Taylor & Francis
Taylor & Francis Group

SAR and QSAR in Environmental Research

ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: www.tandfonline.com/journals/gsar20

A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm

A.M. Al-Fakih, Z.Y. Algamal, M.H. Lee, M. Aziz & H.T.M. Ali

To cite this article: A.M. Al-Fakih, Z.Y. Algamal, M.H. Lee, M. Aziz & H.T.M. Ali (2019) A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm, SAR and QSAR in Environmental Research, 30:6, 403-416, DOI: 10.1080/1062936X.2019.1607899

To link to this article: https://doi.org/10.1080/1062936X.2019.1607899

View supplementary material

Published online: 24 May 2019.

Submit your article to this journal

Article views: 181

View related articles

View Crossmark data

Citing articles: 3 View citing articles

SAR AND QSAR IN ENVIRONMENTAL RESEARCH
2019, VOL. 30, NO. 6, 403–416
https://doi.org/10.1080/1062936X.2019.1607899

A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm

A.M. Al-Fakih[a,b], Z.Y. Algamal[c], M.H. Lee[d], M. Aziz[a,e] and H.T.M. Ali[f]

[a]Department of Chemistry, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia; [b]Department of Chemistry, Faculty of Science, Sana'a University, Sana'a, Yemen; [c]Department of Statistics and Informatics, University of Mosul, Mosul, Iraq; [d]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia; [e]Advanced Membrane Technology Centre, Universiti Teknologi Malaysia, Johor, Malaysia; [f]College of Computers and Information Technology, Nawroz University, Kurdistan region, Iraq

ABSTRACT

Time-varying binary gravitational search algorithm (TVBGSA) is proposed for predicting antidiabetic activity of 134 dipeptidyl peptidase-IV (DPP-IV) inhibitors. To improve the performance of the binary gravitational search algorithm (BGSA) method, we propose a dynamic time-varying transfer function. A new control parameter, μ, is added in the original transfer function as a time-varying variable. The TVBGSA-based model was internally and externally validated based on Q^2_{int}, Q^2_{LGO}, Q^2_{Boot}, $\mathrm{MSE}_{\mathrm{train}}$, Q^2_{ext}, $\mathrm{MSE}_{\mathrm{test}}$, Y-randomization test, and applicability domain evaluation. The validation results indicate that the proposed TVBGSA model is robust and not due to chance correlation. The descriptor selection and prediction performance of TVBGSA outperform BGSA method. TVBGSA shows higher Q^2_{int} of 0.957, Q^2_{LGO} of 0.951, Q^2_{Boot} of 0.954, Q^2_{ext} of 0.938, and lower $\mathrm{MSE}_{\mathrm{train}}$ and $\mathrm{MSE}_{\mathrm{test}}$ compared to obtained results by BGSA, indicating the best prediction performance of the proposed TVBGSA model. The results clearly reveal that the proposed TVBGSA method is useful for constructing reliable and robust QSARs for predicting antidiabetic activity of DPP-IV inhibitors prior to designing and experimental synthesizing of new DPP-IV inhibitors.

ARTICLE HISTORY
Received 29 January 2019
Accepted 11 April 2019

KEYWORDS
DPP-IV; type 2 diabetes; antidiabetic; BGS algorithm; time-varying transfer function

Introduction

Diabetes mellitus is one of the major degenerative diseases in the twenty-first century [1]. According to the World Health Organization (WHO), an estimated 1.6 million deaths in 2016 were directly caused by diabetes. A number of 422 million adults were estimated with diabetes in 2014. The WHO estimates that diabetes was the seventh major cause of death in 2016 [2]. There are three main types of diabetes mellitus, i.e. type 1, type 2 and gestational diabetes. Type 2 diabetes accounts for over 90% of all diabetic cases globally [1,3–5]. It is one

CONTACT A.M. Al-Fakih ✉ aalfakih2011@gmail.com; M. Aziz ✉ madzlan@utm.my

This article has been republished with minor changes. These changes do not impact the academic content of the article.
Supplementary material for this article can be accessed at: https://doi.org/10.1080/1062936X.2019.1607899.

of a group of metabolic disorders with multiple aetiologies. It is a common disease and known as a non-insulin-dependent diabetes mellitus. It is caused by insulin resistance, inadequate secretion of insulin, chronic hyperglycaemia, increased hepatic glucose production, or glucose intolerance [1,6–8].

Glucagon-like peptide-1 (GLP-1) has an important role in insulin secretion. GLP-1 is released from L-cells of the small intestine as a response to food digestion. Higher activity of GLP-1 results in sustained insulin secretion which regulates an elevated glucose level. In addition, GLP-1 retards gastric emptying, induction of satiety and stimulation, regeneration and differentiation of islet β-cells. Dipeptidyl peptidase-IV (DPP-IV) serine protease presents in many tissues and body fluids, and exists either with membrane bound or soluble enzyme. DPP-IV degrades GLP-1 (GLP-1 [7–36] amide) into inactive GLP [9–36] amide at N-terminus position. Therefore, the inhibition of DPP-IV helps to increase GLP-1 concentration, which leads to an increase in insulin secretion, and thus ameliorate hyperglycaemia in type 2 diabetes [6–12]. Intensive attention has been given to DPP-IV as an important target for the treatment of type 2 diabetes [9,10,12,13].

Several potent antidiabetic drugs including sitagliptin, vildaglipin, saxagliptin and alogliptin are used. They act by inhibiting activity of DPP-IV, improving insulin sensitivity, and repairing β-cells [3,9,14]. Drug design is strongly associated with quantitative structure–activity relationship (QSAR) modelling [15]. QSAR methods have been efficiently applied for drug design by constructing possible relationship between the desired activity of a series of compounds and their molecular structures. Several modelling procedures such as regression and classification can be used in constructing efficient QSAR models for drug design [16,17].

QSAR modelling studies have been conducted and proposed on several datasets of DPP-IV inhibitors. Sharma et al. [14] developed QSAR models using 3D-QSAR comparative molecular field analysis (CoMFA) and comparative molecular similarity indice analysis (CoMSIA) on a set of trifluorophenyl derivatives as DPP-IV inhibitors. According to their results, the models which were developed using structure-based alignment were significant and showed good predictive value with r^2 of 0.963 and 0.934 for CoMFA and CoMSIA, respectively. It was concluded that the findings are useful in designing novel DPP-IV inhibitors. Jiang et al. [10] carried out 3D-QSAR studies on a set of arylmethylamines as DPP-IV inhibitors using CoMFA approach. According to their results, the best-developed model showed r^2 of 0.953 and it might provide useful pharmacophoric features information for designing new potent DPP-IV inhibitors. Patel and Ghate [12] carried out 3D-QSAR analyses using CoMFA and CoMSIA on 36 quinoline and isoquinoline derivatives as DPP-IV inhibitors. The aim of the study was to identify the responsible structural features for antidiabetic activity. According to their results, the best model showed conventional coefficients (r^2) of 0.991 and 0.983, and predicted correlation coefficients (r^2_{pred}) of 0.874 and 0.847 for CoMFA and CoMSIA, respectively. It was concluded that the findings could be useful in designing and predicting new potent DPP-IV inhibitors. Saqib and Siddiqi [13] carried out 3D-QSAR analyses on 45 triazolopiperazine amide derivatives as DPP-IV inhibitors using CoMFA and CoMSIA. CoMFA and CoMSIA contour maps were also used to analyse the structural features of the ligands. According to their results, models with good predictive abilities were developed. The models showed r^2

of 0.868 for both CoMFA and CoMSIA methods, and r^2_{pred} of 0.816 and 0.863 for CoMFA and CoMSIA, respectively. It was concluded that the information obtained from CoMFA and CoMSIA three-dimensional contour maps could be used in designing new antidiabetic agents.

Other QSAR methods have been used for modelling activities of several datasets of DPP-IV inhibitors. Abd El-Karim et al. [7] performed 2D and 3D QSAR modelling using auto QSAR of Schrödinger, QuaSAR of MOE and 3D Field-based QSAR of Schrödinger on a new dataset of tetralin-sulfonamide derivatives as DPP-IV inhibitors. According to their results, the 3D model and the auto QSAR-based 2D QSAR model showed r^2 of 0.920 and 0.890, respectively. The QuaSAR-based 2D model showed r^2 of 0.989. It was concluded that the models revealed high predictive power with a good agreement between the predicted and observed pIC_{50}values. Amini et al. [8] developed QSAR models using multiple linear regression (MLR) and Levenberg Marquardt artificial neural network (LM-ANN) on 33 aminomethyl-piperidones compounds as DPP-IV inhibitors. According to their results, LM-ANN model showed better performance with r value of 0.983 and 0.966 for training and test sets, respectively. It was concluded that the results showed a close agreement between experimental and predicted values of pIC_{50}. Paliwal et al. [11] conducted QSAR analysis on 47 pyrrolidine analogues as DPP-IV inhibitors using MLR and PLS methods. According to their results, the models developed using MLR and PLS methods were robust with r^2(CV) of 0.84 and 0.82, respectively. It was concluded that this analysis helped to find out the role of some descriptors in determining the activity of DPP-IV inhibitors.

In the present study, a dataset of 134 DPP-IV inhibitors was used to carry out our proposed QSAR modelling procedure. We propose a time-varying transfer function to enhance the performance of the binary gravitational search algorithm (BGSA). Our proposed time-varying binary gravitational search algorithm (TVBGSA) modelling method was used in this work to model and predict the activity of the studied dataset as DPP-IV inhibitors. Several evaluation criteria were used and discussed to investigate the performance of our proposed method.

Gravitational search algorithm (GSA)

GSA was first introduced by Rashedi et al. [18]. In GSA, individuals are treated as objects with masses. According to the Newton’s law of gravity and motion, the attraction between the objects is caused by the gravity force, and they move towards the heavier masses objects [18–22]. There are four characteristics of each GSA mass, i.e. position, inertial mass, active gravitational mass, and passive gravitational mass. Position corresponds to the problem solution, while a fitness function is used to determine the other three characteristics. The GSA details are summarized as follows [19,23]:

First, for a system of *N*masses (agents) where the position of *i*th mass is defined by

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots, x_i^n\right), \qquad i = 1, 2, \ldots, N \tag{1}$$

where x_i^d is the ith agent's position in the dth dimension, and n is the search space dimension. Second, the gravitational force ($F_{ij}^d(t)$) which acting on mass i from mass j at time t is defined by

$$F_{ij}^d(t) = G(t)\frac{M_{pi}(t).M_{aj}(t)}{R_{ij}(t)+\varepsilon}\left(x_j^d(t) - x_i^d(t)\right) \tag{2}$$

where M_{aj} is the active gravitational mass related to agent$j$, M_{pi} is the passive gravitational mass related to agent i, $G(t)$ is gravitational constant at time t, ε is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j and defined by

$$R_{ij}(t) = \left\|X_i(t), X_j(t)\right\|_2 \tag{3}$$

Third, for computing acceleration of an agent i, total forces is defined by

$$F_i^d(t) = \sum_{j=1, j\neq i}^{N} rand_j F_{ij}^d(t) \tag{4}$$

where $rand_j$ is a random number in the interval $[0, 1]$. Fourth, according to the total forces, the acceleration of the agent i at time t, and in dth direction, is defined by

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \tag{5}$$

where M_{ii} is the inertial mass of ith agent. Fifth, the next velocity of an agent is computed as a fraction of its present velocity added to its acceleration. The new velocity and position of agent i are given by Equations (6) and (7), respectively:

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \tag{6}$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \tag{7}$$

where $v_i^d(t)$ and $x_i^d(t)$ are the velocity and position in dth dimension of agent i at time t, respectively, and $rand_i$ is an uniform random variable in the interval $[0, 1]$ which adds a randomized characteristic to the search. Finally, after computing the fitness of current population, both gravitational mass and inertial mass were updated by Rashedi et al. [18].

Furthermore, for balancing the GSA's exploration and exploitation, a k_{best} agent is employed. The k_{best} agent is a function of time. It has an initial value, k_0, at the beginning and decreasing with time. In GSA, k_0 is normally set to the total number of agents, N, and k_{best} decreases linearly. Finally, only one agent will apply force to the other agents. Therefore, Equation (4) is modified as follows:

$$F_i^d(t) = \sum_{j\in k_{best}, j\neq i}^{N} rand_j F_{ij}^d(t) \tag{8}$$

where k_{best} is the set that includes the first k agents which have the best fitness value and biggest mass. At the same time, an initial G_0 value is specified to the gravitational constant, G, and will decrease with time as defined by

$$G(t) = G(G_0, \mathrm{t}) \tag{9}$$

Binary gravitational search algorithm (BGSA)

A binary GSA was proposed by Rashedi et al. [24]. BGSA was proposed to solve problems in continuous valued space [25,26]. Some GSA's basic concepts were modified. Each dimension in discrete binary takes only 0 or 1. Moving through a dimension represents that the corresponding variable value varies from 0 to 1 or vice versa. For introducing a binary version of the gravitational algorithm, the force, acceleration and velocity updating procedures may be considered similar to the continuous algorithm Equations (5)–(7). The main difference between continuous and binary GSA is that in the binary algorithm the position updating is a switching between 0 and 1 according to the mass velocity. This is to update the position in which the current bit value is changed with a calculated probability according to the mass velocity. Therefore, in BGSA, the velocity is updated based on Equation (6) and the new position is considered to be either 1 or 0 with the given probability [24,27,28].

Prior to define a transfer function to map the velocity to the position updating probability, these are some basic concepts of GSA to be taken into account. First, a large absolute value of the velocity indicates a not proper current position of the mass which required a great movement to reach the optimum position. Second, a small absolute value of the velocity specifies a close current position of the mass to the optimum position and a small distance is needed to reach the optimum position. Then, when the optimal solution is found, the velocity becomes zero. Therefore, the following concepts should be considered in order to implement the BGSA. First, a large absolute value of velocity must result in a high probability of a mass position change in reference to its previous position which is from 0 to 1 or vice versa. Second, a small absolute value of the velocity must result in a small probability of the position change. For example, a zero value of the velocity shows a good mass position and must not be changed [24].

The proposed method

In BGSA, a proper probability function should be defined based on the mentioned concepts. For a small $\left|v_i^d\right|$, the changing probability, x_i^d, must be near zero and it must be high for a large $\left|v_i^d\right|$. The $S\left(v_i^d\right)$ function was defined to transfer v_i^d into a probability function. The $S\left(v_i^d\right)$ function is bounded within interval $[0, 1]$ and increases with increasing $\left|v_i^d\right|$. The $S\left(v_i^d\right)$ was defined according to the following equation:

$$S\left(v_i^d(t)\right) = \left|\tanh\left(v_i^d(t)\right)\right| \tag{10}$$

When the $S\left(v_i^d\right)$ function is calculated, the agents' movement will be according to the rule specified in the following equation:

$$\begin{aligned} &\text{if } rand < S\left(v_i^d(t+1)\right) \text{ then } x_i^d(t+1) = \text{complement}\left(x_i^d(t)\right) \\ &\text{else } x_i^d(t+1) = x_i^d(t) \end{aligned} \tag{11}$$

In order to achieve a good converge rate, velocity was limited, $\left|v_i^d\right| < v_{\max}$.

In optimization algorithm, it is expected that the focus of the early stages of the implementation the algorithm will be on exploration to avoid falling into the local point, but in later stages of implementation, the algorithm focuses more on exploitation to improve the quality of the solution [29,30]. As in Islam et al. [29] and Mafarja et al. [30], in this paper, a dynamic transfer function is proposed to improve the BGSA. In our proposed time-varying transfer function, a new control parameter μ is added in the original transfer function. This μ is a time-varying variable which starts with a large value and gradually decreases over time. The proposed μ is defined as

$$\mu = \frac{\mu_{\max}}{t^2} \tag{12}$$

where $\mu_{\max}$ is the maximum value of the control parameter μ. Accordingly, the proposed transfer function is defined as

$$S_{TV}\left(v_i^d(t)\right) = \left|\tanh\left(\frac{v_i^d(t)}{\mu}\right)\right| \tag{13}$$

It is obvious that the proposed function can return the higher probability than the original transfer function for the same value of velocity. Additionally, the proposed function can converge to be a vertical line when iteration increasing.

Experimental

Dataset

A dataset of 134 compounds (DPP-IV inhibitors) and their antidiabetic activity, IC_{50} (nM), were collected from the literature [14,31]. The inhibition of DPP-IV by the studied compounds was used as the endpoint to determine the antidiabetic activity of these compounds, measured by the concentration of the compound, IC_{50}, that caused 50% inhibition of DPP-IV. The IC_{50} activities were converted into their corresponding pIC_{50} (the logarithm of reciprocal of IC_{50}). The pIC_{50} values of the compounds used ranged from 5.009 to 8.959 with 45 compounds in (5.009–6.000), 45 compounds in (6.066–7.000), and 44 in (7.009–8.959). The distribution of pIC_{50} values over the collected dataset illustrates that the dataset is unbiased with respect to the dependent variable (biological endpoint) and thus should not lead to data-biased QSAR models. The data were randomly divided into 94 compounds (70%) as a training dataset and 40 compounds (30%) as a test dataset. The pIC_{50} values of the training and test compounds are distributed among the range of whole dataset's pIC_{50}. The compounds' structures and pIC_{50} values are given in Table S1 (Supplementary material). The training dataset was used for constructing the QSAR model and the test dataset was used for the evaluation of the model's performance, based on several evaluation criteria.

Molecular descriptor calculation

The molecular structures of the compounds were sketched using Chem3D software (CambridgeSoft Corporation, Cambridge, MA, USA). The structures were optimized using the molecular mechanics (MM2) method implemented in Chem3D software, and then using the molecular orbital package (MOPAC) module implemented in the same Chem3D software. DRAGON software (version 6.0) was used to generate 4885 molecular descriptors based on the optimized molecular structures [32]. To include consistent and useful descriptors, preprocessing steps were performed as follows. First, descriptors that had constant or zero values for all compounds were excluded. Second, the remaining descriptors were further refined by removing those in which 70% of their values were zeros. After that, descriptors with a relative standard deviation of less than 0.001 were removed. In addition, the correlation of the remaining descriptors was examined to omit multicollinearity by removing those that were highly correlated ($r_{ij} \geq 0.90$). Finally, 1048 descriptors remained for constructing the QSAR model.

Evaluation criteria

Several evaluation criteria were performed to provide a satisfactory evaluation of the used modelling methods in constructing an efficient QSAR model. The used criteria for the training dataset were mean-squared error of the training dataset (MSE_{train}) and leave-one-out internal validation ($Q^2{}_{\text{int}}$), which are defined by

$$MSE_{\text{train}} = \frac{\sum_{i=1}^{n_{train}} \left(y_{i,train} - \hat{y}_{i,train}\right)^2}{n_{train}} \tag{14}$$

and

$$Q^2{}_{\text{int}} = 1 - \left[\frac{\sum_{i=1}^{n_{train}} \left(y_{i,train} - \hat{y}_{i,train}\right)^2}{\sum_{i=1}^{n_{train}} \left(y_{i,train} - \bar{y}\right)^2}\right] \tag{15}$$

respectively.

The test dataset was used to validate the models by computing mean-squared error of the test dataset (MSE_{test}) and the external validation ($Q^2{}_{\text{ext}}$), which are defined by

$$MSE_{\text{test}} = \frac{\sum_{i=1}^{n_{test}} \left(y_{i,test} - \hat{y}_{i,test}\right)^2}{n_{test}} \tag{16}$$

and

$$Q^2{}_{\text{ext}} = 1 - \left[\frac{\sum_{i=1}^{n_{test}} \left(y_{i,test} - \hat{y}_{i,test}\right)^2}{\sum_{i=1}^{n_{test}} \left(y_{i,test} - \bar{y}_{train}\right)^2}\right] \tag{17}$$

respectively, where n_{train} and n_{test} represent the training and test sample sizes, the $y_{i,train}$, $y_{i,test}$, $\hat{y}_{i,train}$, and $\hat{y}_{i,test}$ stand for the pIC_{50} values of the training dataset, test dataset, and their corresponding predicted pIC_{50} values, while $\bar{y}$ and $\bar{y}_{train}$ represent the mean of all the pIC_{50} values and the mean of the training pIC_{50} values, respectively.

Experimental setting

There are five control parameters in our proposed time-varying transfer function: the number of agents (n_a) (population size), the maximum number of iterations (t_{max}), the maximum value of the control parameter μ of Equation (12), the gravitational constant (G_0), and the descending coefficient (a) of the $G(t) = G_0 \times \exp(-a \ \times t / t_{max})$. The specific parameter values are outlined in Table 1. The position for each agent is a vector of 0 and 1 values with size equals the number of the descriptors. Initially, the positions were randomly generated from a uniform distribution between 0 and 1. Further, the best fitness function that can combine the minimum error and the minimum number of selected descriptors is preferable. The fitness function used in BGSA to evaluate each agent position is defined as

$$\text{fitness} = 0.9 \times \text{MSE} + 0.1 \times \left(\frac{d - \tilde{d}}{d} \right) \tag{18}$$

where MSE is the prediction error obtained, d represents the number of descriptors in the dataset, and $\tilde{d}$ represents the number of selected descriptors.

Experimental results

Prediction evaluation

To demonstrate the usefulness of the proposed TVBGSA method (Equation (13)), comparative experiments with the original BGSA (Equation (10)) was carried out. The results of the prediction evaluation of the constructed QSAR models using TVBGSA and BGSA are listed in Table 2. From Table 2, it is obvious that the performance of the TVBGSA is much better compared to the BGSA, in terms of the number of selected descriptors. The proposed method, TVBGSA, selects five descriptors out of 1048 descriptors compared with nine selected descriptors by BGSA. The names of the selected descriptors and their descriptions for both methods are presented in Table 3.

Table 1. Parameters setting for BGSA.

Parameter	*Value*
n_a	50
t_{max}	1000
G_0	1
a AP	0.5
μ_{max} in Equation (12)	3

Table 2. Prediction evaluation criteria values for the training and test datasets.

		Training set				*Testing set*	
Methods	*No. of descriptors*	MSE_{train}	Q^2_{int}	Q^2_{LGO}	Q^2_{Boot}	MSE_{test}	Q^2_{ext}
TVBGSA	5	3.571	0.957	0.951	0.954	4.306	0.938
BGSA	9	6.691	0.923	0.918	0.915	7.328	0.912

Table 3. Names of the selected descriptors and their descriptions.

Method	*Descriptor name*	*Group type*	*Description*
TVBGSA	MATS4p	2D autocorrelations	Moran autocorrelation of lag 4 weighted by polarizability
	JGI3	2D autocorrelations	Mean topological charge index of order 3
	Mor07m	3D-MoRSE descriptors	Signal 07/weighted by mass
	SpMAD_EA(bo)	Edge adjacency indices	Spectral mean absolute deviation from edge adjacency mat. weighted by bond order
	RDF110p	RDF descriptors	Radial Distribution Function – 110/weighted by polarizability
BGSA	P_VSA_MR_2	P_VSA-like descriptors	P_VSA-like on Molar Refractivity, bin 2
	SpMaxA_EA	Edge adjacency indices	Normalized leading eigenvalue from edge adjacency mat.
	MW	Constitutional indices	Molecular weight
	JGI3	2D autocorrelations	Mean topological charge index of order 3
	Mor07m	3D-MoRSE descriptors	Signal 07/weighted by mass
	SpMAD_EA(bo)	Edge adjacency indices	Spectral mean absolute deviation from edge adjacency mat. weighted by bond order
	SdO	Atom-type E-state indices	Sum of dO E-states
	Mor24p	3D-MoRSE descriptors	Signal 24/weighted by polarizability
	Eig09_AEA(bo)	Edge adjacency indices	Eigenvalue n. 9 from augmented edge adjacency mat. weighted by bond order

It is clear from Table 2 that the TVBGSA is superior to BGSA in terms of prediction performance for the training data. The TVBGSA yields higher Q^2_{int} of 0.957, Q^2_{LGO} of 0.951 and Q^2_{Boot} of 0.954, and lower $\text{MSE}_{\text{train}}$, compared to the obtained by BGSA. In addition, depending on the test dataset, TVBGSA reveals greater value for Q^2_{ext} of 0.938 and lesser MSE_{test} value compared to the obtained by BGSA. The enhancement of TVBGSA, in terms of MSE_{test}, over BGSA is 41.23%. Furthermore, for the test dataset, the predictive ability of TVBGSA (Q^2_{ext}= 0.938), is better than the obtained value of 0.912 by BGSA. Overall, the results clearly demonstrate that the proposed TVBGSA is effective in modelling high-dimensional QSARs for antidiabetic activity of DPP-IV inhibitors. The TVBGSA not only improved the prediction performance, but also identified a smaller subset of descriptors compared with BGSA.

Descriptor interpretation

The interpretation of the selected descriptors by the TVBGSA method gives an insight into the related physicochemical properties. MATS4p (Moran autocorrelation of lag 4 weighted by polarizability) and JGI3 (mean topological charge index of order 3) belong to 2D autocorrelation descriptors. These descriptors are based on the molecular topology with the consideration of chemical information by specified weights of the molecule atoms [33]. The MATS4p descriptor is related to polarizability property of molecule atoms. The JGI3 descriptor belongs to the topological charge indices which were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [34]. The Mor07m descriptor (signal 07/weighted by mass) belongs to 3D-MoRSE descriptors (3D-Molecule representation of structures based on electron diffraction). The 3D-MoRSE descriptors are geometrical descriptors that were derived from the 3D spatial coordinates of a molecule. The 3D-MoRSE code considers a molecular transform, a proposal based on electron diffraction studies that were used to prepare theoretical scattering curves

[34,35]. The SpMAD_EA(bo) descriptor (spectral mean absolute deviation from edge adjacency mat. weighted by bond order) belongs to edge adjacency indices which encode information about the connectivity between graph edges [34]. The RDF110p descriptor (Radial Distribution Function – 110/weighted by polarizability) belongs to RDF descriptors. These descriptors are calculated from the radial distribution function of an ensemble of a number of atoms that can be interpreted as the probability distribution of finding an atom in a spherical volume [34,35].

Y-randomization test

The TVBGSA model was further validated by applying the Y-randomization test [36]. It was applied in order to ensure that the predictive power of the TVBGSA model was not based on chance. This test randomly shuffled the pIC_{50} values several times and applied TVBGSA each time. Then, $Q^2{}_{\text{int}}$ was calculated each time. If all the obtained values were less than the true $Q^2{}_{\text{int}}$ value of the constructed QSAR model by TVBGSA, then the constructed QSAR model was not due to chance correlation, indicating that the proposed TVBGSA could lead to an acceptable method using the training dataset. Figure 1 shows the results of the Y-randomization test for 500 times of $Q^2{}_{\text{int}}$ values. It is clear from Figure 1 that the $Q^2{}_{\text{int}}$ values are in the range of 0.0525 to 0.3991. In comparison to the true $Q^2{}_{\text{int}}$ value of TVBGSA ($Q^2{}_{\text{int}} = 0.957$), these values indicate that the QSAR model of the antidiabetic activity (pIC_{50}) of DPP-IV inhibitors by TVBGSA is not due to chance correlation or structural dependence of the training dataset.

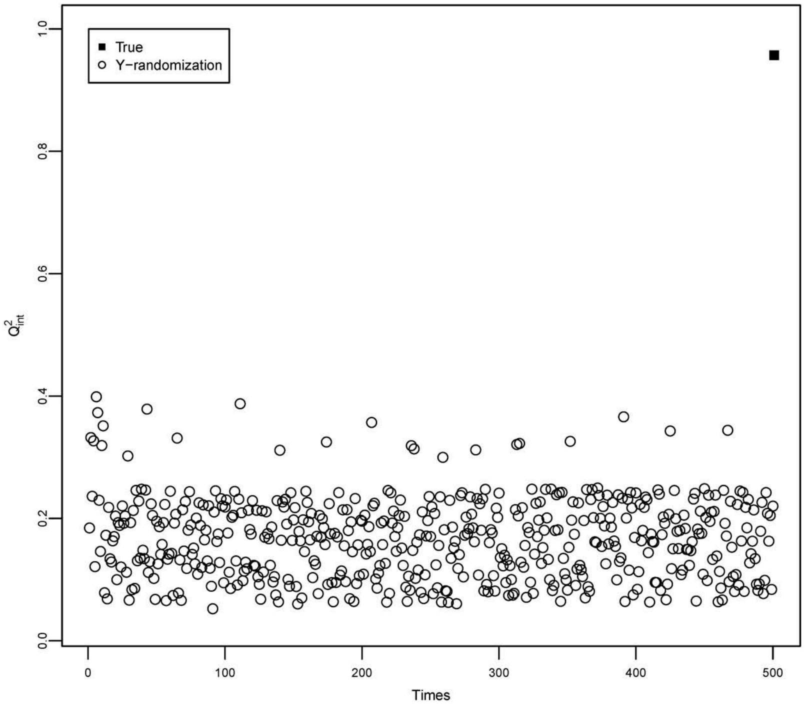

Figure 1. Y-randomization test for TVBGSA over 500 times.

Robustness performance

To further evaluate the ability of TVBGSA to construct a robust QSAR model, the leverage approach was used as an applicability domain (AD) assessment. AD is defined as 'a theoretical region in chemical space, defined by the model descriptors and modelled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors' [37]. Figure 2 displays the Williams plot of the leverage values against the standardized residuals for each compound for the TVBGSA model (the dotted line indicates the leverage threshold, while the dashed line represents the standardized residual limits). The influential compound can be detected when its leverage value is greater than the leverage threshold ($h^* = 3(p+1)/n$) where p is the number of the selected descriptors in the final QSAR model, and n represents the number of compounds. It is obvious from Figure 2 that no compounds have a standardized residual higher than the limit ±3, which can be considered to be outliers, or with a high leverage value. Thus, as it is clearly demonstrated in Figure 2, the results confirm that the QSAR model constructed using TVBGSA is reliable and robust.

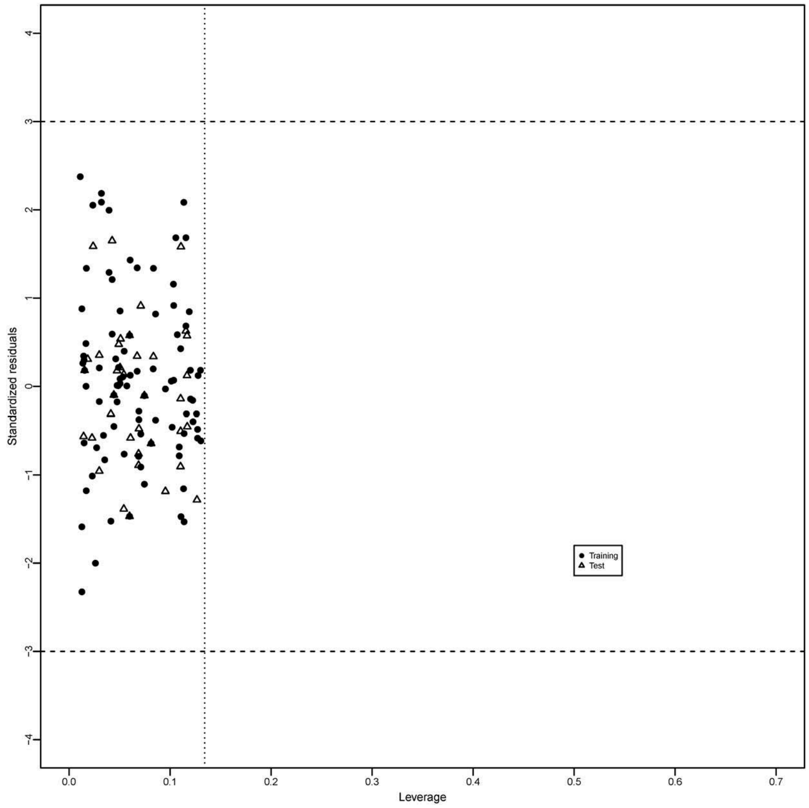

Figure 2. Williams plot for the training and test datasets of TVBGSA.

Conclusion

TVBGSA modelling method was proposed for the prediction of the antidiabetic activity (pIC_{50}) of DPP-IV inhibitors. The results derived from the internal validation criteria, i.e. Q^2_{int}, Q^2_{LGO}, Q^2_{Boot} and MSE_{train} of the training dataset and the external validation criteria, i.e. Q^2_{ext} and MSE_{test} of the test dataset prove the better predictive power of TVBGSA model compared to BGSA. In addition, the results obtained by the AD assessment and Y-randomization test confirm that TVBGSA model is reliable, robust and not due to chance correlation. In conclusion, the current study proposes TVBGSA as a useful modelling approach to be used for constructing reliable and robust QSARs for predicting antidiabetic activity of DPP-IV inhibitors prior to designing and experimental synthesizing new DPP-IV inhibitors. Although TVBGSA results yielded significantly better performance, it has a limitation. Its performance is fully depending on the maximum value of the control parameter.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Ministry of Education Malaysia through the Fundamental Research Grant Scheme and the Universiti Teknologi Malaysia through the Professional Development Research University grant number [R.J130000.7113.04E93].

ORCID

Z.Y. Algamal http://orcid.org/0000-0002-0229-7958
M.H. Lee http://orcid.org/0000-0002-3700-2363

References

[1] G.A.R.Y. Suaifan, M.B. Shehadeh, R.M. Darwish, H. Al-Ijel, and V. Abbate, *Design, synthesis and in vivo evaluation of novel glycosylated sulfonylureas as antihyperglycemic agents*, Molecules 20 (2015), pp. 20063–20078. doi:10.3390/molecules201119676.

[2] World Health Organization, *Global Health Estimates: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2016*, World Health Organization, Geneva, 2018. Available at https://www.who.int/en/news-room/fact-sheets/detail/diabetes. (Accessed on 21 March 2019).

[3] K. Lin, Z. Cai, F. Wang, W. Zhang, and W. Zhou, *Synthesis and biological evaluation of xanthine derivatives on dipeptidyl peptidase 4*, Chem. Pharm. Bull. 61 (2013), pp. 477–482. doi:10.1248/cpb.c12-01046.

[4] P. Zimmet, K.G.M.M. Alberti, and J. Shaw, *Global and societal implications of the diabetes epidemic*, Nature 414 (2001), pp. 782–787. doi:10.1038/414782a.

[5] V.K. Vyas, H.G. Bhatt, P.K. Patel, J. Jalu, C. Chintha, N. Gupta, and M. Ghate, *CoMFA and CoMSIA studies on C-aryl glucoside SGLT2 inhibitors as potential anti-diabetic agents*, SAR QSAR Environ. Res. 24 (2013), pp. 519–551. doi:10.1080/1062936X.2012.751553.

[6] J.H. Ahn, W.S. Park, M.A. Jun, M.S. Shin, S.K. Kang, K.Y. Kim, S. Dal Rhee, M.A. Bae, K.R. Kim, and S. G. Kim, *Synthesis and biological evaluation of homopiperazine derivatives with β-aminoacyl group*

as dipeptidyl peptidase IV inhibitors, Bioorg. Med. Chem. Lett. 18 (2008), pp. 6525–6529. doi:10.1016/j.bmcl.2008.10.076.

[7] S.S. Abd El-Karim, M.M. Anwar, Y.M. Syam, M.A. Nael, H.F. Ali, and M.A. Motaleb, *Rational design and synthesis of new tetralin-sulfonamide derivatives as potent anti-diabetics and DPP-4 inhibitors: 2D & 3D QSAR, in vivo radiolabeling and bio distribution studies*, Bioorg. Chem. 81 (2018), pp. 481–493. doi:10.1016/j.bioorg.2018.09.021.

[8] Z. Amini, M.H. Fatemi, and S. Gharaghani, *Hybrid docking-QSAR studies of DPP-IV inhibition activities of a series of aminomethyl-piperidones*, Comput. Biol. Chem. 64 (2016), pp. 335–345. doi:10.1016/j.compbiolchem.2016.08.003.

[9] S. Nordhoff, M. López-Canet, B. Hoffmann-Enger, S. Bulat, S. Cerezo-Gálvez, O. Hill, C. Rosenbaum, C. Rummey, M. Thiemann, V.G. Matassa, P.J. Edwards, and A. Feurer, *From lead to preclinical candidate: Optimization of β-homophenylalanine based inhibitors of dipeptidyl peptidase IV*, Bioorg. Med. Chem. Lett. 19 (2009), pp. 4818–4823. doi:10.1016/j.bmcl.2009.06.036.

[10] C. Jiang, S. Han, T. Chen, and J. Chen, *3D-QSAR and docking studies of arylmethylamine-based DPP IV inhibitors*, Acta Pharm. Sin. B 2 (2012), pp. 411–420. doi:10.1016/j.apsb.2012.06.007.

[11] S. Paliwal, D. Seth, D. Yadav, R. Yadav, and S. Paliwal, *Development of a robust QSAR model to predict the affinity of pyrrolidine analogs for dipeptidyl peptidase IV (DPP- IV)*, J. Enzym. Inhib. Med. Chem. 26 (2011), pp. 129–140. doi:10.3109/14756361003777057.

[12] B.D. Patel and M.D. Ghate, *3D-QSAR studies of dipeptidyl peptidase-4 inhibitors using various alignment methods*, Med. Chem. Res. 24 (2015), pp. 1060–1069. doi:10.1007/s00044-014-1178-7.

[13] U. Saqib and M.I. Siddiqi, *3D-QSAR studies on triazolopiperazine amide inhibitors of dipeptidyl peptidase-IV as anti-diabetic agents*, SAR QSAR Environ. Res. 20 (2009), pp. 519–535. doi:10.1080/10629360903278677.

[14] M. Sharma, S. Jain, and R. Sharma, *Trifluorophenyl-based inhibitors of dipeptidyl peptidase-IV as antidiabetic agents: 3D-QSAR COMFA, CoMSIA methodologies*, Network Model. Anal. Health Inform. Bioinform. 7 (2018), pp. 1. doi:10.1007/s13721-017-0163-8.

[15] E. Estrada, *On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research*, SAR QSAR Environ. Res. 11 (2000), pp. 55–73. doi:10.1080/10629360008033229.

[16] Z.Y. Algamal, M.K. Qasim, and H.T.M. Ali, *A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine*, SAR QSAR Environ. Res. 28 (2017), pp. 415–426. doi:10.1080/1062936X.2017.1326402.

[17] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, *Random forest: A classification and regression tool for compound classification and QSAR modeling*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 1947–1958. doi:10.1021/ci034160g.

[18] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, *GSA: A gravitational search algorithm*, Inf. Sci. 179 (2009), pp. 2232–2248. doi:10.1016/j.ins.2009.03.004.

[19] B. Xing and W.-J. Gao, *Gravitational search algorithm, in Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms*, Springer International Publishing, Cham, 2014, pp. 355–364.

[20] E. Rashedi, E. Rashedi, and H. Nezamabadi-Pour, *A comprehensive survey on gravitational search algorithm*, Swarm Evol. Comput. 41 (2018), pp. 141–158. doi:10.1016/j.swevo.2018.02.018.

[21] V.K. Bohat and K.V. Arya, *An effective gbest-guided gravitational search algorithm for real-parameter optimization and its application in training of feedforward neural networks*, Knowledge-Based Syst. 143 (2018), pp. 192–207. doi:10.1016/j.knosys.2017.12.017.

[22] B. Mohseni Bababdani and M. Mousavi, *Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives*, Chemom. Intell. Lab. Syst. 122 (2013), pp. 1–11. doi:10.1016/j.chemolab.2012.12.002.

[23] S. Mirjalili, S.Z. Mohd Hashim, and H. Moradian Sardroudi, *Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm*, Appl. Math. Comput. 218 (2012), pp. 11125–11137.

[24] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, *BGSA: Binary gravitational search algorithm*, Nat. Comput. 9 (2010), pp. 727–745. doi:10.1007/s11047-009-9175-3.

[25] A.A. Ibrahim, A. Mohamed, and H. Shareef, *Optimal power quality monitor placement in power systems using an adaptive quantum-inspired binary gravitational search algorithm*, Int. J. Electr. Power Energy Syst. 57 (2014), pp. 404–413. doi:10.1016/j.ijepes.2013.12.019.

[26] B. Ji, X. Yuan, Z. Chen, and H. Tian, *Improved gravitational search algorithm for unit commitment considering uncertainty of wind power*, Energy 67 (2014), pp. 52–62. doi:10.1016/j.energy.2014.02.014.

[27] X. Yuan, B. Ji, S. Zhang, H. Tian, and Y. Hou, *A new approach for unit commitment problem via binary gravitational search algorithm*, Appl. Soft. Comput. 22 (2014), pp. 249–260. doi:10.1016/j.asoc.2014.05.029.

[28] T. Baghgoli, M. Mousavi, and B. Mohseni Bababdani, *Descriptor selection evaluation of binary gravitational search algorithm in quantitative structure-activity relationship studies of benzyl phenyl ether diamidine's antiprotozoal activity and Chalcone's anticancer potency*, Chemom. Intell. Lab. Syst. 182 (2018), pp. 31–40. doi:10.1016/j.chemolab.2018.08.007.

[29] M.J. Islam, X. Li, and Y. Mei, *A time-varying transfer function for balancing the exploration and exploitation ability of a binary PSO*, Appl. Soft. Comput. 59 (2017), pp. 182–196. doi:10.1016/j.asoc.2017.04.050.

[30] M. Mafarja, I. Aljarah, A.A. Heidari, H. Faris, P. Fournier-Viger, X. Li, and S. Mirjalili, *Binary dragonfly optimization for feature selection using time-varying transfer functions*, Knowledge-Based Syst. 161 (2018), pp. 185–204. doi:10.1016/j.knosys.2018.08.003.

[31] A. Nitta, H. Fujii, S. Sakami, M. Satoh, J. Nakaki, S. Satoh, H. Kumagai, and H. Kawai, *Novel series of 3-amino-N-(4-aryl-1,1-dioxothian-4-yl)butanamides as potent and selective dipeptidyl peptidase IV inhibitors*, Bioorg. Med. Chem. Lett. 22 (2012), pp. 7036–7040. doi:10.1016/j.bmcl.2012.09.099.

[32] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, *DRAGON, Software Version 6.0, Talete Srl*, software, Milan, Italy, 2010. available at http://www.talete.mi.it/.

[33] V. Consonni, R. Todeschini, and M. Pavan, *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors*, J. Chem. Inf. Comput. Sci 42 (2002), pp. 682–692.

[34] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd. ed., Vol. 41, Wiley-VCH, Weinheim, Germany, 2009.

[35] J. Caballero and M. Fernández, *Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks*, J. Mol. Model. 12 (2006), pp. 168–181. doi:10.1007/s00894-005-0014-x.

[36] C. Rücker, G. Rücker, and M. Meringer, *y-Randomization and its variants in QSPR/QSAR*, J. Chem. Inf. Model. 47 (2007), pp. 2345–2357. doi:10.1021/ci700157b.

[37] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701. doi:10.1002/(ISSN)1611-0218.