CS 677 High-performance Computing
Programming Assignment #4

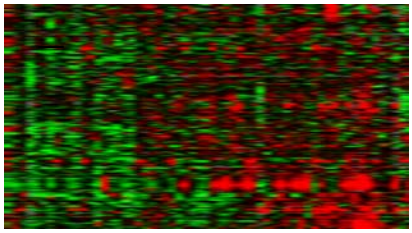
Distributed Computation via MPI: using Microarray Analysis for Genome Wide Association Studies

Specification

The basic idea is to develop an MPI-based message-passing application that analyzes raw microarray data to identify differentially expressed genes between two user-defined groups of patient samples.

Background

The ongoing development of cDNA microarray technology has facilitated the simultaneous measurement and comparison of gene expression on a genomic scale. One technique uses the fluorescent intensity ratios of differentially expressed genes to characterize gene expression patterns. These patterns, in turn, can reveal the gene sets that underlie a particular phenotype or that represent a regulatory gene defect.



For example, a molecular profiling study might attempt identification of gene expression signatures to distinguish groups of samples based on various specified parameters (e.g. differentiation state, tumor stage, tumor grade).

Discrimination Algorithm

To determine expression signatures, individual genes are scored and ranked based on an appropriate discrimination metric. The algorithm first separates individual gene expression values based on the supplied sample grouping; then the Student's t-statistic is calculated and used to identify significant discriminators (i.e. genes that significantly distinguish between the two sample groups). An effective form of the t-statistic is:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{(\sigma_1)^2}{n_1}\right) + \left(\frac{(\sigma_2)^2}{n_2}\right)}}$$

where n_1 and n_2 are the sizes of the two groups, x_1 and x_2 are the means of the two groups, and σ_1 and σ_2 are the sample standard deviations of the two groups. The significance of the resulting t-statistic can be determined via permutation analysis;

that is, by comparison to a distribution of t-statistics generated by randomly permuting the sample groupings.

For each gene a distribution of randomly generated t-statistics is created. This distribution has a mean and a standard deviation, and the significance of a discriminating gene can be determined by comparing its t-statistic to the distribution of t-statistics generated by random permutation of samples (see diagram below left). If the initial t-statistic is 3 standard deviations from the mean of the random distribution, then with 97% confidence this gene is a discriminator.

$$D = \frac{|t_S - \mu_D|}{\sigma_D}$$

Each gene is ranked by its discrimination score (D, above right), where t_S is the t-statistic of the reference sample, μ_D is the mean of the random distribution, and σ_D is the standard deviation of the random distribution. A gene's D-score is equivalent to the actual t-statistic's *z*-score with respect to a distribution of random t-statistics.

Example

Consider the following gene expression matrix, where the rows are genes and the columns represent measured intensity values for six samples. The myc and Met genes appear to be differentially expressed between the diseased and normal sample groups.

Gene	name	disease1	disease2	disease3	normal1	normal2	normal3
gene1	myc	0.1	0.2	0.3	2.3	3.1	2.5
gene2	PDGF	1.1	1.1	1.3	1.2	1.1	1.0
gene3	Met	4.5	5.5	6.5	2.0	1.9	0.9

The question is: what gene best distinguishes between the normal and disease groups? To begin, a student's t-test is performed between the disease group values vs. the normal group values for each gene (e.g. for the **myc** gene, calculate the t-statistic between (0.1, 0.2, 0.3) and (2.3, 3.1, 2.5)). The results appear in the table below:

name	T-statistic
myc	-9.84
PDGF	0.76
Met	5.77

It appears that the **myc** gene is better than the **Met** gene at distinguishing between the two groups. However, that conclusion is only valid if the gene expression values of the population of every group constituted a normal distribution. We cannot make that assumption. Therefore we must "construct" the distribution for each gene, and see where each gene's t-statistic falls within that distribution. For each gene, a large set of random sample groupings are constructed (of the same sizes as the original reference grouping). One random sample grouping for the **myc** gene might be (disease1, normal1, disease2) vs. (disease3, normal2, normal3). For each random

sample, a t-test is performed and a t-statistic is generated and used to construct the random distribution. Finally, the original (reference) t-statistic is compared with the random permutation distribution to compute the D-score as described above.

Microarray Data

Use the anonymized microarray patient data from the National Cancer Institute (NCI-60) to identify genes that best serve as discriminants for renal cancer ("RE"). There are eight diseased samples, out of 60 total patients. There are 4550 total genes.

Note: an empty cell indicates that data does not exist for that gene for that patient; you must adjust your statistical calculations accordingly to account for the difference(s) in group size.

Note 2: the original data set is formatted as an Excel spreadsheet. I have performed some pre-processing and created an alternate spreadsheet you may choose to use:

- Format is Comma-Separated Values (.csv)
- Only includes Gene ID and intensity values
- Removes Gene 4538X (disease sample size of 1)

Requirements:

Create an OpenMPI-based solution using C/C++ or Python. Your solution should:

- □ Distribute the computation workload
  - o Note that the computation of each gene's t-stat and subsequent permutation analysis is independent of other genes
- □ Compute the t-statistic for each gene in the specified dataset
  - o Use two user-defined groups (Renal vs. Control)
- □ Calculate the discriminant ranking (D-score) for each gene based on 1000 permutations of the sample data.
- □ Sort the genes by D-score
  - o Answer the question: which genes best identify the Renal Cancer group?
- □ Conduct a performance evaluation (i.e. Speedup).

As usual:

- □ Submit your source code, program output (top 10 genes), and performance analysis/discussion.
- □ Be prepared to discuss your results.