Nova Biotechnologica *et* Chimica

A machine learning approach in predicting mosquito repellency of plant – derived compounds

Jose Isagani B. Janairo✉, Gerardo C. Janairo and Frumencio F. Co

De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines

Article info	Abstract
	The increasing prevalence of mosquito – borne diseases has prompted intensified efforts in the prevention of being bitten by the vector. Among the various strategies of vector control, the application of repellents provides instant and effective protection from mosquitoes. However, emerging concerns regarding the safety of the widely used repellent, DEET, has led to initiatives to explore natural alternatives. In order to fully realize the potential of natural repellents, focusing on the discovery of natural compounds eliciting repellency is of paramount importance. In this paper, machine learning was utilized to establish association between the mosquito repellent activity of 33 natural compounds using 20 chemical descriptors. Individually, the descriptors had insignificant monotonic relationship with the response variable. But when optimized, the formulated model through boosted trees regression exhibited reliable predictive ability ($r^2_{train} = 0.93$, $r^2_{test} = 0.66$, $r^2_{overall} = 0.87$). The findings presented have also introduced new descriptors that exhibited association with repellency through ensemble learning such as heat capacity, *Log P*, entropy, enthalpy, Gibb's free energy, energy, and zero-point energy.

Introduction

Vector – borne diseases are serious health burdens, which account for one – sixth of the illnesses suffered by the global population (WHO 2014). In particular, mosquitoes are carriers of dreaded diseases such as malaria, dengue, zika, and chikungunya. Aside from being a health concern, mosquito – borne diseases are also associated in aggravating poverty (Suaya *et al.* 2009). Considering the negative multi-faceted impact of these diseases, the prevention of being bitten by the vector is of paramount importance. Among the various strategies available to prevent being bitten by mosquitoes, the application of repellents is considered safe, and provides instant and effective protection. Repellents also serve as the first line of defense in cases where the mosquito-borne disease, such as zika, has no established therapies and personal protection is the best approach to alleviate the disease burden (Gulland 2016; Wong *et al.* 2016). Repellents exert their effect by preventing the binding of attractant odors to Odorant – Binding Proteins (OBPs), that leads to the disruption of signal transduction pathways related to odor recognition (Pellegrino *et al.* 2011). OBPs play a critical role in odor perception given that they act as carriers to the bound odor across the mucus barrier to initiate the physiological olfactory response (Pelosi 1994). In addition, repellents have also been shown to block the electrophysiological responses of sensory neurons of mosquitoes toward attractive odors (Ditzen *et al.* 2008).

✉ *Corresponding author:* jose.isagani.janairo@dlsu.edu.ph

The synthetic repellent, N-N-diethyl-meta-toluamide (DEET) is a widely used insect repellent, and is considered to be the most - broad spectrum (Katz *et al.* 2008). However, issues concerning the toxicity of DEET to humans (Diaz 2016) as well as being a pollutant to the environment (Dsikowitzky *et al.* 2014) have led to initiatives to explore natural repellents as alternatives. However, most natural repellents come in the form of crude botanical extracts, such as lemongrass (Oyedele *et al.* 2002), and thyme (Park *et al.* 2005). It is therefore evident that identification of the compounds eliciting repellency can significantly enhance the formulation and efficacy of repellents. A good example is citronellal, which is a major component of several botanical repellents (Trongtokit *et al.* 2005). When tested individually, citronellal exhibited promising repellency dose (RD$_{50}$) against the malaria vector *Anopheles gambiae*, which is comparable with DEET (Omolo *et al.* 2004). Thus, natural compound screening is crucial in the discovery of more nature-derived mosquito repellents.

Economic and practical considerations are major impediments in compound screening, since testing each compound within a botanical extract for repellency is tedious, and entails serious financial support. A viable strategy in increasing the efficiency of drug discovery activities involves computation – driven approaches. Doing so can help identify promising leads, group together compounds with similar mechanisms of action, or provide other relevant information (Miszta *et al.* 2013). However, data – driven studies about mosquito repellents have mostly been confined to DEET and DEET-like compounds (Katritzky *et al.* 2006; Natarajan *et al.* 2008; Katritzky *et al.* 2008). Thus, employing data science and statistical strategies to examine natural repellents will accelerate and promote discovery and development for these natural compounds. In this study, a predictive model for mosquito repellency against *Anopheles gambiae* of natural compounds was established though ensemble learning. The results presented may help accelerate the identification of repellent leads from natural sources, and provide novel insights regarding their biological activities.

Experimental

The repellent activities (RD$_{50}$) against *A. gambiae* of the natural compounds examined in this study were taken from Omolo *et al.* (2004). RD$_{50}$ refers to the minimum concentration of the compounds

Table 1. Summary of transformations carried out on selected variables in order to minimize variations due to scale.

Variable	Transformation	Range of Values
RD$_{50}$ (response variable)	$Log\ (1/RD_{50})$	2.23 to 5.00
Energy [kJ/mol]	$Log\ (\text{-Energy})$	6.01 to 6.24
HOMO [eV]	none	-6.90 to -5.23
LUMO [eV]	none	-1.86 to 1.60
Chemical Potential [eV]	none	-4.22 to -2.19
Hardness [eV]	none	2.49 to 4.30
Electrophilicity [au]	none	0.59 to 3.40
Dipole [debye]	none	0.05 to 4.34
Solvation [kJ/mol]	none	-17.13 to 11.12
Weight [g/mol]	$Log\ (\text{Weight})$	2.13 to 2.34
Area [A^2]	$Log\ (\text{Area})$	2.26 to 2.44
Volume [A^3]	$Log\ (\text{Volume})$	1.25 to 2.42
Polar Surface Area, PSA [A^2]	$Log\ (\text{PSA+1})$	0 to 1.40
Ovality	none	1.24 to 1.47
Log P	none	1.71 to 4.28
Polarizability [kJ/mol]	$Log\ (\text{Polarizability})$	1.73 to 1.79
Zero Point Energy, *ZPE*	$Log\ (ZPE)$	2.71 to 2.98
Enthalpy, *H*	$Log\ (\text{-}H)$	2.59 to 2.82
Heat Capacity, *Cv*	$Log\ (Cv)$	2.21 to 2.44
Entropy, *S*	$Log\ (S)$	2.57 to 2.69
Gibb's Free Energy, *G*	$Log\ (\text{-}G)$	2.59 to 2.82

needed to elicit repellency to half of the mosquito population. Thus lower RD_{50} values indicate better repellency. The most stable conformers of the 33 compounds of the library were determined using the Merck Molecular Force Field (MMFF) in Spartan '10 (Wavefunction, Inc.). After the most stable conformers were identified, the optimum geometries, and molecular descriptors were obtained through Density Functional Theory B3LYP / 6-311++G**. The 33 compounds are: camphene, beta-pinene, p-cymene, alpha-terpinene, gamma-terpinene, alpha-terpinolene, alphe-pinene, perillyl alcohol, cis-verbenol, cis-carveol, geraniol, alpha-terpineol, eugenol, terpinen-4-ol, linalool, citronellal, perillaldehyde, camphor, verbenone, fenchone, carvone, caryophyllene oxide, limonene oxide, 1,8-cineole, alpha-fenchyl alcohol, borneol, myrtenol, geranyl acetate, thujone, myrtenal, aromadendrene, 4-isopropylbenzaldehyde.

The molecular descriptors energy, HOMO energy, LUMO energy, dipole moment, solvation energy, molecular weight, area, volume, polar surface area, ovality, *Log P*, polarizability, zero-point energy (*ZPE*), constant volume heat capacity at 298 K (*Cv*), enthalpy, entropy, and Gibb's free energy were directly obtained from the DFT calculations. The quantum chemical descriptors of chemical potential, electronegativity, chemical hardness, chemical softness, and electrophilicity were calculated using various reactivity equations founded on DFT, as reviewed by Kaya and Kaya (2015).

The calculated descriptors were transformed in order to minimize their variations in terms of scale and magnitude. The transformations carried out are summarized in Table 1 (raw and transformed datasets are available in the Supporting Information).

The transformed descriptors were then subjected to correlation analysis using R version 3.5.0 (R Core Team 2018). All transformed descriptors, including the response variable were tested for normality of distribution using the Shapiro-Wilk test. Following this test, isotonic relationships that may exist between the transformed descriptors and the response variable, were evaluated using the appropriate correlation function. After correlation analysis, the transformed descriptors were used to formulate a predictive model for mosquito repellent activity using machine learning. For support vector machine (SVM) regression, the radial basis function (RBF) kernel was utilized. As a result, the best gamma and C parameters were automatically selected. 75 % of the data set was

Table 2. Correlation analysis summary between the transformed response variable and the transformed descriptors.

Descriptor correlated with RD_{50}	Shapiro-Wilk (p-value)	Spearman's rho	p-value	Kendall's tau	p-value
Energy	<0.01	0.189	0.293	0.154	0.209
HOMO	0.014	-0.368	0.035	-0.240	0.052
LUMO	0.033	-0.202	0.259	-0.130	0.292
Chemical Potential	<0.01	-0.327	0.063	-0.226	0.065
Hardness	0.071	–	–	–	–
Electrophilicity	<0.01	0.236	0.187	0.163	0.183
Dipole	0.253	–	–	–	–
Solvation	0.079	–	–	–	–
Weight	<0.01	0.217	0.236	0.187	0.153
Area	<0.01	0.246	0.167	0.171	0.163
Volume	<0.01	0.228	0.203	0.1444877	0.239
PSA	<0.01	–	–	–	–
Ovality	<0.01	–	–	–	–
Log P	0.069	–	–	–	–
Polarizability	<0.01	0.231	0.196	0.156	0.204
ZPE	<0.01	0.038	0.833	0.030	0.804
H	<0.01	0.190	0.290	0.156	0.204
Cv	<0.01	0.254	0.154	0.190	0.121
S	<0.01	0.276	0.120	0.205	0.094
G	<0.01	0.189	0.292	0.152	0.215

used for training, while the remaining 25 % served as the test set. The results were validated by applying a 10-fold cross validation. For random forest regression, the random test data proportion was set to 0.30, and 0.5 for the subsample proportion. The stopping parameters were set as follows: minimum n cases = 5, maximum n cases = 10, minimum n child in node = 5, maximum n of nodes = 100. For the boosted trees regression, the learning rate was set to 0.1, with the following conditions: number of additive terms = 200, random test data proportion = 0.30, subsample proportion = 0.50. The stopping parameters were set as follows: minimum n of cases = 5, maximum n of levels = 10, minimum n in child node = 1, maximum n of nodes = 3. For both the random forest, and boosted trees regression, 60 % of the data was dedicated to train the algorithm while the remaining 40 % was used for model testing. The predictive ability of the models was assessed based on the goodness of fit between the observed and predicted values of the RD_{50} for the compounds.

Results and Discussion

Predictive modelling is an important component of drug design and discovery. Predictive models establish association with the response variable, usually the biological activity, with a set of molecular properties or descriptors using statistical tools. Common statistical methods used in establishing association is multiple linear regression, but the utilization of machine learning has recently gained traction due to its versatility and effectivity in establishing predictive models (Gertrudes *et al.* 2012). Correlation analysis was used to aid in the selection of transformed descriptor to be included in the formulation of the predictive models. A prerequisite to correlation analysis is the test for normality distribution, which was done through the Shapiro-Wilk test. In this test, a *p*-value greater than significance level of 0.05 indicates a normal distribution of the data. The transformed response variable had a *p*-value of $1.16e^{-5}$, indicating a non-normal distribution. This means that the usual Pearson correlation is not applicable for analysing isotonic relationships with the

response variable. The Shapiro-Wilk test was further conducted for the transformed variables, and the results are shown in Table 2. The descriptors hardness, dipole, solvation had *p*-values greater than 0.05, indicating a normal distribution, while the other descriptors had non-normal distribution. Thus, hardness, dipole, and solvation were excluded in the succeeding correlation analysis using Spearman, and Kendall. These aforementioned correlation tests are used for data with a non-normal distribution, wherein the data are converted into ranks prior to correlation. As a consequence, PSA, ovality, and *Log P* were also excluded from the correlation analysis since ties in the rankings of the compounds were observed. Hence, the remaining descriptors were subjected to both Spearman, and Kendall correlation tests. The results of the tests showed that the transformed descriptors had insignificant isotonic relationship with the response variable, as demonstrated by *p*-values greater than 0.05. As a consequence of these results, other approaches in selecting relevant descriptors to predict the mosquito repellency should be explored, such as backward elimination (Dudek *et al.* 2006). Thus, all 20 transformed descriptors were initially used to formulate the predictive models.

The next part of the study involves identifying the suitable algorithm for the given data set, since the type of machine learning algorithm affects the performance of the predictive model. A model to predict the repellent activities of the 33 natural compounds using the 20 descriptors was preliminarily established through boosted trees, support vector machine, and random forest. The three aforementioned machine learning techniques are different supervised predictive algorithms, each possessing different characteristics. Boosted regression trees is a sequential model-building classifier wherein weak predictors in the previous model are given more weight or boosted, to improve prediction performance in the next model. Random forest is a bagging algorithm wherein bootstrapped samples with replacement are taken from the data set and predictions are obtained from each of these samples. Final predictions are then obtained from the average predictions of these samples. This way, the model variance may be improved but not

the predictive power. Boosting such as boosted regression trees reduces bias and therefore improves prediction while bagging such as random

forest reduces variance. Random forests, unlike boosted trees, are less sensitive to outliers. Boosted trees has another advantage in terms of speed and

Table 3. Comparison of the performance of various machine learning algorithms in predicting the repellent activity of the 33 natural compounds using all calculated molecular descriptors.

	Boosted Trees			Support Vector Machine			Random Forest		
	Train	**Test**	**Overall**	**Train**	**Test**	**Overall**	**Train**	**Test**	**Overall**
Mean Square Error	0.134	0.133	0.133	0.487	0.271	0.428	0.455	0.244	0.372
Mean Absolute Error	0.274	0.262	0.269	0.647	0.472	0.599	0.562	0.452	0.519
Mean Relative Squared Error	0.014	0.014	0.014	0.044	0.024	0.038	0.045	0.024	0.037
Mean Relative Absolute Error	0.090	0.088	0.089	0.193	0.141	0.179	0.177	0.142	0.164
Correlation Coefficient	0.862	0.616	0.807	0.431	0.393	0.413	0.508	0.316	0.351
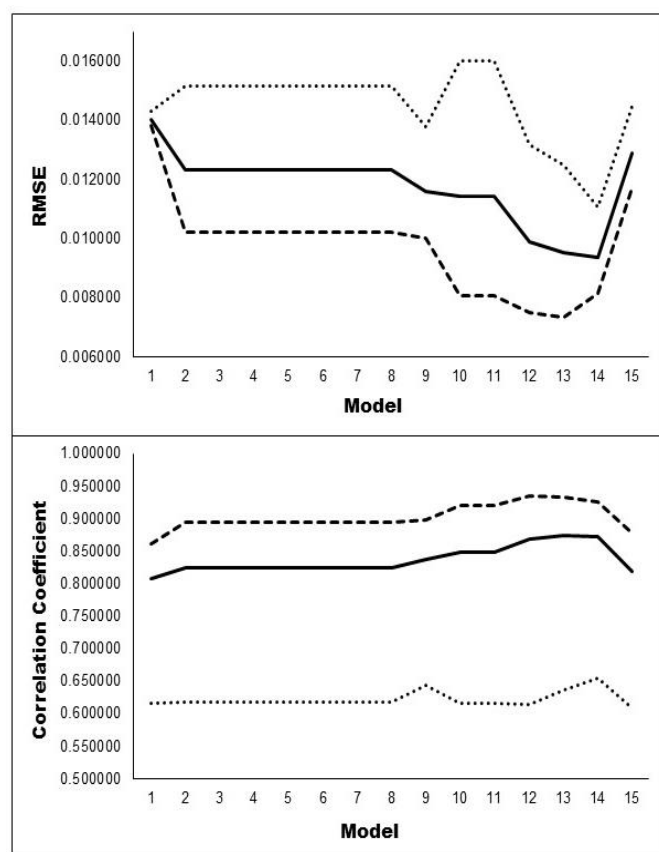
Fig. 1. Performance of the predictive models based on relative mean squared error (RMSE, top graph) and correlation coefficient (bottom graph). The broken line represents values for the training set, the dotted line for the test set, and the solid line for the overall set.

simplicity over random forests since the former performs better with smaller trees than the latter. Boosted trees is also insensitive under monotone transformations of the predictors such as logarithmic transformation. In SVM, optimal

hyperplanes are constructed in a high-dimensional space for classification or regression. For a good separation, a hyperplane with the largest distance from the nearest training data point is chosen in order to achieve a lower error rate. SVM, like boosted trees, is also susceptible to overfitting that could be avoided by properly tuning its hyperparameters.

The formulated predictive models utilized transformed descriptors and response variables (Table 1) in order to have a narrower range of values among the variables. Various goodness of fit measures between the observed and predicted repellent activity values were used for model diagnostics, as shown in Table 3. It is thus evident that boosted trees regression (using gradient boosting algorithm) is the ideal algorithm to be used for this particular dataset since it had the lowest deviations, and strongest correlations between the observed and predicted values. The adeptness of boosted trees for this particular dataset can be possibly rooted in its robustness to outliers, irrelevant variables, correlated variables, and log-transformed predictors (Hastie *et al.* 2008).

Model optimization was thus confined to boosted tree regression, wherein a stepwise back elimination of the predictors was conducted based on the calculated predictor importance. This is necessary since the results of the correlation analysis (Table 2) failed to show strong association between the descriptors and the response variable. For example, the least important descriptor for the full-descriptor model 1 was hardness. Hence

in model 2, only 19 descriptors were used for the modelling since hardness was removed. This process was repeated until the remaining descriptors had importance scores of greater than or equal to 0.90 (The predictors and their corresponding importance for each model is available in the supporting information). The resulting performance of the formulated predictive models following this optimization method is shown in Fig. 1.
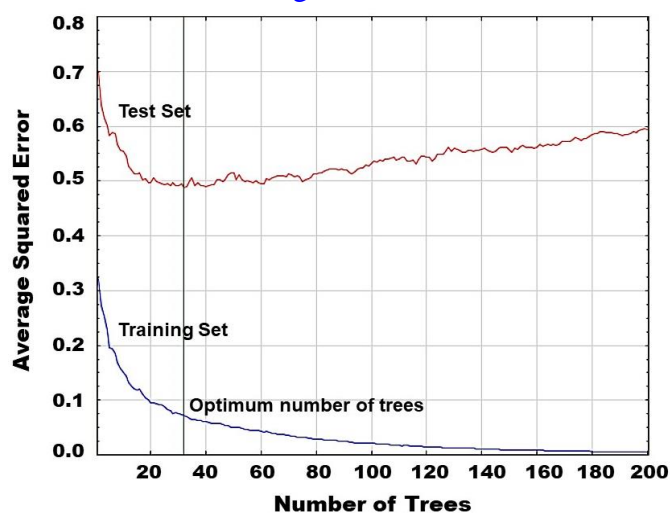

Fig. 2. Summary of boosted trees showing the optimum number of trees prior to data overfitting.

The best model was judged to be model 14, since the remaining descriptors all had importance scores of greater than 0.90 (Table 4). Model 14 also exhibited the lowest RMSE values, and highest correlation coefficients in all sets (training, test, and overall sets). Model 14 was found optimal in terms of average squared error using 32 trees for the test set (Fig 2). The formulated regression models were validated using 40 % of the data. While no general rule exists regarding the optimum division of the data between training and testing (Roy *et al.* 2008), a 60–40 split seems appropriate in this case to balance training and validation of the relatively small population of the data set.

The most important chemical descriptor is *Log P*, which relates the hydrophobicity of the molecule. This result resonates with the property of OBPs that possess a hydrophobic cavity to which odors are bound (Murphy *et al.* 2013). Majority of the descriptors are thermodynamic properties of the molecules, which are known to heavily influence ligand binding to proteins (Bostrom *et al.* 1988).

Table 4. Descriptors of model 14 used to predict the repellent activity of 33 natural compounds. All descriptors had importance scores of greater than 0.90.

Transformed Descriptor	Predictor Importance Score
Log P	1.00000
Log (Cv)	0.991466
Log (PSA +1)	0.973521
Log (S)	0.956767
Log (-Energy)	0.938961
Log (-H)	0.938961
Log (-G)	0.938961
Log (ZPE)	0.927269

Aside from establishing a satisfactory predictive model for mosquito repellency, the present findings have introduced new descriptors that showed association with repellency within the context of ensemble learning. Previous quantitative structure – activity relationship models reported the descriptors of boiling point, molecular surface area, total charge of the other substituents on the cyclic backbone, and dipole moment were correlated with the repellent activity of a set of terpenoid compounds (Wang *et al.* 2008). Another model showed that the repellent activity of a class of terpenoid compounds showed correlation with the LUMO energy, minimum valence of the O atom, principal moment of inertia, and the shadow area of the repellent (Song *et al.* 2013). Molecular topological indices have also been demonstrated to exhibit association with mosquito repellency (Garcia-Domenech *et al.* 2010). It is interesting to note that the present study has provided new descriptors in which repellency can be predicted, which are heat capacity, *Log P*, entropy, enthalpy, Gibb's free energy, energy, and *ZPE*. Past studies that attempted to predict mosquito repellency mostly applied multiple linear regression on a compound library with limited diversity. The presented results thus demonstrate the utility of machine learning in bioactivity predictive modelling. Individually, the descriptors had insignificant correlation with repellency. But when these descriptors were processed through boosted trees, a satisfactory predictive model was constructed. It should be noted that further validation of the model is still needed. Although the algorithm already exhibits reliable predictive power, increasing the general applicability and robustness of the model requires incorporating

more compounds to the data set. However, this can be challenging at this point considering that most of the reports on *A. gambiae* repellency focus on crude botanical extracts (Odalo *et al.* 2005; Deletre *et al.* 2013), or use a different metrics to measure repellency, such as oviposition activity index (Kweka *et al.* 2010), and % repellency (Logan *et al.* 2010).

Conclusions

A model to predict mosquito repellency of natural compounds for the *Anopheles gambiae* using boosted trees regression was established. DFT calculated descriptors were used to predict repellency, wherein *Log P*, heat capacity, polar surface area, entropy, energy enthalpy, Gibb's free energy, and zero-point energy are new descriptors found to be associated with repellency. Overall the findings presented are expected to promote and accelerate the discovery of natural mosquito repellents, as well as to forge a deeper understanding on the molecular properties responsible for eliciting mosquito repellency.

Acknowledgements

References

Bostrom J, Norrby P-O, Lilijefors T (1998) Conformational energy penalties of protein-bound ligands. J. Comput. Aided Mol. Des. 12: 383-396.

Deletre E, Martin T, Campagne P, Bourget D, Cadin A, Menut C, Bonafos R, Chandre F (2013) Repellent, irritant and toxic effects of 20 plant extracts on adults of malaria vector *Anopheles gambiae* mosquito. Plos One. 8: e82103.

Diaz JH (2016) Chemical and plant-based insect repellents: efficacy, safety, and toxicity. Wilderness Environ. Med. 27: 153-163.

Ditzen M, Pellegrino M, Vosshall LB (2008) Insect odorant receptors are molecular targets of the insect repellent DEET. Science. 319: 1838-1842.

Dsikowitzky L, Dwiyitno, Heruwati E, Ariyani F, Irianto HE, Schwarzbauer J (2014) Exceptionally high concentrations of the insect repellent N,N-diethyl-m-toluamide (DEET) in surface waters from Jakarta, Indonesia. Environ. Chem. Lett. 12: 407-411.

Dudek AZ, Arodz T, Galvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb. Chem. High Throughput Screen. 9: 213-228.

Garcia-Domenech R, Aguilera J, El Moncef A, Pocovi S, Galvez J (2010) Application of molecular topology to the prediction of mosquito repellents of a group of terpenoid compounds. Mol. Divers. 14: 321-329.

Gertrudes JC, Maltarollo VG, Silva RA, Oliveira PR, Honorio KM, da Silva AB (2012) Machine learning techniques and drug design. Curr. Med. Chem. 19: 4289-4297.

Gulland A (2016) Zika virus is a global public health emergency, declares WHO. BMJ. 352: i657.

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning: data mining, inference, and prediction. 2nd edition, Springer, New York.

Katritzky AR, Wang Z, Slavov S, Tsikolia M, Dobchev D, Akhmedov NG, Hall CD, Bernier UR, Clark GG, Linthicum KJ (2008) Synthesis and bioassay of improved mosquito repellents predicted from chemical structure. Proc. Natl. Acad. Sci. USA. 105: 7359-7364.

Katritzky AR, Dobchev DA, Tulp I, Karelson M, Carlson DA (2006) QSAR study of mosquito repellents using Codessa Pro. Bioorg. Med. Chem Lett. 16: 2306-2311.

Katz TM, Miller JH, Herbert AA (2008) Insect repellents: historical perspectives and new developments. J. Am. Acad. Dermatol. 58: 865-871.

Kaya S, Kaya C (2015) A new method for calculation of molecular hardness: a theoretical study. Comput. Theor. Chem. 1060: 66-70.

Kweka EJ, Lyatuu EE, Mboya MA, Mwang'onde BJ, Mahandre AM (2010) Oviposition deterrence induced by *Ocimum kilimandscharicum* and *Ocimum suave* extracts to gravid *Anopheles gambiae* s.s. (Diptera: Culicidae) in laboratory. J Global Infect Dis. 2: 242-245.

Logan JG, Stanczyk NM, Hassanali A, Kemei J, Santana AEG, Ribeiro KAL, Pickett JA, Luntz AJM (2010) Arm-in-cage testing of natural human-derived mosquito repellents. Malar. J. 9: 239.

Miszta P, Basak SC, Natarajan R, Nowak W (2013) How computational studies of mosquito repellents contribute to the control of vector borne diseases. Curr. Comput. Aided Drug. Res. 9: 300-307.

Murphy EJ, Booth JC, Davrazou F, Port AM, Jones DN (2013) Interactions of *Anopheles gambiae* odorant-binding proteins with a human derived-repellent: implications for the mode of action of N,N-diethylbenzamide (DEET). J. Biol. Chem. 288: 4475-4485.

Natarajan R, Basak SC, Mills D, Kraker JJ, Hawkins DM (2008) Quantitative structure-activity relationship modeling of mosquito repellents using calculated descriptors. Croat. Chem. Acta. 81: 333-340.

Odalo JO, Omolo MO, Malebo H, Angira J, Njeru PM, Ndiege IO, Hassanali A (2005) Repellency of essential oils of some plants from the Kenyan coast against *Anopheles gambiae*. Acta Tropica. 95: 210-218.

Omolo MO, Okinyo D, Ndiege IO, Lwande W, Hassanali A (2004) Repellency of essential oils of some Kenyan plants against *Anopheles gambiae*. Phytochemistry. 65: 2797-2802.

Oyedele AO, Gbolade AA, Sosan MB, Adewoyin FB, Soyelu OL, Orafidiya OO (2002) Formulation of an effective mosquito – repellent topical product from Lemongrass oil. Phytomedicine. 9: 259-262.

Park B-S, Choi W-S, Kim J-H, Kim K-H, Lee S-E (2005) Monoterpenes from thyme (*Thymus vulgaris*) as potential mosquito repellents. J. Am. Mosq. Control Assoc. 21: 80-83.

Pellegrino M, Steinbach N, Stensmyr MC, Hansson BS, Vosshall LB (2011) A natural polymorphism alters odour and DEET sensitivity in an insect odorant receptor. Nature. 478: 511-514.

Pelosi P (1994) Odorant binding proteins. Crit Rev Biochem Mol Biol. 29: 199-228.

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roy PP, Leonard JT, Roy K (2008) Exploring the impact of size training sets for the development of predictive QSAR models. Chemometr. Intell. Lab Syst. 90: 31-42.

Song J, Wang Z, Findlater A, Han Z, Jiang Z, Chen J, Zheng W, Hyde S (2013) Terpenoid mosquito repellents: a combined DFT and QSAR study. Bioorg. Med. Chem. Lett. 23: 1245-1248.

Suaya JA, Shepard DS, Siqueira JB, Martelli CT, Lum LC, Tan LH, Kongsin S, Jiamton S, Garrido F, Montoya R, Armien B, Huy R, Castillo L, Caram M, Sah BK, Sughayyar R, Tyo KR, Halstead SB (2009) Cost of dengue in eight countries in the Americas and Asia: a prospective study. Am. J. Trop. Med. Hyg. 80: 846-855.

Trongtokit Y, Rongsriyam Y, Komalamisra N, Apiwathnasorn C (2005) Comparative repellency of 38 essential oils against mosquito bites. Phytother. Res. 19: 303-309.

Wang Z, Song J, Chen J, Song Z, Shang S, Jiang Z, Han Z (2008) QSAR study of mosquito repellents from terpenoid with a six-member-ring. Bioorg. Med. Chem. Lett. 18: 2854-2859.

Wong SS-Y, Poon RW-S, Wong SC-Y (2016) Zika virus infection – the next wave after dengue? J. Formos. Med. Assoc. 115: 226-242.

World Health Organization (2014) A global brief on vector – borne diseases. Geneva: WHO Press.