Journal of Molecular Graphics and Modelling 71 (2017) 88–95

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Topical Perspectives

Comprehensive comparison of twenty structural characterization scales applied as QSAM of antimicrobial dodecapeptides derived from Bac2A against *P. aeruginosa*

Yu Qian [1], Ya'nan Liang [1], Wanqian Liu, Guizhao Liang *

Key Laboratory of Biorheological Science and Technology, Ministry of Education, School of Bioengineering, Chongqing University, Chongqing 400044, PR China

ARTICLE INFO

Article history:
Received 25 December 2015
Received in revised form 2 November 2016
Accepted 6 November 2016
Available online 9 November 2016

Keywords:
Antimicrobial peptides (AMPs)
Quantitative sequence-activity modeling (QSAM)
Structural characterization
Factor analysis scale of generalized amino acid information (FASGAI)

ABSTRACT

Informative numerical characterizations of amino acid residues are essential for quantitative sequence-activity modeling (QSAM). To date, a variety of structural characterization methods based on local amino acids have been proposed. However, limited detailed reports are available using same datasets and modeling methods to compare the ability to characterize structures of amino acids. Here, we evaluate the characterization capability of 20 descriptor sets on a set of antimicrobial peptides (AMPs) derived from Bac2A against *P. aeruginosa*. Results display the models by FASGAI, z-scales, VHSE, DPPS, HESH and ProtFP descriptors present qualified predictive capability. Moreover, the structural characterization of the studied AMPs should involve the hydrophobic, bulky and electronic properties of amino acids; besides, the secondary structural information should not be ignored. In parallel, the FASGAI-based model exhibits a more robust prediction than other models, and reasonably describe the structure-activity relationship of the studied dodecapeptides, which is in line with the reported experimental observations. This work provides references for methods of structural characterization as applied in QSAM of AMPs against *P. aeruginosa*.

1. Introduction

Quantitative structure-activity relationship (QSAR) can reflect the influence of molecular properties of on their activities [1–5]. QSAR for peptides is termed as quantitative sequence-activity modeling (QSAM), which has been successfully applied in a number of peptides, including bitter tasting dipeptides [6], tachykinins [7], ACE inhibitors [8], etc. Similar to QSAR, the first step of QSAM is to map the structural characteristics of target sequences into numerical variables [9]. To date, a number of descriptors, such as FASGAI [10], z-scales [11], NNAAIndex [6], ISA-ECI [12], MS-WHIM [13], SZOTT [14], ST-scales [15], T-scales [16], VHSE [17], VSTV [18], GRID [19], DPPS [20], BLOSUM [21], HESH [22], Lin's scales [23], ProtFP(8) [24], QTMS (ADFQ) [25], SVRG [26], SVWG [27] and VSW [28], have been proposed. At present, structural numerical characterizations of amino acids have achieved significant progress. These descriptors are generally based on 2D and 3D methods. However, limited detailed comparison of these methods is available using same datasets and modeling methods.

Antimicrobial peptides (AMPs) have the broad-spectrum killing ability not only to gram negative bacteria but also to gram positive bacteria. They can kill microorganisms rapidly and directly

Abbreviations: QSAM, Quantitative sequence-activity modeling; QSAR, Quantitative structure-activity relationship; AMPs, Antimicrobial peptides; GA, Genetic algorithm; PLS, Partial least squares; PCA, Principal component analysis; FASGAI, Factor analysis scale of generalized amino acid information; ISA, Isotropic surface area; ECI, Electronic charge index; WHIM, Weighted holistic invariant molecular; SZOTT, Scores of 0D, 1D, 2D and 3D; ST-scales, Structural topology scales; T-scales, Topological scales of amino acids; VHSE, Principal component score vector of hydrophilicity, steric, and electronic properties; VSTV, Principal component score vector of structural and topological variables; NNAAIndex, Index of natural and nonnatural amino acids; ProtFP, Protein fingerprint; DPPS, The divided physicochemical property scores; HESH, Hydrophobic, electronic, steric, and hydrogen; QTMS, Quantum topological molecular similarity; SVRG, Principal component scores vector of radial distribution function and geometrical descriptors; SVWG, Principal component scores vector of WHIM and GETWAY descriptors; GETWAY, Geometry, topology, and atom-weights assembly; VSW, vector of principal component scores; R^2, The cumulative multiple correlation coefficient; Q^2_{cv}, The cross-validated square of cumulative multiple correlation coefficient by a leave-one-out procedure; Q^2_{ext}, The multiple correlation coefficient of prediction for test set; VIP, Variable importance of projection.

* Corresponding author.
E-mail address: gzliang@cqu.edu.cn (G. Liang).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jmgm.2016.11.003
1093-3263/

with low toxicity to host. Normally, AMPs are marked as being short (with 10–50 amino acids), positively charged owing to the existence of basic amino acids, and hydrophobic with more than 30 percent hydrophobic amino acids in their sequences [29,30]. Because of their structural diversities, it is not feasible to design all the potential AMPs and confirm their activities by experimental approaches [30]. By view of that, QSAM has been an effective tool for design and screening of new AMPs [31–33].

The aim of this work was to investigate the characterization capability of 20 sets of amino acid descriptors (FASGAI, z-scales, NNAAIndex, ISA-ECI, MS-WHIM, SZOTT, ST-scales, T-scales, VHSE, VSTV, GRID, DPPS, BLOSUM, HESH, Lin's scales, ProtFP, QTMS, SVRG, SVWG and VSW), on a total of 196 AMPs derived from Bac2A against *P. aeruginosa*. Through comparison, the FASGAI-based model exhibited high prediction results relative to other models. Moreover, the structure-activity relationship of AMPs represented by the model were in line with the previous experimental observations. These results would provide methodology references for QSAM of AMPs against *P. aeruginosa*.

2. Principles and methods

2.1. Dataset

The dataset used for QSAM of AMPs contained a total of 196 dodecapeptides (Table S1) [34]. The antimicrobial activity was expressed as the relative EC_{50} of the measured EC_{50} of *P. aeruginosa* to that of Bac34 (VRLRIRVAVIRA). Here, the EC_{50} values were converted to negative logarithm (pEC_{50}) scales as dependent variables. The training set including 157 AMPs (No.1- No.157 in Table S1) was used to establish the model and the test set including 39 AMPs (No.158- No.196 in Table S1) was used to validate the external predictive power of the model.

2.2. Structural characterization

Structures of these dodecapeptides were characterized by 20 different descriptor sets, including FASGAI [10], z-scales [11], NNAAIndex [6], ISA-ECI [12], MS-WHIM [13], SZOTT [14], ST-scales [15], T-scales [16], VHSE [17], VSTV [18], GRID [19], DPPS [20], BLOSUM [21], HESH [22], Lin's scales [23], ProtFP(8) [24], QTMS (ADFQ) [25], SVRG [26], SVWG [27], and VSW [28] (Table S2-S21). All the descriptors used in this work are local descriptors, e.g. 2D or 3D structural and property parameters, orthogonal binary codes, and principal properties, etc., to sequentially characterize sequences and structures of peptides based on their amino acid compositions and positions. Each amino acid was separately represented by each descriptor set, thus the structural features of any peptide were characterized by simply constructing a $12 \times n$ variable matrix, where 12 is the number of the amino acids in the studied dodecapeptides, and n is the number of each descriptor sets mentioned above. For instance, FASGAI descriptors involved 6factors, thus a dodecapeptide was characterized by $12 \times 6 = 72$ variables. All the script for characterization of 20 descriptor sets was compiled by C^{++} language.

2.3. Variable selection

The variables related to the activities of AMPs [35,36] were selected by genetic algorithm-partial least squares (GA-PLS) [37]. In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and the internal predictive ability, respectively. The fitness of each chromosome was evaluated by the internal predictive ability of the PLS model derived from a binary bit pattern. The internal predictive performance of the model was expressed in terms of a cross-validated square of cumulative multiple correlation coefficient (Q^2_{cv}) by a leave-one-out procedure as follows:

$$Q^2_{cv} = 1 - \sum_{i=1}^{training} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{training} (y_i - \bar{y}_{tr})^2 \quad (1)$$

where y_i and $\hat{y}i$ represent the observed value and the predicted value of the dependent variable, respectively; $\bar{y}_{tr}$ indicates the mean observed value of the dependent variable for the training set; the summations run over all samples in the training set.

The empirical parameters in the GA-PLS method were set as follows: the number of populations was 200, the maximum number of generations was 200, the generation gap was 0.8, the crossover frequency was 0.5, and the mutation rate was 0.005. The GA-PLS program was compiled by Matlab 6.1.

2.4. Partial least squares modeling

The PLS regression, which has the desirable property that the precision of the model parameters was improved with the increasing number of relevant variables and observations [38–40], was used to correlate the variables with the relative EC_{50} values. The PLS algorithm consists of outer relations (X and Y block individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \quad (2)$$

$$x_{im} = \sum_{a=1}^{A} u_{ia} c_{am} + g_{im} \quad (3)$$

The t and u latent variables are correlated through the inner relation given below, which leads to the estimation of the y from the x.

$$\hat{U} = bt \quad (4)$$

2.5. Model validation and evaluation

The predictive capability of the models were evaluated by the coefficient of determination the cumulative multiple correlation coefficients by regression modeling (R^2, Eq. (5)), leave-one-out internal validation (Q^2_{cv}, Eq. (1)), and external validation (Q^2_{ext}, Eq. (6)). Generally, the QSAR models proved to be qualified as $Q^2_{cv} > 0.500$ and $R^2 > 0.600$ [41].

$$R^2 = 1 - \sum_{i=1}^{training} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{training} (y_i - \bar{y}_{tr})^2 \quad (5)$$

$$Q^2_{ext} = 1 - \sum_{i=1}^{test} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{test} (y_i - \bar{y}_{tr})^2 \quad (6)$$

where y_i and $\hat{y}i$ represent the observed value and the predicted value of the dependent variable; $\bar{y}_{tr}$ indicates the mean observed value of the dependent variable for the training set; the summations run over all samples in the training set for R^2 and in the test set for Q^2_{ext}.

3. Results and discussion

3.1. Six models by FASGAI, z-scales, VHSE, DPPS, HESH and ProtFP(8) descriptors exhibit qualified prediction

Just as QSAR modeling, structural characterization is the core of QSAM of peptides. Methods of structural characterization in QSAM