Open Access

Approximation by Cubic Splines Leads to Highly Specific Discovery by Microarrays

Jerry L. Bona[1] and Hassan M. Fathallah-Shaykh*[,1,2]

[1]*Department of Mathematics, Statistics, & Computer Science, The University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607, USA;* [2]*Department of Neurological Sciences, Section of Neuro-Oncology, Rush University Medical Center, 1725 West Harrison Street, Chicago, IL, 60612, USA*

Abstract: Genome-scale microarray datasets are noisy. We have previously reported an algorithm that yields highly specific genome-scale discovery of states of genetic expression. In its original implementation, the algorithm computes parameters by globally fitting data to a function containing a linear combination of elements that are similar to the Hill equation and the Michaelis-Menten differential equation. In this essay, we show that approximation by cubic splines yields curves that are closer to the datasets, but, in general, the first derivatives of the cubic splines are not as smooth as the derivatives obtained by global fitting. Nonetheless, little variation of the first derivative is seen in the area of the curve where the Cutoff Rank is computed. The results demonstrate that piece-wise approximation by cubic splines yields sensitivity and specificity equal to those obtained by global fitting.

INTRODUCTION

The genomes of several organisms have recently been sequenced [1-3]. Microarray studies are increasingly being used to explore biological causes and effects and even to diagnose diseases; however, the data resulting from microarray analysis is very noisy and the patterns of expression and molecular signatures of microarrays are not always reproducible [4-6].

MASH is a mathematical algorithm that yields highly accurate predictions of states of genetic expression (up or down-regulation) from the genome-scale profiling of two samples [7]. "Accuracy" in this context refers to the very low incidence of false discovery rates delivered by the procedure; the term "false discovery rate" indicates the number of false positives (i.e. genes whose states of genetic expression are falsely discovered to be differentially expressed) divided by the total number of true negatives (i.e. genes whose states of expression are truly not differentially expressed between samples). Specifically, the false discovery rates of MASH and MIDAS (http://www.tm4.org/midas.html) in same-to-same comparisons using 19K microarrays are 1/192,000 versus 1,347/192,000 measurements, respectively (see [7]).

MASH computes parameters by globally fitting the data to a function including a linear combination of elements that are similar to the Hill equation and the Michaelis-Menten differential equation. The term "global fitting" is used in the sense of fitting the entire dataset to a single equation. We will refer to this function as "the equation" (see Fig. **1**). The Hill equation describes the fraction of the enzyme saturated by the ligand as a function of ligand concentration. It is of the form, $\theta = \frac{[L]^n}{(K_A)^n + [L]^n}$, where θ, $[L]$ K_A, and n indicate the fraction of ligand binding sites filled, the ligand concentration, the ligand concentration producing half occupancy, and the Hill coefficient respectively. The Michaelis-Menten equation describes the kinetics of enzymes. It is of the form, $V = V_{\max}\frac{[S]}{K_m + [S]}$; where V, K_m, and $[S]$ indicate reaction velocity, a constant, and substrate concentration.

Approximation by polynomials would be much simpler, but we found that global determination of something like (1) using a polynomial required a very large number of parameters and the resulting curve fits were poor (i.e. not close to the datasets). There is another, well-known alternative, however; approximation by piecewise polynomials (spline approximation for example).

Our goal in the present essay is to use piecewise polynomial approximation in place of (1) and to compare the results to the MASH algorithm. An advantage of piecewise polynomial approximation is that it can be implemented more or less automatically. The disadvantage at the outset is that many more parameters are needed to determine such an approximation. We swill show that both methods have advantages but that the final predictions by either are virtually identical.

MATERIALS AND METHODS

Microarrays

Normal brain RNA is obtained by pooling RNA from human occipital lobes harvested from 4 individuals with no known neurological disease whose brains are frozen less than 3 hours postmortem [7-10]. The quality of RNA is assayed by gel electrophoresis and only high quality RNA is processed. Microarray chips whose trade names are 1.7K and

*Address correspondence to this author at the University of Alabama at Birmingham, Department of Neurology, Faculty Office Towers 1020, 510 20th Street South, Birmingham, AL 35294-3410, USA; Tel: 205-934-1432; Fax: 205-975-7546; E-mail: hfathallah@msn.com

1875-0362/08 2008 Bentham Open

19K are purchased from the Ontario Cancer Institute (Ontario, Canada). The 1.7K and 19K microarrays contain 1,920 and 19,200 cDNAs spotted in duplicates, respectively. The experimental design, which includes dye swapping as described later, generates four replicate measurements per gene and sample [8, 9]. Each slide contains two replicate adjacent spots. The Cy3/Cy5 design generates two ratios. The Cy5/Cy3 design generates two additional ratios. The total is four replicate ratios with dye-swapping. RNA used in spike-in experiments is transcribed from the same Arabidopsis cDNA spotted on the microarray slides.

Software Specifications

The mathematical analysis, to be described presently, is performed using functions written in Matlab including the spline and optimization toolboxes (Mathworks, Natick, MA). The outcome is a new piece of software called MASH2 that uses smoothing splines (called MASHsm below). It is freely available to academics for noncommercial use. To obtain executable software, please send a request by e-mail. The datasets may be downloaded from http://www.rushu.rush.edu/neurosci/Fathallah.html.

RESULTS AND DISCUSSION

Definitions

The state of genetic expression of a spot in sample A vs. sample B assayed by cDNA arrays is measured by the ratio of the background-subtracted intensities of sample A divided by background-subtracted intensities of sample B. A ratio $r > 1$ ($\log_2(r) > 0$) is taken to mean upregulation of the gene in sample A as compared to B. The terms 'genes,' 'spots,' 'symmetrical,' and 'rank,' are defined in terms of the 1.7K arrays; these terms are also applicable to other microarrays. As mentioned above, the 1.7K microarray contains 1,920 cDNAs or controls, here referred to as *genes*, spotted in duplicated to a total of 3,840 *spots*. The term *symmetrical* image refers to the two images, corresponding to the Cy3 and Cy5 fluorescent dyes, generated from a single microarray slide. Background-subtracted spot intensities are sorted in ascending order, thereby assigning a *Rank* to every spot. For instance, a spot whose rank is 3000 has a higher background-subtracted spot intensity than do all spots whose ranks are less than 3000.

Dye swapping refers to experiments where the Cy3 and Cy5 dyes are swapped between the 2 samples; this is done to annul confounding variables introduced by heterogeneous fluorescence of the Cy5 and Cy3 molecules. Each microarray slide generates a set of symmetrical Cy3/Cy5 images that yield 2 replicate ratios. Thus, each dye swapping dataset generates 4 replicate ratios.

The Datasets and Rationale

The true negative datasets are assembled by comparing the pool of brain RNA to itself (same-to-same). The goal of the same-to-same comparisons is to form an appreciation of the experimental noise, generated in the main by technical artifacts, independent of biological heterogeneity. In this phase of the experiment, normalized expression ratios $\neq 1$ ($\log_2 \neq 0$) are false positive (noise) because the Cy3/Cy5 symmetrical images contain identical genetic information. The artifactual measurements may be caused by several factors including slide-to-slide differences, variations in the reverse transcription reactions, hybridization, labeling, and variations due to the laser. The datasets are normalized by a nonlinear method described elsewhere [7].

The *same-to-same* comparisons include 18 and 20 experiments that generate a total of 9 and 10 dye swapping data sets using the human 1.7K and 19K microarrays, respectively. The experiments are paired by consecutive order. The goal is to filter the largest number of non-zero same-to-same expression ratios originating from technical noise.

The 1.7K microarray includes 64 genes of Arabidopsis cDNA. The true positive datasets include 4 sets of spike-in dye swapping experiments using 1.7K microarrays, where 1 ng of Arabidopsis RNA is added to one sample but not the other. In this design, all 64 genes of Arabidopsis cDNA serve as true positives. MASH, in its original implementation, detects the states of genetic expression of 26/64 Arabidopsis cDNA; its sensitivity is thus 41% [7].

Approximation and Computing the CR

Fig. (**1a**) is a typical plot of the expression $g(x)$ that approximates the log-transformed and sorted background-subtracted spot intensities [7]. The equation fits not only our data of 60/60 1.7K data sets, 200/200 human 19K data sets (38,400 spots on 2 separate slides P1 and P2), but also microarray datasets acquired in independent laboratories. Specifically, $g(x)$ fits all 266 curves resulting from the 133 publicly available arrays from the lymphoma study by Alizadeh *et al*. (R-square > 0.99) [11] (see [7] for details). Each curve-fit generates a unique set of parameters $(a_1,...,a_{19})$ which determine the function in (1). The parameters are obtained *via* lsqcurvefit (MATLAB, optimization toolbox), which solves the nonlinear curve-fitting problem in the least squares sense.

The equation is constructed to fit the microarray datasets that consist of 3 parts: 1) an initial segment where spot intensities rise rapidly, 2) a second almost 'linear' section associated with small increments, and 3) a final 'exponentially-growing' phase (see [7] for details on the construction of the equation and the curve fits). The rate of increase of the first segment is maximal at the point of inflection that corresponds to the maximum of $g'(x)$ in that segment (Fig. **1b**). The 1.7K chips contain 'buffer' spots containing no cDNA, which are expected to generate the lowest intensities caused by nonspecific binding of the probes to glass or buffer. The y-coordinate at the Inflection Rank corresponds to a small background-subtracted intensity, ranging from 50–150, most probably generated by non-specific probe binding. We tentatively conclude that the majority of intensities whose ranks are smaller than the point of inflection are likely caused by nonspecific binding of the probe.

Let n be the total number of spots. The *Cutoff Rank* (*CR*) is defined as the rank such that:

$$g'(CR) = \delta_{optimal} * \frac{g(n)}{n}$$

where $\delta_{optimal} = 0.36$ has been established empirically to optimize sensitivity without lowering specificity [7]. The CR is

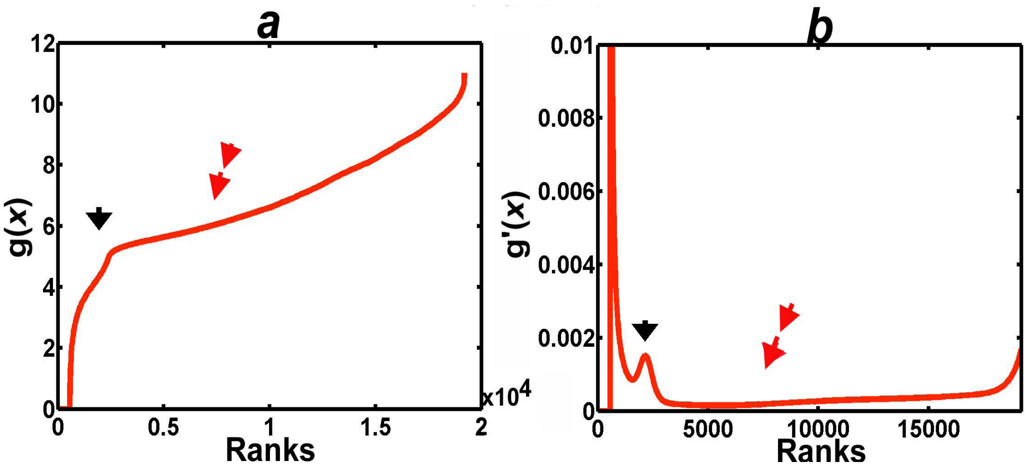


Fig. (1). ***Typical plot and computing the CR.*** **(a)** shows a plot of the function:

$$g(x)=\left(\frac{a_1*x}{x+a_2}+\frac{x}{ns-x+a_3}-a_4\right)*a_5*\left(\frac{1}{1+\left(\frac{a_7}{x}\right)^{a_6}}+\frac{a_8}{1+\left(\frac{a_{10}}{x}\right)^{a_9}}-\frac{a_{11}}{x+a_{12}}\right)*\left(1+\frac{a_{13}}{1+\left|1-\frac{a_{15}}{x}\right|^{a_{14}}}\right)+\left(\frac{1}{1+\left(\frac{a_{17}}{x}\right)^{a_{16}}}-a_{18}\right)*a_{19} \quad (1)$$

that approximates the log-transformed and sorted background-subtracted expression levels. (**b**) shows the plot of its first derivative (see [7]). Here, the independent variable *x* refers to rank. Single arrows point to the point of inflection and the double arrows point to the region of the curve where the CR is located.

located approximately at the junction of the second and third segments of the curves (see Fig. **1**).

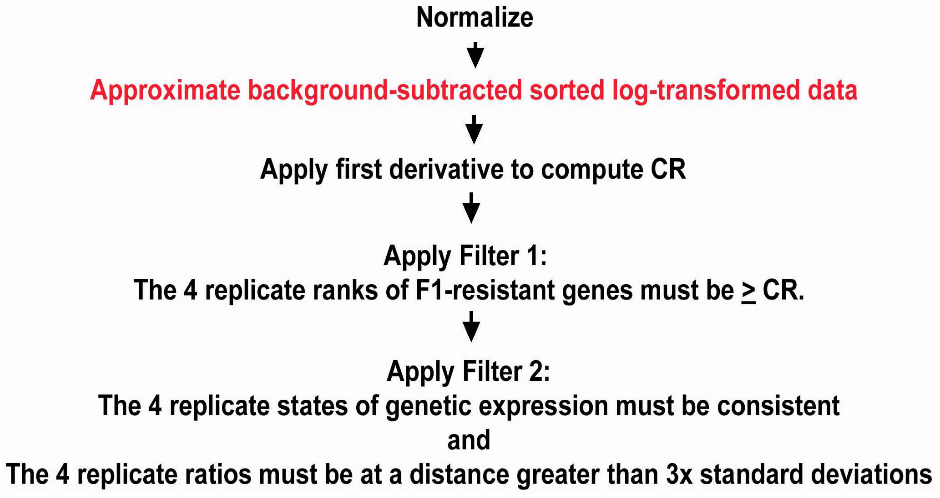
Outline of the MASH Procedure

MASH first computes the Cutoff Rank (CR) from the derivative of the equation. MASH includes two filters, F1 and F2 (Fig. **2**). A spot is sensitive to F1 if *both* its symmetrical ranks in SO1 and SO2 are less than the CR. To be resistant to F1, either Cy3 or Cy5 images of the spot must contain enough signals so that at least one of the symmetrical ranks is larger than the CR [7].

The second filter (F2) of MASH consists of two rules. The first Rule (F2a) requires that all four replicate ratios consistently show up- or down-regulation; i.e. all 4 replicate $\log_2$(ratios) > 0 or all four < 0. Recall that each dye-swapping experiment generates 4 replicate spot. The second rule of F2 (F2b) demands that all unfiltered $\log_2$ (ratios) must be outside the interval whose endpoints are ± 3 * the largest of the four replicate standard deviations of F1-resistant $\log_2$(ratios). Genes filtered by F1 or F2 have their mean $\log_2$(ratio) reset to 0.

Approximation by Splines

An alternative to determining the approximation rendered by the function displayed in (1), is to use standard approximation procedures on the same-to-same and spike-in datasets. Lagrangian interpolation, projection methods, finite-element approximation, etc. (see [12]) all come to mind as plausible alternatives to the generalized Padé-type ap-



Normalize

Approximate background-subtracted sorted log-transformed data

Apply first derivative to compute CR

Apply Filter 1:
The 4 replicate ranks of F1-resistant genes must be ≥ CR.

Apply Filter 2:
The 4 replicate states of genetic expression must be consistent
and
The 4 replicate ratios must be at a distance greater than 3x standard deviations

Fig. (2). ***Diagram summarizing the filters.*** MASH computes the CR from the derivative of the function in (1). MASHls and MASHsm compute the CR from the derivatives of smoothing splines and least square approximation by splines, respectively. All the other steps in the three procedures are the same.

proximation represented by the form (1). A little experimentation convinces one that interpolation is not a good idea. This is because the data is rather noisy and the MASH test for gene expression relies on the derivative of the approximation. Finite element methods using relatively smooth elements appear on the surface to have a better chance of predicting as well as does using MASH with (1). For the present investigation, we fixed on cubic splines for our finite-element space. We used both minimizing the sum of squares of the Euclidean distance as the approximation criteria (least-square, LS, approximation) and demanding maximal smoothness of the approximation (smoothing cubic applies or Sm approximation). The use of splines together with these more global optimizing criteria appears to be a good compromise between the desire for pointwise accuracy and the demand that the approximation have a well-behaved derivative.

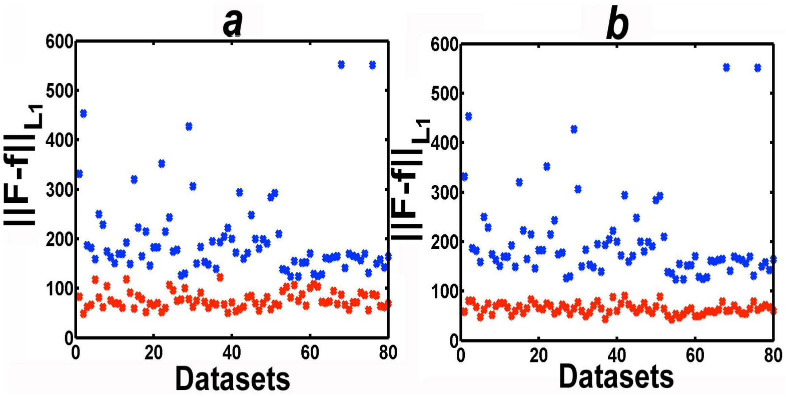


Fig. (3). ***Norms of the difference: spline curves are closer to the data***. Let *n* be the total number of spots on the microarray. Let f*(x)* represent a line connecting all the log-transformed and sorted background-subtracted expression levels in a dataset. Let F*(x)* be function that approximates the data. Then:

$$\|\mathrm{F}-\mathrm{f}\|_{\mathrm{L}_1} = \int_1^n |\mathrm{F}(x)-\mathrm{f}(x)|dx \cdot$$

The approximation F is given by the function (1) (blue, (**a**) and (**b**)), by splines using least square approximation (red, (**a**)), or smoothing splines (red, (**b**)). The averages of the L_1 norms of the differences between f and the F generated by the function (1), least square approximation splines, and smoothing splines are 199.1, 76.4, and 64.4, respectively. The *x*-axis refers to eighty 19K images.

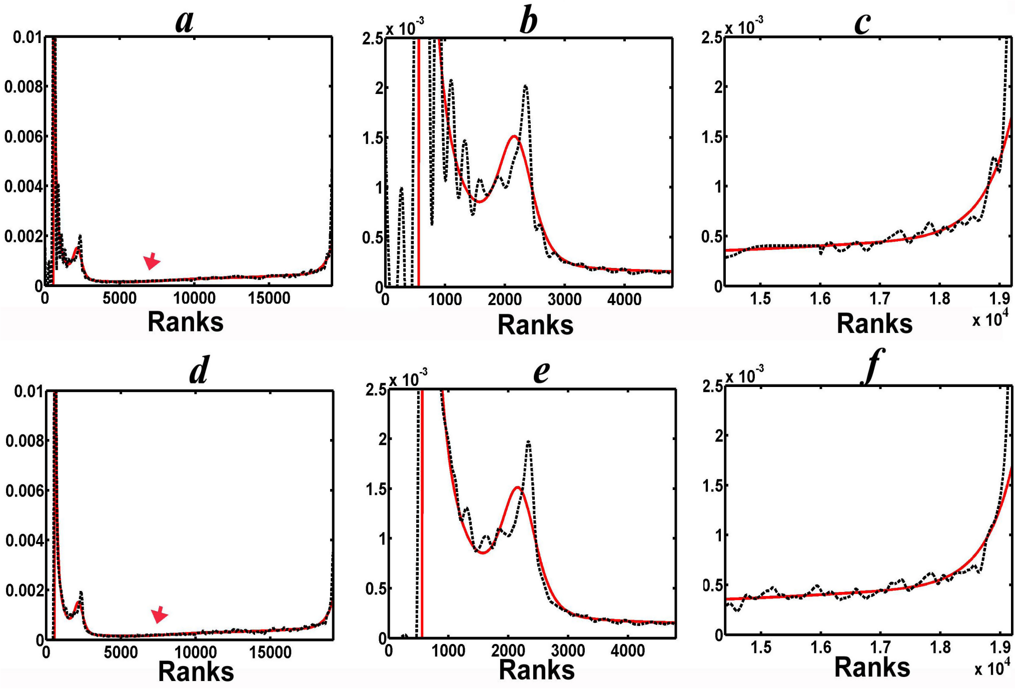


Fig. (4). Plots of the first derivatives: the curves of the first derivatives of the spline approximations are not as smooth as those of the approximation generated by (1). The same dataset is approximated globally by (1) (black line), splines using least square approximation (**a-c**, red), or smoothing splines (**d-f**, red). The graphs (**b-c**) and (**e-f**) show zoom-ins of (**a**) and (**d**), respectively. Notice that the plots show variations of the first derivatives of the splines in low and high ranks around the curves of the derivatives of equation (1). In addition, variation is more pronounced for splines generated by least square approximation as compared to smoothing splines (**b** and **e**). Nonetheless, little variation is seen in the area of the curve where the CR is computed (red arrow).

As compared to the use of (1), the curves generated by the Sm and Ls procedures are closer to the data as evidenced by smaller errors of difference. Fig. (**3**) shows the averages of the L_1-norms of the errors made using (1), splines generated by least square approximation, and smoothing splines to be 199.1, 76.4, and 64.4, respectively. However, the graphs depicting the derivatives of (1) are smoother than those of spline approximations in the sense of having fewer small-scale oscillations (see Fig. **4**). Specifically, the plots show variations of the curves of the first derivatives of the splines at low and high ranks. Fortunately, little variation is seen in the area of the curve where the CR is computed (see again Fig. **4**).

MASH2

Next, we carry out the same methods as MASH except for the step of approximating the data where we use smoothing cubic splines and least square approximation by cubic splines instead of (1); the algorithms are abbreviated MASHsm and MASHls, respectively (Fig. **2**). The goal is to compare the specificity and sensitivity of MASHsm and MASHls to the original implementation of MASH.

Table 1. Approximation by Splines Yields Highly Specific Discovery

	BRAIN vs. BRAIN (19K)	BRAIN vs. Brain (1.7K)
MASH	1/192000	1/17280
MASHls	1/192000	2/17280
MASHsm	0/192000	0/17280

The false discovery rate is computed from nine 1.7K and ten 19K *same-to-same* dye-swapping experiments. The results reveal that the false discovery rates of MASH, MASHsm, and MASHls are similar (Table **1**). Sensitivity is assayed by the percent Arabidopsis genes discovered from the best of four replicate spike-in- experiments, where 1 ng Arabidopsis RNA is added to one RNA sample but not the other. MASHsm sensitivity is equal to MASH (41%, 26/64); MASHls has a sensitivity of 39% (25/64). Receiver Operating Characteristics (ROC) is the standard approach to evaluate the sensitivity and specificity of diagnostic procedures (see [13]). The ROC curve plots the sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold is varied. The ROC can also be represented by plotting the fraction of true positives vs. the fraction of false positives (false discovery rate). MASH, MASHls, and MSHsm generate the empiric ROC areas of 0.703, 0.695, and 0.703, respectively. All 3 have an accuracy rate of 99.8% (Table **2**). MASHsm is also applied to analyze four same-to-same datasets from an independent laboratory [14]. Each dataset includes 710 “genes” spotted in duplicates to a total of 1420 spots. The false discovery rate is 0 per 2,840 genes.

Table 2. Sensitivity and ROC Analysis

	Sensitivity (%)	Empiric ROC Area	Accuracy (%)
MASH	41	0.703	99.8
MASHls	39	0.695	99.8
MASHsm	41	0.703	99.8

ROC estimates a curve, which describes the inherent tradeoff between sensitivity and specificity of a diagnostic test. The area under the ROC curve is important for evaluating diagnostic procedures because it is the average sensitivity over all possible specificities [23-25]. Eng, J. (n.d.). ROC analysis: web-based calculator for ROC curves. Retrieved [08/10/06], from http://www.rad.jhmi.edu/roc

DISCUSSION

Cubic splines generate curves that are closer to the datasets (Fig. **3**), but their first derivatives are not as smooth as the derivatives of the equation (1) (see again Fig. **4**). The findings also reveal that piece-wise approximation by cubic splines generates the same sensitivity and specificity as global approximation by the equation. On retrospect, this is not surprising because the CR is located in segments where the derivatives show little variation (Fig. **4**). The specificity of MASHsm, MASHls, and MASH are significantly better than other state-of-the-art methods (see [7]). Specifically, MIDAS (http://www.tm4.org/midas.html) specificity is 1,347/192000 and 170/1728 for the same 19K and 1.7K microarray datasets, respectively, whereas its sensitivity is the same as MASH (41%). MIDAS includes the Locfit (LOWESS) normalization [15, 16], standard deviation regularization [17], iterative linear regression normalization [15], iterative log mean centering normalization [18], ratio statistics normalization and confidence interval checking (confidence range at 99%) [19], low intensity filter, slice analysis [15, 16], and flip dye consistency checking [15, 17]. Despite all these safeguards, MASH, MASHsm, and MASHls have the same sensitivity as MIDAS [7]. Finally, to belabor the obvious, accurate, affordably obtained, knowledge of states of genetic expression is a powerful tool that has many applications in biology and medicine [8-10, 20-22].

REFERENCES

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science*, vol. 270, pp. 467-470, October 1995.

[2] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays", *Nat. Biotechnol.*, vol. 14, pp. 1675-1680, December 1996.

[3] J.L. DeRisi, V.R. Iyer, and P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, vol. 278, pp. 680-686, October 1997.

[4] R. Kothapalli, S.J. Yoder, S. Mane and T.P. Loughran T.P. Jr., "Microarray results: how accurate are they?", *BMC Bioinformatics*, vol. 3, p. 22, August 2002.

[5] E.E. Ntzani and J.P. Ioannidis, "Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment", *Lancet*, vol. 362, pp. 1439-1444, November 2003.

[6] P.K. Tan, T.J. Downey, E.L. Spitznagel Jr., P. Xu, D. Fu, D.S. Dimitrov, R.A. Lempicki, B.M. Raaka, and M.C. Cam, "Evaluation of gene expression measurements from commercial microarray platforms", *Nucleic Acids Res.*, vol. 31, pp. 5676-5684, October 2003.

[7] H.M. Fathallah-Shaykh, B. He, L.J. Zhao, and A. Badruddin, "Mathematical algorithm for discovering states of expression from direct genetic comparison by microarrays", *Nucleic Acids Res.*, vol. 32, pp. 3807-3814, July 2004.

[8] H.M. Fathallah-Shaykh, B. He, L.J. Zhao, H. Engelhard, L. Cerullo, T. Lichtor, R. Byrne, L. Munoz, K. Von Roenn, G. Rosseau, R. Glick, S. Chen, and F. Khan, "Genomic expression discovery

predicts pathways and opposing functions behind phenotypes", *J. Biol. Chem.*, vol. 278, pp. 23830-23833, April 2003.

[9] H.M. Fathallah-Shaykh, M. Rigen, L.J. Zhao, K. Bansal, B. He, H. Engelhard, L. Cerullo, K. Von Roenn, R. Byrne, L. Munoz, G. Rosseau, R. Glick, T. Lichtor, and E. DiSavino, "Mathematical modeling of noise and discovery of genetic expression classes in gliomas", *Oncogene*, vol. 21, pp. 7164-7174, October 2002.

[10] H.M. Fathallah-Shaykh, "Genomic Discovery reveals a molecular system for resistance to ER and oxidative stress in cultured glioma", *Arch. Neurol.*, vol. 62, pp. 233-236, February 2005.

[11] A.A. Alizadeh, M.B. Eisen, E. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson Jr., L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armittage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L. Staudt, "Distinct types of diffuse late B-cell lymphomas identified by gene expression profiling", *Nature*, vol. 403, pp. 503-511, February 2000.

[12] C. de Boor, *A practical guide to splines*. New York: Speinger-Verlag, 2001.

[13] J.A. Swets and R.M. Pickett, *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press, 1992.

[14] B.A. Rosenzweig, P.S. Pine, O.E. Domon, S.M. Morris, J.J. Chen, and F.D. Sistare, "Dye bias correction in dual-labeled cDNA microarray gene expression measurements", *Environ. Health Perspect.*, vol. 112, pp. 480-487, March 2004.

[15] J. Quackenbush, "Microarray data normalization and transformation", *Nat. Genet.*, vol. 32(Suppl.): pp. 496-501.

[16] I.V. Yang, E. Chen, J.P. Hasseman, W. Liang, B.C. Frank, S. Wang, V. Sharov, A.I. Saeed, J. White, J. Li, N.H. Lee, T.J. Yeatman, and J. Quackenbush, "Within the fold: assessing differential expression measures and reproducibility in microarray assays", *Genome Biol.*, vol. 3, p. research0062, October 2002.

[17] Y. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed, "Normalization of cDNA microarray data; a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acids Res.*, vol. 30, p. e15, February 2002.

[18] H. Causton, J. Quackenbush, and A. Brazma, *Microarray Gene Expression Data Analysis: A Beginner's Guide.* Blackwell Publishing, 2003.

[19] Y. Chen, E.R. Dougherty, and M.L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images", *J. Biomed. Optics*, vol. 2, pp. 364-374, October 1997.

[20] H.M. Fathallah-Shaykh, "Logical networks inferred from highly specific discovery of transcriptionally regulated genes predict protein states in cultured gliomas", *Biochem. Biophys. Res. Commun.*, vol. 336, pp. 1278-1284, November 2005.

[21] F. Vaince, J. Bona, and H.M. Fathallah-Shaykh, "Microarray data analysis: current practices and future directions", *Curr. Pharmacogenomics*, vol. 4, pp. 209-218, September 2006.

[22] H.M. Fathallah-Shaykh, "Microarrays: applications and pitfalls", *Arch. Neurol.*, vol. 62, pp. 1669-1672, November 2005.

[23] J.A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques", *Invest. Radiol.*, vol. 14, pp. 109-121, March-April 1979.

[24] C.E. Metz, "Methodology in radiologic imaging", *Invest. Radiol.*, vol. 21, pp. 720-733, September 1986.

[25] N.A. Obuchowski, "Receiver operating characteristic curves and their use in radiology", *Radiology*, vol. 229, pp. 3-8, October 2003.

Received: July 10, 2008 Revised: August 12, 2008 Accepted: August 17, 2008

© Bona and Fathallah-Shaykh; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.