QSPR-Perturbation Models for the Prediction of B-Epitopes from Immune Epitope Database: An Interesting Route for Predicting *"in silico"* New Optimal Peptide Sequences and/or Boundary Conditions for Vaccine Development

Severo Vázquez-Prieto *, Esperanza Paniagua and Florencio M. Ubeira
Laboratorio de Parasitología, Departamento de Microbiología y Parasitología, Facultad de Farmacia, Universidad de Santiago de Compostela, Campus Vida, Santiago de Compostela 15782, Spain
* Author to whom correspondence should be addressed; E-Mail: severovazquezprieto@gmail.com; Tel.: +34-881-815004; Fax: +34-981-593316.

Received: 28 October 2015 / Accepted: 28 October 2015 / Published: 2 December 2015

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. The results indicate that this type of approach may constitute an interesting route for predicting *"in silico"* new optimal peptide sequences and/or boundary conditions for vaccine development.

Keywords: Epitopes; Vaccine design; Perturbation theory; QSAR/QSPR models; Markov Chains

1. Introduction

The immune epitope database (IEDB) contains data related to antibody and T cell epitopes for humans, non-human primates, rodents, and other animal species (1). 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.

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. 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.

In a recent work, González-Díaz *et al*. (2) have developed 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 1,600 different molecular descriptors (λ_i) that may be generalized and used to solve QSPR problems in chemical structures (3). 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*.

2. 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 (Table 1). 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*. (2), the output of the models $\lambda(\varepsilon_{ij})_{new}$ 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 $\lambda(\varepsilon_{ij})_{ref}$ is the scoring function λ of the efficiency of the initial process ε_{ij}. The function $\lambda(\varepsilon_{ij})_{ref} = 1$, 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. $\lambda(\varepsilon_{ij})_{ref} = 0$ if otherwise. The perturbation terms $\Delta\lambda_{cj} = \lambda(m_q)_{ref} - \lambda(m_i)_{new}$ 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 $\Delta\Delta\lambda_{cj} = \Delta\lambda_{cj\text{-ref}} - \Delta\lambda_{cj\text{-new}} = [\lambda(m_q)_{ref} - {}^*\lambda(c_{qr})_{ref}] - [\lambda(m_i)_{new} - {}^*\lambda(c_{ij})_{new}]$ 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 condition.

The models obtained here are very stable and robust, yielding values of accuracy, sensitivity and specificity > 90% for both training and cross-validation series. These models are not able to improve the model developed by González-Díaz *et al*. (2). 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.

Table 1. The best QSPR-perturbation models found in this work.

Atomic polarizability (α)	$\lambda(\varepsilon_{ij})_{new} = -4.683 \cdot \lambda(\varepsilon_{ij})_{ref} - 44.099 \cdot \Delta\alpha_{seq} + 2.667 \cdot \Delta\Delta\alpha_{ho} + 16.482 \cdot \Delta\Delta\alpha_{so} - 21.668 \cdot \Delta\Delta\alpha_{ip} + 47.096 \cdot \Delta\Delta\alpha_{tq} + 2.0103$ $N = 155169 \quad Rc = 0.91 \quad U = 0.18 \quad p < 0.01$
Partition coefficient (P)	$\lambda(\varepsilon_{ij})_{new} = -4.345 \cdot \lambda(\varepsilon_{ij})_{ref} - 98.689 \cdot \Delta P_{seq} + 7.741 \cdot \Delta\Delta P_{ho} + 30.378 \cdot \Delta\Delta P_{so} - 7.073 \cdot \Delta\Delta P_{ip} + 69.851 \cdot \Delta\Delta P_{tq} + 1.851$ $N = 155169 \quad Rc = 0.89 \quad U = 0.21 \quad p < 0.01$
Polarity (Pol)	$\lambda(\varepsilon_{ij})_{new} = -4.846 \cdot \lambda(\varepsilon_{ij})_{ref} - 708.845 \cdot \Delta Pol_{seq} + 37.565 \cdot \Delta\Delta pol_{ho} + 206.803 \cdot \Delta\Delta Pol_{so} - 204.545 \cdot \Delta\Delta Pol_{ip} + 661.274 \cdot \Delta\Delta Pol_{tq} + 2.084$ $N = 155169 \quad Rc = 0.92 \quad U = 0.16 \quad p < 0.01$

3. Materials and Methods

The same database recently utilized by González-Díaz *et al*. (2) was used in the present study. The calculation of the molecular descriptors was implemented in the in-house program MARCH-INSIDE (4), 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) (5). 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 (α), partition coefficient (P), and polarity (Pol). The construction of this type of models has been explained in detail before (2); therefore, only the general equation is presented:

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

Here, $\lambda(\varepsilon_{ij})_{new}$ 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 $\lambda(\varepsilon_{qr})_{ref}$ 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.

An LDA was carried out using the STATISTICA 6.0 software (6). 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 (7).

4. Conclusions

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 an interesting 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 study was supported by grants AGL2011-30563-C03 (Ministerio de Ciencia e Innovación, Spain) and GPC2014/058 (Xunta de Galicia, Spain).

Author Contributions

S.V.P. conceived and designed the study, analysed and interpreted the data and wrote the paper. All authors discussed the results and implications and commented on the manuscript at all stages.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Vita, R.; Zarebski, L.; Greenbaum, J.A.; 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), D854-862.
2. González-Díaz, H.; Pérez-Montoto, L.G.; Ubeira, F.M. 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.
3. Todeschini, R.; Consonni, V. 2008. Handbook of Molecular Descriptors. Wiley-VCH, Weinheim.
4. 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.

5. González-Díaz, H.; Arrasate, S.; Sotomayor, N.; Lete, E.; Munteanu, C.R.; Pazos, A.; Besada-Porto, L.; Ruso, J.M. 2013b. MIANN models in medicinal, physical and organic chemistry. Curr. Top. in Med. Chem. 13, 619-641.
6. StatSoft.Inc. 2002. STATISTICA (data analysis software system), version 6.0. www.statsoft.com.
7. Hill, T.; Lewicki, P. 2006. STATISTICS: Methods and Applications: A Comprehensive Reference for Science, Industry and Data Mining. StatSoft, Tulsa.