Quantitative Structure Activity Relationship Modeling for Protein Chromatography Separation System

by

Min Li

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the degree of
MASTER OF Information Technology

Approved:

____
Curtis M. Breneman, Thesis Adviser

____
Steven M. Cramer, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

May 2007

ABSTRACT

The Quantitative Structure Activity Relationship (QSAR) research has widely been applied in many fields especially in cheminformatics. Through building QSAR models we can better understand the physical mechanism of the underlying phenomena and predict the potential candidates. The QSAR Applicability Domain (AD) issues are extensively discussed in this thesis. QSAR Applicability Domain estimation is extremely useful for the dataset in that it gives us the scope and limitation of the models. Four QSAR AD estimation methods based on the training set are introduced and applied to the protein chromatography dataset. Molecule solubility and molecule atom type are introduced as well. They are very important steps we should take before QSAR modeling.

Three machine learning methods, Kernel Partial Least Square (KPLS) Regression, Support Vector Classification (SVC) and Binary Classification Trees, are widely used in QSAR modeling. They are also introduced in this thesis. For Gaussian Kernel PLS, we applied different Gaussian Kernel sigma to different types of descriptors. It is proven to be better than normal kernel PLS for the protein chromatography dataset.

Virtual high throughput screening, a technique widely applied to drug discovery, was adapted to finding novel selective displacers for displacement chromatography. Kernel PLS, Support Vector Classification and binary classification tree models are generated based on the protein chromatography training set and work well on training set. These models are then applied to predict a commercial catalog – Matrix Scientific Catalog. Some of predicted molecules will be tested further. After the parallel batch

screening experiments, these molecules will be added to the training set. New QSAR models will be built from the new dataset and predict new molecules.