Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets

Subhash C. Basak\*a, Subhabrata Majumdarb

*1 Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA;* E-Mail: sbasak@nrri.umn.edu

*2 School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA;* E-Mail: majum010@umn.edu

*\* Author to whom correspondence should be addressed;* E-Mail: sbasak@nrri.umn.edu; Tel.: +1 (218)720-4230; Fax: +1 (218)720-4328.

**Abstract:** Variation in high-dimensional data is often caused by a few latent factors, and hence dimension reduction or variable selection techniques are often useful in gathering useful information from the data. In this paper we consider two such recent methods: Interrelated two-way clustering and envelope models. We couple these methods with traditional statistical procedures like ridge regression and linear discriminant analysis, and apply them on two data sets which have more predictors than samples (i.e. *n* << *p* scenario) and have several types of molecular descriptors. One of these datasets consists of a congeneric group of Amines while the other has a much diverse collection compounds. The difference of prediction results between these two datasets for both the methods supports the hypothesis that for a congeneric set of compounds, descriptors of a certain type are enough to provide good QSAR models, but as the data set grows diverse including a variety of descriptors can improve model quality considerably.

**Keywords:** Envelope models, hierarchical quantitative structure-activity relationship (HiQSAR), interrelated two-way clustering; linear discriminant analysis, mutagenicity, ridge regression, topological indices, congenericity principle, diversity begets diversity principle

Running title: QSAR on homogeneous vs. diverse datasets