f CHEMICAL ENGINEERING TRANSACTIONS
VOL. 56, 2017
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Jiří Jaromír Klemeš, Peng Yen Liew, Wai Shin Ho, Jeng Shiun Lim

ISBN 978-88-95608-47-1; **ISSN** 2283-9216
DOI: 10.3303/CET1756174

Quantitative Structure-Activity Relationship Model for Antioxidant Activity of Flavonoid Compounds in Traditional Chinese Herbs

Mismisuraya Meor Ahmad[a,b,c], Sharifah Rafidah Wan Alwi*[,a,b], Rosmahaida Jamaludin[d], Lee Suan Chua[b,e], Azizul Azri Mustaffa[a,b]

[a]Process Systems Engineering Centre (PROSPECT), Research Institute of Sustainable Environment, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b]Faculty of Chemical and Energy Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.
[c]School of Bioprocess Engineering, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
[d]Department of Management Science and Design, Razak School, Universiti Teknologi Malaysia, 50450 UTM Kuala Lumpur Malaysia
[e]Metabolites Profiling Laboratory, Institute of Bioproduct Development, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
syarifah@utm.my

There are many diseases related to the excessive amount of free radical in human body produced by various metabolic functions. The generation of free radicals can be controlled by the presence of antioxidants. One of the largest phytochemical groups in herbs which commonly exhibit antioxidant activity is flavonoid. The structure of flavonoid compounds can be represented by various types of descriptors generated by the DRAGON software. A series of flavonoids with their antioxidant activities values in 2,2'-azinobis-3-ethylbenzothiazoline-6-sulphonic acid (ABTS) assays from traditional Chinese herbs is employed as the data set. The aim of the study is to develop reliable quantitative structure-activity relationship also known as QSAR models of flavonoid compounds using the combination of forward stepwise as variable selection method and multiple linear regression (MLR) analysis. The suitable dimensional block of descriptors and significant descriptors that contribute to the antioxidant properties of flavonoids are identified. The performance of the QSAR models are reported as r^2_{calc} and are validated using cross-validation (r^2_{cv}), the external test set (r^2_{pred}) and Y-randomisation (r^2_r) to confirm the reliability of the model. Based on the findings, the developed models are robust and reliable and able to explain 78 % variance of antioxidant activity. From the QSAR models, two selected descriptors that significantly affect the antioxidant activity are topological indices (PW5) and 2D-autocorrelations (JGI4). Both of them belong to the 2-dimensional (2D) block of descriptors. This finding proves that the simpler 2D descriptors appear to be sufficient and beneficial information and perform better in building predicted model than 3D descriptors.

1. Introduction

Free radicals in the human body are produced through various metabolic functions and react against foreign invaders by triggering the immune system. However, the excessive amount of free radicals that resulted from many factors such as environment pollution, lifestyle, smoking, ultraviolet (UV) radiation and *etc.* may endanger livelihood and ultimately cause numerous diseases (Kumar and Pandey, 2013). The presence of antioxidants can control the generation of free radicals from continuous attacks to the human system. Antioxidants are molecules which can react with free radicals and terminate the reaction chain from further propagation.
One of the largest phytochemicals known as secondary plant metabolites in herbs which commonly exhibit antioxidant activity is flavonoid compounds (Mustafa et al., 2010). Flavonoid has more than 5,000 compounds (Tanwar and Modgil, 2012). The common flavonoids structure consists of two aromatic rings which are A and B rings, interconnected by three carbon atom in C rings as illustrated in Figure 1. Flavonoids can be categorised

Please cite this article as: Ahmad M.M., Wan Alwi S.R., Jamaludin R., Chua L.S., Mustaffa A.A., 2017, Quantitative structure-activity relationship model for antioxidant activity of flavonoid compounds in traditional chinese herbs, Chemical Engineering Transactions, 56, 1039-1044 DOI:10.3303/CET1756174

into several classes, namely flavanols, flavonols, chalcones, flavones, flavanones and isoflavones. Each class resulted in the difference of oxidation and pattern of substitution of the C ring, while individual compounds within a class vary in the pattern of substitution of the A and B rings (Tsuji et al., 2013).

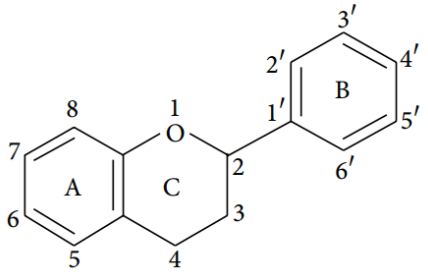

Figure 1: Basic structure of flavonoids

QSAR is a method to find empirical relationship between biological activities as dependent variables and molecular structure of compounds (descriptors) as independent variables. The QSAR model is capable of assessing the correlation between structural attributes of a series of molecules required and their activities. The main factors has been considered to develop the reliable and efficient flavonoids QSAR models is identify the suitable variable selection methods to select the most significant molecular descriptors. For examples, stepwise regression (Mitra et al., 2011), genetic algorithm (GA) (Bakhtiyor et al., 2005) and genetic function approximation (GFA) (Sivakumar and Prabhakar, 2011) are employed in developing QSAR model for flavonoids. Stepwise regression method is frequently used by many researchers because it can avoid collinearities between the descriptors (Baati et al., 2013). The aim of this paper is to develop reliable QSAR model of flavonoid from traditional Chinese herbs using the combination of forward stepwise as variable selection method and multiple linear regression (MLR) analysis. Consequently, the suitable dimensional block of descriptors and significant descriptors that contribute to the antioxidant properties of flavonoids are identified. The 0-Dimensional (0D) descriptor represents chemical formula and 1-Dimensional (1D) descriptor covers substructure analysis including structural fragments of a molecule, functional groups or substituent of interest present in the molecule. The 2-Dimensional (2D) descriptor whereas considers the relation between connected atoms in the molecules, in terms of the presence and nature of the chemical bond. The compounds in 2D representation were characterised in the molecular graph. In contrast, 3-Dimensional (3D) descriptors consider electronic descriptors of molecules. The performance of the QSAR models are reported as r^2_{calc} and are validated using cross-validation (r^2_{cv}), the external test set (r^2_{pred}) and Y-randomisation (r^2_r) to confirm the reliability of the model.

2. Methodology

2.1 Data sets arrangement

A series of flavonoids with their antioxidant activities values in 2,2'-azinobis-3-ethylbenzothiazoline-6-sulphonic acid (ABTS) assays from traditional Chinese herbs was used as data set (Cai et al., 2006). Trolox Equivalent Antioxidant Capacities (TEAC) was used to measure antioxidant activities. In this paper, the antioxidant activities of flavonoids were expressed in molar (M) and then converted to negative logarithmic scale (-Log TEAC (M)). There are 39 flavonoids compounds from six classes, namely flavonols (11), flavones (9), isoflavones (6), flavanols (5), chalcones (4) and flavanones (4). These compounds were split into 28 compounds for the training set and 11 compounds for test set based on the activity based-ranking method. The entire data set was divided based on the flavonoid classes and rearranged in descending order of their antioxidant activity prior to data splitting into training and test sets in the ratio of 2 : 1. The selection of training set enables to capture all the features of the test set (Baati et al., 2013). The training set was used for model development and test set for model validation.

2.2 Model development

Four steps are involved in developing the QSAR model. Firstly, the 2D representation of the flavonoid compounds was drawn by using ChemDraw Ultra 7.0 software (CambridgeSoft, 2002). After that, the structures were converted into 3D representation by using Chem3D Pro 7.0 software (CambridgeSoft, 2002). The structure of compounds was optimised by minimising energy using Molecular Mechanics (MM2) method. The MM2 method was applied because it considers the interaction among the atoms including electrons and orbital. The file format for descriptors calculation was MOL2. Descriptors were generated using DRAGON 6.0 software in

different dimension block (Mauri et al., 2006). The total number of generated descriptors was high and can be reduced by using objective and subjective variable selection methods.
In objective variable selection method, highly correlated and redundant descriptors were carried out using DRAGON 6.0 software while subjective variable selection method was performed in the PLS Toolbox 7.9.5 (Eigenvector_Research_ Inc., 2010) with MATLAB R2013a (Mathwork_Inc., 2013) to further reduced the number of descriptors in order to obtain highly informative descriptors. In this study, the QSAR model was developed by using forward stepwise as subjective variable selection method and multiple linear regression (MLR) analysis (Pourbasheer et al., 2014). This regression analysis was used to weight the relative contributions of the selected descriptors toward the QSAR model. The suitable dimensional block of descriptors and significant descriptors that contribute to the antioxidant properties of flavanoids are identified.
The performance of the QSAR models are reported as r^2_{calc} and are validated using cross-validation (r^2_{cv}), the external test set (r^2_{pred}) and Y-randomisation (r^2_r) to confirm the reliability of the model. The r^2_{calc} value measures how close the experimental data tracks fitted the regression line and thus quantifies any variation in the predicted data with respect to the experimental data. The acceptable value of r^2_{calc} was more than 0.6 (Kar et al., 2014). The low value of root-mean-square-error of calibration (RMSEC) and root-mean-square-error of prediction (RMSEP) was also used to evaluate the performance of QSAR model.

2.3 Stepwise Regression Method

Stepwise regression is the common subjective variable selection methods. This method is transparent but not suitable for non-linear data and more than 1,000 descriptors (Hewitt et al., 2007). The approach is based on probabilities to include or exclude a particular variable based on partial F-values. In forward stepwise regression method, the variable was selected and added to the model one at a time until no improvement is obtained and then the variable was manually removed one by one based on F-value (Pourbasheer et al., 2014).

2.4 Model Validation

The developed QSAR model was validated by using leave-one-out cross validation technique in internal validation where the value of the cross-validated squared correlation coefficient (r^2_{cv}) was calculated. In this technique, each of the compounds of the training set was removed once and the model was built with the remaining compounds. The activity of the removed compound was predicted using the model developed. This technique was repeated until all the compounds were removed at least once and then the prediction activity data can be obtained for the entire training set compound (Pogorzelska et al., 2015).
The external validation was also employed. It involves the prediction of the activity of the compound that was not used in the developed model by calculating predicted correlation coefficient (r^2_{pred}) (Gao et al., 2016). If a value of r^2_{pred} greater than the stipulated value of 0.5, it reflects the efficient prediction for the test set by the developed model. r^2_{pred} may not indicate the predicted activity value thoroughly (Mitra et al., 2011) because it was dependent on the sum of squared difference between the experimental activity data of the test set and the training set compounds. Compounds with a wide range of activity data may give a large value of the r^2_{pred}. To prevent this error, the r^2_m metric with a threshold value of 0.5 was calculated using Eq(1) (Mitra et al., 2010) where the r^2 and r^2_o were referring to the squared correlation coefficient values between the observed and predicted activity data with and without the intercept.

$$r^2_m = r^2\left(1 - \sqrt{(r^2 - r^2_o)}\right) \quad (1)$$

Validation of the developed models was also done by using Y-randomisation/scrambling technique to ensure that the models did not merely capture noise and to assess if the models were the result of chance correlations (Goodarzi et al., 2013). This technique was rebuilding the models using shuffled or randomised activities of the training sets. Then, the evaluation of predictive accuracy of the resultant models was compared with the original model to identify either the robust and reliable QSAR model was produced (Nowaczyk and Kulig, 2012).

3. Results and Discussion

The number of descriptors generated in DRAGON software for each dimension block was high as shown in Table 1. The variable selection method should be applied to reduce the number of descriptors (Campos and Melo, 2014). Based on Table 1, the number of generated descriptors using DRAGON software was reduced considerably through objective variable selection method before it can be further analysis in the PLS Toolbox 7.9.5 with MATLAB R2013a. For example, 4,885 from 0D-3D dimensional block of generated descriptors was reduced to 456 after 4,429 descriptors were removed by using the objective variable selection method.

Table 1: Number of descriptors from different dimensional block of descriptors

Dimension block of descriptors	Descriptors generated (DRAGON software)	Descriptors removed through objective variable selection Method	Remaining descriptors for further analysis
0D-3D	4,885	4,429	456
0D-2D	3,717	3,518	199
3D	1,106	865	241

The remaining descriptors were further analysis in the forward stepwise as subjective variable selection method combined with the MLR analysis to identify the most significant descriptors for each dimension block of descriptors. Table 2 exhibits the detailed statistical parameters in different dimension block of descriptors that were directly obtained using PLS Toolbox 7.9.5 with MATLAB R2013a software. The robust and reliable developed QSAR model to describe the relationship between the structures of flavonoids and their antioxidant activity was chosen based on the high values of r^2_{calc}, r^2_{cv}, r^2_{pred}, r^2_m as well as having the least number of descriptors as possible (Fernandez et al., 2005).

-LogTEAC(M) = 3.7436 + (62.8886 x PW5) + (-132.4030 x JGI4) (2)

The developed QSAR model as shown in Eq(2) with two significant descriptors were selected in forward stepwise and MLR method from the 0D-3D as well as 0D-2D dimensional block of descriptors. This QSAR model produced high statistical parameter values. The two significant descriptors selected were from the topological indices (PW5) and 2D-autocorrelations (JGI4) descriptors. It is interesting to note that, the selection of significant descriptors in 0D-3D and 0D-2D dimensional block was similar with the high value of r^2_{calc} with 0.78 and low value of RMSEC with 0.39. The value of r^2_{calc} which exceeds the stipulated value of 0.6 indicated that the predicted activity data was closely fitted with experimental data (Kar et al., 2014).

Moreover, the calculated value for internal validation using leave-one-out cross validation technique, r^2_{cv} was 0.72 and the predicted value, r^2_{pred} was 0.76. This value shows that developed QSAR can avoid over fitted since the value of r^2_{calc} was higher than r^2_{cv} and has good agreement with Nowaczyk and Kulig (2012). The values of $r^2_{m(loo)}$, 0.78 and $r^2_{m(pred)}$, 0.71 were more than the stipulated value of 0.5. This indicates that the predicted antioxidant activity values were also in close proximity to the experimental antioxidant activity values. In the case of the model validation using Y-randomisation/scrambling, the average value of r^2_r for ten randomisation runs was 0.06. It can be concluded that the developed QSAR model using Eq(2) was robust enough and free of the chance correlation.

The developed QSAR model using 3D dimensional block of descriptors was not fulfilling the statistical parameters requirement especially low in the prediction performance, 0.43. Reliable and robust QSAR model cannot be constructed. This result also demonstrated that 3D dimensional block of descriptors was not suitable to represent the correlation between a series of flavonoid structures and antioxidant activity.

Table 2: The statistical parameter values of the developed QSAR model in different block of descriptors

Dimension block of descriptors	Descriptors	r^2_{calc}	r^2_{cv}	r^2_{pred}	RMSEC	RMSECV	RMSEP
0D-3D 0D-2D	JGI4; PW5	0.78	0.72	0.76	0.39	0.45	0.53
3D	Mor17v; H0s	0.62	0.49	0.43	0.52	0.60	0.79

According to Eq(2), the regression coefficient value of PW5 and JGI4 descriptors play an important role in the contribution to the antioxidant properties of flavonoids. All the predicted values for training and test sets of antioxidant activity in ABTS assays of flavonoids using Eq(2) were plotted against the experimental activity as illustrated in Figure 2. The graph shows that the points were scattered around the line of fit. This again implicated the predictive efficacy of the developed QSAR model.

Eq(2) signifies the importance of the PW5 descriptor where it represents a positive coefficient. This suggests that the antioxidant activity of flavonoids (TEAC(M)) reduced with an increase in the value of PW5 descriptor. PW5 is a topological index that considers the shape of molecules as molecular properties in the variations of compounds (Randic and Razinger, 1995). The shape of molecules with specific kind of branching was also selected as a significant descriptor in the QSAR model developed by Mitra et al. (2010). The variation in branching structural features of flavonoids was also considered by Ray et al. (2007) to develop the specific

QSAR model. PW5 refers to the proportions of path/walk in length 5 from molecular Randic shape index. Randic (2001) characterises shape index for a molecular graph by considering both paths and walks of different length within a graph, and then making the proportions of the number of path and the number of walk the same length.

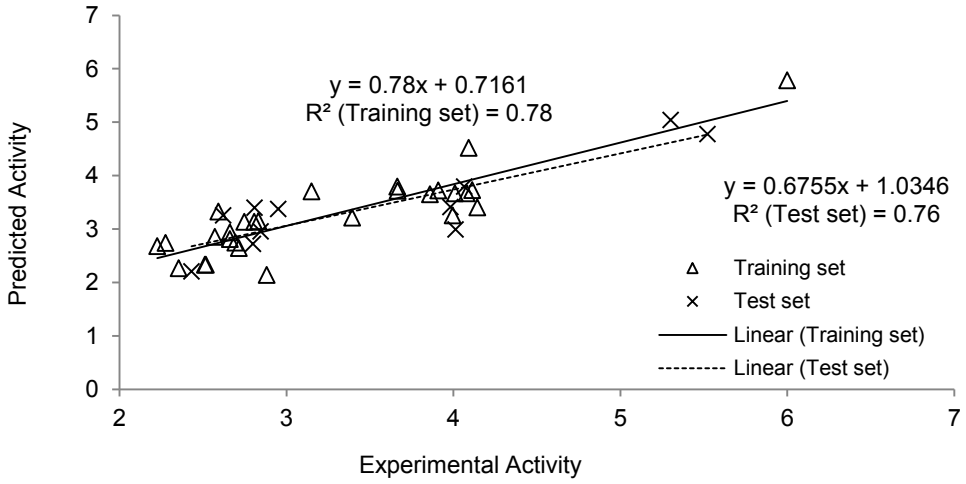

Figure 2: The predicted against experimental antioxidant activity plot using QSAR model - Eq(2)

JGI4 descriptor bears a negative coefficient where it recommends that the antioxidant activity of the molecules (TEAC(M)) increase with an enhancement in the value of this descriptor. This suggests that some unique charge distribution was needed for increase antioxidant activity in ABTS assay. JGI4 refers to mean topological charge index of order 4 from 2D autocorrelation. This type of descriptors but in the different order, JGI5 and JGI8 were also selected and showed the highest coefficient in the QSAR model developed by Farkas et al. (2004). The graph of H-depleted molecular structure was used to represent topological charge index for atoms in a molecule (Nowaczyk and Kulig, 2012).

4. Conclusion

The robust and reliable QSAR model is successfully developed to explain the relationship between a series of flavonoids from traditional Chinese herbs and their antioxidant activities. The significance and robustness of the developed QSAR models have been confirmed by leave-one-out cross validation, external validation and Y-randomisation/scrambling techniques. The two significant descriptors were PW5 and JGI4. Both of them belong to the 2-dimensional (2D) block of descriptors. The low value of specific topological indices of molecules (PW5) leads to high antioxidant activity value (TEAC(M)), while the high value of mean topological charge index (JGI4) gives an enhancement effect on the antioxidant activity value (TEAC(M)). This finding proves that the simpler 2D descriptors appear to be sufficient and beneficial information and perform better in building predicted model than 3D descriptors.

Acknowledgments

The authors gratefully acknowledge financial support from Ministry of Higher Education Malaysia under IPTA Academic Training Scheme (SLAI) Universiti Malaysia Perlis (UniMAP), GUP grant (Q.J130000.2509.12H88) and Universiti Teknologi Malaysia (UTM).

References

Baati N., Nanchen A., Stoessel F., Meyer T., 2013, Quantitative Structure-Property Relationships for Thermal Stability and Explosive Properties of Chemicals, Chemical Engineering Transactions 31, 841-846.

Bakhtiyor F.R., Nasrulla D.A., Syrov V.N., Jerzy L., 2005, A Quantitative Structure-Activity Relationship (QSAR) Study of the Antioxidant Activity of Flavonoids, QSAR & Combinatorial Science 24, 1056-1065.

Cai Y., Sun M., Xing J., Luo Q., Corke H., 2006, Structural-radical scavenging activity relationship of phenolic compounds from traditional chinese medical plants, Life Science 78 (25), 2872-2888.

CambridgeSoft, C., 2002, ChemDraw Ultra. 2002, Massachusetts, USA: ChemOffice.

Campos L.J.D., Melo E.B.D., 2014, Modeling Structure-Activity Relationships Of Prodiginines With Antimalarial Activity Using GA/MLR And OPS/PLS, Journal of Molecular Graphics and Modelling 54, 19-31.

Eigenvector_Research_ Inc., 2010, PLS_Toolbox. Washington, USA.
Farkas O., Jakus J., Héberger K., 2004, Quantitative Structure – Antioxidant Activity Relationships of Flavonoid Compounds, Molecules 9 (12), 1079-1088.
Fernandez M., Caballero J., Helguera A.M., Castro E.A., Gonzalez M.P., 2005, Quantitative structure-antioxidant activity relationships to predict differential inhibition of aldose reductase by flavonoid compounds, Bioorganic & Medical Chemistry 13, 3269-3277.
Gao X., Han L., Ren Y., 2016, In Silico Exploration of 1,7-Diazacarbazole Analogs as Checkpoint Kinase 1 Inhibitors by Using 3D QSAR, Molecular Docking Study, and Molecular Dynamics Simulations, Molecules 21 (5), 1-15.
Goodarzi M., Funar-Timofei S., Heyden Y.V., 2013, Towards better understanding of feature selection or reduction techniques for quantitative structure-activity relationship models, Trends in Analytical Chemistry 42, 49-63.
Hewitt M., Cronin M.T.D., Madden J.C., Rowe P.H., Johnson C., Obi A., Enoch S.J., 2007, Consensus QSAR Models: Do the Benefits Outweigh the Complexity?, Journal of Chemical Information and Modelling 47 (4), 1460-1468.
Kar S., Gajewicz A., Puzyn T., Roy K., 2014, Nano-quantitative structure-activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells, Toxicology in Vitro 28 (4), 600-606.
Kumar S., Pandey A.K., 2013, Chemistry and Biological Activities of Flavonoids: An Overview, The Scientific World Journal 2013, ID 162750, DOI: 10.1155/2013/162750
Mathwork_Inc., 2013, Mathworks, Natick, Massachusetts, United States.
Mauri A., Consonni V., Pavan M., Todeschini R., 2006, DRAGON Software: An Easy Approach to Molecular Descriptor Calculations, Communications in Mathematical and in Computer Chemistry 56, 237-248.
Mitra I., Saha A., Roy K., 2010, Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, Molecular Simulation 36 (13), 1067-1079.
Mitra I., Saha A., Roy, K., 2011, Chemometric QSAR Modeling and In Silico Design of Antioxidant No Donor Phenols, Scientia Pharmaceutical 79 (1), 31-57.
Mustafa R.A., Abdul Hamid A., Mohamed S., Bakar F.A., 2010, Total phenolic compounds, flavonoids, and radical scavenging activity of 21 selected tropical plants, Journal Food Science 75 (1), 28-35.
Nowaczyk A., Kulig K., 2012, QSAR Studies on a Number of Pyrrolidin-2-one Antiarrhythmic Arylpiperazinyls, Medical Chemistry Research 21 (3), 373-381.
Pogorzelska A., Slawinski J., Brozewicz K., Ulenberg S., Baczek T., 2015, Novel 3-Amino-6-chloro-7-(azol-2 or 5-yl)-1,1-dioxo-1,4,2-benzodithiazine Derivatives with Anticancer Activity: Synthesis and QSAR Study, Molecules 20 (12), 21960-21970.
Pourbasheer E., Aalizadeh R., Tabar S.S., Ganjali M.R., Norouzi P., Shadmanesh J., 2014, 2D and 3D-QSAR study of Hepatitis C Virus NS5B Polymerase inhibitors by CoMFA and CoMSIA methods, Journal of Chemical Information and Modelling 54 (10), 2902-2914.
Randic M., 2001, Novel Shape Descriptors for molecular Graphs, Journal of Chemical Information and Modelling 41, 607-613.
Randic M., Razinger M., 1995, Molecular Topographic Indices, Journal of Chemical Information and Modelling 35 (1), 140-147.
Ray S., Sengupta C., Roy K., 2007, QSAR modeling of antiradical and antioxidant activities of flavonoids using electrotopological state (E-state) atom parameters, Central European Journal of Chemistry 5 (4), 1094-1113.
Sivakumar P.M., Prabhakar P.K., Doble M., 2011, Synthesis, antioxidant evaluation and quantitative structure-activity relationship studies of chalcones, Medical Chemistry Research 20 (4), 482-492.
Tanwar B., Modgil R., 2012, Flavonoids: dietry occurence and health benefits, Spatula DD 2 (1), 59-68.
Tsuji P.A., Stephenson K.K., Wade K.L., Liu H., Fahey J.W., 2013, Structure-activity analysis of flavonoids: direct and indirect antioxidant, and antiinflammatory potencies and toxicities, Nutrition Cancer 65 (7), 1014-1025.