A Multi-Strip Algorithm and Its Application to Gene Characterization Using DNA-Array Data

Gilad Lerman[1], Joseph McQuown[1], and Bud Mishra[1,2*]

[1] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY, USA 10012.

[2] Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY, USA 11724.

October 13, 2003

ABSTRACT **A fast adaptive multiscale algorithm has been devised to characterize a random set of points spanning a high dimensional Euclidean space, but concentrated around special lower dimensional subsets. It has been adapted to analyze gene expression data from microarray experiments. We present here the simplest version of this "multi-strip" algorithm applied to a set of points in $\mathbb{R}^D$ concentrated around a line. The algorithm characterizes this set by finding a strip around the *principal axis* of the set, so that it isolates *deviating* points from the *main bulk* of points enveloped by the strip. The algorithm generalizes to computing a strip around a best L^2 d-plane, where $1 \leq d < D$, or even fitting a strip around a d-dimensional Lipschitz graph. We establish various estimates for its performance. When applied to gene-expression data, the algorithm can be thought of as estimating the local statistics (means, standard deviations, tail distributions, etc.) as a function of the entire expression range. Genes with abnormal differential expression values can be identified and given biological interpretations based on the local deviations in their statistics. By avoiding rigid local segmentations (as in segmental nearest neighbor normalization) or non-adaptive global estimates, the algorithm achieves a superior performance.**

1 Introduction

Microarray and gene-chip technologies provide an approach for characterizing transcriptional properties of thousands of genes and studying their interactions simultaneously under many different experimental conditions. However, in many applications the key problem has been statistical noise in the transcriptional data, varying from experiment to experiment and attributable to non-specific hybridization, cross-hybridization, competition, diffusion of the target on the surface, base-specific structural variations of the probe, etc. A better understanding of this noise can come from the kinetic analysis of the base-pairing, denaturing, and diffusion processes. In the absence of detailed knowledge how to deconvolve the measurement data, it is hard to distinguish properly between specific clusters of genes, based on expression intensities data. The purpose of identification (combined with normalization) methods is to compare expression intensities from multiple experiments, and distinguish between a stable subset of genes whose behaviors could be expected to be already well-modeled (so-called housekeeping genes, rank-invariant genes, or genes with constant expression), and a subset of genes deviating from the stable model (so-called non-housekeeping genes, regulated genes or differentially expressed genes). See [17].

The identification process creates a statistical model of the "main bulk" of the genes (i.e., the stable subset) either through a global statistical analysis of transcriptional expression intensities of all the data or through a local statistical analysis of similar statistics as a function of the expression range. The genes deviating from the statistics computed via initial identification are then subjected to further analysis to determine their biological characteristics in response to the experimental condition (see e.g. [1]). In the simplest conceivable setting, one may consider thousands of genes monitored under two different experimental conditions (c_1 and c_2), and the data in a 2-D Euclidean space thought to consist of average over expression intensities for a gene (g) versus a measure of its relative expression intensities. Such a measure of the relative expression intensities may take the

*To whom correspondence should be addressed. E-mail: mishra@nyu.edu

form of *expression ratio* (ER), *logarithm of expression ratio* (LER), *differential expression ratio* (DE), etc. For instance, if the intensity values are $e_{c_1,g}$ and $e_{c_2,g}$, then they may be described by a point

$$\left\langle \frac{\ln e_{c_1,g} + \ln e_{c_2,g}}{2}, \ln \frac{e_{c_2,g}}{e_{c_1,g}} \right\rangle \in \mathbb{R}^2.$$

Implicit in our approach is the assumption that for a large stable subset of genes any one of these measures of relative expression intensities varies randomly about a mean value from experiment to experiment in a way which may depend on the different mean values. For instance, we may model the log ratio (LER) to have a normal distribution with a variance depending on the local average intensities:

$$\ln \frac{e_{c_2,g}}{e_{c_1,g}} \sim \mathcal{N}\Big(0, \sigma(e_g)^2\Big), \tag{1}$$

where e_g is estimated by $(\ln e_{c_1,g} + \ln e_{c_2,g})/2$. In this setting, the area defined by $|y| \le 3\sigma(x)$ may describe a strip containing 99.73% of the housekeeping genes.

In general, we thus aim to separate, by a compact region, the genes belonging to a stable set (e.g., housekeeping genes) from the other genes that respond unambiguously to the change in experimental conditions. The boundary of this region will be referred to as the "strip," and devising an algorithm to compute it efficiently and accurately becomes an interesting mathematical problem.

Formally, we consider the following mathematical problem: Given a set of points in $\mathbb{R}^D$ concentrated around a line, find a strip around the *principal axis* of the set, so that it isolates *deviating* points from the main bulk of points. For this problem, we propose a fast multiscale algorithm and establish some estimates for the quality of the computed strip.

We can easily extend the above mathematical problem to finding a strip around a best L^2 d-plane, where $1 \le d < D$. A more general version of our algorithm (following mathematical ideas of [7, 3, 8]) can be shown to even fit a d-dimensional Lipschitz graph (or chord-arc curve when $d = 1$) and a strip around it. The later generalization can be used, when $d = 1$, in order to both normalize the genes' expression intensities and identify differentially expressed genes; it will be discussed in a future publication. The algorithm described here is used only for identification, assuming the data is normalized around the principal axis (see e.g. [17]).

Our algorithm constructs three different strips in a multiscale fashion. For the first strip A, we show how at different scales the algorithm controls both the number of points outside it and also the rate of change of that strip in the direction of the principal axis (a measure of the strip's complexity). The second strip R maintains at different scales and locations the same ratio between the number of points outside the strip and the total number of points. The third strip S estimates adaptively the *standard deviation* of the points (more precisely it estimates adaptively the second moments of the distances of the points from the principal axis). This multiscale approach is capable of balancing between overfitting at small scales and underfitting at large scales.

2 Algorithm and Methods

2.1 Description of Algorithm

Input, preprocessing and output

The main input to the algorithm is a set $E = \{x_i\}_{i=1}^N$ of N points in $\mathbb{R}^D$ where $N \ge D$. Additional input includes the following predefined parameters: ℓ_0 (integer), n_0 (integer), α_i, $i = 0, 1, 2$ (reals), δ_0 (real), c_0 (real) and C_1 (real, $C_1 > 1$). We discuss our choice of all parameters in the full paper [9]. The parameters α_i, $i = 0, 1, 2$, are set by the user according to the expected ratio of differentially expressed genes over the total number of genes.

The algorithm initially stores the set E in an $N \times D$ data matrix A, whose rows correspond to the D-dimensional vectors in E. It then performs the following operations (we maintain the notation E and A for the transformed set and matrix): First, it shifts each row of A by the center of mass of the set. Second, it computes "the principal axis", $L \equiv L_E$, of the data set (recall that the principal axis of E is the line spanned by the top right singular vector of the shifted matrix A). Third, it rotates the set so that its principal axis coincides with the x axis. The algorithm then fixes an interval $Q_0 = [a_0, b_0)$ of nearly minimal length containing the projection of E onto L.

The output of our algorithm includes three different strip functions: A, R and S. These are real-valued functions defined on Q_0. The algorithm evaluates them for all points in $P_L E$, where P_L denotes the projection operator from $\mathbb{R}^D$ onto L. The envelopes of the strips are obtained by rotating the graphs of the corresponding functions around the x-axis (the line L).

Basic Notation and Definitions

We use the following notation and definitions in describing the main part of the algorithm.

We denote by P_L the projection operator from $\mathbb{R}^D$ onto L (the principal axis of E).

If K is a subset of $\mathbb{R}^D$, we denote by $|K| \equiv |K \cap E|$ the number of points of E in K. If Q is an interval, we denote

by $\ell(Q)$ its length. We denote by χ_Q the indicator function of Q:

$$\chi_Q(x) = \begin{cases} 1, & \text{if } x \in Q; \\ 0, & \text{otherwise.} \end{cases}$$

The algorithm operates on generalized dyadic grids, which depend on a fixed rule $\mathcal{R}$ for partitioning an interval $[a, b)$ into two subintervals: $[a, m)$ and $[m, b)$, where $m = \mathcal{R}([a, b))$. We use either the median rule: $\mathcal{R}(Q) = P_L(\text{median of } \widetilde{Q})$ (see below for the definition of $\widetilde{Q}$) or the symmetric rule (equivalently midpoint rule): $\mathcal{R}([a, b)) = \frac{a+b}{2}$. The generalized grids $\mathcal{D}_j(Q_0) \equiv \mathcal{D}_j^{\mathcal{R}}(Q_0)$ are formed as follows. If $j = 0$, then $\mathcal{D}_0(Q_0) = \{Q_0\}$. If $j > 0$, $Q = [a, b)$ is an interval in $\mathcal{D}_j(Q_0)$ and $m = \mathcal{R}([a, b))$, then set

$$Q_L(Q) := [a, m) \quad \text{and} \quad Q_R(Q) := [m, b).$$

Define

$$\mathcal{D}_{j+1}(Q_0) = \bigcup_{Q \in \mathcal{D}_j(Q_0)} \Big(Q_L(Q) \bigcup Q_R(Q) \Big),$$

and

$$\mathcal{D}(Q_0) = \bigcup_{j=0}^{\ell_0} \mathcal{D}_j(Q_0).$$

If Q is an interval in $\mathcal{D}(Q_0)$, we define its extensions $\widehat{Q}$ and $\widetilde{Q}$ to $\mathbb{R}^D$ by the formula:

$$\widehat{Q} = \{x \in \mathbb{R}^D : P_L x \in Q\},$$

and

$$\widetilde{Q} = \begin{cases} \{x \in \widehat{Q} : \ \text{dist}(x, L) \le c_0 \cdot \ell(Q)\}, & \text{if } \ Q \subsetneq Q_0; \\ \widehat{Q}_0, & \text{if } \ Q = Q_0. \end{cases}$$

Define the "top" part of $\widetilde{Q}$ as follows:

$$T(\widetilde{Q}) = \widetilde{Q} \setminus (\widetilde{Q}_L \cup \widetilde{Q}_R).$$

If R is any set contained in $\widehat{Q}$, then define

$$\sigma_R = \left(\frac{1}{|R|} \sum_{x_i \in R} \text{dist}^2(x_i, L) \right)^{\frac{1}{2}} \quad \text{and} \quad \beta_R = \frac{\sigma_R}{\ell(Q)}.$$

If $Q \in \mathcal{D}(Q_0) \setminus \{Q_0\}$, then denote by P_Q the dyadic parent of Q according to the grid $\mathcal{D}(Q_0)$ and also define $P_{Q_0} := Q_0$.

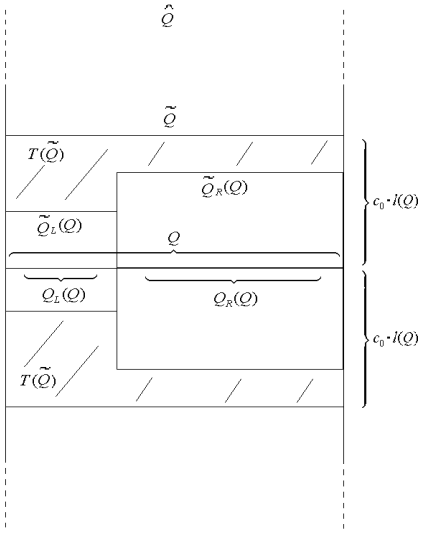

Figure 1: Illustration of different parts assigned to the interval Q.

The stopping time construction

The description of the algorithm can now be completed by assigning its stopping time criteria. For each $Q \in \mathcal{D}(Q_0)$ we define

$$f_Q = \frac{|T(\widetilde{Q})|}{|\widehat{Q}|} \quad \text{and} \quad F_Q = \sum_{\substack{Q' \in \mathcal{D}(Q_0) \\ Q' \supseteq Q}} f_{Q'}.$$

The algorithm computes F_Q with a top-bottom procedure: First, it initializes $F_Q \equiv 0$ for all $Q \in \mathcal{D}(Q_0)$. Then, it applies the reduction formula (from coarse levels to fine levels):

$$F_Q = F_{P_Q} + f_Q.$$

While proceeding from top to bottom levels, the algorithm stops at an interval $Q' \in \mathcal{D}(Q_0)$ (together with all of its descendants in $\mathcal{D}(Q_0)$) if and only if one of the following conditions is satisfied:

1. $F_{Q'} > \alpha_0$. (2)
2. $|\widetilde{Q}'| < n_0$.
3. $\beta_{\widetilde{Q}'} > \delta_0$ (optional).
4. $|\widehat{Q}' \setminus \widetilde{Q}'| > \alpha_1 \cdot |\widetilde{Q}'|$ (optional) . (3)

The first stopping time condition is the crucial one for controlling the number of points outside the different strips (mainly A). The second one is necessary in order to have valid estimates in each interval. The third one allows us to

control the "complexity" of the strip A (see Proposition 2.3). The fourth one is used to obtain the last equations of both Propositions 2.1 and 2.2. The last two stopping conditions may be ignored by setting $\delta_0 = c_0$ and $\alpha_1 = 1$, respectively. A detailed discussion of the stopping time criteria can be found in the full paper [9].

We denote

$$\mathcal{Q} = \{Q : Q \text{ is a stopping time interval in } \mathcal{D}(Q_0)\}.$$

We partition $\mathcal{Q}$ into two different disjoint sets of "good" and "bad" intervals respectively:

$$\begin{aligned}\mathcal{G} =& \{Q : Q \in \mathcal{Q},\ |\widetilde{Q}| \geq n_0 \ \text{ and } \ \beta_{\widetilde{Q}} \leq \delta_0\},\\ \mathcal{B} =& \mathcal{Q} \setminus \mathcal{G}.\end{aligned}$$

The strips A, R and S

We describe piecewise constant versions of the different strip functions. They all use the stopping time criteria described earlier, but differ in the manner they select the parameters to determine the stopping time intervals.

In order to assign A, the algorithm computes for each interval $Q \in \mathcal{Q}$ the following number:

$$\gamma_{\widetilde{Q}} = \begin{cases} \min\{C_1 \cdot \sigma_{\widetilde{Q}},\ c_0 \cdot \ell(Q)\}, & \text{if } \ Q \in \mathcal{G};\\ \min\{C_1 \cdot \sigma_{\widetilde{P}_Q \cap \widehat{Q}},\ c_0 \cdot \ell(Q)\}, & \text{otherwise .}\end{cases}$$

It then sets the values of A as follows:

$$A(x) = \sum_{Q \in \mathcal{Q}} \gamma_{\widetilde{Q}} \cdot \chi_Q(x), \quad \text{for all } \ x \in P_L E\,. \tag{4}$$

The algorithm computes the strip R, so that at each stopping time interval Q it leaves a fraction of size α_2 of the points outside the strip. More precisely, if $Q \in \mathcal{Q}$, then

$$|x : x \in \widehat{Q} \text{ and } \operatorname{dist}(x, L) \geq R(P_L x)| = \lfloor \alpha_2 \cdot |\widehat{Q}| \rfloor \approx \alpha_2 \cdot |\widehat{Q}|,$$

where the "floor function" $\lfloor x \rfloor$ denotes the largest integer smaller or equal to x.

The algorithm computes the strip S as follows:

$$S(x) = \sum_{Q \in \mathcal{Q}} \sigma_{\widehat{Q}} \cdot \chi_Q(x).$$

Note that this strip estimates locally (and adaptively) the square root of the second moments of the distances of the points of E to the line L.

By multiplying S by a certain constant, we obtain an approximate version of R which is less sensitive to noise. More precisely, set $C_\sigma \equiv C_\sigma(\alpha_1) := \sqrt{2} * \operatorname{erfinv}(\alpha_2)$, where erfinv is the inverse Erf function (error function for normal distribution). If the assumption stated in equation (1) is correct, then the strip $C_\sigma \cdot S$ leaves out a fraction of size α_2.

The strips A, R and S constructed above are all piecewise constant functions. However, it is possible to derive smooth strip functions as follows: First, generate many instances of the corresponding piecewise constant function according to different grids. Then average these piecewise constant functions over all the instances. The complete details appear in the full paper [9].

It is possible to apply the stopping time construction twice or to reiterate the whole algorithm. The resulting strips are supposed to be less sensitive to highly deviating points than the original strips.

Lastly, we remark that for gene expression data, we prefer using the smoothed version of the strip $C' \cdot S$ (usually $C' = C_\sigma(\alpha_2)$) with the specific constants described in the full paper [9] and without reiteration.

Analysis of the strips

By appropriate choice for the stopping time criteria, we control at different scales the number of points outside the strip A as well as the rate of change of A in the direction of the line L. We also remark on the relation between the strip A and the strips R and $C_\sigma \cdot S$. We only state the main results. Additional results and proofs are available in the full paper [9].

Denote the set of ancestors of intervals in $\mathcal{Q}$ by $\mathcal{P}$. That is,

$$\mathcal{P} = \{P : P \in \mathcal{D}(Q_0) \text{ and } \exists\, Q \in \mathcal{Q} \text{ such that } Q \subseteq P\}.$$

For any given interval $Q \in \mathcal{P} \setminus \mathcal{Q}$, define the number of points in $\widetilde{Q}$ outside the strip A as

$$m_{\widetilde{Q}}(A) := |\{x : x \in \widetilde{Q} \ \text{ and } \ \operatorname{dist}(x, L) \geq A(P_L x)\}|.$$

Similarly, define

$$m_{\widehat{Q}}(A) := |\{x : x \in \widehat{Q} \ \text{ and } \ \operatorname{dist}(x, L) \geq A(P_L x)\}|$$

We estimate these numbers as follows:

Proposition 2.1 *For any $Q \in \mathcal{P} \setminus \mathcal{Q}$:*

$$\frac{m_{\widetilde{Q}}(A)}{|\widetilde{Q}|} \leq \alpha_0 + \frac{1}{C_1^2} \ \ \textit{and} \ \ \frac{m_{\widehat{Q}}(A)}{|\widehat{Q}|} \leq \alpha_1 + \frac{1}{C_1^2}.$$

Extensive numerical experiments lead us to conclude that the numbers $m_{\widetilde{Q}}(A)$ do not depend on the constant C_1 (es-

pecially for large scale intervals, e.g. Q_0). Indeed, define

$$\mu_{\widetilde{Q}}(A) := \sum_{\substack{Q'\in\mathcal{Q} \ \&\ Q'\subseteq Q \\ C_1\cdot\sigma_{\widetilde{Q}'}<c_0\cdot\ell(Q')}} \left|\left\{x : x\in E,\ P_L x\in Q' \quad \text{and} \right. \right.$$
$$\left.\left. c_0\cdot\ell(Q')\geq \operatorname{dist}(x,L)\geq C_1\cdot\sigma_{\widetilde{Q}'}\right\}\right|$$

and note the following property:

Proposition 2.2 *If there exists a constant* $C'\gtrapprox 1$ *so that*

$$\mu_{\widetilde{Q}}(A)\leq\left(1-\frac{1}{C'}\right)\cdot m_{\widetilde{Q}}(A),$$

then

$$m_{\widetilde{Q}}(A)\leq C'\cdot\alpha_0\cdot|E| \quad \textit{and} \quad m_{\widehat{Q}}(A)\leq C'\cdot\alpha_1\cdot|E|. \quad (5)$$

Our algorithm controls at different scales the rate of change of the strip A in the direction of the line L, which we view as a complexity of that strip. We formulate this property more precisely as follows:

Proposition 2.3 *Assume that for any* $Q\in\mathcal{Q}$: $\beta_{\widetilde{P}_Q}\approx\beta_{\widetilde{P}_Q\cap\widehat{Q}}$ *and that the grids are symmetric (midpoint rule). If* Γ *is any one of the curves obtained by intersecting the strip obtained by the function* A *together with a* D*-plane containing the line* L*, then*

$$\ell(\Gamma\cup\widehat{Q})\leq(1+C_1\cdot\delta_0)\cdot\ell(Q) \quad \textit{for any} \quad Q\in\mathcal{P}\setminus\mathcal{Q}. \quad (6)$$

The above estimates hold for the strip A. However, note that the strips $C_1\cdot S$ and A are quite similar (recall that the values of the functions A and S depend on the input constant C_1). Indeed, the strip A is obtained by first thresholding the points outside $\cup_{Q\in\mathcal{Q}}\widetilde{Q}$, and then estimating $C_1\cdot\sigma_{\widetilde{Q}}$ for each $Q\in\mathcal{Q}$. Whereas, the strip S estimates $C_1\cdot\sigma_{\widehat{Q}}$ for each $Q\in\mathcal{Q}$. The similarity of A and S thus follows from the stopping time condition stated in equation (3), which controls locally the differences between $\widetilde{Q}$ and $\widehat{Q}$ (there is an additional assumption which is necessary for that similarity; see [9]). The similarity of R and $C_\sigma\cdot S$ has been discussed in the previous section, together with the assumptions under which it holds.

3 Results and Discussion

We examined the performance of the multi-strip algorithm with three different data sets: (i) A synthetic *in silico* gene expression data set, generated under a mixture model combining a stable set of genes with a small number of deviating gene expressions. (ii) An experimental *in vitro* gene expression data set derived from the megaplasmid pSOL1 deficient *C. acetobutylicum* strain M5 relative to WT [17]. (iii) Finally, a gene expression data set examining the sex-biased genes of *D. melanogaster* [13].

Synthetic Gene Expression Data

For the purpose of testing the algorithm, we rely on two-dimensional synthetic data sample from several types of Gaussian mixture distributions. We use the synthetic data for demonstration and algorithm development purposes only and in no way suggest that one could convincingly argue the optimality of an algorithm based on these limited experiments alone. The choice of two dimensions can be extended to multiple sample gene-chip experiments in higher dimensions.

The data is simulated as follows. First, we create an i.i.d. sample of 5000 points from a mixture of bivariate normal distributions concentrated around the x-axis. We denote this mixture distribution by F_0. Next, indices of 50 up regulated and 50 down regulated genes are randomly chosen. Last, we convolve the distributions of both up and down regulated genes with a similar mixture of Gaussians with means in the upper half plane and lower half plane, respectively. The resulting distributions are denoted by F_{up} and F_{down}, respectively. We specify the complete parameters in [9].

We denote the class of "stable" genes sampled from the distribution F_0 by St, the class of up-regulated genes, sampled from the distribution F_{up} by Up, the class of down-regulated genes, sampled from the distribution F_{down} by Do and finally the set of differentially expressed genes ($Do\cup Up$) by Df. After running the multiscale algorithm, we identify the gene expressions that lie outside the strip $C_\sigma\cdot S$ as differentially expressed and refer to them as positives (or P). Similarly, we refer to the genes inside the strip as negatives (or N). The set of true (T) and false (F) positives and negatives are set as follows: $TP := Df\cap P$, $FP := St\cap P$, $TN := St\cap N$ and $FN := Df\cap N$. We define the sensitivity Sns, the specificity Spc and the error Er as follows:

$$Sns=\frac{|TP|}{|Df|},\quad Spc=\frac{|TN|}{|St|} \quad \text{and} \quad Er=\frac{1}{2}\cdot\left(\frac{|FP|}{|St|}+\frac{|FN|}{|Df|}\right).$$

We use an ROC curve, shown in Figure 3, to demonstrate how well the strip $C_\sigma\cdot S$ separates the differentially expressed genes for different choices of the parameter α_2 (for the actual quality of separation of these genes in the data set, see [9]). The area below the piecewise linear ROC curve is 0.78. The error Er is minimized when $\alpha_2=0.11$. Figure 3 shows the synthetic data set together with the strip $C_\sigma\cdot S$, where $\alpha_2=0.11$.

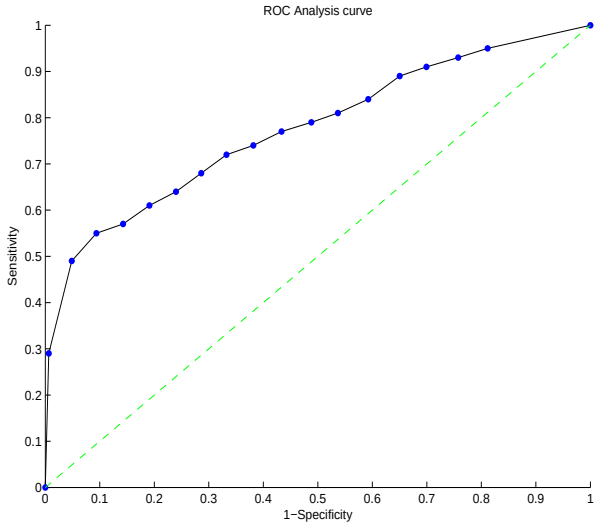

Figure 2: *ROC curve for separating the differentially expressed genes in the synthetic data by the strip $C_\sigma \cdot S$. The blue dots correspond to different values of α_2.*

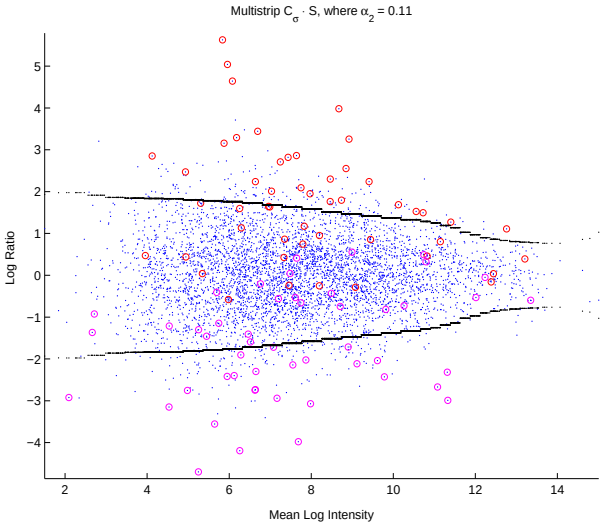

Figure 3: *Synthetic data set with a multistrip. "Stable" genes are denoted by blue dots, up regulated genes are denoted by red circles and down regulated genes are denoted by magenta circles. The multistrip (in black) is $C_\sigma \cdot S$, where $\alpha_2 = 0.11$.*

C. acetobutylicum Gene Expression Data and comparison with SNNLerm Algorithm

Yang et al. [17] have developed what they call a segmental nearest neighbor method of LERs (SNNLerm) for gene expression normalization and identification. They divide the log mean intensity range into a fixed number of equidistant intervals and compute the mean and standard deviation of LERs for each interval using only nearest neighbor genes. The value of the strip function ("mask") in each interval is determined by the standard deviation. They also assign confidences to the points in each intervals. They concluded that their identification method is superior to other methods (conditioned on using the SNNLerm normalization).

We compare the SNNLerm identification algorithm with our algorithm using the glass slide arrays of tissue samples taken from the megaplasmid pSOL1 deficient *C. acetobutylicum* strain M5 relative to WT [17]. Strain M5 is isogenic to WT but lacking the pSOL1 plasmid. Only 169 out of the 178 pSOL1 genes are included in the glass slides. The pSOL1 genes are expected to be expressed with a broad range of levels in WT, but unexpressed in M5. Therefore the expression ratios of these genes should be characterized as non-differentially expressed and even down-regulated. Of course, this classification depends on whether such a deviating gene is actually expressed in WT or not. We used six glass arrays (see [9] for details), which were chosen by Yang et al. [17] to produce Table 1 (see [17, page 1126]). We were not able to reproduce the same table and thus analyzed each slide separately. After pre-filtering and normalizing each slide by the initial part of the SNNLerm algorithm we ran both identification algorithms. We used the strip $C_\sigma \cdot S$ for the multiscale algorithm (with the corresponding parameters specified in [9]). In order to be able to compare between the two algorithms, we have determined the value of α_2 in order to obtain the same average fraction (averaged over the six slides) of pSOL1 genes identified by both algorithm as differentially expressed over the total number of those genes.

We use the error of identification specified in [17, equation (9)]. More specifically, we denote the set of pSOL1 genes in each experiment by Df and the complementary set by St. We identify the gene expressions that lie outside the assigned strip (or with confidences greater than 95.5% when using the SNNLerm algorithm) as differentially expressed and refer to them as positives (or P). We use the notation P, N, TP, FP, TN and FN as in the previous section. Also denote by DU the points of the set Df, which the given algorithm identified as up regulated (that is, above the strip). We define the the error $\widetilde{Er}$ as follows:

$$\widetilde{Er} = \frac{1}{2} \cdot \left(\frac{|FP|}{|St|} + \frac{|FN| + |DU|}{|Df|} \right).$$

We summarize the results in Table 1. We remark that

Numerical Results	Slide 422	Slide 424	Slide 783	Slide 784	Slide 786	Slide 805
Total Count						
$\|Df\|$	118	127	51	144	119	136
$\|St\|$	655	645	551	742	653	706
SNNLerm						
$\|FP\|$	58	47	38	34	37	41
$\|FN\|$	106	115	47	107	95	111
$\|DU\|$	1	1	1	0	0	1
$\|TP\|$	12	12	4	37	24	25
$\widetilde{E}r$	0.498	0.493	0.505	0.394	0.427	0.441
Multiscale						
$\|FP\|$	61	43	38	36	32	41
$\|FN\|$	103	112	47	109	96	108
$\|DU\|$	1	1	1	0	0	1
$\|TP\|$	15	15	4	35	23	28
$\widetilde{E}r$	0.487	0.478	0.505	0.403	0.428	0.430

Table 1: *Comparison of SNNLerm and the Multistrip method for identification of C. acetobutylicum pSOL1 genes in six slides of M5-WT experiments.*

Df is less than 169 due to pre-filtering of pSOL1 genes with high background noise. The multiscale algorithm performs better than the SNNLerm algorithm for slides numbers: 422, 424, 805, while SNNLerm performs better for slide number: 784. The two algorithms are comparable for slides numbers: 783 and 786. Unlike the SNNLerm algorithm, the multiscale algorithm is adaptive. In particular, parameter values are independent of the types of microarray experiments (glass, vinyl, plastic).

D. melanogaster Gene Expression Data and Sex-Biased Genes

Lastly, we apply the multiscale algorithm to detect sex-biased genes of Drosophila melanogaster using one of the many experiments of Parisi et al. [13]. In this experiment tissue is taken from adult male versus adult female flies without having removed their reproductive organs (slide is available from the Gene Expression Omnibus under accession GSM2456).

Global gene expression in Drosophila melanogaster has been reported to have an elevated transcription of X-chromosome genes in males due to a dosage-compensation mechanism. However, it has been suggested that, unlike in the somatic cells, there is no dosage compensation in the germ line and this hypothesis can be tested by comparing expression data in males against expression data in females (of both somatic, germ line and mixed cells).

In order to distinguish between male-biased and female-biased genes and also due to the non-symmetric nature of the data, we implement a slight variation of the multiscale algorithm. That is, we run the algorithm twice for the two sets of genes in the two half planes bisected by the diagonal of the data. We use this line instead of the principal axis and thus avoid the initial transformation of the algorithm (specific details are in [9]).

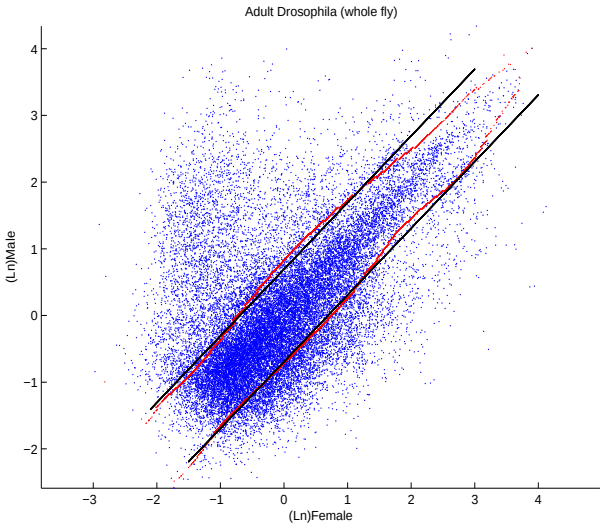

Figure 4: *Logarithmic intensities of Drosophila melanogaster whole adult fly, male vs. female. The two fold strip is in black and the multistrip is in red.*

Parisi et al. used the threshold $\ln 2$ to determine the differentially expressed genes (two fold approach). In order to compare their constant strip with ours, we set for each subset (in each half plane) α_2, so that the number of genes outside both strip are the same. For the sake of simplicity, we used the strip R. The resulted strip together with the two fold strip are shown in Figure 3.

Some Concluding Remarks

The multiscale algorithm is a robust, efficient and mathematically innovative way to adaptively analyze data without prescribing assumptions to the data when little prior information is available. Thus, this and other such priorless approaches depart from conventional statistical methods as well as Bayesian methods in that we have no longer access to a model, or fitting to a model through optimization of a likelihood, expectation, or related functions (e.g., MCMC, EM or MLE methods). Even empirical Bayes methods [5] cannot

reconcile the problems of non-specific hybridization, cross-hybridization, competition, target diffusion, probe-specific complications, etc., that happen at the local level. Any algorithm that pre-determines the localities of the expression level also undermines analysis. In any case, through local spatial adaptability, the focus of this multiscale procedure becomes a low-complexity representation of the structure in the data without ascribing parametric distributions, see Jones [7], David and Semmes [3] and Lerman [8]. Furthermore, the complexity of the representation is provably bounded by a "competitive factor" with respect to the best possible representation. Other algorithmic examples of similar approach include CART (Breiman et al. [2]), MARS (Friedman [6]), MART, variable bandwidth kernel methods (Muller and Stadtmuller [11]), etc.

Our application of this approach to gene expression data is decidedly a natural one; nonetheless, an important one, as it resolves many important difficulties in comparing poorly understood variations in gene-expression measurements from experiment to experiment. We may compare our algorithm to other techniques for defining and elucidating genes with putative differential expression as well as methods for normalization and experimental control. See Li [10], Dudoit and Yang [4], Efron et al. [5], Garrett and Parmigiani [14]. Yang et al. [17] and Newton and Kendziorski [12]. We focused on three important datasets (one synthesized and two experimental) and concluded that multi-scale approach in its most skeletal form captures the local variations extremely well, even when it has no direct way of modeling the nature of the variation. There are several further modifications that remain to be explored: extension to higher dimension, asymmetry in the data sets, low dimensional variations in the "principal axis", shrinkage approaches to handle sparsely populated regions, etc.

We thank Peter Jones, Fang Cheng and Yi (Joey) Zhou for many helpful discussions; E. Terry Papoutsakis and Carles Paredes for their help in interpreting the data appearing in their original paper on pSOL1 genes; and finally, Mark Green and IPAM (UCLA) for inviting us to take part in their bioinformatics and computational biology meetings, where discussions of similar topics stimulated our research. The work reported in this paper was supported by grants from NSF's Qubic program, NSF's ITR program, Defense Advanced Research Projects Agency (DARPA), Howard Hughes Medical Institute (HHMI) biomedical support research grant, the US department of Energy (DOE), the US air force (AFRL), National Institutes of Health (NIH) and New York State Office of Science, Technology & Academic Research (NYSTAR).

References

[1] B. W. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. (2003) "A comparison of normalization methods for high density oligonucleotides array data based on variance and bias." *Bioinformatics*, **19**(2):185–93.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. (1983) *CART: Classification and Regression Trees.* Wadsworth, NY.

[3] G. David,S. Semmes.(1993). *Analysis of and on uniformly rectifiable sets*, volume 38 of American Mathematical Society, Providence, RI.

[4] S. Dudoit, Y.H. Yang, T.P. Speed, and M.J. Callow. (2002) "Statistical Methods for Identifying Differentially Expressed Genes in Replicated CDNA Microarray Experiments." *Statistica Sinica*, **12**(1):111–139.

[5] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. (2001) "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association*, **96**:1151–1160.

[6] J. Friedman. (1992) "Multivariate adaptive regression splines." *Annals of Statistics*, **19**: 1–67.

[7] P. W. Jones. (1990) "Rectifiable sets and the traveling salesman problem." *Invent. Math.*, **102**(1): 1–15.

[8] G. Lerman. (2003) "Quantifying curvelike structures of measures by using L_2 Jones quantities." *Comm. Pure App. Math.*, **56**(9): 1294–1365.

[9] G. Lerman, J. McQuown and B. Mishra. In preparation, preliminary preprint available at `http://www.cs.nyu.edu/cs/faculty/mishra/`.

[10] C. Li and W. Wong. (2001) "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection." *Proceedings of the National Academy of Sciences.*, **98**(1): 31–36.

[11] H. Muller and U. Stadtmuller. (1987) "Variable Bandwidth Kernel Estimators of Regression Curves." *Annals of Statistics*, **15**(1): 182–201.

[12] M. Newton, C. Kendziorski, C.S. Richmond, and F.R. Blattner. (2001) "On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data." *Journal of Computational Biology*, **8**: 37–52.

[13] M. Parisi, R. Nuttall, D. Naiman, G. Bouffard, J. Malley, J. Andrews, S. Eastman, and B. Oliver. (2003) "Paucity of Genes on the *Drosophila* X Chromosome Showing Male-Biased Expression." *Science*. **299**(5607):697-700.

[14] G. Parmigiani and S. Garrett. (2003) *The Analysis of Gene Expression Data*, chapter 16. Springer-Verlag, New York.

[15] J.M. Ranz, C.I. Castillo-Davis, C.D. Meiklejohn, and D.L. Hartl. (2003) "Sex-Dependent Gene Expression and Evolution of the *Drosophila* Transcriptome." Science. **300**(5626):1742-1745.

[16] G. Terrell and D. Scott. (1990) "Variable Kernel Density Estimation." Technical Report 7, Rice University.

[17] H. Yang, H. Haddad, C. Tomas, K. Alsaker, and E.T. Papoutsakis. (2002) "A Segmental Nearest Neighbor Normalization and Gene Identification Method Gives Superior Results for DNA-Array Analysis." *Proc. Natl. Acad. Sci. USA*. **100**(3):1122-1127.