

Computer Science 549 – Computational Biology

Prof. Steven Skiena

Fall 2005

Homework 3

Due Tuesday, November 22, 2005

November 8, 2005

Each of the problems should be solved on a separate sheet of paper to facilitate grading. Limit the solution of each problem to one sheet of paper unless otherwise stated. Many of these problems are deliberately open-ended and vague; do the best you can on them but don't make yourself crazy. Please don't wait until the last minute to look at the problems.

You must do this assignment in groups of 2 to 3 people.

1. Gain access to some kind of clustering program/environment. If you are familiar with Matlab, you can use that. Otherwise download the R Statistical Library from www.r-project.org/. Read through the "Introduction to R" document available under "Manuals" to orient you to the package.

Experiment with the clustering commands in your environment. In particular, compare the results of average, complete, and single-link clustering methods. A dissimilarity matrix and instructions on how to use the clustering commands will be posted. at <http://www.cs.sunysb.edu/~skiena/549>.

Write a one page report on your experiences.

2. Do a cluster analysis of one data set of interest to you using the R Statistical library. You may pick whatever you want, but possible choices include:
 - Microarray data from the Stanford Microarray Database at <http://genome-www5.stanford.edu/>. Perhaps the Gasch, et.al data (AGASCH) under "Public Login" is interesting. Either cluster genes by similarity or experimental conditions (microarrays) by similarity.
 - Stock market and financial data – can you cluster companies into natural groups based on how their stock has performed over time? Historical data is available from <http://kumo.swcp.com/stocks/> and finance.yahoo.com

Your job is to identify an appropriate distance function to measure similarity/dissimilarity among the objects, and then cluster them.

Experiment with different clustering algorithms and different cost functions. Do you get clusters that reflect some sense of reality? How does changing parameters and algorithms effect the quality of the clusters?

Write a three page paper on your experiments.