Prediction of AHAS Inhibition by Sulfonylurea Herbicides Using Genetic Algorithm and Artificial Neural Network

Asiye Saidi
Department of Chemistry, Shahid Bahonar University of Kerman, 22 Bahman Boulevard, Kerman 76167, Iran
Asive_saidi@yahoo.com

Mohammad Mirzaei
Department of Chemistry, Shahid Bahonar University of Kerman, 22 Bahman Boulevard, Kerman 76167, Iran
M_mirzaei@mail.uk.ac.ir

Behnam Mohseni
Department of Chemistry, Shahid Bahonar University of Kerman, 22 Bahman Boulevard, Kerman 76167, Iran
Behchem48@yahoo.com

Abstract

Acetohydroxyacid synthase (AHAS; EC 2.2.1.6) catalyzes the first common step in branched-chain amino acid biosynthesis. This enzyme is inhibited by several chemical classes of compounds and this inhibition is the basis of action of the sulfonylurea herbicides. Here we calculate the negative logarithm inhibition constant (pKi) of 64 sulfonylurea analogs as inhibitors of pure recombinant Arabidopsis thaliana AHAS using quantitative structure–activity relationship (QSAR). Suitable set of molecular descriptors were calculated and the important descriptors were selected by genetic algorithm and stepwise multiple regression methods. These variables serve as inputs to generated neural networks. After optimization and training of the networks, they were used for the calculation of pKi for the prediction set. Comparison between obtained results showed the superiority of genetic algorithm over stepwise multiple regression method in feature-selection. For network that used the genetic algorithm for feature-selection, there are very good agreements between calculated and experimental pKi for data set. The correlation coefficient between calculated and experimental values of pKi for training and prediction set are 0.944 and 0.914, respectively.

Keywords: QSAR, genetic algorithm, artificial neural network, acetohydroxyacid synthase, sulfonylurea
Introduction

Since their discovery in the 1910s, sulfonylureas have emerged as a group of herbicides with many innovative natures. Their level of activity is unprecedented and may be up to 1-2 times that of conventional herbicides. Compared with other herbicides, sulfonylureas have much lower use range and are more rapidly degraded in soil (Penmetsa et al., 2005; Corcia et al., 2005). They have the attributes of low application rates, environmental safety, good crop selectivity and low mammalian toxicity. Following an extensive synthetic program led by Levitt and colleagues (Sauers and Levitt, 1984) the first sulfonylurea herbicide chlorsulfuron was developed. Since that time, a large number of other sulfonylurea herbicides have been identified and are now applied widely (Beyer et al., 2005).

The general features of most active compounds are an ortho-substituted aromatic ring attached to the sulfur atom, and a heterocyclic ring substituted in both meta positions and attached to the distal nitrogen atom of the sulfonylurea bridge. This heterocyclic ring is either a pyrimidine or triazine. The mode of action of sulfonylureas started to become clear when it was discovered that sulfometuron methyl is a potent inhibitor of bacterial acetohydroxyacid synthase (AHAS; EC 2.2.1.6), the enzyme that catalyzes the first common step in branched-chain amino acid biosynthesis (LaRossa and Schloss, 1984). Contemporary, Ray showed that chlorsulfuron inhibits plant AHAS (Ray, 1984). Since then, other sulfonylurea herbicides have been shown to inhibit AHAS, and it is widely accepted that inhibition of this enzyme is the mode of action of sulfonylureas as well as several other families of herbicides (Duggleby and Pang, 1990). An important property of the use of such inhibitors is that there is no AHAS counterpart in humans and other animals. Despite of the large numbers of papers on AHAS published in recent years on the interaction between herbicides and this enzyme, there are some aspects of the inhibition that remain puzzling (Chipman et al., 1990). Thus, it is important to determine which structural features of the herbicides are responsible for the enzyme inhibition. This is essential for the design on new herbicides since its properties may be predicted prior to synthesis and consequently the design may, in this way, be guided by the results of calculations. Quantitative structure activity relationship (QSAR) modeling has shown to be very effective for this purpose. This approach provides information that is useful for molecular design and medicinal chemistry (Schultz et al., 1990). The QSAR models are mathematical equations which relate chemical structure of compounds to a wide variety of their physical, chemical, biological and technological properties. If we could elucidate in detail how these properties are determined by structure, we can predict such properties simply from the molecular structure. The main task of QSAR is to obtain a reliable statistical model for the prediction of activities or properties of new chemical substances and analytical systems. Nowadays, QSAR models are rapidly developing and have been widely used by chemists for predicting different chemical and physical properties of different types of molecules. In the case of herbicides there are lots of QSAR studies (Diaz and Delgado, 1990; Wang et al., 1990; Roy and Paul, 1990). Surprisingly, considering the similarity of the compounds applied to develop the models, they involve different number and types of descriptors, complicating the physical interpretation. On the other hand, the models are based on empirically derived descriptors which limit their application to new or developing chemicals. Duggleby et al. (1990) performed comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSiA) analyses of a new family of sulfonylurea herbs. Roy and Paul (1990) performed docking and 3D QSAR analysis for sulfonlurea derivatives. They considered linear relationships but the commonly used multiple linear regression (MLR) will fail to develop an appropriate QSAR model when the nonlinear phenomenon is significant to some extent within the data investigated; therefore nonlinear modeling techniques such as artificial neural networks (ANN) were necessary to be introducing for building an accurate and reliable QSAR model. ANN has recently gained much popularity for calibrating the nonlinear relationships (Wythoff, 1990; Stefan, 1990). Therefore, ANN has become an important modeling tool for building QSAR models (Marengo et al., 1990; Fatemi and Goudarzi, 1990).

In the present paper, we represent a QSAR model for the prediction of inhibition constant of diverse sulfonylurea herbicides consist of monosubstituted, bridge modification, pyrazole, pyridine, thiazole and disubstituted sulfonylureas, using artificial neural network and genetic algorithm. As far as we are aware, this is the first QSAR study using a hybrid method to the prediction of inhibition constant of sulfonylurea herbicides.
Methods

Data set

The structures of a diverse set of sulfonylurea herbicides as well as their negative logarithm inhibition constant (pKi) reported in the literature (Wang et al., 1988). The value of pKi ranged between 2.68 and 4.84 for sulfonylurea and Chlorimuron ethyl herbicides, respectively. The molecules in data set randomly divided into three sets; training, test and prediction set which each of them consisting of 64, 12, 18 member, respectively.

Structural descriptors

To obtain a QSAR model, compounds are represented by theoretical molecular descriptors. In order to compute the structural descriptors, the structures of all herbicides were drawn with HyperChem program and were optimized using the semi-empirical quantum method AM1 of the HyperChem program (Dewar et al., 1994). After geometry optimization, HyperChem output files were used by the Dragon program as input to calculate molecular descriptors. In order to reduce redundant and non-useful information, prescreening of descriptors were carried out in the following way; (1) constant or near constant descriptors were eliminated, and (2) among those descriptors whose intercorrelations exceeded 0.9, the most suitable and interpretable ones were kept while the others were deleted. The remaining descriptors were used to generate the QSAR models. These parameters encoded different aspects of the molecular structure. Since the number of descriptors considered is large, a suitable feature selection method should be combined with a proper feature mapping technique. In the present work we have considered stepwise multiple regression and genetic algorithm as feature-selection tools and ANN was employed for feature mapping.

Genetic algorithm

Variable selection is always one of the most important steps in developing a QSAR model, which is especially important when one is required to deal with a large variable set. Genetic algorithm (GA) is a stochastic optimization method that has been inspired by evolutionary principles (Holland, 1996). The different aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores distinct regions in parameter space (Cartwright, 1997). For the moment, one of the best available tutorial on variable selection using GA published by Leardi and Gonzalez (1997). In the present paper, GA optimization method was tried following the studies of Rogers and Hopfinger (1986) and Luke (1992) with a few minor modifications. In this GA an individual of the population is represented by the string of bits that encoding the selected feature. The first step in a GA is to create a gene pool of n individuals. Each individual (chromosome) contain some descriptors that in the first generation are chosen randomly from a common list and in a way such that no two individuals can be fixed that contain exactly the same set of descriptors. The fitness of each individual in this generation is appointed by a user specified fitness function. In the next step reproduction take place, which individuals are selected probabilistically on the basis of their fitness scores and serve as parents. The selection strategy that applied in this program was random selection method. Next step is a crossover, that each of parent contributes a random selection of half of its descriptors and an offspring is constructed by combining these two halves of genetic code. Therefore, the generated offspring contains characteristics from both of its parents. Finally, this offspring is subjected to a random mutation in one of its gene, i.e. one descriptor is replaced by another. This selection crossover mutation process is repeated until all of the n parents in the gene pool are replaced by their offspring. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until a desired number of generations or target fitness score is reached. In our GA program that was written by MATLAB one modification is made. This is the inclusion of elitism, which protect the fittest individual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. In original studies, the fitness function of the individual was determined by a function related to the residual error in the regression analysis of the training data. Here we try to use varieties of fitness functions, which are proportional to the residual error of the prediction set and the number of selected variables according to the following equation:
Fitness = (1 − w) (1/MSE_t) + w (1/m)

In this equation MSE_t is mean square error of test set (Tan et al., 2016), m is the number of variables in the represented model and w is a parameter between 0 and 1 that implies the weights of m in the value of fitness. In fact the value of w determines the number of variables exist in the chromosome. Some experiments were applied using different value for w. Results obtained showed that for small value of w the number of variables in the fittest individual was high and on the other hand if the value of w was to be high the number of variables in the best chromosome was small. Hence after some experiments the value of w was set to be 0.9. It is worth noting that the parameter of w was determined in a preliminary study, before the overall genetic algorithm optimization has been carried out. Here for the calculation of the fitness of each chromosome a nonlinear model was constructed using variables consists in each chromosome separately ANN and the value of MSE_t was calculated using this model. This procedure was applied for each chromosome separately.

Artificial neural network

Artificial neural network (ANN) is generally applied as a technology offering an alternative way to simulate ambiguous and complex problems. The importances of using neural networks in process modeling are that they have learning and generalization abilities as well as nonlinearity. Numerous applications of ANN have been known in pattern recognition, materials modeling, data analysis and property prediction (Swarup, 2016; Cai et al., 2018). A neural network is a computational structure, consisting of a number of highly interconnected processing units called neurons. The neurons are connected to each other by weighted links over which signals can pass. Each neuron receives multiple inputs from other neurons in proportion to their connection weights and generates a single output, which may be propagated to several other neurons (Swingler, 2016). Among different types of ANN models, back-propagation (BP) algorithm, an iterative gradient algorithm, is so popular that it has been used for the present work. BP neural network consists of an input layer, one or more hidden layers and output layers. In order to train the network using the BP algorithm, the differences between the ANN output and its desired values are calculated, and the values of weights and biases modified by using these error terms. In the present work, an ANN program was written in MATLAB in our laboratory. This network was feed-forward fully connected that has three layers with sigmoidal transfer function. The network inputs are selected descriptors, the signal of the output node represents the pKi values for sulfonylurea herbicides and the number of nodes in the hidden layer would be optimized. The levenberg–Marquardt (LM) algorithm is one of the most efficient learning algorithms for neural networks (Kanzow et al., 2018). The advantages of using LM algorithm are that specifying momentum or rate is not necessary and training processes are much more rapid. Therefore, in this work LM algorithm was used to develop the nonlinear model. The ANNs cannot be able to select important descriptors that would be used as its inputs; therefore it is necessary to apply a variable selection method. In this work, we use genetic algorithm and stepwise multiple regression feature-selection methods for these purposes. Then the optimized network was trained using training set for the adjustment of weights and biases values. It is known that a neural network can become over-fitted. An over-fitted network has usually learned the stimulus pattern it has seen perfectly, but cannot give an accurate prediction for unseen stimuli, and it is no longer able to generalize. There are several methods for overcoming this problem. One method is to use a test set to evaluate the prediction power of the network during its training. Since the test error is not a good estimate of the generalization error, the prediction potential of the model was evaluated on a third set of data, named the prediction set. The compounds in the prediction set were not used during the training process and were reserved to evaluate the predictive power of the generated ANN.

Results and discussion

Nonlinear model
The data set and corresponding observed and predicted values of the pKi of all molecules studied. For the selection of the most important descriptors both genetic algorithm and stepwise multiple regression techniques were used. Then these descriptors were used as inputs for generated ANNs. In other hand, two separate ANNs were constructed that used these descriptors as inputs and their outputs are pKi values of the molecules. These models referred as GANN and stepwise-NN, respectively. Applied GA contained a population of \( \cdot \cdot \cdot \) individuals, which evolved for \( \cdot \cdot \cdot \) generations. Then by comparison between the fitness values of individuals, the best model was chosen. The process of the genetic algorithm is shown in Fig. \( \cdot \) for all the generations from the beginning to the end of the process. The best fitness plot for the GA maps the gradual convergence of the best fitness values of successive generations towards the final optimum value. It indicates that for this case study, after \( \cdot \cdot \cdot \) generations, the optimal results can be obtained.

![Figure 1](image.png)

**Fig. 1**. The results of genetic algorithm for \( \cdot \cdot \cdot \) generations.

Table 1 shows the names of descriptors of the ANNs models that their descriptors were chosen by stepwise-NN and GANN methods. Although the numbers of descriptors in two models are identical but they are differ from each other. The minimum MSEs to control over-fitting appear after \( \cdot \) epochs. As can be seen from Fig. 2, after \( \cdot \) iterations MSE of test set increases while MSE of training set decreases or changes slightly. Table 2 shows the architecture and specification of the optimized network. The statistical parameters obtained by these models for the training and prediction set were shown in Table 3. These simulations demonstrated some significant differences between two networks. It can be seen from this table that statistical results of the GANN are better than other method. Also these results reveal that the GA is superior method for feature-selection in this QSAR study. The predicted values of the pKi using GANN model for data set were shown in Fig. 3 shows a plot of the GAANN calculated versus the experimental values of pKi for the data set molecules. Correlation coefficient of \( \cdot \cdot \cdot \cdot \) for this plot confirms the suitability of the ANN model to predict of permeability coefficient. Results obtained reveals that there are some nonlinear relation between the inhibition constant of sulfonylurea herbicides and the selected structural molecular descriptors.

| Table 1. Definitions and notations of descriptors for stepwise-NN and GANN. |
a: The notations are based on Dragon software.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>stepwise-NN notation</th>
<th>GANN Descriptor</th>
<th>GANN notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest eigenvalue</td>
<td>BEHm</td>
<td>Randic-type eigenvector</td>
<td>VRA1</td>
</tr>
<tr>
<td>Superpendentic index</td>
<td>SPI</td>
<td>Moran autocorrelation</td>
<td>MATS10</td>
</tr>
<tr>
<td>Average eigenvector coefficient</td>
<td>VEZ</td>
<td>Radial distribution function</td>
<td>RDF1.2m</td>
</tr>
<tr>
<td>Complementary information content</td>
<td>CIC</td>
<td>Molecular representation of structure</td>
<td>Mor1.u</td>
</tr>
<tr>
<td>Structural information content</td>
<td>SIC</td>
<td>Autocorrelation</td>
<td>Hve</td>
</tr>
<tr>
<td>Second Mohar index</td>
<td>TI</td>
<td>Number of ethers</td>
<td>nROR</td>
</tr>
<tr>
<td>Salvation connectivity index</td>
<td>Xsol</td>
<td>No. of X—CR—X</td>
<td>C-0.31</td>
</tr>
</tbody>
</table>

Table 1. Architectures of the optimized ANN.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSEt</th>
<th>MSEp</th>
<th>Rt</th>
<th>Rp</th>
<th>Ft</th>
<th>Fp</th>
</tr>
</thead>
<tbody>
<tr>
<td>GANN</td>
<td>8.848</td>
<td>8.22</td>
<td>8.744</td>
<td>8.78</td>
<td>14.22</td>
<td>48</td>
</tr>
<tr>
<td>stepwise-NN</td>
<td>8.229</td>
<td>8.88</td>
<td>8.719</td>
<td>8.462</td>
<td>26.2</td>
<td>19</td>
</tr>
</tbody>
</table>

t: training set
p: prediction set

Table 1. Statistical parameters obtained using stepwise-NN and GANN models.

![Graph showing the Mean Square Error (MSE) over epochs for MSEtrain and MSEtest](image)
Model validation

In spite of good accuracy and apparent mechanistic appeal, QSAR models should pass rigorous validation tests to be useful as reliable screening tools. Y-randomization test is a tool used in validation of QSAR models, whereby the performance of the original model in data description is compared to that of models built for permuted (randomly shuffled) response, based on the original descriptor pool and the original model building procedure. The Y-scrambling procedure was performed to ensure that there is not any chance correlation in data matrix (Tropsha et al., 2882). The mean value of R after Y times Y-scrambling was 0.934, which disapproved the chance correlation probability. The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R), but is mainly their possibility of predictive application. For this reason internal validation, leave one out cross-validation (LOO) and leave n out (LNO) were applied on GANN model which resulted in square cross validated correlation coefficient $R_{cv}^2$ and $Q^2$ are (0.934) and (0.917) respectively, which confirmed good predictive ability of this model.

Interpretation of descriptors

In this work, quantitative relationships between inhibition constants of sulfonylurea herbicides and their structural descriptors were investigated by using nonlinear approach. As shown in Table 1, seven descriptors appeared in the GANN model. VRA is among the topological descriptors. MATS belongs to the autocorrelation descriptors. RDF is a type of radial distribution function (RDF) descriptors that are based on the distance distribution in the molecule. 2D Molecule Representation of Structures based on Electron diffraction (Mor) is derived from Infrared spectra simulation using a generalized scattering function. He belongs to the GATEWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors. nROR descriptor describes the number of aliphatic ether functions. C-821 Descriptor gives the number of specified atom-centered fragments present in molecule (Docherty and Kulpa, 2882; Soltzberg and Wilkins, 1799; Consonni et al., 2882; Viswanadhan et al., 1772).

To determine the order of importance of descriptors in GANN model, the sensitivity analysis was performed. According to this method, the differences between the mean-square error (MSE) of the complete data set and the MSE were obtained when the ith variable is excluded from the trained network (MSEi), and were shown as Mdiffi:

$$Mdiffi = MSE_{i} - MSE$$

It is obvious that the most important variable is the one that leads to the highest value of Mdiffi. The values of Mdiffi for GANN model were calculated and plotted in Fig. 8. As it can be seen in this figure, the orders of importance of selected molecular descriptors are: $H_{e} > MATS > nROR > C-821 > Mor > RDF > VRA$. According to the sensitivity analysis results, among these descriptors, the GANN model has the least and most sensitivity to VRA and $H_{e}$ descriptors.
respectively. This means that H\textsuperscript{2}e is the most effective parameter in inhibition of AHAS by the sulfonylurea herbicides.

![Fig. 4. Sensitivity analysis results.](image)

**Conclusion**

For sulfonylurea analogs, various approaches of pKi prediction have been proposed. Roy and Paul presented two predictive equations for \( \xi \) sulfonylurea derivatives but we used \( \gamma \) sulfonylurea herbicides. On the other hand, our data set are more diverse and therefore our model is more general than their models. Also in this work, ANN was applied to consider the effect of nonlinear phenomena in the inhibition behavior of the compounds, but they calculated linear models. Comparisons of the results of these two works which are presented in Table 4 indicate the superiority and better fitting of the GANN results obtained in this work. The superiority of this model accomplishes two messages. First, the evolutionary programming of genetic algorithm is very effective in the selection of the best descriptors, second, the strength of ANN in its ability to allow for flexible mapping of the selected features.

**Table 4:** Statistical parameters for pKi of Training Set.

<table>
<thead>
<tr>
<th>Method</th>
<th>( n )\textsubscript{Training}</th>
<th>( R^\text{2} )\textsubscript{Training}</th>
<th>( Q^\prime )</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( n )</td>
<td>( R^\text{2} )</td>
<td>( Q^\prime )</td>
<td>( F )</td>
</tr>
<tr>
<td>B</td>
<td>( n )</td>
<td>( R^\text{2} )</td>
<td>( Q^\prime )</td>
<td>( F )</td>
</tr>
</tbody>
</table>

A: this work  
B: Roy and Paul work

\( n \): number of molecules

**References**


Cartwright, H.M., \( (1^\text{IA}) \). Applications of artificial intelligence in chemistry, Oxford University Press, Oxford.


