AbstractID: 5070 Title: Data Mining and Knowledge Discovery Methodology Development using Predicting Transmembrane Proteins by Synergy of Computational and Molecule Biophysics Approaches as an example

Motivation: We consider developing data mining and knowledge discovery methodology in medical physics is important. We use predicting membrane proteins as an example, because such proteins account for roughly one third of proteins encoded in Human Genome and comparative genomes, and play crucial roles ranging from signal transduction to energy metabolism, but their structures can not be determined well by experimental methods. Since Human Genome had been completely sequenced, function and structure of proteins coded for by Human Genome are not known well. In our attempts to construct methods for automated large-scale structural and functional annotation of genes and proteins in Human Genome and comparative genomes, the identification of transmembrane proteins is thus an important, but also very difficult task. We used a broad array of techniques that combined computational and biomedical physics approaches for the above task.

Results: We developed a hybrid unsupervised-supervised classifier using new variants of the Self-Organizing Feature Map algorithms, in combination with novel Feature Generation, Feature Selection and Ensemble Methods such as Boosting with bagging and boosting with confidence Information. Our combined computational and biomedical physics approaches proved beneficial by showing larger areas in our predictor's ROC curves. We will discuss the effectiveness our synergic approaches, and also provide comparisons to more traditional classifiers such as neural networks, decision trees and support vector machines. The framework that we developed can be broadly applied to other medical physics problems and marks a beginning of methodology development, not only in molecular biophysics and genomics, but also in medical physics.