Details of the Simulated Data Sets*
Supplement to the paper "Validating Clustering in Gene Expression Data"(to appear in Bioinformatics)

Ka Yee Yeung, David R. Haynor, Walter L. Ruzzo

November 17, 2000

The simulated data set in Figure 1 has 420 genes and 17 conditions, and contains 5 clusters. The modeling of the simulated data set is an effort to show the power of our methodology, and is a preliminary effort to model gene expression data sets. We do not claim that it is the correct model for gene expression data sets.

Let $D(i,j)$ be the simulated expression level of gene i and condition j in the simulated data set with five clusters in Figure 1. Let $D(i,j) = \delta_j + \lambda_j * (\alpha_i + \beta_i \phi(i,j))$, where $\phi(i,j) = \sin(\frac{2\pi j}{8} - \frac{2\pi k}{5})$. α_i represents the average expression level of gene i, which is chosen according to the standard normal distribution. β_i is the amplitude control for gene i, which is chosen according to a normal distribution with mean 2 and standard deviation 0.3. $\phi(i,j)$ models cyclic time series data, two cycles in this case. A cluster is modeled with a phase shift. k is the cluster number, which is chosen according to Zipf's Law [Zipf 1949], which allows us to model clusters with different sizes. λ_j is the amplitude control of condition j, chosen according to the normal distribution with mean 1 and standard deviation 0.1. δ_j represents an additive experimental error, chosen according to the normal distribution with mean 0 and standard deviation 0.1.

*We would like to thank Lue Ping Zhao at the Fred Hutchinson Cancer Research Center for suggestions of modeling gene expression data.