ON COMPARING THE CLUSTERING OF REGRESSION MODELS METHOD WITH K-MEANS CLUSTERING

L.X. Qin[1] and S.G. Self[2]

[1]*Memorial Sloan-Kettering Cancer Center, New York, USA*
[2]*Fred Hutchinson Cancer Research Center, Seattle, USA*

Email: *qinl@mskcc.org*

Gene clustering is a common question addressed with gene expression data. Methods proposed for this problem include algorithmic clustering methods, such as K-means clustering and hierarchical clustering, and model-based clustering methods, such as the mixture of multivariate normal method. These methods base the clustering directly on the observed measurements. Qin and Self (Biometrics 2006) proposed a new model-based clustering method, the clustering of regression models (CORM) method, which bases the clustering of genes on their relationship to covariates. This method explicitly models different sources of variations and complements regression-based per-gene analysis. In this paper, we discuss connections and differences between the CORM method and K-means clustering. We show that the CORM method tends to seek a partition of genes that has stable cluster centers across samples. We use simulated data to compare the performance of the clustering of linear models (CLM) method, K-means clustering, and an extended K-means clustering with respect to efficiency and robustness to model misspecification. Simulation results show that the CLM method outperforms K-means clustering and the extended K-means clustering when the assumed regression model is true and is robust to certain model misspecifications. We also use a microarray dataset to demonstrate a scenario where only the CORM method is appropriate.