Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 90 (2008) 15–24
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Quantitative structure activity relationship study on EC50 of anti-HIV drugs

Hongzong Si [a,*], Shuping Yuan [a,1], Kejun Zhang [b], Aiping Fu [a,1], Yun-Bo Duan [a,1], Zhide Hu [c]

[a] *Institute for Computational Science and Engineering, Laboratory of New Fibrous Materials and Modern Textile, the Growing Base for State Key Laboratory, Qingdao University, Qingdao, Shandong 266071, China*
[b] *Department of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China*
[c] *Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, China*

Received 8 March 2007; received in revised form 6 June 2007; accepted 26 June 2007
Available online 7 July 2007

Abstract

A quantitative model was developed to predict the EC50 of nucleoside by the gene expression programming (GEP). Each kind of compound was represented by several calculated structural descriptors involving constitutional, topological, geometrical, electrostatic and quantum-chemical features of the compound. The GEP method produced a nonlinear quantitative model of the five-descriptor with a correlation coefficient and a mean error of 0.91 and 0.41 for the training set, 0.63 and 0.67 for the test set, respectively. It is shown that the GEP predicted results are in good agreement with experimental ones, better than those of the support vector machine.

Keywords: Gene expression programming (GEP); Quantitative structure activity relationship (QSAR); Support vector machine (SVM); Heuristic method; Human immunodeficiency virus (HIV); Nucleoside

1. Introduction

The acquired immunodeficiency syndrome (AIDS) which is caused by the human immunodeficiency virus (HIV) [1,2] has become a serious global threat to human health and life. The progress in the HIV biology has provided detailed knowledge of molecular events in the replication cycle of the HIV lymphotropic virus-I (HIV-I). Those kinds of knowledge are required to develop effective antiviral agents and strategies to eliminate the HIV replication. Current understanding of molecular events in the HIV life cycle proposes seven steps: viral entry, reverse transcription, integration, gene expression, assembly, budding, and maturation. From the viewpoint of the theory, every stage in the viral life cycle may be served as a potential target for designing anti-HIV agents and therapies [3]. The MT4 cell, a human T cell, carrying human T cell lymphotropic virus-I (HTLV-I), is easily infected by the HIV-I. The HIV-I reverse transcriptase (HIV-I RT) is an attractive target for the drug therapy of AIDS because it is essential for the HIV replication while not required for normal host cell replication. This multifunctional enzyme can transcript the RNA genome of HIV-I into DNA, which is subsequently integrated into the host cell's genome. The nucleoside reverse transcriptase inhibitors (NRTIs) are specific in against HIV-I and do not inhibit host cell polymerases. Furthermore, these drugs have low cytotoxicity and few side effects. NRTIs that lead to termination of DNA synthesis during the HIV-I replication comprise a major class of clinically available antiretroviral drugs [4]. The level of serum concentration of nucleoside plays an important role for effective replication of HIV-I. The EC50 is the concentration to achieve 50% of maximum effect and is a good index to evaluate effects of drugs.

To improve the quality of the 'research' compounds, efficient methods are required in early drug discovery to search key factors influencing drug serum concentration and to understand the mechanisms responsible for drug EC50. At the same time, with the growth of combinatorial chemistry methods in drug discovery, a large number of candidate compounds are synthesized and screened in parallel for in vitro pharmacological activity, which has dramatically increased the demand for efficient models to predict EC50 [5].

The quantitative structure-property/activity relationship (QSPR/QSAR) method is based on the assumption that the variation of the behavior of the compounds, as expressed by

* Corresponding author. Tel.: +86 532 85950786; fax: +86 532 85950768.
E-mail address: sihz03@126.com (H. Si).
[1] Tel.: +86 532 85950786; fax: +86 532 85950768.

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.06.011

many measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors [6]. This method can be used for the prediction of the properties of new compounds. It can also be applied to identify and describe important structural features of the molecules that are relevant to variations in molecular properties. Computational models are useful because they rationalize a large number of experimental observations and therefore save times and money in the process of drug design. In addition, they are useful in areas of the design of virtual compound libraries, the computational optimization of compounds, and the design of combinatorial libraries with appropriate absorption, distribution, metabolism and excretion properties. It is generally assumed that the physicochemical descriptors of drug molecules are useful for predicting EC50. Consequently, the QSPR/QSAR has been successfully established to predict EC50 [7].

To increase the accuracy, artificial intelligence techniques have been applied to the QSPR/QSAR analysis since the late 1980s [8,9]. An appropriate method plays a key role in building the QSPR/QSAR model. The GEP [10] is a genotype/phenotype system that involves computational programs with different sizes and shapes of expression trees encoded in linear chromosomes of the fixed length. The genetic encoding used in the GEP allows a totally unconstrained interplay between chromosomes and expression trees.

In the GEP, the implementation of different genetic operators is extremely simple because of the existence of a truly functional and autonomous genome. Systems with different evolutionary behaviors can be easily simulated in the GEP. Therefore, the GEP has been successfully used to predict the evaporation estimation [11] and the cement strength [12].

In the present work, the GEP is generalized to set up the anti-HIV drugs QSAR model based on the descriptors of the diverse data set. These descriptors are calculated from the molecular structures alone by the software CODESSA. Then five descriptors are selected as inputs by the heuristic method (HM). In order to investigate the influence of different descriptors on EC50, the HM is used to build several multivariable linear models. Based on this, we developed a new QSAR model to explore the EC50 of the drugs with diverse structures. The GEP is compared with another powerful non-linear method of SVM. It is shown that the GEP predicted results are better than those of SVM in both training set and test set. To our knowledge, it is the first time that the QSAR method is used for predicting the EC50 of NRTIs on the basis of the molecular structural descriptors.

2. Theories and methods

2.1. Data set

48 drugs of NRTIs are taken from the database of http://www.niaid.nih.gov/ and are listed in Table 1. The data set is randomly separated into a training set of 36 compounds and a test set of 12 compounds. The training set is used to build the model and the test set is employed to evaluate the prediction ability of the model. The leave-one-out (LOO) cross-validation is performed for the whole training set.

2.2. Calculation of the descriptors

All of the molecules are drawn into Hyperchem [13] and pre-optimized using the MM+ molecular mechanics force field. Then a more precise optimization is done with the semi-empirical AM1 method in MOPAC [14]. The molecular structures are optimized using the Polak–Ribiere algorithm until the root mean square gradient reaches 0.01. The MOPAC output files are introduced to CODESSA program to calculate five classes of the descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), topological (Wiener index, Randic indices, Kier–Hall shape indices, etc.), geometrical (moments of inertia, molecular volume, molecular surface area, etc.), electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.), and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.) [15].

2.3. Development of linear model by HM [16]

Once the molecular descriptors are generated, the HM in CODESSA is used to pre-select the descriptors and build the linear model. The advantages of the HM are the high speed and no software restrictions on the size of the data set. The HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. The details of selecting descriptors are as follows: First of all, all descriptors are checked to ensure that values of each descriptor are available for each structure. Descriptors for which values are not available for every structure in the data are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and the insignificant descriptors are removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. The details of validating intercorrelation are (a) all quasi-orthogonal pairs of structural descriptors are selected from the initial set. Two descriptors are considered orthogonal if their intercorrelation coefficient r_{ij} is lower than 0.1; (b) CODESSA uses the pairs of orthogonal descriptors to compute the biparametric regression equations; (c) to an MLR model containing n descriptors, a new descriptor is added to generate a model with $n+1$ descriptors if the new descriptor is not significantly correlated with the previous n descriptors; step (c) is repeated until MLR models with a prescribed number of descriptors are obtained. The goodness of the correlation is tested by the square of coefficient regression (R^2), square of cross-validate coefficient regression (R^2_{CV}), the F-test (F), and the standard deviation (s^2). From the above processes, five descriptors are selected from descriptors pool and the linear model is produced by the HM. The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality [17].

However, the factors influencing on the EC50 of drugs are complex and not all of them are in linear correlation with the