Chapter 12
Application of Machine Learning in Drug Discovery and Development

Shuxing Zhang
The University of Texas at M.D. Anderson Cancer Center, USA

ABSTRACT

Machine learning techniques have been widely used in drug discovery and development, particularly in the areas of cheminformatics, bioinformatics and other types of pharmaceutical research. It has been demonstrated they are suitable for large high dimensional data, and the models built with these methods can be used for robust external predictions. However, various problems and challenges still exist, and new approaches are in great need. In this Chapter, the authors will review the current development of machine learning techniques, and especially focus on several machine learning techniques they developed as well as their application to model building, lead discovery via virtual screening, integration with molecular docking, and prediction of off-target properties. The authors will suggest some potential different avenues to unify different disciplines, such as cheminformatics, bioinformatics and systems biology, for the purpose of developing integrated in silico drug discovery and development approaches.

INTRODUCTION

Drug discovery and development is regarded as one of the most complex research areas encompassing many disciplines (Cohen et al., 2004; Gomeni et al., 2001; Martin, 1991; Tropsha & Zheng, 2001). This might be part of the reasons why it is extremely expensive and time-consuming, and why the current pharmaceutical business is hitting a wall with its severely stalled discovery engine (Tralau-Stewart et al., 2009). In seeking efficient and costly effective approaches, computer-aided methods are gaining more and more attentions, and recent years have witnessed dramatic progress in computer-aided drug design (Cohen et al., 2004; Gomeni et al., 2001; Martin, 1991; Tropsha & Zheng, 2001). This is partially due to the significant advances in the analysis of explosively growing biological and chemical data using modern machine learning techniques (Mitchell et

DOI: 10.4018/978-1-61520-911-8.ch012

al., 1989; Mjolsness & DeCoste, 2001; Schneider & Downs, 2003). *Machine learning* has been referred to the development of algorithms that improve their performance in pattern recognition, classification, regression and prediction based on the models derived from existing data. Therefore, it is closely related to fields such as data mining, pattern recognition, theoretical computer science, and many other areas (Mitchell et al., 1989). For instance, algorithms for classification have been used frequently to identify active and inactive compounds, while regression approaches are applied to the training and prediction of continuous data. Despite being widely used in other biomedical research such as bioinformatics, we will focus herein on its application to small molecule drug discovery and development.

Currently there is a variety of existing implementations of machine learning. However, it is often difficult to assess the usefulness and limitations of a particular method for the problems at hand (Mitchell et al., 1989; Schneider & Downs, 2003). In drug discovery and development, machine learning has been widely used in *quantitative structure-activity relationship (QSAR),* ligand-based *virtual screening, in silico ADMET* (absorption, distribution, metabolism, excretion, and toxicity) studies, and many other areas (King et al., 1992; King et al., 1996; Mitchell et al., 1989; Mjolsness & DeCoste, 2001; Schneider & Downs, 2003). Different QSAR approaches have been developed during the past few decades (Hansch et al., 1963; Klein et al., 1986; Kubinyi, 1986). In a generalized classification or regression problem, modern QSAR are characterized by the use of multiple descriptors of chemical structures combined with the application of both linear and non-linear optimization techniques, and a strong emphasis on rigorous model validation to afford robust and predictive QSAR models (Tropsha, 2005; Tropsha, 2006). *Molecular descriptors* are used for representing structural and physicochemical properties of compounds. More than 3000 thousand descriptors have been developed to date, ranging from constitutional descriptors, such as molecular weight, to more complex 2D and 3D descriptors representing different topologic, geometric, connectivity, and physicochemical properties (Li et al., 2007). Frequently used descriptors in QSAR modeling include constitutional descriptors (e.g., counts of atoms, bonds, etc.), property-based descriptors (e.g., logP), BCUT descriptors, topological descriptors, geometrical descriptors, electrostatic, quantum chemical descriptors, thermodynamic descriptors, and many others. These descriptors can be calculated by several popular programs such as DRAGON (Tetko et al., 2005), Molconn-Z (Kellogg, 2002), MOE (Chemical Computing Group, Quebec, Canada), CODESSA (http://www.codessa-pro.com/index.htm), ADMET Predictor (Simulation Plus, Lancast, CA), JOELib (http://www.ra.cs.uni-tuebingen.de/software/joelib/), and PowerMV (http://www.niss.org/PowerMV/).

Once descriptors are obtained for the molecules, statistical modeling techniques are required to establish correlation between the descriptors and activities. For instance, a comprehensive review of the application of neural network in a variety of QSAR problems has been presented, which discussed how NNs can be applied to the prediction of physicochemical and pharmacokinetic properties (Baskin et al., 2008). In addition, SVM was found to yield improved performance compared to multiple linear regressions (MLR) and radial basis functions (RBF) (Yao et al., 2004). Various version of SVM programs have been developed and they were used in calculating the activity of enzyme inhibitors as well as in many other studies of similar types (Duch et al., 2007). In addition to the activity prediction of molecules, QSAR models are also frequently used in virtual screening for hit discovery (Hansch & Fujita, 1995; Kubinyi, 1990; Tropsha & Golbraikh, 2007). Virtual screening is usually applied to the identification of those that are potentially active in the biological tests of interest. The ultimate goal is to reduce the molecules to be tested from a