Chemometrics and Intelligent Laboratory Systems 127 (2013) 35–42

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Quantitative structure–activity relationship study of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors by genetic algorithm feature selection and support vector regression

Yong Cong [a], Bing-ke Li [a], Xue-gang Yang [a], Ying Xue [a,b,*], Yu-zong Chen [b,c], Yi Zeng [d]

[a] Key Lab of Green Chemistry and Technology in Ministry of Education, College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China
[b] State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610064, People's Republic of China
[c] Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore
[d] Key Laboratory of Advanced Scientific Computation of Sichuan Province, Xihua University, Chengdu 610039, People's Republic of China

ARTICLE INFO

Article history:
Received 31 December 2012
Received in revised form 19 May 2013
Accepted 25 May 2013
Available online 1 June 2013

Keywords:
QSAR
Neuraminidase inhibitors
Support Vector Machine (SVM)
Partial Least Square (PLS)
Genetic algorithm (GA)

ABSTRACT

The quantitative structure–activity relationship (QSAR) for the prediction of the activity of two different scaffolds of 108 influenza neuraminidase A/PR/8/34 (H1N1) inhibitors was investigated. A feature selection method, which combines Genetic Algorithm with Partial Least Square (GA–PLS), was applied to select proper descriptor subset for QSAR modeling in a linear model. Then Genetic Algorithm-Support Vector Machine coupled approach (GA–SVM) was first used to build the nonlinear models with nine GA–PLS selected descriptors. With the SVM regression model, the corresponding correlation coefficients (R) of 0.9189 for the training set, 0.9415 for the testing set and 0.9254 for the whole data set were achieved respectively. The two proposed models gained satisfactory prediction results and can be extended to other QSAR studies.

1. Introduction

Influenza is a major respiratory infection associated with significant morbidity in the general population and mortality in elderly and high-risk patients [1]. Understanding the molecular and cell biology of influenza is highly important for suppressing this widely spread disease. Influenza is an RNA virus that contains two major surface glycoproteins, neuraminidase (NA) and hemagglutinin. NA can cleave the a-ketosidic connections of sialic acid and nearby sugar residues. The removal of sialic acid lowers the viscosity of the virus, thus permitting the entry of the virus into epithelial cells. NA also destroys hemagglutinin on the virus surface allowing the emergence of progeny virus units from infected cells [2]. Inhibition of the viral NA has been shown to be effective in reducing viral replication and has been validated as a strategy for treating influenza virus infection [3].

In recent years, there have been significant efforts in the design of inhibitors of NA. Two NA inhibitors, Zanamivir (from GlaxoSmithKline and Biota) and oseltamivir (from Hoffman La Roche and Gilead Sciences), have been approved by the Food and Drug Administration (FDA) in the USA for the treatment and prevention of influenza [4–6]. Though Zanamivir displays an excellent antiviral activity when administered intranasally, it is less effective when delivered systemically. It has very low oral bioavailability and is rapidly eliminated by renal excretion [7]. Oseltamivir is orally active, but it has been reported to cause vomiting and nausea. Current therapeutic measures have only provided limited control of influenza. The developments of vaccines for influenza virus are of restricted usefulness as they are not susceptible to the high mutability of the virus. Effective chemotherapy for influenza virus is also limited due to newly discovered drug resistance in mutant strains. Therefore, there is still a great need to design and identify new agents for the chemotherapy of influenza virus infection and formulate effective drugs for systemic administration.

Structure-based drug design has facilitated the discovery of anti-influenza drug. Crystal structure of influenza A neuraminidase was solved in the year 1983 [8] and its complex with its natural substrate sialic acid was reported in 1992 [9]. Based on the structure–activity relationship (SAR) study, Kim and colleagues synthesized series of novel carbocyclic influenza neuraminidase inhibitors and evaluated these small compounds for influenza A/PR/8/34 (H1N1) and B/Lee/40 neuraminidase inhibitory activity. The information of original NA crystal structure has proved valuable in their discovery and development of zanamivir and oseltamivir. Later, several flavonoids and biflavonoids have been reported as showing anti-influenza virus activity by inhibiting NAs [10–14]. Flavonoids (substituted phenyl-benzopyranes) are low molecular weight compounds that are widespread in the plant kingdom,

* Corresponding author at: Key Lab of Green Chemistry and Technology in Ministry of Education, College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China. Tel.: +86 28 85418330.
E-mail address: yxue@scu.edu.cn (Y. Xue).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.05.012

are relatively easy to synthesize and show several interesting biological activities in enzymatic systems.

Structure–activity relationship (SAR) has been successfully applied in the discovery of NA inhibitors. Among these methods, the quantitative structure–activity relationship (QSAR) has been demonstrated to be an effective computational tool in understanding the correlation between the structures of molecules and their activities. Its advantage over other methods lies in the fact that QSAR model can directly indicate the structural factors which play an important role in the determination of the activity and the descriptors used to build the models can be calculated from the structure alone which are independent on any experimental properties. Verma and Hansch [15] have developed 17 QSARs for different sets of compounds to understand chemical–biological interactions governing their activities toward influenza neuraminidase, among which four models (Nos. 2, 3, 4, and 6) studied about carbocyclic derivatives for their influences on H1N1 inhibiting activities. The carbocyclic derivative dataset used is from Kim's research group. Du and colleagues performed the inhibitory activity assays of 35 flavonoid compounds against influenza A/PR/8/34 (H1N1) neuraminidase [16], and then constructed the three-dimensional QSAR model with these 35 flavonoids to explore the structural requisites for NA inhibitory activity. A meaningful QSAR model with R^2 of 0.5968, Q^2 of 0.6457, and Pearson-R value of 0.8679, was obtained [17]. In addition to those work, the other 2D and 3D QSAR studies on benzoic acids, cyclopentanes, isoquinolines, pyrrolidines and miscellaneous compounds for their inhibitory activities against NAs were also reported [18–22]. However, most of these QSARs have been developed and tested by using no more than ~45 compounds and were always confined to the same scaffold compounds. Study about structure–activity relationship based on more compounds with greater diversity in their structures and activities should be helpful in providing reference information for receptor-based and ligand-based anti-influenza drug design.

Previous chemoinformatic methods, including multiple linear regression (MLR), heuristic method (HM), principal component regression (PCR) and partial least squares (PLS) were used in linear QSAR modeling. Recently, the Support Vector Machine (SVM) and the Radial Basis Function network (RBF) have been consistently shown to have excellent performance for predicting various pharmacodynamic, pharmacokinetic and toxicological properties of compounds of diverse structures due to their flexibility in modeling nonlinear cases [18,23]. Therefore, it is of interest to test the usefulness and performance of SVM as potential tool for the prediction of NA inhibitors.

In the present study, we will perform a QSAR study based on the collected 108 compounds with carbocyclic and flavonoid scaffolds, which have clear inhibitory activity against influenza virus strain A/PR/8/34 (H1N1) reported so far in the literature [24–31]. Two coupled approaches, in which the Genetic Algorithm is combined respectively with Partial Least Squares (GA–PLS) and the Support Vector Machine (GA–SVM), are first used to predict the activity of 108 NA inhibitors together. The GA–PLS strategy is also applied to select the proper molecular descriptors in our work. The aim of this investigation is to establish a new and reliable QSAR model for predicting the activity of NA inhibitory compounds and to explore the correlation between the structures of molecules and their NA inhibitory activities in order to provide reference information for anti-influenza drug design. Because of being built based on a larger dataset and diversity structures of NA inhibitors, the model could be more comprehensive.

2. Materials and methods

2.1. Data sets

A total of 108 influenza neuraminidase A/PR/8/34 (H1N1) inhibitors used in this work were collected from recently published papers [24–31]. The majority of the tested inhibitors are efficient H1N1 inhibiting agents showing IC_{50} values from 0.3 nM to 591,030 nM. The 2D structure of each compound was drawn by ChemDraw [32] and subsequently converted into 3D structure by Corina [33], and then followed by optimization using AM1 method. The resulted 3D structure of each compound was manually checked to ensure that the chirality of the chiral agent is properly generated and no structure of compounds was duplicated. We further separated them into the training set (80 compounds) and test set (28 compounds) based on their similarity and distribution in the chemical space. The chemical space is defined by the commonly used structural and chemical descriptors [34], and each compound was distributed in a special position of the chemical space. The compound similarity can be identified by using the Tanimoto coefficient sim(i,j) [35]

$$sim(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}\left(x_{dj}\right)^2 - \sum_{d=1}^{l} x_{di}x_{dj}} \quad (1)$$

where l is the number of molecular descriptors. A compound i is considered to be similar to a compound j in the data set if the corresponding sim(i,j) value is greater than a cutoff value. The cutoff values for similarity compounds are typically in the range of 0.80–0.95 [36]. A stricter cutoff value of 0.93 was used in this study. If the value of similarity between two compounds is greater than 0.93, we randomly put these two compounds into training set and testing set, respectively.

2.2. Molecular descriptors

Molecular descriptors have been routinely used for quantitative description of structural and physicochemical features of molecules in QSAR studies. In this work, a total of 189 molecular descriptors were used, this set of descriptors was manually selected from more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or irrelevant to the prediction of pharmaceutical agents [37]. These descriptors, given in Table 1, include 18 descriptors in the class of simple molecular properties (such as molecular weight and number of rotatable bonds), 27 descriptors in the class of molecular connectivity and shape (such as molecular connectivity indices and molecular kappa shape indices), 97 descriptors in the class of elector-topological state (such as electro-topological state indices), 22 descriptors in the class of quantum chemical properties (such as atomic charges and molecular dipole moment), and 25 descriptors in the class of geometrical properties (such as solvent accessible surface area and hydrophobic region). The 189 descriptors were computed from the optimized 3D structure of each compound using our own designed molecular descriptor computing program, but not all of these descriptors are essential for the QSAR modeling. In order to decrease interferes of multicollinearity before GA–PLS feature selection, we preprocessed these 189 molecular descriptor set. The procedure includes: (1) removing descriptors which have the identical value for more than 90% of the samples; (2) removing descriptors with relative standard deviation less than 0.05; (3) for each pair of descriptors with Pearson correlation coefficient [38] over 0.9, only one descriptor, which has the higher correlation with the inhibitory activity, remained. After that procedure, only 67 molecular descriptors remained and further selected using the GA–PLS feature selection method.

2.3. Feature selection with GA–PLS strategy

The GA–PLS-based coupling method applied to select the suitable features in our work was implemented in the PLS-Genetic Algorithm Toolbox [39]. This program is an optimization tool based on the GA