Combined 4D-Fingerprint and Clustering Based Membrane-Interaction QSAR Analyses for Constructing Consensus Caco-2 Cell Permeation Virtual Screens

OSVALDO A. SANTOS-FILHO,[1] ANTON J. HOPFINGER[2,3]

[1]Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, 2733 Heather Street, Vancouver, British Columbia, Canada V5Z 3J5

[2]College of Pharmacy, MSC09 5360, 1 University of New Mexico, Albuquerque, New Mexico 87191-0001

[3]The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045

Received 5 October 2006; revised 26 December 2006; accepted 1 January 2007

Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/jps.21086

ABSTRACT: A set of 30 structurally diverse molecules, for which Caco-2 cell permeation coefficients were determined, formed the training set for construction of Caco-2 cell permeation models based upon membrane-interaction (MI) QSAR analysis and a new QSAR method called 4D-fingerprint QSAR analysis. The descriptor terms of the 4D-fingerprints equation are molecular similarity eigenvalues, and this set of descriptors is being evaluated as a potential "universal" QSAR descriptor set. The 4D-fingerprint model suggests that Caco-2 cell permeation is governed by the spatial distribution of hydrogen bonding and nonpolar groups over the molecular shape of a molecule. Moreover, a complementary resampling of the original Caco-2 cell permeation training set, followed by the construction of several "clustered" MI-QSAR models, led to a consensus model consistent in interpretation with the 4D-fingerprint model. © 2007 Wiley-Liss, Inc. and the American Pharmacists Association J Pharm Sci 97:566–583, 2008

Keywords: Caco-2 cell permeation; membrane; QSAR; universal 4D-fingerprint descriptors; drug absorption; molecular similarity; cluster analysis

INTRODUCTION

Drug discovery programs generally focus on the development of orally active drugs. This preferred route of drug administration is often considered an absolute requirement by those who market drugs.

One *in vitro* model that has been shown to correlate with the transport of a drug across the intestinal epithelial cell barrier is a Caco-2 cell monolayer.[1] Caco-2 cells, a well-differentiated intestinal cell line derived from human colorectal carcinoma, display many of the morphological and functional properties of the *in vivo* intestinal epithelial cell barrier. Consequently, Caco-2 cell models are used regularly for the determination of cellular transport properties. That is, in both industry and academia, Caco-2 cell models are surrogate markers for *in vivo* intestinal permeability in humans.[2]

The increasing use of Caco-2 cell screening for oral bioavailability has led to an increasing interest in trying to understand the mechanism of cellular permeability of this screen. While it seems relatively clear that passive diffusion usually underlies transport, understanding more about Caco-2 cell transport would:

Correspondence to: Osvaldo A. Santos-Filho (Telephone: 604-875-4111 ext. 61518; Fax: 604-875-4013; E-mail: osvaldo@interchange.ubc.ca)

Journal of Pharmaceutical Sciences, Vol. 97, 566–583 (2008)

(1) Define the appropriate range of applications and solute chemistries for which this screen is applicable.
(2) Identify the key properties of the organic solutes (drugs) responsible for cellular permeation behavior.
(3) Permit construction of possible quantitative structural activity relationship, QSAR, models to use as virtual screens to predict Caco-2 cell permeation, and, more generally, oral bioavailability.

Recently, there has been a surge in computational efforts to compute ADME properties, including Caco-2 cell permeation of structurally diverse compounds and drugs.[3] Such computational approaches remain focused on modeling structurally diverse solute data sets by dealing with only the properties of the solutes. Researchers have adopted a "philosophy" to get around the limitations inherent to performing a QSAR analysis on a structurally diverse data set. The number of intramolecular solute properties (descriptors) computed is made as large as possible. Then a data reduction method is employed as part of the data-fitting process in constructing the QSAR model. The idea is that if enough solute features are included in the trial properties set, the key intramolecular solute properties for describing multiple mechanisms of action (permeation) can be captured and built into the QSAR model without data overfitting.

Of course, there is no way to know if the right set of intramolecular solute features are included in the QSAR descriptor pool, let alone if any set of intramolecular solute descriptors exist that can capture the requisite mechanistic information by themselves. Moreover, once data reduction is performed, it becomes exceedingly difficult to interpret the resulting QSAR model. Within this context, it can be argued that some type of structure-based QSAR approach is needed to meaningfully handle the chemical and structural diversity of the solutes commonly encountered in constructing ADME-property QSAR models.

We have developed a method called *membrane-interaction (MI) QSAR* analysis, where structure-based design is combined with classic intramolecular QSAR analysis to model chemically and structurally diverse compounds interacting with cellular membranes.[4–6] MI-QSAR analysis has proved effective in generating models for several ADMET properties, such as predicting eye irritation by organic molecules,[4,5] predicting Caco-2 cell permeability coefficients,[7] predicting blood–brain barrier partition coefficients,[8] and characterizing skin penetration processes of organic molecules.[9]

The basic assumption of MI-QSAR analysis is that the phospholipid regions of a cellular membrane constitute the "receptor" required in structure-based design. This correspondingly permits the incorporation of structural and chemical diversity into a training set. A set of *membrane-solute intermolecular properties* are determined and added to a set of "classic" *intramolecular* solute QSAR descriptors to enlarge the trial QSAR descriptor pool. Ostensibly, this provides the information needed to incorporate chemical and structural diversity into the QSAR analysis.

In a previous article,[7] a training set of 30 structurally diverse molecules, whose permeability coefficients across the cellular membranes of Caco-2 cells were measured, was used to construct significant MI-QSAR models of Caco-2 cell permeation. According to that work, cellular permeation was found to depend primarily upon the aqueous solvation free energy (solubility) of the drug, the extent of drug interaction with a phospholipid model (dimyristoylphosphatidylcholine, DMPC) monolayer, and the conformational flexibility of the solute within the model membrane. Part of the work reported in this article is an extension of that study, and now includes training set clustering[10,11] to better define groups of chemically similar molecules in the diverse training set.

However, the issue of expanding the trial descriptor set used to build the QSAR model has also been addressed in this study. As part of the 4D-QSAR paradigm, we have developed a method of 4D molecular similarity analysis.[12,13] The components that go into the numerical estimates of molecular similarity are eigenvalues from a conformationally averaged distance matrix of a molecule. These eigenvalues have been shown to be useful and, perhaps, "universal" descriptors and are now referred to as 4D-fingerprints.[14] The 4D-fingerprint descriptors have been employed in this study to build Caco-2 cell permeation models which can be "combined" with the MI-QSAR models to provide a consensus virtual screening model.

MATERIALS AND METHODS

Caco-2 Cell Permeation Coefficients

The dependent variable in this MI-QSAR analysis is the Caco-2 cell permeability coefficient, $P_{\text{caco-2}}$.