Semester Project
Cmp Sci 4370/5370 Biological Data Science

Description

- Given a set of discretized gene expression values for diseased cases and normal controls, identify a genetic pattern consisting of:
  - A subset of genes
  - The expression directions for each of the genes in the subset
- Such that the percentage of cases with the genetic pattern minus the percentage of controls with the pattern is maximized
- This selection must be fully automated in your code
  - Do not rely on visual inspections of heatmaps

Project Definition

- Let *p* be a 2-dimensional matrix of integral constants
  - Number of rows = number of genes
  - Number of columns = *C* + *N* (number of diseased cases + normal controls)
  - Represents the discretized expression data (1 = high, 0 = neutral, -1 = low)
- Let *u* and *d* be 1-dimensional arrays of binary variables
  - Number of variables = number of genes, *G*
  - u_i = 1 if and only if gene *i* is in the selected expression pattern with high expression direction
  - d_i = 1 if and only if gene *i* is in the selected expression pattern with low expression direction
- Let *p* be 1-dimensional arrays of binary variables
  - Number of variables = *C* + *N* (number of diseased cases + normal controls)
  - Values are set to 1 if and only if the corresponding individual carries the entire expression pattern represented by *u* and *d*
  - For each individual *j*, $1 \le j \le C+P$, p_j = 1 if and only if:

$$\sum_{i=1}^{G} c_{ij} u_i = \sum_{i=1}^{G} u_i$$

AND

$$\sum_{i=1}^{G} -c_{ij} d_i = \sum_{i=1}^{G} d_i$$

- Objective is to maximize *J*:

$$J = \frac{\sum_{j=1}^{C} p_j}{C} - \frac{\sum_{j=C+1}^{C+N} p_j}{N}$$

- This is Youden's J statistic

Data Input Format

- One header row with individual IDs
- One header column with gene IDs
- *C*+*N* data columns
- *G* data rows
- Data consists of four possible values: 1, 0, -1, -2
  - 1 = high
  - 0 = neutral

    - -1 = low
    - -2 = missing data
- All values are tab separated

- Example:

IDs	D1	D2	C1	C2	C3
G1	-1	0	1	0	1
G2	1	1	-1	1	-1
G3	0	-1	0	0	-1
G4	-1	-1	0	-1	-1

Data Output Format

Expression Pattern:
G2 High
G4 Low
Cases with Pattern:
D1 D2
Controls with Pattern:
C2
J = 0.667

Other Requirements

- Command line arguments:
  ./pathway inputFileName numGenes numCases numControls outputFileName

- Must compile and run on clark server
- Strongly recommend C/C++
    - Other languages might not be feasible for large datasets
- Signed Academic Integrity statement must be submitted in order to receive a grade for project

Rubric

Item	**Points**
Compiles and runs with correct command line arguments on the clark server	25
Programming style	15
Correctly formatted output file	5
Logic and creativity	55