GProcessor 2.0a
Linux User's Manual

Zhong Guan, PhD
Hongyu Zhao's Lab of Statistical Genetics
Department of Epidemiology and Public Health
Yale University School of Medicine

GenePix™ **Pro** of **Axon Instruments, Inc.** is one of the most widely-used microarray analysis software. However, the built-in normalization function of **GenePix**™ **Pro** is just a simple linear normalization. Many research show that the most commonly used fluorescent dyes, say Cy3 and Cy5, are relatively unstable and may have label effect on channel intensities. This kind of label effect cannot be accounted for by the simple linear normalization. The most efficient nonlinear normalization method which can deal with the label effect is Lowess fit method which was originally proposed by William S. Cleveland as a statistical method. Another efficient statistical analysis method which can be used to analyze microarray data is the analysis of variance (ANOVA) method. We can get the fold change estimates and also account for the label effect and some other effects by the ANOVA procedure. This program ***GProcessor*** reflects our efforts to apply the above methods to **GenePix** output. Furthermore, this program also inherits some features of the program ***GPmerge***.

Input Data Format and Requirements

The input data must be given by **GenePix** Results (GPR) files which are in ATF - Axon Text File format (*.atf) which are created by any version of **GenePix**™ **Pro** software. All the GPR files should have the same **GenePix** Array List (GAL) file, the same number of channels(colors) and the same colors and treatments although the ratio formulae may differ.

Features

1. **Choosing a ratio type**:
   In **GenePix**, there are 5 different forms of ratio calculations for user to select, *i.e.*, *Ratio of Medians*, *Ratio of Means*, *Median of Ratios*, *Mean of Ratios* and *Regression Ratio*. ***GProcessor*** will prompt the user to choose one of these methods. The selected ratios are normalized with the **GenePix** calculated normalization factor according to the selected method. In fact, **GenePix** simply divides the calculated ratios by the geometric average value of the ratios of all the genes.

2. **Lowess fit Normalization**:
   If a user choose one of the first two ratio types, *i.e.*, *Ratio of Medians*, *Ratio of Means*, he/she will have an option to normalize the channel intensities and the ratios by using the Lowess fit method. Optionally, you may want to change some parameters used for Lowess Fit Normalization: f is used to specify the amount of smoothing; f is the fraction of points used to compute each fitted value; the larger the value of f, the smoother the fitted values become; choosing f between 0.2 and 0.8 should serve most purposes. If you have no idea which value to use, try f = 0.5. *nsteps* is the number of iterations in the robust fit; if *nsteps* = 0, the non-robust fit is returned; setting *nsteps* equal to 2 should serve most purposes. Since microarray data usually has very huge size, the default value for *nsteps* is set to be 3. *delta* is a non-negative parameter which may be used to save computations; if data size is less than 100, set *delta* equal to 0.0; Otherwise the default *delta* value is set to be 1% of the range of the abscissa points on the scatter plot.

3. **Ratios Output**:
   The ratios of the selected type for all the genes are given in the file *Ratios.txt*. If the ratio type is *Ratio of Medians* or *Ratio of Means*, then the ratios are normalized by Lowess fit method, otherwise, the ratios are normalized by the simple normalization method using the **GenePix** normalization factor according to the selected ratio.

4. **ANOVA Procedure**:
   Again if the user choose to use one of the first two ratio types, *i.e.*, *Ratio of Medians*, *Ratio of Means*, the ANOVA procedure is another

option for the user. Assuming that there is no dye effect or the dye effects have been adjusted by performing the Lowess normalization. The user can try and select any one of the following models:

- Model 1: $\log(y_{ikg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ikg}$;
- Model 2: $\log(y_{ikg}) = \mu + A_i + V_k + G_g + (VG)_{kg} + \epsilon_{ikg}$;

where y_{ikg} is the background-subtracted intensity of gene g in i-th array under k-th treatment, μ is the overall mean, A_i is the effect of i-th array, V_k the effect of k-th variety (treatment), G_g the effect of g-th gene, $(AG)_{ig}$ and $(VG)_{kg}$ are corresponding interactions, and ϵ_{ikg} is the random error.

If there are dye effects, Lowess Fit Normalization should be performed in order to do ANOVA. For the selected model, you are given the ANOVA table which contains the F-ratios and the corresponding P-values together with the multiple R^2. Once the user decide to use one of the models, the ANOVA results will be saved in a file *AnovaTable.txt*. This file contains the ANOVA table and the estimated ratio fold changes together with the P-values flag values. The flag value tell the user the number of "Bad" spots found in the replicates for each gene.

5. **Other data processing**:
GProcessor calculates the *Mean of Ratios*, *Median of Ratios*, *the Mean of log2 ratio* for each gene using all good replicated spots from different arrays. The values of the mean of ratios, the mean of log2-ratios and the median of ratios of merged data sets are given in the output file *Merge.txt*. CV's (coefficient of variation defined as the standard deviation divided by the mean) for all useful replicated spots was calculated and listed as a column in output file *Merge.txt*. Users can use this information to check the quality of their spots or even their slides. For each gene, the *P-value* of *T-statistic* is calculated based on all the good replicated spots. This P-value serves only as a reference of significance level of differential expression. In *Merge.txt* you can also find the Gene Number, Gene Name, ID, the mean of ratios for each data set, the standard deviations of overall mean ratio and overall mean log2-ratio for each gene, the number of spots actually used to calculate the overall merged mean and median of ratio, and the number

of outliers.

6. **Outliers detecting and Quality controlling**:
   A very simple outlier detecting algorithm was incorporated in ***GProcessor***. Those spots which leads the large difference between the mean of ratios and the median of ratios are defined as outliers and removed from the T-test procedure. The number of outliers for each gene is listed in *Merge.txt*. Moreover, for some spots the channel intensities after background subtraction may be less than 0, we eliminate these spots also from the overall ratio calculation. The number of spots actually used in ratio calculation for each gene is also listed in *Merge.txt*. Besides the flags given by the original GPR file which are 100 (Good), 0(Not flagged), -50(Not Found), -75(Absent) and -100(Bad), respectively, we introduce one more flag value -25(Not Good) to a feature-indicator(spot) in any of the following situations: (i) the ratio of intensity over background is less than constant C (user can choose C in a range from 1.0 to 5.0) for all the channels, (ii) the *Saturation* (the values given in the "F# % Sat." column of the GPR file). GPR is greater than 10% for any of the channels, (iii) the feature intensity if lower than the user specified lowest acceptable value.

7. **Ratio formula changing**:
   In some microarray experiments, dye exchange was conducted. In this case a user can define the *positive* dye configuration, then the dye configuration after exchange can be called *negative*. For those slides with *negative* dye configuration, a user can click *Option* button of **GenePix™ Pro 1.0** to change the ratio formula. For example, for a *positive* slides select W_1/W_2 and for a *positive* array select W_2/W_1. After changing the formula, click *Analysis* button of **GenePix™ Pro 3.0** to extract data from these *negative* slides, the data should be exported as *.gpr* files. The *.gpr* files obtained in this way from *negative* slides, together with those *.gpr* files from *positive* slides, can be used to ***GProcessor*** for pooling purposes.

   GProcessor will automatically detect the Ratio formulations and use this information to calculated corresponding ratios. If the ratio formulae in the headers of GPR files are incorrect, you can either use **GenePix™ Pro** to modify it or input the ratio formulae manually.

8. **Correlation Coefficients**:

In output file *CorrCoef.txt*, different kinds of correlation coefficients are summarized in table formats. These include the correlation coefficients both for intensity and for ratio measurements across different data sets or across different slides. Users can use these information to examine and compare their data.

9. **Different version GPR files**:
   We noticed that the header format of a GPR file of **GenePix™ Pro** 3.0.0.x-3.0.5.x is different from that of **GenePix™ Pro** 3.0.6.x. Furthermore, **GenePix™ Pro** 4.0 GPR file not only has different header but also has more data columns than the previous versions since it can deal with up to four colors. This version (2.0a) of ***GProcessor*** can automatically detect the different versions of GPR files and treat them differently. So the users can use all the GPR files available as input of ***GProcessor***. Of course, the user is responsible for making sure that all the GPR files of different versions use the same gene list file.

Usage

1. Save the executable file *gprocessor* in a directory. If necessary, you may have to change the mode of this file so that it is executable. Although it is not necessary, we recommend that you put the executable file *gprocessor* and all the .gpr files to be processed in the same directory.

2. In the directory where the executable is saved, type *./gprocessor*, then the program will be activated.

3. Enter the proper parameters when prompted. The total computation takes only few seconds.

4. ***GProcessor*** has several output files: "AnovaResult.txt", "CorrCoef.xls", "MergeData.txt", "Cluster.txt", "T-test.txt", "SamDataPaired.xls" and "SamData1Class.xls". All these files are in tab delimited text format.

Remark

- Always use **original** GPR files for ***GProcessor***.
  Although the current version (2.0a) also works for GPR file which have

been opened and re-saved by programs like **MS Excel**, we recommend not to open and re-save the GPR file with **MS Excel** before using them for ***GProcessor***, since **Microsoft® Excel** will change the format of GPR file.

Release and Version

The current version is ***GProcessor*** 2.0a, which was released on July 8, 2003. ***GProcessor*** was designed for the convenience of YMD(Yale Microarray Database) and HHMI Biopolymer/W.M. Keck Foundation Biotechnology Resource Laboratory at Yale University users to process replicated microarray data sets obtained from **GenePix™ Pro** image analysis software. No warranty is expressed or implied.

The ***GProcessor*** 2.0a was released as a package in zip format which can be downloaded from http://zhao.med.yale.edu/. The package includes ***GProcessor*** executable program, a *User's Manual* in PDF format, two replicated data sets *p1.gpr* and *p2.gpr*. User can use these two files to test the program, in each data set there are 1 replicated spots.

The ***GProcessor*** 2.0a was tested for original GPR file generated from GenePix™ Pro 3.0.0.98, 3.0.5.56, 3.0.6.52,3.0.6.73 and 4.0 (the most recent version). We believe that it works for all GenePix™ pro versions range from 3.0.0.x to 3.0.6.x and 4.0. If you have GPR files which don't work with ***GProcessor*** 2.0a, please send email to zhong.guan@yale.edu and attach your GPR files for the test purpose.

Contacts

Please send bugs report and source code request to zhong.guan@yale.edu. For comments, suggestions and critics please contact hongyu.zhao@yale.edu, kenneth.williams@yale.edu, janet.hager@yale.edu and zhong.guan@yale.edu

References

References for Lowess Normalization:

1. Yang, Y. H. et al., Normalization for cDNA Microarray Data SPIE BiOS 2001, San Jose, California, January 2001.

2. Dudoit, S. et al., Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments; Technical Report #578, Aug. 2000, Dept. of Statistics, UC Berkeley

3. Tseng, G. C. et al., Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variation and Assessment of Gene Effects. Neucleic Acids Research, 2001, Vol.29, No 12, 2549-1557.

Reference for ANOVA method:

4. Kerr, M. K. et al., Analysis of Variance for Gene Expression Microarray Data, J. Comput. Biol., 7, 819-837.

Last Update

July 8, 2003.