European Journal of Medicinal Chemistry 45 (2010) 49–54

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

Quantitative structure-activity relationships studies of CCR5 inhibitors and toxicity of aromatic compounds using gene expression programming

Weimin Shi, Xiaoya Zhang, Qi Shen*

Department of Chemistry, Zhengzhou University, Zhengzhou 450052, China

ARTICLE INFO

Article history:
Received 23 February 2009
Received in revised form
27 July 2009
Accepted 10 September 2009
Available online 16 September 2009

Keywords:
Quantitative structure-activity relationship
Gene expression programming
CCR5 inhibitor
Aromatic compounds

ABSTRACT

Quantitative structure-activity relationship (QSAR) study of chemokine receptor 5 (CCR5) binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas and toxicity of aromatic compounds have been performed. The gene expression programming (GEP) was used to select variables and produce nonlinear QSAR models simultaneously using the selected variables. In our GEP implementation, a simple and convenient method was proposed to infer the K-expression from the number of arguments of the function in a gene, without building the expression tree. The results were compared to those obtained by artificial neural network (ANN) and support vector machine (SVM). It has been demonstrated that the GEP is a useful tool for QSAR modeling.

1. Introduction

Acquired immunodeficiency syndrome is a set of symptoms and infections resulting from the damage to the human immune system which caused by the human immunodeficiency virus (HIV), and has become a major worldwide epidemic [1]. In the cell-entry stage of HIV replication, a protein called gp120 on the envelope of the virus binds to CD4, a protein found on the surface of some white blood cells. The CD4 protein acts as a receptor for gp120, "unlocking" the cell and allowing the virus to enter. In addition to CD4, HIV also needs a co-receptor in order to enter the host cells: CCR5 and CXCR4 [2,3]. CCR5 is a chemokine receptor present in different cells, especially in T cells, macrophages, monocytes, dendritic cells and microglia. Several chemokine receptors can function as viral coreceptors, but CCR5 is likely the most physiologically important coreceptor during natural infection. Moreover, at least half of all infected individuals harbor only CCR5-using viruses throughout the course of infection. The fact that CCR5 is the most commonly used coreceptor by HIV in early infection makes it a highly attractive target for antiretroviral therapy. A number of new experimental HIV drugs, called entry inhibitors, have been designed to interfere with the interaction between CCR5 and HIV [4,5]. However, many challenges have hindered the development of a viable CCR5 inhibitor, including potential side effects, toxicities, drug interactions, and resistance. One could not, however, confirm that the compounds designed would always possess good binding affinity to CCR5, while the synthesis and testing of these compounds on CCR5 coreceptor are time-consuming and expensive. Consequently, it is of interest to develop a prediction method for biological activities before the synthesis. Quantitative structure-activity relationship models have been built using the experimental data accumulated. Using such an approach one could predict the activities of newly designed compounds before a decision is being made whether these compounds should be really synthesized and tested

Various modeling techniques have been widely used in QSAR studying, such as multiple linear regression (MLR) [6], partial least squares (PLS) [6], artificial neural networks (ANN) [7,8] and support vector machine (SVM) [9,10]. The ANN and SVM can incorporate nonlinear relationships between descriptors and activity and often produce superior QSAR models compared to models derived by the more traditional approach MLR and PLS. However, artificial neural networks also have disadvantages. Disadvantages include its "black box" nature, greater computational burden, proneness to over-fitting and the empirical nature of model development. ANN does not give explicit knowledge representation in the form of rules, or some other easily interpretable form. The model is implicit, hidden in the network structure and optimized weights between the nodes. SVM has been found useful in handling QSAR tasks in case of the high dimensionality and has been recommended as a popular approach to efficiently treating this particular data structure. However, there is increasing evidence that variable selection is also essential for successful SVM analysis and the lack of variable selection can also spoil the SVM performance.

* Corresponding author. Tel.: +86 371 67781024; fax: +86 371 67767200.
E-mail address: shenqi@zzu.edu.cn (Q. Shen).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.09.022

Gene expression programming (GEP) [11,12], a relatively new evolutionary algorithm, can also be used as an excellent data mining and modeling technique. GEP invented by Ferreira in 1999 was developed from genetic algorithms (GA) and genetic programming (GP). The one main difference between the three algorithms resides in the nature of the individuals. The individuals in GA are linear strings of fixed length (chromosomes) and the individuals in GP are nonlinear entities of different sizes and shapes. However GEP stores the individuals as linear chromosomes of fixed length which are then presented as expression trees with different sizes and shapes for evaluation. Compared to GA and GP, the advantages of GEP are that the chromosomes are relatively small and easy to manipulate. The successful combination of two methods based on different entities enables GEP to enjoy both the simplicity of GA and the flexibility of GP. GEP is gaining attention due to its ability to discover the underlying data relationships and express them mathematically. GEP has been successfully applied in regression, optimization and classification [13–15]. Several QSAR works using GEP have been published [16]. However, in these works variable selection was separated from QSAR modeling by GEP. Variables were selected by some other methods followed by use in GEP for QSAR modeling. As examples of application of the GEP algorithm, CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas [17–19] and toxicity of aromatic compounds to chlorella vulgaris[20] were predicted by GEP. In our GEP implementation, a simple and convenient method is proposed to infer the K-expression from the number of arguments of the function in a gene, without building the expression tree. The GEP was used to select variables and produce nonlinear QSAR models simultaneously using the selected variables. The variables selection and modeling were incorporated in GEP and nonlinear models were expressed mathematically. The results were compared to those obtained by ANN and SVM. It has been demonstrated that the GEP is a useful tool for QSAR modeling.

2. Materials and methods

2.1. CCR5 receptor binding affinity data

A set of 79 substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas as CCR5 receptor, whose binding affinity are reported by Leonard et al [17], was used to test the performance of the GEP in QSAR. Molecular structure of 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas are represented in Fig. 2. The affinity for I-labeled RANTES (regulated on activation normal T-cell expressed and secreted) to Chinese hamster ovary (CHO) cells expressing human CCR5 were expressed as IC_{50} and have been converted to logarithmic scale [pIC_{50}(mM)]. The data set of 79 substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas was randomly divided into two groups with 59 compounds used as training set for developing regression models and remaining 20 compounds used as the validation set in the prediction of CCR5 receptor binding affinity.

A total of 147 molecular descriptors were calculated for substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas including structural, spatial, thermodynamic, electronic, quantum mechanical descriptors, and E-State indices. Structural descriptors include the molecular weight (MW), the number of rotatable bonds (Rotbonds) and the number of hydrogen bond (Hbond acceptor). The spatial descriptors [21,22] used involve radius of gyration (RadOfGyration), density, molecular surface area, principal moment of inertia (PMI), molecular volume, and shadow indices. The thermodynamic descriptors [23] were taken describing the hydrophobic character (logP: logarithm of the partition coefficient in octano/water. atom-type-based AlogP descriptors: log of the partition coefficient atom type value), refractivity (MolRef: molar refractivity), heat of formation, Hf) and the dissolution free energy for water and octanol (Fh2o: desolvation free energy for H_2O; Foct: desolvation free energy for octanol). The electronic descriptors taken were concerning surperdelocalizability (Sr), atomic polarizabilities (Apol), and the dipole moment (Dipole). Electrotopological-state indices [24,25] (E-State indices) used involve S-aasC, S-aaN, S-aaCH etc. All these molecular descriptors were generated using Cerius^{2}3.5 software system on a Silicon Graphics R3000 workstation. Besides the aforementioned nearly calculated molecular descriptors, 18 variables used by Leonard et al [17] were also included in the list of the candidate variables.

The descriptor analysis involves the detection and removal of those structural descriptors which exhibit high pair-wise correlations with other descriptors, or which contain little discriminatory information. Pairs of descriptors that are highly correlated ($r \geq 0.950$) encoded similar information, and one of them should be removed. Descriptors that contain a high percentage ($\geq$90%) of identical values are also discarded. Thus, only 66 of total descriptors were remained. All these 66 molecular descriptors were generated using Cerius^{2}3.5 soft system on Silicon Graphics R3000 workstation and were scaled into (0, 1.0) for analysis.

2.2. Toxicity data of aromatic compounds

A total of 65 aromatic chemicals representing several mechanisms of toxic action, whose acute aquatic toxicity data were determined by Netzeva et al [20], was considered in this study. The data set is chemically heterogeneous and includes phenols, anilines, nitrobenzenes, and benzaldehydes as well as compounds with more than one functional group on the benzene ring. The toxicity was expressed as $\log(1/EC_{50})$, which were determined in a biochemical assay utilizing the alga Chlorella ulgaris in the logarithmic phase of their growth cycle were used. Among 65 aromatic chemicals, 50 randomly selected samples were used as training set and the remaining 15 samples as the prediction set.

Each sample is described by 38 molecular descriptors for QSAR modeling, including structural, spatial, thermodynamic, electronic, quantum mechanical descriptors, and E-State indices. Seven variables used by Netzeva et al [20] were also included in the list of the candidate variables. All these 38 molecular descriptors were scaled into (0, 1.0) for analysis.

2.3. Theory of gene expression programming

The GEP originated in 1999 by Ferreira [11,12] and is based on the ideas of GA and GP. Similar to GA and GP, GEP uses populations of individuals, selects them according to fitness, and introduces genetic variation using one or more genetic operators. However, unlike GA and GP, the individuals in GEP are encoded as linear strings of fixed length (the genome or chromosomes) which are afterwards expressed as nonlinear entities of different sizes and shapes (expression trees). The theory of GEP is composed of representation of the candidate solution, genetic operators and the definition of fitness function.

2.3.1. The gene and chromosome of GEP

The chromosome of GEP is composed of one or more genes which are expressed as a linear, symbolic string of some fixed size. Each gene is divided into two parts, a head and a tail. The head is of a predetermined length (h) and contains symbols for both functions and terminals (variables and constants). The tail can only contain terminals and its length (t) is computed as

$$t = h(n-1) \quad (1)$$