A Topological Substructural Approach for the Prediction of P-Glycoprotein Substrates

MIGUEL ANGEL CABRERA,[1] ISABEL GONZÁLEZ,[2] CARLOS FERNÁNDEZ,[2] CARMEN NAVARRO,[2] MARIVAL BERMEJO[2]

[1]Department of Drug Design, Centre of Chemical Bioactive, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba

[2]Department of Pharmacy and Pharmaceutical Technology, University of Valencia, Burjassot 46100, Valencia, Spain

Received 17 January 2005; revised 14 June 2005; accepted 16 June 2005

Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/jps.20449

ABSTRACT: A topological substructural molecular design approach (TOPS-MODE) has been used to predict whether a given compound is a P-glycoprotein (P-gp) substrate or not. A linear discriminant model was developed to classify a data set of 163 compounds as substrates or nonsubstrates (91 substrates and 72 nonsubstrates). The final model fit the data with sensitivity of 82.42% and specificity of 79.17%, for a final accuracy of 80.98%. The model was validated through the use of an external validation set (40 compounds, 22 substrates and 18 nonsubstrates) with a 77.50% of prediction accuracy; fivefold full cross-validation (removing 40 compounds in each cycle, 80.50% of good prediction) and the prediction of an external test set of marketed drugs (35 compounds, 71.43% of good prediction). This methodology evidenced that the standard bond distance, the polarizability and the Gasteiger–Marsilli atomic charge affect the interaction with the P-gp; suggesting the capacity of the TOPS-MODE descriptors to estimate the P-gp substrates for new drug candidates. The potentiality of the TOPS-MODE approach was assessed with a family of compounds not covered by the original training set (6-fluoroquinolones), and the final prediction had a 77.7% of accuracy. Finally, the positive and negative substructural contributions to the classification of 6-fluoroquinolones, as P-gp substrates, were identified; evidencing the possibilities of the present approach in the lead generation and optimization processes. J Pharm Sci 95:589–606, 2006

Keywords: TOPS-MODE; P-glycoprotein; P-gp; molecular modeling; *in silico*; QSAR

INTRODUCTION

P-glycoprotein (P-gp), the product of the multidrug resistance (MDR) gene, is an ATP-dependent efflux transporter that affects the absorption, distribution and excretion of some clinically important drugs.[1] Overexpression of this protein may result in MDR and is a major cause of the failure of cancer chemotheraphy, and decrease efficacy of antibiotics and antiviral agents.[2] Because of the broad impact of this drug efflux transporter on *in vivo* disposition and pharmacokinetics, identification of compounds that are P-gp substrates can aid the optimization and the selection of new drug candidates.[3]

Several *in vitro* screening assays such as the monolayer efflux, ATPase activity, and calcein-AM fluorescence assays have been suggested to classify compounds as P-gp substrates.[2–5] Each of these assays provides different information and has advantages and disadvantages.[6] Nevertheless, methods that increase (low cost and fast-speed methodologies) the identification of P-gp substrates are therefore useful during early drug development.

Many "*in silico*" approach have been directed to develop of computational methods for the P-gp

Correspondence to: Miguel Angel Cabrera (Telephone: 53-42-281473, 53-42-281192; Fax: 53-42-281130; E-mail: macabrera@cbq.uclv.edu.cu; migue@gammu.com)

Journal of Pharmaceutical Sciences, Vol. 95, 589–606 (2006)

substrate prediction, considering that the molecular mechanism of the P-gp mediated transport is not well understood and a high-resolution structure of P-gp is still unavailable.[7] Although different physicochemical properties such as lipophilicity, hydrogen bonding ability, molecular weight, and surface area have been used to find a relationship between these features and the biological activity,[5,6,8,9] more complex descriptors and powerful methods of molecular modeling or multivariate analysis are necessary.[7,10]

Among the molecular descriptors described in the literature, one with a more wide application has been the topological descriptor. From the topological representation, the graph-theoretical methods have become one of the most important tools for the mathematical description of the molecular structure. The TOPS-MODE approach,[11] based on the calculation of the spectral moments of the bond matrix of molecular graph, has been used to generate graph-theoretical descriptors, expressing physical and biological properties in terms of substructural features of molecules. This approach has been successfully applied to different QSPR studies.[12–17]

Taking into consideration the above mentioned, this stydy describes a quantitative structure-activity relationship (QSAR) analysis on a dataset of 163 compounds assayed for P-gp activity. The TOPS-MODE approach was used in the generation of a discriminant function by a Linear Discriminant Analysis (LDA) that permits the classification of drugs (P-gp substrate and nonsubstrate). Furtherly, the "*in silico*" model was assessed by the use of different validation tests and their predictive potentiality was evaluated with a family of compounds not covered by the original training set (6-fluoroquinolones). Finally, the positive and negative substructural contributions to the classification of 6-fluoroquinolones, as P-gp substrates, were identified; demonstrating the role of the present approach in the lead generation and optimization processes.

EXPERIMENTAL SECTION

TOPS-MODE Approach

The TOPS-MODE approach is based on the calculation of the spectral moments of the so-called bond matrix,[11] whose theoretical basis has been described in previous reports.[18–22] Nevertheless, an overview of this approach will be given below.

The bond matrix is defined as a square and symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, being two bonds adjacent if they are incident to a common atom. The spectral moments of the edge adjacency matrix are defined as the traces, that is, the sum of the main diagonal, of the different powers of such a matrix. In order to apply the present approach to the structure-property-relationship, the following steps should be followed:

1. To select an adequate training set with great structural diversity.
2. To draw the hydrogen depleted molecular graphs for each molecule of the training set.
3. To differentiate the molecular bonds with appropriate weights.
4. To compute the spectral moments of the bond matrix for each molecule of the data set.
5. To find a quantitative structure-property relationship by using a regression analysis:

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + \ldots\ldots + a_k\mu_k + b \quad (1)$$

   where, P is the studied property, in our case, the P-gp activity (substrate or nonsubstrate), μ_k is the kth spectral moment, and the a_k's are the coefficients obtained by the linear regression.
6. To test the predictive capacity of the computational model by cross-validation procedures and an external prediction sets.
7. Finally, to compute the contribution of the different substructures in order to determine their quantitative contribution to the P-gp activity of the studied molecules.

Although, it is known that other readily interpretable descriptors have been used to predict if a compound is or not a P-gp substrate, in the present study are employed the spectral moment. The election is related with the advantages of this approach such as: it gives the possibility of finding structural contributions to the property studied in an explicit way, for example the contributions can be found for groups and fragments present in molecules for which experimental values of the corresponding properties are not available; also the spectral moments can be obtained weighting the molecular bonds with physicochemical parameters and thus, the TOPS-MODE descriptors measure the degree of concentration of these