Chemometrics and Intelligent Laboratory Systems 135 (2014) 157–165

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Replacement based non-linear data reduction in radial basis function networks QSAR modeling

Atefe Malek-Khatabi [a], Mohsen Kompany-Zareh [a,b,*], Somayeh Gholami [a], Saeed Bagheri [a]

[a] Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran
[b] Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 3 October 2013
Received in revised form 22 March 2014
Accepted 6 April 2014
Available online 18 April 2014

Keywords:
QSAR
Radial basis function networks
Replacement method
Human immunodeficiency virus type 1
K-means clustering
Cluster analysis

ABSTRACT

A combination of radial basis function network (RBFN) and replacement method (RM) is introduced for a better description of quantitative structure activity relationship models (QSAR). RBFN–RM provides a way to choose the informative centers in order to reduce the volume of data and to increase the prediction ability, without eliminating any of the descriptors. This method was applied for predicting the activity of a series of 1-[2-hydroxyethoxymethyl]-6-(phenylthio) thymine] (HEPT) derivatives, as non-nucleoside reverse transcriptase inhibitors (NNRTIs). Prediction ability of RBFN–RM was compared to combinations of cluster analysis (CA) and *K*-means clustering with RBFN (RBFN–CA and RBFN–*K*-means). The Q^2 value for RBFN–RM, RBFN–*K*-means, and RBFN–CA was calculated as 0.9766, 0.7965, and 0.7084, respectively, which showed the merit of RBFN–RM. The method was applied on Selwood and GABA data sets, as well. To check the stability of the RM procedure, for each data set, the models were validated by using different arrangements of calibration and validation sets. Using any of the calibration and validation arrangements for HEPT, Selwood, and GABA data sets the estimated correlation values, r, for calculated versus actual activities in the validation sets were higher than 0.96.

1. Introduction

Quantitative structure activity relationship (QSAR) is based on congenericity principle which describes the physiological activity of a substance as a function of its composition and constitution [1]. The main steps involved in QSAR include data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development and finally model performance evaluation. This method develops a way for prediction of activity and describes important structural features of molecules that are responsible for variations in molecular activities [2]. Development of QSAR based computational methods, which allow one to generate relationships between chemical structure and biological activity, have effectively improved the drug design process [3].

Human immunodeficiency virus type 1 (HIV-1) is responsible for one of the greatest problems of mankind in its recent history which was called acquired immunodeficiency syndrome (AIDS) and currently extensive works are going to block its replication. The reverse transcriptase (RT) of HIV-1 is an essential enzyme to catalyze the conversion of the viral RNA into proviral DNA, so it is an attractive target for antiviral therapy of AIDS [4,5]. A large number of compounds have been synthesized to detect various active sites on this enzyme [6,7]. HEPT derivatives are the first non-nucleoside reverse transcriptase inhibitor (NNRTI) analogs shown to have potent anti-HIV activity [8]. NNRTIs are anti-HIV-1-RT specific compounds that exhibit low cytotoxicity and produce few side effects. The common feature of NNRTIs is that they consist of aromatic moieties and bind to a hydrophobic pocket near the polymerase catalytic site [9] and inhibit the ability of the enzyme to perform normal RT functions [10]. A large number of HEPT derivatives have been synthesized [11,12] and extensively studied [9]. The use of NNRTIs in combination therapy with NRTIs and with protease inhibitors is currently the best method for controlling HIV infections, but there is a further need for development of NNRTIs and design of new and more effective drugs. The main advantage of QSAR model is the independency of considered descriptors to experimental activity which means that they can be calculated from structure alone. Therefore, whether a compound was synthesized or not, its activity can be predicted from an established reliable QSAR model. There are several recent QSAR studies on different data set such as 1-[2-hydroxyethoxymethyl]-6-(phenylthio) thymine] (HEPT) derivatives which are based on MLR, PLS and neural networks [2,13–23].

Multiple linear regression (MLR) and partial least squares (PLS) are two methods that have typically been used in QSAR study of RT inhibitors [17]. Kireevand co-workers have used MLR to relate the RT inhibitory activity of 87 analogs of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)

* Corresponding author at: Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran.
E-mail address: kompanym@iasbs.ac.ir (M. Kompany-Zareh).

http://dx.doi.org/10.1016/j.chemolab.2014.04.005
0169-7439/

thymine (HEPT) [21]. Luco and Ferretti have developed a QSAR model based on MLR and PLS methods for the anti-HIV activity of large group of HEPT derivatives [17]. In another report it has been demonstrated that PLS is a powerful tool for improving the interpretability of the data and also for activity prediction [20].

There has been an increasing trend toward the use of multivariate calibration techniques such as PLS and MLR to interpret a QSAR model [17,20,21]. One drawback of these multivariate calibration approaches is the assumption of a linear relationship between the biological activity and descriptors. In accordance to that biological phenomena considered non-linear by nature, contribution of some of the parameters to RT inhibition properties can also be non-linear [24]. Applying neural networks to interpret QSAR models can be a proper solution to overcome the non-linear intrinsic mapping of biological activities [20].

Recently, artificial neural networks (ANNs) have gained great popularity in QSAR/QSPR researches due to their flexibility in modeling the non-linear problems [2,13,14,20,22,23,30,33]. These methods are particularly useful in cases where it is difficult to define an exact mathematical model for describing a specific structure–property relationship [25,26]. They have been widely used to predict physico-chemical properties based on calculated descriptors. Most of the previous works have been trained the neural networks using back-propagation learning algorithm, which has some disadvantages such as local minima, slow convergence, time-consuming non-linear iterative optimization, difficulty in explicit optimum network configuration, etc. [27]. In contrast, radial basis function network (RBFN) allows modeling of non-linear data using a simple linear regression of the non-linearly transformed data using least squares method in such a way to guarantee an optimal (unique) solution. As the least squares operation is performed on the transformed data, RBFN is not a simple linear modeling. It has advantages of small training times and global minimum of error surface during training. Furthermore, optimization of its topology and learning parameters is easy to implement and has a simple structure [28,29]. RBFN has been comprehensively applied to pattern and speech recognition, signal processing, robot control and modeling of multivariable systems. In addition, these methods have been widely applied in chemistry researches such as multivariate calibration, classification and QSPR studies [17,30–33].

In many of ANN based QSAR's models, variable selection was applied for preprocessing the data before being introduced into the network. The goal is to remove the uncertain and uninformative variables and to keep a small set of informative and predictive descriptors. In the radial basis algorithm three layers construct a network which named input, hidden layer and output layer. The optimum number of centers (neurons) in the Gaussian (RBF) function hidden layer and the appropriate spread or scaling parameter should be optimized in the construction of a neural network model. Centers in hidden layer of RBFN are the points in the data space from which distances of samples are considered and are non-linearly transformed. Therefore, center selection is an approach for size reduction of data without any descriptor elimination. It has been tried with various methods such as random subset selection, *K*-means clustering, orthogonal least squares learning algorithm, RBFN-PLS and genetic algorithm (GA) [2,13,16,22,23]. In this study center selection was performed with no descriptor elimination and all of the raw data were introduced into the network. In order to optimize the number of centers, replacement method (RM) was used. A goal in this work is to compare the performance of the models formed from clustering based selection of RBF centers to models from application of replacement to RBF centers selection.

The results show that RBFN–RM has got better prediction in comparison to most of the previously reported approaches and in some cases has a comparable result with the best of them such as GA. Although, GA is a really powerful method but it has many adjustable parameters that should be taken into account and tune of these parameters during implementation of the GA, making its execution much more laborious and complicated [34]. In other words, RM as a very simple method with prediction ability as good as GA, was introduced to select the certain set of centers for training the RBFN in a simple, rapid, high performance and reproducible way.

2. Theory

2.1. Radial basis function networks

Neural networks, natural or artificial, are systems of high number of interconnected information processing neurons. Artificial neural networks (ANNs) emulate the function of brain and are the frequently used networks in QSAR. Radial basis function network (RBFN) is a class of ANNS and its application has increased rapidly in the last few years. This is due to the particular advantages of RBFN such as better approximation capability, simpler network structure, short learning time and not get stuck in local minimum [23]. A three layer RBFN was applied in this study, which was written by the authors in MATLAB® software. This type of networks consists of an input layer, a hidden layer and an output layer. Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer. The input layer only distributes the input vectors to the hidden layer. The hidden layer of RBFN consists of a number of RBFN neurons (n_h) that each hidden layer neuron represents a radial basis function (RBF). The RBF in neurons is among the Gaussian function and is utilized for non-linear transformation of the input data. A Gaussian function is characterized by a center vector (c_j) and spread (r_j). By measuring the Euclidean distance between input vector (x_i) and the radial basis function center (c_j) (as a distance that named φ matrix), non-linear transformation in the hidden layer was implemented as given below:

$$h_j(x_i) = exp\left(-\left\|x_i - c_j\right\|^2 \Big/ r_j^2\right) \tag{1}$$

Where h_j, represent the output of the *j*th RBFN neuron. The operation of the output layer is linear as shown in Eq. (2).

$$y_k(x_i) = \sum_{j=1}^{n_h} w_{k_j} h_j(x_i) + b_k \tag{2}$$

Where $y_k(x_i)$ is the *k*th output neuron for the input vector x_i, w_{kj} is the weight connection between the *k*th output and *j*th hidden layer neuron and b_k is the bias.

From Eqs. (1) and (2) it can be concluded that designing RBFN models involve optimizing centers, spread, and weights. The spread of radial basis function can either be chosen the same for all the neurons, as was applied in this study, or can be supposed differently for each neuron, that the first way is common in most of researches [2,14,22]. The adjustment of the connection weights between hidden and output layer was performed by using a least squares solution as the simplest regression approach after center and spread selection of the radial basis functions. Also with optimization of centers, spread and weights the bias in prediction of activities will be optimized to its minimum possible value. The performance of the RBFN was evaluated in terms of Q^2 parameter which was calculated by Eq. (3) [35]:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 \Big/ n_{EXT}}{\sum_{j=1}^{n_{TR}} \left(y_j - \bar{y}_{TR}\right)^2 \Big/ n_{TR}} \tag{3}$$

Where n_{EXT} and n_{TR} represent the number of external set (test set) and training set samples, respectively. The y_i, $\hat{y}_i$, y_j and $\bar{y}_{TR}$ are the