Int J Pept Res Ther (2016) 22:445–450
DOI 10.1007/s10989-016-9524-x

QSPR-Perturbation Models for the Prediction of B-Epitopes from Immune Epitope Database: A Potentially Valuable Route for Predicting "In Silico" New Optimal Peptide Sequences and/or Boundary Conditions for Vaccine Development

Severo Vázquez-Prieto[1,2] · Esperanza Paniagua[2] · Florencio M. Ubeira[2] · Humberto González-Díaz[3,4]

Accepted: 14 March 2016 / Published online: 19 March 2016

Abstract In the present study, three different physicochemical molecular properties for peptides were calculated using the program MARCH-INSIDE: atomic polarizability, partition coefficient, and polarity. These measures were used as input parameters of a linear discriminant analysis (LDA) in order to develop three different quantitative structure–property relationship (QSPR)-perturbation models for the prediction of B-epitopes reported in the immune epitope database (IEDB) given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. The accuracy, sensitivity and specificity of the models were >90 % for both training and cross-validation series. The statistical parameters of the models were compared to the results achieved with the electronegativity QSPR-perturbation model previously reported by González-Díaz et al. (J Immunol Res. doi:10.1155/2014/768515, 2014). The results indicate that this type of approach may constitute a potentially valuable route for predicting "in silico" new optimal peptide sequences and/or boundary conditions for vaccine development.

✉ Severo Vázquez-Prieto
severovazquezprieto@gmail.com

[1] Instituto de Ciencias Biomédicas, Facultad de Ciencias de la Salud, Universidad Autónoma de Chile, Temuco, Chile

[2] Laboratorio de Parasitología, Departamento de Microbiología y Parasitología, Facultad de Farmacia, Universidad de Santiago de Compostela, Campus Vida, 15782 Santiago De Compostela, Spain

[3] Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain

[4] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

Keywords Epitopes · Vaccine design · Perturbation theory · QSAR/QSPR models · Markov chains

Introduction

The immune epitope database (IEDB: http://www.iedb.org) contains data related to antibody and T cell epitopes for humans, non-human primates, rodents, and other animal species (Vita et al. 2010). This system registers an important amount of information about the molecular structure and the experimental conditions (c_{ij}) in which different *i*-th molecules were determined to be immune epitopes or not. With the availability of these types of databases (Gao and Kurgan 2014), epitope prediction using computational methods has emerged as a promising approach for developing peptide-based vaccines. Such techniques allow for screening among large numbers of possible immune-active peptides in order to find those likely to induce an immune response to a particular cell type, providing a fast and cost-effective way to identification of potential candidates for vaccine development (Du et al. 2007; Chen et al. 2007).

Quantitative structure–activity/property relationship (QSAR/QSPR) methods let transform molecular structures into numeric molecular descriptors (λ_i) and find relationships between these structures and their biological activity. Consequently, these techniques are widely used today to predict the properties of complex molecular systems, including peptides, proteins, RNAs, drug-protein complexes, and protein–protein complexes (see, e.g., Bermúdez et al. 1999; Agüero-Chapín et al. 2005; Du et al. 2005; Galindo et al. 2006; Chou and Shen 2008; Du et al. 2008a, b; Prado–Prado et al. 2008; Chou 2009; Du et al. 2009; Rodríguez-Soca et al. 2009; Viña et al. 2009; Wei et al. 2009; Toropov et al. 2012;

Toropova et al. 2015). Likewise, QSAR/QSPR methods have been successfully used in immunoinformatics to predict the propensity different molecular structures have for playing different roles in immunological processes (see, e.g., Doytchinova et al. 2004; Estrada et al. 2004; Gerberick et al. 2004; Xiao and Segal 2005; Bhasin et al. 2006; Barh et al. 2010; Bremel and Homan 2010; Díez-Rivero et al. 2010; Roberts and Patlewicz 2010; Bi et al. 2011; Martínez-Naves et al. 2011; Tenorio-Borroto et al. 2012; Fagerberg et al. 2013; Patlewicz et al. 2013).

On the other hand, perturbation theory comprises methods that add "small" variation terms to the mathematical description of problems with known solutions in order to find an appropriate solution for related problems with no known solutions. Accordingly, this theory has been widely used in all branches of knowledge, including bio-molecular sciences. The reader may see the interesting review by González-Díaz et al. (2013a) on this topic. In the same work, the authors also formulated a general-purpose perturbation theory for multiple-boundary QSAR/QSPR problems. Subsequently, this new modeling method was applied by González-Díaz et al. (2014) to develop an electronegativity QSPR-perturbation model for B-epitopes reported in IEBD able to predict the probability of occurrence of an epitope after a perturbation in the peptide sequence (m_i), source organism (*so*), host organism (*ho*), immunological process (*ip*), and experimental technique (*tq*) used.

In principle, there are more than 1600 different molecular descriptors (λ_i) that may be generalized and used to solve QSPR problems in chemical structures (Todeschini and Consonni 2008). In the present study, three different physicochemical molecular properties for peptide sequences reported in IEDB were calculated in order to develop three different QSPR models able to predict the efficiency of a new peptide as B-epitope given perturbations in m_i, *so*, *ho*, *ip*, and *tq*. The statistical parameters of the models were compared to the results achieved by the model developed by González-Díaz et al. (2014).

Materials and Methods

Calculation of Molecular Descriptors for Peptides

The same database recently utilized by González-Díaz et al. (2014) was used in the present study. The data contains variations in >50,000 peptides determined in experimental assays with boundary conditions involving >500 source organisms, >50 host organisms, >10 biological process, and >30 experimental techniques (González-Díaz et al. 2014). The calculation of the molecular descriptors was implemented in the program MARCH-INSIDE (González-Díaz et al. 2007), which makes use of a Markov Chain method to calculate the *k*-th mean values of different physicochemical molecular properties ${}^k\lambda(m_i)$ for *i*-th molecules (m_i). These ${}^k\lambda(m_i)$ values are calculated as an average of atomic properties (λ_i) for all atoms in the peptide molecule and its neighbors placed at a topological distance $d \leq k$. The parameter k is called the parameter of the Markov Chain, the natural power of the Markov matrix. In this work, the average value of all atomic polarizabilities ${}^k\alpha(m_i)$, partition coefficients ${}^kP(m_i)$, and polarities ${}^kPol(m_i)$ for all δ_i atoms connected to the *i*-th atom ($i \rightarrow j$) and their neighbors placed at a distance d $\leq$ 5 was calculated for all peptides (González-Díaz et al. 2013b):

$$ {}^k\lambda(m_i) = \frac{1}{6}\sum_{k=0}^{5} {}^k\lambda_j = \frac{1}{6}\sum_{k=0}^{5}\sum_{i\rightarrow j}^{\delta_i} p_k(\lambda_j)\cdot\lambda_j \quad (1) $$

The probabilities ${}^kp(\lambda_j)$ for the atomic properties in question were calculated using a Markov Chain model for the gradual effects of the neighboring atoms at different distances in the molecular backbone, as has been explained in detail in González-Díaz et al. (2013b).

Derivation of the QSPR-perturbation Models

In a recent work, González-Díaz et al. (2014) have applied the perturbation theory to the QSPR peptide prediction problem and formulated an electronegativity QSPR-perturbation model able to predict the probability of occurrence of a B-epitope after a variation in the structure and/or the boundary conditions of a peptide of reference. Therefore, the theoretical foundations of the method are not detailed here. In the present work, three new QSPR-perturbation models for prediction of B-epitopes reported in IEDB were developed using different types of molecular descriptors $\lambda(m_i)$ to codify structural information: atomic polarizability, partition coefficient, and polarity. The construction of this type of models has been explained in detail before (González-Díaz et al. 2014); therefore, only the general equation is presented:

$$ \lambda(\varepsilon_{ij})_{\text{new}} = {}'c_0\cdot\lambda(\varepsilon_{qr})_{\text{ref}} + \sum_{j=1}^{4} {}'d_{ij}\cdot\Delta\Delta\lambda_{ijqr} + {}'e_0 \quad (2) $$

Here, in line with González-Díaz et al. (2014), $\lambda(\varepsilon_{ij})_{\text{new}}$ is the efficiency function as epitope of a new peptide obtained after a change in the structure and/or the boundary conditions $c_j \equiv (c_0, c_1, c_2, c_3 \ldots c_n)$ of a peptide of reference. The set of boundary conditions used here are the same reported in IEDB: c_0 = the specific peptide; c_1 = the organism that expresses the peptide (so_j); c_2 = the host organism exposed to the peptide (ho_j); c_3 = the immunological process (ip_j); and c_4 = the experimental technique (tq_j). The variable $\lambda(\varepsilon_{qr})_{\text{ref}}$ refers to a known efficiency function as epitope of a peptide of reference experimentally

determined under a set of c_j boundary conditions. The function $\lambda(\varepsilon_{ij})$ was defined as a discrete value function for classification purpose: $\lambda(\varepsilon_{ij}) = 1$ for epitopes reported in the conditions c_j and $\lambda(\varepsilon_{ij}) = 0$, when otherwise. The values c_0 and d_{ij} are the coefficients obtained for the linear discriminant analysis (LDA) classification functions. The variational perturbation terms $\Delta\Delta\lambda_{ijqr}$ account both for the deviation of the molecular descriptors of all amino acids in the sequence of the new peptide with respect to the peptide of reference and with respect to all boundary conditions. The constant e_0 represents the independent term of the model (González-Díaz et al. 2014). The expanded formula of the models is given below:

$$\lambda(\varepsilon_{ij})_{\text{new}} = {}'c_0 \cdot \lambda(\varepsilon_{qr})_{ref} + \sum_{j=1}^{4} {}'d_{ij} \cdot \left((\lambda_i - \lambda_j) - (\lambda_q - \lambda_r)\right) + {}'e_0 \quad (3)$$

Statistical Analysis

An LDA was carried out using the STATISTICA 6.0 software (StatSoft.Inc. 2002). In the absence of a true external data set, the original data set was randomly divided into two series, a training series for model development and a cross-validation series for model validation (75 and 25 % of the data set, respectively). A forward stepwise strategy was used for variable selection, and the statistical significance of the models was determined by calculating the canonical correlation coefficient (R_c) and U-statistic. The accuracy, specificity, and sensitivity for the training and cross-validation series were also examined (Hill and Lewicki 2006). In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang 1995). However, of these three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in Chou (2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., Zhang et al. 2008; Esmaeili et al. 2010; Mohabatkar 2010; Sahu and Panda 2010; Khosravian et al. 2013; Mohabatkar et al. 2013). However, to reduce the computational time, the independent dataset test was adopted in this study.

Results and Discussion

In the present work, three different QSPR-perturbation models were developed, one for each class of molecular descriptor calculated with the software MARCH-INSIDE: atomic polarizability (α), partition coefficient (*P*), and polarity (*Pol*). The following were the best QSPR-perturbation models found:

Polarizability-perturbation model:

$$\lambda(\varepsilon_{ij})_{new} = -4.683 \cdot \lambda(\varepsilon_{ij})_{ref} - 44.099 \cdot \Delta\alpha_{seq} + 2.666 \cdot \Delta\Delta\alpha_{ho} + 16.482 \cdot \Delta\Delta\alpha_{so} - 21.668 \cdot \Delta\Delta\alpha_{ip} + 47.096 \cdot \Delta\Delta\alpha_{tq} + 2.0103$$
$$N = 155169 \quad Rc = 0.91 \quad U = 0.18 \quad p < 0.01 \quad (4)$$

Partition coefficient-perturbation model:

$$\lambda(\varepsilon_{ij})_{new} = -4.345 \cdot \lambda(\varepsilon_{ij})_{ref} - 98.689 \cdot \Delta P_{seq} + 7.741 \cdot \Delta\Delta P_{ho} + 30.378 \cdot \Delta\Delta P_{so} - 7.073 \cdot \Delta\Delta P_{ip} + 69.851 \cdot \Delta\Delta P_{tq} + 1.851$$
$$N = 155169 \quad Rc = 0.89 \quad U = 0.21 \quad p < 0.01 \quad (5)$$

Polarity-perturbation model

$$\lambda(\varepsilon_{ij})_{new} = -4.846 \cdot \lambda(\varepsilon_{ij})_{ref} - 708.845 \cdot \Delta Pol_{seq} + 37.565 \cdot \Delta\Delta pol_{ho} + 206.803 \cdot \Delta\Delta Pol_{so} - 204.545 \cdot \Delta\Delta Pol_{ip} + 661.274 \cdot \Delta\Delta Pol_{tq} + 2.084$$
$$N = 155169 \quad Rc = 0.92 \quad U = 0.16 \quad p < 0.01 \quad (6)$$

In these equations, *N* is the number of cases used to train the models, R_C is the canonical correlation coefficient, and U is the Wilk's lambda or U-statistic. In line with González-Díaz et al. (2014), the output of the models $\lambda(\varepsilon_{ij})_{\text{new}}$ is a real value function that scores the propensity with which a new peptide obtained after perturbation of the initial conditions acts as B-epitope. On the other side, the first input term $\lambda(\varepsilon_{ij})_{\text{ref}}$ is the scoring function λ of the efficiency of the initial process ε_{ij}. The function $\lambda(\varepsilon_{ij})_{\text{ref}} = 1$, if the *i*-th peptide could be experimentally demonstrated to be a B-epitope in the assay of reference (ref) carried out in the conditions c_j. $\lambda(\varepsilon_{ij})_{\text{ref}} = 0$ if otherwise. The perturbation terms $\Delta\lambda_{cj} = \lambda(m_q)_{\text{ref}} - \lambda(m_i)_{\text{new}}$ are the difference in the mean value of the molecular property in question for all amino acids in the sequence of the peptide of reference. The independent variables $\Delta\Delta\lambda_{cj} = \Delta\lambda_{cj\text{-ref}} - \Delta\lambda_{cj\text{-new}} = [\lambda(m_q)_{\text{ref}} - {}^*\lambda(c_{qr})_{\text{ref}}] - [\lambda(m_i)_{\text{new}} - {}^*\lambda(c_{ij})_{\text{new}}]$ quantify values of the conditions of the new assay *cj*-new that represent perturbations with respect to the initial conditions c_{ij}-ref of the assay of reference. The quantities ${}^*\lambda(c_{ij})$ and ${}^*\lambda(c_{qr})$ are the average values of the mean values $\lambda(m_i)$ and $\lambda(m_q)$ of the molecular property in question for all new and reference peptides in IEDB that are epitopes under the *j*-th or *r*-th boundary

Table 1 Detailed training and cross-validation results for the different QSPR models developed in this work

Molecular descriptor	Training series[a]				Parameters	Cross-validation series[a]			
	0	1		%		%		0	1
Polarizability	83699	3568	0	95.91	Sp	95.97	0	27736	1164
	5486	62416	1	91.92	Sn	91.6	1	1859	20267
			Total	94.17	Ac	94.08	Total		
Partition coefficient	83396	3871	0	95.56	Sp	95.67	0	27648	1252
	5821	62081	1	91.43	Sn	91.09	1	1971	20155
			Total	93.75	Ac	93.68	Total		
Polarity	83902	3365	0	96.14	Sp	96.28	0	27826	1074
	5268	62634	1	92.24	Sn	91.87	1	1798	20328
			Total	94.44	Ac	94.37	Total		

Sp specificity, *Sn* sensitivity, *Ac* accuracy

[a] Rows: observed classifications; columns: predicted classifications

condition (González-Díaz et al. 2014). The variational perturbation terms $\Delta\Delta\lambda_{cj}$ resemble terms typical of perturbation theory and moving average functions used in Box-Jenkins models in time series (Box and Jenkins 1970; González-Díaz et al. 2013a). This type of information has been recently incorporated inside QSAR/QSPR models (Speck-Planche et al. 2013a, b, c; Vázquez-Prieto et al. 2014).

The models obtained here are very stable and robust, yielding values of accuracy, sensitivity and specificity >90 % for both training and cross-validation series (see Table 1). The present results are excellent compared with other similar models in the literature including moving average or perturbation models (Speck-Planche et al. 2012a, b; González-Díaz et al. 2013a). These models are not able to improve the model developed by González-Díaz et al. (2014) in terms of specificity (97 and 97.1 %), sensitivity (93.6 and 93.3 %), and accuracy (95.5 and 95.4 %) for both training and cross-validation series respectively. However, the results obtained are very similar and the values of different statistical parameters demonstrate the high significance of the models, validating the consistency of the method. Thus, the information obtained from the four different types of QSPR-perturbation models developed to date may be combined to increase the likelihood of a correct prediction of new epitopes or the optimization of known peptides towards computational vaccine design (González-Díaz et al. 2014).

Because user-friendly and publicly accessible web-servers represent the future direction for developing more practically useful models, simulated methods and predictors (Chou and Shen 2009), efforts shall be made in the future work to provide a web-server for the method presented in this paper, as done in a series of recent papers (see, e.g., Guo et al. 2014; Lin et al. 2014; Liu et al. 2014; Qiu et al. 2014a, b; Xu et al. 2014).

Conclusions

In conclusion, this work has demonstrated that atomic polarizability, partition coefficient, and polarity values calculated with MARCH-INSIDE seem to also be good molecular descriptors for finding QSPR-perturbation models which are able to predict the results of variations in peptide sequences and experimental assay boundary conditions reported in IEBD. Consequently, this type of approach may constitute a potentially valuable route for predicting "in silico" new optimal peptide sequences and/or boundary conditions for vaccine development. In addition, this study may serve as a basis for building better and more reliable models in the future (e.g., consensus QSPR models). This computational technique is by no means aimed at replacing experimentation but rather helps us to somewhat rationalize this process, while at the same time reducing costs in terms of material resources and time.

Acknowledgments This work was supported by Grants GPC2014/058 from the Xunta de Galicia and AGL2011-30563-C03 from the Ministerio de Ciencia e Innovación, Spain.

Compliance with Ethical Standards

Conflict of interest Severo Vázquez-Prieto, Esperanza Paniagua, Florencio M. Ubeira, and Humberto González-Díaz declare that they have no conflict of interest.

Human and Animal Rights This article does not contain studies with human participants or animals performed by any of the authors.

References

Agüero-Chapín G, Varona-Santos J, de la Riva GA, Antunes A, González-Villa T, Uriarte E, González-Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy

M, Schoenberger SP, Stewart S, Surko P, Way S, Wilson S, Sette A (2005) The design and implementation of the immune epitope database and analysis resource. Immunogenetics 57:326–336

Barh D, Misra AN, Kumar A, Vasco A (2010) A novel strategy of epitope design in *Neisseria gonorrhoeae*. Bioinformation 5:77–82

Bermúdez CI, Daza EE, Andrade E (1999) Characterization and comparison of *Escherichia coli* transfer RNAs by graph theory based on secondary structure. J Theor Biol 197:193–205

Bhasin M, Reinherz EL, Reche PA (2006) Recognition and classification of histones using support vector machine. J Comput Biol 13:102–112

Bi J, Song R, Yang H, Li B, Fan J, Liu Z, Long C (2011) Stepwise identification of HLA-A*0201-restricted $CD8^+$ T-cell epitope peptides from herpes simplex virus type 1 genome boosted by a steprank scheme. Biopolymers 96:328–339

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco

Bremel RD, Homan EJ (2010) An integrated approach to epitope analysis II: a system for proteomic-scale prediction of immunological characteristics. Immunome Res 6(1):1

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428

Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6:262–274

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor Biol 273:236–247

Chou KC, Shen HB (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. Nat Sci 2:63–92

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Díez-Rivero CM, Chenlo B, Zuluaga P, Reche PA (2010) Quantitative modeling of peptide binding to TAP using support vector machine. Proteins 78:63–72

Doytchinova IA, Guan P, Flower DR (2004) Quantitative structure-activity relationships and the prediction of MHC supermotifs. Methods 34:444–453

Du QS, Mezey PG, Chou KC (2005) Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. J Comput Chem 26:461–470

Du QS, Wei YT, Pang ZW, Chou KC, Huang RB (2007) Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction. Protein Eng Des Sel 20:417–423

Du QS, Huang RB, Chou KC (2008a) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–260

Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008b) Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). J Comput Chem 29:211–219

Du QS, Huang RB, Wei YT, Pang ZW, Du LQ, Chou KC (2009) Fragment-based quantitative structure-activity relationship (FB-QSAR) for fragment-based drug design. J Comput Chem 30:295–304

Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263:203–209

Estrada E, Patlewicz G, Gutierrez Y (2004) From knowledge generation to knowledge archive. A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitization. J Chem Inf Comput Sci 44:688–698

Fagerberg T, Zoete V, Viatte S, Baumgaertner P, Alves PM, Romero P, Speiser DE, Michielin O (2013) Prediction of cross-recognition of peptide-HLA A2 by melan-a-specific cytotoxic T lymphocytes using three-dimensional quantitative structure-activity relationships. PLoS One 8(7):e65590

Galindo JF, Bermúdez CI, Daza EE (2006) tRNA structure from a graph and quantum theoretical perspective. J Theor Biol 240:574–582

Gao J, Kurgan L (2014) Computational prediction of B cell epitopes from antigen sequences. Methods Mol Biol 1184:197–215

Gerberick GF, Ryan CA, Kern PS, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA (2004) A chemical dataset for evaluation of alternative approaches to skin-sensitization testing. Contact Derm 50:274–288

González-Díaz H, Molina-Ruiz R, Hernández I (2007) MARCH-INSIDE version 3.0 (MARkov CHains INvariants for SImulation & DEsign). Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es

González-Díaz H, Arrasate S, Gómez-San Juan A, Sotomayor N, Lete E, Besada-Porto L, Ruso JM (2013a) New theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. Curr Top Med Chem 13:1713–1741

González-Díaz H, Arrasate S, Sotomayor N, Lete E, Munteanu CR, Pazos A, Besada-Porto L, Ruso JM (2013b) MIANN models in medicinal, physical and organic chemistry. Curr Top Med Chem 13:619–641

González-Díaz H, Pérez-Montoto LG, Ubeira FM (2014) Model for vaccine design by prediction of B-epitopes of IEDB given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. J Immunol Res. doi:10.1155/2014/768515

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529

Hill T, Lewicki P (2006) STATISTICS: methods and applications: a comprehensive reference for science, industry and data mining. StatSoft, Tulsa

Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett 20:180–186

Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42:12961–12972

Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou KC (2014) iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9:e106691

Martínez-Naves E, Lafuente EM, Reche PA (2011) Recognition of the ligand-type specificity of classical and non-classical MHC I proteins. FEBS Lett 585:3478–3484

Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17:1207–1214

Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med Chem 9:133–137

Patlewicz G, Ball N, Booth ED, Hulzebos E, Zvinavashe E, Hennes C (2013) Use of category approaches, read-across and (Q)SAR: general considerations. Regul Toxicol Pharmacol 67:1–12

Prado-Prado FJ, González-Díaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg Med Chem 16:5871–5880

Qiu WR, Xiao X, Lin WZ, Chou KC (2014a) iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. J Biomol Struct Dyn 6:1–12

Qiu WR, Xiao X, Lin WZ, Chou KC (2014b) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. Biomed Res Int 2014:947416

Roberts DW, Patlewicz GY (2010) Updating the skin sensitization in vitro data assessment paradigm in 2009-a chemistry and QSAR perspective. J Appl Toxicol 30:286–288

Rodríguez-Soca Y, Munteanu CR, Prado-Prado FJ, Dorado J, Pazos Sierra A, González-Díaz H (2009) Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. J Proteome Res. doi:10.1021/pr900827b

Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem 34:320–327

Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN (2012a) Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. Eur J Pharm Sci 47:273–279

Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN (2012b) In silico discovery and virtual screening of multi-target inhibitors for proteins in Mycobacterium tuberculosis. Comb Chem High Throughput Screen 15:666–673

Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN (2013a) Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. AntiCancer Agents Med Chem 13:791–800

Speck-Planche A, Kleandrova VV, Cordeiro MN (2013b) New insights toward the discovery of antibacterial agents: multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. Eur J Pharm Sci 48:812–818

Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN (2013c) Multi-target inhibitors for proteins associated with Alzheimer: in silico discovery using fragment-based descriptors. Curr Alzheimer Res 10:117–124

StatSoft.Inc. (2002) STATISTICA (data analysis software system), version 6.0. www.statsoft.com

Tenorio-Borroto E, Penuelas Rivas CG, Vasquez Chagoyan JC, Castanedo N, Prado-Prado FJ, García-Mera X (2012) ANN multiplexing model of drugs effect on macrophages; theoretical and flow cytometry study on the cytotoxicity of the antimicrobial drug G1 in spleen. Bioorg Med Chem 20:6181–6194

Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley, Weinheim

Toropov AA, Toropova AP, Raska I Jr, Benfenati E, Gini G (2012) QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. Struct Chem 23:1891–1904

Toropova MA, Veselinović AM, Veselinović JB, Stojanović DB, Toropov AA (2015) QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. Comput Biol Chem 59:126–130

Vázquez-Prieto S, González-Díaz H, Paniagua E, Vilas R, Ubeira FM (2014) A QSPR-like model for multilocus genotype networks of *Fasciola hepatica* in Northwest Spain. J Theor Biol 343:16–24

Viña D, Uriarte E, Orallo F, González-Díaz H (2009) Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharm 6:825–835

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res 38(Database issue):D854–D862

Wei H, Wang CH, Du QS, Meng J, Chou KC (2009) Investigation into adamantane-based M2 inhibitors with FB-QSAR. Med Chem 5:305–317

Xiao Y, Segal MR (2005) Prediction of genomewide conserved epitope profiles of HIV-1: classifier choice and peptide representation. Stat Appl Genet Mol Biol. doi:10.2202/1544-6115.1158

Xu R, Zhou J, Liu B, He YA, Zou Q, Wang X, Chou KC (2014) Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. J Biomol Struct Dyn 28:1–11

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34:565–572