Journal of Computational Science 11 (2015) 112–120

Contents lists available at ScienceDirect

Journal of Computational Science

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

An in silico screening study and design of potent cognition agents

Hossein Tavakoli[a], Jahan B. Ghasemi[a,b,*]

[a] Drug Design in Silico Lab., Faculty of Chemistry, K N Toosi University of Technology, Tehran, Iran
[b] Faculty of Chemistry, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 7 April 2015
Received in revised form 15 July 2015
Accepted 16 September 2015
Available online 26 October 2015

Keywords:
Quantitative structure–activity relationship (QSAR)
Enhanced replacement method
In silico screening
Absorption, distribution, metabolism, excretion and toxicity (ADMET) studies
Design of histamine H_3 antagonists

ABSTRACT

The binding affinities of human histamine H_3 antagonists were predicted using an enhancement replacement method – partial least squares (ERM-PLS) model and the results were compared with those of genetic algorithm (GA)-PLS and stepwise linear regression (SW)-PLS models. Based on this accurate, robust and reliable ERM-PLS model, an in silico screening study was performed, which resulted in some new and potent cognition agents, some of which show equal or even better affinity and ADMET properties than the previous structures. The effects of structural moieties on the activity of the compounds were discussed.

1. Introduction

Histamine is one of the most important biogenic amine neurotransmitters, involved in the modulation of various functions in the central nervous system (CNS) [1]. Histamine H_3 receptors (HH3Rs) are primarily localized in the CNS neurons, and lesser in the specific peripheral tissues. On the histaminergic neurons, HH3R, as an autoreceptor, it controls the release of histamine, while on the non-histaminergic neurons, it regulates the release of multiple important neurotransmitters such as acetylcholine [2], norepinephrine [3], dopamine [4], and serotonin [5] (Scheme 1). The antagonists of HH3Rs enhance the concentration of the cerebral histamine. Therefore, they play very essential roles in the treatment of the cognition disorders (attention deficit and hyperactivity disorder, Alzheimer's disease, and schizophrenia); sleep diseases (hypersomnia and narcolepsy); and energy homeostasis (obesity) [6].

In human in vivo studies, accurate determination of the binding affinities (pK_i) of drug-like H_3 antagonists is particularly a main interest. In the absence of the experimental data, especially in too complex biological systems, a predictive quantitative structure activity relationship (QSAR) is a good remedy [7–12]. In these systems, selecting the best combination of variables that resulted in a predictive QSAR model is crucial. Furthermore, in spite of developing of various sophisticated machine learning methods during the last two decades [13,14], partial least squares (PLS) [15], as a strong factor-based method, is the main implementation learner in the molecular modeling softwares for relating chemical and biological spaces together.

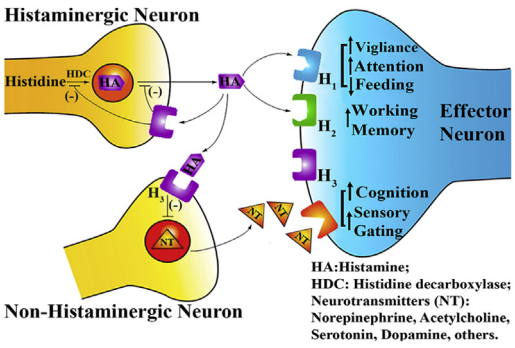
In this study, we compared the performance of an unsystematic variable selection algorithm, genetic algorithm (GA) [16], with those of two systematic search algorithms, enhanced replacement method (ERM) [17,18] and stepwise (SW) variable selection [19]. To develop a predictive model, a series of elite descriptors that extracted by these algorithms were used as the inputs of the PLS method. The validity of the models was examined according to the organisation for economic co-operation and development (OECD) principles [20,21]. Finally, to explore novel biologically active structures, a comprehensive in silico screening study was established by introducing different substituents to the common structure of compounds or the most active compound in the series. Because a drug-like compound should bear excellent ADMET properties, as well as, proper biological activity, the out of range compounds were filtered out from the final candidate set.

2. Data set and methods

According to the OECD guideline, a standard QSAR model should be associated with a defined end point, an unambiguous

* Corresponding author at: Drug Design in Silico Lab., Faculty of Chemistry, K N Toosi University of Technology, Tehran, Iran.
E-mail address: Jahan.ghasemi@gmail.com (J.B. Ghasemi).

http://dx.doi.org/10.1016/j.jocs.2015.09.007
1877-7503/



Scheme 1. Schematic representation of H_3 receptor function.

algorithm, a defined applicability domain, appropriate measures of goodness-of-fit, robustness, predictive power, and a mechanistic interpretation [21].

2.1. A defined end point

An ideal QSAR model must be developed based on some well-defined experimental data, which generated by an identical protocol [21]. Therefore, in this study we used a series of 74 quinoline compounds containing the histamine H3 antagonists, all of which were assessed by a standard protocol, displacement of [3H]-N-α-methyl histamine, using cloned human H_3 receptors [22]. All the biological activities are the average of three independent measurements and their standard errors are below 0.25. The logarithmic values of the biological activities (nM) were modeled as the end points.

2.2. Optimization of the molecules and molecular descriptors generation

The structures were sketched in the ChemOffice Ultra version 11.0 [23] and optimized geometrically with the molecular mechanics force field (MM+) in the hyperchem pro 8.05 package [24]. The semi-empirical quantum chemical calculations were also performed according to the AM1 (Austin Model 1) to a gradient norm limit of 0.01 kcal $Å^{-1}$. Various molecular descriptors were generated by the Dragon software [25], and constant or near constant variables and some of which had a pair correlation value greater than 0.97 were removed. The remaining 613 descriptors were used for modeling and further analysis. The data were auto-scaled prior to any analysis to give equal importance to all the variables.

3. Theatrical backgrounds

3.1. Variable selection and regression algorithms

3.1.1. Enhancement replacement method

ERM is an improved version of the effective variable selection algorithm, replacement method (RM) [14,15]. In each step, RM tries to minimize the relative errors of the coefficients of a least-squares model by systematically replacing the variable with the greatest standard deviation in its coefficient with all of which were not selected previously.

To minimize the error of model in the modified replacement method (MRM), the variable with the largest error is replaced, even if that replacement is not accompanied by a smaller error. ERM contains the sequence of RM-MRM-RM algorithms [17].

3.1.2. Genetic algorithm

The genetic algorithm was inspired by the natural selection and evolution processes. It uses selection, cross-over and mutation operations to find the best combination of the variables by minimizing the fitness function [16].

3.1.3. Stepwise multiple linear regression

The SW algorithm tries to decrease the total error of model using a statistical *F*-test and systematically adding or removing variables to/from the model based on their statistical significance in the multi-linear regression [19].

3.1.4. Partial least squares regression

The PLS method attempts to find those latent variables (LVs), which not only capture the greatest amount of variance in **X**, but also maximize the correlation with **y**; i.e. PLS attempts to maximize the covariance between **X** and **y** [15].

The concept of variable importance in the projection (VIP) is computed based on the contribution of a variable to the variance captured by the PLS dimensions. A variable with VIP score greater than 1.0 is considered as important variable, while an unimportant variables has a VIP score less than 0.8 [26].

3.2. Model validation (goodness-of-fit, robustness, predictive ability) and models comparison

Validation is one of the main aspects in QSAR studies. In this study, the cross-validation procedure, validation through a test set, Tropsha criteria and the y randomization test were used to examine the validity of the constructed models [27,28].

Akaike criterion (AIC) is a proper metric for evaluating the quality of a model. AIC is given by:

$$\mathrm{AIC} = \left(\sum_{i=1}^{n} \mathbf{res}_i^2\right) \cdot \frac{n+d}{(n-d)^2} \tag{3}$$

where n refers to the number of training set samples and $\mathbf{res}_i$ is the vector of the differences between the experimental and predicted biological activities. The smallest AIC value is the most useful model [29].

4. Results and discussion

4.1. Comparison of the QSAR models

The Kennard–Stone algorithm [30] was used to divide the data into two subsets of 60 and 14 compounds as the training and test sets respectively. Fig. 1 shows the training and test samples in the principle components space.

In an ERM-PLS model, the number of variables that is selected by ERM and participated in the PLS should be optimized. To address this, the number of variables that maximizes simultaneously the r^2 values of the training, LOO-CV and test sets of the PLS model was chosen as the optimum number of the variables. According to

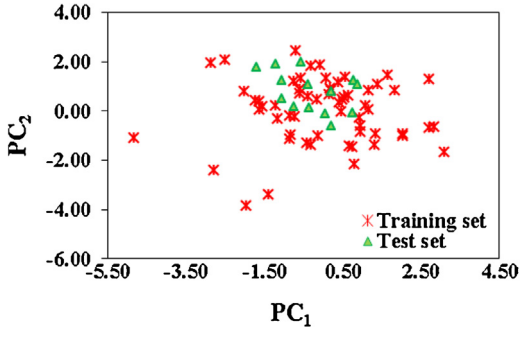


Fig. 1. The data are uniformly split to two subsets by Kennard–Stone algorithm.