Running GeneProgram on the demo input files

This document describes how to run GeneProgram on the provided demo files. For more detailed information please see the published GeneProgram paper and supplemental material, as well as source code (in GeneProgram.jar).

These instructions assume that you are running Windows XP and Java 1.6.0; other configurations will require changes to the control files (see below).

Begin by creating a directory called c:\GeneProgram and unzip GeneProgram.zip to this directory.

Running the algorithm will then require three steps:

1. Preprocessing data
2. Generating samples
3. Generating the summary output files

1 Preprocessing data

1.1 Required input files

Three input files are required:

1. Main data file = demo_values.txt

This file contains the floating point expression data. The file is tab delimited. The first column contains the gene names and subsequent columns contain the experiments. The first row provides experiment labels. Values must be positive floating point values or NaN (not a number, to indicate missing data or below threshold expression). The demo file contains data for 500 genes and 42 experiments.

2. Modifier patterns data file = demo_mod1.txt

This file contains the modifier patterns (e.g., 6 possible temporal patterns). It has the same format as the main data file, except values must be integers. A value of -1 indicates missing data or below threshold expression (equivalent to NaN in the main data file).

3. Modifier pattern names = demo_mod_names.txt

This file contains labels for the modifier patterns (e.g., E+, M+, L+, etc.) The format is a tab delimited row of names.

1.2 Discretizing data

Go to the command prompt and type `cd c:\GeneProgram`. Then type `discretize`. This will run a batch file that performs preprocessing on the data. The preprocessing class uses a text control file called "control_discretize.txt". The file specifies parameters to be used for preprocessing. The basic parameters are:

--inputFileName	Main data file name
--baseOutputFileName	File prefix for outputting discretized data
--threshold	Threshold (absolute value) to consider gene differentially expressed
--numLevels_end	Number of discretization levels to use
--baseModifierFileName	File prefix for modifier pattern data and names

2 Generating samples

After the data has been discretized, a large number of samples must be generated. To begin the sampler, type `sample` at the command prompt. The demo generates 100,000 burn-in samples and 10,000 additional samples, which may take several hours to run (depending on your processor speed). As the sampler runs, it will output two numbers per line: the sample number, and the number of expression programs discovered for that sample. The sampler will also output intermediate "snap-shot" files in the output directory every 1,000 samples.

The sampler control files that specify parameters are called "control_burnin.txt" and "control_sample.txt". The basic parameters are:

--mainOutFileName	File name prefix for output snapshot files
--dataFileName	Main data file name
--modifierFileName	File name prefix for modifier pattern data and names
--iters	Number of iterations to run the sampler
--burnin	Number of iterations to burn-in the sampler
--snapShotInterval	Interval for saving snapshot files
--useGroups	Use tissue groups in the model
--groupStartIter	Number of iterations after which group sampling should begin

3 Generating the summary output files

Once samples have been collected, GeneProgram can create summary output (recurrent expression programs and consensus tissue groups). To generate summary files, type `summary` at the command prompt.

The demo data will generate approximately 50 recurrent expression programs and 15 consensus tissue groups.

The main file of interest is “c:\GeneProgram\summary\demoREPs_use.txt”. Each REP has an identifying row, a list of tissue usage statistics, and a list of genes

The first identifying row has five columns: (1) the REP identifier, (2) the percentage of samples the REP occurs in, (3) the mean generality score, (4) the number of tissues that significantly use the REP, and (4) the number of`genes in the REP

The second row lists the tissues that significantly use the REP (REPUse > 0.25). The third row gives the "REPUse" score, which is a weighted score of the extent to which the tissue uses all the genes in the REP. The fourth row gives the "tissueUse" score, which is the percentage of the tissue's differentially expressed genes that come from the REP. The fifth row indicates the percentage of samples in which the tissue significantly used the REP (REPUse > 0.25).

The rows labeled E+, M+, L+, E-, M-, L- indicate the percentage of samples in which a given tissue used the REP with a particular temporal pattern.

After these rows, the genes in the REP are listed. The left-most column is the empirical mean expression level. The next two columns are the gene name and gene description.

The file “c:\GeneProgram\summary\demo_groups.txt” reports the consensus tissue groups (each row is a group, with tissues separating by tabs).

The file “c:\GeneProgram\summary\demoREPs_biological_process.txt” reports the GO categories (if any) for which each REP is significantly enriched.