Supporting Information

Evaluation and Discovery of novel synthetic chalcone derivatives as anti-inflammatory agents

Jianzhang Wu, Jianling Li, Yuepiao Cai, Yong Pan, Faqing Ye, Yali Zhang, Yunjie Zhao, Shulin Yang, Xiaokun Li, Guang Liang *

Content： Quantitative structure-activity relationship (QSAR) study for synthetic chalcones

Materials and Methods

Descriptors calculation and selection

The molecular structures of all the chalcones were built with Maestro (Version 9.1 Schrödinger, LLC). The full geometry optimization for the investigated molecules was carried out with MOPAC2009 version 9.0 [1]. All the calculations were based on the semi-empirical Parameterized model 6 (PM6) method [2]. The molecular descriptor computing were performed on MODEL (Molecular Descriptor Lab), a web-based server for computing structural and physicochemical features of compounds, according to the methods described in the literature [3]. The descriptors studied here contain the constitutional descriptors, physicochemical descriptors, topological descriptors, geometrical descriptors, charge (electronic) descriptors, and quantum chemistry descriptors. The optimized geometry of molecules was uploaded to MODEL. After the calculation of the molecular descriptors, about 4000 molecular descriptors based on molecular 3D structure were obtained. The constant descriptors in all molecules were eliminated, pairs of variables with a correlation coefficient greater than 0.85 were considered as intercorrelated, and one of each correlated pair was deleted.

Multiple linear regression (MLR) analysis

MLR analysis was performed using R program, a powerful tool for statistical computing and graphics, to derive QSAR models. The biological data used in this study were their TNF-α- or IL-6-inhibitory rates when compared to LPS alone group. Compounds with negative values were abandoned because of their pro-inflammatory activities. The inhibition rates against TNF-α and IL-6 release, named as $IR_{TNF-\alpha}$ and IR_{IL-6} respectively, were used as dependent variables in the linearization procedure. Stepwise multiple linear regression (Stepwise-MLR) was used to select the significant descriptors. The most relevant descriptors were used as independent variables.

Validation of the models

Validation of the lineal models is required for testing the predictive ability and generalizing the methods by cross-validation. The leave-one-out (LOO) procedure was employed. When a data point was removed from the analyzed set, the regression was recalculated, and then the predicted value for that point was compared to its actual value. This process was repeated until each datum had been omitted once and then the sum of squares of these deletion residuals could be used to calculate q^2, an equivalent statistic to R^2.

Table S1. The calculated descriptors in this work

Variables	Descriptor Class	Description
R-GETAWAY1	Geometrical properties	R-GETAWAY descriptors R6 (weighted by atomic van der Waals radii)
R-GETAWAY2	Geometrical properties	R-GETAWAY descriptors RT+ (weighted by atomic sanderson electronegativity radii)
3D-MoRSE	Geometrical properties	Unweighted 3D-MoRSE descriptors I(2)
Moran	Topological Indices	Moran autocorrelation–lag 4 (weighted by atomic electronegativity)
EEVA	Quantum chemistry properties	Electronic Eigenvalue Descriptors (-39.25ev)
EVA	Quantum Chemistry properties	EVA Descriptors (754.00)

Results

One-variable and multivariable regressions between the different activities and the abundant descriptors were studied. To minimize the effect of co-linearity and avoid over-fitting, the number of variables was limited to three. The attempt on both one-variable and two-variable models were found to be unsuccessful for anti-IL-6 and anti-TNF-α activities of these chalcones. The statistically significant models were only obtained with three variables for anti-IL-6 activities (Eq.1 and Eq.2 in Figure 3). The involved molecular

descriptors and their corresponding definition were listed in Table S1. However, no significantly statistical correlation between anti-TNF-α activities and the molecular descriptors was obtained.

Eq. 1 was the best QSAR model obtained with the GETAWAY descriptors, which had a high adjusted squared regression coefficient (R^2_{adj}=0.72) and was able to describe more than 74% of the variance in the experimental activity. All variables in Model 1 were the geometrical molecular descriptors. Two of those, R-GETAWAY1 and R-GETAWAY2, are belong to GETAWAY descriptor family (GEometric, Topological and Atomic Weighted AssemblY descriptors), proposed by Consonni [4] and derived from the Molecular Influence Matrix (MIM). R-GETAWAY descriptors were from the *influence/distance matrix* **R** where the elements of the MIM were combined with those of the geometry matrix. R-GETAWAY descriptors can be calculated with different atomic weightings, such as van der Waals volume, electronegativity, atomic mass, and polarizability. Several positive features of these descriptors have been confirmed by recent surveys. The models based on these descriptors illustrated good predictive potentials. R-GETAWAY1 and R-GETAWAY2 have opposite influence in the model, increasing R-GETAWAY1 or decreasing R-GETAWAY2 would enhance the anti-inflammatory activities of the chalcone analogues. The correlation analysis shows that a correlation coefficient of 0.635 between R-GETAWAY2 and IR_{IL-6} in this model. Furthermore, R-GETAWAY2 weights according to atomic Sanderson electronegativity radii, indicating that the electronegativity plays an important role in the anti-inflammatory activities of these compounds. Another variable is 3D-MoRSE descriptor (3D-Molecule Representation of Structures based on Electron diffraction), which is developed from the idea of obtaining information from the 3D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves [5].

Attempts were also made to construct linear models with other descriptors. Eq. 2 was obtained. Despite its squared regression coefficient (0.73) was close to that of Eq.1, but the variables in Model 2 were completely different from those in Model 1. Quantum chemistry descriptors and topological indices were employed to construct model 2. The 3D structure-sensitive EVA descriptor (EigenVAlue descriptors) is derived from fundamental mid- and near-infrared range molecular vibrational frequencies. It is usually used as a fundamental molecular property reliably and easily characterized from the potential energy function [6]. EEVA descriptors (Electronic EigenVAlue descriptors) are vector-descriptor proposed as a modification of EVA. Semi-empirical molecular orbital energies, i.e. the eigenvalues of the Schrödinger equation, are used instead of the vibrational frequencies of the molecule [6]. Both are quantum chemistry descriptors. The third variable,

Moran, is an autocorrelation descriptor based on the autocorrelation function.

Based on the statistic analysis, two satisfactory models for anti-IL-6 activities were obtained. In spite of the similar R^2 they have, the correlation between compound activities and structure information could be illustrated from different perspectives by using various descriptors. R^2 was a common parameter to test the validity of models, while it could be increased artificially by adding more variables (descriptors). To better test the validity of the models, the q^2 values were adopted as a more reliable statistic parameter, which means the square of the correlation coeffcient of the cross-validation and are calculated from LOO test. The q^2 values could be used as a measure of the predictive ability of a regression equation. The q^2 values of equation 1 and equation 2 are 0.71 and 0.70, respectively, indicating the stability of these models.

The scatter plot of predicted versus experimental values is illustrated in Figure 3. As can be seen, the equation 1 and equation 2 have modest quality and the variables used in these equations can explain the variance in the anti-IL-6 activities of chalcone derivatives.

Reference

1. Stewart, J. J. P. Stewart Computational Chemistry, Colorado Springs, USA, **2009.** (Http://OpenMOPAC.net)
2. Stewart, J. J. P. Optimization of parameters for semi-empirical methods V: modification of NDDO approximations and application to 70 elements. *J Mol Modeling*, **2007**, 13, 1173-1213
3. Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z. MODEL - Molecular Descriptor Lab: A Web-Based Server for Computing Structural and Physicochemical Features of Compounds. *Biotechnol. Bioeng,* **2007,** 97, 389-396.
4. Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors, *J Chem Inf Comput Sci*, **2002**, 42, 682-692
5. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D Space, *J Chem Inf Comput Sci*, **1996**, 36, 1030-1037
6. Todeschini, R.; Consonni, V.; Handbook of Molecular Descriptors, first ed. by Wiley-VCH, Mannheim, Gemany, **2000**.