Presented at the Research Congress 2013
De La Salle University Manila
March 7-9, 2013

QSAR MODELS FOR PREDICTING TOXICITIES OF MICROCYSTINS IN CYANOBACTERIA USING GETAWAY DESCRIPTORS

[2]Alex A. Tardaguila, [2]Jennifer C. Sy, and [1]Eric R. Punzalan

[1]Chemistry Department, De La Salle University, 2401 Taft Ave., Malate 1004, Manila [2]Physical Science Department, Pamantasan ng Lungsod ng Maynila, Intramuros 1002, Manila

Abstract: In this study, four QSAR models predicting LD_{50} of microcystins (MCs) were generated. The structures of the compounds were obtained from the literature. The data were divided into two sets: training set (N=20) and test set (N=3). All 3-D structures of these MCs were optimized by semi-empirical method, PM3 prior to calculations of 197 GETAWAY molecular descriptors. Multiple linear regression (MLR) using stepwise method was applied to determined significant descriptors and QSAR models. The method generated five significant descriptors and four QSAR models. The predictive powers of these models were evaluated by applying the following statistical parameters for the training set and test set: Pearson R (R), coefficient of determination (R^2), leave-one-out Q^2 (LOO Q^2), R^2_M, internal R^2 prediction (R^2_{pred}), root-mean-square error for prediction (RMSEP), goodness-of-fit chi squared test (χ^2). Internal validation of Model 4 (N = 20) revealed that it has the highest predictive power (R = 0.953; R^2 = 0.908; LOO Q^2 = 0.838; R^2_M = 0.743; RMSEP = 1.288, χ^2= 0.324). External validation (using the test set, N=3) showed that Model 4 (*LN LD_{50} = 52.554 (±5.803) - 2.589 (±1.409)H3e - 10.113 (±1.960) R8e + 245.321(±70.521)R2v-10.396(±2.460)H4p+317.169 (±121.689)R8p+*) has the highest predictive power (R = 0.925; R^2 = 0.855; R^2_{pred} = 0.780; R^2_M = 0.530, RMSEP = 0.233; χ^2= 0.037).

Keywords: QSAR, multiple linear regression, microcystins

1. INTRODUCTION

Microcystins (MCs) are cyclic nonribosomal peptides produced by cyanobacteria, and possess the generalized structure, cyclo (-D-Ala-**X**-D-MeAsp-**Y**-Adda-D-Glu-Mdha-) (Botes et al., 1982, 1984). They are cyanotoxins which are harmful to plants, animals and humans. Microcystins can strongly inhibit protein phosphatases type 1 (PP1) and 2A (PP2A) (MacKintosh et al., 1990). Furthermore, they cause serious damage to the liver (Chorus et al., 2000). To date, there are 86 microcystins reported in the literature, but only 36 have been evaluated for their toxicities (LD_{50}) (Zurawell, Chen, Burke, & Prepas, 2005). The difficulty in the establishment of their toxicities may be due to the separation, purification and scarcity of the MCs prior to laboratory testing. However, evaluating the toxicities of compounds is important because it gives us insight about the possible effects of unknown compounds to public health.

health.

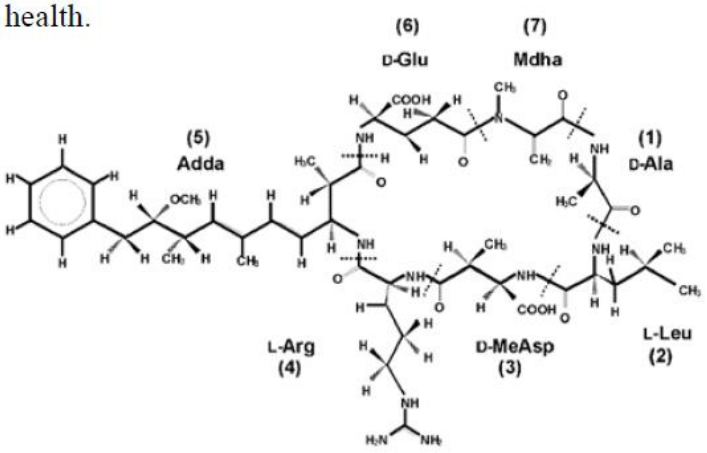

Figure 1. Structure of Microcystin-LR: Position (1) D-alanine (2) L-leucine (3) D-*erythro*-b-methylaspartic acid (4)L-Arganine (5) formula (2*S*,3*S*,8*S*,9*S*)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyldeca-4,6-dienoic acid (6) glutamic acid (7) *N*-methyldehydroalanine (Zurawell, Chen, Burke, & Prepas, 2005)

Alternatively, quantitative structure-activity relationship (QSAR) models allow us to predict toxicities using molecular descriptors of these compounds without undergoing tedious animal and laboratory testing (Verma, J, Khedkar, V. M., and Coutinho, C. E., 2010). Currently, there are more than 3000 molecular descriptors that are used in QSAR studies (Todeschini, R and Consonni,V.,2009). These are categorized into different types: 0D, 1D, 2D, and 3D molecular descriptors (Todeschini, R. et. al., 1994; Xue, L., Bajorath, J., 2000).

GETAWAY (GEometry, Topology, and Atom Weights AssemblY) is a set of 3D molecular descriptors that try to match 3D molecular geometry provided by the molecular influence matrix and atom relatedness by topology with chemical information by different atomic weighting schemes such as unit weights, mass, polarizability, electronegativity. GETAWAY descriptors have low or no degeneracy at all, which avoids getting the same value for a descriptor for more than one compound sharing the same structural features. The molecular influence matrix, **H** is defined by

$$\mathbf{H} = \mathbf{M} \cdot (\mathbf{M}^{T} \cdot \mathbf{M}) \cdot \mathbf{M}^{T} \qquad \text{(Eq. 1)}$$

where **M** is the molecular matrix. The resultant **A x A** matrix is invariant to rotation of the molecular coordinates. The diagonal elements $\mathbf{h}_{ij}$ are termed leverages and represent the influence of each atom in determining the shape of the molecule. Each off diagonal element $\mathbf{h}_{ij}$, represents the degree of of accessibility of the j'th atom to interactions with the i'th atom (V. Consonni and R. Todeschini, 2002).

In the present work, we report our preliminary QSAR models for predicting LD_{50} of microcystins found in cyanobacteria using GETAWAY descriptors, obtained from multiple linear regression (MLR) method.

2. METHODOLOGY

The structures of the 24 MCs were obtained from the literature (Zurawell, Chen, Burke, & Prepas, 2005). The data were divided into two sets: training set (N=20) and test set (N=4). All 3-D structures of these MCs were optimized by semi-empirical method, PM3 using Hyperchem (Hypercube Inc.) prior to calculations of 3D molecular descriptors, GETAWAY, using Dragon (Talete SRL). Multiple linear regression (MLR) using stepwise method was applied to determined significant descriptors. The predictive power of these models were evaluated by applying the following statistical parameters for the training set and test set: Pearson R (R), coefficient of determination (R^2), leave-one-out Q^2 (LOO Q^2), R^2_M, internal R^2 prediction (R^2_{pred}), root-mean-square error for prediction (RMSEP), goodness-of-fit chi squared test (χ^2). All statistical analyses were performed using SPSS software.

3. RESULTS AND DISCUSSION

In order to predict the toxicity of different MCs, GETAWAY descriptors were calculated to characterize the structural features of 20 compounds. The five descriptors (out of 197) selected by stepwise MLR method were employed to generate QSAR models for predicting LD_{50} of MCs. These 5 descriptors are listed in Table 1. These descriptors belong to H-GETAWAY descriptors, which have been calculated from the molecular influence matrix H, and R-GETAWAY descriptors, which are from the influence/distance matrix, R, where the elements of the molecular influence matrix (H) are combined with those of the geometry matrix (V. Consonni and R. Todeschini, 2002).

Table 1. GETAWAY Descriptors

GETAWAY Descriptors	Name
H3e	H autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities
R8e	R autocorrelation of lag 8 / weighted by atomic Sanderson electronegativities
R2v+	R maximal autocorrelation of lag 2 / weighted by atomic van der Waals volumes
H4p	H autocorrelation of lag 4 / weighted by atomic polarizabilities
R8p+	R maximal autocorrelation of lag 8 / weighted by atomic polarizabilities

Table 2. Pearson correlation matrix of parameters used in Model 1-4

	H3e	R8e	R2v+	H4p	R8p+
H3e	1				
R8e	-0.285	1			
R2v+	0.153	-0.151	1		
H4p	0.806(**)	-0.350	0.481(*)	1	
R8p+	0.053	0.210	0.590(**)	0.395	1

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Model 1 has H3e and R8e as significant descriptors; these explain 70.5% variance in LN LD_{50} of MCs. The two descriptors are not significantly correlated (Table 2), which indicates that the H3e and R8e are independent to each other. The coefficients of H3e and R8e have negative signs which show inverse relationship with the LN LD_{50}.

Model 1

LN LD_{50} = 58.228(±8.897) -7.365(±1.168)H3e - 7.277 (±2.726) R8e N =20, F = 20.306, SE =0 .375 (P < 0.001) R = 0.840 , R^2 = 0.705

Model 2 has H3e, R8e, R2v+, as significant descriptors and these explain 78.8% variance in LN LD_{50} of MCs. These three descriptors are uncorrelated.

Model 2

LN LD_{50} = 55.981 (±7.829) -7.664 (±1.028)H3e - 6.596(±2.726) R8e + 183.892(±73.553)R2v+ N =20, F = 19.802, SE =0.214 (P < 0.001) R = 0.888, R^2 = 0.788

Model 3 has four significant descriptors, H3e, R8e, R2v+, and H4p. These explain 86.3% of variance in LN LD_{50}. However, H4p is correlated with H3e, and H4p is correlated with R2v+. These indicate that these variables have multiplicative interaction effect on LN LD50.

Model 3

LN LD_{50} = 50.442 (±6.765) -3.991 (±1.533)H3e - 7.643 (±2.020) R8e + 316.970 (±76.462)R2v+ -7.336(±2.546)H4p N =20, F = 23.709, SE =0.147 (P < 0.001) R = 0.929, R^2 = 0.863

Model 4 has five significant descriptors, H3e, R8e, R2v+, H4p, and R8p+. The descriptors explain 90.8% of variance in LN LD_{50}. Like in model 3, model 4 has correlated variables, R2v+ and R8p+. The descriptors H3e, R8e and H4p are inversely related to LN LD50. On the other hand, R2v+ and R8p+ are directly related to LN LD_{50}.

Model 4

LN LD_{50} = 52.554 (±5.803) - 2.589 (±1.409)H3e - 10.113 (±1.960) R8e + 245.321(±70.521)R2v -10.396 (±2.460)H4p+317.169 (±121.689)R8p+ N =20, F = 27.652, SE =0.106 (P < 0.001) R = 0.953, R^2 = 0.908

We used Models 1-4 to predict the toxicities of the MCs in the test set (N=4) without considering the multiplicative interaction. Table 3 shows the predicted and observed LN LD_{50} for the MCs in the training set and test set.

Table 3. Observed and Predicted LN LD_{50} of Microcystins using Model 1-4

	Microcystins (MCs) (ref)	Obs. LN LD_{50}	Pred. LN LD_{50}(1)	Pred. LN LD_{50}(2)	Pred. LN LD_{50}(3)
1	**Microcystin-YR ***	**4.2485**	**4.9968**	**5.0618**	**4.7909**
2	Microcystin-YM(O)	4.0254	4.3356	4.3923	3.7680
3	Microcystin-YA.	4.1744	4.3010	4.2997	4.1326
4	Microcystin-M(O)R	6.6201	5.6849	6.5486	6.5699
5	**Microcystin-LY****	**4.4998**	**4.5103**	**4.2649**	**3.9018**
6	Microcystin-LR	3.9120	4.7024	4.6556	4.4217
7	Microcystin-LA	3.9120	3.5801	3.8253	3.9020
8	Microcystin-FR	5.5215	5.2193	5.6787	5.5668
9	Microcystin-AR.	5.5215	6.2821	6.0735	5.5820
10	[L-MeAla7]Microcystin-LR	4.4427	4.3358	4.2927	4.1601
11	**[D-MeAla7]microcystin-LR***	**4.6052**	**4.3872**	**4.3452**	**4.3531**
12	[DMAdda5]Microcystin-LR	4.5539	4.4179	4.3088	4.4306
13	[Dha7]Microcystin-RR	5.1930	5.6711	5.3596	5.3855
14	[Dha7]Microcystin-LR	5.5215	4.8894	4.8003	4.8951
15	[D-Asp3]Microcystin-RR	5.5215	5.5136	5.2455	5.1848
16	[D-Asp3]Microcystin-LR	3.9120	4.8050	4.7564	4.7856
17	**[D-Asp3]Microcystin-HtyR***	**5.4381**	**4.5720**	**4.9329**	**5.3871**
18	[D-Asp3, Dha7]Microcystin-LR	5.4161	5.1019	5.0098	5.3555
19	[D-Asp3, ADMAdda5]Microcystin-LR	5.0752	4.4468	4.4140	4.6309
20	[D-Asp3, (E)-Dhb7]Microcystin-RR	5.5215	5.1886	5.0815	5.3627
21	[D-Asp3, (E)-Dhb7]Microcystin-LR	4.2485	4.2598	4.1882	4.4499
22	[D-Asp3, (E)-Dhb7]Microcystin-HtyR	4.2485	4.4558	4.4617	4.6823
23	[ADMAdda5]Microcystin-LR	4.0943	4.5197	4.2982	4.1776
24	[6(Z)-Adda5]Microcystin-LR.hin	7.0901	6.8145	6.8348	7.0817

Pred. LN LD_{50}(4)

4.4156 4.1490 4.1798 6.7029 **3.9005** 4.3372 3.7319 5.2904 5.3098 4.0752 **4.3296** 4.5885 5.4707 5.0481 5.1340 4.5069 **5.1592** 5.3225 4.9994 5.3406 4.2604 4.5748 4.2319 7.2712

*Test Set (N = 4) ** Removed from Test Set

The predictive power of these models were evaluated by applying the following statistical parameters for the training set and test set: Pearson R (R), coefficient of determination (R^2), leave-one-out Q^2 (LOO Q^2), R^2_M, internal R^2 prediction (R^2_{pred}), root-mean-square error for prediction (RMSEP), goodness-of-fit chi squared test (χ^2). The values for each model are summarized in Table 4.

Internal validation of the training set (N = 20), revealed that the four models are generally have good prediction power ($R^2 > 0.6$, LOO $Q^2 > 0.6$, R^2_M.>0.5, RMSEP <1, $\chi^2 < 0.5$) (Veerasamy1, R., et. al., 2011). Among the four models, model 4 has the highest prediction power. Furthermore, the usefulness of the model was tested by its ability to predict the toxicities of MCs in a test set. These compounds in the test set were not used for generating QSAR models.

Table 4. Predictive Statistics for Training Set

Model	R	R^2	LOO Q^2	R^2_M	RMSEP	χ^2
1	0.840	0.705	0.605	0.573	1.245	
2	0.888	0.788	0.712	0.630	1.273	
3	0.929	0.863	0.898	0.699	1.264	
4	**0.953**	**0.908**	**0.838**	**0.743**	**1.288**	
	0.977	0.888	0.788	0.712	0.630	0.737
	0.929	0.863	0.898	0.699	1.264	0.499
	0.953	**0.908**	**0.838**	**0.743**	**1.288**	**0.324**

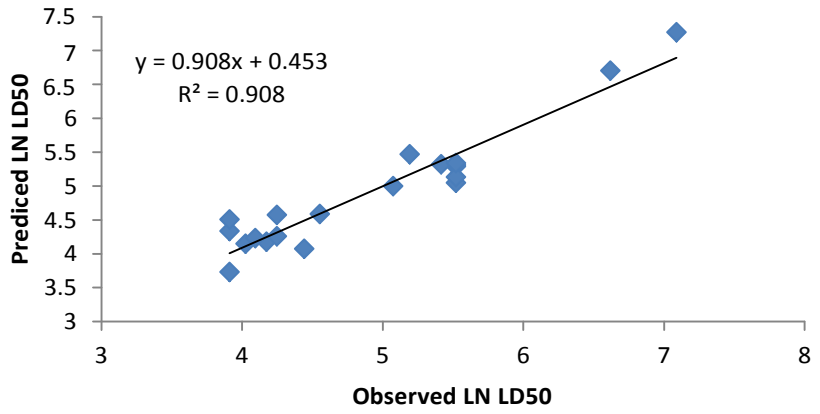

Figure 2. Plot of Predicted versus Observed LN LD_{50} of Microcystins (N=20) using Model 4

For external validation, we applied the four models to predict the toxicities of four MCs in the test set. Initial calculations revealed that MC 5 is poorly predicted by the model. This might be due to the structural features of MC 5 that were not similar to that of MCs in the training set. Hence, we decided to remove it from the test set. The remaining three MCs in the test set were re-evaluated

using the four QSAR models. Results show that model 4 successfully predicted the LN LD_{50} of the three MCs (Table 5).

Table 5. Predictive Statistics for Microcystins (1, 11, 17) in the Test Set

Model	R	R^2	R^2_{pred}	R^2_M	RMSEP	χ^2
1	-0.497	0.247	-0.647	0.033	0.116	0.280
2	0.058	0.003	-0.194	6.11×10^{-6}	0.146	0.217
3	0.744	0.553	0.563	0.184	0.279	0.084
4	**0.925**	**0.855**	**0.780**	**0.530**	**0.233**	**0.037**

5. CONCLUSIONS

The QSAR models for the toxicities of 20 microcystins have been obtained by MLR method and using three-dimensional descriptors, GETAWAY. The best MLR model has five descriptors . These descriptors are related to the three-dimensional distribution of electronegativities, van der Waals volumes and polarizabilities of atoms in MC. The selected descriptors effectively discriminate substituents of different amino acids composition of MCs. This model successfully predicted the toxicities of MCs in the test set (N=3).

6. REFERENCES

Botes, D. P., Kruger, H., and Viljoen, C. C. 1982. Isolation and characterization of four toxins from the blue-green alga, *Microcystis aeruginosa*. *Toxicon, 20*, 945–954.

Botes, D. P., Tuinman, A. A., Wessels, P. L., Viljoen, C. C., Kruger, H., Williams, D. H., Santikarn, S., Smith, R. J., and Hammond, S. J. (1984). The structure of cyanoginosin-LA, a cyclic heptapeptide toxin from the cyanobacterium *Microcystis aeruginosa*. *J. Chem. Soc. Perkin Trans., 1*, 2311–2318.

Chorus, I., Falconer, I. R., Salas, H. J., and Bartram, J. (2000). Health risks caused by freshwater cyanobacteria in recreational waters. *J. Toxicol. Environ. Health B*, *3*, 323–347.

Consonni V, Todeschini R, Pavan M. Gramatica, P. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J Chem Inf Comp Sci*., *42*, 693-705.

Consonni, V. and Todeschini, V.(2001). *Rational Approaches to Drug Design*. Prous Science, Barcelona.

HyperChem. Hypercube, Inc.: Waterloo, ON, Canada

MacKintosh, C., Beattie, K. A., Klumpp, S., Cohen, P., and Codd, G. A. (1990). Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. *FEBS Lett. 264*,187–192.

Talete SRL Dragon for Windows Version 5.4 (2006)

Todeschini, R and Consonni,V.(2009). *Molecular Descriptors for Chemoinformatics* (2 volumes). Wiley-VCH.

Todeschini, R., Lasagni, M., Marengo, E. (1994). New molecular descriptors for 2D- and 3D-structures. Theory. *J. Chemom.*, *8*, 263–273.

Xue, L., Bajorath, J. (2002). Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb.Chem. High Throughput Screening*, *3*, 363–372.

Verma, J, Khedkar, V. M., and Coutinho, C. E.(2010). 3D-QSAR in Drug Design - A Review. *Current Topics in Medicinal Chemistry 10*, 95-115.

Zurawell,R.W. Chen, H., Burke,J. M., Prepas, E. E. (2005). Hepatotoxic Cyanobacteria: A Review of the Biological Importance of Microcystins in Freshwater Environments. *Journal of Toxicology and Environmental Health, Part B, 8*, 1–37.

Veerasamy1, R., Rajak, H., Jain, A. Sivadasan, S., Varghese, C. P., and Agrawal, R. K. (2011). Validation of QSAR Models - Strategies and Importance. *International Journal of Drug Design and Discovery*, *2*, 511-519