Journal of Saudi Chemical Society (2016) **20**, 282–290

King Saud University

Journal of Saudi Chemical Society

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm–multiple linear regressions

Abolghasem Beheshti [a], Eslam Pourbasheer [a], Mehdi Nekoei [b,*], Saadat Vahdani [c]

[a] *Center of Excellences in Electrochemistry, Faculty of Chemistry, University of Tehran, Iran*
[b] *Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran*
[c] *Department of Chemistry, Islamic Azad University, Quchan Branch, Quchan, Iran*

Received 9 November 2011; accepted 18 July 2012
Available online 12 September 2012

KEYWORDS

QSAR;
Genetic algorithm;
Multiple linear regression;
Antimalarial;
Urea derivatives

Abstract A quantitative structure–activity relationship (QSAR) was performed to analyze antimalarial activities of 68 urea derivatives using multiple linear regressions (MLR). QSAR analyses were performed on the available 68 IC_{50} oral data based on theoretical molecular descriptors. A suitable set of molecular d
escriptors were calculated to represent the molecular structures of compounds, such as constitutional, topological, geometrical, electrostatic and quantum-chemical descriptors. The important descriptors were selected with the aid of the genetic algorithm (GA) method. The obtained model was validated using leave-one-out (LOO) cross-validation; external test set and Y-randomization test. The root mean square errors (RMSE) of the training set, and the test set for GA–MLR model were calculated to be 0.314 and 0.486, the square of correlation coefficients (R^2) were obtained 0.801 and 0.803, respectively. Results showed that the predictive ability of the model was satisfactory, and it can be used for designing similar group of antimalarial compounds.

1. Introduction

Malaria is one of the most prevalent diseases of our planet, which claims millions of lives annually. The health problem caused by malaria, one of the most lethal of the parasitic diseases, is now compounded due to the emergence of strains of plasmodium, which show resistance to the known chemotherapeutic agents. The paucity of new affordable drugs has not only complicated the clinical management of malaria in endemic areas, but has also resulted in an increase in the mortality rate (Rastelli et al., 2003). This situation underscores

* Corresponding author.
E-mail addresses: nekoei_m1@yahoo.com, m_nekoei1356@yahoo.com (M. Nekoei).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jscs.2012.07.019
1319-6103

needs for urgent discovery of new antimalarial agents. The artemisinin and its derivatives or their combinations have now replaced the chloroquine (CQ) and other quinoline antimalarials, especially in the endemic areas. However lower abundance and high cost of artemisinin and related products motivate the medicinal chemists to search for new chemical pharmacophores which may prove effective as antimalarial. Although there are several experimental methods available for screening the estrogenic activity of chemicals (e.g., in vivo and in vitro assay tests), and these all have also been carried out using receptors and other biological materials of human, rat, mouse, and calf origin at least (Hill, 1972), they are costly, time-consuming, and can potentially produce toxic side products from the experimental methods used today. The efficient way to obtain a complete set of the data, without necessity of performing expensive laboratory experiments is the application of the quantitative structure–activity relationship (QSAR) techniques. The QSAR is one of the most important areas in chemometrics, and is a valuable tool that is used extensively in drug design and medicinal chemistry (Hansch et al., 1990; Manly et al., 2001; Pourbasheer et al., 2010, 2011). Chemical and biological effects are related closely to molecular physico-chemical properties, which can be calculated or predicted by their structure using various kinds of methods (Burger and Abraham, 2003). Once a reliable QSAR model is established, we can predict the activities of molecules, and know which structural features play a significant role in the biological process. The advances in QSAR studies have widened the scope of rational drug design as well as the search for the mechanisms of drug actions. Many different methodologies, such as multiple linear regression (MLR), partial least squares (PLS), heuristic method (HM), and different types of artificial neural networks (ANN), can be applied for QSAR development. Genetic algorithm (GA) has gained great popularity in QSAR research. The GA–MLR method, developed by Rogers and Hopfinger (Rogers and Hopfinger, 1994), is employed in a statistical analysis to select the relevant descriptors and to generate different QSAR models. Sensitivity analysis of QSAR models is then performed, and the best model developed can be used for predicting test set molecules that were not included in the training set molecules. Randomization tests performed on the model at various intervals of confidence levels ensure its proper validation. The main aim of the present work is to establish a new QSAR model for predicting antimalarial activity of 68 urea derivatives using GA–MLR technique.

2. Data set and methods

2.1. Data set

In this study, a data set of 68 urea derivatives was collected from the literature (Madapa et al., 2009a,b; Mishra et al., 2009). The chemical structures and antimalarial activity (IC_{50}) of these 68 molecules are presented in Table 1. The IC_{50} values were converted into its logarithmic scale $pIC_{50} = -\log (IC_{50})$, to reduce the skewness of the data set, which was then used for subsequent QSAR analysis as the response variable. It is essential to assess the predictive power of QSAR models by using a test set of molecules according to the following criteria: (1) the antimalarial activity values of the test set should span the training set several times; (2) the biological assay methods for both the training set and test set should be the same or comparable; (3) the test set should represent a balanced number of both active and inactive molecules for uniform sampling of the data set. The remaining molecules are taken as the training set in order to create an efficient QSAR model.

2.2. Descriptor calculation

All of the molecules were drawn into the HyperChem (Version 7.0 Hypercube, Alberta, Canada) software and pre-optimized using the MM+ molecular mechanics force field. Then a more precise optimization was performed with the semi-empirical AM1 method in MOPAC (Stewart, 1989). Descriptors were calculated using the CODESSA (Katritzky et al., 1994) and DRAGON software package (Todeschini et al., 2003) which include: constitutional, topological, geometrical, electrostatic, charged partial surface area, quantum-chemical, molecular orbital and thermodynamic descriptors. Before commencing with the development of the QSAR model, the correlation matrix of about 457 descriptors was calculated and highly correlated descriptors, with correlation values above 0.9, were removed. Furthermore, descriptors with constant values as well as those with poor correlation the antimalarial activity were discarded; some descriptors having zero value were also discarded. Finally, remained descriptors were considered for statistical fitting using the GA–MLR method.

2.3. Genetic algorithm for descriptor selection

Genetic algorithms (GAs) are governed by biological evolution rules (Hunger and Huttner, 1999; Aires-de-Sousa et al., 2002; Ahmad and Gromiha, 2003). The GAs, which are based on the principles of Darwinian evolution, have emerged as robust optimization and search methods (Holland, 1975). In a GA feature selection procedure, potential solutions for the problem being studied are subsets of molecular descriptors. They are represented as data structures called chromosomes, which are binary strings of length N (the total number of available features), with a zero or one in position i indicating the absence or presence of feature i in the set. The initial population of chromosomes is usually generated randomly. After that, GA runs in cycles. The fitness of each chromosome is evaluated by the fitness function. The fitness function used here was the leave-one-out, cross-validated correlation coefficient (Q^2_{LOO}). New chromosomes are then created by genetic operators such as crossovers and mutations. Crossover occurs when two parent chromosomes exchange parts of their corresponding elements. Mutations induce sporadic alterations of randomly selected chromosome elements. In each cycle, a new chromosome (feature set) is produced either by mutation or crossover on the selected parents, and it is compared with the worst member of the existing population. If the new one is better, it becomes a member of the population, and the original worst one is discarded; if not, the new one is discarded, and GA goes into next generation with the population unchanged. The genetic algorithm cycle is repeated until a satisfactory descriptor set is found or a pre-set limit of generation is reached. The GA program was written in Matlab 6.5.