A QSAR study of some cyclobutenediones as CCR1 antagonists by artificial neural networks based on principal component analysis

Shahlaei M., Fassihi A., Saghaie L., Arkan E., Pourhossein A.

Department of Medicinal Chemistry, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, Department of Medicinal Chemistry, School of Pharmacy and Isfahan Pharmaceutical Sciences Research Center, Isfahan University of Medical Sciences, Isfahan, Department of Medical Nanotechnology, School of Advanced Medical Technologies, Tehran University of Medical Sciences, Tehran, Department of chemical engineering, Science and Research branch, Islamic Azad university, Kermanshah, Iran.

ABSTRACT

Background and the purpose of the study: A quantitative structure activity relationship (QSAR) model based on artificial neural networks (ANN) was developed to study the activities of 29 derivatives of 3-amino-4-(2-(2-(4-benzylpiperazin-1-yl)-2-oxoethoxy) phenylamino) cyclobutenedione as C-C chemokine receptor type 1 (CCR1) inhibitors.

Methods: A feed-forward ANN with error back-propagation learning algorithm was used for model building which was achieved by optimizing initial learning rate, learning momentum, epoch and the number of hidden neurons.

Results: Good results were obtained with a Root Mean Square Error (RMSE) and correlation coefficients ($R^2$) of 0.189 and 0.906 for the training and 0.103 and 0.932 prediction sets, respectively.

Conclusion: The results reflect a nonlinear relationship between the Principal components obtained from calculated molecular descriptors and the inhibitory activities of the investigated molecules.

Keywords: Quantitative Structure Activity Relationship, inhibitory activity, feed-forward ANN, PCA.

INTRODUCTION

The chemokine proteins are a class of small molecules that play a significant role in leukocyte-trafficking during immune response (1). CCR1, as one of the chemokine receptors, is expressed on a number of human cells such as monocytes, macrophages, dendritic cells, and T cells (2). A large number of studies have provided strong evidences for a significant role of the chemokines, RANTES (Regulated upon Activation, Normal T cell Expressed and presumably Secreted), and MIP-1a (Macrophage inflammatory protein-1) in chronic inflammatory diseases. Because MIP-1a and RANTES are agonists for CCR1, antagonists for this protein may be helpful in treatment of these diseases (3).

QSAR has become a very well known discipline in the drug discovery researches (4-6). The basis of such relationships is the assumption that the variation of bioactivity of molecules, as expressed by $pIC_{50}$, can be regressed with changes in molecular descriptors. Development of QSAR involves selection of most informative independent variables to describe different sets of molecules and the application of various algorithms, such as multiple linear regression or ANN to construct the QSAR model. The advantage of ANN is it inherent power to reveal non-linear relationships between independent and dependent variables in the derivation of the QSAR models.

A quantitative structure activity relationship (QSAR) model based on artificial neural networks (ANN) was developed to study the activities of 29 derivatives of 3-amino-4-(2-(4-benzylpiperazin-1-yl)-2-oxoethoxy) phenylamino cyclobutenedione as C-C chemokine receptor type 1 (CCR1) inhibitors (7).

METHODS

Biological and chemical data from 29 derivatives of 4-3-amino-4-(2-(4-benzylpiperazin-1-yl)-2-oxoethoxy)phenylamino)cyclobutenedione were used in this study (Table 1) (7). All the structures were drawn and optimized using the semiempirical quantum-chemical routine of AM1 implemented in
Table 1. The structures and biological activities of compounds

| Compound | R<sub>1</sub> | Observed pIC<sub>50</sub> | Predicted pIC<sub>50</sub> |
|----------|--------------|-------------------------|--------------------------|
| 1<sup>+</sup> | | 7.585 | 7.802 |
| 2 | | 6.455 | 6.321 |
| 3 | | 7.107 | 7.125 |
| 4 | | 6.568 | 6.628 |
| 5 | | 7.013 | 7.117 |
| 6 | | 5.630 | 5.739 |
| 7 | | 5.536 | 5.234 |
| 8 | | 6.657 | 6.661 |
Table 1. (Cont.) The structures and biological activities of compounds 9-13

![Chemical structure of compounds 9-13](image)

| Compound | $R_2$ | $R_3$ | Observed $pIC_{50}$ | Predicted $pIC_{50}$ |
|----------|-------|-------|---------------------|----------------------|
| 9        | H     | H     | 6.036               | 6.012                |
| 10       | H     | Br    | 7.602               | 7.243                |
| 11       | H     | F     | 6.795               | 6.462                |
| 12       | H     | Me    | 7.045               | 7.001                |
| 13       | F     | F     | 6.853               | 7.192                |

Table 1. (Cont.) The structures and biological activities of compounds 14-27

![Chemical structure of compounds 14-27](image)

| Compound | $R_4$ | Observed $pIC_{50}$ | Predicted $pIC_{50}$ |
|----------|-------|---------------------|----------------------|
| 14       | Et    | 7.142               | 7.118                |
| 15       | Pr    | 6.769               | 7.097                |
| 16       | CH$_2$Ph | 6.096           | 6.083                |
| 17       | H     | 8.000               | 7.976                |
| 18       |       | 6.795               | 6.786                |
| 19       |       | 7.200               | 7.203                |
| 20       |       | 7.698               | 7.970                |
| 21       |       | 7.744               | 7.693                |
| 22       |       | 7.193               | 7.192                |
In order to model the biological activities of the studied compounds, four classes of descriptors were calculated: constitutional, geometrical, topological, and functional group using Dragon (9) on the minimal energy conformations. The data set was divided into the training and the testing sets based on Kennard and Stone algorithm (10). Principal component analysis was used to compress a pool of descriptors into principal components (PCs) as new variables. After that, a model using a nonlinear regression model, artificial neural network, was constructed to make a relationship between PCs and the pIC₅₀. A feed-forward ANN with error back-propagation learning algorithm was applied for model building.

RESULT AND DISCUSSION

Results
PCA was performed on the data set that gives 14 significant PCs (% variance explained > 1). Fourteen

Table 1. (Cont.) The structures and biological activities of compounds 14-27

| Compound | Rᵢ | Observed pIC₅₀ | Predicted pIC₅₀ |
|----------|----|----------------|-----------------|
| 23       |    | 7.920          | 7.713           |
| 24       |    | 7.301          | 7.273           |
| 25<sup>a</sup> |    | 7.823          | 7.426           |
| 26       |    | 7.376          | 6.794           |
| 27<sup>b</sup> |    | 7.585          | 7.798           |

Table 1. (Cont.) The structures and biological activities of compounds 28-29

| Compound | Rᵢ | Observed pIC₅₀ | Predicted pIC₅₀ |
|----------|----|----------------|-----------------|
| 28       | H  | 8.154          | 8.248           |
| 29       | Me | 7.119          | 7.285           |

<sup>a</sup>pIC₅₀=-log(IC₅₀)
<sup>b</sup>Compounds selected as test set
Table 2 The result of principal component analysis on the total descriptors.

| Component | Eigenvalues | % of Variance Explained | Cumulative % |
|-----------|-------------|-------------------------|--------------|
| 1         | 470.394     | 39.495                  | 39.495       |
| 2         | 138.563     | 11.634                  | 51.130       |
| 3         | 127.783     | 10.729                  | 61.859       |
| 4         | 79.828      | 6.702                   | 68.561       |
| 5         | 61.826      | 5.191                   | 73.752       |
| 6         | 42.604      | 3.577                   | 77.330       |
| 7         | 36.975      | 3.104                   | 80.434       |
| 8         | 34.61       | 2.906                   | 83.340       |
| 9         | 30.600      | 2.569                   | 85.910       |
| 10        | 23.659      | 1.986                   | 87.896       |
| 11        | 18.673      | 1.567                   | 89.464       |
| 12        | 17.002      | 1.427                   | 90.892       |
| 13        | 14.264      | 1.197                   | 92.089       |
| 14        | 13.344      | 1.120                   | 93.210       |

Figures 1. The first two components (PC1, and PC2) from the principal component analysis of the 29 studied molecules.

PCs with their eigenvalues are shown in the table 2. Therefore, the next steps of study were restricted to these 14 PCs. Plotting of first PC vs. second PC showed none of the compounds is outlier (Fig. 1). Clarification of the theory of the artificial neural networks in details has been adequately described elsewhere (11) and some relevant remarks is presented.

Back propagation artificial neural network includes three layers. The first layer namely input layer has N_I neurons, and function of this layer is reception of information (i.e. inputs) which transfers them to all neurons in the next layer called the hidden layer that their number are indicated by N_H. The neurons in the hidden layer calculate a weighted sum of the inputs that is subsequently transformed by a linear or non-linear function. The last layer is the output layer and its neurons handle the output from the network and it calculate response vector. The function of synapse is connection of input layer to hidden layer and hidden layer to output layer. The manner in which each node transforms its input depends on the “weights” and bias of the node, which are modifiable.

BA-ANN network was trained with the training set of molecules using a back propagation algorithm followed by conjugate gradient descent in the second stage (4-7). RMSECV was then employed as tool to select optimum value of various parameters (6).
Figure 2. Optimization of the number of principal components (PCs) to enter the ANN.

Figure 3. Optimization of number of neurons in hidden layer (A), momentum (B), and, Learning rate (C).
Table 3 Various statistical parameter for developed PC- ANN model

|                | \( R^2 \) | \( \text{RMSE}^a \) | \( \text{PRESS}^b \) | \( R^2 - R_0^2 / R^2 \) | \( R^2 - R_{\text{PRESS}}^2 / R^2 \) | \( K^d \) | \( k' \) | \( R_s^j \) |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Training set   | 0.906          | 0.189          | 0.752          | -0.102         | -0.102         | 1.001          | 0.997          | 0.630          |
| Test Set       | 0.932          | 0.103          | 0.388          | -0.072         | -0.070         | 1.003          | 0.994          | 0.690          |

\( R^2 \) = Square Regression coefficient
\( \text{RMSE} = \text{Root mean square error} \)
\( \text{PRESS} = \text{predicted error sum of square for training set} \)

\( k = \sum y_i ^2 / \sum y_i ^2 
\)

\( k' = \sum y_i ^2 / \sum y_i ^2 

Table 4 Structures and details of the proposed molecules as novel CCR15 inhibitors.

| Compound | R     | Predicted pIC_{50} |
|----------|-------|--------------------|
| S1       |       | 8.112              |
| S2       |       | 8.082              |
| S3       |       | 7.962              |
| S4       |       | 8.004              |

Figure 2 shows the effect of the different number of PCs on predictability of developed model. On the based of this figure, the ANN has the highest degree of predictability when the number of PCs is 4. The parameters of network which should be optimized are learning rate, number of neuron in hidden layer, momentum .The optimal values for these parameters as it is shown in figures 3A-C are 1.2, 9, and 0.9, respectively.

The predicted activity of the ANN calculated values of pIC_{50} versus the experimental values are shown in figure 4 and reported in table 1. As it was expected, the calculated values are in good agreement with the experimental values. Various statistical criteria for ANN model were calculated and reported in table 3 (5-8). The external predictability of a proposed model was generally tested using test sets and \( R^2_{CV} \). The satisfactory
prediction of the values of the inhibitory activity of the test set compounds demonstrates the efficacy of the QSAR in predicting the activities of external molecules.

Discussion

In the developed model, a network including a fully connected three layer, feed-forward ANN model trained with a back-propagation learning algorithm was used. The input of the network was the eigenvalue ranked PCs, the number of them which entered neural network varied from 1 to 14, of which 4 PCs of them were selected as input of networks. By using this number of PCs the best results on the basis of lowest $RMSECV$ in the output of network were obtained. Since there are no exact theoretical principles for choosing the appropriate network topology, before the training of the network, the adjustable parameters as number of nodes in the hidden layer, transfer function, learning rate and etc. were optimized. The values resulting from hidden layer are transferred to the last layer, which contains a single neuron representing the predicted activity. For output layer a linear transfer function was chosen. Various ANN architectures were run with the four selected PCs as input. In each run, the neuron architecture and parameters were optimized to reach the lowest $RMSECV$ as the performances of the resulted models.

According to the criteria proposed by Tropsha and Roy (4-6), for testing the reliability and the robustness of QSAR models, the obtained model is very predictive (Table 3).

As a final point, one could dispute that what does the developed model mean to medicinal chemists? As discussed above, the calculated PCs have meaning physicochemically, but they may be employed for building statistical models which help the medicinal chemist limit the number of compounds to be synthesized. For instance, medicinal chemist can propose a training set comprised of molecules which have the characters of two or more chemical classes with the smallest amount of similarity. Then the model can be used to predict the activity of his proposed molecules. Therefore, the QSAR model was used to estimate inhibitory activities of a few suggested compounds.

The general structures of four suggested compounds and also their calculated activities are reported in table 4. The suggested compounds are combination of the most potent compounds of table 1. The relative high predicted activity of the tested compounds suggest further study such as synthesis of other compounds with such chemical structures.

CONCLUSION

The main objective of this study was to define and establish a QSAR model to predict bioactivity of a series of 3-amino-4-(2-(2-(4-benzylpiperazin-1-yl)-2-oxoethoxy) phenylamino) cyclobutenedione derivatives as novel CCR1 antagonists without any knowledge of the under study system. Various theoretical calculated molecular descriptors were applied to calculate PCs. Calculated PCs were used to make model of the relationship between the molecule structures of the studied compounds and the corresponding bioactivities. The study showed that the calculated PCs as input variable to network can improve the predictive ability of the neural networks. Moreover, the suggested QSAR model was based on nonlinear ANN approach, which can be employed to simulate any kinds of complex correlation or function relationship in a given multivariable system. i.e., ANN approach is more appropriate for modeling where no clearly defined mathematical model for a system is available. Bioactivity is one of the most important properties for a given compound. Therefore, accurate, well-organized and intelligent QSAR model for the bioactivity will be influential for drug design and development.
REFERENCES

1. Schall T, The chemokines. In: Thompson A, ed. The Cytokine Handbook. San Diego, Academic Press; 1994; 419-460.
2. Xie YF, Sircar I, Lake K, Komandla M, Ligsay K, Li J, Xu K, Parise J, Schneider L, Huang D, Liu J, Sakurai N, Barbossa M, Jack R. Identification of novel series of human CCR1 antagonists. Bioorg Med Chem Lett, 2008; 18:2215-2221.
3. Liang M, Rosser M, Ng H, May K, Bauman J, Islam I, Ghannam A, Kretschmer P, Pu H, Dunning L, Snider R, Morrissey M, Hesselgesser J, Perez H, Horuk R. Species selectivity of a small molecule antagonist for the CCR1 chemokine. Eur J Pharmacol, 2000; 389:41-49.
4. Saghaie L, Shahlai M, Fasshi A, Madadkar-Sobhani A, Ghollivand M, Pourhossein A. QSAR Analysis for Some Diaryl-substituted Pyrazoles as CCR2 Inhibitors by GA-Stepwise MLR. Chem Biol Drug Des, 2011; 77:75-85.
5. Arkan E, Shahlai M, Pourhossein A, Fakhri K, Fasshi A. Validated QSAR analysis of some diaryl substituted pyrazoles as CCR2 inhibitors by various linear and nonlinear multivariate chemometrics methods. Eur J Med Chem, 2010; 45:3394-3406.
6. Shahlai M, Sabet R, Ziari MB, Moemifard B, Fasshi A, Karbakhsh R. QSAR study of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2 using LS-SVM and GRNN based on principal components. Eur J Med Chem, 2010; 45:4499-4508.
7. Xie YF, Lake K, Ligsay K, Komandla M, Sircar I, Nagarajan G, Li J, Xu K, Parise J, Schneider L, Huang D, Liu J, Dines K, Sakurai N, Barbossa M, Jack R. Structure–activity relationships of novel, highly potent selective, and orally active CCR1 antagonists. Bioorg Med Chem Lett, 2007; 17:3367–3372.
8. Hyperchem. Hyperchem, Molecular Modeling System. In: Developed by Hyper Cube I, ed.: Hyper Cube, Inc. and Auto Desk, Inc.
9. Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON software 2.1 ed. Milano, Italy; 2002.
10. Kennard R, Stone L. Computer Aided Design of Experiments. Technometrics 1969; 11:137-148.
11. Haykin S. Neural Network. Englewood Cliffs, NJ: Prentice Hall; 1994.