CS 426 Introduction to Computational Biology

Homework 6
(due Friday December 5, 2003)

1. Implement from scratch (de novo) the k-means clustering algorithm
2. Use it on the *Saccharomyces cerevisiae* (baker's yeast) expression data provided on the website.
    a. Run the algorithm for k =100 number of clusters
    b. Use both Euclidian distance and Pearson correlation metrics. Compare the clusters obtained using both.
    c. For missing values use only the available data.
3. Interpret the results
    a. Pick the five "best" nontrivial clusters (at least 5 members.
    b. Draw the expression profile of each gene in the cluster overlapping one another.
    c. Try to see if the clusters selected make sense based on the description of genes
4. What to turn in: Submit electronically the code, a printed report containing the answers for 3 and a code printout.