Malaysian Journal of Analytical Sciences (MJAS)
Published by Malaysian Analytical Sciences Society
ANALIS
Persatuan Sains Analisis Malaysia

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) STUDY OF NEWLY SYNTHESIZED CARBONYL THIOUREA DERIVATIVES ON *Acanthamoeba* sp.

(Kajian Kuantitatif Hubungan Struktur-Aktiviti (QSAR) Terhadap Terbitan Karbonil Tiourea Hasilan Sintesis Baru Terhadap *Acanthamoeba* sp.)

Maizatul Akma Ibrahim[1]*, Nor Hafizah Zakaria[1], Mohd Sukeri Mohd Yusof[2]

[1]Department of Plant Science, Kulliyyah of Science,
International Islamic University Malaysia, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia
[2]Faculty of Science and Marine Environment,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

**Corresponding author: maizatulakma@iium.edu.my*

Received: 3 December 2021; Accepted: 23 March 2022; Published: 27 June 2022

Abstract

This research aims to build a mathematical quantitative structure-activity relationship (QSAR) model, which could relate the relationship between newly-synthesized carbonyl thiourea derivatives with their anti-amoebic activities. Therefore, in this study, inhibition concentration of 50% cells population (IC_{50}) was evaluated for 44 carbonyl thiourea derivatives on pathogenic *Acanthamoeba* sp. (Hospital Kuala Lumpur isolate). QSAR analysis was conducted using the obtained IC_{50} data with additional 4 thiourea compounds of the same group from our previous work by applying three linear regression techniques namely stepwise-MLR, GA-MLR, and GA-PLS. Results showed that these thiourea derivatives are positively active against the tested *Acanthamoeba* sp. with IC_{50} values ranging from 2.56 to 7.81 µg/mL. From the evaluation of the obtained models, the GA-PLS technique is found to be the best model due to its best predictive ability. The final equation of GA-PLS model gave good statistical output with values of r^2 = 0.827, r^2_{cv} = 0.682 *RMSEC*=0.047, *RMSECV*=0.064, and r^2_{test} =0.790 and *RMSEP*=0.051. Y-randomization test has confirmed that the model did not occur from the chance of correlation with r^2 = 0.015-0.372. Small residual with values less than 0.25 from the prediction in the test set proves the robustness of the model. The information generated from this study will provide an insight into designing a new lead compound from carbonyl thiourea containing highly potential anti-amoebic properties.

Keywords: thiourea derivatives, *Acanthamoeba* sp., IC_{50}, anti-amoebic activity, QSAR models

Abstrak

Kajian ini mensasarkan pembinaan model matematik kuantitatif hubungan struktur-aktiviti (QSAR) yang memberi hubungkait antara terhadap terbitan karbonil tiourea hasilan sintesis baru dengan aktiviti anti-amebik. Oleh itu, kepekatan penghambatan 50% populasi sel (IC_{50}) dikaji ke atas 44 sebatian karbonil tiourea terhadap *Acanthamoeba* sp. (Hospital Kuala Lumpur isolat) berstatus patogenik. Analisa hubungan struktur-aktiviti kuantitatif dijalankan menggunakan data IC_{50} yang diperolehi bersama 4 sebatian tambahan kumpulan sama dari hasil kerja kami sebelum ini, dengan mengaplikasi tiga teknik regresi linear, iaitu stepwise-MLR, GA-MLR dan GA-PLS dijalankan. Hasil kajian menunjukkan bahawa sebatian tiourea ini aktif secara positif terhadap *Acanthamoeba* sp.yang diuji dengan nilai IC_{50} antara 2.56 hingga 7.81 µg/mL. Penilaian terhadap semua model QSAR yang dibina dalam kajian ini menunjukkan teknik GA-PLS adalah model yang terbaik kerana kemampuan ramalannya yang terbaik. Persamaan akhir untuk model GA-PLS menunjukkan output statistik yang baik dengan nilai r^2 = 0.827, r^2_{cv} = 0.682, *RMSEC* = 0.047, *RMSECV* = 0.064, r^2_{test} = 0.790 dan *RMSEP* = 0.051. Ujian perawakan-y mengesahkan bahawa model tersebut tidak terhasil secara kebetulan dengan r^2 = 0.015-0.372. Baki kecil dengan nilai kurang dari 0.25 dari ramalan set ujian membuktikan kekuatan model tersebut. Data terbina dari kajian ini akan memberi maklumat untuk mencipta sebatian penting baru dari tiourea karbonil yang mempunyai aktiviti anti-amebik yang berpotensi tinggi.

Kata kunci: tiourea terbitan, *Acanthamoeba* sp., IC_{50}, aktiviti anti-amebik, model QSAR

Introduction

Acanthamoeba is a pathogenic protozoan that is ubiquitously found in the environment. It can cause *Acanthamoeba* keratitis, which commonly occurs in contact lens users [1]. There are several antimicrobial agents such as chlorhexidine gluconate and polyhexamethylene biguanide that have been used to treat the disease but according to Vontobel et al. [2], both agents do not readily penetrate the cornea of the eyes, which require months of topical administration, making them ineffective for keratitis treatment. Reports showed resistance of *Acanthamoeba* towards these antiseptics, especially during the later stage of infection, which makes it difficult to be treated [3]. Therefore, the development of new potential antiamoebic agents is demanded to overcome these problems. Thiourea, an organosulfur compound has been reported to contain diverse biomedical benefits such as antibacterial, anticancer, antifungal, anti-inflammatory, antithyroid, herbicidal, and antitubercular properties [4, 5, 6]. Naz et al. evaluated the antibacterial activity of thiourea derivatives and found that these compounds significantly inhibited several pathogenic bacteria including *E. faecalis*, *P. aeruginosa*, *S. typhi*, and *K. pneumoniae* [7]. In another study, Keche and Kemble also reported the antimicrobial activity of novel thiourea compounds against several selected bacteria and fungi and they revealed that these compounds have promising antimicrobial activity [8]. Nevertheless, there is still lacking literature regarding the anti-amoebic properties of thiourea derivatives on *Acanthamoeba*. Therefore, a cytotoxicity test was conducted in this study to determine the lC$_{50}$ values for the newly synthesized thiourea derivatives on *Acanthamoeba*.

A QSAR study was applied to analyze the molecular structures of the synthesized compounds to correlate with their anti-amoebic activities and allow for the effects of compounds of interest to be predicted [9]. In this study, QSAR methods were used to quantitatively study the relationship between the presented synthesized carbonyl thiourea analogs with their anti-amoebic activity. The approach applied multiple linear regression (MLR) and partial least square (PLS) to build the QSAR equation models. Conventional stepwise methods and genetic algorithm (GA) were employed to choose the best descriptors subset in the model development. Internal and external validations were carried out to evaluate the robustness of the generated models. From the information gained in this study, optimized thiourea-based compounds that work best against pathogenic *Acanthamoeba* could be predicted and developed as new agents to treat *Acanthamoeba* keratitis.

Materials and Methods

In this study, a set of carbonyls thiourea derivatives consisting of 40 compounds were synthesized and characterized at the Faculty of Science and Marine Environment, Universiti Malaysia Terengganu. These synthesized thiourea derivatives were confirmed by spectral studies of Fourier Transformation Infrared (FT-IR) spectroscopy and ^{1}H and ^{13}C Nuclear Magnetic Resonance (NMR). The preparation of M1–M44 compounds is based on the synthesis method of the previous study with additions of compounds labeled as M7, M8, M25, and M26 attained from the same work [10]. The other synthesized thiourea and their molecular weights are listed in Table 1.

Table 1. Molecular structures of the newly-synthesized carbonyl thiourea derivatives

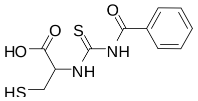
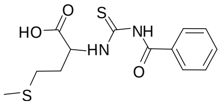
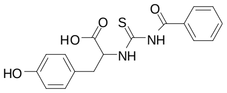
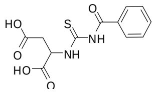
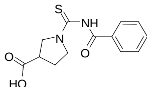
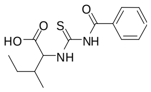
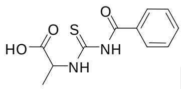
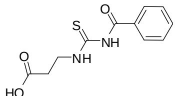
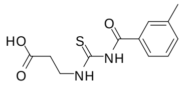
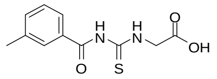
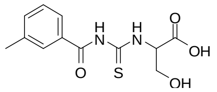
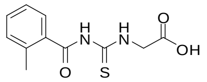
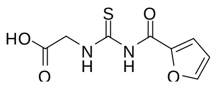
Code	**Chemical Name**	**Molecular Weight**	**Structure**
M1	2-(3-benzoyl thioureido)-3-mercapto propanoic acid	284.35	
M2	2-(3-benzoyl thioureido)-4-(methylthio)butanoic acid	312.41	
M3	2-(3-benzoyl thioureido)-3-(4-hydroxyphenyl)propanoic acid	344.38	
M4	2-(3-benzoyl thioureido)succinic acid	296.30	
M5	1-(benzoyl carbamothioyl) pyrrolodine-3-carboxyl acid	278.33	
M6	2-(3-benzoyl thioureido)-3-methyl pentanoic acid	294.37	
M7	2-(3-benzoyl thioureido)propanoic acid	252.29	 [8]
M8	3-(3-benzoyl thioureido)propanoic acid	252.29	 [8]
M9	3-[3-(3-methylbenzoyl) thioureido] propanoic acid	266.32	
M10	2-[3-(3-methylbenzoyl) thioureido] acetic acid	252.29	
M11	3-hydroxy-2-[3-(3-methylbenzoyl) thioureido]propanoic acid	282.32	
M12	2-[3-(2-methylbenzoyl)thioureido] acetic acid	252.29	
M13	2-(3-furan-2-carbonyl thioureido) acetic acid	228.23	

Table 1 (cont'd). Molecular structures of the newly-synthesized carbonyl thiourea derivatives

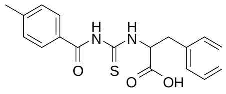
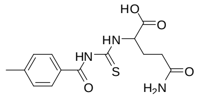
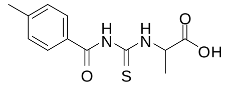
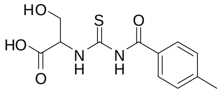
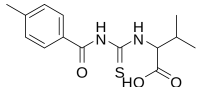
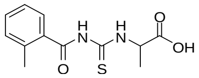
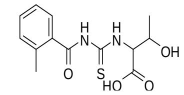
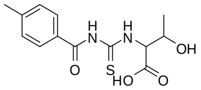
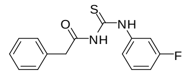
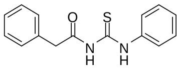
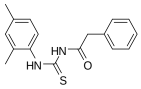
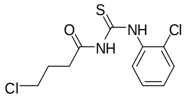
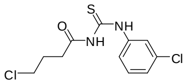
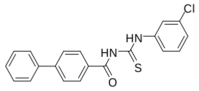
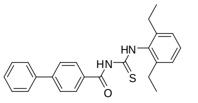
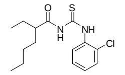
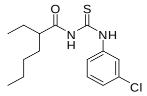
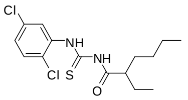
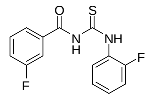
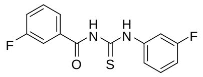
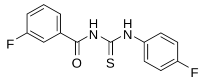
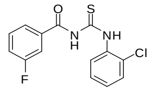
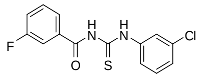
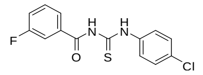
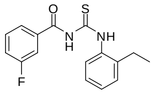
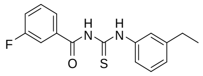
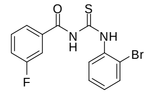
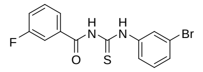
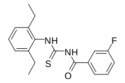
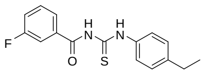
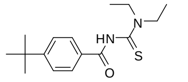
Code	Chemical Name	Molecular Weight	Structure
M14	2-[3-(4-methylbenzoyl) thioureido]-3-phenylpropanoic acid	342.41	
M15	5-amino-2-[3-(4-methylbenzoyl) thioureido]-5-oxopentanoic acid	323.37	
M16	2-[3-(4-methylbenzoyl) thioureido] propanoic acid	266.32	
M17	3-hydroxy-2-[3-(4-methylbenzoyl) thioureido]propanoic acid	282.32	
M18	3-methyl-2-[3-(4-methylbenzoyl) thioureido]butanoic acid	294.37	
M19	2-[3-(2-methylbenzoyl) thioureido] propanoic acid	266.32	
M20	3-hydroxy-2-[3-(2-methylbenzoyl) thioureido]butanoic acid	296.34	
M21	3-hydroxy-2-[3-(4-methylbenzoyl) thioureido]butanoic acid	296.34	
M22	*N*-(3-fluorophenyl)-*N'*-phenylacetyl thiourea	288.34	
M23	*N*-phenyl-*N'*-phenylacetyl thiourea	270.35	
M24	*N*-(2,4-dimethylphenyl)-*N'*-phenyl acetylthiourea	298.40	
M25	*N*-(2-chlorophenyl)-*N'*-(4-chlorobutanoyl) thiourea	291.20	 [8]
M26	*N*-(3-chlorophenyl)-*N'*-(4-chlorobutanoyl) thiourea	291.20	 [8]
M27	*N*-(3-chlorophenyl)-*N'*-(biphenyl-4-yl) carbamoyl thiourea	366.86	

Table 1 (cont’d). Molecular structures of the newly-synthesized carbonyl thiourea derivatives

Code	Chemical Name	Molecular Weight	Structure
M28	*N*-(2,6-diethylphenyl)-N’-(biphenyl-4-yl)carbamoyl thiourea	388.53	
M29	*N*-(2-chlorophenyl)-*N’*-(2-ethyl hexanoyl) thiourea	312.86	
M30	*N*-(3-chlorophenyl)-*N’*-(2-ethyl hexanoyl) thiourea	312.86	
M31	*N*-(2,5-dichlorophenyl)-*N’*-(2-ethyl hexanoyl) thiourea	347.40	
M32	*N*-(3-fluorobenzoyl)-*N’*-(2-fluoro phenyl) thiourea	292.30	
M33	*N*-(3-fluorobenzoyl)-*N’*-(3-fluoro phenyl) thiourea	292.30	
M34	*N*-(3-fluorobenzoyl)-*N’*-(4-fluoro phenyl) thiourea	292.30	
M35	*N*-(3-fluorobenzoyl)-*N’*-(2-chloro phenyl) thiourea	308.76	
M36	*N*-(3-fluorobenzoyl)-*N’*-(3-chloro phenyl) thiourea	308.76	
M37	*N*-(3-fluorobenzoyl)-*N’*-(4-chloro phenyl) thiourea	308.76	
M38	*N*-(3-fluorobenzoyl)-*N’*-(2-ethyl phenyl) thiourea	302.37	
M39	*N*-(3-fluorobenzoyl)-*N’*-(3-ethyl phenyl) thiourea	302.37	
M40	*N*-(3-fluorobenzoyl)-*N’*-(2-bromo phenyl) thiourea	353.21	
M41	*N*-(3-fluorobenzoyl)-*N’*-(3-bromo phenyl) thiourea	353.21	
M42	*N*-(3-fluorobenzoyl)-*N’*-(2,6-diethyl phenyl) thiourea	330.42	
M43	*N*-(3-fluorobenzoyl)-*N’*-(4-ethyl phenyl)thiourea	302.37	
M44	*N*-(4-*tert*-butylbenzoyl)-*N’*-diethyl phenyl thiourea	292.44	

Determination of lC$_{50}$ values

Thiourea derivatives were prepared and chlorhexidine gluconate (Raza Manufacturing, Malaysia) was used as the positive control whereas 10^4 cells/mL of healthy *Acanthamoeba* sp. without any treatment was used as the negative control [10]. The plates were incubated at 30°C for 72 hours. After incubation, the staining process was performed using the eosin dye method [11]. The absorbance was measured at 490 nm using an ELISA plate reader (Tecan, Australia). The readings were plotted in GraphPad Prism software (version 5.03) (San Diego, USA) to give a non-linear sigmoidal dose-response curve in which the cytotoxicity study was expressed as 50% cytotoxic dose (IC$_{50}$). T-test (SPSS, version 11.5, USA) was used to compare mean values between untreated and treated cultures, and $p < 0.05$ is considered statistically significant.

Data set selection for QSAR study

The data set that contains thiourea derivatives with anti-amoebic activity is shown in Table 2. The IC$_{50}$ values in the molar (M) unit were converted to pIC$_{50}$ for the convenience of computational work. The data were divided into training and test set. The training set is comprised of 30 thiourea compounds while the test set consisted of the remaining 14 compounds. Compound M7 was later removed from the data set as it was detected to be an outlier throughout the model building by all selected methods of stepwise multiple linear regression (stepwise-MLR), genetic algorithm multiple linear regression (GA-MLR), and genetic algorithm partial least square (GA-PLS).

Table 2. Functional groups of carbonyl thiourea analogs with their anti-amoebic activity

Code	**R$_1$**	**R$_2$**	**IC$_{50}$ (μM)**	**pIC$_{50}$ (M)**
M1	$C_4H_8O_2S$	C_8H_8O	9.00	5.05
M2	$C_6H_{12}O_2S$	C_8H_8O	8.63	5.06
M3	$C_{10}H_{12}O_3$	C_8H_8O	9.65	5.02
M4	$C_5H_8O_4$	C_8H_8O	13.35	4.87
M5	$C_5H_{10}O_2$	C_8H_8O	11.06	4.96
M6	$C_7H_{14}O_2$	C_8H_8O	13.24	4.88
M7	$C_4H_8O_2$	C_8H_8O	10.85	4.97
M8	$C_4H_8O_2$	C_8H_8O	11.59	4.94
M9	$C_4H_8O_2$	$C_9H_{10}O$	19.26	4.72
M10	$C_9H_{10}O$	$C_3H_6O_2$	22.35	4.65
M11	$C_9H_{10}O$	$C_4H_8O_3$	18.52	4.73
M12	$C_9H_{10}O$	$C_3H_6O_2$	19.58	4.71
M13	$C_3H_6O_2$	$C_6H_6O_2$	20.16	4.70
M14	$C_9H_{10}O$	$C_{10}H_{12}O_2$	14.47	4.84
M15	$C_9H_{10}O$	$C_6H_{11}NO_3$	17.04	4.80
M16	$C_9H_{10}O$	$C_4H_8O_2$	21.79	4.66
M17	$C_9H_{10}O$	$C_4H_8O_3$	18.75	4.73
M18	$C_9H_{10}O$	$C_6H_{12}O_2$	17.20	4.76
M19	$C_9H_{10}O$	$C_4H_8O_2$	19.95	4.70
M20	$C_9H_{10}O$	$C_5H_{10}O_3$	15.58	4.81
M21	$C_9H_{10}O$	$C_5H_{10}O_3$	16.18	4.79

Table 2 (cont'd). Functional groups of carbonyl thiourea analogs with their anti-amoebic activity

Code	R_1	R_2	IC_{50} (μM)	pIC_{50} (M)
M22	$C_9H_{10}O$	C_7H_7F	14.65	4.83
M23	$C_9H_{10}O$	C_7H_8	19.47	4.71
M24	$C_9H_{10}O$	C9H12	19.80	4.70
M25	C_5H_9ClO	C_7H_7Cl	23.84	4.62
M26	C_5H_9ClO	C_7H_7Cl	26.81	4.57
M27	$C_{14}H_{12}O$	C_7H_7Cl	16.31	4.79
M28	$C_{14}H_{12}O$	$C_{11}H_{16}$	14.14	4.85
M29	$C_9H_{18}O$	C_7H_7Cl	20.93	4.68
M30	$C_9H_{18}O$	C_7H_7Cl	21.52	4.67
M31	$C_9H_{18}O$	$C_7H_6Cl_2$	18.62	4.73
M32	C_8H_7FO	C_7H_7F	23.46	4.63
M33	C_8H_7FO	C_7H_7F	22.28	4.65
M34	C_8H_7FO	C_7H_7F	20.91	4.68
M35	C_8H_7FO	C_7H_7Cl	20.78	4.68
M36	C_8H_7FO	C_7H_7Cl	20.23	4.69
M37	C_8H_7FO	C_7H_7Cl	20.03	4.70
M38	C_8H_7FO	C_9H_{12}	19.74	4.70
M39	C_8H_7FO	C_9H_{12}	19.29	4.72
M40	C_8H_7FO	C_7H_7Br	18.41	4.74
M41	C_8H_7FO	C_7H_7Br	17.98	4.75
M42	C_8H_7FO	$C_{11}H_{16}$	19.02	4.72
M43	C_8H_7FO	C_9H_{12}	20.93	4.68
M44	$C_{12}H_{16}O$	C_5H_{12}	21.55	4.67

Structure entry and molecular modeling

Structure entry and molecular modeling were first carried out by acquiring a three-dimensional (3D) representation of the thiourea derivative molecules. The compounds' two-dimensional (2D) molecular structures were generated using ChemDraw Ultra (version 10.0). The structures were then converted to 3D structures using Chem3D Ultra 10.0. The 3D molecules of thiourea compounds were analyzed with the MM2 method and MOPAC (Chem3D Ultra 10.0) at default settings to acquire stable molecular structures with the lowest energy.

Descriptors generation

A total of 1661 molecular descriptors of optimized thiourea molecules were computed with DRAGON software (version 5.2). In this analysis, only 0D, 1D, and 2D descriptors were used while the 3D descriptors were excluded since molecular structures of the compounds were not uniformly aligned.

Feature selection

The descriptors that did not contain relevant information for the model development were eliminated from the set. The next step was to find the best subset from the remaining descriptors. Stepwise multiple regression and Genetic Algorithm (GA) were used in this step using MATLAB version 7.6.0 (The Mathworks Inc.) software with PLS Toolbox version 5.2.2 (Eigenvector Research Inc.). Meanwhile, Genetic Algorithm GUI [genalg] function was used for GA selection.

Model development

For the model development, routines were performed in MATLAB with PLS Toolbox whereby MLR and PLS analyses were applied. The approaches for modeling utilize integrated stepwise with MLR (stepwise-MLR), GA with MLR (GA-MLR) and GA with PLS (GA-PLS). Statistical significance of the final model was characterized by the squared correlation coefficient, r^2, root mean square error of calibration, *RMSEC*, and root mean square error of cross validation, *RMSECV*. The obtained outputs were converted [regcon] to equation models. The regression coefficient in the equation indicates the significance of an individual descriptor. A plot of predicted versus experimental activity was evaluated to examine the goodness of fit for the generated models. A plot of residuals versus predicted values was used to detect outliers in the data set.

Model validation

The model validation process was also conducted by MATLAB with the PLS toolbox. Both internal and external validations were conducted in this study. Cross-validation by the leave-one-out (LOO) method was applied to provide a rigorous internal check on the built models. This validation process was accomplished by validating the models from their statistical outputs of cross validated squared correlation coefficient r^2_{cv}, squared correlation coefficient of test set, r^2_{test} and root mean square error of prediction, *RMSEP*. Data scrambling technique by the y-randomization test was performed to ensure that the constructed models were not the result of mere chance correlations. On the other hand, external validation was also performed involving the prediction of activity for compounds from the test set.

Results and Discussion

Inhibition concentration for half of cells population (IC_{50} values)

Thiourea derivatives that exhibited the best anti-amoebic activity were 2-(3-benzoylthioureido)-3-mercaptopropanoic acid and 2-(3-benzoylthioureido)-4-(methylthio)butanoic acid labeled as M1 and M2, respectively. These two thiourea compounds showed the lowest IC_{50} value suggesting that they provide the best intrinsic anti-amoebic activity. These two compounds are amino acid types of derivatives that could be recognized by the presence of hydroxyl (-COOH) with an amine group in the molecule. In general, all derivatives of amino acid groups in this study showed lower IC_{50} values compared to other compounds in the series. This indicates that amino acids could enhance the activity of the thiourea derivatives. Xu et al. [12] supported this finding by highlighting that in general, amino acid derivatives of the compounds could exhibit a variety of biological properties. Meanwhile, Hauck et al. [13] emphasized that amino acid derivatives in the compounds contribute to a hydrophilic moiety, which gives high selectivity toward receptors. Ibrahim et al. [8] that came out with thiourea derivatives labeled as M7, M8, M25, and M26 also suggested that the mechanism of action for the proposed thiourea derivatives toward the protozoan parasite *Acanthamoeba* should focus on the hydrophobicity of thiourea molecules to explain their actions. The suggested drug-receptors for the compounds' main target in the amoeba cells are the transport proteins that are distributed throughout the cell membrane.

Chlorhexidine gluconate was used as a positive control as it is a general biocidal effective against a wide variety of microorganisms [14]. The chlorhexidine-treated amoeba in the experiment exhibited a slightly lower IC_{50} of 6.30 ± 0.49 μM. From the T-test analysis, absorbance readings from the untreated and treated cells showed statistical significance (p <0.05). Thiourea in its basic structure has one sulfur atom. A sulfur atom has six valence electrons, and its electronic configuration is similar to oxygen [15]. Most sulfur-containing organics exhibit a low order of toxicity. However, their toxicity may be enhanced by substitution in the molecules. Patnaik [16] explained that an -SH group attached to a benzene ring imparts greater toxicity to the molecule than that attached to an alkyl group. Therefore, the thiourea derivatives were synthesized with at least one benzene ring as an attachment in this study with an attempt to increase their activity on the tested cells.

Table 3. IC$_{50}$ values of 45 carbonyl thiourea derivatives compound on *Acanthamoeba* sp.

Code	**Chemical Name**	**IC$_{50}$ µg/mL**
M1	2-(3-benzoylthioureido)-3-mercaptopropanoic acid	2.56 ± 0.46
M2	2-(3-benzoylthioureido)-4-(methylthio)butanoic acid	2.70 ± 0.27
M3	2-(3-benzoylthioureido)-3-(4-hydroxyphenyl)propanoic acid	3.32 ± 0.18
M4	2-(3-benzoylthioureido)succinic acid	3.96 ± 0.26
M5	1-(benzoylcarbamothioyl)pyrrolodine-3-carboxyl acid	3.08 ± 0.34
M6	2-(3-benzoylthioureido)-3-methylpentanoic acid	3.90 ± 0.20
M7	2-(3-benzoylthioureido)propanoic acid	2.74 ± 0.42
M8	3-(3-benzoylthioureido)propanoic acid	2.92 ± 0.24
M9	3-[3-(3-methylbenzoyl)thioureido]propanoic acid	5.13 ± 0.59
M10	2-[3-(3-methylbenzoyl)thioureido]acetic acid	5.64 ± 0.63
M11	3-hydroxy-2-[3-(3-methylbenzoyl)thioureido]propanoic acid	5.23 ± 0.41
M12	2-[3-(2-methylbenzoyl)thioureido]acetic acid	4.94 ± 0.20
M13	2-(3-furan-2-carbonylthioureido) acetic acid	4.60 ± 0.61
M14	2-[3-(4-methylbenzoyl)thioureido]-3-phenylpropanoic acid	4.95 ± 0.49
M15	5-amino-2-[3-(4-methylbenzoyl)thioureido]-5-oxopentanoic acid	5.51 ± 0.47
M16	2-[3-(4-methylbenzoyl)thioureido]propanoic acid	5.80 ± 0.20
M17	3-hydroxy-2-[3-(4-methylbenzoyl)thioureido]propanoic acid	5.29 ± 0.10
M18	3-methyl-2-[3-(4-methylbenzoyl)thioureido]butanoic acid	5.06 ± 0.40
M19	2-[3-(2-methylbenzoyl)thioureido]propanoic acid	5.31 ± 0.22
M20	3-hydroxy-2-[3-(2-methylbenzoyl)thioureido]butanoic acid	4.62 ± 0.43
M21	3-hydroxy-2-[3-(4-methylbenzoyl)thioureido]butanoic acid	4.80 ± 0.48
M22	*N*-(3-fluorophenyl)-*N’*-phenylacetylthiourea	4.23 ± 0.95
M23	*N*-phenyl-*N’*-phenylacetylthiourea	5.26 ± 0.55
M24	*N*-(2,4-dimethylphenyl)-*N’*-phenylacetylthiourea	5.91 ± 0.51
M25	*N*-(2-chlorophenyl)-*N’*-(4-chlorobutanoyl)thiourea	6.94 ± 0.79
M26	*N*-(3-chlorophenyl)-*N’*-(4-chlorobutanoyl)thiourea	7.81 ± 0.34
M27	*N*-(3-chlorophenyl)-*N’*-(biphenyl-4-yl)carbamoylthiourea	5.98 ± 0.79
M28	*N*-(2,6-diethylpehnyl)-N’-(biphenyl-4-yl)carbamoylthiourea	5.49 ± 0.38
M29	*N*-(2-chlorophenyl)-*N’*-(2-ethylhexanoyl)thiourea	6.55 ± 0.56
M30	*N*-(2-chlorophenyl)-*N’*-(2-ethylhexanoyl)thiourea	6.73 ± 0.30
M31	*N*-(2,5-dichlorophenyl)-*N’*-(2-ethylhexanoyl)thiourea	6.47 ± 0.70
M32	*N*-(3-fluorobenzoyl)-*N’*-(2-fluorophenyl)thiourea	6.86 ± 0.36
M33	*N*-(3-fluorobenzoyl)-*N’*-(3-fluorophenyl)thiourea	6.51 ± 0.27
M34	*N*-(3-fluorobenzoyl)-*N’*-(4-fluorophenyl)thiourea	6.11 ± 0.17
M35	*N*-(3-fluorobenzoyl)-*N’*-(2-chlorophenyl)thiourea	6.42 ± 0.28
M36	*N*-(3-fluorobenzoyl)-*N’*-(3-chlorophenyl)thiourea	6.25 ± 0.35
M37	*N*-(3-fluorobenzoyl)-*N’*-(4-chlorophenyl)thiourea	6.18 ± 0.32
M38	*N*-(3-fluorobenzoyl)-*N’*-(2-ethylphenyl)thiourea	5.97 ± 0.63
M39	*N*-(3-fluorobenzoyl)-*N’*-(3-ethylphenyl)thiourea	5.83 ± 0.49
M40	*N*-(3-fluorobenzoyl)-*N’*-(2-bromophenyl)thiourea	6.50 ± 0.52
M41	*N*-(3-fluorobenzoyl)-*N’*-(3-bromophenyl)thiourea	6.35 ± 0.71
M42	*N*-(3-fluorobenzoyl)-*N’*-(2,6-diethylphenyl)thiourea	6.29 ± 0.21

Table 3 (cont'd). IC_{50} values of 45 carbonyl thiourea derivatives compound on *Acanthamoeba* sp.

Code	Chemical Name	IC_{50} µg/mL
M43	*N*-(3-fluorobenzoyl)-*N'*-(4-ethylphenyl)thiourea	6.33 ± 0.44
M44	*N*-(4-*tert*-butylbenzoyl)-*N'*-diethylphenylthiourea	6.30 ± 0.59
M45	*N*-(3-fluorobenzoyl)-*N'*-(2-chlorophenyl)thiourea	6.42 ± 0.28

QSAR study

QSAR utilizes linear regression of statistical analysis to build mathematical equation models, which could elucidate the relationship for molecular structures of the compounds with their potential biological activities. QSAR will also help to create a preliminary hypothesis regarding the mechanism of action by investigating the compounds on a particular biological system. Through this approach, it is assumed that the compounds that fit in a QSAR model are acting with the same mechanism of action [17].

Development of QSAR model by stepwise-multiple linear regression (stepwise-MLR) analysis

In order to build a QSAR model, feature selection was primarily conducted to discard unnecessary variables from the pool of 202 descriptors. Stepwise multiple regression chose 7 descriptors that fit best with the equation model, which were MATS1m, GATS5m, GATS3e, ESpm01d, ESpm05d, JGI1, and JG12. To avoid the overfitting problem, the proportion of the descriptors to the compound was maintained in a 5:1 ratio of the thumb rule. The descriptors set were maintained to contain five or fewer variables that contain the best information to represent the model. This was done by single exclusion practice of each variable from the selected obtained descriptors. The variables were tested for their significance by removing each one of them and their predictive power was tested each time by using the proposed model. From this technique, two variables, GATS5m and GATS3e, were found to be insignificant for the model and were removed. From the verification by correlation matrix, a high degree of correlation was found between descriptor ESpm01d and ESpm05d with a value of > 0.8. ESpm05d was chosen from the two since it was more significant by giving a statistically better model with higher predictive ability. The regression model was constructed with four descriptors namely MATS1m, ESpm05d, JGI1, and JGI2. According to Hadanu et al. [18], the best model could be selected based on the value of correlation coefficient (r), squared correlation coefficient (r^2), Standard Error of Estimation (SE), degree of freedom (F) or Predictive Residual Sum of Square (PRESS). Frimayanti et al. [19] proposed that the accepted value range for QSAR models is r^2 > 0.6 and r^2_{cv} > 0.5. The final statistical output for the generated model of stepwise-MLR is shown in Table 4.

Table 4. Statistics for a model developed using the stepwise-MLR method

Statistical Output	Value
r^2	0.732
r^2_{cv}	0.522
RMSEC	0.058
RMSECV	0.080

r^2 = squared correlation coefficient;
r^2_{cv} = cross validated squared correlation coefficient;
RMSEC= root mean square error of calibration;
RMSECV= root mean square error of cross validation

Based on the statistical evaluation, high r^2 and r^2_{cv} values with very low *RMSEC* and *RMSECV* indicate that the constructed model is statistically robust to predict the activity of other compounds outside the training set. This statistical outcome shows that the model is capable to elucidate 73.2% of the variance in anti-amoebic activity.

The final QSAR equation model given by the stepwise-MLR method is shown in Equation 1.

The generated regression model shows the order of significance for used descriptors is as follow: ESpm05d > JGI1 > MATS1m > JGI2. The observed and calculated or predicted activity by using this model was compared and it showed that the predicted pIC_{50} values did not greatly differ from the pIC_{50} values obtained from the ex periments with residual values less than 0.95. The observed values are in the range of 4.57 to 5.05 M while the predicted values range from 5.50 to 5.54 M.

$$pIC_{50} = - 0.0145*MATS1m + 0.1013* ESpm05d - 0.0733*JGI1- 0.0090*JGI2 + 4.7592 \quad (1)$$

Validation of QSAR regression model from stepwise multiple linear regression (stepwise-MLR)

The statistical output of model validation is shown in Table 5. It proves that the developed model has a high predictive ability. It was given that r^2_{test} was 0.739, which is above 0.5 with an *RMSEP* value of 0.062, which is low enough for roots mean square error of prediction.

The built of the QSAR models must be properly validated prior to use for interpreting and predicting biological responses of non-investigated compounds. Qin et al. [20] emphasized the importance of vigorous validation of QSAR models despite the high fitting accuracy for the training set and apparent mechanistic appeal. The model equation was later used to predict thiourea analogs' anti-amoebic activity in the test set data. The comparison showed that the observed values are in the range of 4.63 to 5.05 M. Meanwhile, the range of the predicted values was 5.49 to 5.56 M. The residuals obtained gave values less than 0.90. This indicates the robustness and high predictive power of the built model. The robustness of a QSAR model should be also validated by a y-randomization test to ensure that the statistical output in the original model, which gave high r^2 and r^2_{cv} values, were not merely from a chance correlation or structural dependency of the training set. If a true QSAR relationship existed with the real dependent variable, the results for the *y*-random runs should be very low [21]. Y-randomization test was run 20 times and gave r^2 values between 0.052 to 0.233, which shows low r^2 and r^2_{test} (Table 6). Thus, it is concluded that the generated model by stepwise-MLR was not obtained from a random chance of correlation.

Table 5. Statistics of prediction for stepwise-MLR model

Statistical Output	**Value**
r^2_{test}	0.739
RMSEP	0.062

r^2_{test} = correlation coefficient of test;
RMSEP= root mean square error of prediction

Table 6. Y- randomization test of stepwise-MLR model

Iteration	r^2	r^2_{test}	Iteration	r^2	r^2_{test}
1	0.055	0.016	11	0.181	0.052
2	0.180	0.010	12	0.248	0.034
3	0.143	0.083	13	0.152	0.007
4	0.059	0.007	14	0.147	0.042
5	0.089	0.146	15	0.089	0.334
6	0.182	0.011	16	0.060	0.015
7	0.055	0.034	17	0.110	0.233
8	0.052	0.006	18	0.233	0.004
9	0.129	0.115	19	0.082	0.025
10	0.182	0.003	20	0.141	0.001

r^2 = squared correlation coefficient;
r^2_{test} = correlation coefficient of test

Development of QSAR model by genetic algorithm multiple linear regression (GA-MLR) analysis

The regression model was constructed with five descriptors namely RBN, MATS2m, EEig07x, JGI1, and N-070. The correlation matrix was investigated to detect highly correlated variables that would be less important for the model, which detected 135 that represents ESpm05d to be highly correlated with 198, representing N-070 by giving a value higher than 0.80 from the correlation matrix. The analysis was executed again to find out which variable was more significant than others by removing each of them from the model. The final analysis concluded that N-070 was more significant compared to ESpm05d and, therefore, it was included in the model. The statistical output of the final model is shown in Table 7. Statistical result of the constructed GA-MLR model shows a high r^2 value that concludes the equation could explain 84.8% of the variance in the compounds' activity. Meanwhile, r^2_{cv} is also quite high with a value of 0.767. *RMSEC* and *RMSECV* obtained are very low with values of 0.044 and 0.055, respectively. The statistical output shows that the GA-MLR method was a good technique to obtain a good model. This result also proves that the hybrid GA-MLR approach produced a better QSAR model compared to the stepwise-MLR method for this study.

The final QSAR equation model of the GA-MLR method is elaborated in Equation 2 in which it shows a significant order of the five selected variables as follows; N-070 > MATS2m > JGI1 > EEig07x > RBN. The variables of MATS2m, JGI1, and N-070 were shown to work reversibly with pIC_{50} while RBN and EEig07x both worked positively with pIC_{50}. The predicted pIC_{50} values were compared to the experimental data. The data revealed that the predicted values ranging from 4.93 to 5.21 M and the experimental values ranging from 4.57 to 5.05 M. Low residual values in the range of less than 0.55 from the activity prediction proves that the constructed model was robust.

$$pIC_{50} = 0.0301*RBN - 0.0576*MATS2m + 0.0445*EEig07x - 0.0529*JGI1 - 0.0896*N\text{-}070 + 4.7602 \qquad (2)$$

Table 7. Statistics for the model developed using GA-MLR method

Statistical Output	**Value**
r^2	0.848
r^2_{cv}	0.767
RMSEC	0.044
RMSECV	0.055

r^2 = squared correlation coefficient;
r^2_{cv} = cross validated squared correlation coefficient;
RMSEC= root mean square error of calibration;
RMSECV= root mean square error of cross validation

Validation of QSAR regression model from genetic algorithm multiple linear regression (GA-MLR)

The statistical result in Table 8 confirms the robustness of the generated GA-MLR model with acceptable values of r^2_{test} and *RMSEP*. Tropsha et al. [22] suggested that a proposed QSAR model is considered predictive if it satisfies the condition of $r^2_{test} > 0.6$. The predicted pIC_{50} values were compared to the experimental data. The predicted values ranged from 4.92 to 5.20 M and the experimental values ranged from 4.63 to 5.06 M. The residuals gave values less than 0.45 suggesting that the model is robust and has high predictive power.

Validation by y-randomization was run 20 times and the test showed that the squared correlation coefficient, r^2 were in the range of 0.004 to 0.300 (Table 9). The difference between randomized models compared to the original GA-MLR statistical representation was significant. Therefore, this proves that chance correlation was negligible in the model development.

Table 8. Statistics of prediction for GA-MLR model

Statistical Output	**Value**
r^2_{test}	0.777
RMSEP	0.057

r^2_{test} = correlation coefficient of test;
RMSEP= root mean square error of prediction

Table 9. Y- randomization test of GA-MLR model

Iteration	r^2	r^2_{test}	**Iteration**	r^2	r^2_{test}
1	0.151	0.008	11	0.227	0.457
2	0.047	0.240	12	0.176	0.288
3	0.233	0.099	13	0.226	0.064
4	0.130	0.113	14	0.127	0.122
5	0.046	0.066	15	0.212	0.126

Table 9 (cont'd). Y- randomization test of GA-MLR model

Iteration	r^2	r^2_{test}	Iteration	r^2	r^2_{test}
6	0.167	0.015	16	0.218	0.045
7	0.087	0.051	17	0.083	0.081
8	0.176	0.002	18	0.230	0.010
9	0.068	0.143	19	0.215	0.001
10	0.300	0.002	20	0.004	0.195

r^2 = squared correlation coefficient;
r^2_{test} = correlation coefficient of test

Development of QSAR model by genetic algorithm partial least square (GA-PLS) analysis

The third method of QSAR in this study applied GA in combination with the PLS technique. The selected variables used MATS2m, MATS3m, EEig09d, JGI6, and N-070. The correlation matrix of the selected variables showed that no higher correlation than 0.8 for the variables was used hence this explains the consistency to be used in the model. The statistical output for the GA-PLS model is shown in Table 10. The statistical result proves the reliability of the built QSAR model from GA-PLS regression with squared correlation coefficient, r^2 of 0.827 and cross-validated, r^2_{cv} of 0.682. This result is in accordance with a study by Edraki et al. [23] in QSAR analysis of 3,5-bis (arylidene)-4-piperidone derivatives in cytotoxicity models. The statistical output using the GA-PLS model revealed the r^2 of 0.86 and r^2_{cv} of 0.66. The model also gave low root mean square error values for *RMSEC* and *RMSECV*.

Table 10. Statistics for the model developed using the GA-PLS method

Statistical Output	Value
r^2	0.827
r^2_{cv}	0.682
RMSEC	0.047
RMSECV	0.064

r^2 = squared correlation coefficient;
r^2_{cv} = cross validated squared correlation coefficient;
RMSEC= root mean square error of calibration;
RMSECV= root mean square error of cross validation.

This statistical output of the model proves to be able to explain 82.7% of the variance in the experimental activity and r^2_{cv}, which manifests good predictive ability. The best combination of a selected variable by the GA-PLS method is illustrated in Equation 3.

The order of variables are as follows; EEig09d > N-070 > JGI6 > MATS2m > MATS3m in which this QSAR model was used to predict the activity of compounds in the training set. The values of calculated compared to observe activity. The predicted values were in the range of 4.72 to 4.82 M and the experimental values were in the range of 4.63 to 5.06 M. The residuals observed were compared to the predicted values, which were less than 0.25. This explains the robustness of the model. However, the predictive ability of the model was further evaluated in external validation to ensure that the QSAR model also has a good predictive ability for compounds that were not included in the training set.

$$pIC_{50} = -0.0457*MATS2m + 0.0242*MATS3m + 0.0638*EEig09d - 0.0511*JGI6 - 0.0609*N\text{-}070 + 4.7554 \qquad (3)$$

Validation of QSAR regression model from genetic algorithm partial least square (GA-PLS)

The result gave good statistical output with r^2_{test} and *RMSEP* of 0.790 and 0.051, respectively (Table 10). The robustness of the GA-PLS QSAR model was also assessed through its predictive power for the comparison of the observed to predicted activity. The predicted values were in the range of 4.72 to 4.82 M while the experimental values were in the range of 4.57 to 5.05 M. Small residual with values less than 0.25 proved the robustness of the equation model, which could be used to predict other compounds that were not included in the model development. The test of y-randomization was run 20 times and confirmed that the model does not occur merely by random chance correlation of statistics since it gave low squared correlation coefficients, r^2 in the range of 0.015-0.372 (Table 11).

Table 10. Statistics of prediction for GA-PLS method

Statistical Output	**Value**
r^2_{test}	0.790
RMSEP	0.051

r^2_{test} = correlation coefficient of test;
RMSEP= root mean square error of prediction

Table 11. Y- randomization test of GA-PLS model

Iteration	r^2	r^2_{test}	**Iteration**	r^2	r^2_{test}
1	0.078	0.006	11	0.087	0.085
2	0.350	0.158	12	0.055	0.007
3	0.123	0.001	13	0.118	0.008
4	0.084	0.088	14	0.172	0.007
5	0.015	0.004	15	0.210	0.003
6	0.136	0.183	16	0.167	0.090
7	0.026	0.021	17	0.120	0.077
8	0.066	0.021	18	0.148	0.088
9	0.372	0.148	19	0.079	0.001
10	0.074	0.010	20	0.041	0.016

r^2 = squared correlation coefficient;
r^2_{test} = correlation coefficient of test

Comparisons of constructed QSAR models

QSAR mathematical models from three different techniques namely stepwise-MLR, GA-MLR, and GA-PLS were evaluated and compared (Table 12). A genetic algorithm (GA) is a well-suited approach to the problem of variable selection and optimization. GA performs its optimization by comparing root-mean-square error of cross validation, *RMSECV* of proposed models as the measure of fitness [24]. The hybrid approach (GA-MLR) which combines GA with MLR may be useful in the derivation of highly predictive and interpretable QSAR models [17]. Therefore, this approach, as well as the conventional stepwise-MLR technique, was applied in this study. The method of stepwise-MLR gave four selected variables that are well fitted in the equation model. The model gave good statistical output with high r^2 and r^2_{cv} and low *RMSEC* and *RMSECV*. This indicates that the constructed model has good fitness and is robust. On the other hand, the model obtained from the GA-MLR method contains five descriptors and produced a statistically better model compared to the stepwise-MLR. This showed that GA-MLR is a good technique to obtain a better QSAR model in this study. However, the weaknesses of using the MLR technique are that the data are often crude, imprecise, and strongly collinear. These imply that this traditional regression technique, which assumes the selected descriptors to be exact and 100% relevant and independent of each other, will not always work well. Thus, in situations where many strongly collinear descriptors and biological responses operate together, data analytical methods, other than the classical MLR techniques, must be used [25].

Table 12. Summary of constructed models

	Stepwise-MLR	**GA-MLR**	**GA-PLS**
Statistical output:			
r^2	0.732	0.848	0.827
r^2_{cv}	0.522	0.767	0.682
RMSEC	0.058	0.044	0.047
RMSECV	0.080	0.055	0.064
r^2_{test}	0.739	0.777	0.790
RMSEP	0.062	0.057	0.051
Y-random (r^2)	0.055-0.248	0.004-0.300	0.015-0.372
No. of descriptors	4	5	5
Residual in train. set	$\leq$ 0.93	$\leq$ 0.54	$\leq$ 0.23
Residual in test set	$\leq$ 0.87	$\leq$ 0.42	$\leq$ 0.24
Descriptors:			
1	MATS1m	RBN	MATS2m
2	ESpm05d	MATS2m	MATS3m
3	JGI1	EEig07x	EEig09d
4	JGI2	JGI1	JGI6
5	-	N-070	N-070

r^2 = squared correlation coefficient;
r^2_{test} = correlation coefficient of test;
r^2_{cv} = cross validated squared correlation coefficient;
RMSEC= root mean square error of calibration;
RMSECV= root mean square error of cross validation

PLS is well suited to overcome overfitting and multicollinearity problems. It has been used to alleviate the effect of multicollinearity and to prevent overfitting by reducing the dimension size [tran 26]. This approach has also successfully come out with a statistically robust model of five variables that were better compared to the model developed from stepwise MLR and GA-MLR techniques based on the predictive ability. PLS also gave much lower residual values in both training and the test set with < 0.23 and < 0.24, respectively compared to the MLR models that gave higher residual values from their prediction. The descriptors or variables, which demonstrate to be important, are those that have been used more than once from the three approaches. It is suggested that these variables namely JGI1, MATS2m, and N-070 are found to be the influential factors and contributors in the models' development for this study and could possibly have a significant role in modulating the anti-amoebic activity for the thiourea derivative compounds. These variables are in the groups of topological charge indices, 2D-autocorrelations, and atom centered fragments.

GA, MLR, and PLS were also implemented by many authors in their studies for prediction in high-dimensional linear regressions [27,28]. GA is suitable to solve optimization and variable selection problems while MLR yields models that are simpler and easier to interpret compared to PLS because these methods perform regression on latent variables that do not have physical meaning [28]. Due to the co-linearity problem in MLR analysis, one may remove the collinear descriptors before MLR model development. MLR equations can describe the structure-activity relationships but some information will be discarded in the MLR analysis. On the other hand, factor analysis–based methods such as PLS regression can handle the collinear descriptors and, therefore, PLS analysis provides a better analysis with a highly predictive QSAR model [29]. In a previous study, the QSAR model built using PLS and GA-MLR methods showed a remarkable coefficient of determination (r^2) in predicting the anticholinergic side effects of drugs in lower urinary tract infection (PLS: R^2=0.808 and GA-MLR: R^2 =0.804) [26]. A series of 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives were subjected to QSAR methods using factor analysis-based multiple linear regression (FA-MLR), principal component regression (PCR), and partial least squares combined with genetic algorithm for variable selection (GA-PLS). The result revealed that GA-PLS showed the most significant QSAR model with 96% and 91% predicted variances in the pIC_{50} data (compounds tested against *S. aureus*) [28]. The studies proved that GA-PLS is a reliable QSAR model in predicting the biological activities of chemical compounds based on mathematical and statistical relations.

Descriptors interpretation

In this study, all applied approaches offered a few similar descriptors or variables for the equations, which are JGI1, MATS2m, and N-070. The finding suggests that they might be the most influential descriptors that contribute to the QSAR models and provide a significant role in the anti-amoebic activity for the thiourea derivatives. Meanwhile, the selected and best QSAR model, which was obtained from the GA-PLS technique, gave 5 variables that include EEig09d, N-070, JGI6, MATS2m, and MATS3m. These variables belong to 2D autocorrelation (Moran autocorrelation - lag2/weighted, and -lag3/weighted by atomic masses), edge adjacency indices (Eigenvalue 09 from edge adjacency matrix weighted by dipole moments), topological charge indices (mean topological charge index of order 6), and atom-centered fragment class (Ar-NH-Al). According to Speck-Planche et al. [30], atom-centered fragments are simple molecular descriptors that are defined as the number of specific atom types in a molecule. They are calculated from molecular composition and atom connectivities. Each type of atom in the molecule is described in terms of its neighboring atoms. The atom-centered fragment descriptors have been demonstrated to be very useful descriptors and have been employed in previous QSAR studies. Meanwhile, Viswanadha et al. [31] reported that these atom-centered fragments descriptors provide important information about hydrophobic and dispersive interactions, which are involved in biological processes such as transport and distribution of compounds through cells membrane, as well as the information about the compound's receptor interactions. From the variables that have been selected in the model of this study, it could be concluded that the atom-centered fragment plays a major role in providing the information on the thiourea molecules' preliminary penetration into *Acanthamoeba* through the cells membrane.

On the other hand, another representative descriptor or variable selected in the generated model of this study is topological structural descriptors, which are a representation of a molecular structure that arises from the chemical identity of each atom. In this study, two Moran autocorrelation descriptors were used in the GA-PLS model. These descriptors describe hydrophobicity scale, average flexibility index, polarizability parameter, the free energy of amino acid solution in water, residue accessible surface area, amino acid residue volume, steric parameters, and relative mutability [32]. In the meantime, physical interpretation of Burden eigenvalues and topological charge indices is difficult because they condense a large amount of structural and property information into a single number [33]. However, these descriptors have been extensively used in medicinal chemistry [34]. Todeschini and Consonni [35] reviewed that due to the relevancy and complexity of amino acid chains and macromolecules, some descriptors were defined to represent amino acid chains and sequences of amino acids. Amino acid properties were modeled for example by connectivity indices, substituent descriptors, charge descriptors, and principal properties. Raychaudury et al. [36] used a topological descriptor to characterize the size and shape of the side chains in the amino acids, which was based on a graph-theoretical approach applied to root weighted molecular graphs with hydrogen included. Substantially, it was relevant with the result obtained from analysis in this study that concentrated on topological descriptors in the model since the thiourea compound series contained 21 amino acids of the side chain of the molecules. Todeschini et al. [37] explained the interpretability of descriptors whereby it is important to take into account that model response is frequently the result of a series of complex biological or physicochemical mechanisms. Therefore, it is very difficult and reductionist to ascribe to the mechanistic meaning of the selected molecular descriptors in a QSAR model. Furthermore, it must also be highlighted that in multivariate models, even though the interpretation of a singular molecular descriptor can be certainly useful, it is only the combination of a selected set of descriptors that is able to model the studied biological end-point. It was also stressed that in QSAR modeling, attention should be focused on the model quality through its predictive ability.

Conclusion

QSAR models were constructed from the obtained IC_{50} values of 44 carbonyl thiourea analogues toward *Acanthamoeba* sp. (HKL isolate) ranging from 2.56 to 7.81 µg/mL. Linear regression techniques, stepwise-MLR, GA-MLR, and GA-PLS were applied to build the best QSAR model to correlate the compounds with their anti-amoebic activity. Three equation models were successfully generated with the hybrid GA-PLS approach that gave the best output with a statistically robust model of five variables. Thus, this GA-PLS QSAR model could be applied in the future for the development of a new lead compound based on carbonyl thiourea with optimized anti-amoebic activity to design a new anti-amoebic agent for A*canthamoeba* keratitis disease.

Acknowledgement

The authors would like to acknowledge and thank International Islamic University Malaysia (IIUM), Universiti Malaysia Terengganu (UMT), and Universiti Teknologi Malaysia (UTM) for providing the space and facilities to conduct this work. This paper is especially dedicated to the research team, Professor Dr. Nakisah Mat Amin and Professor Mohamed Noor Hasan, who has retired from UMT and UTM.

References

1. Kang, H., Sohn, H. J Park, A. Y. et al. (2021). Establishment of an *Acanthamoeba* keratitis mouse model confirmed by amoebic DNA amplification. *Scientific Report,* 11: 4183.

2. Vontobel, S. F, Abad-Villar, E. M., Kaufmann, C., Zinkernagel, A. S., Hauser, P. C. et al. (2015). Corneal Penetration of Polyhexamethylene Biguanide and Chlorhexidine Digluconate. *Journal of Clinical and Experimental Ophthalmology,* 6: 430.
3. Turner, N. A., Russell, A. D., Furr, J. R. and Lloyd, D. (2000). Emergence of resistance to biocides during differentiation of Acanthamoeba castellanii. *Journal of Antimicrobial Chemotherapy*, 46: 27-34.
4. Lorenzo-Morales, J., Khan, N. A. and Walochnik, J. (2015). An update on Acanthamoeba keratitis: diagnosis, pathogenesis and treatment. *Parasite*, 22: 10.
5. Chen, S., Wu, G. and Zeng, H. (2005). Preparation of high antimicrobial activity chitosan-Ag$^+$ complex. *Carbohydrate Polymers,* 60: 33-38.
6. Alcolea, V., Plano, D., Karelia, D. N., Palop, J. A., Amin, S., Sanmartín, C. *et al.* (2016). Novel seleno- and thio-urea derivatives with potent in vitro activities against several cancer cell lines. *European Journal of Medical Chemistry*, 113: 134-144.
7. Sadeghian-Rizi, S., Sakhteman, A. and Hassanzadeh, F. (2016). A quantitative structure-activity relationship (QSAR) study of some diaryl urea derivatives of B-RAF inhibitors. *Research in Pharmaceutical Sciences*, 11(6): 445-453.
8. Naz, S., Zahoor, M., Umar, M., Alghamdi, S., Sahibzada, M. and UlBari, W. (2020). Synthesis, characterization, and pharmacological evaluation of thiourea derivatives. *Open Chemistry*, 18(1): 764-777.
9. Keche, A. P. and Kamble, V. M. (2019). Synthesis and anti-inflammatory and antimicrobial activities of some novel 2-methylquinazolin-4(3H)-one derivatives bearing urea, thiourea and sulphonamide functionalities, *Arabian Journal of Chemistry*, 12(7): 1522-1531.
10. Asadollahi-Baboli, M. and Dehnavi, S. (2018). Docking and QSAR analysis of tetracyclic oxindole derivatives as α-glucosidase inhibitors. *Computational Biology and Chemistry*, 76: 283-292.
11. Ibrahim, M. A., Yusof, M. S. and Amin, N. M. (2014). Anti-amoebic properties of carbonyl thiourea derivatives. *Molecules*, 19(4): 5191-5204.
12. Wiji Prasetyaningrum, P., Bahtiar, A. and Hayun, H. (2018). Synthesis and cytotoxicity evaluation of novel asymmetrical mono-carbonyl analogs of Curcumin (AMACs) against Vero, HeLa, and MCF7 Cell Lines. *Scientia Pharmaceutica*, 86(2): 25.
13. Xu, Q., Deng, H., Li, X. and Quan, Z. S. (2021). Application of amino acids in the structural modification of natural products: A review. *Frontiers in Chemistry,* 9: 650569.
14. Hauck, M., Jürgens, S. R. and Leuschner, C. (2010). Effect of amino acid moieties on metal binding in pulvinic acid derivatives and ecological implications for lichens producing these compounds. *The Bryologist*, 113(1): 1-7.
15. Ahmadi, S. and Habibpour, E. (2017). Application of GA-MLR for QSAR modeling of the arylthioindole class of tubulin polymerization inhibitors as anticancer agents. *Anticancer Agents in Medical Chemistry*, 17(4): 552-565.
16. Rahman, F. U., Bibi, M., Khan, E., Shah, A. B., Muhammad, M., Tahir, M. N., Shahzad, A., Ullah, F., Zahoor, M., Alamery, S. et al. (2021). Thiourea derivatives, simple in structure but efficient enzyme inhibitors and mercury sensors. *Molecules,* 26: 4506.
17. Patnaik, P. (2007). A comprehensive guide to the hazardous properties of chemical substances: Thiourea. Wiley-Interscience, Hoboken, New Jersey: pp. 904.
18. Saxena, A. K. and Prathipati, P. (2003). Comparison of MLR, PLS and GA-MLR in QSAR analysis. *SAR QSAR Environmental Research*, 14(5-6): 433-445.
19. Hadanu, R., Mastjeh, S., Mustofa, J., Sholikhah, E. N., Wijayanti, M. A. and Tahir, I. (2007). Quantitative structure-activity relationship analysis (QSAR) of antimalarial 1,10-phenanthroline derivatives compounds. *Indonesian Journal of Chemistry*, 7(1): 72-77.

20. Frimayanti, N., Yam, M. L., Lee, H. B., Othman, R., Zain, S. M. and Rahman, N. A. (2011). Validation of quantitative structure-activity relationship (QSAR) model for photosensitizer activity prediction. *International Journal of Molecular Sciences*, 12(12), 8626–8644.
21. Qin, L., Zhang, X., Chen, Y., Mo, L., Zeng, H. and Liang, Y. (2017). Predictive QSAR Models for the toxicity of disinfection byproducts. *Molecules***,** 22(10): 1671.
22. Ravichandran, V., Mourya, V. K. and Agrawal, R. K. (2009). Prediction of anti-HIV activity of phenyl ethyl thiourea (PET) derivatives: QSAR approach. *Digest Journal of Nanomaterials and Biostructures,* 4: 213-221.
23. Tropsha, A., Gramatica, P. and Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR and Combinatorial Science*, 22: 69-77.
24. Edraki, N., Das, U., Hemateenejad, B., Dimmock, J. R. and Miri, R. (2016). Comparative QSAR analysis of 3,5-bis (arylidene)-4-piperidone derivatives: The development of predictive cytotoxicity models. *Iranian Journal of Pharmaceutical Research,*15(2): 425-437.
25. Katoch, S., Chauhan, S. S. and Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools Application,* 80**:** 8091–8126.
26. Filzmoser, P., Gschwandtner, M. and Todorov, V. (2012). Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*, 26: 42-51.
27. Tran, T. N., Afanador, N. L., Buydens, L. M. C. and Blanchet, L. (2014). Interpretation of variable importance in partial least squares with significance multi-variate correlation (sMC). *Chemometrics and Intelligent Laboratory System,* 138: 153-160.
28. Yuyama, M., Ito, T., Arai, Y., Kadowaki, Y., Iiyama, N. et al., (2020). Risk prediction method for anticholinergic action using auto-quantitative structure–activity relationship and docking study with molecular operating environment. *Chemical and Pharmaceutical Bulletin*, 68(8): 773-778.
29. Sabet, R. and Fassihi, A. (2008). QSAR study of antimicrobial 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives using different chemometric tools. *International Journal of Molecular Sciences*, *9*(12): 2407-2423.
30. Deeb, O., Hemmateenejad, B, Jaber, A., Garduno-Juarez, R., Miri, R. (2007). Effects of the electronic and physicochemical parameters on the carcinogenecis activity of some sulfa drug using QSAR analysis based on genetic-MLR & genetic-PLS. Chemosphere, 67: 2122-2130.
31. Speck-Planche, A., Kleandrova, V.V., Luan, M.F., and Cordeiro, N.D.S. (2011). Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *European Journal of Medicinal Chemistry*, 46: 5910-5916.
32. Viswanadhan, V. N., Reddy, M. R., Bacquet, R. J. and Erion, M. D. (1993). Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases. *Journal of Computational Chemistry,* 14: 1019-1026.
33. Horne, D. S. (1988). Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers,* 27: 451-477.
34. Helguera, A.M., Natália, M., Cordeiro, D.S., González, M.P., Pérez, M.A.C., Ruiz, R.M., and Castillo, Y.P. (2007). QSAR modeling for predicting carcinogenic potency of nitroso-compounds using 0D-2D molecular descriptors. *11th International Electronic Conference on Synthetic Organic Chemistry*, 1-30 November 2007.
35. Jain, H.K., and Agrawal, R.K. (2006). QSAR analysis of indomethacin derivatives as selective COX–2 inhibitors. *Internet Electronic Journal of Molecular Design*, 5: 224-236.
36. Todeschini, R., and Consonni, V. 2000. Handbook of Molecular Descriptors: Methods and Principles in Medicinal Chemistry. Volume 11, pp. 11. Weinheim: Wiley-VCH.

37. Raychaudhury, C., Banerjee, A., Bag, P., and Roy, S. (1999). Topological shape and size of peptides: Identification of potential allele specific helper T cell antigenic sites. *Journal of Chemical Information and Computer Sciences,* 39(2): 248-254.
38. Todeschini, R., Consonni, V., and Gramatica, P. (2009). Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R. eds. Comprehensive Chemometrics: Chemical and biochemical analysis, pp. 129-172. Oxford: Elsevier.