BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 8 2006, pages 943–949
doi:10.1093/bioinformatics/btl033

Gene expression

A new summarization method for affymetrix probe level data

Sepp Hochreiter*[,1,2], Djork-Arné Clevert[1] and Klaus Obermayer[1]

[1]Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany and [2]Institute of Bioinformatics, Johannes Kepler Universität Linz, 4040 Linz, Austria

Received on October 7, 2005; revised on December 13, 2005; accepted on January 30, 2006
Advance Access publication February 10, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: We propose a new model-based technique for summarizing high-density oligonucleotide array data at probe level for Affymetrix GeneChips. The new summarization method is based on a factor analysis model for which a Bayesian maximum a posteriori method optimizes the model parameters under the assumption of Gaussian measurement noise. Thereafter, the RNA concentration is estimated from the model. In contrast to previous methods our new method called 'Factor Analysis for Robust Microarray Summarization (FARMS)' supplies both *P*-values indicating interesting information and signal intensity values.

Results: We compare FARMS on Affymetrix's spike-in and Gene Logic's dilution data to established algorithms like Affymetrix Microarray Suite (MAS) 5.0, Model Based Expression Index (MBEI), Robust Multi-array Average (RMA). Further, we compared FARMS with 43 other methods via the 'Affycomp II' competition. The experimental results show that FARMS with default parameters outperforms previous methods if both sensitivity and specificity are simultaneously considered by the area under the receiver operating curve (AUC). We measured two quantities through the AUC: correctly detected expression changes versus wrongly detected (fold change) and correctly detected significantly different expressed genes in two sets of arrays versus wrongly detected (*P*-value). Furthermore FARMS is computationally less expensive then RMA, MAS and MBEI.

Availability: The FARMS **R** package is available from http://www.bioinf.jku.at/software/farms/farms.html

Contact: hochreit@bioinf.jku.at

Supplementary information: http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps

1 INTRODUCTION

The microarray technique is currently one of the most successful experimental tools in microbiological research. It extracts a gene expression profile from a tissue sample and, therefore, supplies the expression state of tens of thousands of genes. Microarray experiments can be used to infer metabolic pathways, to characterize protein–protein interactions or to extract target genes for developing therapies for various diseases (e.g. cancer). One of the leading microarray chip technologies (GeneChips) has been developed by Affymetrix and is considered here.

A GeneChip contains probe sets of 10–20 probe pairs representing unique genes. Each probe pair consists of two oligonucleotides of length 25, namely the perfect match (PM) and the mismatch (MM) probe. The perfect match probe is the exact complement of a 25 bp subsequence in the target gene. It is supposed to bind a labeled RNA (hybridization) which is obtained from the gene's mRNA in the tissue sample. The mismatch is identical to the perfect match except that one base is changed at the center position of the oligonucleotide leading to lower affinity to the gene's labeled RNA. Mismatches are supposed to detect non-specific hybridization.

The data recorded with the microarray technique are characterized by high levels of noise induced by the preparation, hybridization and measurement processes. Noise originates from chip fabrication tolerances, tolerances in the efficiency of RNA extraction and reverse transcription, background intensity fluctuations, non-uniform target labeling, temperature fluctuations, pipette errors, hybridization efficiency and scanning deviations. Also biological effects may disturb the target signal in the data, e.g. tissue samples from the same experimental condition may not show equal levels of RNA.

In order to analyze and evaluate GeneChip data from an experiment with multiple arrays, the data preprocessing at probe-level is a crucial step. An expression summary value is calculated using a four-step procedure. (1) 'Background correction', which removes the unspecific background intensities of the scanner images; (2) 'normalization', which reduces the undesired non-biological differences between chips and normalizes the signal intensity of the arrays; (3) 'PM correction', which removes non-specific signal contributions such as unspecific binding or cross-hybridization from the PM probes and (4) 'summarization', which combines the multiple preprocessed probe intensities to a single expression value. Errors introduced in one of these steps may corrupt further processing, e.g. spurious correlation with target conditions may appear especially for few tissue samples (arrays) and large number genes. For new chip generations with more genes on a chip the probability of detecting random correlations increases and summarization techniques will become even more important. The probable number of random correlations is the number of genes multiplied by the probability of a random correlation for independent measurement noise. Recently the new generation of HGU_133+2 GeneChips has been introduced by Affymetrix which provides the coverage of the entire human genome on a single array. Here one chip contains more than 54 000 probe sets and 1 300 000 distinct oligonucleotides.

In this paper we focus on new techniques for summarization. The summarization method which comes with an Affymetrix scanner is the Affymetrix Microarray Suite 5.0 [MAS 5.0, Aff, (2001); Hubbell *et al*., 2002]. The two best known approaches to improve MAS 5.0 are the Model Based Expression Index [MBEI, Li and

*To whom correspondence should be adderessed.

Wong (2001)] and the Robust Multi-array Average [RMA, Irizarry *et al*. (2003a, b); Bolstad *et al*. (2003)]. The Affymetrix Microarray Suite 5.0 (http://www.affymetrix.com/support/technical/manuals.affx) provides a 'present call' for each gene to indicate whether the measurement is likely to contain signal rather than noise but disregards information available at the summarization step. In addition the relevance of a gene in a certain experimental setting is usually determined by how strongly it is expressed at the one or the other condition. This, however, may not be the best way to evaluate the chip data, because even if a signal is present and strong but Gaussian distributed, its 'information content' may be low [see Friedman and Tukey, 1974; Friedman and Stuetzle, 1981; Huber, 1985] and it may not be useful to distinguish between conditions. Here we propose a summarization method which supplies noise corrected measurement values and improved present calls for genes as well as quantitative measures for the 'relevance' of a gene in a given context. Benchmark results using datasets from the open challenge 'Affycomp II' http://affycomp.biostat.jhsph.edu, Cope *et al*., 2004 and the 'golden spike-in' dataset from Choe *et al*. (2005) show that FARMS performs better than state-of-the-art methods like MAS 5.0, MBEI and RMA.

2 FACTOR ANALYSIS FOR ROBUST MICROARRAY SUMMARIZATION (FARMS)

2.1 The model

2.1.1 The basic model Our approach to the summarization problem is based on a linear model with Gaussian noise. Denote the actually observed and to zero mean normalized log-PMs by $\boldsymbol{x}$ and the normalized log-RNA concentration in the hybridization mixture by z. Then we assume that the log-observations $\boldsymbol{x}$ depend on the true log-concentration z via

$$\boldsymbol{x} = \lambda z + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{x}, \lambda \in \mathbb{R}^n \tag{1}$$

and

$$z \sim \mathcal{N}(0,1), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}). \tag{2}$$

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multidimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ [$\mathcal{N}(0,1)$ is the one-dimensional standard Gaussian]. z is usually called a 'factor'. $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ is the diagonal noise covariance matrix while $\boldsymbol{\epsilon}$ and z are statistically independent. According to the model, the observation vector $\boldsymbol{x}$ is Gaussian distributed as shown in the following equation:

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \lambda\lambda^T + \boldsymbol{\Psi}). \tag{3}$$

Consequently, the PMs are log-normal distributed. The λ_j are the shape-parameters of the log-normal distribution for each PM_j. To introduce individual shape-parameter for the PMs is justified by the findings in Li and Wong (2001), where the authors found that probes of the same probe-set may have different response to the same RNA amount. In Li and Wong (2001) these probe-effects were consistent over various arrays which implies specific binding characteristics of the probes. However, λ_j subsumes also signal contributions via signal strength σ as seen in text before Equation (9), where we set $\lambda_j = \sigma + \tau_j$. Large signal leads to large σ which scales up the shape-parameter which in turn results in a more heavy tail and allows for higher PM values carrying a signal. In the following we will motivate our model assumptions and then describe how to use factor analysis to infer the 'summarized' values z from the multiple observations $\boldsymbol{x}$ for each array and gene.

2.1.2 Using PM values only and the assumption of Gaussian noise In this section we want to justify the model assumption, that the vector $\boldsymbol{x}$ is Gaussian distributed. In Naef *et al*. (2002) replicate experiments on different arrays were made and the PM values as well as the PM − MM values were analyzed. The authors found that the PM values ('PM') have lower noise at low intensity than PM minus MM ('PM − MM') whereas for intermediate and high intensities the noise levels for PM and PM − MM were similar. Therefore we will use in our model only PM measurements.

Naef *et al*. (2002) also found that the distribution p_{diff} of the difference $\log(\mathrm{PM}_x) - \log(\mathrm{PM}_y)$ (x and y denote arrays of replicate measurements) is Gaussian, where the width depends on the intensity of the probe. Let p_{pm} be the distribution of log(PM). If p_{diff} is Gaussian and the distribution p_{pm} symmetric around a mean value μ, then p_{pm} is a Gaussian. This can be derived by setting w.l.o.g. $\mu = 0$ (note that the difference of the log-PMs is considered) and

$$\begin{aligned} p_{\mathrm{diff}}(a) &= \int_{-\infty}^{\infty} p_{\mathrm{pm}}(b)\, p_{\mathrm{pm}}(a+b)db \\ &= \int_{-\infty}^{\infty} p_{\mathrm{pm}}(b')\, p_{\mathrm{pm}}(a-b')d(b'), \end{aligned} \tag{4}$$

where $b' = -b$ and where we used $p_{\mathrm{pm}}(-b') = p_{\mathrm{pm}}(b')$. Fourier transformation of both sides yields

$$\mathcal{F}(p_{\mathrm{diff}})(a) = (\mathcal{F}(p_{\mathrm{pm}})(a))^2. \tag{5}$$

Because the Fourier transformation of a Gaussian is a Gaussian and the square root of a Gaussian is also Gaussian, the above statement holds.

Freudenberg *et al*. (2004) also found log-transformed data are normally distributed using a probe-wise Shapiro–Wilk test. Using the Affymetrix HGU133A latin square dataset (cf. Section 3), we confirmed that the log-transformed perfect matches are closer to a Gaussian distribution than the original perfect matches (Fig. 1). In conclusion, the assumption of a Gaussian distribution for the $\log(\mathrm{PM}_x)$ values seems to be justified.

2.1.3 The factor model assumptions In this section we motivate our linear ansatz λz from Equation (1), where z is interpreted as the logarithm of the true amount of mRNA in the tissue sample. Consider one gene, N arrays i—one for each tissue sample—and n perfect matches PM_{ij}, $1 \leq j \leq n$, on each array i. For each array we have a true (ideal) signal s_i indicating the logarithm of the amount of mRNA from this gene which is present in the tissue sample. Let z_i be the signal s_i normalized to mean zero and variance 1, that is

$$s_i = z_i \sigma + \mu, \sigma > 0. \tag{6}$$

Now we assume that for each PM_{ij} the signal deviates by τ_j and γ_j from the true values σ and μ giving

$$S_{ij} = z_i(\sigma + \tau_j) + \mu + \gamma_j, \tag{7}$$

where we assume that both the τ_j and the γ_j a distributed with zero mean. The value $\sigma + \tau_j$ determines the variance of the j-th measurement PM_{*j} and $\mu + \gamma_j$ its mean, i.e. we assume that each oligonucleotide corresponding to PM_j has its own characteristics (e.g. hybridization efficiency or crosstalk). Adding the measurement

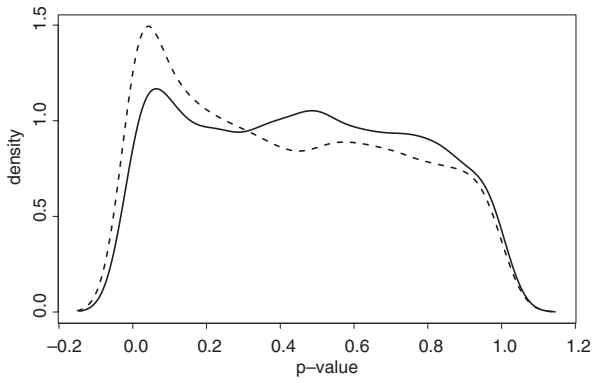

Fig. 1. Estimated density of *P*-values from the Shapiro–Wilk test for normality using 10 000 randomly selected PM intensities and 42 arrays from Affymetrix HGU133A latin square data. The continuous and dashed lines indicate the result for the $\log_2$-transformed and the original PMs, respectively. The deviation from a uniform distribution of the *P*-values indicates the deviation from Gaussian distributions. The $\log_2$-transformed PMs are closer to a Gaussian.

noise ϵ to S_{ij} gives

$$\log(\mathrm{PM}_{ij}) = S_{ij} + \epsilon_{ij} = z_i(\sigma + \tau_j) + \mu + \gamma_j + \epsilon_{ij}, \tag{8}$$

where ϵ_{ij} is a zero mean Gaussian (non-zero mean is accounted for by γ_j). The values τ_j, γ_j and the standard deviation of the ϵ_{ij} may depend on the gene's signal intensities for the arrays. This takes the findings in Chudin *et al*. (2001); Naef *et al*. (2002); and Tu *et al*. (2002) into account, that the variance of the noise depends on the signal strength. Therefore, estimated values are only valid for the measurements under considerations, i.e. the actual signal strength.

If we set $\lambda_j = \sigma + \tau_j$ and normalize the observation $\boldsymbol{x}$ to zero mean by subtracting

$$\begin{aligned} \tfrac{1}{N}\sum_{i=1}^{N}\log(\mathrm{PM}_{ij}) &= (\sigma+\tau_j)\,\tfrac{1}{N}\left(\sum_{i=1}^{N} z_i\right) \\ &\quad + \mu + \gamma_j + \tfrac{1}{N}\left(\sum_{i=1}^{N}\epsilon_{ij}\right) \\ &\approx \mu + \gamma_j = \mu_j, \end{aligned} \tag{9}$$

where the approximation is due to the zero mean assumptions then we arrive at Equation (1), the basic model. According to the model assumptions, $z \sim \mathcal{N}(0,1)$ [Equation (2)], our approach is best suited for genes with strong Gaussian distributed signal or for genes with low signal intensities (small σ), because the Gaussian noise is superimposed on the weak signal. The Gaussian signal assumption is justified for the majority of genes which are independent of the conditions, however it is not justified for the genes conveying a non-Gaussian signal. It will turn out that the model-based approach also provides good results for non-Gaussian distributions of z, because the non-Gaussianity of z has only a minor impact on the model likelihood as we will see at the end of Subsection 2.2.2.

2.2 Estimation of model parameters and signal

We now describe how to estimate the true signal strengths based on the data model of Section 2.1. The procedure consists of three steps:

(1) normalization of the observations to zero mean [cf. Equation (9)]
(2) the maximum a posteriori factor analysis to estimate model parameters λ_j in order to calculate σ and
(3) recovering the true signals s_i [Equation (6)] from z_i,

which we will describe in the following text.

2.2.1 Normalization of the observations In order to fulfill model assumptions, the log-PM values are normalized to zero mean by subtracting $\mu_j = \mu + \gamma_j$ which is estimated using Equation (9).

2.2.2 Maximum a posteriori factor analysis The Bayesian posterior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\})$ of the model parameters $(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ given the dataset $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is proportional to the product of the observation's likelihood $p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})$ of data $\{\boldsymbol{x}\}$ given the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}$ multiplied by the prior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ (e.g. DeGroot, 1970):

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \propto p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\ p(\boldsymbol{\lambda}, \boldsymbol{\Psi}). \tag{10}$$

For the prior we assume that $p(\boldsymbol{\lambda}, \boldsymbol{\Psi}) = p(\boldsymbol{\lambda})$, i.e. that the prior for the factor loadings $\boldsymbol{\lambda}$ is independent from the prior for $\boldsymbol{\Psi}$ and that the latter is uninformative (i.e. flat). The prior for $\boldsymbol{\lambda}$ is $p(\boldsymbol{\lambda}) = \prod_{j=1}^{n} p(\lambda_j)$ and for $p(\lambda_j)$ we choose the rectified Gaussian distribution $\mathcal{N}_{\mathrm{rect}}(\mu_\lambda, \sigma_\lambda)$ (see Hinton and Ghahramani, 1997) given by

$$\lambda_j = \max\{y_j, 0\} \ \text{ with } \ y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda). \tag{11}$$

σ_λ is chosen proportional to the mean of the variance $\mathrm{Var}(x_{*j})$ of the observations to allow the factor to explain the data variance, that is

$$\sigma_\lambda^2 = \rho\,\frac{1}{n}\sum_{j=1}^{n}\mathrm{Var}(x_{*j}). \tag{12}$$

The prior reflects the facts that

(1) the observed variance in the data is often low which makes high values of λ_j unlikely,
(2) a chip typically contains many more genes with constant signal ($\lambda_j \sim 0$) than genes with variable signal (large value of λ_j),
(3) negative values of λ_j are not plausible, because that would mean that increasing mRNA concentrations lead to smaller signal intensities.

The two hyperparameters ρ and μ_λ allow quantifying different aspects of potential prior knowledge. For example, μ_λ near zero assumes that most genes do not contain a signal and introduces a bias for λ-values near zero (items 1 and 2 from above).

The second factor of the posterior is the likelihood which is according to Equation (3)

$$p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \prod_{i=1}^{N}\mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi})\ (\boldsymbol{x}_i), \tag{13}$$

where $\mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi})\ (\boldsymbol{x}_i)$ is the distribution's density evaluated at $\boldsymbol{x}_i$.

Following Rubin and Thayer (1982), we estimate the parameters of the factor analysis model with the expectation-maximization (EM) algorithm of Dempster *et al*. (1977) modified to maximize the Bayesian posterior, Equation (10), of the model parameters given the data. The EM procedure estimates $\boldsymbol{\lambda}$, $\boldsymbol{\Psi}$ and the posterior values for z for every $\boldsymbol{x}$. Analogous to the EM algorithm for

maximum likelihood, the EM algorithm maximizes a lower bound of the log-posterior

$$-\frac{1}{2}\boldsymbol{\sigma}_{\lambda}^{-2}(\boldsymbol{\lambda}-\mu_{\lambda}\mathbf{1})^{T}(\boldsymbol{\lambda}-\mu_{\lambda}\mathbf{1}) + \frac{nN}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{N} E_{z_i|x_i}((\boldsymbol{x}_i - \boldsymbol{\lambda}z_i)^{T}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\lambda}z_i)), \quad (14)$$

where $\boldsymbol{x}$ is already normalized to mean zero and

$$z_i \,|\, \boldsymbol{x}_i \sim \mathcal{N}(\mu_{z_i|x_i}, \sigma^2_{z_i|x_i}),$$
$$\mu_{z_i|x_i} = (\boldsymbol{x}_i)^{T}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}+\boldsymbol{\Psi})^{-1}\boldsymbol{\lambda} \quad \text{and} \quad (15)$$
$$\sigma^2_{z_i|x_i} = 1 - \boldsymbol{\lambda}^{T}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}+\boldsymbol{\Psi})^{-1}\boldsymbol{\lambda}.$$

A detailed derivation of both the lower bound and the complete EM algorithm can be found in the supplementary information (http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps).

Note that the maximum a posteriori factor analysis is also able to extract non-Gaussian signals. The likelihood covariance matrix is $\boldsymbol{\lambda}\boldsymbol{\lambda}^{T} + \boldsymbol{\Psi}$, therefore increasing the diagonal elements of $\boldsymbol{\Psi}$ would lead to a larger decrease of the likelihood than increasing one eigenvalue via $\boldsymbol{\lambda}\boldsymbol{\lambda}^{T}$ (note that scaling a non-Gaussian to variance one increases $\boldsymbol{\lambda}$). Reason for the larger decrease of the likelihood in the first case is the cumulative effect of increasing n eigenvalues of the covariance matrix. Therefore, explaining data variance by a non-Gaussian factor has higher likelihood than explaining it by n measurement noise corrections.

2.2.3 Estimation of the true signals Finally we need to recover the 'true' signal s_i from the estimated values z_i, i.e. we need to estimate σ and μ in Equations (6) and (8). For each perfect match we have

$$\sigma = \lambda_j - \tau_j \;\text{ and }\; \mu = \mu_j - \gamma_j. \quad (16)$$

We determine σ and μ with the least squares fit, which is unbiased because we assumed in Subsection 2.1.3 that both τ_j and γ_j are drawn from a distribution with zero mean:

$$\sigma = \text{argmin}_{\tilde{\sigma}} \sum_{i=1}^{n}(\lambda_j - \tilde{\sigma})^2 = \frac{1}{n}\sum_{j=1}^{n}\lambda_j, \quad (17)$$

$$\mu = \text{argmin}_{\tilde{\mu}} \sum_{j=1}^{n}(\mu_j - \tilde{\mu})^2 = \frac{1}{n}\sum_{j=1}^{n}\mu_j. \quad (18)$$

The 'true' signal is then computed as

$$s_i = \sigma\, z_i\, f + \mu, \quad (19)$$

where f is a factor which compensates for the reduction of variance during preprocessing and factor analysis (some of the data variance is explained by the noise). The value of f is empirically determined on toy data for different normalization procedures: 2.0 for quantile normalization and 1.5 for cyclic loess (see Section 3.2 for the normalization procedures). Note that the factor f does not influence the AUC-values which we used to evaluate the different methods in Section 3.

We call the new summarization procedure which has been described 'Factor Analysis for Robust Microarray Summarization' (FARMS).

2.3 Extraction of the relevant genes

Using factor analysis we estimated the 'true' signals s_i. Their actual strengths, i.e. the value of σ, can be taken as a measure of the potential relevance of a gene in a given experimental setting: high value of σ indicates more relevant genes. A complementary and in several cases even better criterion, however, can be derived via the factor z and its distribution across arrays. Following the idea of projection pursuit of Friedman and Tukey (1974); Friedman and Stuetzle (1981); Huber (1985) interesting or 'relevant' variables are often not Gaussian distributed. This assumption is especially true for most microarray experiment designs, where genes are of interest if their expression levels are correlated with different experimental conditions. Often two conditions must be distinguished, thus genes which show a bimodal rather than a Gaussian distribution are of interest because they may be correlated with the conditions. But also for a larger number of conditions one would expect that non-Gaussianity is a good indicator for relevance. A quantitive measure can be obtained by a test of Gaussianity for the estimated variables z through the Shapiro–Wilk test (more robust in the case of a small sample size than the Kolmogorov–Smirnov test). FARMS is especially suited for this test because it assumes a Gaussian signal, thus violating this assumption indicates a strong signal. Genes can be ranked according to their σ-values or according to their non-Gaussianity, and the top candidates can then be investigated further.

3 EXPERIMENTS AND RESULTS

3.1 Datasets

For the following benchmarks we use four well-known evaluation datasets denoted by (A), (B), (C) and (D) which were produced by controlled experiments with known target expression values or known mutual relations. The first three datasets are from the open challenge 'Affycomp II' (http://affycomp.biostat.jhsph.edu/, Cope *et al*., 2004) whereas the fourth dataset is known as the 'golden spike-in' dataset from Choe *et al*. (2005).

Dataset A. This dataset is the original assessment dataset in Cope *et al*. (2004). It consists of two sub datasets with the Affymetrix human HGU95A array: the spike-in experiments and the dilution experiments.

For the first, spike-in dataset A1, the concentration of RNA for 14 genes, the so-called spike-in genes, was artificially controlled by adding RNA with predefined concentrations to the hybridization mixture. The 'latin square design' contained 20 experiments with different RNA concentrations of the 14 spike-in genes chosen from {0.0, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0, 256.0, 512.0, 1024.0} pM. For each experiment two replicate arrays were prepared except one with only two replicates. The datasets consist of 59 arrays stored in 'CEL' files. A 'CEL' file gives the 75 percentile pixel intensity of each spot, i.e. each gene in the array image.

The second, dilution dataset A2 from GeneLogic uses two tissue samples, human liver (HL) and human central nervous system (CNS), from which the RNAs were hybridized to the 75 HGU95A_v2 arrays. The dataset is based on changing dilutions (concentrations) and combinations of RNA taken from the two different tissues. Arrays are hybridized to a mixture of HL and CNS where the amount of RNA taken from each source is one from the six values {1.25, 2.5, 5.0, 7.5, 10.0, 20.0} µg. Each dilution

experiment is replicated five times and each replicate was evaluated on a different scanner.

Dataset B. This dataset is the first part of the new assessment from http://affycomp.biostat.jhsph.edu/. It is identical to dataset A1 but separately listed because of the separate Affycomp evaluation results.

Dataset C. This dataset is the second part of the new assessment from http://affycomp.biostat.jhsph.edu/. It is based on a 'latin square' experimental design which consists of 42 HGU133A arrays, with 42 spike-in genes with RNA concentrations from {0.0, 0.0125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0, 256.0, 512.0} pM. Here three spike-in genes of the same concentration were combined in order to create three replicates for each experiment.

Dataset D. Recently Choe *et al*. (2005) supplied a dataset consisting of six Affymetrix DrosGenome1 chips. This dataset mimics a common used microarray experimental setting, where two samples, i.e. a treatment and a control sample are compared in order to identify differentially expressed genes. The array can detect 3860 known individual RNA samples together with 2551 RNA samples as controls (and background) where the latter have the same concentration in all experiments. A total of 1309 RNAs samples mimic the differentially expressed genes, these RNAs were split into 8 subsets of about 80 to 180 RNAs. Each subset differs by one predefined relative concentration change from {1.2, 1.5, 1.7, 2.0, 2.5, 3.0, 3.5, 4.0} between the spike-in and control sample. Finally, the spike-in and control sample were hybridized in triplicates.

3.2 Benchmark details

We compare our method, FARMS, to the three best known summarization methods MAS5, MBEI and RMA as well with the 43 methods which participated at the challenge 'Affycomp II' (as of October 7, 2005). Microarray Suite (MAS) 5.0 is a non-parametric algorithm implemented by Affymetrix (Aff, 2001; Hubbell *et al*., 2002). The Model Based Expression Index [MBEI, Li and Wong (2001)] is like the Robust Multi-array Average [RMA, Irizarry *et al*. (2003a, b); Bolstad *et al*. (2003)] a model-based approach (software packages are available at http://www.dchip.org or www.bioconductor.org).

FARMS does not use background correction and uses either quantile normalization (Bolstad *et al*., 2003) or cyclic loess (Yang *et al*., 2002; Dudoit *et al*., 2002). FARMS uses quantile normalization as default normalization procedure because it is computational efficient. It does not apply PM corrections and uses PMs only. For all experiments with FARMS we set $\rho = 1/8$, $\mu_\lambda = 0$ and $f = 2.0$ for quantile normalization and $f = 1.5$ for cyclic loess. The maximal cycles for factor analysis were fixed to 100 and factor analysis was terminated if the λ-update vector has length smaller than 0.00001.

RMA can be improved through advanced background correction leading to a method called GCRMA (Wu *et al*., 2004, Available at http://ideas.repec.org/p/bep/jhubio/1001.html). GCRMA has lower performance on datasets A–C with respect to the AUC-values than FARMS as can be seen in the supplementary information but is superior to RMA. For our FARMS method we did not use background correction, however in future studies we want to investigate whether background correction can improve our FARMS method especially whether the GCRMA background corrections is suitable.

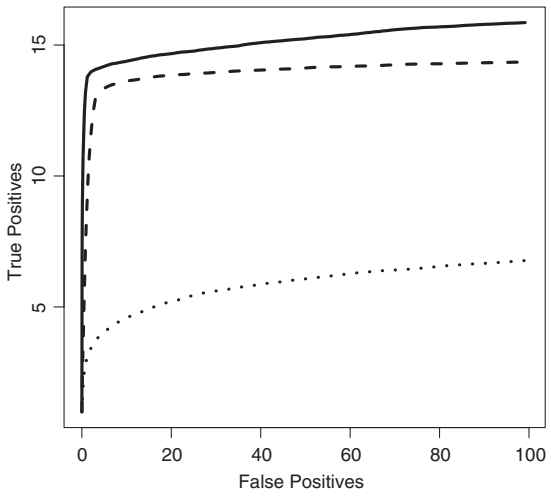

Fig. 2. ROC curves for all fold changes in dataset A1. ROC curve for FARMS with quantile normalization (solid line) is always above the ROC curve for RMA (dashed line) and MAS 5.0 (dotted line), therefore FARMS is better than RMA and MAS 5.0 for all false positive rates.

3.3 Results

For the evaluation of datasets A, B and C, we participated at the 'Affycomp II' challenge (http://affycomp.biostat.jhsph.edu/, Cope *et al*., 2004). For the complete challenge results see Tables 1–3 in the supplementary information (http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps).

3.3.1 AUC fold changes We think that from all challenge results the area under the curve (AUC) criterion is best suited to measure the quality of a summarization method. The AUC criterion is the area under the receiver operating characteristics (ROC) curve which plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) and serves a quality measure for classification methods. The AUC criterion can be applied here by defining gene classes: for a pair of arrays class 1 genes are the genes for which expression value differences exceed a certain relative factor (fold change). Now the output of a summarization method can be interpreted as classification by computing the class membership of genes based on the predicted expression values. We prefer the AUC criterion over other measures provided by 'Affycomp II' evaluation because it is independent of scaling of the results (log-expression values) and trades sensitivity against specificity. Other quality measures from the 'Affycomp II' evaluation focus either on sensitivity or specificity and are often not scaling independent. The AUC is computed for different fold changes, i.e. for different thresholds for being in class 1. Figures 2–4 show the fold change ROC curves for A1, C and D, respectively. Table 1 gives the corresponding AUC for datasets A–D. Note, that dataset D is especially suited to generate precise ROC curves because of the large number of defined RNAs. Except for dataset A, FARMS has the best AUC performance of the 43 competitors of the 'Affycomp II' challenge (the challenge method which has higher AUC values than FARMS in dataset A has lower AUC values for datasets B and C).

FARMS with quantile normalization is best for datasets A–B, whereas FARMS with cyclic loess is best for dataset D. However, both FARMS methods show higher performance than all its

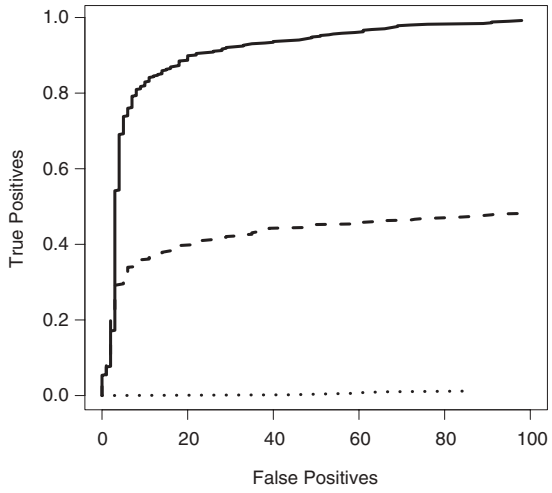

Fig. 3. ROC curve for fold changes of low intensity genes in dataset C. ROC curve for FARMS with quantile normalization (solid line) is always above the ROC curve for RMA (dashed line) and MAS 5.0 (dotted line) therefore FARMS is better than RMA and MAS 5.0 for all false positive rates.

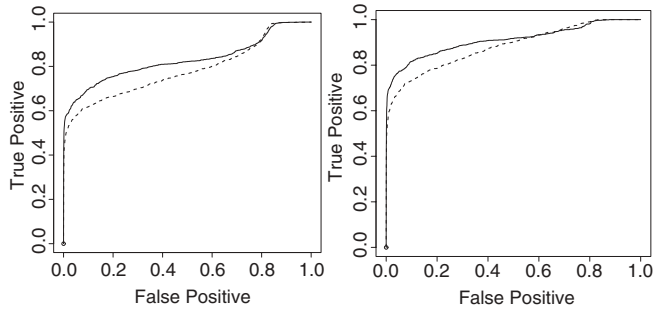

Fig. 4. Dataset D ROC curves for fold changes ⩾1.2 (left) and ⩾1.7 (right) for RMA (dashed line) versus FARMS (solid line). In both cases FARMS performs better than RMA as its ROC curve is above RMA's.

competitors (RMA, MAS and MBEI). FARMS shows a large improvement over RMA for small signal changes: for dataset A and fold change 2 the AUC value is 0.54 for RMA, 0.84 for FARMS (quantile normalization) and 0.78 for FARMS (cyclic loess), for dataset B the low intensity AUC is 0.51 for RMA, 0.89 for FARMS (quantile) and 0.80 FARMS (cyclic loess), and for dataset C the low intensity AUC is 0.57 for RMA, 0.94 for FARMS (quantile) and 0.91 for FARMS (cyclic loess). The AUC for random guessing is 0.5 and the maximal AUC is 1.0, therefore the improvement of FARMS over RMA is considerable.

3.3.2 AUC P-values Above AUCs for fold changes assess the quality of summarization methods with respect to the identification of differentially expressed genes in a pair of arrays. Here we want to go one step further and determine the quality of summarization methods with respect to the identification of significant differentially expressed genes in two conditions. To perform a significance test to the expression values in two conditions is a common experimental setting in biology and in medicine, therefore we evaluate the quality of different summarization methods by wrongly detected significant differences and missed differences.

Table 1. AUC results for fold changes for datasets A–D

AUC	FARMS q	FARMS l	RMA	MAS 5.0	MBEI	1	2	mean
FC Dataset A								
all	0.89	0.85	0.82	0.36	0.67	**0.91**	0.86	0.71
=2	0.84	0.78	0.54	0.07	0.17	**0.91**	0.69	0.42
1 Dataset B								
Low	**0.89**	0.80	0.51	0.07	0.21	0.74	0.68	0.44
Med	0.97	0.95	0.91	0.00	0.43	**0.98**	0.97	0.65
High	**0.97**	0.94	0.64	0.00	0.16	0.95	0.94	0.48
Mean	**0.91**	0.84	0.60	0.05	0.26	0.79	0.75	0.49
1 Dataset C								
Low	**0.94**	0.91	0.57	0.09	—	0.76	0.61	0.48
Med	**0.99**	**0.99**	0.91	0.00	—	0.95	0.95	0.64
High	**1.00**	**1.00**	0.96	0.00	—	0.99	0.99	0.61
Mean	**0.95**	0.93	0.65	0.06	—	0.81	0.66	0.44
FC Dataset D								
≥1.2	0.72	**0.74**	0.70	0.52	0.49	—	—	—
≥1.7	0.90	**0.91**	0.88	0.64	0.59	—	—	—

We compare FARMS with RMA, MAS 5.0 and MBEI and for dataset A–C also with 43 competitors from the *affycomp* Bioconductor Project benchmark where the best ('1'), the second best ('2') and the mean results ('mean') are given (as of October 7, 2005). FARMS results are reported for quantile normalization ('q') and for cyclic loess ('l'). The table reports AUC values for different fold changes ('FC', datasets A and D), i.e. detection of different concentrations changes, as well as different signal intensities ('I', datasets B and C). The best result is marked bold.

Analogous to the AUC for fold changes we define an AUC for *P*-values. Class 1 genes are the genes which have by design different expression values in the two conditions. A summarization method classifies a gene as being differently expressed in the two conditions if the *P*-value of a test is below a given threshold (we set it to 0.05). This allows us to compute the ROC curve.

A significance test, a modified *t*-test, for differentially expressed genes for microarray experiments with two conditions was suggested by Tusher *et al*. (2001). In the modified *t*-test a small positive constant ('fudge-constant') is added to the denominator to prevent genes with small variance from being selected as significant. According to Cui and Churchill (2003) we set the 'fudge-constant' to the 90th percentile of the standard deviation of all genes.

Datasets B and C encompass 19 and 14 experimental conditions, respectively, with 3 replicates for each condition. This leads to 171 and 91 experimental condition pairs (only unique variations), respectively, with 6 arrays (3 for each condition) for each experimental setting. The above-mentioned modified *t*-test is applied to these 171 (dataset B) and 91 (dataset C) experimental settings. The average AUC of all ROC curves for *P*-values is given in Table 2. For dataset B the average AUC for RMA is larger than for FARMS but the difference is not significant as confirmed by Wilcoxon-rank-sum test ($P = 0.19$). For dataset C FARMS shows significantly by ($P = 0.00027$) better results than RMA. Most reliable are the results on dataset D, where the number of defined RNAs is large. However, for dataset D there is only one experiment so that the Wilcoxon-rank-sum test cannot be applied, but the large number of spike-in genes allows to perform another test, the conservative McNemar test. It confirmed that FARMS performed significantly by better ($P = 0.000002$) than its competitors.

Table 2. AUC results for *P*-values for datasets B–D

	FARMS q	FARMS l	RMA	MAS 5.0	MBEI
Dataset B					
AUC [$e = 171$]	**0.955**	**0.955**	0.948	0.772	0.670
Dataset C					
AUC [$e = 91$]	**0.975**	**0.974**	**0.981**	0.892	0.875
Dataset D					
AUC [$e = 1$]	**0.802**	**0.823**	0.767	0.286	0.397

We compare FARMS with RMA, MAS 5.0 and MBEI on *e* experiments (one experiment consists of 6 arrays—3 arrays for each condition). Results which are significantly better than others are marked bold, where mutual differences between bold results are not significant.

Table 3. Computational time in (s) for dataset A2

	FARMS	RMA	MAS 5.0	MBEI
Computational time	**246**	472	1323	957

3.3.3 Computational time: The computational time of FARMS (quantile normalization), RMA, MAS 5.0 and MBEI is listed in Table 3. FARMS is the fastest method.

In conclusion FARMS performs better than all competitors with respect to the AUC criterion for fold changes as well as for *P*-values and was the fastest method.

4 CONCLUSION

We have presented a new method called FARMS for summarization of gene expression data obtained from Affymetrix chips. The new method outperforms known methods both with respect to sensitivity and specificity, i.e. detects more signals while being more robust against measurement noise. Further it is faster than the competitors.

ACKNOWLEDGEMENTS

The authors express their gratitude for the funding by the Anna-Geissler- and the Monika-Kutzner-Stiftung.

Conflict of Interest: none declared.

REFERENCES

Microarray Suite User Guide. (2001) Affymetrix, version 5 edition.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Choe,S.E. *et al*. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol*., **6**, R16.1–R16.16.

Chudin,E. *et al*. (2001) Assessment of the relationship between signal transcript concentration for Affymetrix GeneChip arrays. *Genome Biol*., **3**, research0005.1–0005.10.

Cope,L.M. *et al*. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Cui,X. and Churchill,G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.1–210.10.

DeGroot,M.H. (1970) *Optimal Statistical Decisions*. McGraw-Hill, NY.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–22.

Dudoit,S. *et al*. (2002) Statistical methods for identifying genes with differential expression in replicate cDNA microarray experiments. *Stat. Sin*., **12**, 111–139.

Freudenberg,J. *et al*. (2004) Comparison of preprocessing procedures for oligonucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. *Meth. Inform. Med*., **43**, 434–438.

Friedman,J.H. and Stuetzle,W. (1981) Projection pursuit regression. *J. Am. Stat. Assoc*., **76**, 817–823.

Friedman,J.H. and Tukey,J.W. (1974) A Projection Pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput*., **23**, 881–890.

Hinton,G.E. and Ghahramani,Z. (1997) Generative models for discovering sparse distributed representations. *Pilos. Trans. R. Soc. B*, **352**, 1177–1190.

Hubbell,E. *et al*. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.

Huber,P.J. (1985) Projection pursuit. *Ann. Stat*., **13**, 435–525.

Irizarry,R.A. *et al*. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, 1–8.

Irizarry,R.A. *et al*. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Li,C. and Wong,W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Naef,F. *et al*. (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol*., **3**, research0018.1–0018.11.

Rubin,D. and Thayer,D. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.

Tu,Y. *et al*. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl Acad. Sci. USA*, **99**, 14031–14036.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wu,Z., Irizarry,R., Gentleman,R., Murillo,F.M. and Spencer,F. (2004) A model based background adjustment for oligonucleotide expression arrays. Johns Hopkins University Dept. of Biostatistics Working Paper Series 1001, Berkeley Electronic Press.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.