ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(25), DOI: 10.17485/ijst/2019/v12i25/145108, July 2019*

Calculation of Stability Constant of Metal-thiosemicarbazone Complexes using MLR, PCR and ANN

Nguyen Minh Quang[2,3], Phạm Nu Ngoc Han[1], Nguyen Thi Ai Nhung[2] and Pham Van Tat[1*]

[1]Faculty of Science and Engineering, Hoa Sen University, Ho Chi Minh City, Vietnam
[2]Department of Chemistry, University of Sciences, Hue University, Hue City, Vietnam
[3]Faculty of Chemical Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract

Objectives: In this work, the stability constants log β_{11} of complexes between thiosemicarbazone and metal ions were predicted based on the modeling of Quantitative Structure and Property Relationship (QSPR). **Methods:** The QSPR models have been developed by using Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Artificial Neural Network (ANN). **Findings:** The results of QSPR models building have provided very positive results through the statistical values of validation. The QSPR models were cross-validated based on critical statistics. The quality of the QSPR models was exhibited by the statistical standards as the $QSPR_{MLR}$ model: R^2_{train} = 0.9446, R^2_{adj} = 0.939, Q^2_{LOO} = 0.9262, SE = 0.529 and F_{stat} = 160.817; $QSPR_{PCR}$ model: R^2_{train} = 0.949, R^2_{adj} = 0.942, Q^2_{CV} = 0.928, MSE = 0.292, RMSE = 0.540 and F_{stat} = 134.617; $QSPR_{ANN}$ model with architecture I (7)-HL(10)-O(1): R^2_{train} = 0.986, Q^2_{CV} = 0.984 and R^2_{test} = 0.983. **Applications:** Obviously, the results from this work could serve for designing new thiosemicarbazone derivatives that are helpful in the fields of analytical chemistry, pharmacy and environment.

Keywords: Artificial Neural Network, Multivariate Linear Regression, Principle Component Regression, QSPR Models, Stability Constants logβ_{11}, Thiosemicarbazone

1. Introduction

Bonding of metal ions with thiosemicarbazone ligands in aqueous solution plays an important role as in recent studies[1,2] as well as in studies of biological processes[3]. Efforts have been made to design new ligands that can be selectively linked to a metal ion and allow metal ion extraction[1,3]. Now there are many empirical data related to the stability constants of thiosemicarbazone-metal complexes collected[4–16]. In addition, this provides a good opportunity to develop quantitative relationships between the structure and stability constants of complexes that can be used to design new thiosemicarbazone ligands that bind to metal ion[17,18]. There are continuous publications in the literature showing that the development of QSPR models to predict stability constants of complexes using multivariate techniques is a good choice[1].

On the other side the QSPR models of the stability constants of the metal-thiosemicarbazone complexes were preceded for a lot of practical applications. The molecular activities of thiosemicarbazone compounds and their complexes were used to support the analytical chemistry[19] and medicinal areas[3]. The metal-thiosemicarbazone complexes are being applied in the medicinal areas for antibacterial, antifungal, anti-malarial, antitumor and antiviral activity[20–22]. Furthermore, they are also used to catalyze for chemical reactions[23] and in the environmental area[24]. In these cases the applicability of QSPR models in

*Author for correspondence

practice is not simply due to not enough of the information needed to describe the molecules and the details of the calculation method.

To solve the problems outlined above, we proceed to establish the relationships between the Quantitative Structure and Properties (QSPR) related to the stability constant ($\log\beta_{11}$) of metal-thiosemicarbazone complexes Ni^{2+}, Co^{2+}, Mo^{6+}, Cu^{2+}, Mn^{2+}, Zn^{2+}, Ag^{+}, Pb^{2+}, Fe^{2+} and Zn^{2+}. All of these models are based on molecular descriptors of complexes, resulting only from 2D and 3D molecular calculations. Some 3D molecular descriptors were calculated using semi-empirical quantum chemistry with new PM7 and PM7/sparkle. Here we report new QSPR models for the $\log\beta_{11}$ stability constants of 10 transition metal ions with a set of diverse thiosemicarbazone ligands in aqueous solution at 298 K and 0.1 M ion strength[4-16]. The models were cross-validated using an external validation process

2. Materials and Methods

2.1 Data Sets

The experimental stability constants ($\log\beta_{11}$) for the M:L complexes of transition (Ni^{2+}, Co^{2+}, Mo^{6+}, Cu^{2+}, Mn^{2+}, Zn^{2+}, Ag^{+}, Pb^{2+}, Fe^{2+} and Zn^{2+}) metal ions with different thiosemicarbazone ligands in aqueous solution were taken from the published literature[4-16] at standard range 298 K to 323 K, pH 4 to 10 and an average of ionic strength I of 0.1 M. The constants $\log\beta_{11}$ have been also adjusted to the temperature 298 K to 323 K. Complex structures of thiosemicarbazone ligands and metal ions, as well as respective $\log\beta_{11}$ constants derived from the published literature were converted into 2D and 3D structures–Input Data waited on as input database of QSARIS program[25]. The 74 structures of training set involve the metal-thiosemicarbazone complexes containing 19 (Ni^{2+}), 16 (Co^{2+}), 29 (Cu^{2+}), 7 (Mn^{2+}), 1 (Zn^{2+}), 1 (Mo^{6+}) and 1 (Ag^{+}), as given in Table 1. The $\log\beta_{11}$ values vary in the ranges from 6.489 to 11.210 (Ni^{2+}), from 6.382 to 10.590 (Co^{2+}), from 6.179 to 14.560 (Cu^{2+}). The test set includes the 8 transition metal ions Cu^{2+}, Ni^{2+}, Fe^{2+}, Pb^{2+}, Zn^{2+}, Co^{2+}, Mn^{2+} and Ag^{+}, as shown in Table 2. Thiosemicarbazone ligand and metal-ligand complex structures are shown in Figure 1[19].

The difference between the experimental $\log\beta_{11}$ constants may be found for the same complexes of the different authors could have relatively high values, as shown in Table 1. If a lot of $\log\beta_{11}$ constants are available for a ligand, then the most recent values or value consistent with the different experimental methods are chosen. Thus, 19 (Ni^{2+}), 16 (Co^{2+}), 29 (Cu^{2+}), 7 (Mn^{2+}), 1 (Zn^{2+}), 1 (Mo^{6+}) and 1 (Ag^{+})-thiosemicarbazone complexes are taken for the QSPR modeling of the $\log\beta_{11}$ constants, as exhibited in Figure 2. The $\log\beta_{11}$ constants of complexes alter in various ranges, as shown in Table 3. The normal distribution of $\log\beta_{11}$ values for complexes depicted the characteristics of the dataset, as shown in Figure 2.

The metal-thiosemicarbazone complexes are generated by the following reaction between a metal ion (M) and a thiosemicarbazone ligand (L) in an aqueous solution:

$$pM + qL \rightleftharpoons M_pL_q \qquad (1)$$

The reaction occurs this step with p = 1 and q = 1; the stability constant β_{11} is calculated by the following expression:

$$\beta_{11} = \frac{[ML]}{[M][L]} \qquad (2)$$

Table 1. The $\log\beta_{11,exp}$ constants for ML complex types of 20 various ligands with 7 metal ions inaqueous solution. And the experimental range between minimal $\log\beta_{11,exp}$ and maximal $\log\beta_{11,exp}$ values

No	Thiosemicarbazone ligand				Metal ions	$\log\beta_{11}$	Ref.
	R_1	R_2	R_3	R_4			
1	H	H	H	$-C_6H_2(OH)_2OCH_3$	Ni^{2+}	6.4886	[28]
2	H	H	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Ni^{2+}	10.890	[33]
						11.210	[34]
3	H	H	H	$-C_9H_8NO$	Ni^{2+}	7.709	[39]
4	H	H	H	$-C_{10}H_6OH$	Ni^{2+}	9.600	[38]
5	H	H	H	$-C_6H_2(OCH_3)_2OH$	Cu^{2+}	6.2355	[31]
6	H	$-C_6H_5$	$-CH_3$	$-CCH_3{=}N{-}OH$	Cu^{2+}	6.179	[32]
						7.7559	[32]

7	H	H	$-CH_3$	$-CH=N-NHC_6H_5$	Cu^{2+}	11.700	[34]
						12.300	[34]
8	H	H	H	$-C_9H_5NOH$	Cu^{2+}	14.560	[35]
9	H	H	H	$-C_6H_3(OH)OCH_3$	Cu^{2+}	9.030	[37]
						9.830	[32]
10	H	H	H	$-C_9H_8NO$	Cu^{2+}	7.796	[39]
11	H	H	H	$-C_{10}H_6OH$	Cu^{2+}	9.780	[38]
12	H	H	H	$-C_6H_2(OH)_2OCH_3$	Co^{2+}	6.3820	[29]
13	H	$-CH_3$	$-CH_3$	$-CH=N-NHC_6H_5$	Co^{2+}	10.300	[34]
						10.590	[34]
14	H	H	H	$-C_{10}H_6OH$	Co^{2+}	7.890	[38]
						9.000	[38]
15	H	H	H	$-C_9H_8NO$	Co^{2+}	7.251	[39]
						8.34	[39]
16	H	$-CH_3$	$-CH_3$	$-CH=N-NHC_6H_5$	Mn^{2+}	9.770	[34]
						10.050	[34]
17	H	H	H	$-C_{10}H_6OH$	Mn^{2+}	4.660	[38]
						5.670	[38]
18	H	$-C_2H_5$	H	$-C_9H_5NOH$	Zn^{2+}	6.130	[35]
19	H	H	H	$-C_6H_4-N-(CH_3)_2$	Ag^{+}	17.200	[36]
20	H	H	H	$-C_6H_2(OCH_3)_2OH$	Mo^{6+}	6.3365	[30]

Table 2. The log β_{11} stability constants of 10 complexes of external test set are validated by the models

Ligand				ion	$QSPR_{MLR}$		$QSPR_{PCR}$		$QSPR_{ANN}$		$\log\beta_{11,exp}$
R_1	R_2	R_3	R_4		$\log\beta_{11,cal}$	*ARE*, %	$\log\beta_{11,cal}$	*ARE*, %	$\log\beta_{11,cal}$	*ARE*, %	
H	H	CH_3	$-CH=N-NHC_6H_5$	Co^{2+}	11.754	15.01	11.729	14.76	11.513	12.65	10.22[33]
H	H	CH_3	$-CH=N-NHC_6H_5$	Mn^{2+}	9.448	4.27	9.462	4.13	10.863	10.06	9.87[34]
H	H	H	$-C_6H_5$	Ag^{+}	17.231	11.17	17.355	11.97	15.743	1.57	15.5[36]
H	H	H	$-C_6H_5$	Cu^{2+}	16.982	4.06	17.108	3.34	16.556	6.46	17.7[54]
H	H	H	$-C_6H_3(OH)OCH_3$	Pb^{2+}	7.163	9.70	7.159	9.63	6.078	6.93	6.53[37]
H	H	H	$-C_6H_3(OH)OCH_3$	Fe^{2+}	7.864	2.26	7.79	1.30	7.020	8.71	7.69[37]
H	H	H	$-C_6H_3(OH)OCH_3$	Co^{2+}	8.551	6.62	8.487	5.82	8.228	2.60	8.02[37]
H	H	H	$-C_6H_3(OH)OCH_3$	Ni^{2+}	9.045	4.56	8.993	3.97	9.109	5.31	8.65[37]
H	H	H	$-C_{10}H_6OH$	Pb^{2+}	7.234	10.10	7.175	9.20	6.442	1.95	6.57[38]
H	H	H	$-C_{10}H_6OH$	Zn^{2+}	8.467	18.09	8.408	17.26	7.045	1.74	7.17[38]
					MARE,%	8.58	*MARE*,%	8.14	*MARE*,%	5.80	

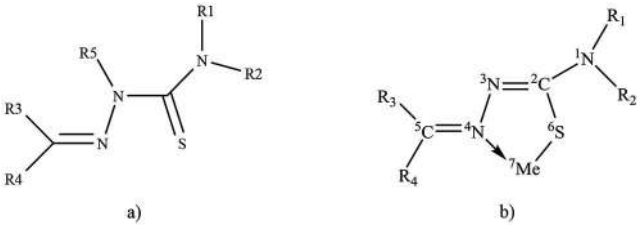

Figure 1. Molecular structure: a) Thiosemicarbazone ligand and b) Metal-thiosemicarbazone complex.

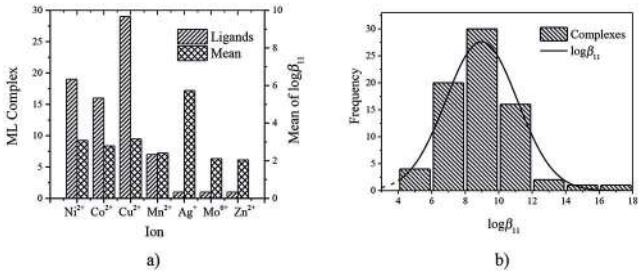

Figure 2. The metal-thiosemicarbazone complexes in dataset: a) Number of complexes and mean values of $\log\beta_{11}$; b) Normal distribution curve of stability constants $\log\beta_{11}$.

2.2 Molecular Descriptors

Optimized structures are wielded for calculating the molecular descriptors. The 2D and 3D molecular descriptors are calculated by coding in different forms of a molecular structure. The molecular descriptors include physico-chemistry descriptor LogP; the 2D descriptors xp3, xp5, xvch8, SaasC, nelem and nrings; the 3D descriptors such as ABSQ, Ovality and Surface[25]. BIOVIA Draw 2017 R2 program was used to re-construct the 2D structures of molecules. The topological and quantum descriptors were calculated on Lenovo W540 PC using the MOPAC2016 a semi-empirical level PM7 and PM7/sparkle[26] and QSARIS program.

The molecular descriptors of each descriptive group were used as initial molecular descriptors in the QSPR model to construct the various QSPR models using different techniques. The predictor selection is one of the most important steps in QSPR modeling. For the following QSPR modeling, the stability constant log β11 was transformed as the dependent variable. The $QSPR_{MLR}$ models were constructed by the predictor selection technique using Genetic Algorithm (GA) of QSARIS[25] and forward technique of REGRESS program[27] on Add-ins in MS-EXCEL[25]. Besides the $QSPR_{PCR}$ model was also built on a XLSTAT2016 program[28]. The Artificial Neural Network model $QSPR_{ANN}$ was developed by using the Neural Network tool in the MATLAB 2016 program[29].

2.3 Regression Analysis

The dataset of 74 complexes is used as a training set. A 74-molecules set was performed for MLR and PCR regression analysis as a constructing method of the QSPR model. The QSPR models were generated by using $\log\beta_{11}$ constant values as dependent variable and different descriptors as independent variables. The cross-validation limit with correlation value is set at 0.7; the descriptors in the final equation are selected by combining regression technique and genetic algorithms to have $QSPR_{MLR}$ and $QSPR_{PCR}$ models. The statistical methods used to evaluate QSPR models include the number of compounds in the regression model, the regression coefficients R^2, the adjusted R^2_a, the number of descriptors in the model k, F-test for statistical significance, cross-correlation coefficient Q^2_{cv}, predictive correlation coefficients R^2_{pred}, and Standard Error (SE).

2.3.1 MLR Analysis

The Multiple Linear Regression analysis (MLR) is used to construct a linear relationship between a dependent variable y ($\log\beta_{11}$) and independent variables *x* (molecular description)[30].

Multiple Regression Analysis (MLR) was used to estimate the regression coefficient (R^2) by least-squared fitting; the Sum of Squared Residual (SSR) values of observed and predicted values are minimized[27,28]. The linear model can produce a linear approximation in relation to all observed data points[30,31]. In linear regression, the dependent variable ($\log\beta_{11}$) y depends on the molecular descriptors, x. The regression equation has the form:

$$y = \sum_{i=1}^{k} b_i x_i + c \qquad (3)$$

Here y is dependent variable $\log\beta_{11}$, the regression coefficient, b_i corresponds to the molecular descriptors, x_i; and c is a constant.

2.3.2 PCR Analysis

Principal Component Regression analysis (PCR) was used to evaluate data based on the correlation between dependent variables and independent variables[28].

Principal Component Regression analysis is used to find the appropriate structure in data sets. The purpose of this method is to transfer the correlative variables, replacing the original descriptors with the new descriptors called the Principal Components (PC). These PCs are related to each other and are built in the simple linear combination of the original variables. This technique turns the data into a new set of axes so that the first few axes reflect most of the variables in the data. The first PC (PC1) is determined by the maximum variance of the entire dataset. The second PC2 (PC2) is the direction that describes the maximum variance in the orthogonal subspace of PC1[25]. The next components are orthogonal to the previously selected components and describe the maximum remaining variance. By drawing data on a new set of spindles, it can automatically detect the basic structure. The value of each point is rotated in a given axis, called the PC value. PCA selects a new axis set for data[12,28,32]. Those are selected in descending order of the data. The purpose of PCR is to evaluate the dependent variables on the basis of the selected principal components of the independent variables.

2.4 Artificial Neural Network

The Artificial Neural Network (ANN) receives the processed input information that is capable of communicating by transmitting information through interconnected neurons, weighted connections. Some of their basic features should be emphasized initially[33,34]. Therefore, ANN is a Multi-layer Perceptron (MLP); MLP can have many hidden layers in architectural style I(k)-HL(m)-O(n):

- Input variables I(k): x_1, x_2, ..., x_k.
- The connection in the network takes place in each neuron. Each connection is determined by the weight between neurons i and j.
- The output variable is composed of n neurons (O (n)): y_j corresponds to a neuron.
- An additional external error variable b (bias) for each neuron.
- The training process follows the rules of propagation in the network, defining the effect through comparison with input variables from outside x_k.
- The activation function sigmoid is chosen, the process is evaluated by the correlation between the input variable and the output variable y_j of a neuron. The hyperbolic sigmoid function can be used as a transfer function in the input and output data sets. It is given in[33,34]:

$$a = \tan sig(n) = \frac{2}{\left(1+e^{-2n}\right)^{-1}} \quad (4)$$

In the current article, the number of hidden layers and the appropriate epoch has been carefully checked with trial and error. We used a feed-forward neural network with the Levenberg-marquest learning algorithm to train it[35–37]. This algorithm seems to be the fastest method for training medium-sized feed-forward neural networks. The training of the ANN neural network model is performed until the average squared error (MSE_{ANN}) is minimized followed by the comparison of the network output with the actual values of the output obtained from the test results[38]. The training process of a neural network consists of adjusting the weights and deviations of the network to optimize neural network performance. The efficiency function for feed-forward neural networks is based on the average square error of the ANN model (MSE_{ANN}). The average squared error between the output of the network (y_i) and the target output (t_i) is given by the following formula[37]

$$MSE_{\text{ANN}} = \frac{1}{n}\sum_{1}^{n}\left(t_i - y_i\right)^2 \quad (5)$$

2.5 Validation of QSPR Model

The optimum method to assess the quality of regression models is to perform the internal assessments for QSPR models. The validation was mainly done by a Leave-one-out (LOO) cross-examination, when an observation ($\log\beta_{11}$) value was excluded from the training set and the training data was divided into subsets of size are equal[25]. The model was constructed using these subsets and the dependent variable value of the data point was not included in the defined subset, which is a predicted value. The predicted averages will be the same for R^2_{train} and Q^2_{LOO} (the value of the correlation coefficient is cross-validated) as all data points would be considered sequentially as predicted values in the LOO subset. The same procedure is repeated after removing another object until all objects have been discarded once. The LOO cross-validation leads to statistically significant patterns

for each regression model[28]. R^2_{train} was used the following formula:

$$R^2_{train} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

The QSPR models screened are based on the values Q^2_{LOO} for cross-validation set, R^2_{test} for the test set. These values are calculated by using formula (6) to validate for all models[25,27,39-42].

The adjusted R^2 value (R^2_{adj}) is the coefficient of significance to determine the number of internal variables for QSPR models. The value of R^2_{adj} can also be negative if the data set does not have a sufficient number of observations *n*. This coefficient is only counted if the user is not fixed to the model[27-28]. R^2_{adj} is defined by formula:

$$R^2_{adj} = R^2_{train} - \left(\frac{k-1}{N-1}\right)(1 - R^2_{train}) \quad (7)$$

The R^2_{adj} value is used to calibrate R^2_{train}, taking into account the number of independent variables used in the model. Average square error (*MSE*)[27] is determined by following formula:

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - k - 1} \quad (8)$$

Here $\hat{y}$, y, and $\bar{y}$ are predicted, actual and mean values of the $\log\beta_{11}$. The values of the square root of errors (RMSE) are calculated from the square root of the MSE values.

The predictability of the QSPR models can be validated by using the average absolute values of the relative errors MARE (%):

$$MARE,\% = \frac{\sum_{i=1}^{n} ARE_i,\%}{n} \quad (9)$$

Where ARE (%) is the absolute value of the residuals. They are calculated by the following formula:

$$ARE,\% = \frac{\left|\log \beta_{11,exp} - \log \beta_{11,cal}\right|}{\log \beta_{11,exp}} 100 \quad (10)$$

Where *n* is the number of test substances; $\log\beta_{11,exp}$ and $\log\beta_{11,cal}$ are the experimental and calculated stability constants.

3. Results and Discussion

3.1 Regression Analysis

3.1.1 $QSPR_{MLR}$ Modeling

The $QSPR_{MLR}$model was gone forwardfor modeling $\log\beta_{11}$ stability constants of the ML complexes of 7 metal ions 19 (Ni^{2+}), 16 (Co^{2+}), 29 (Cu^{2+}), 7 (Mn^{2+}), 1 (Zn^{2+}), 1 (Mo^{6+}) and 1 (Ag^{+}) with 74 structurally diverse thiosemicarbazone ligands, including 74 log β_{11} stability constant values (Table 3). The $QSPR_{MLR}$ and $QSPR_{PCR}$ models were built for 7 transition metal ions using 2D and 3D molecular descriptors. Student T-test method was used to compare RMSE and R^2 values at 95% confidence level. Also, correlation coefficients of $QSPR_{MLR}$ model are the multiple correlation R_{train} of 0.9719 and cross-validation correlation Q_{LOO} of 0.9624, as shown in Figure 3b.

The selected subsets for $QSPR_{MLR}$ models are presented in Table 4. The descriptors k was varied in the range 1 to 8. The changing descriptors have led to the various changes in RMSE, R^2_{train}, Q^2_{LOO}, RMS_{train} and RMS_{CV} values, as shown in Table 4. During the modeling process, the dataset is split randomly into the training and test subset, in which the training subset contains about 80% of initial data set. The $QSPR_{MLR}$ models are cross-evaluated by the Leave-one-out method through the statistical value Q^2_{LOO}. The statistical parameters such as values R^2_{train}, Q^2_{LOO}, and RMSE are used to select a best subset. Therefore the best model has highest R^2_{train} and Q^2_{LOO} values and lowest RMSE value with suitable number k.

In Table 4 the molecular descriptors are refined preliminarily using genetic algorithm. From the descriptors of subsets, the $QSPR_{MLR}$ model was re-built with the forward technique for the REGRESS system[21] on Add-Ins MS-EXCEL.

From Table 4, the best subset with k = 7 is selected for $QSPR_{MLR}$ modeling, as shown in bold:

$\log\beta_{11}$ = 53.803 - 7.024×nelem- 0.070×CosmoArea+ 0.534×xvp3- 8.185×MaxNeg +8.065×Hmin- 70.721×xch10 + 0.371×SsCH3 (11)

With n = 74, R^2_{train} = 0.9446; Q^2_{LOO} = 0.9262; and p-values < 0.05; F-stat = 160.8173, RMS = 0.5292.

The results in Table 4 showed that the k value goes up to 8 then the values R^2_{train} and Q^2_{LOO} are not increased. Thus the statistical values change specifically the values $RMSE_{train}$ and $RMSE_{CV}$ go up. Therefore, the k value goes up to 8 then the statistical value changed insignificantly. Therefore, the best subset of descriptors with k = 7 is selected for $QSPR_{MLR}$ modeling in Equation (11). The best $QSPR_{MLR}$ model in bold is shown in Table 4 [25,27].

3.1.2 $QSPR_{PCR}$ Modeling

The best $QSPR_{MLR}$ model (11) based on 7 molecular descriptors, as listed in Table 4. In this work we have also approached to construct the QSPRPCR model by using this dataset with 8 molecular descriptors, as given in Table 4. This model was constructed from the results of the Principal Components Analysis (PCA). Similarly, the $QSPR_{PCR}$ modeling process is implemented by the training set containing original data of 80% and the remaindered is the test set. The $QSPR_{PCR}$ model is also validated by statistical values R^2_{train}, Q^2_{LOO}, explained variance and RMSE. The change of principal components in $QSPR_{PCR}$ model influences the RMSE values. The increment of the components caused the decrement of RMSE values for training and validation process, respectively, as exhibited in Figure 4. So the best $QSPR_{PCR}$ model consists of 7 principal components. It can be transformed into a $QSPR_{PCR}$ model of the original-molecular descriptors, as shown in Equation (12).

The Principal Component Regression (PCR) equation is depicted for $QSPR_{PCR}$ modeling with statistical values, as following Equation (12):

$$\log\beta_{11} = 54.718 - 7.011 \times nelem - 0.0721 \times CosmoArea + 0.544 \times xvp3 - 7.040 \times MaxNeg + 7.944 \times Hmin - 79.413 \times xch10 + 0.352 \times SsCH3 \quad (12)$$

With n = 74; R^2_{train} = 0.949; Q^2_{CV} = 0.928;MSE = 0.292; RMSE = 0.540; F_{stat} = 134.617.

The $QSPR_{PCR}$ model (12) is statistically significant. This equation has the explained variance of 94.9% in the stability constants, as influenced in Figure 4. From Equations (11), (12) the change of the log β_{11} stability constant could be explained by the molecular descriptors. The statistically importance of the molecular descriptors in the QSPR model can be used in the seeking direction of new complexes. Consequently, the modeling results may orientate the design of new thiosemicarbazone ligands based on the structural descriptors to obtain the higher log β_{11} stability constants.

Table 3. 74 metal-thiosemicarbazone complexes (n), minimum ($\log\beta_{11}$, min) and maximum (log β_{11},max) constants of the stability constants from the selected data for QSPR model

No	Metal ion	Number of complexes, n	log $\beta_{11,min}$	log $\beta_{11,max}$
1	Ni^{2+}	19	6.49	11.21
2	Co^{2+}	16	6.38	10.59
3	Cu^{2+}	29	6.18	14.56
4	Mn^{2+}	7	4.66	10.05
5	Ag^{+}	1	17.20	17.20
6	Mo^{6+}	1	6.34	6.34
7	Zn^{2+}	1	6.13	6.13

Table 4. The multidimensional $QSPR_{MLR}$ models obtained with based on forward regression technique and a Genetic Algorithm to select the suitable subsets. The best model is in bold

k	Molecular descriptors in QSPR models	R^2_{train}	Q^2_{LOO}	RMS_{train}	RMS_{CV}
1	nelem	0.4988	0.4704	1.5242	1.5560
2	nelem; cosmoarea	0.7180	0.7018	1.1512	1.1676
3	nelem; cosmoarea; xvp3	0.8535	0.8007	0.8359	0.9546
4	nelem; cosmoarea; xvp3; Maxneg	0.8853	0.8291	0.7448	0.8838
5	nelem; cosmoarea; xvp3; Maxneg; Hmin	0.9017	0.8406	0.6947	0.8538
6	nelem; cosmoarea; xvp3; Maxneg; Hmin; xch10	0.9339	0.9057	0.5738	0.6564
7	**nelem; cosmoarea; xvp3; Maxneg; Hmin; xch10; SsCH3**	**0.9446**	**0.9262**	**0.5292**	**0.5809**
8	nelem; cosmoarea; xvp3; Maxneg; Hmin; xch10; SsCH3; dipole	0.9446	0.9183	0.5332	0.6110

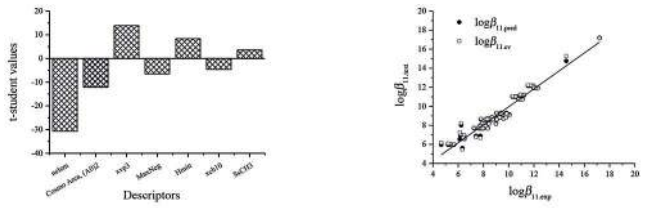

Figure 3. a) A student's t test values are to compare the R^2 values; b) Correlation between predicted vs. experimental $\log\beta_{11}$ values resulting from the $QSPR_{MLR}$ model with values R^2_{train} of 0.9446 and Q^2_{LOO} of 0.9262.

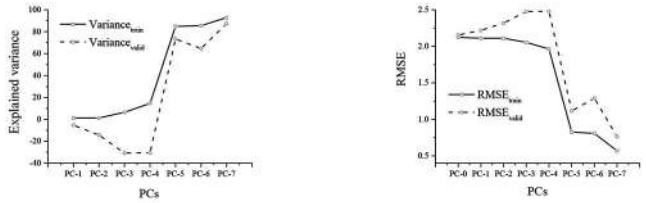

Figure 4. The influence of component change for quality of $QSPR_{PCR}$ model: a) The percentage of explaining variance vs. principal components; b) The RMSE values vs. principal components.

3.2 Construction of $QSPR_{ANN}$ *Model*

For the development of $QSPR_{ANN}$ model, the Artificial Neural Network was also approached in this work. The Artificial Neural Network with the training set of 74 complexes using back-propagation algorithm was implemented, as given in Table 1. The neural network can be constructed for prediction of log β_{11} stability constant values of external test set, as shown in Table 2.

The different iterations of the training process and the change of neurons of hidden layer could create the several $QSPR_{ANN}$ models I(k)-HL(m)-O(n). In Table 2, we have shown the best $QSPR_{ANN}$ model with architecture I(7)-HL(10)-O(1). The developed $QSPR_{ANN}$ model based on the significant descriptors statistically of $QSPR_{MLR}$ and $QSPR_{PCR}$ models.

Therefore the Neural Network architecture I(7)-HL(10)-O(1) consists of the molecular descriptors nelem, cosmo area, xvp3, MaxNeg, Hmin, xch10 and SsCH3 as 7 neurons on the input layer; the output layer O(1) has 1 neuron as the stability constant log β_{11}; the hidden layer HL(10) includes 10 neurons. This three-layer neural network is trained by quick-propagation algorithm combining Levenberg-marquardt algorithm. The transfer function hyperbolic sigmoid tangent is used to train this neural network I(7)-HL(10)-O(1). The others are used in the training process as learning rate of 0.01, the momentum of 0.9, the convergent goal of 10^{-10} and the residual function is RMSE. The $QSPR_{ANN}$ model I(7)-HL(10)-O(1) has the statistical values R^2_{train} of 0.9860, Q^2_{CV} of 0.9840, and R^2_{test} of 0.9830. These results indicate that $QSPR_{ANN}$ model I(7)-HL(10)-O(1) is better than $QSPR_{MLR}$ and $QSPR_{PCR}$ models. So the $QSPR_{ANN}$ modeling could explain the variation 98.6% in the data set; and the $QSPR_{MLR}$ and $QSPR_{PCR}$ models explain the variation 94.5% and 94.9%, respectively. The $QSPR_{ANN}$ model I(7)-HL(10)-O(1) exhibited a better fitness between the predicted and the experimental values. This may also be found in the statistical values ARE, % and MARE, %, as shown in Table 2.

3.3 External Validation

QSPR models must be tested for external validation criteria. The authors recommended that in addition to the cross-validated (Q^2_{CV}) value. The multiple-correlation coefficients R have been determined from the experimental and the predicted stability constant values for an external test set must be close to 1. In this study, we used an external data set of 10 metal-thiosemicarbazone complexes from the experimental literature to test the applicability of the constructed QSPR models, as given in Table 2. The QSPR models satisfied the criteria.

The MARE, % values of the QSPR models are also calculated, respectively, as shown in Table 2, indicating that the $QSPR_{ANN}$ model appeared the highest predictability and the predicted log β_{11} stability constant values resulting from model $QSPR_{ANN}$ are very close to the experimental values. In addition, the one-way ANOVA method is used to compare the discrepancy between the experimental and predicted log β_{11} stability constant values resulting from three QSPR models. Accordingly, the discrepancy between them is insignificant ($F = 0.068598 < F_{0.05} = 2.866266$). Thus, we can use the QSPR models to estimate the log β_{11} stability constant of new complexes.

4. Conclusion

We conclude that the QSPR modeling of transition metal complex was implemented by incorporating the multivariate regression and the Artificial Neural Network. The QSPR models were constructed successfully by the

selected molecular descriptors by the Genetic Algorithm and forward-regression technique. The stability $\log\beta_{11}$ constants of metal-thiosemicarbazone complexes generated by the $QSPR_{MLR}$, $QSPR_{PCR}$ and $QSPR_{ANN}$ models are a good agreement with experimental data.

The developed QSPR models are statistically satisfactory. The applicability of these QSPR models promised to predict accurately the stability constants of the complexes between new thiosemicarbazone ligands with metal ions. The above results indicated that the $QSPR_{ANN}$ model has the best predictability.

5. References

1. Comprehensive coordination Chemistry II. Applications of Coordination Chemistry. 2003. https://b-ok.org/book/550101/d9957e/
2. Catalysis with Metal Complexes. 2012. https://www.amazon.in/Homogeneous-Catalysis-Metal-Complexes-Fundamentals-ebook/dp/B00BWZ2NDO
3. Coordination Chemistry II. Bio-coordination Chemistry. 2003. https://www.elsevier.com/books/comprehensive-coordination-chemistry-ii/mccleverty/978-0-08-043748-4
4. Hymavathi M, Viswanatha C, Devanna N. A sensitive and selective chromogenic reagent using 2-hydroxy 3, 5-dimethoxy benzaldehyde thiosemicarbazone (HDMBTSC) for direct and derivative spectrophotometric determination of Molybdenum (VI). International Journal of Mathematics and Mathematical Sciences. 20114; 2(1):43–8.
5. Hymavathi M, Viswanatha C, Devanna N. Direct and derivative spectrophotometric determination of Copper (II) using a sensitive and selective chromogenic organic reagent 2-hydroxy 3,5-dimethoxy benzaldehyde thiosemicarbazone (HDMBTSC). World Journal of Pharmaceutical Sciences. 2014; 3(8):1688–95.
6. Gomaa EA, Ibrahim KM, Hassan NM. Evaluation of thermodynamic parameters (conductometrically) for the interaction of Cu (II) ion with 4-phenyl-1-diacetyl monoxime -3- thiosemicarbazone (BMPTS) in (60% V) ethanol (EtOH-H_2O) at different temperatures. International Journal of Engineering Science. 2014; 3(1):44–51.
7. Aljahdali M, EL-Sherif AA. Synthesis, characterization, molecular modeling and biological activity of mixed ligand complexes of Cu (II), Ni (II) and Co (II) based on 1,10-phenanthroline and novel thiosemicarbazone. Inorganica Chimica Acta. 2013; 407:58–68.
8. El-Karim ATA, El-Sherif AA. Potentiometric, equilibrium studies and thermodynamics of novel thiosemicarbazones and their bivalent transition metal (II) complexes. Journal of Molecular Liquids. 2016; 219:914–22.
9. Rogolino D, Cavazzoni A, Gatti A, Tegoni M, Pelosi G, Verdolino V, Fumarola C, Cretella D, Petronini PG, Carcelli M. Anti-proliferative effects of Copper (II) complexes with Hydroxyquinoline-thiosemicarbazone ligands. European Journal of Medicinal Chemistry. 20177; 128:140–53.
10. Jimenez MA, Luque De Castro MD, Valcarcel M. Potentiometric study of Silver (l)-thiosemicarbazonates. Microchemical Journal. 1980; 25:301–8.
11. Garg BS, Jain VK. Determination of thermodynamic parameters and stability constants of complexes of biologically active o-vanillinthiosemicarbazone with bivalent metal ions. Thermochimica Acta. 1989; 146:375–9.
12. Sahadev, Sharma RK, Sindhwani SK. Thermal studies on the chelation behaviour of biologically active 2-hydroxy-1-naphthaldehyde thiosemicarbazone (HNATS) towards bivalent metal ions: A potentiometric study. Thermochimica Acta. 1992; 202:291–9.
13. Sarkar K, Garg BS. Determination of thermodynamic parameters and stability constants of the complexes of p-MITSC with transition metal ions. Thermochimica Acta. 1987; 113:7–14.
14. Multivariate Data Analysis. 2010. https://is.muni.cz/el/1423/podzim2017/PSY028/um/_Hair_-_Multivariate_data_analysis_7th_revised.pdf
15. Applied Linear Regression. 2005. https://www.amazon.com/Applied-Linear-Regression-Sanford-Weisberg/dp/0471663794
16. Chemometrics tools in QSAR/QSPR studies. 2015. https://www.sciencedirect.com/science/article/abs/pii/S0169743915001641
17. QSPR/QSAR-based Perturbation Theory approach and mechanistic electrochemical assays on carbon nanotubes with optimal properties against mitochondrial Fenton reaction experimentally induced by Fe2+-overload. 2017. http://agris.fao.org/agris-search/search.do?recordID=US201700139080
18. Singh RB, Garg BS, Singh RP. Analytical applications of thiosemicarbazones and semicarbazones: A review. Talanta. 1978; 25(11-12):619–32.
19. Complexes with Schiff base Ligands: Synthesis, characterization, antimicrobial studies. 2013. https://pdfs.semanticscholar.org/c453/8073262490e37ae6cc73f64ba4e7136c5891.pdf
20. Rajendran M, Panneerselvam A, Periasamy V, Grzegorz MJ. Palladium (II) pyridoxal thiosemicarbazone complexes as efficient and recyclable catalyst for the synthesis of propargylamines by a three-component coupling reactions in ionic liquids. Polyhedron. 2016; 119:300–6.

21. Ezhilarasi. Synthesis characterization and application of salicylaldehyde Thiosemicarbazone and its metal complexes. International Journal of Research in Chemistry and Environment. 2012; 2(4):130–48.
22. Stewart JPP. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. Journal of Molecular Modeling. 2013; 19:1–32.
23. Modern analytical chemistry. 1999. https://www.amazon.in/Modern-Analytical-Chemistry-David-Harvey/dp/0072375477
24. Excel for chemists. 2011. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118093955
25. Essential regression and experimental design for chemists and engineers. 1998. http://www.oocities.org/siliconvalley/network/1032/
26. Excel for scientists and engineers: Numerical methods. 2007. http://library.aceondo.net/ebooks/Computer_Science/Excel_for_Scientists_and_Engineers-Numerical_Methods[Wiley](2007)[1].pdf
27. Avogadro 1.2.0. 2016. https://avogadro.cc/releases/avogadro_120/
28. Hymavathi M, Viswanatha C, Devanna N. A study on synthesis of novel chromogenic organic reagent 3,4-dihydroxy-5-methoxy benzaldehyde thiosemicarbazone and spectrophotometric determination of Nickel (II) in presences of Triton X-100. Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2014; 5(5):625–30.
29. Principal component analysis. http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf
30. Kisi O. Multi-layer perceptrons with Levenberg-marquardt training algorithm for suspended sediment concentration prediction and estimation. Hydrological Sciences Journal. 2004; 49:1025–40.
31. Vogl TP, Mangis JK, Rigler AK, Zink WT, Alkon DL. Accelerating the convergence of the back propagation method. Biological Cybernetics. 1998; 59:257–63.
32. Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal on Applied Mathematics. 1963; 11(2):431–41.
33. Hagan MT, Menhaj M. Training feed-forward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks. 1994; 5(6):989–93.
34. Neural Network Design. 1996. https://hagan.okstate.edu/NNDesign.pdf
35. Verikas A, Bacauskiene M. Using Artificial Neural Networks for process and system modeling. Chemometrics and Intelligent Laboratory Systems. 2003; 67:187–91.
36. Amemiya T. Selection of regressors. International Economic Review. 1980; 21:331–54.
37. Elements of continuous multivariate analysis. 1969. https://www.amazon.in/Elements-Continuous-Multivariate-Analysis-Dempster/dp/0201014858
38. Eriksson L, Johansson E, Kettaneh-Wold N, Kettaneh-Wold S. Multi- and megavariate data analysis: Principles and applications. Journal of Chemometrics. 2001; 16(5):261–2.
39. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–4.
40. Golbraikh A, Tropsha A. Beware of q2. Journal of Molecular Graphics and Modeling. 2002; 20:269–76.
41. Jimenez MA, Luque De Castro MD, Valcarcel M. Titration of Thiosemicarbazones with Cu (II) and vice versa by use of a copper selective electrode in Acetone-Water Mixture: Determination of the conditional formation constants of the cupric Thiosemicarbazonates. Microchemical Journal. 1985; 32:166–73.
42. Hymavathi M, Viswanatha C, Devanna N. A study on synthesis of novel chromogenic organic reagent 3,4-dihydroxy-5-methoxy benzaldehyde thiosemicarbazone and spectrophotometric determination of Cobalt (II) in presences of Triton X-100. Journal of Chemical and Pharmaceutical Research. 2014; 6(7): 2787–91.