Discovering New Competivive Dengue DEN2 NS2B/NS3 Inhibitors Using Similarity Searching

Neni Frimayanti, Sharifuddin M. Zain and Noorsaadah Abd. Rahman
Department of Chemistry, Faculty of Science
University of Malaya 50603 Lembah Pantai, Kuala Lumpur-Malaysia
nenifrimayanti@gmail.com, smzain@um.edu.my, noorsaadah@um.edu.my

***Abstract*—Several flaviviruses are important human pathogens, including dengue virus, a disease against which neither a vaccine nor specific antiviral therapies currently exist. QSAR study was carried out with the purpose of searching new competitive dengue inhibitors with similar properties to the existence inhibitors (i.e. data set). The approach began with the development of rigorously validated QSAR model obtained using multiple linear regression analysis (MLRA) with conventional correlation coefficient (r2) value of 0.82 and cross-validated correlation coefficient (r2CV) value of 0.65 and partial least squares (PLS) technique with r2 value of 0.82 and r2CV value of 0.74. The model showed a good correlative and predictive ability having a predictive correlation coefficient (r2pred) of 0.80. The validated QSAR models were then employed in mining the database which consisted of 45,917 compounds. The degree of similarity (based on Euclidean distance and Tanimoto coefficient) between the compounds probed from the data set and those in the database were calculated using the same set of descriptors in the QSAR model. A total of 7 compounds were short-listed and finally the inhibition constant of these compounds were calculated and predicted to be competitive dengue inhibitors.**

Keywords- QSAR; dengue DEN2 NS2B/NS;, Euclidean distance; Tanimoto coefficient

I. INTRODUCTION

Dengue virus (DV) and West Nile virus (WNV) are closely related members of the flaviviridae family. Dengue is a serious emerging disease which has become a global health burden in the recent decades. This virus is transmitted to man by a domestic mosquito, *Aedes aegypti* as the principal vector although some other species such as *Aedes albopictus* are also important [1]. WNV is transmitted by *Culex* mosquitos [2] and it is widely distributed around the world. In humans, WNV infections are usually asymptomatic or may cause a mild flu-like illness for a few days. This infection is as the West Nile fever. Currently, there are no approved vaccines to fight these diseases. In addition, there is also no known anti-viral therapy for dengue fever.

Both DV and WNV have a positive single strand RNA which codes for a single polyprotein precursor. The polyprotein precursor contains three structural proteins (C (capsid), M (membrane) and E (envelope) and seven non structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) [3, 4]. The DEN2 NS2B/NS3 and WNV NS2B/NS3 serine protease hold promise as a target for therapeutic intervention with small molecule as drugs [5].

The increasing spread and severity of DENV and WNV infections emphasize the importance of drug discovery strategies that is efficient and cost-effective. There are many approaches in developing new compounds that exhibit certain biological activities. In this paper, we present a method based on application of quantitative structure activity relationship (QSAR) models to screen a large chemical database.

QSAR study is based on a numerical description of molecular structure and uses statistics to obtain quantitative correlation to its properties. This methodology assumes that suitable sampling for these structural descriptors would provide all the information needed to understand their properties [6, 7]. Based on the premise that activity is related to the structure, these models can then be used to predict the activity of compounds not included in the model development stage.

In this study, we have developed QSAR models correlating structural characteristics of some panduratin A derivatives and pyrazole derivative with their inhibition constant value. The models were then applied to mining a larger chemical database to discover a set of credible competitive dengue DEN2 NS2B/NS3 inhibitors.

II. RESEARCH METHODOLOGIES

A. Data Set for Analysis

Structures of 30 inhibitors used in this study were obtained from the literature [5, 8, 9] and from previous results in our lab. Compounds with a variety of K_i value (μM) were used as data set to develop the QSAR model. The data set was divided into a training set (20 compounds) for QSAR model development and a prediction set (10 compounds) for model validation. The training set selection was performed by first sorting the list in increasing value of biological activity. Next, the list of compounds were divided into three groups (i.e. group I consisting of compounds nos. 1-10, group II with compounds nos. 11-20 and group III consisting of compounds nos. 21-30). The compounds in group I and III were assigned to the training set and compounds in group II were assigned to the prediction set. This method was chosen to produce more representative samples in the training set.

B. Development of QSAR Model

Molecular structures of those ligands were sketched using the ChemDraw 6.0 software (Cambridge) while Corina in TSAR 3.3 (Accelrys) software package was used to convert the structures into their 3D conformation. The geometries of these molecules were optimized using the Cosmic module of TSAR. The calculation were terminated when the energy difference or the energy gradient become smaller than 1 x 10^{-5} and 1 x 10^{-10} kcal/mol, respectively. Molecular descriptors were generated using TSAR 3.3 (Accelrys) for each compound. 316 descriptors were obtained from this calculation. These descriptors were then reduced to a smaller set of descriptors. These set should be information rich but as small as possible.

Correlation matrix was applied to select the best subset of descriptors to be included in the model by eliminating descriptors that are highly correlated with each other. Next step involves scaling descriptors which requires a thorough manipulation since there may be underlying relationship between these descriptors and it may not be possible to foresee the effects of this process. Range scaling also helps to avoid disproportional weightings of descriptors upon the Euclidean distance calculations in multidimensional descriptors space. The scaling was calculated as follow:

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Where, y_i is the scaled value, x_i is the original value while min (x) is the minimum of the collection of x object and max (x) is the maximum of the collection of x objects.

The selected descriptors were used to build the QSAR model. In this study, QSAR models were developed using multiple linear regression analysis (MLRA) and partial least squares (PLS) technique. The main goal of QSAR model development is to find the best set of descriptors which will produce a stable QSAR model and have the ability to predict properties of unknown compounds.

For the MLRA technique, values for *F*-to-enter and *F*-to-leave were set to 4. Cross-validation analysis was performed using the leave-one-out (LOO) method where one compound is removed from the dataset and its activity is calculated using the model derived from the rest of the dataset. The cross-validated r^2 and conventional r^2 that resulted in the lowest error of prediction were chosen. Unless otherwise stated, the default values for the other QSAR parameters were used.

The last step in QSAR model development is model validation. It is important to evaluate the robustness and the predictive capacity or validity of the model before using the model on the interpretation and prediction of the biological activity. The biological activities of compounds in the predicted set were calculated using the model produced by the training set.

C. Database Mining and Similarity Searching

Thirty compounds in the data set with inhibition constant values (K_i) were selected as probes to calculate the degree of similarity with compounds in a larger database. Euclidean distance was employed to measure this similarity using the same set of descriptors that appeared in the QSAR model. The degree of similarity (i.e. Euclidean distance) can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{n=1}^{N} (X_{in} - X_{jn})^2} \quad (2)$$

where, d_{ij} is the distance between any two compounds, X_{in} and X_{jn} are the values of the *n*th descriptor for compounds *i* and *j*. Compounds in the database within the chosen similarity cutoff value (0.5 Euclidean distance units in a multidimensional space) were considered as hits and can be further investigated. Another similarity concept, the Tanimoto similarity coefficient, was also applied. For real-valued properties the Tanimoto similarity is defined as:

$$\text{Tanimoto} = \frac{\sum_{i=1}^{i=N} X_{iA} X_{iB}}{\sum_{i=1}^{i=N} (X_{iA})^2 + \sum_{i=1}^{i=N} (X_{iB})^2 - \sum_{i=1}^{i=N} X_{iA} X_{iB}} \quad (3)$$

Where, X_{iA} is the value of property i of molecule A, and X_{iB} is the value of property i of molecule B.

The concept of applicability domain or similarity threshold that is specific to each particular QSAR model was applied to the database mining [10, 11] to avoid making prediction for compounds that differ substantially from the training set molecules. The similarity threshold was calculated as follow:

$$D_T = \bar{y} + Z\sigma \quad (4)$$

Where $\bar{y}$ the average Euclidean distance between each compound, σ is the standard deviation of these Euclidean distances and *Z* is an arbitrary parameter to control the significance level (0.5). If the distance of a particular compound from the probe molecules is less than this threshold value, the prediction is considered reliable.

III. Results and Discussion

A. QSAR Modeling

The equation for the QSAR model is:

Log $1/K_i$ = -1.19 * verloop B3 (subst. 2) – 2.20 * inertia moment 2 length – 0.92 * vamp LUMO – 0.78 * vamp polarization YZ + 4.01 (5)

The cross-validated coefficient (r^2CV) defines the goodness of prediction whereas the non-cross-validated conventional correlation coefficient (r^2) indicates the goodness of fit of a QSAR model [12]. The F test value stands for the degree of statistical confidence. The statistical output of the MLRA model is presented in Table 1, a cross validated coefficient of 0.65 was obtained using the leave one out cross validation procedure.

TABLE I. STATISTICAL OUTPUT OF MLRA MODEL

Statistical output	Value
R^2	0.82
Cross validation r^2 (CV)	0.65
F- value	20.67
F-probability	3.42e-007
Standard error of estimate *(SEE)*	0.37
Residual sum of square (RSS)	2.46
Predictive sum of square (PRESS)	4.86

The best QSAR model developed using MLRA technique has r^2 of 0.82 and r^2 (CV) of 0.65. This indicates a very good internal predictive capability of the developed model. The model also exhibited a non cross-validated correlation coefficient of 0.82. The high value of this parameter adds to its usefulness as a predictive tool.

From the QSAR model as presented above, the electrostatic parameter (i.e. Vamp LUMO and Vamp polarization YZ) are negatively correlated with the inhibitory constant. The energy of LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule toward attack by nucleophiles. Both HOMO and LUMO energies have been considered as indicators of drug activity [13]. The inhibitory activity improves with an increase in the electrostatic parameter. The statistical significance of the parameter derived from QSAR model is presented in Table 2 and brief descriptions about those descriptors are shown in Table 3.

The development of QSAR model by using MLRA technique can be accepted, if the models have r^2(CV) greater than 0.5 and r^2 greater than 0.6 [14]. Thus with r^2 value of 0.82 and r^2 *(CV)* value of 0.65, the model generated above is presumably capable of predicting the biological activities of compounds which were not included in the model development process. A plot of experimental *vs.* predicted K_i is shown in Figure 1 while a plot of standard residual *vs.* predicted value (residual plot) is presented in Figure 2. These two plots are important to graphically observe the predictive capability of QSAR. Shorter height of the residual and the fact that the training set molecules are on or near the best fit line, as shown in Figure 1, further add to the usefulness of the developed QSAR.

TABLE II. STATISTICAL SIGNIFICANCE OF PARAMETER

Descriptors	Regression coefficient[a]	Jackife SE[b]	Covariance SE[c]	t-value[d]	t-probability[e]
Verloop B3	-1.19	0.53	0.29	-4.15	6.1×10^{-4}
Inertia moment 2 length	-2.20	0.24	0.25	-8.79	6.2×10^{-5}
Vamp LUMO	-0.93	0.22	0.29	-3.18	5.1×10^{-3}
Vamp polarization YZ	-0.78	0.33	0.29	-2.66	1.6×10^{-2}

[a] The regression coefficient for each variable in the equation
[b] An estimate of the standard error of each regression coefficient derived from a Jacknife procedure on the final regression model
[c] Estimate of the standard error of each regression coefficient derived from the covariance matrix
[d] Significance of each variable included in the final model
[e] Statistical significance for t-values

TABLE III. DESCRIPTORS WHICH WERE INCLUDED IN THE MLRA MODEL

Descriptor	Symbol	Explanation
Verloop parameter	Verloop B3	The distance from the axis of the attachment bond, measured perpendicularly to the edge of the substituents.
Molecular attributes	Inertia moment 2 length	Indicates the strength and orientation behaviors of molecule in an electrostatic field.
Electrostatic parameter	Vamp lumo, vamp polarization YZ	Properties of molecule arising from the interaction between a charge probe such as positive unit point reflecting a proton and target molecule.

PLS technique has also been used to develop the QSAR model. PLS also can be used quite effectively as a tool for interpreting QSAR model and that the information extracted is much more detailed [15]. The PLS routine in TSAR stop the iteration if a model has one the following criterion [16, 17]: the lowest value of predictive sum of squares (PRESS) or when PRESS value starts to increase. In this study, the QSAR model generated using PLS has four dimensions. The PLS dimension three was selected as the best PLS QSAR model with r^2 value of 0.82 and r^2 *(CV)* value of 0.74. The highest r^2*(CV)* and the lowest PRESS value of this dimension indicated that this model is more stable and suitable for predicting the inhibitory constant of compounds that were not included in the training set. The statistical output of this PLS model is presented in Table 4.

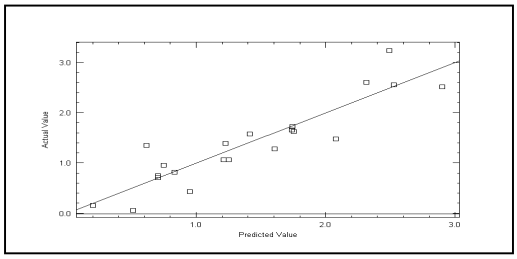


Figure 1. Plot of predicted value *vs.* experimental K_i for MLRA technique

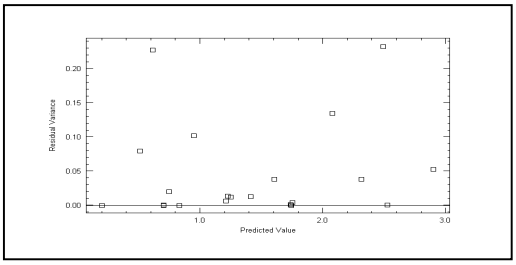


Figure 2. Plot of predicted value *vs.* standard residual for MLRA model

TABLE IV. STATISTICAL OUTPUT OF PLS MODEL

Statistical output	value
Fraction of variance	0.82
Cross validation $r^2(CV)$	0.74
Residual sum of square (RSS)	3.93
Predictive sum of square (PRESS)	5.69

Both models were validated by predicting the inhibitory activity of 10 compounds excluded during the model development process (prediction set). The correlation coefficient (r^2) between predicted and experimental values was also calculated. A predictive correlation coefficient r^2 value of 0.80 obtained for both of these QSAR models the usefulness of the developed QSAR models in predicting activities of molecules not included in its derivation. Alternatively, to further evaluate the significance of the developed model is to test it for statistical stability. For this, the standard error of estimate and predictive residual sum of squares may be employed. Low values of the standard error of estimate (SEE) at 0.37 and predictive residual sum of squares *(PRESS)* for the MLRA model at 4.86 further add to the statistical significance of the developed model. However, the obtained PLS model (i.e. PLS dimension three) has higher *PRESS* value of 5.69. This seemed to indicate the PLS model to be unstable for predicting unknown compounds in the prediction set.

B. Database Mining

A set of compounds in the data set with their inhibitory constant (K_i, μM) were used as the similarity probes for database mining. The degree of similarity based on Euclidean distance and Tanimoto coefficient between compounds in the data set and those in database were calculated using the same set of descriptors used in the QSAR models. Since the limiting value of the distance is 0.5, compounds with the distances higher than 0.5 were rejected and classified as outliers.

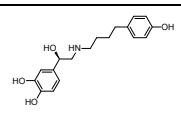
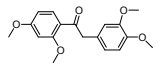
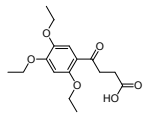
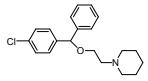
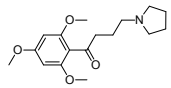
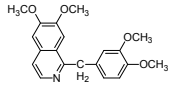
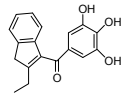
In the first round of screening, out of the 45,917 compounds in the database [18], 526 compounds were found to be within the chosen similarity cutoff value obtained from the 30 probe molecules. These compounds were further subjected to consensus hits criteria (i.e. molecules that consistently appear in both models), reducing the candidate to 469 compounds. Subsequently, the 469 compounds were subjected to the applicability domain criteria (similarity threshold) and the number of possible candidates was further narrowed down to 51 compounds. The inhibition constants of these 51 candidates were predicted by using the two best QSAR models where each model has its specific applicability domain criteria. Using the Tanimoto similarity coefficient, 486 compounds were selected as initial hits and 468 compounds were selected by both QSAR models. Applying the applicability domain, the numbers of possible candidates were further narrowed to 34 compounds.

Prediction of activity should be made within the domain of an appropriately validated QSAR. To achieve this, the applicability domain of a model should be defined, a QSAR model based on the mechanism of action approach tend to rely on expert judgment to define the domain. The applicability domain may be defined by the general properties on a more detailed structural basis for specific toxicities. For a prediction to be valid, the compound must fall within the applicability domain of the models since the applicability domain would prevent predictions for compounds that differ substantially from the training set molecules [19, 20]. Applicability domain also can be used to exclude compounds that are too dissimilar to the training set to make any reliable prediction of their activity. The list of selected compounds earlier (i.e. 51 compounds and 34 compounds) were further refined by finding compounds that appear within both similarity distances. This resulted in as many as 7 compounds (Table 6).

IV. CONCLUSIONS

Quantitative structure activity relationship (QSAR) approach has been used to develop models with high predictive power to predict the inhibition constant (Ki) value of compounds that are not included in the training set. The MLRA technique has been used to develop a good QSAR model which revealed that the inhibitory activity of the DEN2 virus to be predominantly influenced by electrostatic properties. The application of QSAR model in database mining has enabled us to predict potential compounds as new leads as competitive inhibitors for NS2B/NS3 serine protease. The list can be further refined by restricting to compounds which are similar in structure to those in the training set. Currently, laboratory experiments are being carried out to validate these QSAR models.

TABLE V. SELECTED COMPOUNDS WITH THEIR PREDICTED KI VALUE

No	Database ID	Structure	K_i (μM)
1	128470166		17.37
2	Stock1N-10681		28.18
3	041826920		18.62
4	003703762		19.25
5	055837257		17.37
6	000058742		17.78
7	000054035		30.19

ACKNOWLEDGMENT

We thank University of Malaya for financial support through University Grant Research Scheme (UMRG) no. RG012/09BIO.

REFERENCES

[1] WHO. Dengue and Dengue Hemorrhagic fever. World Health Organization. April, 2002 (http://www.who.int/mediacentre/factsheets/fs117/en)

[2] Kanin, W., Somsak, P., Wolfgang, S., Sirirat, K, Homology modeling and molecular dynamics simulations of dengue virus NS2B/NS3 protease: insight into molecular interaction. John Wiley & Sons Ltd. Wiley InterScience, 2009.

[3] Chambers T. J., Nestorowicz A, Amberg S. M., Rice C. M, "Mutagenesis of the yellow fever virus NS2B protein: effects on proteolytic processing, NS2B-NS3 complex formation, and viral replication," J. Virol, vol. 67, 1993, pp. 6797-6807.

[4] Irie, K., Mohan, P. M., Sasaguri, Y., Putnak, R., Padmanabhan, R, "Sequence analysis of cloned dengue virus type 2 genome (New Guinea-C strain)," Gene, vol. 75, 1989, pp. 197-211.

[5] Shyama, S., Sergey, A. S., Boris, I. R., Ananda, H., Ying, S., Alex, Y. S., Nicholas, D. P,"Structure-activity relatinship and improved hydrolytic stability of pyrazole derivatives that are allosteric inhibitors of west nile virus NS2B-NS3 proteinase," Biorg. Med. Chem. Lett, vol. 19, 2000, pp. 5773-5777.

[6] N. Dessalew and P. V. Bharatam, "3D-QSAR and molecular docking study on bisarylmaleimide series as glycogen synthase kinase 3, cyclin dependent kinase 2 and cyclin dependent kinase 4 inhibitors: An insight into the criteria for selectivity," Eur. J. Med. Chem, vol. 42, 2007, pp. 1014–1027.

[7] Leach, A. R., Gillet, V. J, An introduction to chemoinformatics, Kluwer Academic, London; 2003, pp. 77−101.

[8] Tan, S. K., Pippen, R., Yusof, R., Ibrahim, H., Khalid, Noorsaadah, A. R, "Inhibitory activity of cyclohexenyl chalcone derivatives and flavanoids of fingerroot, Boesenbergia rotunda (L), towards dengue-2 virus NS3 protease," Bioorg. Med. Chem. Lett, vol. 16, 2006, pp. 3337-3340.

[9] Vannakambadi, K. G., Nik, M., Ken, J., Chi-Hao, L., Radhakrishnan, P., Krishna, H. M. M, "Identification and characterization of nonsubstrate based inhibitors of the essential dengue and west nile virus proteases," J. Bioorg. Med. Chem, vol. 13, 2005, pp. 257-264.

[10] Kubinyi, H, "QSAR and 3D QSAR in drug design 1. methodology," Drug Discovery Today, vol. 2, 1997, pp. 457-467.

[11] Golbraikh, A., Shen, M., Xiao, Z. Y., Xiao, Y. D., Lee, K. H., Tropsha, A, "Rational selection of training set and test set for the development of validated QSAR models," J. Comput-Aided Mol. Des, vol. 17, 2003, pp. 241-253.

[12] Nigus, D, "QSAR study on aminophenylbenzamides and acrylamides as histone deacetylase inhibitors: an insight into the structural basis of antiproliferative activity," J. Med. Res., vol. 16, 2007, pp. 449-460.

[13] Lohray, B. B., Gandhi, N., Srivastava, B. K., Lohray, V. B, "3D QSAR studies of N-4-arylacrylloylpiperazin-1-yl-phenyl-oxazolidonones: A novel class of antibacterial agents," Bioorg. Med. Chem. Lett, vol. 16, 2006, pp. 3817-3823.

[14] Golbraikh, A., Tropsha, A, "Predictive QSAR modeling diversity sampling of experimental datasets for the training and test set selection," J. Comput-Aided Mol Des., vol. 5, 2002, pp. 231-243.

[15] Tang, K., Li, T, "Comparison of different partial least squares methods in QSAR," Anal. Chem. Acta, vol. 476, 2003, pp. 75-92.

[16] Oxford molecular, TSAR 3.3 for windows reference guide, UK: oxford molecular, Ltd. 2000.

[17] Hawkins, D. M., Basak, S. C., and Shi, X, "QSAR with few compounds and many features," J. Chem. Inf. Comput. Sci,vol. 41, 2001, pp. 663-670.

[18] http://bioinformatics.charite.de (accessed on 19th May 2009).

[19] S. Zhang, L. Wei, K. Bostow, W. Zheng, A. Brossi, K. H. Lee, A. Tropsha, "Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents," J. Comput Aided Mol. Des, vol. 21, 2007, pp. 97-112.

[20] M. Cronin, "Opportunities for Computer Aided Prediction of Toxicity in Drug Discovery," A Report Computational Chemistry, School of Pharmacy and Chemistry. Liverpool, John Moores University. 2002.