BioSystems 169–170 (2018) 5–12

Contents lists available at ScienceDirect

BioSystems

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

Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES

Alla P. Toropova[a,*], Andrey A. Toropov[a], Emilio Benfenati[a], Danuta Leszczynska[b], Jerzy Leszczynski[c]

[a] Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
[b] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, USA
[c] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

ARTICLE INFO

Keywords:
Peptide
Antimicrobial activity
quasi-SMILES
Bioinformatics
Monte carlo method
CORAL software

ABSTRACT

The purpose of this study was the estimation of ability of the so-called optimal descriptors calculated to be a tool to predict the antimicrobial activity of large pool of peptides. Traditional simplified molecular input-line entry system (SMILES) is an efficient tool to represent the molecular structure of different compounds. Quasi-SMILES represents an extension of traditional SMILES. This approach provides the possibility to involve different eclectic conditions related to analyzed endpoint in the modelling process. In addition, the quasi-SMILES can be used to represent structure of peptides via abbreviations of corresponding amino acids. In this study, quasi-SMILES represents sequences of amino acids in peptides that were tested as the basis to predict antimicrobial activity of 1581 peptides. Predictive potential of binary classification for antimicrobial activity for different splits is quite good when it comes to the training, invisible training, calibration, and validation sets. For the external validation sets, the statistical criteria are ranged: (i) sensitivity 0.82–097; (ii) specificity 0.88–0.99; (iii) accuracy 0.87–0.98; and (iv) Matthews correlation coefficient 0.73–0.97. The suggested optimal descriptors calculated with data on composition of amino acids in peptides can be a tool to predict antimicrobial activity of peptides.

1. Introduction

Microorganisms may cause considerable problems for human health and the agricultural industry (Porto et al., 2012). Antimicrobial peptides offer an attractive alternative to traditional drugs (Gabere and Noble, 2017; Speck-Planche et al., 2012; Yousefinejad et al., 2012). In addition, the antimicrobial peptides are an important component of cosmetic industry (Vandebriel and Loveren, 2010). Thus, the mathematical modelling of antimicrobial activity of peptides is a very attractive way to solve problems of human health and agribusiness. Hence, it is unsurprisingly, that databases for antimicrobial activity of various peptides together with different algorithms for prediction of activity of peptides untested in biochemical experiments were suggested. Widely used databases on antimicrobial peptides are DBAASP (Pirtskhalava et al., 2016); CS-AMPPred (Porto et al., 2012); CAMPR3 (Waghu et al., 2016); BACTIBASE (Hammami et al., 2007). The basis of the majority of algorithms, which are aimed to build up predictive models of activity of peptides, are physicochemical and biochemical parameters of amino acids (Yount and Yeaman, 2004; Speck-Planche et al., 2016). The physicochemical data usually used as descriptors to develop models for antimicrobial activity of peptides are polarity, electrostatics charges, 3D geometry, as well as descriptors of quantum mechanics (Pirtskhalava et al., 2016). The above mentioned parameters are usually involved in algorithms of partial least squares (PLS) (Jenssen et al., 2008); artificial neural networks (ANN) (Torrent et al., 2011); random forest (RF) (Breiman, 2001); super vector machine (SVM) (Webb-Robertson, 2009); Nearest Neighbor Algorithm (Wang et al., 2011); Incremental Feature Selection (Gabere and Noble, 2017), and others.

The CORAL software is a conceptual alternative of the above mentioned approaches (Toropova and Toropov, 2017a,b). The software has been used to build up predictive models for (i) organic compounds (Toropov et al., 2017a,b); (ii) nanomaterials (Toropova and Toropov, 2017b); and peptides (Toropova et al., 2015). The CORAL software was developed as a tool to build up quantitative structure – property/activity relationships (QSPRs/QSARs) for traditional organic compounds using simplified molecular input-line entry systems (SMILES) (Weininger, 1988). However, further application of the software

[*] Corresponding author.
E-mail address: alla.toropova@marionegri.it (A.P. Toropova).

https://doi.org/10.1016/j.biosystems.2018.05.003
Received 7 March 2018; Received in revised form 10 May 2018; Accepted 14 May 2018
Available online 22 May 2018
0303-2647/

Table 1
The measure (%) of non-identity of splits into the training, calibration, and validation sets examined in this work.

$$Identity\,(\%) = \frac{N_{i,j}}{0.5{*}(N_i + N_j)}{*}100$$

N_i is the number of substances which are distributed into the set for i-th split;
N_j is the number of substances which are distributed into the set for j-th split.

split	Set	Split 1	Split 2	Split 3	Split 4
1	Training	100*	28.3	22.4	25.8
	Invisible training	100	23.5	24.1	23.8
	Calibration	100	27.3	20.5	23.3
	Validation	100	25.9	22.3	22.3
2	Training		100	24.9	26.5
	Invisible training		100	22.9	28.4
	Calibration		100	28.1	19.0
	Validation		100	27.7	23.8
3	Training			100	25.2
	Invisible training			100	21.6
	Calibration			100	21.7
	Validation			100	30.1
4	Training				100
	Invisible training				100
	Calibration				100
	Validation				100

$$Identity\,(\%) = \frac{N_{i,j}}{0.5{*}(N_i + N_j)}{*}100$$

$N_{i,j}$ is the number of substances which are distributed into the same set for both i-th split and j-th split (set = training, invisible training, calibration, and validation).
N_i is the number of substances which are distributed into the set for i-th split;
N_j is the number of substances which are distributed into the set for j-th split.
Shaded values indicates the diagonal elements, in other words, any split is absolutely identical to itself.

(Toropova et al., 2012; Veselinović et al., 2015; Toropov et al., 2012; Toropova et al., 2015) has shown the ability of the CORAL approach to be a tool of building up predictive models based on all available eclectic information represented by so-called quasi-SMILES (Toropov and Toropova, 2015). Traditional SMILES is a sequence of symbols, which are a representation of the molecular structure. The quasi-SMILES is also a sequence of symbols. However, these symbols are a representation of not only molecular structure, but also of “all available eclectic data” (Toropov and Toropova, 2015). The sequence of amino acids is a version of the quasi-SMILES for the case of building up a predictive model for behavior of antimicrobial peptides.

The aim of this study is building up classification models of antibacterial activity of peptides (i.e. active – inactive) by using the optimal descriptors calculated with sequences of amino acids (which are represented by the one-symbol abbreviations of amino acids).

2. Method

2.1. Data

The experimental data on the antimicrobial activity (1 means active and −1 means inactive) of a large set of peptides was taken from the literature (Speck-Planche et al., 2016). These peptides (n = 1581) were randomly distributed four times into the training (≈25%), invisible training (≈25%), calibration (≈25%), and validation sets (≈25%).

It is to be noted, that fourth split was selected as the best from one hundred random splits that obey the above rules of distributions. This was done in order to confirm the hypothesis, “QSAR is a random event”, i.e. successful and unsuccessful distributions exist (Toropov et al., 2013). Table 1 shows that splits considered in our study are not identical.

2.2. Building up predictive model

In order to build up a classification of peptides into two classes (i) active (+1); and (ii) inactive (-1) so-called semi-correlations described in the literature (Toropova and Toropov, 2017a) have been used. Fig. 1 contains a graphical representation of the traditional correlation and the semi-correlation. The semi-correlation is a special case of the traditional correlations, where dots in coordinates x, y (i.e. observed, predicted) are localized along two parallel lines (Fig. 1).

Fig. 2 contains the general scheme of building up the categorical model for activity of peptides. Peptides are interpreted as sequences of amino acids (represented by the one-symbol abbreviation). These sequences are similar to SMILES, which were used as the basis to build up QSPR/QSAR models for endpoints of traditional molecules (Toropov et al., 2017a,b). The correlation weights for each one-symbol codes are calculated by the Monte Carlo method. The numerical data on the correlation weights of amino acids are results of the optimization with target function defined as the following:

$$TF = R + R' - 0.1 * ABS\,(R - R') \quad (1)$$

where R and R’ are correlation coefficients between the *DCW(T*,N*)* and categorical values antimicrobial activity of peptides (1 for active peptides; and −1 for inactive peptides), for the training and invisible training sets, respectively. The *DCW(T*,N*)* is the optimal descriptor calculated as the following:

$$DCW\,(T^*, N^*) = \Sigma\, CW\,(A_k) \quad (2)$$

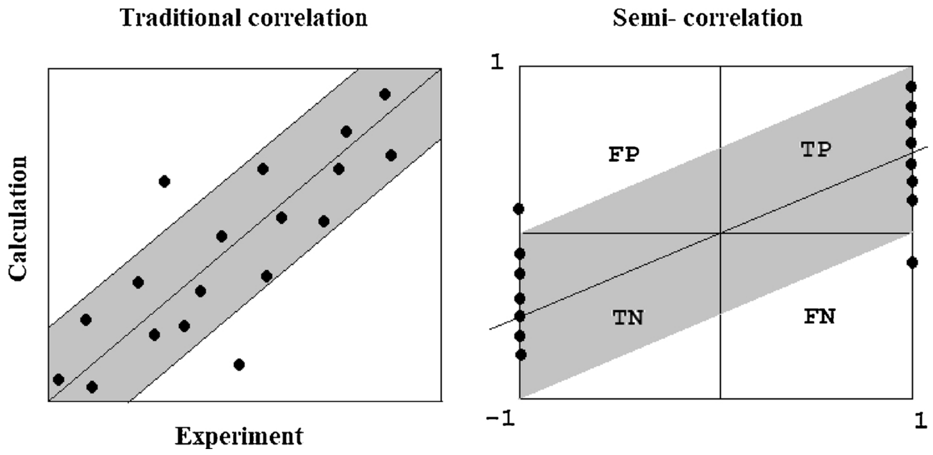

Fig. 1. The comparison of generalized traditional correlation and generalized semi-correlation.