Chapter 14

In Silico Prediction of Blood Brain Barrier Permeability: A Support Vector Machine Model

Zhi Wang
Beijing University of Chemical Technology, China

Aixia Yan
Beijing University of Chemical Technology, China

Jiaxuan Li
Beijing University of Chemical Technology, China

ABSTRACT

The ability of penetration of the blood-brain barrier is an important property for the development of Central Nervous System drugs, which is commonly expressed by logBB (logBB = log(C_{brain}/C_{blood}). In this work, a support vector machine was used to build quantitative models of blood brain barrier permeability. Molecular descriptors for 182 compounds were calculated by ADRIANA.Code and 12 descriptors were selected using the automatic variable selection function in Weka. Based on two common physicochemical descriptors (xlogP and Topological Polar Surface Area (TPSA)) and 10 2D property autocorrelation descriptors on atom pair properties, an SVM regression model was built. The built model was validated by an external test set. The reliable predictions of the test set demonstrate that this model performs well and can be used for estimation of logBB values for drug and drug-like molecules.

INTRODUCTION

In the discovery and development process of a central nervous system (CNS) targeted drug, one important property is the ability of a drug to penetrate the blood-brain barrier (BBB) (Norinder et al., 2002). Usually, BB is defined as the brain-blood concentration ratio of a compound at steady state, which is commonly expressed by logBB (logBB = log(C_{brain}/C_{blood}), which C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively) to present the extent of a drug passing through the blood-brain barrier.
However, it is expensive and time-consuming to obtain the experimental data on blood-brain distribution ratio of a compound. Therefore, it is necessary to use a reliable in silico model to predict logBB for drug candidates.

Numerous models on quantitative prediction of logBB models have been published. Earlier reported models were built by Linear Regression and Multiple Linear Regression (MLR) based on some simple physicochemical descriptors. Young et al. (1988) used the important physicochemical properties (ΔlogP) for brain penetration employing the linear regression method on centrally acting histamine H₂ antagonists. Based on Young’s data set, Van de Waterbeemd and Kansy (1992) investigated the importance of hydrogen bonding on logBB by the MLR method and obtained a good relationship with logBB using the Polar Surface Area (PSA) and molecular volume as descriptors. Calder and Ganellin (1994) also investigated Young’s data set using experimental ΔlogP values and theoretically computed descriptors such as PSA and molecular volume by the MLR method.

In order to get more reliable models, researchers tried to (1) enlarge the data set to extend the chemical space (Abraham et al., 1994; Salminen et al., 1997; Subramanian & Kitchen, 2003; Garg & Verma, 2006); (2) investigate more related descriptors (Lombardo, 1996; Kaliszan & Markuszewski, 1996; Clark, 1999; Ertl et al., 2000; Hutter, 2003; Hou & Xu, 2003; Abraham, 2004; Sun, 2004); and (3) build models by advanced data mining methods such as Partial Least Square (PLS) analysis, Genetic Algorithms and Artificial Neural Networks (Norinder et al., 1998; Luco, 1999; Platts et al., 2001; Iyer et al., 2002; Winkler & Burden, 2004; Yap & Chen, 2005).

In our recent research, it is found that 2D property autocorrelation descriptors have worked well in determining whether a compound may cross the blood-brain barrier or not (Wang et al., 2009). In this work, the 2D property autocorrelation descriptors and some other 2D physicochemical descriptors were used to build a quantitative model for prediction of logBB using a Support Vector Machine (SVM).

**MATERIALS AND METHODS**

**Data Set**

182 compounds with logBB experimental values and SMILES strings were collected from published work by Garg & Verma (2006). The experimental logBB of all these compounds were provided in Appendix. CORINA (CORINA, Molecular Networks GmbH, Erlangen, Germany, http://www.molecular-networks.com) was used to add hydrogen atoms and to compute 3D structures. For each molecule, only a single 3D conformation was generated. Then the whole data set were split into as training set (122 compounds) and test set (60 compounds) randomly. The training set (122 compounds) was subjected to 10-fold cross validation, and the test set (60 compounds) was used as an external validation set.

**Molecular Descriptors**

In our study, a total of 62 descriptors were calculated using ADRIANA.Code (ADRIANA.Code, Molecular Networks GmbH, Erlangen, Germany, http://www.molecular-networks.com), including 6 global molecular descriptors and 56 2D property autocorrelation descriptors.

Global molecular descriptors, expressed by a single value, represent a chemical structure by a structural, chemical or physicochemical feature or property of the molecule. In this study, six important molecular physicochemical descriptors related to brain penetration were calculated. They are molecular weight, number of H-bond donors, number of H-bond acceptors, topological polar surface area (TPSA) (Ertl et al., 2000), octanol/water distribution coefficient (xlogP) (Wang et al., 2000) and mean molecular polarizability (Miller, 1990).
Related Content

Addressing the Challenges of Detecting Epistasis in Genome-Wide Association Studies of Common Human Diseases Using Biological Expert Knowledge
www.igi-global.com/chapter/addressing-challenges-detecting-epistasis-genome/76091?camid=4v1a

Perspective Wall Technique for Visualizing and Interpreting Medical Data
Hela Ltifi, Mounir Ben Ayed, Ghada Trabelsi and Adel M. Alimi (2012). International Journal of Knowledge Discovery in Bioinformatics (pp. 45-61).
www.igi-global.com/article/perspective-wall-technique-visualizing-interpreting/77930?camid=4v1a

Mining Medical Data to Develop Clinical Decision Making Tools in Hemodialysis: Prediction of Cardiovascular Events and Feature Selection using a Random Forest Approach
www.igi-global.com/article/mining-medical-data-develop-clinical/73908?camid=4v1a

Performance Analysis of Microarray Data Classification using Machine Learning Techniques
www.igi-global.com/article/performance-analysis-of-microarray-data-classification-using-machine-learning-techniques/167709?camid=4v1a