Appendix E

MATA: TOOLBOX FOR ANALYZING MICROARRAY DATA

E.1 Introduction

To facilitate the use of the probabilistic framework, a computational tool for analyzing microarray data (MATA) has been developed. This toolbox is a collection of files developed using the MATLAB® language. The following section describes step-by-step how to use MATA on real microarray data (Affymetrix GeneChip®).

E.2 Step-by-step example

E.2.1 Installing the toolbox

First unzip the file Mata.zip (available at www.che.udel.edu/systems/people/cgelmi) to an empty folder in the work directory of MATLAB®. All the analyses will be carried out in this folder and the results will be saved in it.

E.2.2 Preparing the microarray data

Using EXCEL, or another similar program, prepare a text file (*.txt) for the analysis. The file should contain three columns: i) the first column should have the intensity of the genes under the “treated” condition, ii) the second column should have the intensity of the genes under the “reference” or control condition, and iii) the third column should contain a tag or reference number for each gene. This column is

important because it will later help match the intensities with the original gene names. If MATA does not find a third column, it will automatically assign a tag number to each gene using correlative numbers (*i.e.*, 1,2,…,N). The EXCEL file however should contain four columns: the first one should have the genes' names (or ID) and the remaining ones should contain the same information as the text file.

It is important for the data file to be located in the same folder where the toolbox is saved.

E.2.3 Running MATA

In the command *window* of MATLAB® type *groupstag*, and the following splash window will appear:

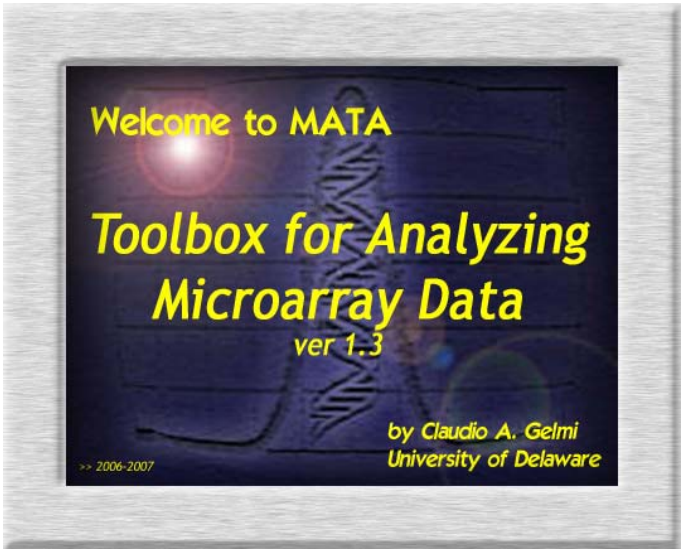

Figure E.1 First screen displayed after writing *groupstag* in the *command window*.

Once the splash window disappears, Figure E.2 will appear asking the user to choose the data file to be analyzed. In this example, the target file is called *f37lR1.txt* and it is located in the folder *example*.

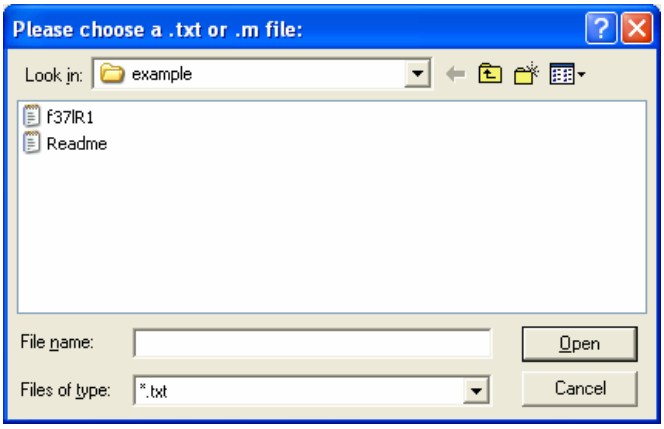

Figure E.2 Menu for selecting the experimental data file.

E.2.4 Basic statistics and diagnostic of the microarray data

After the data file is chosen and the **open** option selected, the first stage of analysis begins. This stage consists of the calculation of some basic statistics followed by a graphical exploratory analysis of the fractional intensity microarray data. A typical display of the *command window* at this point should look like:

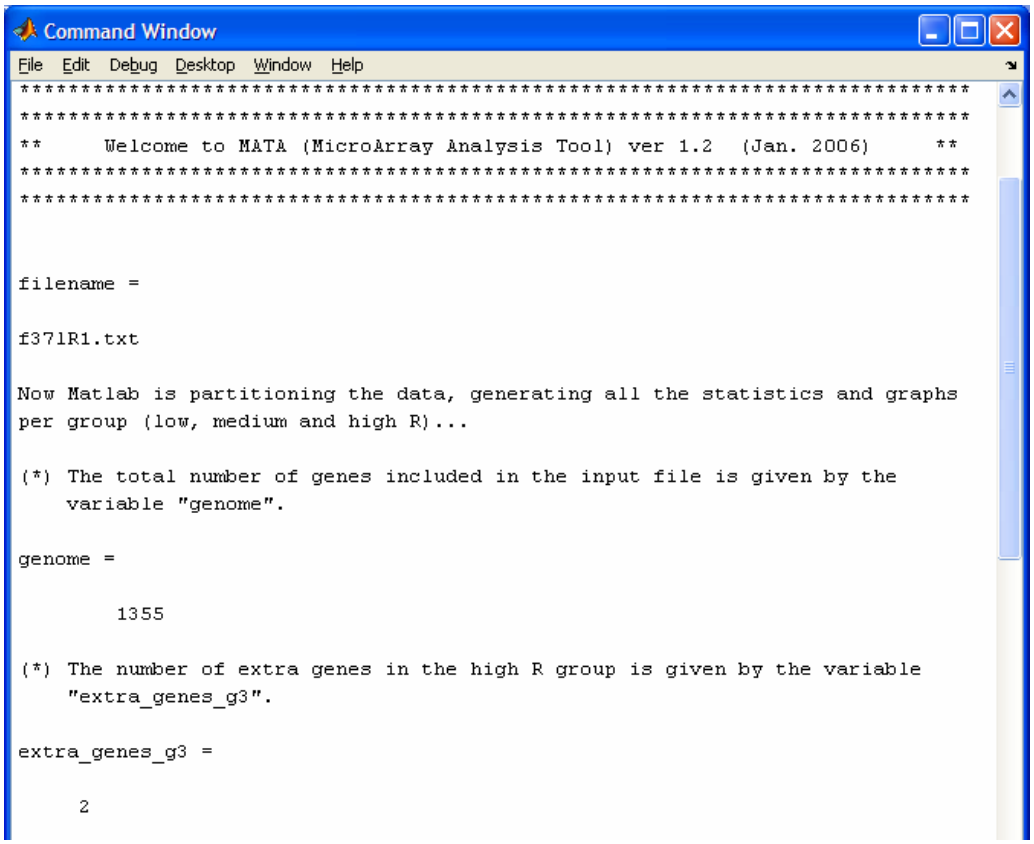

Figure E.3 A typical display of the *command window* after selecting the experimental data (Part 1).

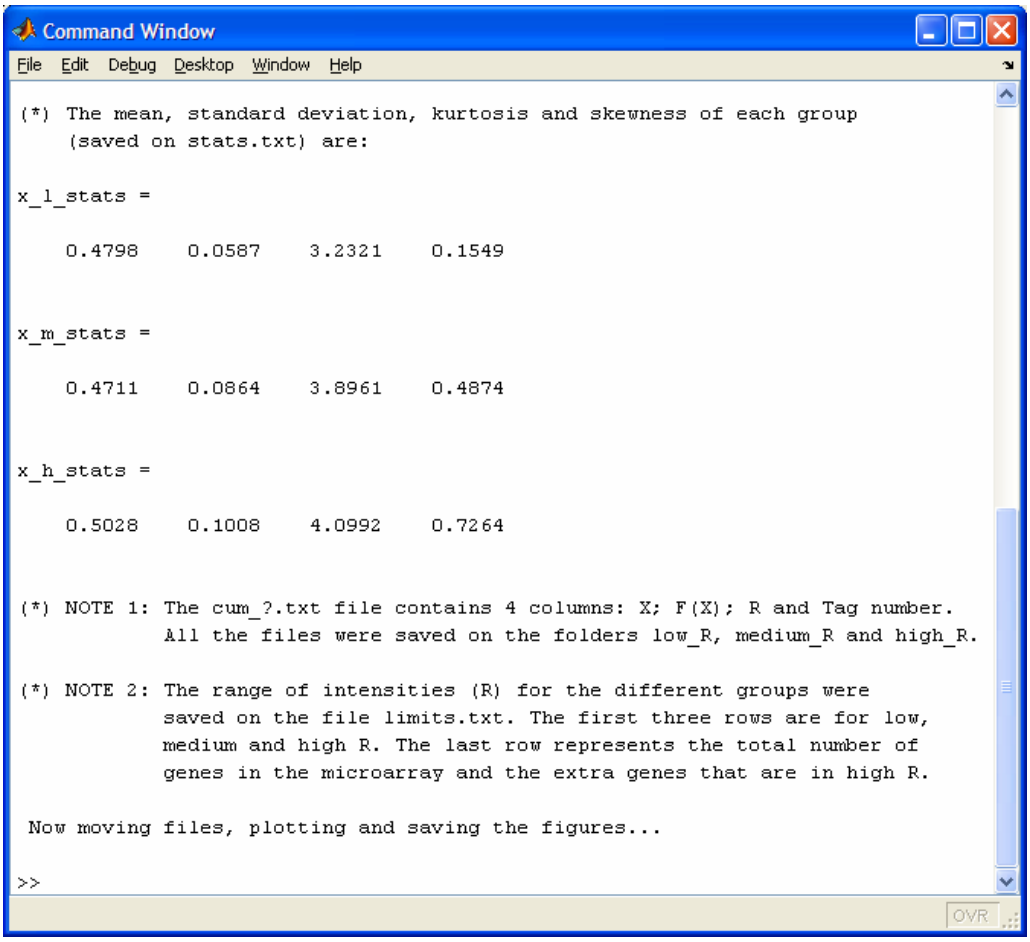

Figure E.4 A typical display of the *command window* after selecting the experimental data (Part 2).

All the displayed information of the fractional intensity data will be saved in two text files:

- *stats.txt* = contains the mean, standard deviation, kurtosis and skewness of the intensity magnitude groups: Low, Medium and High *R* groups;
- *limits.txt* = contains the intensity range of each group along with the total number of analyzed genes.

Both text files are located in the Figure_stats subfolder.

After the statistics are shown on the screen, the following figures are displayed:

1. three histograms of the fractional intensities (Low, Medium, and High *R* groups), and one containing all the fractional intensities;
2. histogram of the individual intensity signal channels (treated and control);
3. log-log intensity scatter plot (treated versus control);
4. MA plot. An MA plot is a 45 degree rotation (with a re-scaling of the abscissa) of the scatter plot of the intensities. The MA compares the log-ratio of two expression intensities versus the mean log-expression of the two, *i.e.*, $M = \log_2(\text{treated}/\text{control})$ versus $A = 0.5 \cdot \log_2(\text{treated} \cdot \text{control})$. MA plots are useful for identifying spots artifacts and for normalization purposes.

All these figures are saved automatically in the Figure_stats subfolder. For this example, the fractional intensities histograms are as follows:

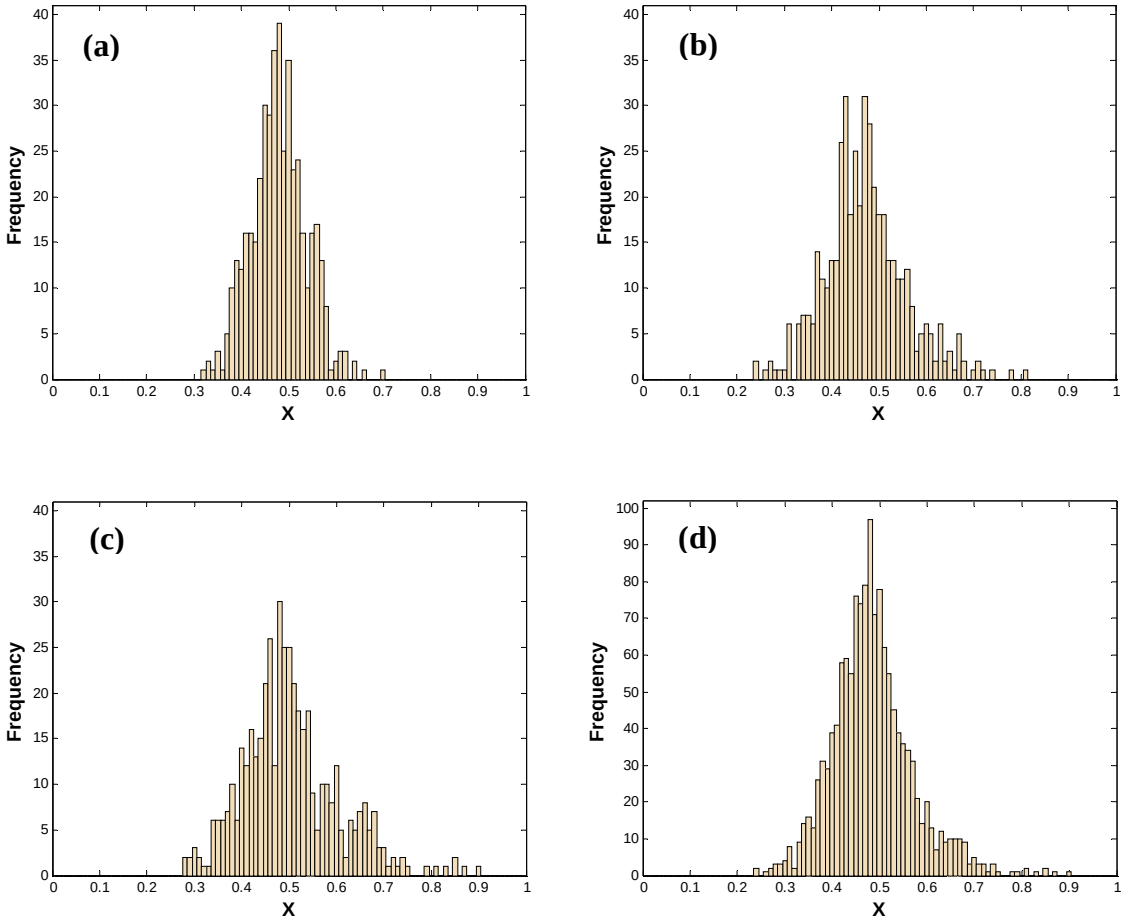

Figure E.5 Histograms of groups: (a) Low *R*, (b) Medium *R*, (c) High *R* and (d) All the fractional intensity data.

Figure E.6 shows a histogram of the individual intensity signal channels:

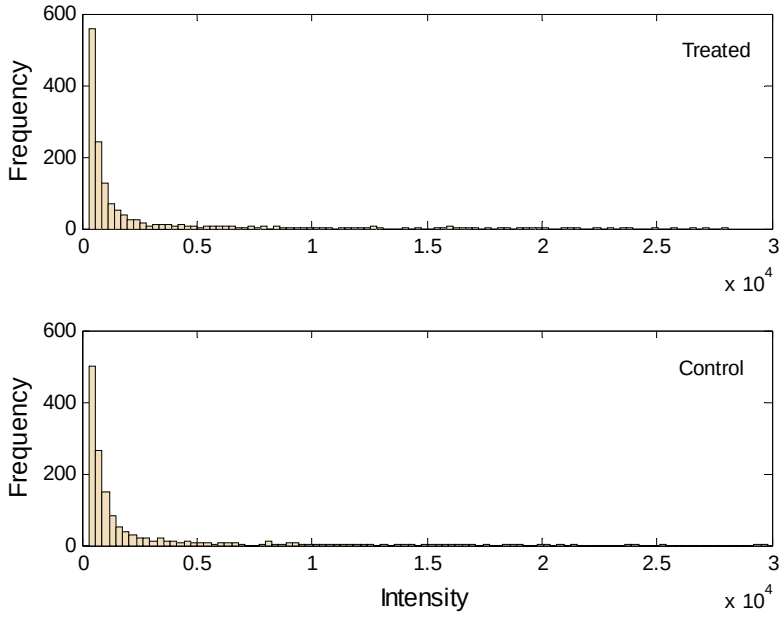

Figure E.6 Histogram of the individual intensity signal channels.

Figure E.7 and Figure E.8 show the log-log intensity scatter plot (treated versus control) and the MA plot of this case study:

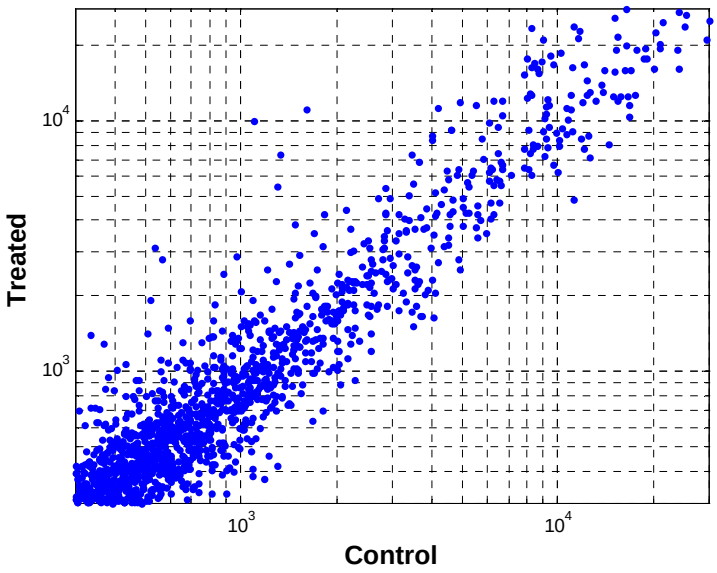

Figure E.7 Log-log intensity scatter plot of treated versus control intensity signals.

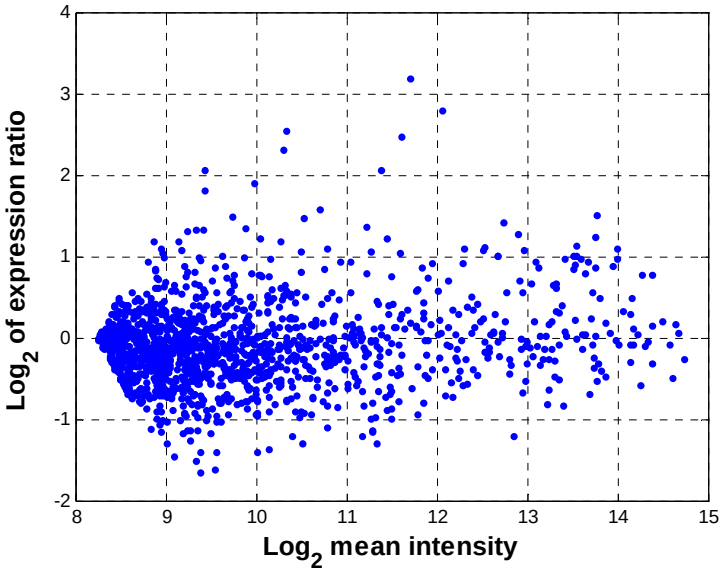

Figure E.8 MA plot.

After all the statistics are displayed and figures are plotted, the folder *example* should look like:

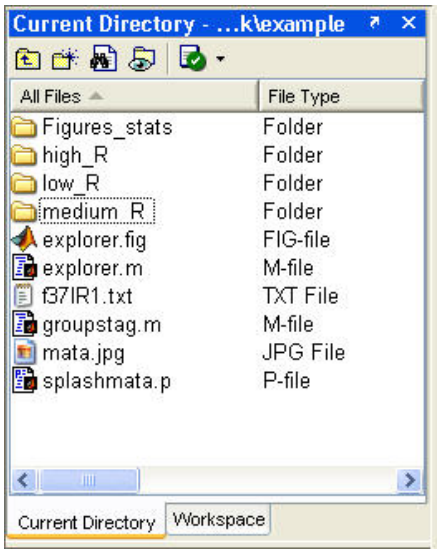

Figure E.9 Files and folders created by MATA after displaying the diagnostic figures.

Three new subfolders have been added to the root folder *example*: low_R, medium_R and high_R. This is because the data has been divided into three groups (Low, Medium and High *R* groups), each according to its total intensity magnitude (for more details see Section 2.1.2.1). In each of these folders all the necessary files have been added automatically in order to perform the Beta distribution fitting.

E.2.5 Fitting the mixture model to the experimental data: Low R group (Part I)

The second stage of analysis begins with the graphical user interface (GUI) **explorer**. The GUI allows the user to choose the best initial guess for the parameters of the Beta distributions. This step increases the chances of MATLAB's

optimization algorithms of finding a global minimum. The *explorer* GUI is shown below:

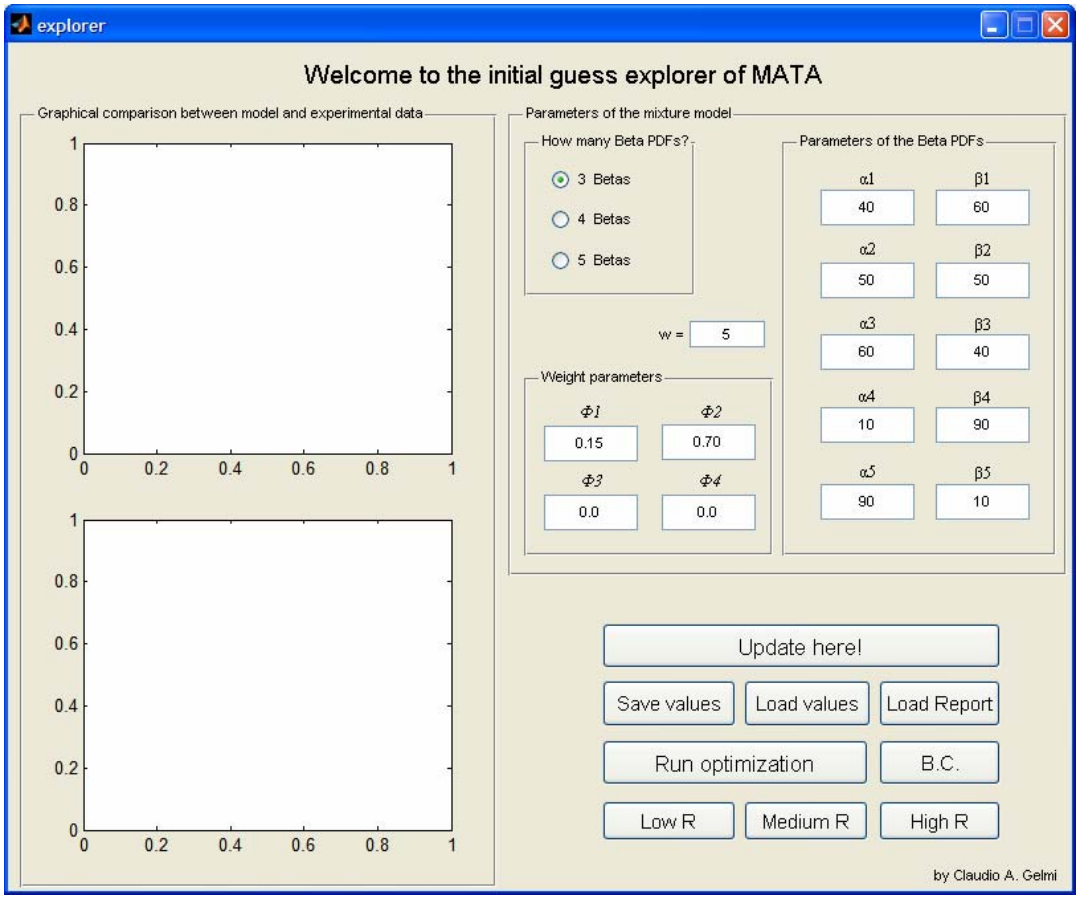

Figure E.10 A view of the GUI explorer of MATA.

To proceed with the fitting of the Beta distributions, the user must choose one of the three folders (low_R, medium_R, high_R) by selecting the respective button on the GUI. In the example under consideration, after choosing the Low *R* folder and selecting the **Update here!** option (note that the GUI opens with default

values to help the user with the selection of the parameters) the GUI looks like Figure E.11:

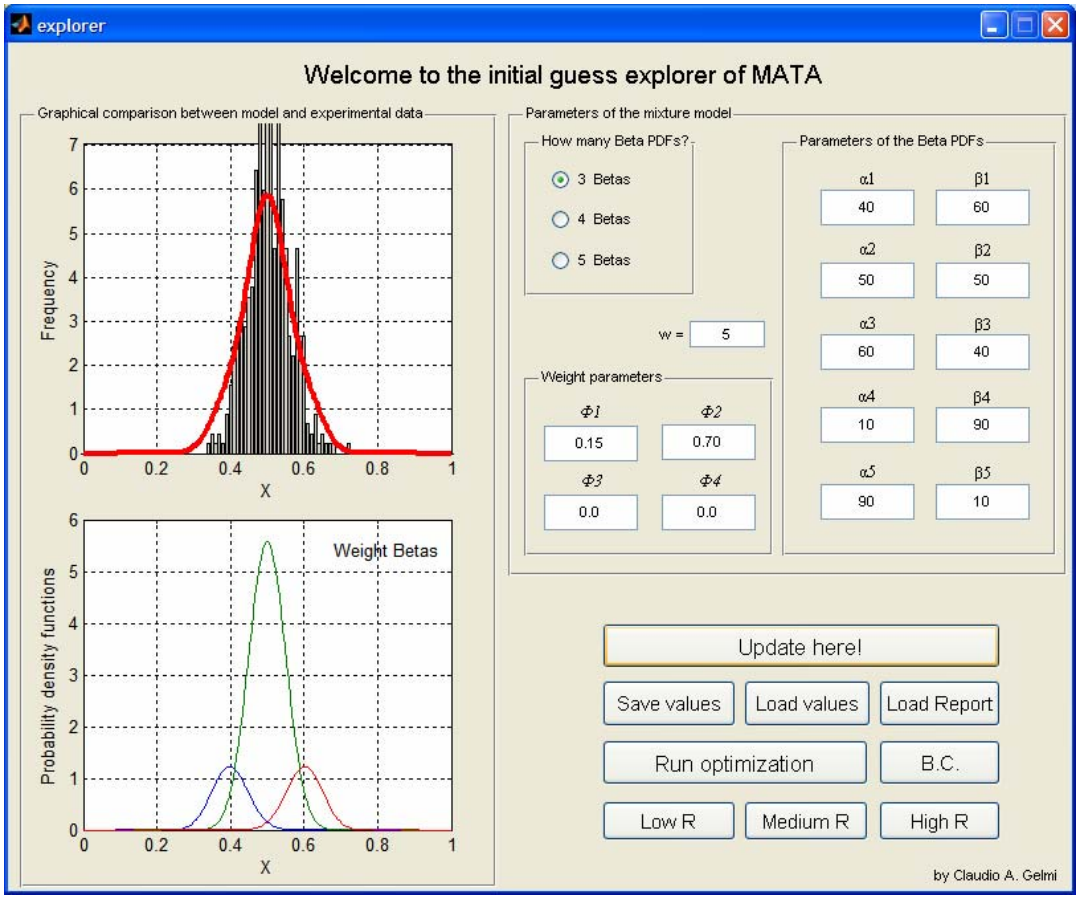

Figure E.11 A view of the initial guess explorer of MATA, after the Update here! option is selected.

In order to obtain a better match between the model and the experimental data (histogram and red curve) the user can change the parameters of the Beta distributions changing the parameters (α_i, β_i). As a guideline to position the distributions, the user can use the fact that the mean of a Beta distribution is given by

$\mu = \alpha/(\alpha+\beta)$. In the study under consideration, after some manual iteration with the parameters, the fitting improves considerably:

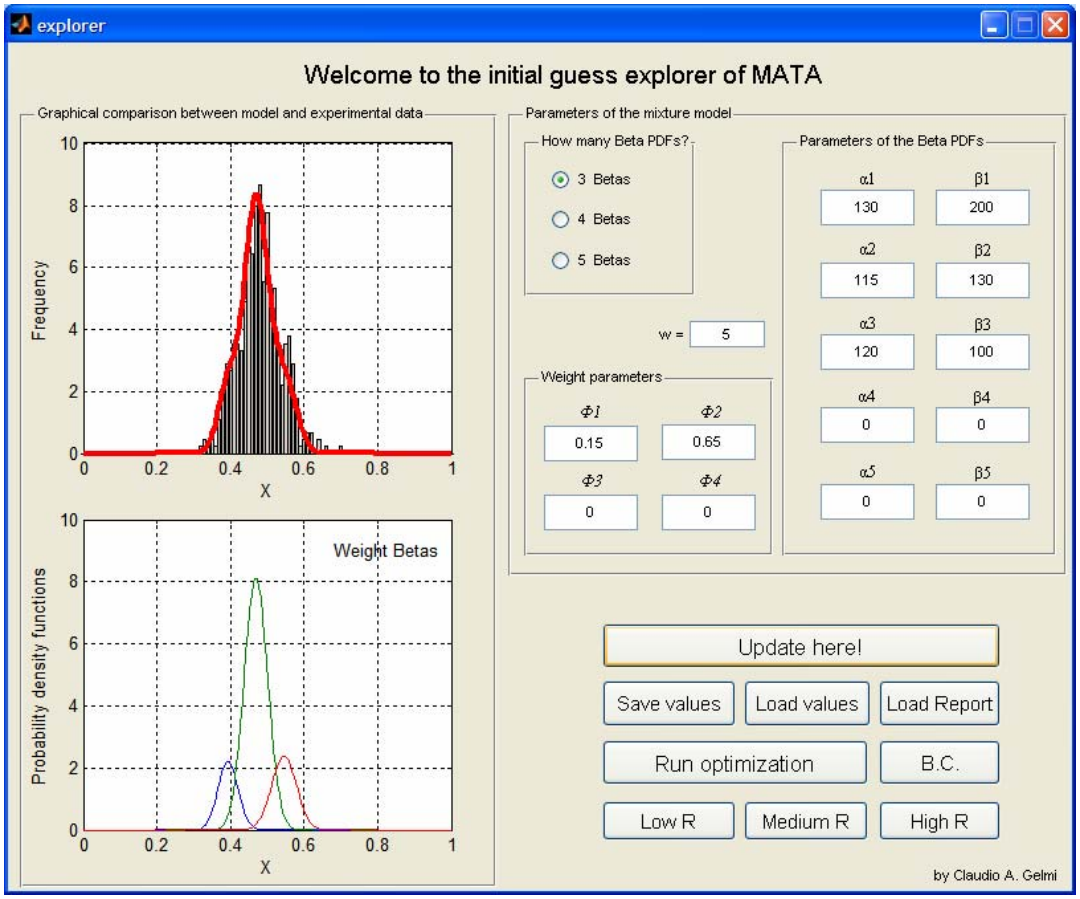

Figure E.12 A view of the initial guess explorer after several manual iterations.

Once the user is satisfied with the new set of parameters, the values can be saved by selecting the **Save values** option (all the GUI values are saved in the init_guess.txt file). From here the user can now allow MATLAB® to find the best set of parameters (Φ_i, α_i and β_i). MATLAB® will perform a two-stage optimization: the

first one uses a random search algorithm and the second a *fminsearch* routine. Note that the *w* parameter in the GUI corresponds to an empirical weight factor that is set (by trial and error) such that the sum of the squared residuals and the sum of the variances of the Beta distributions have the same weight in the objective function. This point later will be discussed later.

Since three Beta distributions are fitted in this example, the weight parameter Φ_3 is already specified as $1-(\Phi_1+\Phi_2)$. This relationship is not reflected in the GUI; it is known internally. The same applies when four ($\Phi_4 = 1-(\Phi_1+\Phi_2+\Phi_3)$) or five Beta distributions ($\Phi_5 = 1-(\Phi_1+\Phi_2+\Phi_3+\Phi_4)$) are fitted.

E.2.6 Fitting the mixture model to the experimental data: Low R group (Part II)

To initiate the two-step optimization, the user has to select **Run optimization** option. After performing several minutes of calculations, MATLAB® displays the following text in the *command window*:

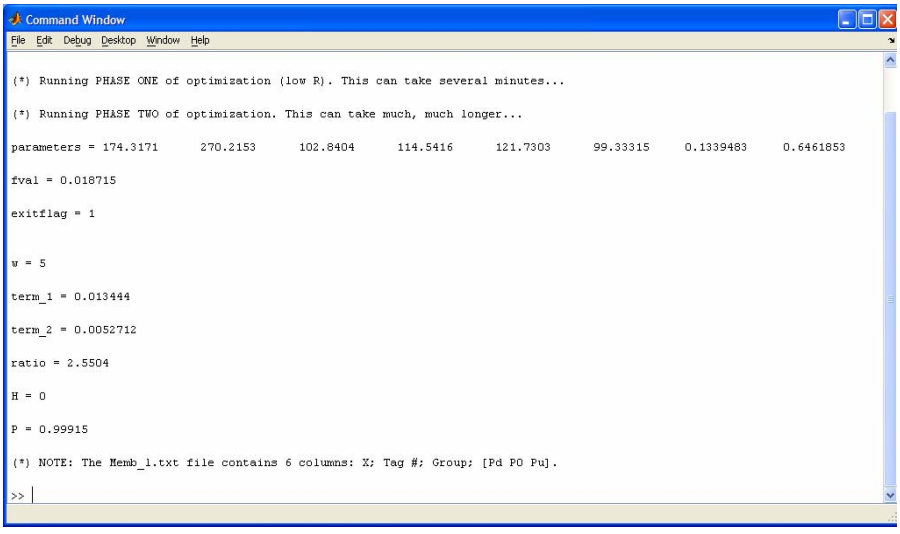

Figure E.13 Typical display after completing the two-step optimization.

The information displayed in Figure E.13 includes:

- *parameters* = the optimum value of the parameters (α_1, β_1; α_2, β_2, α_n, β_n and Φ_1, Φ_2, …Φ_n) found by MATLAB®. These values are also saved in the file *fvalues.txt*;
- *fval* = the final value of the objective function;
- *exitflag* = describes the exit condition of the optimization algorithm (exitflag = 1, the algorithm converged, otherwise exitflag ≠ 1);
- *w* = weight parameter used in the optimization;
- *term_1* = sum of square errors between the ECD and the model;
- *term_2* = sum of the variances of the Beta distributions multiplied by *w*;
- *ratio* = the ratio between *term_1* and *term_2* (the ratio should be close to 1.0);
- *H* = Kolmogorov-Smirnov (K-S) test to determine if the empirical data could have a hypothesized, continuous cumulative distribution function (H = 0, do not reject the null hypothesis at significance level α; H = 1, reject the null hypothesis at significance level α);
- *P* = observed p-value for the parameters (high p-values can be interpreted as strong evidence that the parameters found by optimization belong to the empirical cumulative distribution).

All the information described above is saved in the text file *Report_?.txt* (? = l, m or h).

Figure E.14 shows the mixture model fitted by MATLAB® and how it compares with the experimental data.

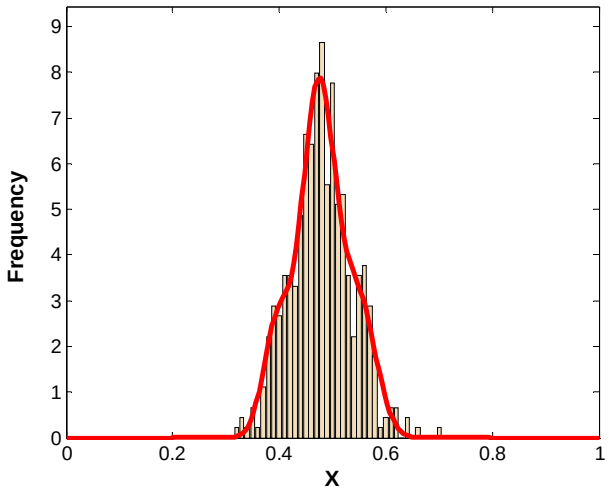

Figure E.14 Comparison between the experimental data and model (Low *R* group).

After the optimization information is shown, a series of plots are displayed, including: i) the error as a function of the number of iterations (provides some indication of the performance of the random search algorithm), ii) the Beta PDFs not weight by the ϕs ("pure Betas" in Figure E.15), iii) the Beta PDFs weight by their respective ϕ ("weight Betas" in Figure E.15), iv) a comparison between histogram and contributing Betas, comparison between cumulative distribution functions, and v) a membership index plot (*a posteriori* probabilities that the genes are up, down or not differentially expressed). The probabilities are saved in the file *memb_?.txt*. This file contains six columns:

1. the fractional intensity of gene *i*;
2. group number (Low *R* = 1; Medium *R* = 2; High *R* = 3);
3. probability of gene *i* of being down-regulated (P_{down});
4. probability of gene *i* of being not differentially expressed (P_{non});

5. probability of gene *i* of being up-regulated (P_{up});
6. a tag or reference number.

In addition, a new GUI is displayed, allowing the user to save the desired figures and files mentioned previously. Here is an example of how to save one of the histograms and the text files *Membership* and *Optim. report*:

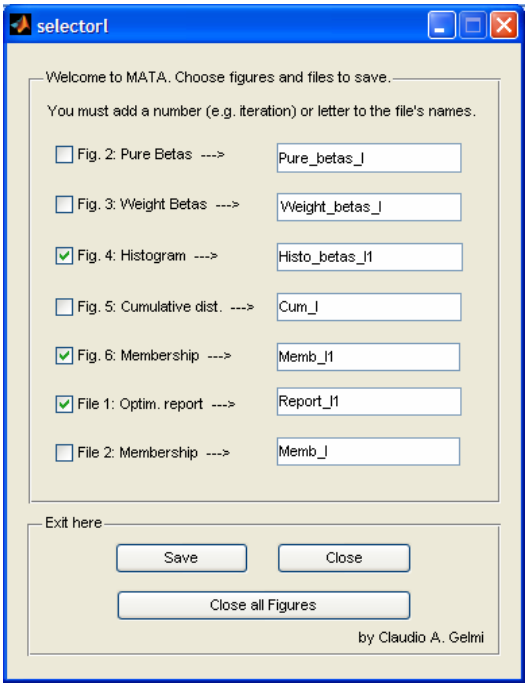

Figure E.15 Menu to assist users to save figures and files.

From the last *command window* (Figure E.13) the ratio between term_1 and term_2 was 2.55. In order to approach the desired 1.0 value, increase the initial *w* parameter (*selector* GUI) to a value of 14 (if the ratio between *term_1* and *term_2* is greater than 1, try increasing *w*; if the ratio is smaller than 1, try decreasing *w*). Once the Report has been loaded and the option of new *w* has been made, the user selects

the **Run optimization** button to proceed with the nonlinear fit. Every time this option is selected, the initial values are saved and can be recovered by selecting the **Load values** option. In order to load the last optimum parameters found by MATLAB®, the user must select the **Load Report** option.

Once a good set of parameters has been found (ratio near 1.0 and a high p-value), the user must check if dye bias correction is needed. In order to check this, the user must select the **B.C.** option on the *explorer* GUI and follow the instructions. For more information, refer to the **Dye Bias Correction** section of this chapter.

With the dye bias correction complete, the analysis of the folder low_R is finished. Analyzing data in the medium_R and high_R folders must still be completed.

E.2.7 How to fit the mixture model to the Medium and High *R* groups

To analyze the Medium and High *R* groups, the procedure that was followed in sections E.2.5 and E.2.6 above should be used. However, it is advisable to use as an initial guess in the *explorer* GUI the parameters found in the previous group (*e.g.*, use as initial values those found in Low *R* group to fit Betas in the Medium *R* group; or use the values found in Medium *R* group to fit Betas in the High *R* group). To use these parameters the user must select the **Load Report** option and move to the next folder selecting the **Medium *R*** button (remember that this example started on the low_R folder); then the user must select **Update here!** If the initial guess is not effective, the user must manually change the parameters until a better fit is found. Then the user must select **Run Optimization** to initiate the optimization.

E.2.8 Consolidating probabilities, confidence index and more

Once the user is satisfied with the fitting and the dye bias correction has been performed (or was not necessary), the computed probabilities of expression status must be matched with the gene probes analyzed via assigned tag numbers. In the original EXCEL file used to prepare the text file supplied to *groupstag* (*.txt), there are four columns of data: i) gene name, ii) intensity of the genes under the "treated" condition, iii) intensity of the genes under the "reference" or control condition, and iv) tag number. The *importool* function within MATA will use this Excel file, the text files generated during the fitting and optimization stages of *groupstag* (*memb_l.txt*, *memb_m.txt*, and *memb_h.txt*), and *Rep_analysis_Matlab.xls* (the user can rename the file to a simpler one), provided in the toolbox, to match and sort the data.

After ensuring that MATLAB® is accessing the MATA directory, type *importool* into the command window. A graphical user interface will appear asking the user to choose how many files to import. As mentioned above, three files were generated during the fitting and optimization stages of *groupstag*, making the default response three. Then, MATLAB® will ask the user to select the three files to use (*memb_l.txt*, *memb_m.txt*, and *memb_h.txt*) one by one. Next, *importool* will request the original EXCEL file described above containing the four columns of raw data. Once selected, MATLAB® will confirm that the user wants to continue with the analysis and upon choosing **YES**, it will request the user's version of *Rep_analysis_Matlab.xls* in which to report the results. MATLAB® will provide an opportunity to import other sorted text files from a previous analysis. The user can then select **NO** if the original three files are the only ones to be analyzed in this run. The program will then execute, providing updates throughout the matching, analysis, and saving processes. Once MATLAB® reports that it is finished with "Done!" in the

command window, the user can access the results on the **Summary** sheet of the user's version of *Rep_analysis_Matlab.xls*.

A typical analysis saved by *importool* in the **Summary** worksheet looks like:

Table E.1 Typical output of the function *repgenes* (Summary worksheet).

# Rep	Probe Set Name	Av. intensity grouping	Fractional intensity (X)			Down-regulated (P_{down})			Not differentially expressed			Up-regulated (P_{up})			c_i
			Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	
3	DRC0040	3.0	0.372	0.378	0.386	0.457	0.529	0.574	0.426	0.471	0.543	0.000	0.000	0.000	0.90
3	DRC0038	3.0	0.530	0.538	0.548	0.013	0.016	0.019	0.980	0.982	0.983	0.001	0.003	0.004	1.00
3	DRC0037	3.0	0.472	0.493	0.507	0.031	0.045	0.068	0.932	0.955	0.969	0.000	0.000	0.000	0.97
3	DRC0036	3.0	0.486	0.517	0.534	0.017	0.028	0.049	0.951	0.971	0.981	0.000	0.001	0.002	0.97
3	DRC0034	3.0	0.473	0.476	0.479	0.058	0.062	0.066	0.934	0.937	0.941	0.000	0.000	0.000	0.99
3	DRC0033	3.0	0.430	0.433	0.439	0.150	0.173	0.186	0.814	0.827	0.850	0.000	0.000	0.000	0.97
3	DRC0032	2.3	0.474	0.486	0.494	0.039	0.073	0.136	0.864	0.926	0.959	0.000	0.001	0.002	0.92
3	DRC0029	3.0	0.496	0.501	0.505	0.032	0.035	0.039	0.961	0.964	0.968	0.000	0.000	0.000	0.99
3	DRC0025	2.0	0.467	0.483	0.513	0.013	0.132	0.203	0.796	0.865	0.980	0.000	0.002	0.007	0.85
3	DRC0020	3.0	0.336	0.349	0.357	0.696	0.748	0.826	0.174	0.252	0.304	0.000	0.000	0.000	0.89
3	DRC0013	1.7	0.485	0.532	0.557	0.001	0.034	0.101	0.830	0.861	0.899	0.000	0.105	0.169	0.88
3	DRC0012	1.0	0.483	0.520	0.562	0.008	0.051	0.112	0.888	0.933	0.966	0.000	0.016	0.048	0.92
3	DRC0010	1.3	0.462	0.476	0.489	0.088	0.163	0.266	0.734	0.837	0.912	0.000	0.000	0.000	0.85
3	DRC0006	3.0	0.415	0.428	0.448	0.120	0.204	0.260	0.740	0.796	0.880	0.000	0.000	0.000	0.89
3	DRC0005	3.0	0.429	0.448	0.459	0.092	0.127	0.189	0.811	0.873	0.908	0.000	0.000	0.000	0.92
3	DRC0004	3.0	0.335	0.435	0.493	0.042	0.312	0.833	0.167	0.688	0.958	0.000	0.000	0.000	0.35
3	DRC0003	3.0	0.331	0.340	0.351	0.734	0.802	0.855	0.145	0.198	0.266	0.000	0.000	0.000	0.90
3	DRC0002	3.0	0.348	0.355	0.365	0.634	0.707	0.756	0.244	0.293	0.366	0.000	0.000	0.000	0.90
3	DRC0001	3.0	0.333	0.345	0.352	0.733	0.772	0.845	0.155	0.228	0.267	0.000	0.000	0.000	0.91
3	DRB0143	2.7	0.626	0.654	0.679	0.000	0.000	0.001	0.004	0.184	0.389	0.610	0.816	0.996	0.69
3	DRB0141	2.3	0.515	0.519	0.521	0.009	0.014	0.022	0.977	0.979	0.981	0.001	0.006	0.010	0.99
3	DRB0139	3.0	0.482	0.487	0.494	0.042	0.049	0.055	0.945	0.951	0.958	0.000	0.000	0.000	0.99

In Table E.1 the first and second columns show the number of replicates available for each gene as well as the gene's name or ID. The third column is the average intensity group (Low R = 1, Medium R = 2, and High R = 3). The next columns show the fractional intensity (x) and the probabilities of expression status for the minimum, maximum and average values. The last column is the confidence index, a measure of the degree of confidence in the calculated probabilities.

E.3 Dye bias correction

As explained in Section 2.1.2.4, if the mean of the Beta distribution that represents the genes that are not differentially expressed is other than 0.5 the user must correct for dye bias. To correct for bias, select **B.C.** in the *explorer* GUI. MATA will confirm this step with the user. If **YES** is selected, the files *cum_?.txt* and *fvalues.txt* are automatically backed up as *cum_?old.txt* and *fvalues_old.txt*.

Figure E.16 depicts the text box that appears when **B.C.** is selected:

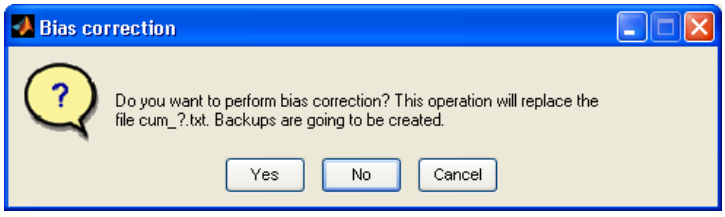

Figure E.16 Bias correction confirmation box.

If the user chooses **YES**, MATA asks for the Beta distribution in the *explorer* GUI that represents the group of genes not differentially expressed. In this case, the second Beta distribution (alpha2, beta2) is the selected one.

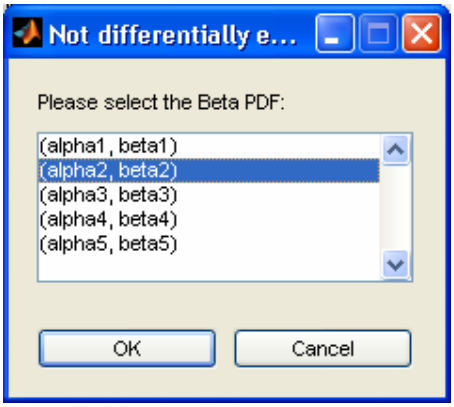

Figure E.17 Menu for selecting the "not differentially expressed" Beta distribution.

Once the parameters are chosen and **OK** is selected, MATLAB® applies the bias correction algorithm and recalculates all the parameters by calling the optimization routine. After the optimization is completed, the final parameters are recovered by selecting **Load Report**, and the figures and files are saved by following the instructions in Section E.2.6.

In Figure E.18, notice that after loading the Report (the **Load Report** option), the parameters alpha2 and beta2 are almost identical. This implies that the mean of the Beta distribution is 0.5, and therefore the dye bias was corrected successfully. In this figure, the middle distribution (bottom graph in Figure E.18) is correctly centered around 0.5 as opposed to the middle distribution seen in Figure E.12. This is another indication that dye bias was necessary and that it was performed successfully.

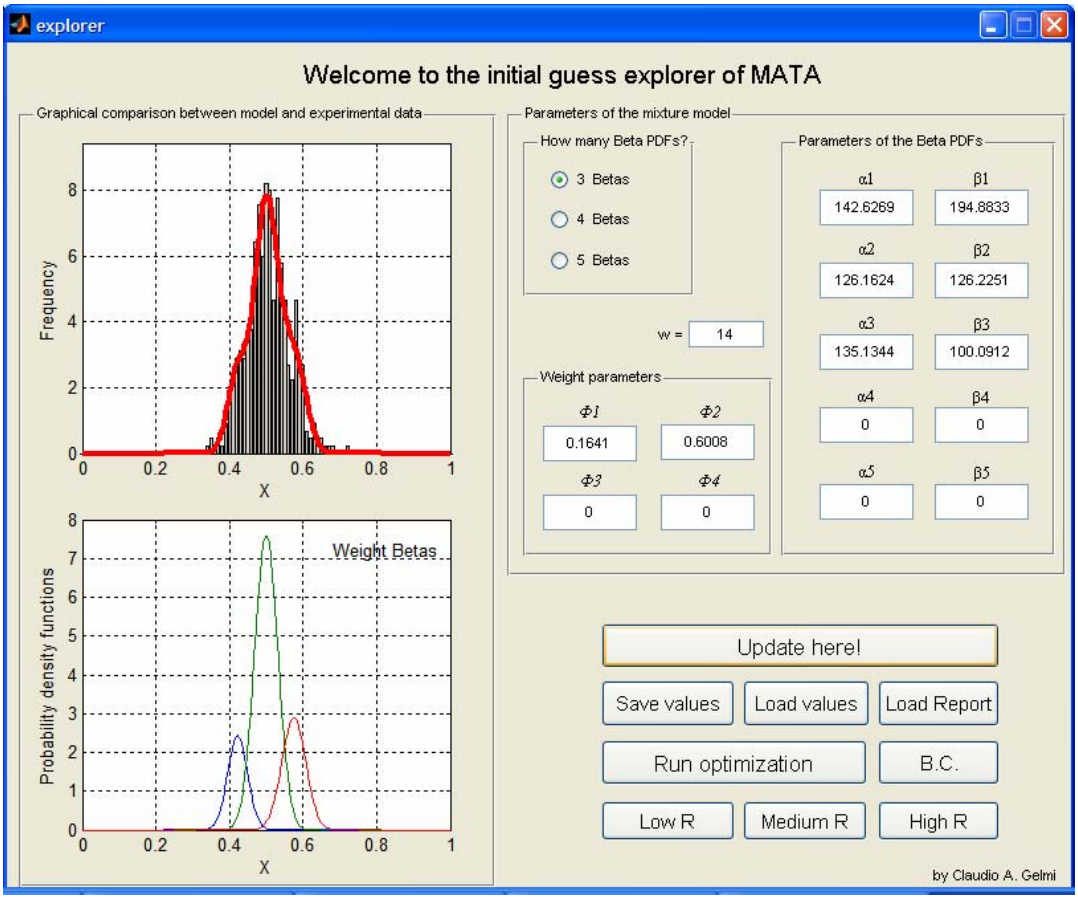

Figure E.18 Screen view of the initial guess explorer after performing dye bias correction.

If MATA detects that the data does not require dye bias correction, the user will see the following message:

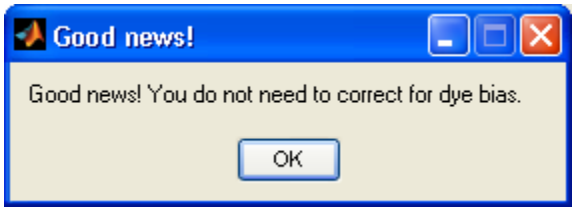

Figure E.19 Text box reporting that dye bias correction is not necessary to perform.