European Journal of Pharmaceutical Sciences 51 (2014) 189–195

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

journal homepage: www.elsevier.com/locate/ejps

Binary classification of chalcone derivatives with LDA or KNN based on their antileishmanial activity and molecular descriptors selected using the Successive Projections Algorithm feature-selection technique

Mohammad Goodarzi [a,*], Wouter Saeys [a], Mario Cesar Ugulino de Araujo [b], Roberto Kawakami Harrop Galvão [c], Yvan Vander Heyden [d]

[a] Department of Biosystems, Faculty of Bioscience Engineering, Katholieke Universiteit Leuven – KULeuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium
[b] Departamento de Quimica, Universidade Federal da Paraiba, CCEN, Caixa Postal 5093, CEP 58051-970 Joao Pessoa, PB, Brazil
[c] Divisao de Engenharia Eletronica, Instituto Tecnologico de Aeronautica, CEP 12228-900 Sao Jose dos Campos, SP, Brazil
[d] Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, 1090 Brussels, Belgium

ARTICLE INFO

Article history:
Received 28 May 2013
Received in revised form 14 September 2013
Accepted 20 September 2013
Available online 30 September 2013

Keywords:
Antileishmanial activity
Successive Projections Algorithm
Linear Discriminant Analysis
Genetic Algorithm
One Nearest Neighbour

ABSTRACT

Chalcones are naturally occurring aromatic ketones, which consist of an α-, β-unsaturated carbonyl system joining two aryl rings. These compounds are reported to exhibit several pharmacological activities, including antiparasitic, antibacterial, antifungal, anticancer, immunomodulatory, nitric oxide inhibition and anti-inflammatory effects. In the present work, a Quantitative Structure–Activity Relationship (QSAR) study is carried out to classify chalcone derivatives with respect to their antileishmanial activity (active/inactive) on the basis of molecular descriptors. For this purpose, two techniques to select descriptors are employed, the Successive Projections Algorithm (SPA) and the Genetic Algorithm (GA). The selected descriptors are initially employed to build Linear Discriminant Analysis (LDA) models. An additional investigation is then carried out to determine whether the results can be improved by using a non-parametric classification technique (One Nearest Neighbour, 1NN). In a case study involving 100 chalcone derivatives, the 1NN models were found to provide better rates of correct classification than LDA, both in the training and test sets. The best result was achieved by a SPA–1NN model with six molecular descriptors, which provided correct classification rates of 97% and 84% for the training and test sets, respectively.

1. Introduction

Chalcones are naturally occurring compounds, which are considered as precursors for the flavonoid synthesis in several plant species. Chemically, chalcones are aromatic ketones consisting of an α-, β-unsaturated carbonyl system joining two aryl rings (Cianci et al., 2008). These compounds are reported to exhibit several pharmacological activities, including antiparasitic (Nielsen et al., 1998; Li et al., 1995), antibacterial (Lin et al., 2002), antifungal (Lopez et al., 2001), anticancer (Pouget et al., 2001; Ducki et al., 2005, 1998), immunomodulatory (Barford et al., 2002), nitric oxide inhibition (Rojas et al., 2002) and anti-inflammatory (Ballesteros et al., 1995; Hsieh et al., 1998) properties. A number of studies has been particularly concerned with antileishmanial activity (Kayser and Kiderlen, 2001; Zhai et al., 1999; Chen et al., 2001).

Leishmaniasis is a widespread parasitic disease caused by protozoan parasites of the genus Leishmania in tropical and subtropical areas around the world. The parasite is transmitted to humans by the insect vector phleobotomine sandfly. Leishmaniasis has been divided into cutaneous (CL), mucocutaneous (MCL) and visceral leishmaniasis or Kala-azar (VL), which can be fatal when untreated (Monzote, 2009). *Leishmania donovani* and *Leishmania infantum* are major agents of VL, while the species L. *major*, *tropica*, *aethiopica*, *braziliensis*, *panamensis*, *amazonensis* and *mexicana* cause CL (Croft and Coombs, 2003).

Leishmaniasis has a considerable influence on the global public health and is endemic in many tropical and subtropical regions. According to World Health Organization (WHO) reports (Croft and Coombs, 2003), not less than 550 million individuals in 88 countries, including some southern European countries, are at risk of infection. The disease affects around 12 million people worldwide with approximately 100,000 deaths yearly (Croft and Coombs, 2003).

The therapy for leishmaniasis still poses serious problems. The first-choice drugs are pentavalent antimonial compounds, which

* Corresponding author. Tel.: +32 16321470; fax: +32 16328590.
E-mail addresses: mohammad.godarzi@gmail.com, Mohammad.Goodarzi@biw.kuleuven.be (M. Goodarzi).

0928-0987/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejps.2013.09.019

were developed before 1960 (Neal et al., 1987) and, in general, require long-term treatment and have severe side effects (Monzote, 2009). The development of drug resistance by the pathogens, especially in HIV–Leishmania co-infected patients, has also aggravated the health problem (Ali, 2002). Therefore, there is an urgent need for the development of new, efficient and safe drugs against leishmaniasis. In this context, valuable information may be obtained by Quantitative Structure–Activity/Property Relationship (QSAR/QSPR) investigations.

QSAR/QSPR studies consist of relating the biological activities of a series of compounds with appropriate molecular descriptors (Goodarzi et al., 2009a,b; Lin et al., 2005; Todeschini et al., 2009; Lei et al., 2009; Caetano et al., 2007). Such relationships may be used to predict the activity/property of new compounds and to design virtual compound libraries. However, the empirical development of a QSAR/QSPR model with good accuracy may be a challenging task. In this context, the selection of appropriate descriptors for use in the model plays a critical role.

The present work is concerned with the development of classification models for the antileishmanial activity of chalcones. More specifically, we compare the performance of two techniques for selection of molecular descriptors, i.e. the Successive Projections Algorithm (SPA) (Pontes et al., 2009, 2005; Gambarra Neto et al., 2009; Moreira et al., 2009) and the Genetic Algorithm (GA) (Galvão and Araújo, 2009; Goldberg, 1989; Jouan-Rimbaud et al., 1995; Leardi, 2001). These selections are initially employed to build Linear Discriminant Analysis (LDA) models. An additional investigation is then carried out to determine whether the results can be improved by using a non-parametric classification technique (One Nearest Neighbour, 1NN) (Jones, 2008; Cover and Hart, 1967; Young, 2001). To the best of our knowledge, this is the first published report that introduces Successive Projection Algorithm (SPA) as methodology for descriptor selection in classification QSAR studies.

2. Experimental

2.1. Data set and descriptors generation

The 2D structures of the molecules were drawn using HyperChem 7 (2007) software (Hypercube, Gainesville, Florida, United States). The final geometries were obtained with the semi-empirical AM1 method in the Hyperchem program. All calculations were carried out at the restricted Hartree–Fock level with no configuration interaction (Young, 2001). The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean squared gradient reached 0.001 kcal mol^{-1}. The resulting geometry was transferred into the Dragon program package (Talete srl, DRAGON for Windows – Software for molecular descriptors calculation, Milano, Italy, 2007) in order to obtain descriptors belonging to Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D molecular Representation of Structure based on Electron diffraction), Molecular Walk Count, BCUT, 2D-Autocorrelation, Aromaticity Index, Randic molecular profile, Radial Distribution Function, Functional group and Atom-Centered Fragment classes (Mauri et al., 2006; Duchowicz et al., 2008; Consonni and Todeschini, 2000).

The calculated descriptors were initially analyzed in order to remove those with constant or near constant values. Moreover, in order to decrease the redundancy in the descriptor data matrix, the correlation between each pair of descriptors was used to detect the presence of collinearity (i.e. $r > 0.9$). Among a collinear descriptor pair, the one presenting the highest correlation with the activity was retained and the other removed from the data matrix.

The experimental values were taken from the literature (Liu et al., 2003). Antimalarial activities of chalcones had been reported in earlier works (with the exception of the 2′-hydroxychalcones), but the antileishmanial activities were reported for the first time in (Liu et al., 2003). The compounds were divided into two classes, according to the antileishmanial activity parameter ED_{50} (μM), as inactive ($ED_{50} > 30$ μM) or active ($ED_{50} < 30$ μM), as was done also in (Liu et al., 2003). Table 1 shows the general structure of the chalcone derivatives with their classified activities.

2.2. Descriptor selection and compound classification

All programs for descriptor selection and compound classification were written by the authors in Matlab software (Matlab Version 7.6; 2008; MathWorks, Natick, MA).

The Successive Projections Algorithm (SPA) is a feature selection method that can be used to minimize multicollinearity problems in Linear Discriminant Analysis (LDA) (Pontes et al., 2009; Gambarra Neto et al., 2009; Moreira et al., 2009). SPA comprises two phases. The first consists of projection operations carried out on the matrix of descriptor values. Such projections are used to generate subsets of descriptors with few multicollinearity. In the second phase, the best subset is selected in order to minimize a cost function associated to the average risk of misclassification for a given validation set. Such a cost is calculated by comparing the Mahalanobis distance (De Maesschalck et al., 2000) of the molecules with respect to their true class, as well as to the closest wrong class (Pontes et al., 2009).

For comparison reasons, a Genetic Algorithm (GA) to select appropriate descriptors was also employed. A standard GA formulation using binary chromosomes was adopted (Galvão and Araújo, 2009; Goldberg, 1989; Jouan-Rimbaud et al., 1995; Leardi, 2001). Each descriptor was associated to a position (gene) in the chromosome. The gene values can be either 1 or 0 (i.e. descriptor is or is not included in the classification model, respectively). The fitness value for each chromosome was calculated as the inverse of the average risk of misclassification, defined as in SPA–LDA. The probability of a given individual being selected for the mating pool was proportional to its fitness (roulette method). One-point crossover and mutation operators were employed with probabilities of 60% and 10%, respectively. The population size was kept constant, with each generation being completely replaced by its descendants. However, the best individual was automatically transferred to the next generation (elitism) to avoid the loss of good solutions. The GA was carried out for 100 generations with 1099 chromosomes each.

The Kennard–Stone (KS) uniform sampling algorithm (Kennard and Stone, 1969) was applied to divide the data into training, validation and test sets with 50, 25 and 25 molecules, respectively. The validation set is employed to guide the selection of variables in SPA–LDA and GA–LDA and the test set in the performance assessment of the resulting LDA models. The molecules applied in the validation and test sets are indicated in Table 1.

Noted that LDA is a parametric classification technique, which implicitly assumes that the objects belonging to each class follow a Gaussian distribution. The training set is employed to estimate the mean vector of each class, as well as a pooled covariance matrix. The resulting decision surfaces are hyperplanes in the feature space (Wu et al., 1996).

The selected descriptors were also employed for classification using the One-Nearest-Neighbour (1NN) rule, a simple non-parametric technique (Jones, 2008; Cover and Hart, 1967). In this technique, no *a priori* assumption regarding the type of probability distribution is made. Each object under analysis is assigned to the class corresponding to its closest neighbour in the training set