Modelling the effect of structural QSAR parameters on skin penetration using genetic programming

K K Chung and D Q Do

Faculty of Pharmacy, University of Medicine and Pharmacy in Ho Chi Minh City, Vietnam
E-mail: dq_duong@yahoo.com

Received 16 June 2010
Accepted for publication 26 September 2010
Published 22 October 2010
Online at stacks.iop.org/ANSN/1/035003

Abstract
In order to model relationships between chemical structures and biological effects in quantitative structure–activity relationship (QSAR) data, an alternative technique of artificial intelligence computing—genetic programming (GP)—was investigated and compared to the traditional method—statistical. GP, with the primary advantage of generating mathematical equations, was employed to model QSAR data and to define the most important molecular descriptions in QSAR data. The models predicted by GP agreed with the statistical results, and the most predictive models of GP were significantly improved when compared to the statistical models using ANOVA. Recently, artificial intelligence techniques have been applied widely to analyse QSAR data. With the capability of generating mathematical equations, GP can be considered as an effective and efficient method for modelling QSAR data.

Keywords: genetic programming, genetic algorithms, statistical method, quantitative structure–activity relationship (QSAR), skin penetration

Classification number: 3.02

1. Introduction

QSAR data are known to be highly complex with nonlinear relationships. Modelling the relationships between chemical structures and biological effects is the target of structure–activity studies [1]. Therefore, with a large number of molecular descriptors, finding an efficient tool to glean knowledge from the relationships of structure–activity in QSAR data is necessary. There are many methods applied routinely for QSAR. In addition to statistical methods [2–4], artificial intelligence techniques have been employed to deal with such data. Recently, several research papers have discussed whether artificial intelligence can deal with the problem of QSAR modelling. Examples include the application of neural networks for discriminating antibacterial activity [5] and for 3D QSAR [6], or an application of neural networks and genetic algorithms for modelling QSAR data [7].

In the previous studies, an alternative artificial intelligence method that was developed based on the Darwinian theory called genetic programming (GP) was successfully employed to modelling pharmaceutical formulation data [8, 9]. In this study, GP was used to analyse QSAR data and a comparison of the ability of GP and statistical methods to model QSAR data was examined for a QSAR data set from a study on the effect of structural QSAR parameters on skin penetration. For the comparison, an assumption was made that in all cases there was no interaction between the molecular descriptors (i.e. all independent variables).

2. Materials and methods

2.1. Materials

The data set used in this work was taken from a study of the effect of structural QSAR parameters on skin penetration. In [3], Ghafourian and Fooladi analysed 20 variables of 17 inputs and 3 outputs—permeability (log $K_p$), skin/water partition (log $K_m$) and diffusion coefficient (log $D/h$), as listed in table 1. Ghafourian and Fooladi analysed 39 compounds. However, in this study, for ease of comparison...
with the statistical results in the literature, only 22 compounds (compound 1–14 and 26–33) were used for evaluating the predictability of GP for all responses. From a total of 22 compounds used for observation, 20 records are used as training data and 2 records are used as test data selected by the Smart Selection function in GP.

2.2. Methods

GP has been known primarily as a learning technique producing mathematical equations [10–12]. It is founded on the basic principle of Darwin’s theory of evolution. In nature, biological structures that are more successful at operating in their given environment have a higher probability of surviving and reproducing. These structures are considered to be the result of Darwinian natural selection operating in an environment over a period of time. With a similar principle of natural selection, Koza introduced GP that attempts and succeeds in applying this evolutionary theory in order to find the best (fittest) or the most appropriate equation (solution) for a problem domain [10].

2.2.1. Representation scheme. GP modifies individuals of the population after a number of runs in every cycle, called a generation, to find the best solution. The structure of individuals in GP is a tree (see figure 1(a)). Possible structures in genetic programming are the set of all compositions of functions and the set of terminals.

The elements in the function set may include:

- arithmetic operations (+, −, *, etc),
- mathematical functions (such as sin, cos, exp and log),
- Boolean operations (such as AND, OR and NOT),
- conditional operators (such as If-Then-Else),
- functions causing iteration (such as Do-Until),
- functions causing recursion,
- any other domain-specific functions that may be defined.

The terminals are typically either a variable or a constant value.

2.2.2. Genetic operations. In the Darwinian principle of reproduction and survival of the fittest, a population always has many generations and the individuals of new generations are the result of the combination of the individuals in the previous generations. In other words, this is a transfer of a set of individuals into a new generation of the population using genetic operations. Crossover, reproduction and mutation are genetic operations for creating the new individuals in each generation (figures 1(b) and (c)). The crossover operation is the most negative atomic charge on the hydrogen bond accepting heteroatoms

\[ q_{\text{b}} \]

the most negative atomic charge on the hydrogen bond accepting heteroatoms

\[ q_{\text{b}} \]

The most negative atomic charge on the hydrogen bond accepting heteroatoms

\[ q_{\text{b}} \]

The most negative electrostatic potential on the solvent accessible surface of molecules

\[ E_{\text{BOM}} \]

Energy of the highest occupied molecular orbital

\[ E_{\text{HOM}} \]

Molecular weight

\[ \log K \]

Permeability coefficient

\[ \log K_m \]

Skin/water partition coefficient

\[ \log D/h \]

The terminals are typically either a variable or a constant value.

\[ \log K \]

Permeability coefficient

\[ \log K_m \]

Skin/water partition coefficient

\[ \log D/h \]

The terminals are typically either a variable or a constant value.

\[ \log K \]

Permeability coefficient

\[ \log K_m \]

Skin/water partition coefficient

\[ \log D/h \]

The terminals are typically either a variable or a constant value.

\[ \log K \]

Permeability coefficient

\[ \log K_m \]

Skin/water partition coefficient

\[ \log D/h \]
risk minimization), AIC (Akaike’s information criterion), MDL (minimum description length) and FPE (final prediction error) [13].

2.2.4. Implementation scheme. The implementation of GP is performed by the following steps:

Step 1: One or more initial populations of individuals are randomly generated with functions and terminals related to the problem domain.

Step 2: The implementation of GP iteratively performs the following steps until the termination criterion has been satisfied:

(i) The fitness value of every individual is estimated according to a selected fitness measure.
(ii) All individuals in the population are sorted based on their fitness values.
(iii) The next generation is produced using the genetic operations
(iv) The termination criterion is checked. If it is not satisfied, the next iteration is performed; if it is satisfied, go to step 3.

Step 3: The result may be a solution to the problem domain.

In particular, the quality of the final model is controlled by either the selection of training data or the values of the control parameters (crossover, mutation, reproduction and fitness function).

2.3. Software tool

The software used was a modified form of that described previously [14], but with additional exponential and other mathematical functions as well as the fitness functions listed. Table 2 lists the values of the control parameters used in the present study.

In order to evaluate the quality of a model generated by GP, the correlation coefficient $R^2$ was computed, with higher values of $R^2$ indicating improved quality of the model,

$$R^2 = \left[ 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \right] \cdot 100,$$
Validation of the predictive power of GP for mathematical records. The predicted value from the model and models were needed to obtain an acceptable model. For this reason, the simplicity of the predictive equation resulted to produce equation (2), only three parameters—the number of hydrogen bonding heteroatoms [OT], the sum of atomic charge on the hydrogen bonding heteroatoms $[q^\text{-b}]$, and the second-order molecular shape index $[\gamma^2 \chi]$ from a total of 17 input variables influence the permeability coefficient $K_p$.

The values of the training parameters, which GP used to produce equation (1), are shown in table 3. For this response, the simplicity of the predictive equation resulted from a maximum of 10 nodes in the training trees. However, because of this small number of tree nodes, the population size required was 10 000 with 500 generations. In addition, maximum values of crossover and mutation in the training models were needed to obtain an acceptable model. For this response, the AIC fitness function was used to generate a high-quality model. The details of the equations from the statistical method are shown below:

$$\log K_p = 3.99 \log [TA] + 4.53 [q^\text{-b}] - 0.762 [OT] - 11.364$$

$$R^2 = 0.97,$$  \hspace{1cm} \text{(2)}

$$\log K_p = -0.786 [OT] + 0.252 [\gamma^2 \chi] - 1.617 [q^\text{-a}] - 5.767$$

$$R^2 = 0.97.$$

As analysed in the literature using MINITAB statistical software [3], equation (2) shows that $\log K_p$ is correlated with the logarithm of total solvent accessible surface area calculated using a probe of 1.4 Å radius ($\log [TA]$), the sum of atomic charge on the hydrogen bonding heteroatoms $[q^\text{-b}]$, and the number of hydrogen bonding heteroatoms [OT] in equation (2). In equation (3), this response is also correlated with the number of hydrogen bonding heteroatoms [OT], the second-order molecular shape index $[\gamma^2 \chi]$ and the sum of atomic charge on the hydrogen bonding heteroatoms calculated by the AM1 method $[q^\text{-a}]$.

When comparing the predictive models between GP and statistical methods, the results of regression analysis for both methods using whole data of 22 structures in figure 2 indicated that there are only minor differences between these techniques. While producing a slightly lower linear $R^2$ value, GP generated a better model with more satisfactory values of slope and intercept coefficients for this response.

### Table 2. The value of control parameters of GP used in this study.

| Training parameter | Value | Training parameter | Value |
|--------------------|-------|--------------------|-------|
| Population size(s) | 10 000 | Crossover          | 0.2–0.5 |
| Generations        | 100–500–1000 | Mutation          | 0.2–0.5 |
| Max. no. nodes     | 10–15–20 | Constants mutation | 0.5 |
| Max. depth         | 10     | Regeneration       | 0.7 |
| Addition functions | Exp    | Fitness type      | AIC, FPE, MDL, MSE and SRM |
| reproduction       | 0.05–0.2 | Test formulations | 2, 10 |

where $\bar{y}$ is the mean value of the dependent variable $y$, $\hat{y}$ is the predicted value from the model and $n$ is the number of records.

### 3. Results and discussion

Validation of the predictive power of GP for mathematical and formulation data has been evaluated and presented elsewhere [8, 9].

#### 3.1. The predicted results of permeability coefficient ($\log K_p$)

The permeability coefficient has the following expression:

$$\log K_p = [q^\text{-b}] [OT] - [\gamma^2 \chi] / 3.651 - 6.457 (R^2 = 0.91).$$  \hspace{1cm} \text{(1)}

From equation (1), the $R^2$ value of the predictive model is considered as a satisfactory result and moreover the predictive equation is extremely simple for the permeability coefficient ($\log K_p$) response. In equation (1), only three parameters—the number of hydrogen bonding heteroatoms [OT], the sum of atomic charges on the hydrogen bonding heteroatoms $[q^\text{-b}]$, and the second-order molecular shape index $[\gamma^2 \chi]$—from a total of 17 input variables influence the permeability coefficient $\log K_p$.

The values of the training parameters, which GP used to produce equation (1), are shown in table 3. For this response, the simplicity of the predictive equation resulted from a maximum of 10 nodes in the training trees. However, because of this small number of tree nodes, the population size required was 10 000 with 500 generations. In addition, maximum values of crossover and mutation in the training models were needed to obtain an acceptable model. For this response, the AIC fitness function was used to generate a high-quality model. The details of the equations from the statistical method are shown below:

$$\log K_p = 3.99 \log [TA] + 4.53 [q^\text{-b}] - 0.762 [OT] - 11.364$$

$$R^2 = 0.97,$$  \hspace{1cm} \text{(2)}

$$\log K_p = -0.786 [OT] + 0.252 [\gamma^2 \chi] - 1.617 [q^\text{-a}] - 5.767$$

$$R^2 = 0.97.$$

As analysed in the literature using MINITAB statistical software [3], equation (2) shows that $\log K_p$ is correlated with the logarithm of total solvent accessible surface area calculated using a probe of 1.4 Å radius ($\log [TA]$), the sum of atomic charge on the hydrogen bonding heteroatoms $[q^\text{-b}]$, and the number of hydrogen bonding heteroatoms [OT] in equation (2). In equation (3), this response is also correlated with the number of hydrogen bonding heteroatoms [OT], the second-order molecular shape index $[\gamma^2 \chi]$ and the sum of atomic charge on the hydrogen bonding heteroatoms calculated by the AM1 method $[q^\text{-a}]$.

When comparing the predictive models between GP and statistical methods, the results of regression analysis for both methods using whole data of 22 structures in figure 2 indicated that there are only minor differences between these techniques. While producing a slightly lower linear $R^2$ value, GP generated a better model with more satisfactory values of slope and intercept coefficients for this response.

#### 3.2. The predicted results of skin/water partition coefficient ($\log K_m$)

The skin/water partition coefficient is determined by the following formula:

$$\log K_m = (-8.906 [\log P_{oct}] - [\gamma^2 \chi]) / (2.026 \log \text{HOMO})$$

$$R^2 = 0.96.$$  \hspace{1cm} \text{(4)}

For this response from GP given in equation (4), the parameters of the octanol/water partition coefficient molecular weight $[\log P_{oct}]$, the second-order molecular shape index $[\gamma^2 \chi]$ and the energy of the highest occupied molecular orbital $\log \text{HOMO}$ correlate with the skin/water partition coefficient ($\log K_m$) response with a high $R^2$ value of 0.96. This result is similar to the prediction for the first response of these QSAR data in that the equation predicted is simple with only three parameters.

As seen in table 4, the AIC fitness function also gave a high quality model with a population size of 10 000 500 generations, as well as the same values of crossover and mutation as shown in table 2.
Figure 2. Scatter plots, linear equations and linear $R^2$ values of log $K_p$ generated from GP equation (1) (▲) and statistical equations (2) (△) and (3) (■).

### Table 4. Values of training parameters of GP producing (4).

| Training parameter | Value   |
|--------------------|---------|
| Population size(s) | 10,000  |
| Max. no. generations | 500    |
| Max. no. nodes     | 15      |
| Max. depth         | 15      |
| Addition functions |         |
| reproduction       | 0.2     |
| Crossover          | 0.5     |
| Mutation           | 0.5     |
| Constants mutation | 0.5     |
| Regeneration       | 0.7     |
| Fitness type       | AIC     |
| Test formulations  | 14, 18  |

### Table 5. Values of training parameters of GP producing (equation (8)).

| Training parameter | Value   |
|--------------------|---------|
| Population size(s) | 10,000  |
| Max. no. generations | 500    |
| Max. no. Nodes     | 15      |
| Max. depth         | 15      |
| Addition functions |         |
| reproduction       | 0.05    |
| Crossover          | 0.5     |
| Mutation           | 0.5     |
| Constants mutation | 0.5     |
| Regeneration       | 0.7     |
| Fitness type       | MDL     |
| Test formulations  | 10, 18  |

The observed log $K_p$ values are presented in the first column, and the predicted log $K_p$ values are given in the second column. The values of $R^2$ are shown in the third column for each equation.

In order to examine the predictability of GP for the skin/water partition coefficient (log $K_m$), regression analysis was applied, and the results of both methods for this response were compared. From figure 3, the results of regression analysis for both methods using whole data of 22 structures show that there are no major differences between the two methods in this case. The linear $R^2$ values of three equations and regression coefficients observed are also similar.

**3.3. The predicted results of diffusion coefficient (log $D/h$)**

The diffusion coefficient log $D/h$ is predicted by GP as follows:

$$\log D/h = -0.49[HT] + ([ESP^{-}] - [q_{s}^{+}])/6.25 - 130.28[q_{s}^{-d}]/[ESP^{-}]/(R^2 = 0.97).$$

It can be seen from equation (8) that the predictive model for this response has extremely high train and test $R^2$ values with four input variables influencing the final results of this response. The diffusion coefficient (log $D/h$) response is correlated with the number of hydrogen bonding hydrogens [HT], the sum of atomic charges on the hydrogen bonding.
heteroatoms calculated by both methods of the AM1 method \([q_a]^-\) and the classical method \([q_a]^-\) [3], and the most negative electrostatic potential on the solvent accessible surface of molecule [ESP] parameters.

GP required the same population size and generations as in the predictions for the previous responses of 10 000 and 500, respectively (see table 5). Apart from the training model using the MDL fitness function and a larger number of nodes in the training model are the same for the GP prediction for the first response of this data set.

In the original paper, Ghafourian and Fooladi [3] showed that this response could also be given by

\[
\log D/h = \log K_p = \log K_m. \tag{9}
\]

Based on equation (9), an alternative equation for this response predicted by GP using the predictions from the log \(K_m\) and log \(K_p\) responses was

\[
\log D/h = \log K_p - \log K_m = [q_a]^- OT - \left[2\alpha\right]/3.651 - 6.457 + (8.906)\log P_{\text{HOMO}} + \left[2\alpha\right]/2.026(E_{\text{HOMO}}). \tag{10}
\]

As reported in the literature [3], the response of the diffusion coefficient \(\log D/h\) was modelled successfully using the statistical method with very high \(R^2\) values, as shown below:

\[
\log D/h = -3.706 - 0.0037[\text{MW}] - 0.609[\text{HT}] + 0.077[\text{ESP}^-] \quad \left(R^2 = 0.99\right) \tag{11}
\]

\[
\log D/h = -4.223 - 0.0356[\alpha] - 0.602[\text{HT}] + 0.065[\text{ESP}^-] \quad \left(R^2 = 0.99\right). \tag{12}
\]

The model predicted by GP in equation (8) demonstrates that there are four parameters from a total of 17 influencing the diffusion coefficient (\(\log D/h\)).

When the predictive model of GP (8) is compared with the statistical results (equations (11) and (12)), both methods indicate that the diffusion coefficient (\(\log D/h\)) is negatively correlated with the number of hydrogen bonding hydrogens [HT] (those connected to the oxygen or nitrogen atoms), and positively correlated with the most negative electrostatic potential on the solvent accessible surface of molecules [ESP], which can be considered as the hydrogen bonding basicity parameter [3]. However, in the statistical result, the molecular weight [MW] influences (\(\log D/h\)), while in the GP result,
\[
\log \frac{D}{h} \text{ is correlated with the sum of atomic charges on the hydrogen bonding heteroatoms } [q_a^{-a}] \text{ and } [q_b^{-b}]\text{ parameters.}
\]

From figure 4 comparing equation (8) from GP to equations (11) and (12), it is apparent that GP delivers a slightly higher quality model for parameter \((\log \frac{D}{h})\) in comparison with the statistical results based on the regression analysis for whole data of 22 records. Although the linear \(R^2\) values of the three models are equal to 0.98, the slope and intercept coefficients in the linear equations of GP are closer to 1.00, while those from the statistical equations are 0.00.

From figure 4, it can also be seen that equation (10) is an improved model with more satisfactory values of slope and intercept coefficients in comparison with the statistical results.

In general, for the diffusion coefficient \((\log \frac{D}{h})\) response, while not generating the same influential parameters as reported in the literature, the quality of prediction of GP for this response is high in both observations on the \(R^2\) value as well as on the regression analysis.

4. Conclusions

For a highly complex problem with nonlinear relationships between chemical structures and biological effects in quantitative structure–activity relationship (QSAR) data, genetic programming, with its advantage of generating mathematical equations, has been shown to be an efficient method. In contrast to statistical approaches, GP requires no assumption of the functional form (e.g. linear or quadratic) to be used to describe the cause-and-effect relationships.

References

[1] Selassie C D 2003 Chapter 1: History of quantitative structure–activity relationships Burger’s Medicinal Chemistry and Drug Discovery Volume 1 Drug Discovery 6th edn, ed D J Abraham (New York: Wiley)
[2] Haibers P D T, Lohanov V S, Katritzky A R, Shah D O and Karelson M 1996 Langmuir 12 1462
[3] Ghafoorian T and Fooladi S 2001 Int. J. Pharm. 217 27
[4] Lee H, Yoom H, Kim Y G, Yoon C N, Jin C, Choi J S, Kim B R and Kim D H 1998 Biochem. Pharmacol. 55 1369
[5] Tomás-Vert F, Pérez-Giménez, Salabert-Salvador M T, García-March F J and Jaén-Oltra J 2000 J. Mol. Struct. (Theochem) 504 249
[6] Hasegawa K, Morikami K, Shiratori Y, Ohtsuka T, Aoki Y and Shimma N 2003 Chemometr. Intell. Lab. Syst. 69 51
[7] Niculescu S P 2003 J. Mol. Struct. (Theochem) 622 71
[8] Do D Q 2006 Genetic programming for formulation design PhD Thesis University of Bradford, UK
[9] Do D Q, Rowe R C and York P 2008 Int. J. Pharm. 351 194
[10] Koza J R 2006 Introduction to Genetic Programming—Advances in Genetic Programming (Cambridge, MA: MIT Press) chapter 2
[11] Koza J R 1997 Future Work and Practical Applications of Genetic Programming—Handbook of Evolutionary Computation. (Bristol: Institute of Physics Publishing and New York: Oxford University Press) pp 1–6
[12] Koza J R 1998 Genetic Programming on the Programming of Computers by Means of Natural Selection Sixth printing (Cambridge, MA: MIT Press)
[13] Weyer E and Kavli T 1994 The ASMOD algorithms: some new theoretical and experimental results Proc. IEEE Colloquium Adv. Neural Netw. Control Syst. 3 1
[14] Zongker D, Punch B and Rand B 1996 Lil-gp Genetic Programming System Michigan State University http://garage.cse.msu.edu/software/lil-gp/