Journal of Molecular Structure 1108 (2016) 168–178

Contents lists available at ScienceDirect

Journal of Molecular Structure

journal homepage: http://www.elsevier.com/locate/molstruc

Quantitative structure–activity relationship of the curcumin-related compounds using various regression methods

Ardeshir Khazaei [a, **], Negin Sarmasti [a], Jaber Yousefi Seyf [b, *]

[a] Faculty of Chemistry, Bu-Ali Sina University, Hamedan 6517838683, Iran
[b] Department of Chemical Engineering, Tarbiat Modares University, 14115-143 Tehran, Iran

ARTICLE INFO

Article history:
Received 27 June 2015
Received in revised form
24 November 2015
Accepted 25 November 2015
Available online 8 December 2015

Keywords:
Cancer
Curcumin
QSAR
Sphere exclusion
MLR
PLS
PCR
SW
GA
SA

ABSTRACT

Quantitative structure activity relationship were used to study a series of curcumin-related compounds with inhibitory effect on prostate cancer PC-3 cells, pancreas cancer Panc-1 cells, and colon cancer HT-29 cells. Sphere exclusion method was used to split data set in two categories of train and test set. Multiple linear regression, principal component regression and partial least squares were used as the regression methods. In other hand, to investigate the effect of feature selection methods, stepwise, Genetic algorithm, and simulated annealing were used. In two cases (PC-3 cells and Panc-1 cells), the best models were generated by a combination of multiple linear regression and stepwise (PC-3 cells: $r^2 = 0.86$, $q^2 = 0.82$, pred_$r^2 = 0.93$, and $r^2_{m\ (test)} = 0.43$, Panc-1 cells: $r^2 = 0.85$, $q^2 = 0.80$, pred_$r^2 = 0.71$, and $r^2_{m\ (test)} = 0.68$). For the HT-29 cells, principal component regression with stepwise ($r^2 = 0.69$, $q^2 = 0.62$, pred_$r^2 = 0.54$, and $r^2_{m\ (test)} = 0.41$) is the best method. The QSAR study reveals descriptors which have crucial role in the inhibitory property of curcumin-like compounds. 6ChainCount, T_C_C_1, and T_O_O_7 are the most important descriptors that have the greatest effect. With a specific end goal to design and optimization of novel efficient curcumin-related compounds it is useful to introduce heteroatoms such as nitrogen, oxygen, and sulfur atoms in the chemical structure (reduce the contribution of T_C_C_1 descriptor) and increase the contribution of 6ChainCount and T_O_O_7 descriptors. Models can be useful in the better design of some novel curcumin-related compounds that can be used in the treatment of prostate, pancreas, and colon cancers.

1. Introduction

Natural plant products have been used throughout human history for various purposes. Turmeric as a natural plant is cultivated in in China, Southeast Asia, India, and other Asian and tropical countries and regions. In addition, it is common in other countries of the world and is recognized by different names in different languages worldwide [1]. Curcumin ((1E,, 6E)-1,7-bis (4-hydroxy-3-methoxyphenyl)-1,6-heptadiene-3,5-dione) or diferuloylmethane is a major constituent of turmeric which has gained remarkable attention recently for its multiple pharmacological activities, including antioxidant, anti-inflammatory, antibacterial, and antiviral activities, as well as therapeutic potential for cancer and Alzheimer's disease [1–3]. Due to its anticancer properties, low molecular weight, and nontoxicity, researcher have focused to develop possible novel curcumin-like anticancer pharmaceuticals. It has been accounted for that the oral bioavailability of curcumin has limited due to its low dissolution rate and poor solubility in water [4] which diminish the clinical usefulness of curcumin. In view of this, numerous curcumine-like compound have been synthesized not only to increase low bioavailability of curcumin, but also to enhance its selectivity and potency in the treatment of human cancers [5–9]. Despite the synthesis of a large number of curcumin-like compounds, a few of them were found to have lower inhibitory concentration (IC_{50}) than curcumin when tested in vitro on cultured cancer cells [10].

In other hand, the synthesis of many compounds to find the desired target molecule is time consuming and costly. In that event, ones can obtain a rapid and cost-effective biological activity by QSAR without necessity of performing laboratory experiments [11]. The quantitative structure–activity relationship (QSAR) relates a set of physico-chemical properties or molecular descriptors to a response-variable which could be any response such as biological

* Corresponding author.
** Corresponding author.
E-mail address: jabneg@gmail.com (J.Y. Seyf).

http://dx.doi.org/10.1016/j.molstruc.2015.11.072
0022-2860/

activity of the chemicals [12,13]. Recently, sixty-one curcumin-related compounds have been synthesized by Wei et al. for their anticancer activity toward cultured prostate cancer PC-3 cells, pancreas cancer Panc-1 cells, and colon cancer HT-29 cells [14]. Although, their structure activity relationship (SAR) have been investigated, but QSAR studies have not been carried out for these compounds. The set of 61 molecules of curcumin-related compounds with anticancer activity is sufficient to obtain the QSAR models with high predictive power. According to the above matter, we developed some statistically significant QSAR models for curcumin-related derivatives with anticancer activity. Models can be useful to further and effective designing novel curcumin-related derivatives.

2. Materials and method

2.1. Data set and structure optimization

A set of 61 molecules of curcumin-related compounds with anticancer property were used for the present QSAR study [14]. Salts molecule and molecules with indeterminate IC_{50} removed from QSAR study. The chemical structure and biological activities (IC_{50}) of these molecules are shown in the Tables 1–3. To reduce the skewness of data set, the IC_{50} values were converted to a logarithmic base (pIC_{50} = -log (IC_{50})). Subsequently pIC_{50} values were used as the response values in the QSAR studies. As we know certain descriptors depend on 3D structure and compound should be represented in their bioactive conformations, but there is not any information on the bioactive conformation. In view of this, we use a minimum energy conformation of the compounds. The compounds were then subjected to conformational analysis and energy minimization using Monte Carlo conformational search with root mean square (RMS) gradient of 0.001 kcal/mol using a Merck Molecular Force Field (MMFF) force field. The Monte Carlo conformational search method is similar to the random incremental pulse search (RIPS) method that generates a new molecular conformation by randomly perturbing the position of each coordinate of each atom in the molecule. Structure with the lowest energy for each compound was generated after energy minimization. Then, optimized structures and corresponding pIC_{50} were imported into the VLife MDS 3.5 software [15]. Vlife MDS is complete molecular modelling software which can perform tasks such as QSAR, combinatorial Library generation, pharmacophore, cheminformatics, docking, etc. The energy-minimized geometries were used for the calculation of the various 2D molecular descriptors (Individual, Chi, ChiV, Path count, ChiChain, ChiVChain, Chainpathcount, Cluster, Pathcluster, Kapa, Element Count, and so on). In addition, alignment independent descriptors were calculated and were added to the descriptor list. A descriptor that is constant for all the molecules will not contribute to QSAR and hence were removed. Training and test set were created by using a sphere exclusion method for choosing uniformly distributed molecules in both sets [16,17]. In the acceptable dissimilarity value or sphere exclusion radius the test set should be interpolative i.e. derived within the min–max range of the train set. Dissimilarity value was set such that test set approximately 20% of the total number of molecules in the study [18].

2.2. Regression and variable selection methods

Feature or variable selection is one of the crucial steps in a QSAR procedure, which known as variable selection technique [19]. In principle, any variable selection method can be coupled with any regression method of choice for building quantitative model. Multiple linear regression (MLR) [20], principal component regression (PCR) [21], and partial least squares (PLS) [22] were used as the regression methods and stepwise, Genetic algorithm (GA), and simulated annealing (SA) were used as the variable methods [23]. In the stepwise (SW) feature selection method, forward-backward (FB) were used and the cross-correlation limit was set at 0.5, the number of variables at 4, and the term selection criteria at r^2. In the GA method, cross correlation limit, population, and number of generations were set at 0.5, 10, and 1000, respectively. In the SA method, maximum temperature, minimum temperature, and cross correlation limit were set at 100, 0.01, and 0.5, respectively. At each radius, GA, SA, and SW as the variable selection methods were used with the MLR, PCR and PLS as the regression methods. Vlife MDS software was used in the procedure of feature selection and regressions.

2.3. Statistical analysis

The performance and quality of the developed models were evaluated by various criteria, for example using r^2 (the squared correlation coefficient), q^2 (cross-validated correlation coefficient), F-test (Fischer's value) for statistical significance, and pred_r^2 (r^2 for the external test set), etc. The main application of a QSAR model is its capability replicated by the model. However, if the following criteria are met a QSAR model will be robust and predictive: $r^2 > 0.6$, $q^2 > 0.6$ and pred_$r^2 > 0.5$ [24]. Validation of the external and internal ability of a model is evaluated by some validation criteria. Most well-known and frequently use criteria are q^2 and pred_r^2 for internal and external validation, respectively.

2.3.1. Internal validation

The cross-validation analysis was performed using the leave-one-out (LOO) method. Internal or cross-validation criterion (q^2) is calculated with the following formula:

$$q^2 = 1 - \frac{\sum_{i=1}^{training}\left(y_i - y_{pred-i}\right)^2}{\sum_{i=1}^{training}\left(y_i - \overline{y}\right)^2} \quad (1)$$

In the above equation, $\bar{y}$ means average activity value of the training dataset while y_i represent observed activity. Where the notation $y_{pred\text{-}i}$ indicates that the response is predicted by a model estimated when the i-th sample was left out from the training set.

Both summations are over all molecules in the training set and hence the predictions were based on the current trial solution, the q^2 obtained indicates the predictive power of the current model.

2.3.2. External validation

When truly external date points aren't available for test of the model prediction ability, original date set are divided into training and test set. The model is built from train set, and then is validated with the test set. pred_r^2 is calculated according to the following equation:

$$R^2_{pred} = 1 - \frac{\sum_{i=1}^{test}\left(y_{pred(test)} - y_{(test)}\right)^2}{\sum_{i=1}^{test}\left(y_{(test)} - \overline{y}_{training}\right)^2} \quad (2)$$

Where $y_{pred(test)}$ and are y_{test} predicted and observed activity for test set and $\bar{y}_{training}$ is the mean activity value of train test. Pred_r^2 is used as a yardstick to assess the ability of a model to predict the unknown activity of the new compounds. Both summations are over all molecules in the test set. The pred_r^2 value is indicative of the predictive power of the current model for the external test set. The statistical significance of the model is determined by the F-test which high value is interest.