The Use of Computational Data Mining Methods for Selecting and Optimizing Lead Compounds

Zsolt Lepp
2008 Dec. 02

Part 1. The use of data mining methods to assist the analysis of molecular modeling studies.
The methods which are used to predict enzymatic activities of ligands could be grouped into two main approaches. (1) QSAR type analyses, using numerical descriptors of some properties of ligands, and relate these properties to enzymic kinetic parameters. (2) directly calculating the binding properties of ligands from various molecular level simulations of the 3D structure of receptor-ligand complexes. The purposes of this work was to search for links between traditional QSAR and molecular level simulations and to find out how the combination of these two approaches could improve the predicting ability of both models. *Ref.:* Connecting Traditional QSAR and Molecular Simulations of Papain Hydrolysis - Importance of Charge Transfer. Z. Lepp, H.Chuman. Bioorganic and Medicinal Chemistry *Bioorganic & Medicinal Chemistry.* **2005,** *13,* 3093- 3105.

Part 2. Creation of virtual screening models for rapid selection of potentially active molecules from large-scale data.
In the last decade the amount of chemical-biological information has been exponentially growing. This makes the use of virtual screening methods more and more necessary to select potential lead molecules and to characterize large-scale chemical libraries. In the frame of this study a set of generally applicable atom-type descriptors were used together with the data mining method called Support Vector Machines (SVM) to create virtual screening models. These models were successfully used to accurately predict potential activities of molecules against a set of antidepressant targets, as well as broader therapeutic areas. Ref.: Screening for New Antidepressant Leads of Multiple Activities by Support Vector Machines. Z. Lepp, T. Kinoshita, H. Chuman. *J. Chem. Inf and Model* **2006**; *46*(1); 158-167.

Part 3. Selecting and characterizing pharmacologically relevant molecular fragments.
In the frame of this research, 240 proteins were organized by using the ligands of these targets. The purpose of the work was to extract significant molecular fragments that might be responsible for the activities of those organic molecules. The maximum common substructures (MCS) of each of the 240 ligand data sets were determined. The dataset can be used to perform various scientific investigations, such as (1) to analyze the characteristic chemical properties of the inhibitors of a protein; (2) to determine which chemical features are common among numerous data sets; (3) creation of prediction methods to find novel active molecular structures; (4) to help the assembly of chemical probes by providing small molecular fragments ordered by pharmacological similarity.

Part 4. Using network analysis methods on the network of molecules determined by molecular similarity, to perform various chemoinformatics tasks.
As many current chemical researches results in a large number of novel molecular structures it is becoming increasingly necessary to efficiently manage large molecular libraries. In the current research a new approach is proposed: analyzing the chemical libraries by creating a network, in which compounds are linked to each other by structural similarity. It has been investigated that how traditional chemoinformatics tasks, such as building structure-activity models, virtual screening, clustering, etc could be accomplished by the means of analyzing such networks.
Ref.: Finding Key Members in Compound Libraries by the Analyses of Molecular Structure Similarity Networks; Z. Lepp, C. Huang, T. Okada; *submitted* ***2008. Oct.***