Joan Cabestany Francisco Sandoval
Alberto Prieto Juan M. Corchado (Eds.)

Bio-Inspired Systems: Computational and Ambient Intelligence

10th International Work-Conference
on Artificial Neural Networks, IWANN 2009
Salamanca, Spain, June 10-12, 2009
Proceedings, Part I

Springer

Volume Editors

Joan Cabestany
Universitat Politècnica de Catalunya - UPC
E.T.S.E. Telecomunicació, Barcelona, Spain
E-mail: cabestan@eel.upc.es

Francisco Sandoval
Universidad de Málaga
E.T.S.I. Telecomunicación, Málaga, Spain
E-mail: fsandoval@uma.es

Alberto Prieto
Universidad de Granada
E.T.S.I. Informática y Telecomunicación, Granada, Spain
E-mail: aprieto@ugr.es

Juan M. Corchado
Universidad de Salamanca
Departamento de Informática, Salamanca, Spain
E-mail: corchado@usal.es

Library of Congress Control Number: Applied for

CR Subject Classification (1998): J.3, I.2, I.5, C.2.4, H.3.4, D.1, D.2

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-642-02477-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-02477-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12695607 06/3180 5 4 3 2 1 0

On Selecting the Best Pre-processing Method for Affymetrix Genechips

J.P. Florido[1], H. Pomares[1], I. Rojas[1], J.C. Calvo[1], J.M. Urquiza[1], and M. Gonzalo Claros[2]

[1] Department of Computer Architecture and Computer Technology, University of Granada, Granada, Spain
{jpflorido,hector}@ugr.es, {irojas,jccalvo,jurquiza}@atc.ugr.es
[2] Department of Molecular Biology and Biochemistry, University of Málaga, Málaga, Spain
claros@uma.es

Abstract. Affymetrix High Oligonucleotide expression arrays, also known as Affymetrix GeneChips, are widely used for the high-throughput assessment of gene expression of thousands of genes simultaneously. Although disputed by several authors, there are non-biological variations and systematic biases that must be removed as much as possible before an absolute expression level for every gene is assessed. Several pre-processing methods are available in the literature and five common ones (RMA, GCRMA, MAS5, dChip and VSN) and two customized Loess methods are benchmarked in terms of data variability, similarity of data distributions and correlation coefficient among replicated slides in a variety of real examples. Besides, it will be checked how the variant and invariant genes can influence on preprocessing performance.

1 Introduction

Microarray technology is a powerful tool used for the high-throughput assessment of gene expression of thousands of genes simultaneously which can be used to infer metabolic pathways, to characterize protein-protein interactions or to extract target genes for developing therapies for various diseases [1]. Several platforms are currently available, including the commonly used high oligonucleotide-based Affymetrix GeneChip® arrays.

As described in [1], an Affymetrix GeneChip contains probe sets of 10-20 probe pairs representing unique genes. Each probe pair consists of two oligonucleotides of 25 bp in length, namely perfect match (PM) probes (the exact complement of an mRNA) and the mismatch (MM) probes (which are identical to the perfect match except that one base is changed at the center position). The MM probe is supposed to distinguish noise caused by non-specific hybridization from the specific hybridization signal, although some researchers recommend avoiding its use [17].

A typical microarray experiment has biological and technical sources of variation [2]. Biological variation results from tissue heterogeneity, genetic polymorphism, and changes in mRNA levels within cells and among individuals due to sex, age, race, genotype-environment interactions and other "living" factors. Biological variation is of interest to researchers as it reflects true variation among experiments. On the other

J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 845–852, 2009.

hand, sample preparation, labeling, hybridization and other steps of microarray experiment can contribute to technical variation, which can significantly impact the quality of array data. Therefore, to that systematic non-biological sources of variation mask real biological variation, significant pre-processing is required and involves four steps for Affymetrix GeneChips: background correction, normalization, PM correction and summarization [15].

In this paper, a comparison of some of the most well known pre-defined pre-processing methods (RMA, GCRMA, MAS5, dChip and VSN) and two customized Loess methods is performed. Accuracy of preprocessing is assessed in terms of data variability, similarity in data distributions and correlation among replicates in a variety of real examples. Besides, one of the pre-processing method will be selected to evaluate how the variant and invariant genes have an influence on the quality metrics compared to the whole set of genes.

Section 2 describes the main pre-processing methods existing in the literature for Affymetrix GeneChips and section 3 describes implementations, data sets and results. Conclusions are drawn in section 4.

2 Pre-processing Affymetrix GeneChips

Instead of describing how does work every pre-processing method, they will be compared in each of the four pre-processing steps [3]:

- *Background correction.* It removes unspecific background intensities of scanner images. Three possible algorithms can achieve this correction: the Robust Multi-chip Average (*RMA*) convolution [5] and the *MAS* [6] and *Gcrma* [7] algorithms.
- *Normalization.* It is intended to reduce most of the non-biological differences between chips providing normalized (comparable) signal intensities for every chip [1]. The following methods were applied: *Scaling (constant)* [6], *Quantile* [4], *Loess* [4], *Invariant Set* [8] and Variance Stabilization Method (*VSN*) [9].
- *PM correction*. PM signal intensity should be adjusted to account for nonspecific signal. There are two possible algorithms: *MAS* [6] and *PMonly* [3].
- *Summarization*. It is the final stage in pre-processing Affymetrix GeneChip data and computes expression values from all within-chip replicates by combining the intensities of the 11-20 probe replicates to produce a single expression value for a gene. There are some well-known methods in the literature: *Median Polish* [5], *Tukey Biweight* [6], *Li-Wong MBEI* expression index [10] and *Avgdiff* [11].

The algorithms described above can be found in the Bioconductor *affy*, *affyPLM* and *vsn* packages [3][9][12], which are R libraries of functions and classes.

3 Experiments and Results

In this section, pre-defined and custom pre-processing methods used in the comparison will be explained as well as the data sets and the quality metrics used in the experiment for evaluating such comparison.

3.1 Implementation of Pre-processing Methods

There are six different functions (Table 1) to calculate all pre-processing methods:

- *expresso()* (*affy* package [3][12]), which provides quite general facilities for computing expression summary values.
- *rma()* (*affy* package [3][12]) to calculate only the RMA method.
- *threestep()* (*affyPLM* package [12]), which provides the user the ability to compute very general expression measures.
- *gcrma()* (*affyPLM* package [12]), which computes the GCRMA method.
- *mas5()* (*affy* package [3][12]) to calculate the MAS5 method.
- *vsnrma()* (*vsn* package [9][12]) to calculate only the VSN method.

Table 1. Summary of microarray data pre-processing methods used in this paper and the R function utilized for its calculation

Methods	*Background Correction*	*Normalization*	*PM correction*	*Summarization*	*Functions* [a]
RMA	rma	quantiles	pmonly	medianpolish	expresso(), rma(), threestep()
GCRMA	gcrma	quantiles	pmonly	medianpolish	gcrma(), threestep()
MAS5	mas	mas	mas	mas	expresso(), mas5()
VSN		vsn	pmonly	medianpolish	expresso(), vsnrma()
dChip		invariantset	pmonly	liwong	expresso()
Loess 1	mas	loess	mas	mas	expresso()
Loess 2		loess	pmonly	avgdiff	expresso()

[a] Note that a pre-processing method can be used with more than one function, usually from different R libraries.

3.2 Data Sets

Five different data sets covering different experimental designs (from control/ treatment to a time series, with different replications) were used:

- *Dilution experiment*. Two sources of cRNA A (human liver tissue) and B (Central Nervous System cell line) have been hybridized to human array (*HGU95A*) in a range of proportions and dilutions. The data is available at the *affydata* package [12] and detailed description can be accessed there.
- *Estrogen*. The package *estrogen* [12] contains 8 Affymetrix *HG-U95Av2* CEL files from an experiment involving estrogen receptor positive (ER+) breast cancer cells [13]. Estrogen effect on time is evaluated.
- *Pig infection.* This experiment contains 10 samples run on *Porcine* Affymetrix Genechip arrays in which 6 pigs were treated with an infectious agent and the remaining ones were not treated. These unpublished data were kindly given by Prof. Juan José Garrido (Department of Genetics, University of Córdoba, Spain).
- *Cell infection*. This experiment was also kindly given by Prof. Juan J. Garrido and consists of 18 Affymetrix *Porcine* CEL files. It contains the infection of 2 cell lines (IPI and IPEC) with a specific pathogen for each one. Each infection is sampled at 0, 2 and 4 h of infection by triplicate. The IPI and IPEC strains have been treated separately since they are different cell lines treated with different pathogens.

– *CLL.* The CLL package [12] contains the chronic lymphocytic leukemia (CLL) gene expression data. The CLL data had 24 samples run on *HG-U95Av2* Affymetrix GeneChip arrays that were either classified as progressive or stable in regards to disease progression.

3.3 Quality Metrics

In order to evaluate the performance of the pre-processing methods, three different metrics have been used:

– *Replicate variability* [14]. It is based on the assumption that expression level of a gene should ideally remain the same across multiple replicated slides. Variability is measured by the mean of the standard deviation over all genes. Smaller mean is indicative of better pre-processing.
– *Kolmogorov-Smirnov (K-S) test* [14]. It is a goodness-of-fit test of two continuous distributions and it is based on the hypothesis that an effective normalization procedure should result in two similar, ideally identical, distributions with a small, ideally zero-valued K-S statistic. On the other hand, two different distributions will generate a large K-S statistic.
– *Spearman Rank Correlation Coefficient* [15]. It is based on the comparison of the correlation coefficient between replicated slides assuming that, given an experiment, the correlation coefficient among replicated slides will be increased after the pre-processing stage.

3.4 Results

Raw expression data. The different pre-processing methods should be compared with the raw expression data in terms of the quality metrics described above. Thus, all the experiments were run with no background correction, no normalization and no PM correction. Just the summarization step is needed to obtain gene expression values. Three different summarization algorithms using the *expresso* function were tested: *median-polish*, *Tukey-Biweight* and *AvgDiff.*

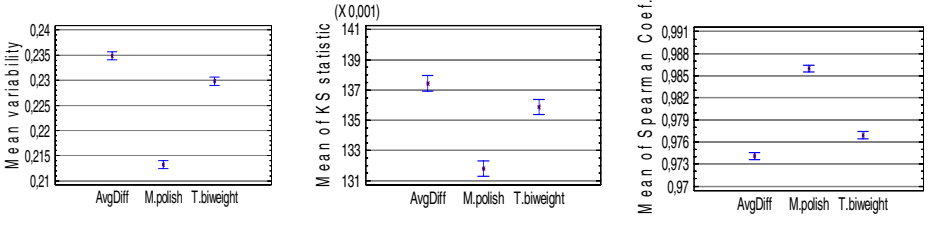

Fig. 1. Means and 95% Least Significant Differences (LSD) Intervals of the different summarization algorithms through the quality metrics

For each quality metric, a two-way Analysis Of Variance (ANOVA) test checked the statistical significance of the results. Fig.1 shows that *median-polish* summarization method performs significantly better ($P < 0.05$). Hence, it has been selected to obtain raw expression values for comparing with all pre-processing methods.

Comparison of functions for pre-defined pre-processing methods. Since there are several pre-defined pre-processing methods (RMA, GCRMA, MAS5 and VSN) that can be used with more than one function (Table 1), selection of one function for each method is a must. dChip pre-processing method could not be included: it does not converge for most of the data sets since a minimum of 10 replicates arrays is recommended [12]. Thus, for each quality metric, a one-way ANOVA test was run for the different functions within a pre-processing method and no statistical difference ($P > 0.05$) among functions for the same method is found. Hence, the decision was taken from the average running time of a total of 10 executions (Fig. 2). As a result, *rma()* performs faster for RMA, *gcrma()* and *threestep()* for GCRMA are practically the same (although the latter will be used due to it provides the user with a great deal of control), *mas5()* is the fastest for MAS5 and *vsnrma()* is the best for VSN.

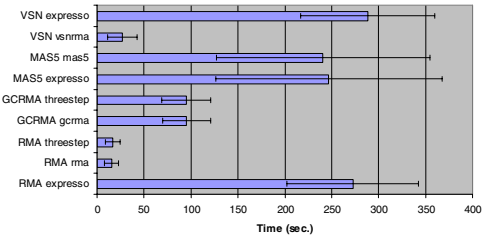

Fig. 2. Average and standard dev. running time of pre-processing functions for all data sets

Comparison of different pre-processing methods. For each quality metric, a two-way ANOVA test was used to check how each pre-processing method performs in all data sets**.** According to the mean variability, the *VSN* method gets the best results (Fig. 3a). *VSN*, *RMA*, *GCRMA* and *Loess 2* are statistically better ($P < 0.05$) than the raw data. The *VSN* performance was not unexpected because it specifically aims to stabilize the variance across the replicated arrays. On the other hand, *MAS5* and *Loess 1* obtain worse results than the raw data. Since both methods are identical but for normalization step, it can be concluded that the normalization method alone cannot account for the pre-processing performance.

With regard to the mean Kolmogorov-Smirnov statistic (Fig. 3b), *RMA* and *GCRMA* get the best results, followed by *Loess 2*, *VSN*, *MAS* and *Loess 1*, and all of them perform significantly better ($P < 0.05$) than raw data. *RMA* and *GCRMA* perform the best since the use of *Quantile* normalization algorithm (see section 2) that forces the empirical distributions in different slides to be identical.

Finally, according to the mean Spearman rank coefficient (Fig. 3c), *RMA*, *VSN* and Raw data performs better with no statistical difference among them. This means that no significant improvement has achieved when pre-processing the data in terms of correlation.

In conclusion, a pre-processing method is considered appropriate when (i) its mean variability is lower than for the raw data; (ii) the mean K-S statistic is lower than for the raw data; and (iii) the Spearman rank coefficient is higher than for the raw data. Only RMA and VSN method seems to fulfill these rules, although in (iii) the difference is not statistically significant.

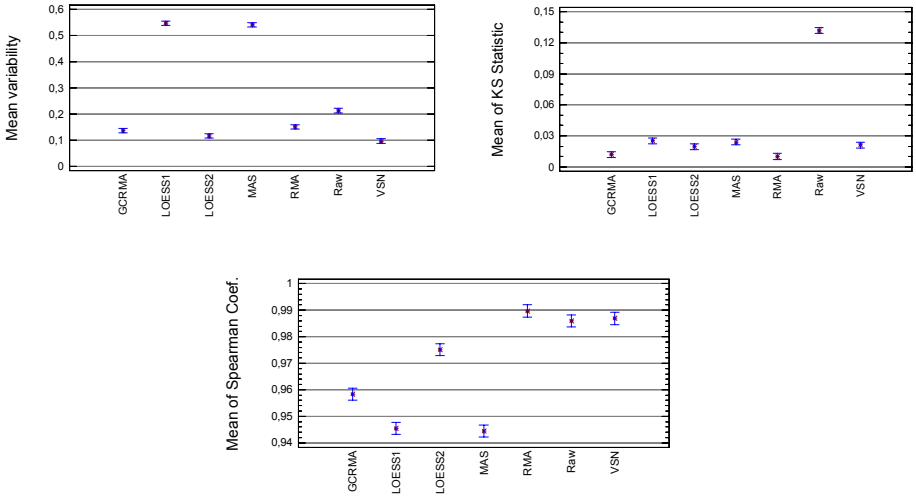

Fig. 3. Means and 95% LSD intervals of the different pre-processing methods through the quality metrics: a) mean variability, b) mean of K-S statistic, c) mean of Spearman Coefficient

The influence of variant and invariant genes with respect the whole set of genes. It would be interesting how the variant and invariant genes can influence on pre-processing performance. Invariant and highly-variant genes were inferred from raw summarized data. Two sets (A and B) of replicated slides must be selected. For each set, the mean value for each gene in all replicates is obtained. Then, a list of interesting genes is obtained according to M = abs(log2(A/B)), so that the *n* genes (1000 in our case) with the highest value of M will be the variant genes and the *n* genes with the lowest value of M, closest to zero, will be the invariant ones. These groups were pre-processed with *RMA*, due to its great performance shown in the previous subsection. Raw summarized data were utilized as controls using these groups as well as the pre-processing with *RMA* and raw summarized data (*median-polish*) using the whole set of genes. Statistical significance was assessed by a three-way ANOVA test.

Mean variability and K-S statistic (Fig. 4a, 4b) reveal that any set of genes pre-processed with RMA is statistically more homogeneous ($P < 0.05$) than raw data (which only have been summarized by *median-polish*). This improvement, as expected due to *quantile* normalization, is more aggresive for the K-S statistic. Moreover, pre-processing whole data is statistically equivalent ($P > 0.05$) to pre-processing only the invariant set of genes, while the difference is clearly significant ($P < 0.05$) when only the variant set is pre-processed. This result is not surprising, since it is assumed that most of the genes in a microarray experiment are non-differentially ones, i.e. they are invariants. The mean Spearman rank coefficient (Fig. 4c) also reveals that pre-processing an invariant set of genes with RMA is equivalent to do the same with the whole data. However, it is noticeable that invariant and all genes datasets do not show a great improvement after the pre-processing while variant genes do it clearly, showing that pre-processing does not break (even it can improve it slightly) the correlation among data replication. Therefore, the variant set

improves the correlation with pre-processing since they are supposed to be more spread before the pre-processing. Hence, it has been confirmed that pre-processing with an invariant dataset is similar to pre-processing the whole chip provided that most of genes in chip are not differentially expressed and it is also confirmed that pre-processed data are preferable to raw data.

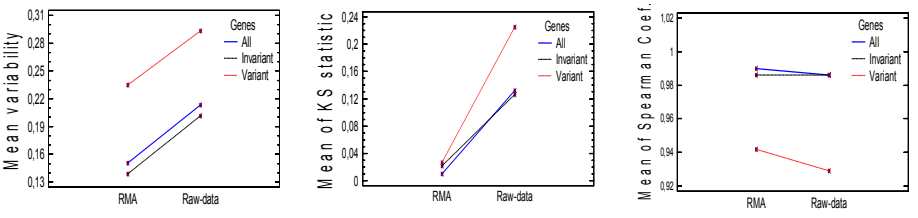

Fig. 4. Interaction plots of methods and genes in terms of the quality metrics: a) mean variability, b) mean of K-S Statistic, and c) mean of Spearman Rank Coefficient

4 Conclusions and Future Work

In this paper, we performed a comparison of some of the most well known pre-processing methods (RMA, GCRMA, MAS5 and VSN) and two custom implementations of LOESS in terms of quality metrics such as data variability, similarity of data distributions, and correlation coefficient among replicated slides. Five real datasets obtained from different laboratories, with different design and with different number of replica were employed, which provide a strong support for the conclusions since they are not linked to particular data. Implemented functions for preprocessing are time-efficient but for *expresso()* and *median-polish* is the summarization method that seems to perform better. According to our three golden rules for an adequate pre-processing method, only VSN and RMA seems to normalize data in such a way that the results are better devoid of experimental and technical errors. Indirectly, it has been shown that *Loess* normalization performance is highly dependent on the rest of pre-processing steps. When an invariant gene set is pre-processed with RMA, its behaviour is similar than when the whole chip is pre-processed with the same method since most of the genes in a microarray are not expected to be differentially expressed. Moreover, most of the genes in a microarray experiment are expected to be correlated even if no pre-processing method is applied.

As a future work, a more complex system is desired: apply supervised learning models such as Artificial Neural Networks with the aim of recognize data patterns to decide the best pre-processing method for a given data set [16]. We are also searching for a function that combines the three quality metrics to obtain an objective way to evaluate which is the best way of pre-processing for a given set of data.

Acknowledgment

This paper has been supported by the Spanish Ministry of Education and Science under project TIN2007-60587 and the FPU research grant AP2007-03009.

References

1. Hochreiter, S., Clevert, D.A., Obermayer, K.: A new summarization method for affymetrix probe level data. Bioinformatics 22(8), 943–949 (2006)
2. Zakharkin, S.O., Kim, K., Mehta, T., et al.: Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics 6, 214 (2005)
3. Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A.: affy-analysis of Affymetrix GEneChip data at the probe level. Bioinformatics 20(3), 307–315 (2003)
4. Bolstad, B.M., et al.: A Comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinf. 19, 185–193 (2002)
5. Irizarry, R.A., Hobbs, B., Collin, F., et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2003)
6. Affymetrix Microarray Suite Users Guide, Affymetrix, Santa Clara, v.5.0 edn. (2001)
7. Wu, Z., Irizarry, R., Gentleman, R., et al.: A model based background adjustement for oligonucleotide expression arrays. Journal of the American Statistical Association (2005)
8. Schadt, E.E., et al.: Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J. Cell Biochem. Suppl., 120–125 (2001)
9. Huber, W., et al.: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(suppl. 1), S96–S104 (2002)
10. Li, C., Wong, W.H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA 98(1), 31–36 (2001)
11. Affymetrix Microarray Suite Users Guide. Affymetrix, Santa Clara, v.4.0 edn. (1999)
12. Gentleman, R.C., et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5(10), Article R80 (2004)
13. Scholtens, D., Miron, A., Merchant, F., et al.: Analyzing factorial designed microarray experiments. Journal of Multivariate Analysis 90, 19–43 (2004)
14. Xiong, H., Zhang, D., Martyniuk, C.J., et al.: Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data. BMC Bioinformatics 9(25) (2008)
15. Lim, W.K., Wang, K., et al.: Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics 23, 282–288 (2007)
16. Rojas, I., Pomares, H., et al.: Analysis of the functional block envolved in the design of radial basis function networks. Neural Processing Letters 12(1), 1–17 (2000)
17. Irizarry, R.A., Bolstad, B.M., Collin, F., et al.: Summaries of Affymetrix GeneChip proble level data. Nucleic Acids 31(4) (2003)