# Support Vector Machines for High Throughput Screening Data Mining: Type I MetAPs Inhibition Study

**Jianwen Fang, Yinghua Dong, Gerald H. Lushington, Qi-Zhuang Ye, and Gunda Georg, Bioinformatics Core Facility/Information and Telecommunication Technology Center/ Molecular Graphics and Modeling Laboratory/High Throughput Screening Laboratory, University of Kansas, Lawrence, KS 66045.**
2006 Annual Kansas City Area Life Sciences Research Day.

This paper reports a successful application of Support Vector Machines (SVM) in mining high throughput screening (HTS) data of a type I Methionine aminopeptidases (MetAPs) Inhibition study. A library with 43,736 small organic molecules was used in the study and 1355 compounds in the library with 40% or higher inhibition activity were considered as active. The dataset was randomly split into a training set and a test set (3:1 ratio). We were able to rank compounds in the test set using their decision values predicted by SVM models that were built on the training set. We defined a novel score $PT_{50}$, the percentage of test set needed to be screened in order to recover 50% of the actives, to measure the performance of the models. With carefully selected parameters, SVM models increased the hit rates significantly and 50% of the active compounds could be recovered by screening just 7% of the test set. We found the size of the training set played a significant role in the performance of the models. A training set with ten thousand member compounds is likely the minimum size required to build a model with reasonable predictive power.

This work is supported by the K-INBRE Bioinformatics Core (NIH grant number P20 RR016475) and NIH Grant RR-P20 RR17708.