Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 41 (2006) 1339–1346
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://france.elsevier.com/direct/ejmech

Short communication

Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors

Sumit Deswal, Nilanjan Roy*

Pharmacoinformatics division National Institute of Pharmaceutical Education and Research, Sector 67, Phase X, 160062 S.A.S. Nagar, Punjab, India

Revised and accepted 3 July 2006
Available online 01 August 2006

Abstract

A quantitative structure activity relationship (QSAR) analysis has been performed on a data set of 42 aryl heterocycle-based thrombin inhibitors. Several types of descriptors including topological, spatial, thermodynamic, information content and E-state indices were used to derive a quantitative relationship between the anti thrombin activity and structural properties. Genetic algorithm based genetic function approximation method of variable selection was used to generate the model. Best model was developed when number of descriptors in the equation was set to five. Highly statistically significant model was obtained with atom type logP descriptors, logP and Shadow_YZ. The model is not only able to predict the activity of new compounds but also explained the important regions in the molecules in a quantitative manner.

Keywords: QSAR; Thrombin inhibitors; Genetic function approximation; Aryl heterocyclic

1. Introduction

Thrombin is a serine protease involved in the conversion of fibrinogen to fibrin [1]. It is one of the important drug targets for cardiovascular disorders like cardiac arrest, ischemic stroke and pulmonary infraction [2]. Older drugs like heparin and warfarin interfere with activity of many clotting factors. These agents suffer from a number of side effects like increased rate of bleeding and need for parenteral administration. In addition these drugs specially warfarin, is involved in a number of drug–drug interactions, which can be problematic in many cases. Antiplatelet drug aspirin does not have a similar thromboprophylactic efficacy to warfarin, with the degree of reduction in risk of stroke being less pronounced [3]. In view of these facts, the agents with specific thrombin inhibitor activity and improved oral bioavailability may be a good alternative for the older drugs. So there is an urgent need to design such agents.

Quantitative structure activity relationships (QSAR) are one of the most important methods in chemometrics, which give information that is useful for drug design and medicinal chemistry [3,4]. A QSAR equation is a mathematical equation that correlates the biological activity to a wide variety of physical or chemical parameters [5]. It started in 1962 when Corwin Hansch first time developed a QSAR model which correlate biological activity to Hammett constant and hydrophobicity [6]. The parameter π which is the relative hydrophobicity of a substituent, was defined by Iwasa et al. in 1964 [7]. Hansch and Fujita combined Hammett and hydrophobic constants to give linear Hansch equation, which later on resulted in the development of Hansch parabolic equation [8]. After that a number of structural, topological, thermodynamic, spatial and other type of descriptors have been developed which can be used for QSAR model generation [9]. There are many examples available in literature in which QSAR models have been used successfully for the screening of compounds for the biological activity [10–13]. Several efforts have been done in past to develop QSAR model for thrombin inhibitors [14–16]. But most of the models lack the high statistical quality. Here we have developed a QSAR model for the aryl heterocycle-based thrombin inhibitors. The behavior of QSAR model is examined with a variety of statistical parameters and the contribution of various descriptors was analyzed.

* Corresponding author.
E-mail address: nilanjanroy@niper.ac.in (N. Roy).

0223-5234/$ - see front matter © 2006 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2006.07.001

2. Experimental

2.1. Data set

The inhibitory activity of the aryl heterocycle-based thrombin antagonist was taken from literature in terms of K_i values [17]. The K_i values were converted to pK_i to get the linear relationship in the equation using following formula:

$$pK_i = -\log K_i$$

Total set of 42 compounds was divided in training and test set of 34 and eight compounds randomly. The structure and actual and predicted activity for both the training and test set compounds are shown in Tables 1–3. Because of the structural uniqueness compounds **1**–**3** and **42** could not accommodated in any table hence shown in Fig. 1.

2.2. Molecular modeling

The X-ray crystal structure of most active compound (**32**) bound with thrombin was extracted from Protein Data Bank (PDB code 1SL3). As the PDB structures do not contain hydrogen atoms, so the hydrogen atoms were attached to structure and it was further energy minimized using semi-empirical AM1 method of energy minimization included in MOPAC 6.0 [18]. Other compounds were built using this structure as template and these structures were also energy minimized using same method.

2.3. Descriptor calculation

E-state indices [19,20], electronic, information content, spatial, structural, thermodynamic and topological descriptors [21] were calculated using the Cerius^2 4.10 software package. Descriptors included in the model are listed and described in Table 4.

2.4. Regression analysis

The total number of descriptors calculated was more than 150 but some of the descriptors were rejected because they contain a value of zero for all the compounds. Further, the inter correlation of descriptors was taken in to account and highly correlated descriptors were grouped together and descriptor with highest correlation with biological activity was taken from the group. From descriptors thus remained, the selection of variables to obtain the QSAR models were carried out using genetic function approximation (GFA) method. GFA is genetics based method of variable selection, which combines Holland's genetic algorithm (GA) with Friedman's multivariate adaptive regression splines (MARS) [22,23]. The GFA method works in the following way: first of all a particular number of equations (set at 100 by default in the Cerius^2 software) are generated randomly. Then pairs of "parent" equations are chosen randomly from this set of 100 equations and "crossover" operations are performed at random. The number of crossing over was set by default at 5000. The goodness of each progeny equation is assessed by Friedman's lack of fit (LOF) score, which is given by following formula

$$\mathrm{LOF} = \mathrm{LSE}/\{1 - (c + dp)/\mathrm{m}\}^2$$

Where LSE is the least-squares error, c is the number of basis functions in the model, d is smoothing parameter, p is the number of descriptors and m is the number of observations in the training set. The smoothing parameter, which controls the scoring bias between equations of different sizes, was set at default value of 1.0 and the new term was added with a probability of 50%. Only the linear equation terms were used for model building, which is set by default in the software. The best equation out of the 100 equations was taken based on the statistical parameters such as regression coefficient, adjusted regression coefficient, regression coefficient cross validation and F-test values.

3. Results and discussion

We first determined the number of descriptors necessary and sufficient for the QSAR equation. Taking a brute force approach, we increased the number of terms in the QSAR equation one by one and evaluated the effect of addition of new term on the statistical quality of model. As the r^2 correlation coefficient can be easily increased by number of terms in the QSAR equation, so we took the cross validation correlation coefficient, q^2 as the limiting factor for number of descriptors to be used in the model. As shown in Fig. 2 the q^2 value increases till the number of descriptors in the equation reached up to 5. When number of descriptors in the equation was 6, there was a decrease in q^2 value of model. So the number of descriptors was restricted to 5. The models with increasing number of descriptors are shown in Table 5 along with the statistical parameters.

After analyzing equation with five descriptors it was found that compound **18** is an outlier. The reason for **18** being found as outlier is its very low activity. It has highest value of K_i and thus lowest value of pK_i among all the compounds. So it was removed from the training set and a new equation was generated as given under:

$$pK_i = -5.28169 + 3.32264 * \mathrm{Atype\_N\_69} + 0.04043 * \mathrm{Shadow\_YZ} + 1.86922 * \mathrm{Atype\_Cl\_89} - 0.44212 * \mathrm{LogP} + 0.94893 * \mathrm{Atype\_C\_20} \quad (1)$$

$$N = 33,\ \mathrm{LOF} = 0.114,\ r^2 = 0.959,\ r^2_{adj} = 0.951,\ \mathrm{F-test} = 125.279,\ \mathrm{LSE} = 0.055,\ r = 0.979,\ q^2 = 0.943,\ r^2_{pred}(8) = 0.504,\ r^2_{pred}(7) = 0.947$$

Where N is number of compounds in training set, LOF is lack of fit score, r^2 is squared correlation coefficient, r^2_{adj}. is square of adjusted correlation coefficient, F-test is a variance-related static which compares two models differing by one or