73%).

For HLA-DQ binding prediction assessment [57], EpiTOP was compared to two artificial neural networks – NetMHCII and NetMHCpan [68, 69]. External test set of 660 peptides was used to assess the performance of the servers in terms of sensitivity, specificity, positive predicted value (ppv), F1-score and Matthews correlation coefficient (MCC) at threshold of pIC_{50} = 5.3. EpiTOP overall outperformed both servers (only in specificity NetMHCII showed lightly better performance).

As a PCM model, EpiTOP has an applicability domain. It is valid only for the target MHC proteins included in the model. However, the ligand space is wider, limited only by the binding peptide length (only nonamers) and by the type of residues (only the 20 naturally occurring amino acids).

CONCLUSION

Proteochemometrics is a rational method for quantitative structure – activity relationship studies, suitable for multi-target drug discovery. It utilizes information from both ligand and protein structures and combines it to derive models with clear physical sense and interpretability. Because of the enormous polymorphism, the MHC proteins of both classes I and II are suitable targets for PMC modeling. The developed models for HLA-DRB1 and HLA-DQ binding prediction showed excellent predictive ability and accuracy. The development of PCM models for other MHC alleles is forthcoming.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Doytchinova, I.A.; Flower, D.R. QSAR and the Prediction of T-Cell Epitopes. *Curr. Proteomics*, **2008**, *5*(2), 73–95.

[2] Ingolotti, M.; Kawalekar, O.; Shedlock, D.J.; Muthumani, K.; Weiner, D.B. DNA vaccines for targeting bacterial infections. *Expert Rev. Vaccines*, **2010**, *9*(7), 747-763.

[3] Janeway, Jr C.A.; Travers, P.; Walport, M.; Shlomchik, M.J. The major histocompatibility complex and its functions. In: *Immunobiology: The Immune System in Health and Disease.* 5th edition. New York: Garland Science; **2001.**

[4] Robinson, J.; Halliwell, J.A.; Hayhurst, J.H.; Flicek, P.; Parham, P.; Marsh, S.G.E. The IPD and IMGT/HLA database: allele variant databases *Nucleic Acids Res.,* **2015**, *43*, D423-431.

[5] Bjorkman, P.J.; Saper, M.A.; Samraoui, B.; Bennett, W.S.; Strominger, J.L.; Wiley, D.C. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, **1987**, *329*(6139), 506-512.

[6] Murthy, V.L.; Stern, L.J. The class II MHC protein HLA-DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding. *Structure*, **1997**, *5*(10), 1385-1396.

[7] Menssen, R.; Orth, P.; Ziegler, A.; Saenger, W. Decamer-like conformation of a nona-peptide bound toHLA-B*3501 due to non-standard positioningof the C terminus. *J. Mol. Biol*., **1999**, *285*(2), 645-653.

[8] Fan, Q.R.; Wiley, D.C. Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor. *J. Exp. Med*., **1999**, *190*(1), 113-123.

[9] Siebold, C.; Hansen, B.E.; Wyer, J.R.; Harlos, K.; Esnouf, R.E.; Svejgaard, A.; Bell, J.I.; Strominger, J.L.; Jones, E.Y.; Fugger, L. Crystal structure of HLA-DQ0602 that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proc. Natl. Acad. Sci. USA*, **2004**, *101*(7), 1999- 2004.

[10] Dai, S.; Murphy, G.A.; Crawford, F.; Mack, D.G.; Falta, M.T.; Marrack, P.; Kappler, J.W.; Fontenot, A.P. Crystal structure of HLA-DP2 and implications for chronic beryllium disease. *Proc. Natl. Acad. Sci. USA*, **2010**, *107*(16), 7425-7430.

[11] Oyston, P.; Robinson, K. The current challenges for vaccine development. *J. Med. Microbiol.*, **2012**, *61*(7), 889-894.

[12] Guidance for Industry. Immunogenicity Assessment for Therapeutic Protein Products. **2014**.

[13] De Groot, A.S.; Moise, L. Prediction of immunogenicity for therapeutic proteins: state of the art. *Curr. Opin. Drug Discov. Devel.*, **2007**, *10*(3), 332.

[14] Tomar, N.; De, R.K. Immunoinformatics: an integrated scenario: Immunoinformatics. *Immunology*, **2010**, *131*(2), 153-168.

[15] Doytchinova, I.A.; Flower, D. R. QSAR and the prediction of T-cell epitopes. *Curr. Proteomics*, **2008**, *5*(2), 73-95.

[16] Flower, D.R.; Macdonald, I.K.; Ramakrishnan, K.; Davies, M.N.; Doytchinova, I.A. Computer-aided selection of candidate vaccine antigens. *Immunome Res*., **2010**, *6*(Suppl. 2), S1.

[17] Patronov, A., Doytchinova, I.A. T-cell epitope vaccine design by immunoinformatics. *Open Biol.*, **2013**, *3*(1), 120139.

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

[19] Suhrbier, A.; Schmidt, C.; Fernan, A. Prediction of an HLA B8-restricted influenza epitope by motif. *Immunology*, **1993**, *79*(1), 171-173.

[20] Hammer, J.; Bono, E.; Gallazzi, F.; Belunis, C.; Nagy, Z.; Sinigaglia, F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J. Exp. Med.*, **1994**, *180*(6), 2353–2358.

[21] Cochlovius, B.; Stassar, M.; Christ, O.; Raddrizzani, L.; Hammer, J.; Mytilineos, I.; Zöller, M. In vitro and in vivo induction of a Th cell response toward peptides of the melanoma-associated glycoprotein 100 protein selected by the TEPITOPE program. *J. Immunol.*, **2000**, *165*(8), 4731–4741.

[22] Zhang, L.; Chen, Y.; Wong, H.-S.; Zhou, S.; Mamitsuka, H.; Zhu, S. TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. *PLoS ONE*, **2012**, *7*(2), e30483.

[23] Doytchinova, I.A.; Blythe, M.J.; Flower, D.R. Additive Method for the Prediction of Protein-Peptide Binding Affinity. Application to

the MHC Class I Molecule HLA-A*0201. *J. Proteome Res.*, **2002**, *1*(3), 263-272.

[24] Guan, P.; Doytchinova, I.A.; Zygouri, C.; Flower, D.R. MHCPred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res.*, **2003**, *31*(13), 3621-3624.

[25] Udaka, K.; Mamitsuka, H.; Nakaseko, Y.; Abe, N. Prediction of MHC class I binding peptides by a query learning algorithm based on hidden Markov models. *J. Biol. Phys.*, **2002**, *28*(2), 183–194.

[26] Nielsen, M.; Lundegaard, C.; Worning, P.; Hvid, C.S.; Lamberth, K.; Buus, S.; Brunak, S.; Lund, O. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **2004**, *20*(9), 1388–1397.

[27] Nanni, L. Machine learning algorithms for T-cell epitopes prediction. *Neurocomputing*, **2006**, *69*(7-9), 866–868.

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

[29] Atanasova, M.; Patronov, A.; Dimitrov, I.; Flower, D.R.; Doytchinova, I. EpiDOCK – a molecular docking-based tool for MHC class II binding prediction. *Protein Eng. Des. Sel.*, **2013**, *26*(10), 631-634.

[30] Sezerman, U.; Vajda, S.; DeLisi, C. Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci*., **1996**, *5*(7), 1272–1281.

[31] Adrian, P.E.; Rajaseger, G.; Mathura, V.S.; Sakharkar, M.K.; Kangueane, P. Types of inter-atomic interactions at the MHC–peptide interface: identifying commonality from accumulated data. *BMC Struct. Biol*., **2002**, *2*, 2.

[32] Rognan, D.; Scapozza, L.; Folkers, G.; Daser, A. Molecular dynamics simulation of MHC–peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry*, **1994**, *33*(38), 11476–11485.

[33] Lapins, M.; Wikberg, J. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinformatics*, **2010**, *11*(1), 339.

[34] van Westen, G.J.P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2011**, *2*(1), 16-30.

[35] Cortés-Ciriano, I.; Ain, Q.U.; Subramanian, V.; Lenselink, E.B.; Méndez-Lucio, O.; IJzerman, A.P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T.E.; van Westen, G.J.P.; Bender, A. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med. Chem. Commun.*, **2015**, *6*(1), 24-50.

[36] Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J.E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta*, **2001**, *1525*(1-2), 180-190.

[37] Prusis, P.; Uhlén, S.; Petrovska, R.; Lapinsh, M.; Wikberg, J.E. Prediction of indirect interactions in proteins. *BMC Bioinformatics*, **2006**, *7*(1), 167.

[38] Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J.E. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.*, **2002**, *61*(6), 1465–1475.

[39] Lapinsh, M. Proteochemometric Mapping of the Interaction of Organic Compounds with Melanocortin Receptor Subtypes. *Mol. Pharmacol.*, **2005**, *67*(1), 50-59.

[40] Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J.E.S. Improved approach for proteochemometrics modeling: application to organic compound--amine G protein-coupled receptor interactions. *Bioinformatics*, **2005**, *21*(23), 4289-4296.

[41] Strömbergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K.; Wikberg, J.E.S.; Komorowski, J.; Hvidsten, T.R. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins Struct. Funct. Bioinforma.*, **2006**, *65*(3), 568-579.

[42] Pérot, S.; Regad, L.; Reynès, C.; Spérandio, O.; Miteva, M.A.; Villoutreix, B.O.; Camproux, A.-C. Insights into an Original Pocket-Ligand Pair Classification: A Promising Tool for Ligand Profile Prediction. *PLoS ONE*, **2013**, *8*(6), e63730.

[43] Wegner, J.K.; Frohlich, H.; Zell, A. Feature Selection for Descriptor Based Classification Models. 1. Theory and GA-SEC Algorithm. *J. Chem. Inf. Model.*, **2004**, *44*(3), 921-930.

[44] Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J.E. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics*, **2008**, *9*(1), 181.

[45] Prusis, P.; Lapins, M.; Yahorava, S.; Petrovska, R.; Niyomrattanakit, P.; Katzenmeier, G.; Wikberg, J.E.S. Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. *Bioorg. Med. Chem.*, **2008**, *16*(20), 9369-9377.

[46] Strömbergsson, H.; Prusis, P.; Midelfart, H.; Wikberg, J.E.; Komorowski, H.J. Proteochemometrics Modeling of Receptor-Ligand Interactions Using Rough Sets. In: *German Conference on Bioinformatics*. Citeseer, **2004**, 85–94.

[47] Strömbergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M.; Wikberg, J.E.S.; Komorowski, J. Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins Struct. Funct. Bioinforma.*, **2006**, *63*(1), 24-34.

[48] De Bruyn, T.; van Westen, G.J.P.; IJzerman, A.P.; Stieger, B.; de Witte, P.; Augustijns, P.F.; Annaert, P.P. Structure-Based Identification of OATP1B1/3 Inhibitors. *Mol. Pharmacol.*, **2013**, *83*(6), 1257-1267.

[49] Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C.P.; Agrawal, R.K. Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov.*, **2011**, *3*, 511–519.

[50] Freyhult, E.; Prusis, P.; Lapinsh, M.; Wikberg, J.E.; Moulton, V.; Gustafsson, M.G. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *BMC Bioinformatics*, **2005**, *6*(1), 50.

[51] Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.*, **2014**, *7*(1), 472-472.

[52] van Westen, G.J.P.; Wegner, J.K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. *PLoS ONE*, **2011**, *6*(11), e27518.

[53] Dakshanamurthy, S.; Issa, N.T.; Assefnia, S.; Seshasayee, A.; Peters, O.J.; Madhavan, S.; Uren, A.; Brown, M.L.; Byers, S.W. Predicting New Indications for Approved Drugs Using a Proteochemometric Method. *J. Med. Chem.*, **2012**, *55*(15), 6832-6848.

[54] Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Model*., **2004**, *44*(1), 1-12.

[55] Vita, R.; Overton, J.A.; Greenbaum, J.A.; Ponomarenko, J.; Clark, J.D.; Cantrell, J.R.; Wheeler, D.K.; Gabbard, J.L.; Hix, D.; Sette, A.; Peters, B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*., **2014**, *43*(Database issue):D405-D412.

[56] Dimitrov, I.; Garnev, P.; Flower, D.R.; Doytchinova, I. Peptide binding to the HLA-DRB1 supertype: A proteochemometrics analysis. *Eur. J. Med. Chem.*, **2010**, *45*(1), 236-243.

[57] Dimitrov, I.; Doytchinova, I. Peptide Binding Prediction to Five Most Frequent HLA-DQ Proteins – a Proteochemometric Approach. *Mol. Informatics*, **2015**, *34*(6-7), 467-476.

[58] Ivanov, S.; Dimitrov, I.; Doytchinova, I. Quantitative Prediction of Peptide Binding to HLA-DP1 Protein. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2013**, *10*(3), 811-815.

[59] Doytchinova, I.A.; Flower, D.R. Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics*, **2003**, *19*(17), 2263-2270.

[60] Dimitrov, I.; Garnev, P.; Flower, D.R.; Doytchinova, I. EpiTOP - a proteochemometric tool for MHC class II binding prediction. *Bioinformatics*, **2010**, *26*(16), 2066-2068.

[61] Toseland, C.P.; Clayton, D.J.; McSparron, H.; Hemsley, S.L.; Blythe, M.J.; Paine, K.; Doytchinova, I.A.; Guan, P.; Hattotuwagama, C.K.; Flower, D. R. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*, **2005**, *1*(4), 82–93.

[62] Todd, J.A.; Bell, J.I.; McDevitt, H.O. HLA-DQβ gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature*, **1987**, *329*(6140), 599-604.

[63] Sollid, L.M.; Markussen, G.; Ek, J.; Gjerde, H.; Vartdal, F. Thorsby, E. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J. Exp. Med.*, **1989**, *169*(1), 345–350.

[64] Spurkland, A.; Celius, E.G.; Knutsen, I.; Beiske, A.; Thorsby, E.; Vartdal, F. The HLA-DQ(alpha 1*0102, beta 1*0602) heterodimer may confer susceptibility to multiple sclerosis in the absence of the

HLA-DR(alpha 1*01, beta 1*1501) heterodimer. *Tissue Antigens*, **1997**, *50*(1), 15-22.

[65] Delgado, J.C.; Hameed, A.; Yunis, J.J.; Bhol, K.; Rojas, A.I.; Rehman, S.B.; Khan, A.A.; Ahmad, M.; Alper, C.A.; Ahmed, A.R.; Yunis, E.J. Pemphigus Vulgaris Autoantibody Response is Linked to HLA-DQB10503 in Pakistani Patients. *Hum. Immunol.*, **1997**, *57*(2), 110-119.

[66] Zanelli, E.; Breedveld, F.C.; de Vries, R.R.P. HLA association with autoimmune disease: a failure to protect? *Rheumatology*, **2000**, *39*(10), 1060-1066.

[67] Morel, P.A.; Dorman, J.S.; Todd, J.A.; McDevitt, H.O.; Trucco, M. Aspartic acid at position 57 of the HLA-DQ beta chain protects against type I diabetes: a family study. *Proc. Natl. Acad. Sci.*, **1988**, *85*(21), 8111–8115.

[68] Nielsen, M.; Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, **2009**, *10*(1), 296.

[69] Hoof I.; Peters B.; Sidney J.; Pedersen L.E.; Sette A.; Lund O.; Buus, S.; Nielsen, M. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, **2009**, *61*(1), 1-13.