Monatsh Chem (2011) 142:943–948
DOI 10.1007/s00706-011-0510-x

ORIGINAL PAPER

Prediction of antileukemia activity of berbamine derivatives by genetic algorithm–multiple linear regression

Mehdi Nekoei · Mahmoud Salimi · Mohsen Dolatabadi · Majid Mohammadhosseini

Received: 14 November 2010 / Accepted: 14 April 2011 / Published online: 19 May 2011

Abstract A quantitative structure–activity relationship study was performed on a data set of 32 berbamine derivatives that possess antileukemia activity. Semiempirical quantum chemical calculation (AM1 method) was used to find the optimum 3D geometries of the studied molecules. A suitable set of molecular descriptors was calculated and genetic algorithm–multiple linear regression was employed to select the descriptors that resulted in the models with the best fit to the data. A multiple linear regression model with five selected descriptors was obtained. The values of statistical measures such as R^2, Q^2, and F obtained for the training set were within acceptable ranges, so this relationship was applied to the test set. The predictive ability of the model was found to be satisfactory, and it can therefore be used to design similar groups of compounds. Also, results suggest that the charge, electronegativity, and atomic van der Waals volumes of the molecules are the main independent factors that contribute to the antileukemia activity of berbamine derivatives.

Keywords QSAR · Genetic algorithm · Multiple linear regression · Berbamine · Antileukemia

M. Nekoei (✉) · M. Mohammadhosseini
Department of Chemistry, Shahrood Branch,
Islamic Azad University, Shahrood, Iran
e-mail: m_nekoei1356@yahoo.com; nekoei_m1@yahoo.com

M. Salimi
Chemical Engineering Department, Faculty of Engineering,
Islamic Azad University, Arak Branch, Arak, Iran

M. Dolatabadi
Department of Chemistry, Faculty of Science,
University of Birjand, Birjand, Iran

Introduction

Gleevec (also called imatinib or STI571), which is an inhibitor of bcr/abl tyrosine kinase [1], has been a remarkable success in the treatment of chronic myelogenous leukemia (CML). However, a significant proportion of the patients who are chronically treated with Gleevec develop resistance [2]. Thus, it is necessary to identify novel inhibitors that are active against Gleevec-resistant mutants of bcr/abl oncoprotein. Recently, Rongzhen Xu et al. [3] reported that berbamine can selectively induce cell death of both Gleevec sensitive- and resistant-Ph$^+$ CML cells. They also reported that berbamine can selectively induce caspase-3-dependent apoptosis of leukemia NB4 cells via the survivin-mediated pathway, suggesting that berbanine may be a novel agent for the treatment of leukemia [4]. Moreover, in order to discover novel antileukemia berbamine derivatives, they described the results of berbamine derivative synthesis as well as their antileukemia activities for imatinib-resistant K562 leukemia cells [5].

Although several experimental methods are available for screening the biological activities of chemicals (e.g., in vivo and in vitro assay tests), and these have also all been carried out using receptors and other biological materials of human, rat, mouse, and calf origin at the very least [6], they are costly, time-consuming, and can potentially produce toxic side products. Quantitative structure–activity relationship (QSAR) studies have been demonstrated to be an effective computational tool for understanding the relationships between the structures of molecules and their properties, such as biological activity [7], physical properties [8], and toxicity [9]. QSAR studies express the biological activities of compounds as a function of their various structural parameters and describe how variations in biological activity depend on changes in chemical structure [10]. If such a relationship can be derived from the structure–activity data, the model equation allows medicinal chemists to say with some confidence which property plays an important role in the mechanism of action of the drug. The success of a QSAR study depends on the appropriate selection of robust statistical methods to elucidate the predictive model and relevant structural parameters to express the essential features within chemical structures.

Nowadays, genetic algorithms (GA) are well known to be interesting and widely used methods of variable selection [11–13]. GA is a stochastic method that is used to solve optimization problems defined by fitness criteria; problems are solved by applying the evolution hypothesis of Darwin and different genetic functions (i.e., crossover and mutation).

In the present work, we used GA for variable selection and model development in the QSAR analysis of berbamine derivatives. Finally, the accuracy of the proposed model was illustrated using leave-one-out (LOO) and leave-many-out (LMO) cross-validations and Y-randomization techniques.

Results and discussion

GA variable subset selection method based MLR was used to select the most important descriptors. The five most significant descriptors according to the GA-MLR algorithm are GGI1, Mor17e, G1v, G3e, and H8v.

Since collinearity between the variables degrades the performance of the MLR-based QSAR models, the correlation between each of the variables used in this study was examined before a multiparametric analysis was undertaken. The correlation matrix itself shows how the descriptors used are correlated. The correlation matrix obtained in the present case is shown in Table 1. From Table 1, it is clear that the correlation coefficient values for all pairs of descriptors are <0.50, which means that the selected descriptors are independent.

Table 1 The correlation coefficient matrix for the descriptors used in this study

	GGI1	Mor17e	G1v	G3e	H8v
GGI1	1				
Mor17e	0.47	1			
G1v	−0.43	0.11	1		
G3e	0.12	0.36	0.15	1	
H8v	0.43	−0.20	−0.37	−0.19	1

In order to build and test the models, a data set of 32 compounds was randomly separated into a training set of 26 compounds. This was used to build the model and a test set of six compounds, which was applied to test the built model. With the selected five descriptors, we built a linear model using the training set data, and Eq. 1 was obtained:

$$\begin{aligned} \mathrm{p}IC_{50} = & -14.46 + 0.12(\pm 0.11)\ \mathrm{GGI1} \\ & - 0.06(\pm 0.11)\ \mathrm{Mor17e}\ + 58.29 \\ & (\pm 0.11)\ \mathrm{G1v}\ + 56.17(\pm 0.11)\ \mathrm{G3e} \\ & + 4.71(\pm 0.11)\ \mathrm{H8v} \end{aligned} \tag{1}$$

$$N = 26, R^2_{\text{train}} = 0.861, Q^2_{\text{LOO}} = 0.784, Q^2_{\text{LMO}} = 0.792,$$
$$F = 24.785, R^2_{\text{test}} = 0.974.$$

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, Q^2_{LOO} and Q^2_{LMO} are the squared cross-validation coefficients for leave-one-out and leave-many-out cross-validation, respectively, and F is the Fisher F statistic. The figures in parentheses are the standard deviations. This model was then used to predict the test set data. The prediction results and relative error percentages (RE%) are given in Table 2. As can be seen from Table 2, the calculated values of pIC_{50} are in good agreement with those of the experimental values, and the variables used in this equation can also be used predict the activity of the molecules with a prediction error percentage of <6%. The predicted values of pIC_{50} for the compounds in the training and test sets using Eq. 1 are plotted versus the experimental values in Fig. 1. A plot of the residuals for the predicted values of pIC_{50} for the training and test sets versus the experimental values is given in Fig. 2. As can be seen, the model did not show any proportional or systematic errors, because the distribution of the residuals on both sides of zero is random.

The model obtained was validated using LOO and LMO cross-validation processes. For the LOO cross-validation, a data point was removed from the set, and the model was recalculated. The predicted activity for that point was then compared to its actual value. This was repeated until each data point had been omitted once. For the LMO cross-validation, 20% of the data points were removed from the dataset and the model was refitted; the predicted values for those points were then compared to their experimental values. Again, this was repeated until each data point had been omitted once. The cross-validation parameters are shown in Eq. 1. The cross-validated correlation coefficient (Q^2) was 0.784 for LOO and 0.792 for LMO. This confirms that the regression model obtained has good internal and external predictive power. However, the small size of the data set may produce an overfitted model. In order to assess the robustness of the model, a Y-randomization test was applied. The dependent variable vector pIC_{50} was

Table 2 Chemical structures and the corresponding observed and predicted pIC_{50} values obtained by the MLR method

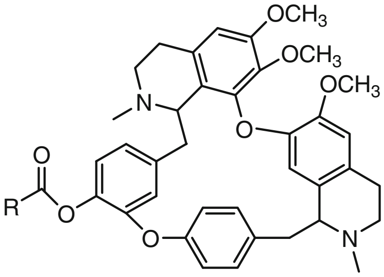
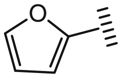
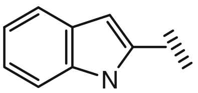
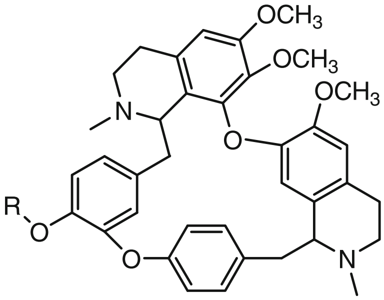
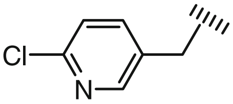
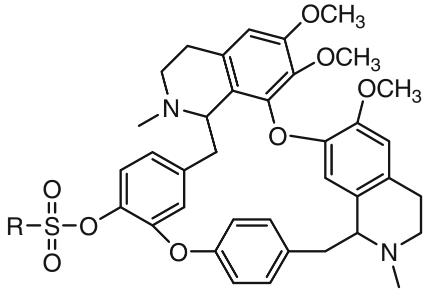
No.	General structure	R	Exp.	Pred.	RE%[b]
1		CH_3	5.18	5.44	5.11
2		$(CH_3)_3C$	6.00	5.90	−1.72
3[a]		$CH_3C{=}CH_3$	5.42	5.58	3.00
4		$ClCH_2$	5.57	5.36	−3.75
5[a]		Ph	5.67	5.79	2.13
6		4-CH_3OPh	6.26	6.25	−0.16
7		3,4,5-$(CH_3O)_3$Ph	6.16	6.17	0.22
8		2-CH_3Ph	6.16	5.95	−3.34
9[a]		4-ClPh	6.01	5.96	−0.90
10		2-ClPh	6.22	6.04	−2.87
11[a]		4-BrPh	6.40	6.33	−1.10
12		2-BrPh	6.11	6.13	0.31
13		3-FPh	5.71	5.67	−0.70
14		4-NO_2Ph	5.67	5.75	1.42
15		3,5-$(NO_2)_2$Ph	5.43	5.71	5.10
16		3-CF_3Ph	6.05	5.87	−2.96
17		3-$ClCH_2$Ph	6.30	6.22	−1.35
18		PhCH_2	5.65	5.72	1.17
19		3,4,6-$(F)_3$PhCH_2	5.52	5.50	−0.41
20			5.56	5.89	5.87
21			6.22	6.18	−0.68
22		H	5.05	5.07	0.43
23		CH_3CH_2	5.50	5.31	−3.45
24		PhCH_2	6.10	6.17	1.21
25[a]		4-BrPhCH_2	6.15	6.17	0.38
26		4-NOPhCH_2	6.44	6.33	−1.66
27		$BrCH_2CH_2CH_2$	5.78	5.83	0.93
28		3,4,5-$(CH_3O)_3$PhCH_2	5.82	5.88	1.09
29		$C_2H_5O(CH_2)_4$	5.39	5.38	−0.20
30			6.40	6.51	1.77

Table 2 continued

No.	General structure	R	Exp.	Pred.	RE%[b]
31[a]	 OCH$_3$ OCH$_3$ OCH$_3$ N O O R–S–O O O N	CH_3	5.32	5.37	0.95
32		Ph	5.70	5.71	0.20

[a] Test set

[b] Relative error percentage

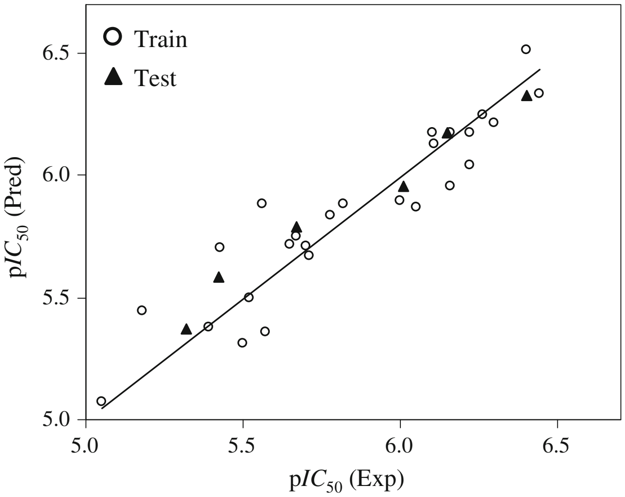


Fig. 1 Predicted versus the experimental pIC_{50} values, obtained by MLR

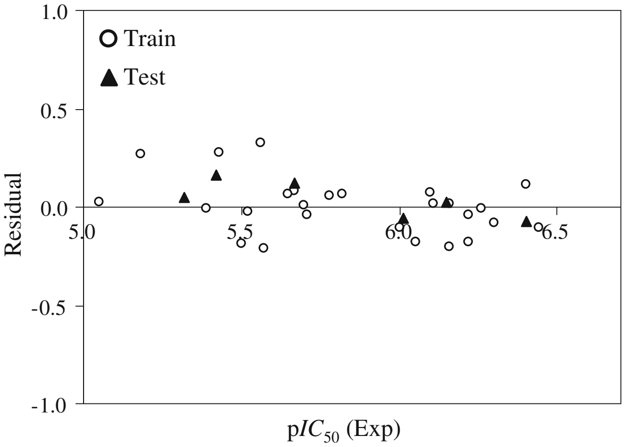


Fig. 2 Residual versus experimental pIC_{50} values, obtained by MLR

Table 3 R^2_{train} and Q^2_{LOO} values after several Y-randomization tests

Iteration	R^2_{train}	Q^2_{LOO}
1	0.077	0.154
2	0.102	0.113
3	0.150	0.047
4	0.027	0.124
5	0.001	0.218
6	0.359	0.046
7	0.047	0.372
8	0.011	0.211
9	0.148	0.089
10	0.066	0.100

randomly shuffled and a new QSAR model was developed using the original variable matrix. The new QSAR model is expected to give low values for R^2_{train} and Q^2_{LOO}. Several random shuffles of the *y*-vector were performed, and the results of this are shown in Table 3. The low R^2_{train} and Q^2_{LOO} values show that the good results in our original model were not due to a chance correlation or a structural dependency of the training set.

Interpretation of descriptors

Besides demonstrating statistical significance, QSAR models should also provide useful chemical insights for drug design. For this reason, an acceptable interpretation of the QSAR results is provided below. By interpreting the descriptors contained in the model, it is possible to gain some insights into factors that are related to the antileukemia activity of berbamine derivatives.

GGI1 (topological charge index of order 1) is one of the topological charge indices that appears in the model. Topological charge indices were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [14, 15]. As is apparent from Eq. 1, the GGI1 coefficient has a positive sign, which indicates that pIC_{50} is directly related to this descriptor.

The second descriptor is Mor17e, which is one of the 3D-MoRSE descriptors. 3D-MoRSE (3D molecule representation of structures based on electron diffraction) descriptors are derived from infrared spectra simulation using a generalized scattering function [16]. This descriptor was proposed as signal 17/weighted by atomic Sanderson electronegativities, which relates to the electronegativity of the molecule. Thus, increasing the electronegativity of the molecule increases its Mor17v value. As can be seen from Eq. 1, the Mor17e descriptor has a negative sign, which indicates that an increase in the electronegativity of the molecule leads to a decrease in its pIC_{50} value.

The third and fourth descriptors are G1v and G3e, which are WHIM descriptors. The WHIM descriptors are based on the statistical indices calculated using the projections of atoms along principal axes. The algorithm consists of performing a principal components analysis on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the atoms. Directional WHIM symmetry descriptors are related to the number of central symmetric atoms (along the *m*th component), the number of unsymmetric atoms, and the total number of atoms of the molecule [16]. The weighting schemes that are used to compute the weighted covariance matrix in these descriptors are atomic van der Waals volumes and atomic Sanderson electronegativities for G1v and G3e. The G1v and G3e descriptors have positive signs, which indicate that pIC_{50} is directly related to these descriptors; therefore, increasing the G1v and G3e values of a molecule increases its pIC_{50} value.

The final descriptor of the GA-MLR model was the H autocorrelation of lag 8/weighted by atomic van der Waals volumes (H8v). This descriptor, a GETAWAY one, is related to the atomic van der Waals volume, the size, and the location of the atom in the molecule. The greater the atomic van der Waals volume, the atom size, and the distance between the atom and the center of the molecule, the greater the descriptor value [16]. This descriptor has a positive influence on the pIC_{50} values.

Summarizing, the charges, electronegativities, and atomic van der Waals volumes of the molecules are the most important influences on the antileukemia activities of berbamine derivatives.

Conclusions

QSAR was applied to the antileukemia activities of berbamine derivatives using the GA-MLR method. The validation procedures utilized in this work (separation of data into independent training and test sets, Y-randomization) illustrated the accuracy and robustness of the QSAR model produced, not only by calculating its fitness for sets of training data but also by testing the predictive ability of the model. Molecular charge, electronegativity, and atomic van der Waals volume were found to be important factors that control the antileukemia activity. The proposed model can identify and aid with the design of novel antileukemia activators.

Data and methodology

Data set

In this study, the data set of 32 berbamine derivatives with antileukemia activity values used for the QSAR analyses was selected from [5]. The antileukemia activity values were expressed as IC_{50} (50% inhibitory concentration) values. Chemical structures and activity data for the complete set of compounds are presented in Table 2. The activity data [IC_{50} (μM)] were converted to the logarithmic scale pIC_{50} (M) [$-\log(IC_{50})$] and then used for subsequent QSAR analyses as the response variables.

Software

A Pentium IV personal computer running the Windows XP operating system was used. Geometry optimization was performed with HyperChem (Version 7.0; Hypercube, Inc.). Dragon 3.0 software was utilized to calculate the molecular descriptors [17]. The GA-MLR regression and other calculations were performed using MATLAB (Version 7.0, Mathworks, Inc.).

Descriptor calculation

The chemical structures of the molecules were built using Hyperchem software. AM1 semi-empirical calculations were used to optimize the 3D geometries of the molecules [18, 19]. To ensure that the resulting structures had optimum geometry, optimization was repeated many times with different starting geometries. The optimization was preceded by the application of the Polak–Ribiere algorithm to reach a root mean square gradient of 0.01. A wide variety of QSAR descriptors have been reported in the literature [20–22]. There has recently been an increase in the use of theoretical descriptors in QSAR studies. Dragon software was used to calculate the descriptors in this research, and a total of 1,497 molecular descriptors belonging to 18 different types of theoretical descriptors were calculated for each molecule.

The calculated descriptors were first analyzed for the existence of constant or near-constant variables, and those detected were removed. In addition, to decrease the redundancy present in the descriptor data matrix, the correlations of the descriptors with each other and with the activities ($\mathrm{p}IC_{50}$) of the molecules were examined, and collinear descriptors (i.e., $r > 0.9$) were detected. Among the collinear descriptors, the one that presented the highest correlation with the activity ($\mathrm{p}IC_{50}$) was retained while the others were removed from the data matrix. The remaining descriptors were then collected in an $n \times m$ data matrix, where $n = 32$ and $m = 415$ are the numbers of compounds and descriptors, respectively.

Multiple linear regression analysis

The multiple linear regression method (MLR) is one of the modeling methods most commonly used in QSAR. MLR regression, a linear technique that can determine the relative importance of descriptors, is usually used to generate QSAR models. The MLR method provides an equation that links the structural features to the $\mathrm{p}IC_{50}$ values of the compounds:

$$\mathrm{p}IC_{50} = a_0 + a_\mathrm{i}d_\mathrm{i} + \cdots + a_\mathrm{n}d_\mathrm{n}, \quad (2)$$

where the intercept (a_0) and the regression coefficients of the descriptors (a_i) are determined using the least-squares method. The d_i are descriptors; the elements of this vector are numerical values for the 3D structure of the molecule.

Genetic algorithm

GAs [23–26] are inspired by biological evolution. When a GA is applied to a subset selection problem, we need a regression method to extract the relationship between the selected descriptors and the dependent variables. Many multiple or multivariate regression methods, such as MLR, principal component regression (PCR), and partial least squares (PLS), can be used for this purpose [27–30]. In this research, GA-MLR was used for variable selection. To select the most relevant descriptors, the evolution of the population was simulated. Each individual of the population, as defined by a chromosome of binary values, represented a subset of descriptors. The number of genes in each chromosome was equal to the number of descriptors. The population of the first generation was selected randomly. A gene took a value of 1 if its corresponding descriptor was included in the subset; otherwise it took a value of zero. The number of genes with a value of 1 was kept relatively low in order to get a small subset of descriptors; that is, the probability of generating 0 for a gene was set to be greater (at least 60%) than 1. The operators used here were crossover and mutation. The probabilities that these operators were used were varied linearly with generation renewal (0–0.1% for mutation and 60–90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations had the same fitness.

References

1. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL (2001) N Engl J Med 344:1031
2. Le Coutre P, Tassi E, Varella-Garcia M, Barni R, Mologni L, Cabrita G, Marchesi E, Supino R, Gambacorti-Passerini C (2000) Blood 95:1758
3. Xu R, Dong Q, Yu Y, Zhao X, Gan X, Wu D, Lu Q, Xu X, Yu X (2006) Leuk Res 30:17
4. Zhao X, He Z, Wu D, Xu R (2007) Chin Med J 120:802
5. Xie J, Ma T, Gu Y, Zhang X, Qiu X, Zhang L, Xu R, Yu Y (2009) Eur J Med Chem 44:3293
6. Hill DL (1972) The biochemistry and physiology of tetrahymena, 1st edn. Academic, New York, p 230
7. Deeb O, Hemmateenejad B (2007) Chem Biol Drug Des 70:19
8. Verma RP, Kurup A, Hansch C (2005) Bioorg Med Chem 13:237
9. Agrawal VK, Khadikar PV (2001) Bioorg Med Chem 9:3035
10. Sammes PG, Taylor JB (1990) Comprehensive medicinal chemistry, vo1 4. Pergamon, Oxford, p 766
11. Depczynski U, Frost VJ, Molt K (2000) Anal Chim Acta 420:217
12. Alsberg BK, Marchand-Geneste N, King RD (2000) Chemom Intell Lab Syst 54:75
13. Jouanrimbaud D, Massart DL, Leardi R, Denoord OE (1995) Anal Chem 67:4295
14. Gilvez J, Garcia R, Salabert MT, Soler R (1994) J Chem lnf Comput Sci 34:520
15. Gilvez J, Garcia-Domenech R, De Juliin-Ortiz V, Soler R (1995) J Chem lnf Comput Sci 35:272
16. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
17. Todeschini R, Consonni V, Pavana M (2003) Dragon 3.0 software. http://www.disat.unimib.it/chm/
18. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
19. Dewar MJS, Dieter K (1986) J Am Chem Soc 108:8075
20. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. RSP-Wiley, Chichester
21. Kostantinova EV (1996) J Chem Inf Comp Sci 36:54
22. Lucasius CB, Kateman G (1993) Chemom Intell Lab Syst 19:1
23. Lucasius CB, Kateman G (1994) Chemom Intell Lab Syst 25:99
24. Hibbert DB (1993) Chemom Intell Lab Syst 19:277
25. Leardi R, Boggia R, Terrile M (1992) J Chemom 6:267
26. Leardi R (1994) J Chemom 8:65
27. Khajehsharifi H, Pourbasheer E (2008) J Chin Chem Soc 55:163
28. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2008) Monatsh Chem 139:1423
29. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2008) Bull Korean Chem Soc 29:833
30. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2009) Monatsh Chem 140:15