PerfectMatch Manual

Version: 2.3

Date updated: Oct 29, 2004

A PC program for Affymetrix microarray data analysis using PDNN model

Li Zhang, Haitao Zhao, James Mitchell & Clift Norris

Department of Biostatistics and Applied Mathematics

The University of Texas MD Anderson Cancer Center

1515 Holcombe Blvd, Box 447

Houston, TX 77030

Table of contents

Section 1. Introduction, installation and usage

PerfectMatch program is designed to use PDNN model for analyzing Affymetrix microarray data. This model assumes two modes of binding on the oligonucleotide arrays: gene specific binding and non-specific binding (cross hybridization). For each probe, the model gives an estimate of two binding energies, one for gene specific binding, and the other for nonspecific binding. Probe binding energy is computed as a weighted sum of stacking energies of nearest-neighbor nucleotides, where the weights depend on the position along the probe. Using the affinity values, the model then estimates gene expression levels through matching the observed probe signals and model-fitted values. For more details of the method, see the manuscript published by Li Zhang *et al.* on Nature Biotechnology, 2003 Jul; 21(7): 818-21. The manuscript is also included in the ***PerfectMtach*** package.

To install the program, unzip the *PM.zip* file and run the *setup.exe* file provided in the package. The program is developed on PC Windows operating system. It does not require user to restart the computer after installation. Note that the package also includes a set of files stored in "data" directory that contains examples of input files needed to run the ***PerfectMatch*** program.

To use the program to analysis Affymetrix arrays, please follow the standard procedure below:

1. Normalization
2. Optimize parameters
3. Estimate gene expression
4. View genes
5. View Cell image
6. Scatter plot

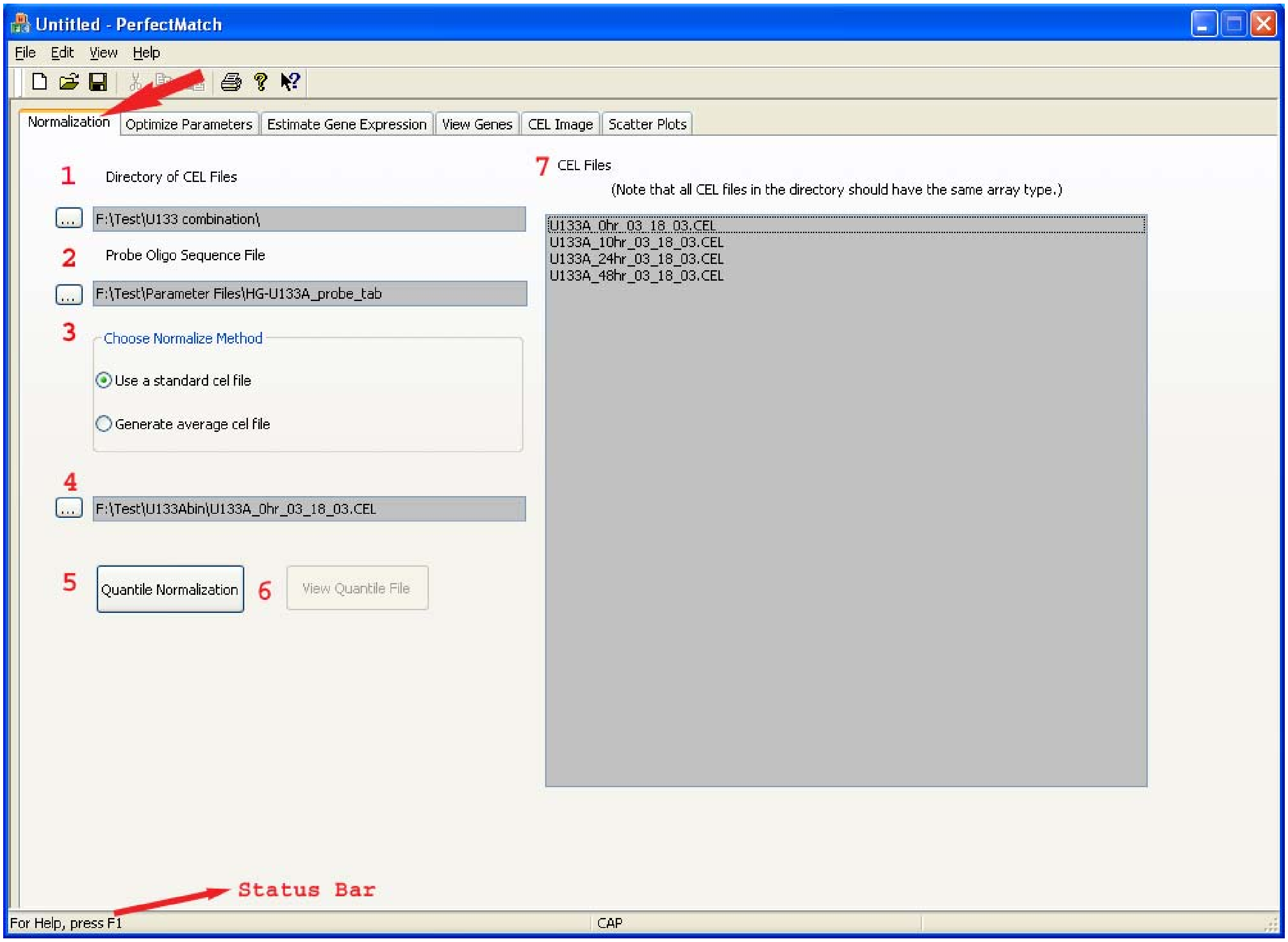
Please note Normalization and Optimize parameters are prerequisites for Estimate gene expression, but their order of execution can change, View genes, View cell image and view scatter plot have to perform after Estimate gene expression is finished.

Affymetrix cel file format:

Recently affymetrix had introduced a new binary file format, the version 2.3 of PerfectMatch can deal with this new format as well as the old text format. For information about affymetrix cell file format please refer to **http://www.affymetrix.com/support/developer/index.affx** in COMMUNITY RESOURCES->File Formats

Section 2. Using the *Normalization* tab to normalize data

From version 2.3 , a new ***Normalization*** tab is introduced dedicate to normalization.



Directory of CEL Files

F:\Test\U133 combination\

Probe Oligo Sequence File

F:\Test\Parameter Files\HG-U133A_probe_tab

Choose Normalize Method

● Use a standard cel file

○ Generate average cel file

F:\Test\U133Abin\U133A_0hr_03_18_03.CEL

CEL Files

(Note that all CEL files in the directory should have the same array type.)

```
U133A_0hr_03_18_03.CEL
U133A_10hr_03_18_03.CEL
U133A_24hr_03_18_03.CEL
U133A_48hr_03_18_03.CEL
```

Specifying input files

Click on **Directory of Cel Files** (1) browser to specify a directory that contain all the *.CEL files to be processed. Note that the program expect that all the CEL files in the directory have the same array type. All the CEL files under the directory will be shown on the left panel under "CEL Files" (7).

Click on **Probe Oligo Sequence File** browser (2) to specify a file containing the probe sequence information of the array. The file can be obtained from Affymetrix website download center. Be sure to obtain the file in tabular format. You can also download it from out website http://odin.mdacc.tmc.edu/~zhangli/PerfectMatch/

Choose Normalization Method (3) **:** click one of the 2 radio buttons to choose normalization method.

- **Use a Standard Cel File** The program will use this file in a quantile normalization procedure. After normalization, all CEL files in the directory will be rescaled to have the exact same PM probe signal intensity distribution as the standard file and the normalized files will be stored in a binary format with ".binCEL" file extension. Please use (4) browser to specify the standard cell file.

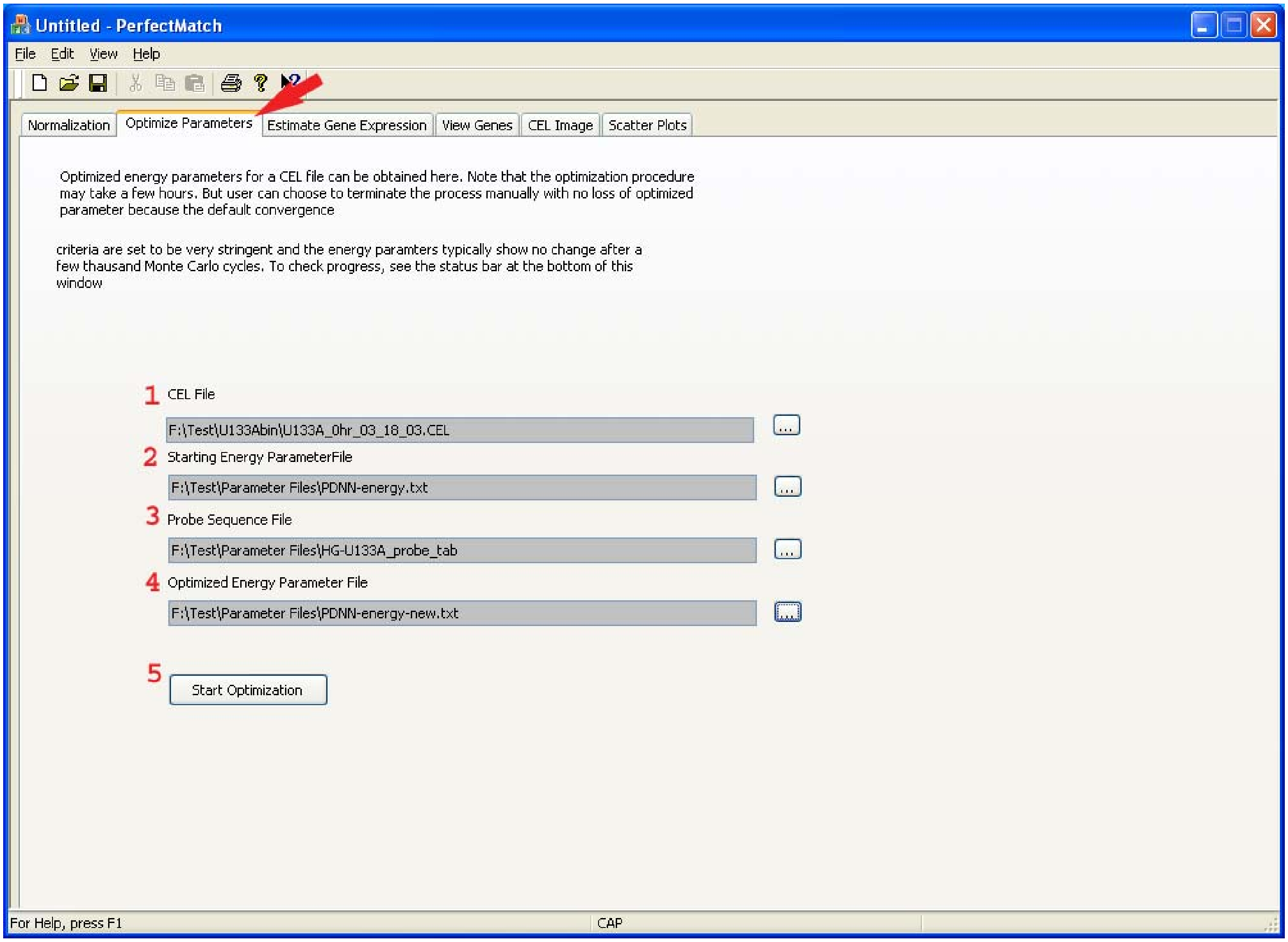
- **Generate average cel file:** The program will sort each cell file and add all the intensity together with the same rank, get the average , produce a average cell file and use it to normalize all the data. Please use (4) browser to specify the output directory and name of the average cell file.

After verifying all the input fields values are correct click on **Quantile Normalization**(5) button to begin normalization, after normalization the **View Quantile File**(6) button will be enabled. The Quantile file record the 2% 25% 50% 75% 98% of sorted intensity readings for each cel file , by comparing quantile,we hope to identify chips with significant errors.

During program execution the **status bar** display the progress .

Section 3. Using the *Optimize Parameter* tab to optimize energy parameters

Optimize parameters tab interface is designed for optimizing the probe binding energy parameters for a particular Cel file using PDNN model. It is recommended that user should use this procedure to obtain the optimized energy parameters for the standard Cel file chosen for quantile normalization.



Optimized energy parameters for a CEL file can be obtained here. Note that the optimization procedure may take a few hours. But user can choose to terminate the process manually with no loss of optimized parameter because the default convergence

criteria are set to be very stringent and the energy paramters typically show no change after a few thausand Monte Carlo cycles. To check progress, see the status bar at the bottom of this window

1 CEL File: F:\Test\U133Abin\U133A_0hr_03_18_03.CEL

2 Starting Energy ParameterFile: F:\Test\Parameter Files\PDNN-energy.txt

3 Probe Sequence File: F:\Test\Parameter Files\HG-U133A_probe_tab

4 Optimized Energy Parameter File: F:\Test\Parameter Files\PDNN-energy-new.txt

Specifying input files:

- Click on **Cel File** (1) browser to specify a CEL files to be processed.
- Click on **Starting Energy parameter File** browser (2) to specify a probe binding energy parameter file. Examples of such files can be found in the "data" directory included in this package.

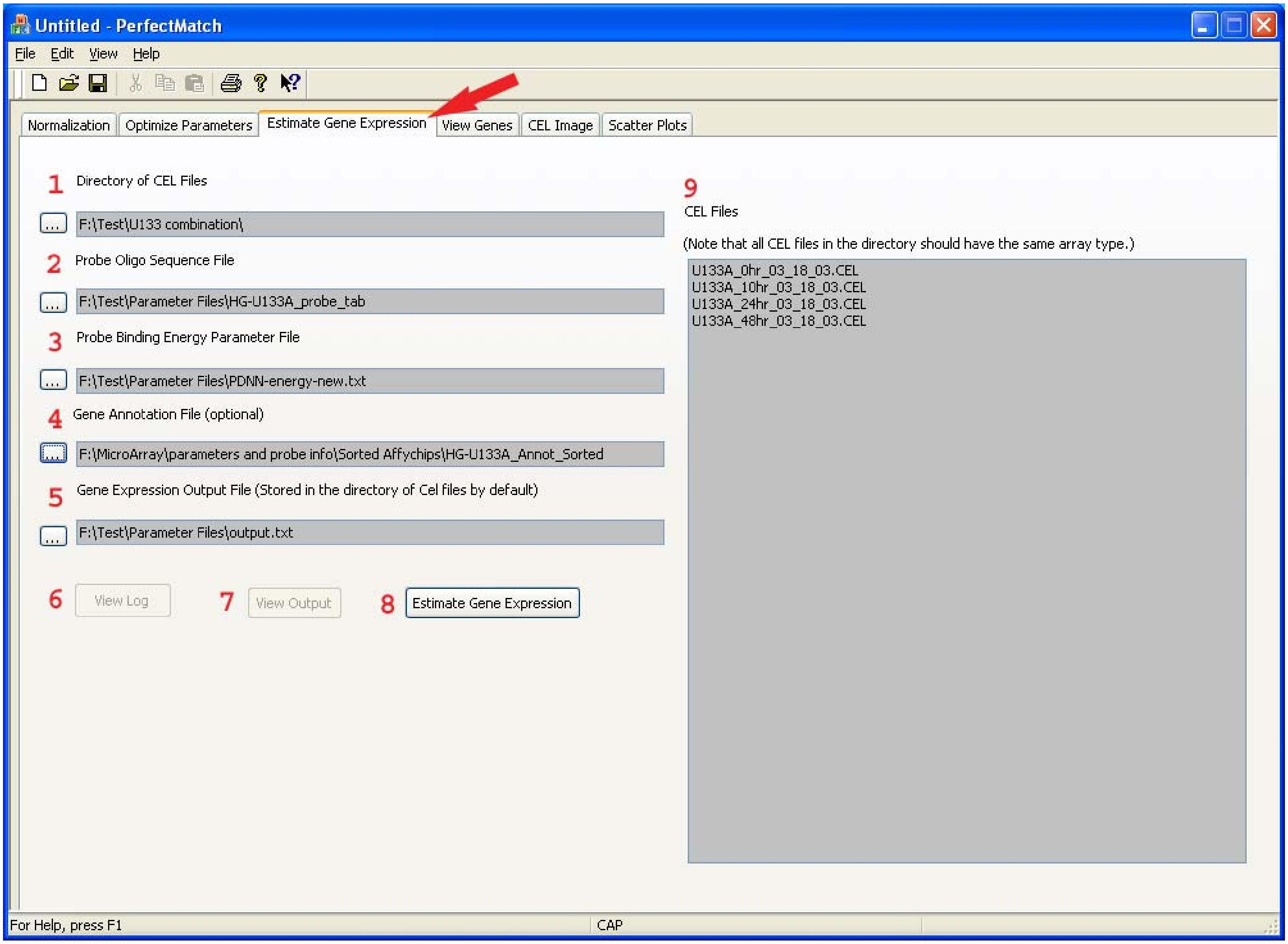
- Click on **Probe Oligo Sequence File** browser (3) to specify a file containing the probe sequence information of the array. The file can be obtained from Affymetrix website download center. Be sure to obtain the file in tabular format.
- Click on **Optimize Parameter File** browser (4) to store optimized parameters.

After verifying all the inputs click **Start Optimization** to begin optimization.

Note: The optimization procedure may take a few hours to converge. But user can choose to terminate the program early if the fitness level shows little progress (check the status bar, which is updated every 100 steps of Monte Carlo cycles.) The program updates the file every 100 Monte Carlo cycles. Therefore, terminating the program early will not lose the optimized energy parameter file.

Section 4. Using the *Estimate Gene Expression* tab to specify input files and estimate gene expression

"***Estimate Gene Expression***" tab interface is designed to specify input files and compute gene expression values in multiple samples stored in a directory.



1 Directory of CEL Files

F:\Test\U133 combination\

2 Probe Oligo Sequence File

F:\Test\Parameter Files\HG-U133A_probe_tab

3 Probe Binding Energy Parameter File

F:\Test\Parameter Files\PDNN-energy-new.txt

4 Gene Annotation File (optional)

F:\MicroArray\parameters and probe info\Sorted Affychips\HG-U133A_Annot_Sorted

5 Gene Expression Output File (Stored in the directory of Cel files by default)

F:\Test\Parameter Files\output.txt

9 CEL Files

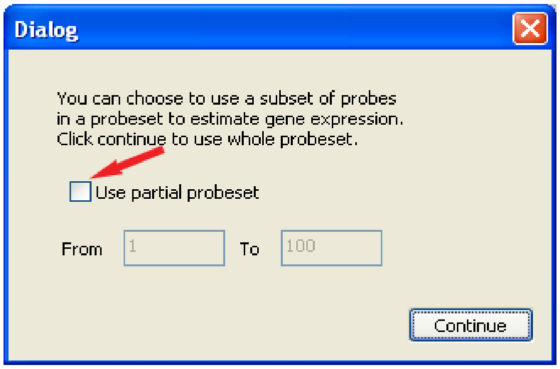
(Note that all CEL files in the directory should have the same array type.)

```
U133A_0hr_03_18_03.CEL
U133A_10hr_03_18_03.CEL
U133A_24hr_03_18_03.CEL
U133A_48hr_03_18_03.CEL
```

Specifying input files

- Click on **Directory of Cel Files** (1) browser to specify a directory that contain all the *.CEL files to be processed. Note that the program expect that all the CEL files in the directory have the same array type. All the CEL files under the directory will be shown on the left panel under "CEL Files"(9).
- Click on **Probe Oligo Sequence File** browser (2) to specify a file containing the probe sequence information of the array. The file can be obtained from Affymetrix website download center. Be sure to obtain the file in tabular format.
- Click on **Probe Binding Energy parameter File** browser (3) to specify a probe binding energy parameter file. Examples of such files can be found in the "data" directory included in this package. The parameter file needs to be optimized for the standard CEL file for best results. Please see directions under " **Optimize Parameters"** tab for how to perform this task.
- Click on **Gene Annotation** browser (4) to specify a file containing functional annotation of the probesets in the array. Examples of such files can be obtained from Affymetrix website. The file should be in tab-delimited format and each line should start with a probeset name. This file optional. Annotation will be included in the output file when specified.
- Click on **Gene Expression Output File** browser (5) to specify a file to store output.

After specifying the input files, user can click "**Estimate Gene Expression**" button (8).

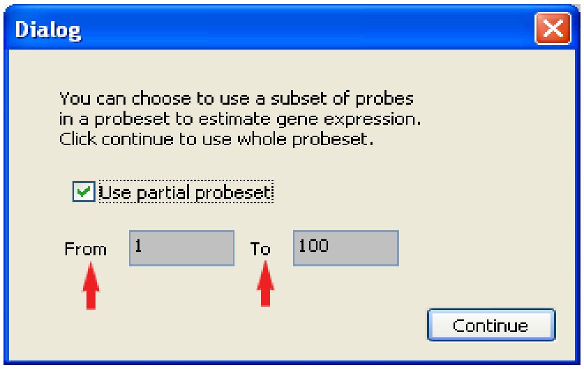


You can choose to use a subset of probes in a probeset to estimate gene expression. Click continue to use whole probeset.

□ Use partial probeset

From 1 To 100

A message box will popup to let you decide whether you want to process full index or just part of the sequence, if you want process full index, leave the **Use partial probeset** unchecked (default) and click **continue**, otherwise check the **Use partial probeset ,** select the from index and to index click on continue, the calculation will begin.



You can choose to use a subset of probes in a probeset to estimate gene expression. Click continue to use whole probeset.

☑ Use partial probeset

From 1 To 100

The "status bar" at the bottom will show the Calculation progress. For each Cel file, four steps will be shown in the status bar: reading Cel file, normalizing data, estimating expression and updating summary output file.

After the calculation is done, status bar will show “Calculation complete” and user can click "View Output"(7) or “View Log"(6) buttons to inspect the output and log files.

Description of the output files

The output file will be produced by the program in a spreadsheet style, with each row representing a probeset (gene) and each column representing a sample. The output file shall contain gene expression levels of all arrays in the CEL file directory. The expression level values are represented on natural logarithm scale. The gene annotations are listed along with the expression levels. Besides this spreadsheet file, the program also automatically generates a log file “**PDNN.log**” and a *.**pdn** files for each associated CEL file. Below is fraction of an example *.pdn file:

File_ID	Probeset	lnN0	lnN-C1.CEL	err_T	corr	P_size	crossPM	avg_Affynity
C1.CEL	1007_s_at	12.3	12.288	0.701	0.9941	16	0.112	0.0043
C1.CEL	1053_at	12.5	12.454	0.83	0.9886	15	0.169	0.0014
C1.CEL	117_at	11.6	11.559	0.554	0.9943	14	0.245	0.0019
C1.CEL	121_at	11.3	11.342	0.307	0.9988	15	0.376	0.0054
C1.CEL	1255_g_at	11.5	11.48	0.806	0.9867	16	0.048	0.0017
C1.CEL	1294_at	11.3	11.314	0.453	0.9963	15	0.233	0.0028
C1.CEL	1316_at	11.9	11.866	0.574	0.9944	16	0.058	0.0033

Note: These columns provide information for evaluating probe performance. Note that it is the third column that is used for gene expression profiling. File_ID is the CEL file associated with this .pdn file; Probeset is the probeset (gene) name; LnN0 is gene expression level on natural log scale before excluding outliers (this column can be ignored); LnN is gene expression level on natural log scale; err_T defines goodness of fit between the model and the observed data of the probeset; corr is correlation coefficient between the observed and model fitted lnPM signals; P_size is the number of probes used in the model fitting; cross_PM is the estimated ratio of non-specific binding signal vs. total probe signal; avg_affynity is average gene specific binding affinity of the probes in a probeset.

PDNN.log file records information for quality controls purposes. Below is an example Log file:

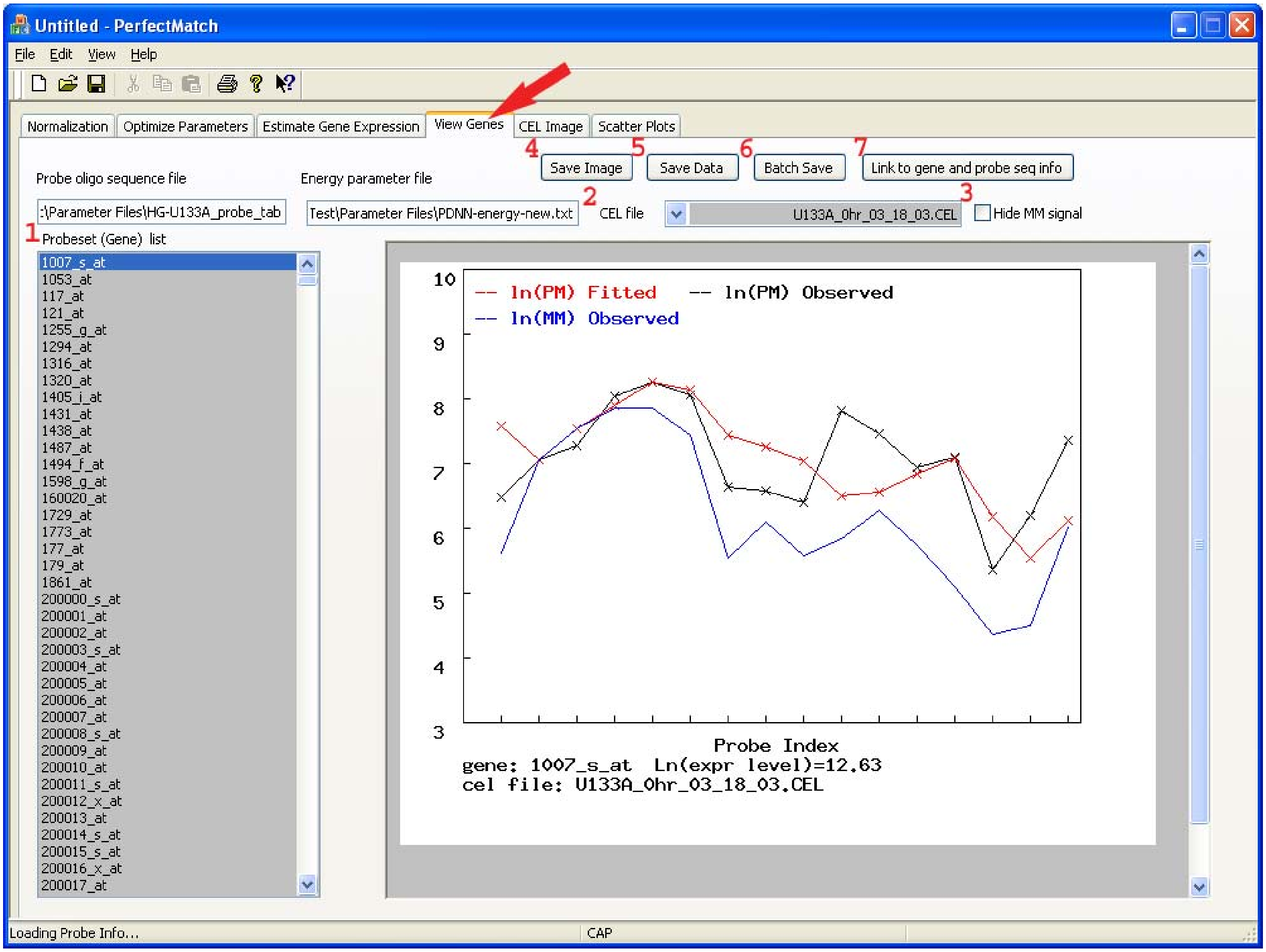
Summary	Num genes	Cross-hyb-const	Background	Fitness	Absent genes	ScalingFactor
C1.CEL	22283	5114.7	192.671	1.303	0	1

Note: “ScalingFactor” and “Absent genes” are not properly computed in the current release. Please ignore the values.

Section 5. Using the *View Genes* tab to inspect probe level data

“***View Genes***” Tab interface is designed to show detailed probe level data along with the model fitted probe signals.

User needs to check if **Probe Oligo Sequence File** and **Energy parameter File** are correctly specified on “Estimate Gene Expression” tab.



Probe oligo sequence file :\Parameter Files\HG-U133A_probe_tab

Energy parameter file Test\Parameter Files\PDNN-energy-new.txt

CEL file U133A_0hr_03_18_03.CEL

□ Hide MM signal

gene: 1007_s_at Ln(expr level)=12.63
cel file: U133A_0hr_03_18_03.CEL

To select or change a gene to be shown, user can select a probeset (gene) from the genes list(1) in the left panel. Probe level data of the chosen probeset in the will be shown in right panel. To change sample, select a CEL file from the CEL file drop down list (2), program will display the predicted data (using PDNN) and the observe data in the display window. MM probe data can be hidden when “Hide MM signal” check box (3) is checked.

Saving data and images

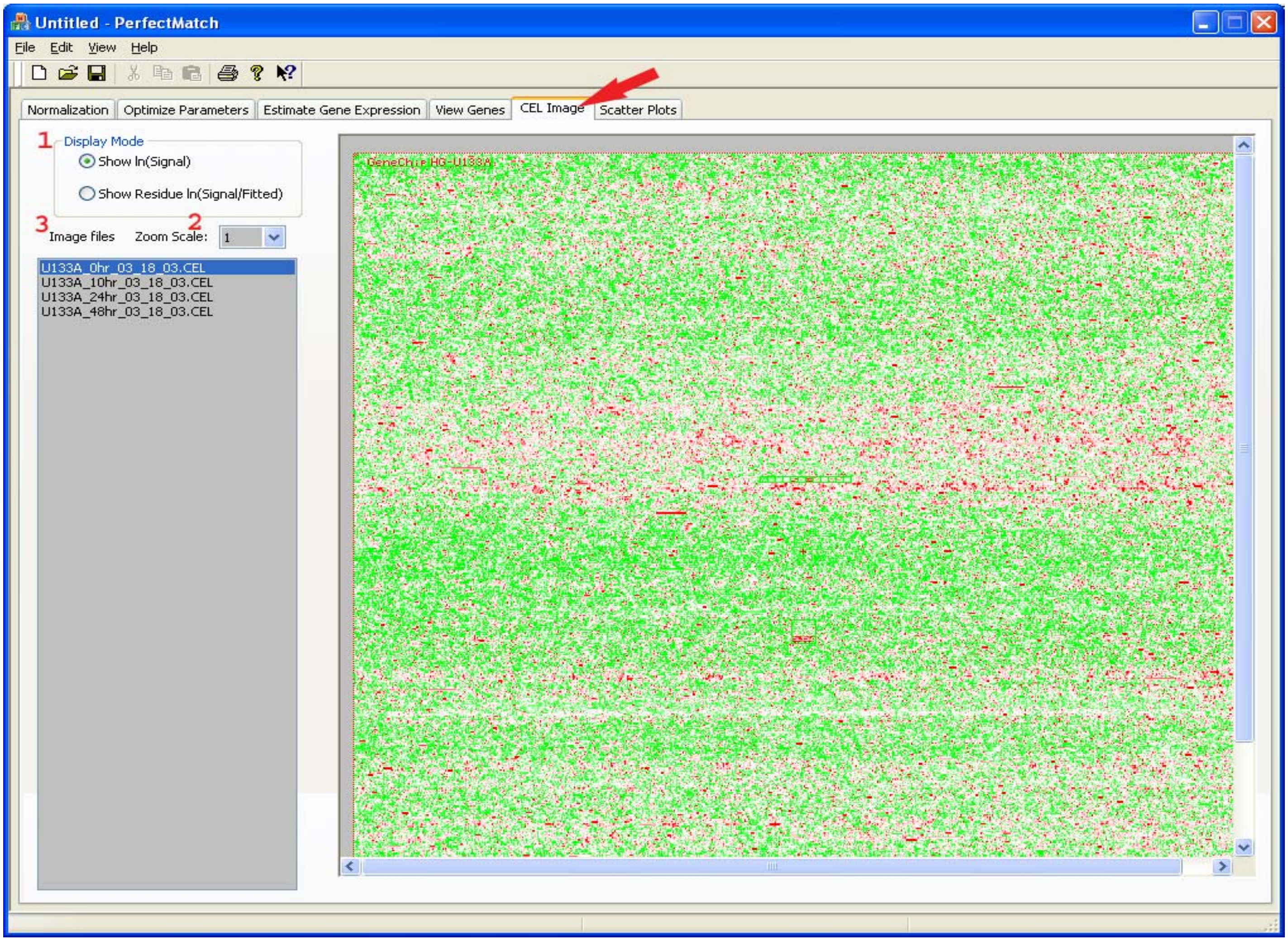
Clicking “Save Image” (4)button will save the graphical image of probe level data in a PNG formatted file. The file name is automatically generated by concatenating the Cel file name and the probeset name. The numerical data shown in the figure can be saved in a text-delimited file when “Save Data” (5)button is clicked. The file name is generated same as the image file but with “TXT” suffix. Clicking “Batch save” (6) button saves the numerical data in all the samples in Cel file list for the same probeset.

Link to information of gene functional information and probeset design.

Clicking “Link to gene and probe seq info” (7)button opens a new browser and direct user to Affymetrix website. The Affymetrix web site provides detailed probe design information as well as functional annotation of the gene. User needs to register on the website to get a login name and password. PerfectMatch program only asks the user to specify user login name and password once.

Section 6. Using the *Cel Image* tab to inspect array images

“Cell Image” Tab interface is designed to show observed microarray data and compare with the model fitted probe signals. Please note there are four fields in this window: Display Mode, Zoom Scale, Image files, and a demo window for displaying graphics.



Display Mode
● Show ln(Signal)
○ Show Residue ln(Signal/Fitted)

Image files

Zoom Scale: 1

U133A_0hr_03_18_03.CEL
U133A_10hr_03_18_03.CEL
U133A_24hr_03_18_03.CEL
U133A_48hr_03_18_03.CEL

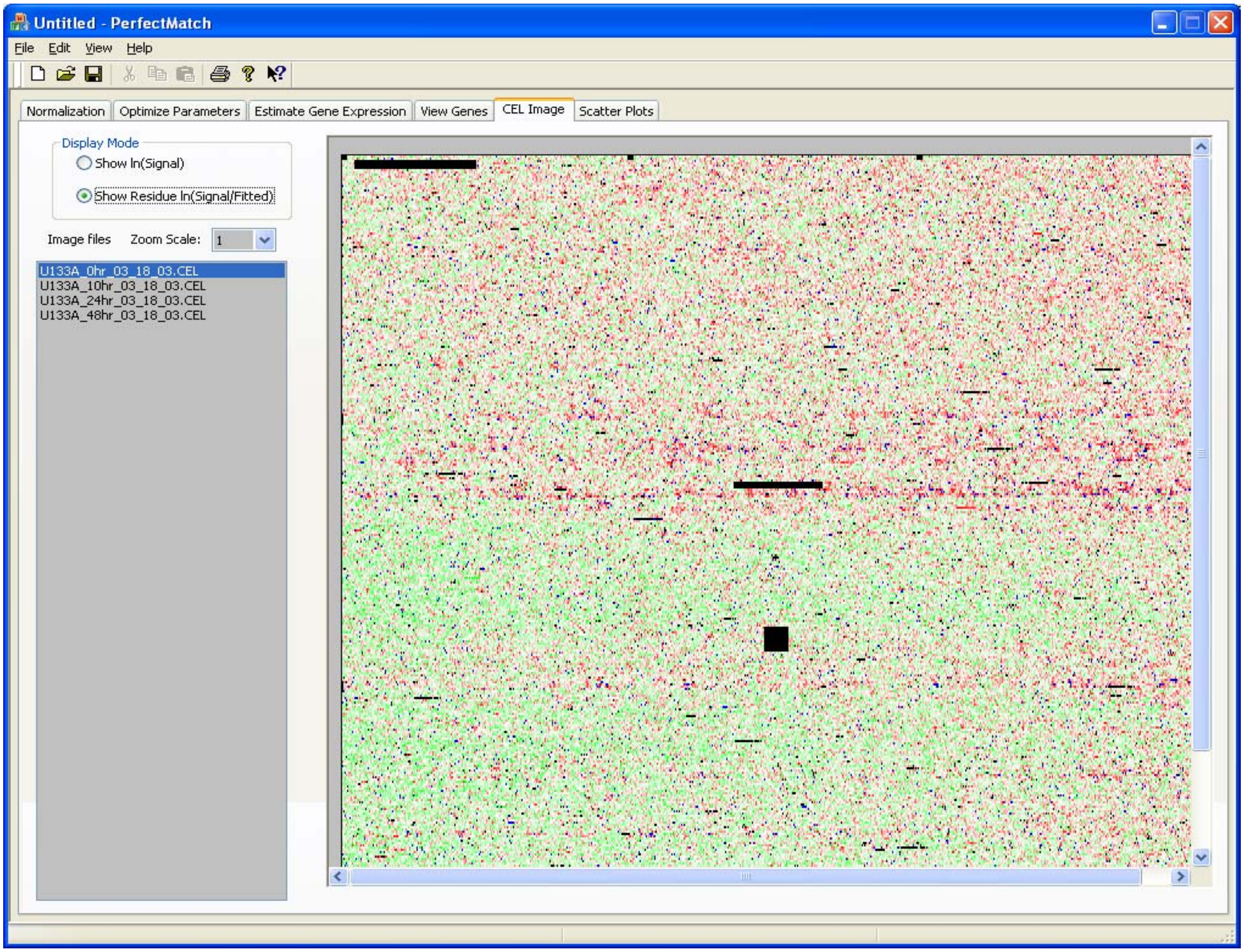
1) Display Mode: you have two modes to choose from: “**Show ln(Signal)**” and “**Show residue ln(Signal/fitted)**”.

“**Show ln(Signal)**” : display the Cel data in ln scale. Below is a Show in(Signal) display, we marker the highest intensity as red , lowest intensity as green , the median as white, everything else falls in between red and green in different depth.

“**Show residue ln(Signal/fitted)**” : display the Ln ration between the original CEL data and the fitted data using PDNN model. As shown in graph below:

In the residue plot, red and green have the same meaning as the show ln(signal) graph. But we add 2 more color in to the display : black and blue.

Black mark all the invalid points in the fitted data. Blue mark all the Outliers in the fitted data.



Display Mode

○ Show ln(Signal)

● Show Residue ln(Signal/Fitted)

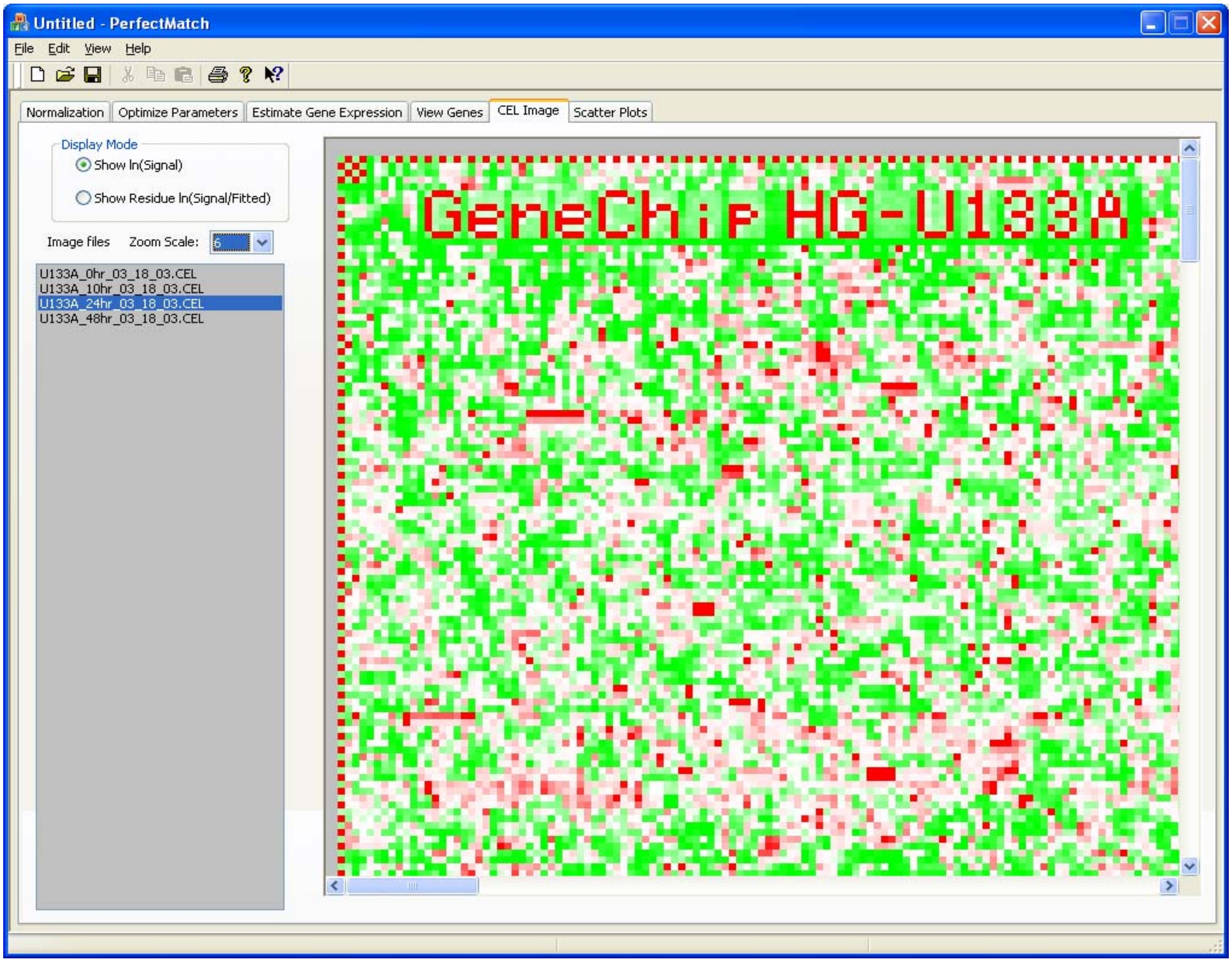
Image files

Zoom Scale: 1

U133A_0hr_03_18_03.CEL
U133A_10hr_03_18_03.CEL
U133A_24hr_03_18_03.CEL
U133A_48hr_03_18_03.CEL

2) **Image files**: List the entire Cel data files which can be displayed, you have to click on one of those to see the image.

3) Scale: you can select a different zoom factor from the drop down box to zoom in and zoom out the image for further study.



Display Mode

● Show ln(Signal)

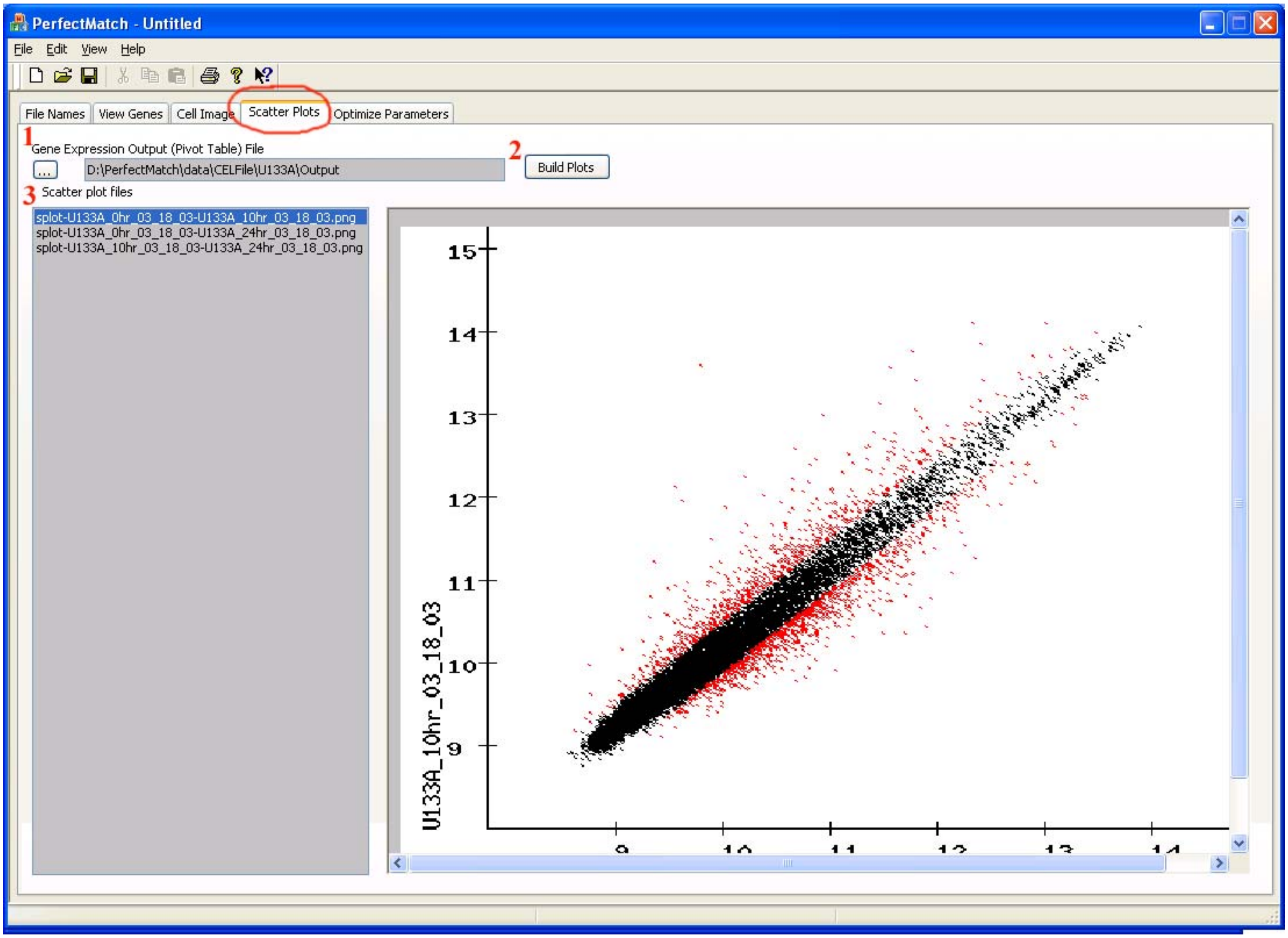
○ Show Residue ln(Signal/Fitted)

Image files Zoom Scale: 6

U133A_0hr_03_18_03.CEL
U133A_10hr_03_18_03.CEL
U133A_24hr_03_18_03.CEL
U133A_48hr_03_18_03.CEL

Section 7. Using the *Scatter Plots* tab to generate scatter plots for multi-array comparison

The “Scatter Plots” Tab are designed to automatically generate scatter plots to compare gene expression values in multiple samples. The program will generate plots for all possible comparisons. User needs to specify the “**Gene Expression Output file**”and click “**Build Plot**” button to generate the plots.



Gene Expression Output (Pivot Table) File

D:\PerfectMatch\data\CELFile\U133A\Output

Scatter plot files

splot-U133A_0hr_03_18_03-U133A_10hr_03_18_03.png
splot-U133A_0hr_03_18_03-U133A_24hr_03_18_03.png
splot-U133A_10hr_03_18_03-U133A_24hr_03_18_03.png

Once the plots are generated, user can select one from the list shown in the left panel to see a plot. The plots are generated in PNG format and stored in the Cel file directory. The off-diagonal points are shown in red, which approximately presents the genes with more two-fold change in gene expression.

- **Release Note on 03/12/2004**

New PerfectMatch (version 2.2) has been released. The Program window can be resized . also modified some bugs .