Evaluation of microarray profile preprocessing methods

Lin Li and Subha Arulselvam
School of Computing
Clemson Univeristy
{ll,sarulse}@clemson.edu

Abstract

Background: One purpose of microarray experiments is to identify differentially expressed genes (DEGs) under different experimental conditions. The preprocessing of microarray is to run quality controls of the image data generated by scanning of a microarray and to convert the image data to a table of expression level values or expression level ratio values for the genes on the microarray. It includes 4 steps: image analysis, background subtraction, normalization, and summarization. There are several algorithms for each stage. In addition, there are different gene selection methods to process the preprocessing result. These methods are categorized into fold-change (FC)-based methods and t-statistic-based methods. A good combination of preprocessing algorithm and gene selection method is critical to achieve optimal performance for detecting DEGs.

Results: We ignored image analysis in the preprocessing, and only compared ten combinations of algorithms for three important steps: background subtraction, normalization, and summarization. The dataset in our experiment is one publicly available spike-in dataset. The evaluation of each combination was based on the area under the receiver operating characteristic (ROC) curve (AUC). We found that mas-constant-avgdiff (in order of background subtraction-normalization-summarization) combination indicates high average of AUC values while rma-invariantset-medianpolish combination indicates low average of AUC values. Higher AUC value indicates higher precision of detecting differential expressed genes.

Conclusion: The result shows that out of the ten combinations in our experiment, the combination of mas-constant-avgdiff leads to highest precision of detecting differential expressed genes.

Availability: A tool which computes AUC value for combinations of preprocessing methods is available at http://mmlab.cs.clemson.edu/881.

BACKGROUND

A microarray is any array (a large number of ordered objects) of biological material, printed on a solid substrate in a "micro" format, which allows many objects to share a relatively small area. It consists of different nucleic acids or protein probes that are

chemically attached to a substrate, which can be a microchip, a glass slide or a microsphere-sized bead. The technology is used as a mean to compare the profiles of two different samples to one another. For instance, with the genome array, one could determine which genes are being turned on or off incorrectly to cause skin cancer by comparing the Ribonucleic acid (RNA) profile of a cancerous skin cell to that of a healthy skin cell. This could help in the discovery of potential drugs.

The microarray image needs a series of preprocessing steps to obtain final data such as the differentially expressed genes. The preprocessing steps are image analysis, background subtraction, normalization, and summarization.

There are many researches on evaluation of these methods individually [1]. B.M.Bolstad et al. propose a paper on the comparison of normalization method [2]. Sepp Hochretier et al. present a new summarization method for affymetrix probe level data in [3]. In addition, many researches focus on the comparison and evaluation of gene selection methods [4][5]. However, no research has been done on evaluation of the combination of the preprocessing steps. Our project focuses on the comparison of combinations of steps exclusive of image analysis involved in microarray preprocessing and aims to find the best combination of steps that would provide higher precision on detecting differentially expressed genes.

In this paper, we compare 10 combinations of preprocessing methods based on the average AUC value. We also give a brief description of the online tool that we develop to automatically compute the AUC value.

METHOD

Figure 1 shows the flow for processing microarray data. In our experiment, we use the pipeline integrated in affy package of Bioconductor [6][11] to preprocess the input microarray. The methods for background subtraction included by affy package are RMA and AMS. The methods for normalization integrated in this package are: constant, contrasts, invariant set, loess, methods, quantiles, qspline, quantlies.robust, and yBatch.methods. The methods for summarization in this package are avgdiff, liwong, MAS, medianpolish, and playerout. In total, there are 90 combinations of the three steps. Note that when we test the combination, we found some of the combinations cannot preprocess the microarray data. We randomly picked 10 combinations that can successfully preprocess the microarray data.

We used 8 gene selection methods in our experiment, which are categorized into two classes [5]. One class is FC-based methods, including weighted average difference (WAD), average difference (AD), fold change (FC), and rank products (RP). The other class is based on t-statistics, which includes moderated t-statistic (modT), significance analysis of microarrays (samT), shrinkage t-statistic (shrinkT), and intensity-based moderated t-statistic (ibmT).

We chose AUC value as the metric for the evaluation. The AUC is a good metric for evaluating the precision of detecting differentially expressed genes because the ROC curve is created by plotting the true positive (TP) rate (sensitivity) against the false positive (FP) rate (1 minus the specificity) [5-8].

The R code for the computation of AUC values is adapted from [5].

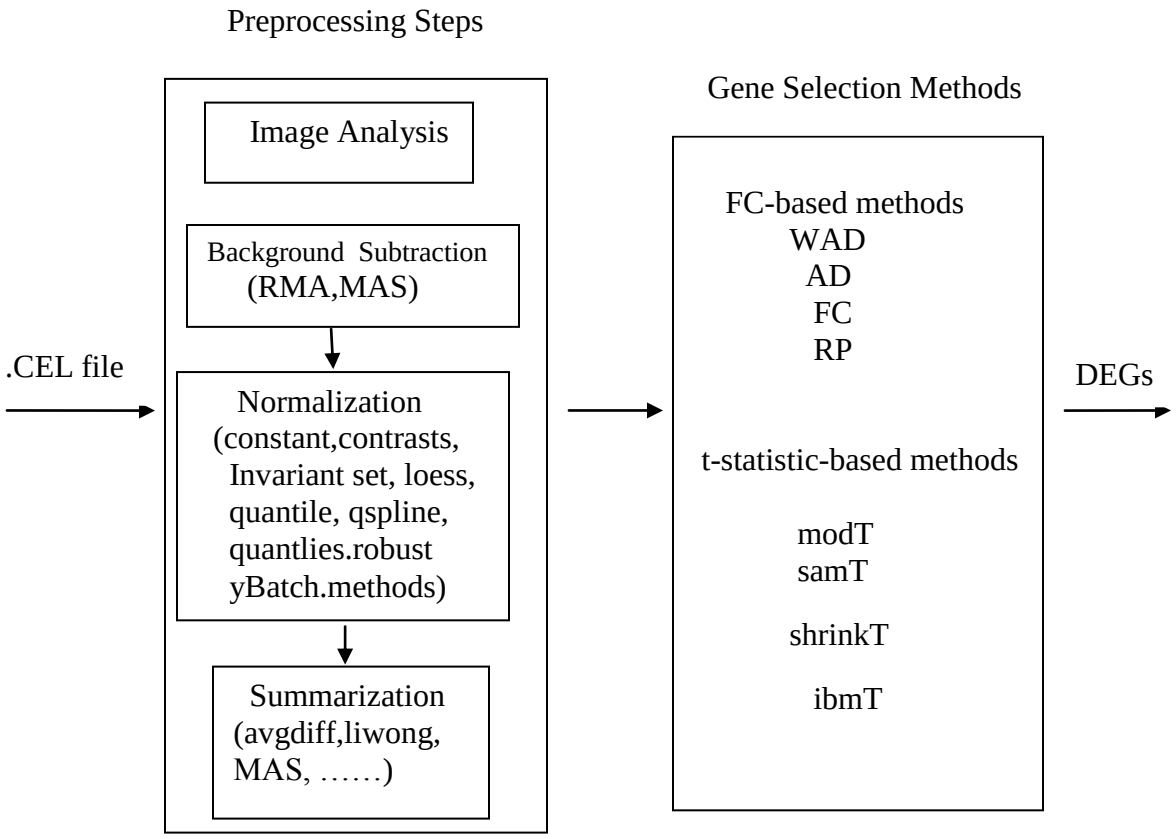

Figure 1. Microarray data processing flow

DATASET

We applied the data processing flow in Figure 1 to a publicly available dataset [9][10]: Spike-in hgu95a Data. The reason we chose this dataset is because there are 16 known DEGs in this dataset. The detail description of the dataset is in [5].

RESULTS

We evaluated the dataset under ten different combinations. Those combinations are shown in Table 1.

Table 1 ten combinations of preprocessing methods for microarray data

Combination	Background subtraction	normalization	summarization
1	rma	constant	avgdiff
2	mas	constant	avgdiff
3	mas	invariantset	medianpolish
4	mas	loess	liwong
5	rma	contrasts	liwong
6	rma	quantiles	liwong
7	mas	quantiles	liwong
8	rma	loess	liwong
9	rma	invariantset	medianpolish
10	mas	quantiles	avgdiff

The AUC values for each combination (1 – 10 in Table 1) using eight gene selection methods are listed in Table 2. It is easy to tell that the combination of rma, invariantset and medianpolish has the lowest average AUC out of the 10 combinations, while the combination of mas, quantiles and avgdiff has the highest average AUC value. Therefore, we can say that out of the 10 combinations, it is better to use the combination of mas, quantiles and avgdiff if AUC value is the metric. Also, ibmT is a good choice for gene selection methods.

Table 2 AUC values for 10 combinations of preprocessing methods

Gene Selection Methods	1	2	3	4	5	6	7	8	9	10
AD	0.9887044	0.9887	0.986479	0.9852944	0.9866376	0.980953	0.9882633	0.9715652	0.9466693	0.9957177
WAD	0.9987213	0.99872	0.9993011	0.9911033	0.9980819	0.994816	0.9930908	0.9901913	0.9991425	0.9992417
FC	0.985344	0.98534	0.9868606	0.9829302	0.9768041	0.974831	0.9863253	0.9626338	0.946709	0.9949395
RP	0.9930908	0.99309	0.9671788	0.9634318	0.981136	0.950684	0.9646461	0.9547086	0.9218973	0.9832722
modT	0.9964512	0.99645	0.9997076	0.9970163	0.9989046	0.998478	0.9980769	0.9964512	0.999207	0.9993606
samT	0.9961538	0.99615	0.9996977	0.9962084	0.9987262	0.998563	0.9975763	0.994206	0.9983991	0.9993011
shrinkT	0.9963719	0.99637	0.9994052	0.9920301	0.9981414	0.994721	0.9946223	0.9887143	0.9982058	0.9992764
ibmT	0.9971253	0.99713	0.9997918	0.9972145	0.9989443	0.998518	0.9982157	0.9961538	0.9996233	0.9994845
Average	0.9939953	0.9939938	0.9923027	0.9881536	0.9921720	0.9864455	0.9901021	0.9818280	0.9762317	0.99632425

ONLINE TOOL

We created an online tool to compute AUC value to evaluate the choice of preprocessing methods. User needs to enter the personal email address to get the computation result. Also, user is required to select the method for background subtraction, normalization and summarization and click the "submit" button to submit the computation request. If there is any error in the computation, user will get an email of notification. Otherwise, once the computation is finished, a text file with all the AUC values will be sent to user. Figure 2 illustrates the interface of our online tool for comparison of microarray preprocessing methods.

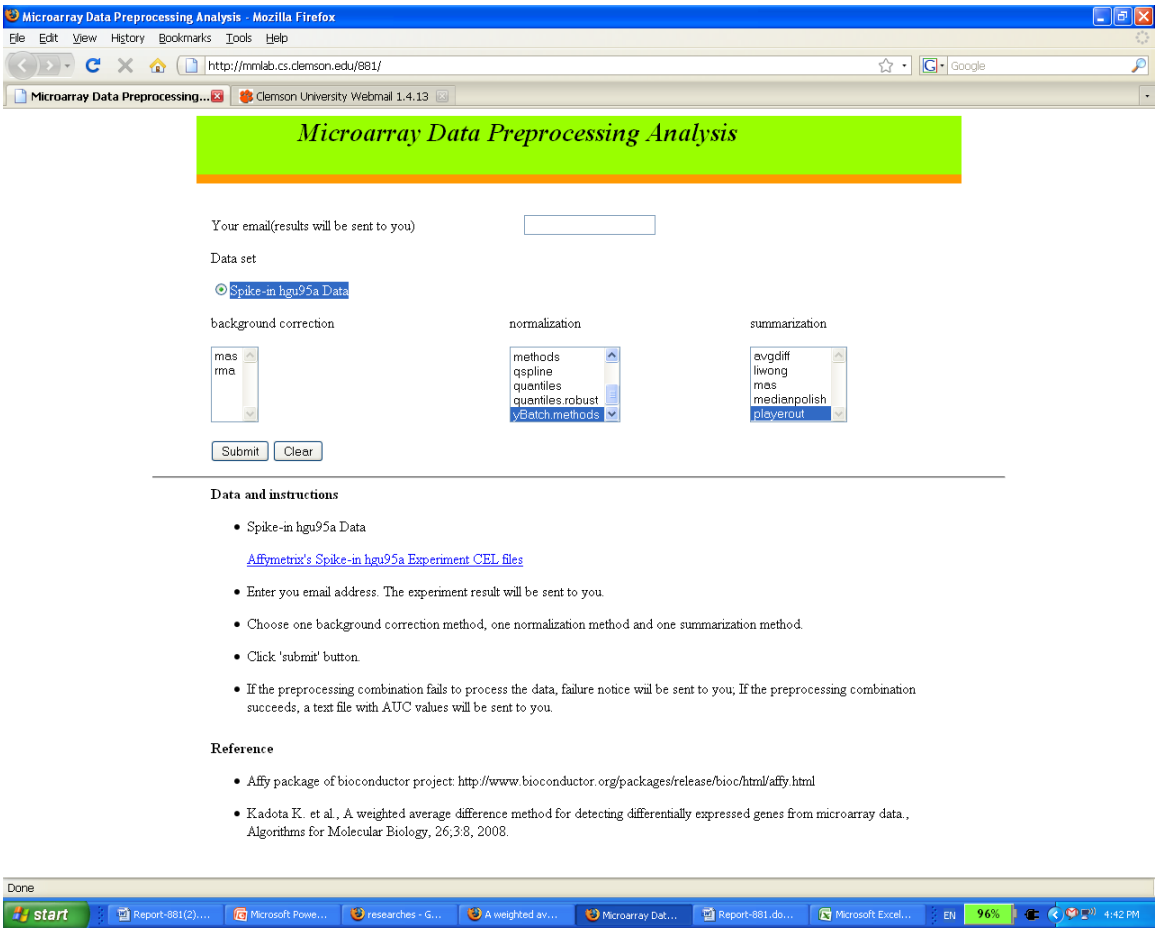

Figure 2. Interface of our online tool for comparison of microarray preprocessing methods

CONCLUSION

In this paper, we present our project on the evaluation of microarray preprocessing methods. The dataset for our project is Spike-in hgu95a data. We also used the pipeline integrated by affy package in Bioconductor to preprocess the microarray dataset. We

compared 10 combinations using AUC value as the metric. The result shows that out of the 10 combinations, the combination of rma, invariantset and medianpolish has the lowest average AUC, while the combination of mas, quantiles and avgdiff has the highest one.

REFERENCES

[1] Irizarry,R.A. et al., *Comparison of Affymetrix GeneChip expression measures*, Bioinformatics, 22, 789–794, 2006.
[2] Bolstad,B.M. et al., *A comparison of normalization methods for high density oligunucleotide array data based on variance and bias*, Bioinformatics, 19, 185–193, 2003.
[3] Hochreiter, S., et al., *A new summarization method for Affymetrix probe level data*, Bioinformatics, 22, 943–949, 2006.
[4] Jeffery I.B., Higgins D.G., Culhane A.C, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data*, BMC Bioinformatics, 7, 359, 2006.
[5] Kadota K. et al., *A weighted average difference method for detecting differentially expressed genes from microarray data,* Algorithms for Molecular Biology, 26, 3:8, 2008.
[6] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*, Biostatistics, 2003, 4:249–264. doi: 10.1093/biostatistics/4.2.249.
[7] Hochreiter S, Clevert DA, Obermayer K, *A new summarization method for Affymetrix probe level data*, Bioinformatics, 2006, 22:943–949. doi: 10.1093/bioinformatics/btl033.
[8] Chen Z, McGee M, Liu Q, Scheuermann RH, *A distribution free summarization method for Affymetrix GeneChip arrays*, Bioinformatics, 2007;23:321–327. doi: 10.1093/bioinformatics/btl609.
[9] Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP, *A benchmark for Affymetrix GeneChip expression measures*, Bioinformatics, 2004, 20:323–331. doi: 10.1093/bioinformatics/btg410.
[10] Affycomp II website. http://affycomp.biostat.jhsph.edu/.
[11] Laurent Gautier et al., Description of affy. Available at http://bioconductor.org/packages/2.4/bioc/html/affy.html.