Which normalization method is best? A platform-independent biologically inspired quantitative comparison of normalization methods

E.P. van Someren
Dept. of Mediametics
Delft Univ. of Technology
Delft, The Netherlands
E.P.vanSomeren@ewi.tudelft.nl

M.J.T. Reinders
Dept. of Mediametics
Delft Univ. of Technology
Delft, The Netherlands

Abstract

Since the introduction of microarray technology, several different normalization techniques have been introduced, but it is still unclear which normalization method is best. We present the first comparative study of normalization methods for both cDNA as well as oligonucleotide arrays that is based on their overall performance on five complementary performance measures.
The presented comparison is unique in that it 1) compares normalization methods with very different outcomes, 2) is applied to two different array platforms, 3) introduces several different (biologically inspired) performance measures and 4) can be applied to any data set.
The results show amongst others that, for cDNA arrays, the well-established lowess-compensation of logratio is not biologically beneficial and that a novel ratio-based normalization (without logarithm) performs best overall. For Affymetrix arrays, we found that Rosetta's Experiment Builder is generally to be preferred.

1. Introduction

Since the introduction of microarray technology, several different normalization techniques have been introduced [1, 2]. When faced with the necessity to normalize you own data, it is still unclear which normalization method is most suited for your laboratory conditions. We present the first comparative study of normalization methods for both cDNA as well as oligonucleotide arrays that is based on their overall performance on five complementary performance measures.

The presented comparison is unique in that it compares normalization methods with very different outcomes and is applied to two different array platforms. The testing methodology is unique because several different performance measures are introduced that need to be satisfied simultaneously and that all quantify biologically desired properties of the data after normalization. Furthermore, the methodology can be applied to virtually all data sets, without requiring one to create data specifically for testing normalization methods.

2 Performance Measures

We defined five different performance measures that described desired properties of the data after normalization. The method that scores well on *all* performance measures will be the method to be preferred.

After normalization, replicate arrays should be more similar to each other than to non-replicates. Therefore, we want to minimize, f^{REP}, the (normalized) sum of the ranks of replicate array pairs after ranking all array pairs based on increasing distance.

After normalization, the trajectories of time-course arrays should be smooth and/or arrays of the same patient, tissue, treatment etc. should be more similar to each other than to arrays that are not related. Therefore, we want to minimize, f^{REL}, the (normalized) sum of the ranks of related array pairs after ranking all array pairs based on increasing distance.

After normalization, gene profiles should match the shape of PCR profiles. Therefore, we want to maximize, f^{COR}, the average correlation of microarray gene expression profiles with corresponding PCR profiles.

After normalization, profiles of housekeeping and unrelated genes should be constant, whereas profiles of known marker and related genes should be significantly regulated. Therefore, we want to maximize, f^{DIF}, the separation in the population of differential expression scores of genes that should be differentially expressed versus genes that should not.

After normalization, gene selections and/or gene clusters

Criterium	ind	lograt	p2	drat	ma	logp2	p1	logp1	best	second	third
f^{REP}	**0.06**	0.42	0.21	0.22	0.33	0.70	0.81	0.83	ind	p2	drat
f^{REL}	**.10**	0.27	0.11	0.13	0.27	0.42	0.56	0.60	ind	p2	drat
f^{COR}	**.98**	.97	.88	.97	.96	.83	-.11	-.08	ind	drat,lograt	-
f^{DIF}	-.56	-.58	-.31	-.23	-.40	**-1.10**	2.34	-0.02	logp2	lograt	ind
Avg. rank	**1.5**	3.25	3.5	3.5	4	4.75	7.5	7.5	ind	lograt	p2,drat

Table 1. Typical example of summarized performance of eight cDNA normalization methods on four different performance criteria on one of the tested data sets. The performance criteria (one on each row) are replicate similarity f^{REP} , related array similarity f^{REL} , correlation with PCR f^{COR} and differential expression between markers and housekeeping/control genes f^{DIF} . The normalization methods are relative induction (ind), logratio ($lograt$), (log of) red channel ($(log)p2$), (log of) green channel (was common reference) ($(log)p1$), balanced ratio ($drat$), lowess-compensated logratio (ma). For each performance criterium the methods are ranked from best (1) to worst (8) (see also last three columns). In addition, the methods (column 2-9) are ordered in the table based on the averaged rank over all criteria, which is also depicted in bottom row.

should be more functionally related. Therefore, we want to maximize, f^{ENR}, (threshold independent) enrichment of gene sets.

3 Results and Discussion

The results on cDNA array data show that the choice on how to combine the two channels is more important than between-array normalization. Furthermore, we found than the well-established lowess-compensation of logratio (ma) is not biologically beneficial and that a novel ratio-based normalization (without logarithm) (ind) performs best overall. Table 1 illustrates a typical result on a murine cDNA array time-series dataset.

For Affymetrix arrays, we found that Rosetta's Experiment Builder [3] is consistently better in reducing systematic bias. However, typically, the more systematic bias is reduced, the more biological information is lost. Figure 1 shows one of the results on a murine Affymetrix array compendium. The results illustrate the typical contradictionary performance on replicate similarity versus Gene Ontology enrichment, i.e. f^{REP} needs to be minimized, while $|f^{ENR}|$ needs to be maximized simultaneously.

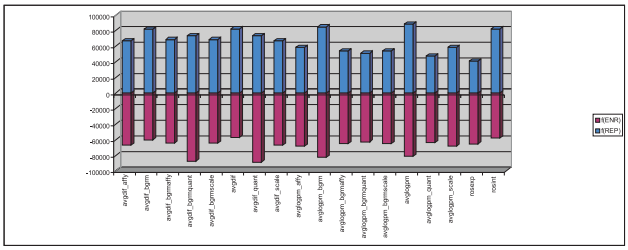

Figure 1. Typical example of 18 Affymetrix normalization methods on two complementary performance criteria on one of the tested data sets. The displayed performance criteria are replicate similarity f^{REP} and GO enrichment f^{ENR}. Normalization methods are Rosetta's Experiment and Intensity Builder versus standard normalization methods based on combined choices for background correction, between-array normalization and how to combine PM and MM values.)

References

[1] Y. Chen, E. Dougherty, and M. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2(4):364–374, October 1997.

[2] D. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6):557–569, 2001.

[3] L. Wong. Data processing and analysis methods in the rosetta resolver system. Technical report, Technical Report, november 2004.