Bioinformatics Advance Access published December 12, 2006

BIOINFORMATICS

Functional Genomics via Multiscale Analysis: Application to Gene Expression and ChIP-on-chip Data

Gilad Lerman [1]*, Joseph McQuown [3], Alexandre Blais [4], Brian D. Dynlacht [4,5], Guangliang Chen [1], Bud Mishra [2,4]

[1]Department of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455 USA., [2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY, USA. 10012., [3]Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794., [4]NYU School of Medicine, 550 First Avenue, New York, NY, USA 10016., [5]NYU Cancer Institute, 550 First Avenue, New York, NY, USA 10016.

Associate Editor: John Quackenbush

ABSTRACT

We present a fast, versatile, and adaptive-multiscale algorithm for analyzing a wide variety of DNA microarray data. Its primary application is in normalization of array data as well as subsequent identification of "enriched targets", e.g. differentially expressed genes in expression profiling arrays and enriched sites in ChIP-on-chip experimental data.

We show how to accommodate the unique characteristics of ChIP-on-chip data, where the set of "enriched targets" is large, asymmetric and whose proportion to the whole data varies locally.

Our software as well as our raw DNA microarray data with PCR validations are freely available at
http://www.math.umn.edu/∼lerman/supp/bioinfo06.
Contact: lerman@umn.edu

1 INTRODUCTION

Microarray analysis is a high-throughput method to measure abundance of multiple species of target DNA by simultaneous hybridization to an array of DNA probes. When the target DNA is cDNA corresponding to gene expression, it measures the transcriptomic state of a cell under an experimental condition. When the target DNA is sampled from genomic DNA, it measures copy-number variations within a genome as population polymorphisms or as chromosomal aberrations in a tumor genome (Pollack *et al*., 1999). Finally, when the target is genomic DNA selected by immunoprecipitation (IP) with a protein, it identifies those regions of genome that interact with proteins, such as transcriptional factors, thus elucidating regulatory genetic controls (Ren *et al*., 2000).

All these applications use comparative methods. In a "two-color" scheme, simultaneous array hybridization detects target DNAs of two different experiments, which are labeled with different fluorescent dyes. Target DNAs that have differential behavior from one experiment to the other are called "enriched," the objects sought after in these high-throughput experiments. Enriched targets are found in two steps: First, the measurements are transformed through a "normalization" step in order to assign similar local means (or medians) to "presumed unenriched" targets. The purpose of this step is to compensate for experimental sources of variation, like dye-specific effects and hybridization unevenness in DNA arrays (Smyth *et al*., 2003; Buck *et al*., 2004). Next, the normalized data is used as a basis for statistical identification of enriched targets that truly differ between the two analyzed DNA samples.

In practice, the targets are measured optically in terms of a raw intensity value, and analyzed after being transformed by a logarithmic function. With just two samples involved, the following transformed data representation is standard: for every target, the logarithms of intensities (according to the two samples) are averaged to create an A value (log of their geometric mean) and subtracted to create an M value (log of their ratio), and then plotted in an M vs. A plot. The majority of targets will belong to a stable unenriched set of targets, and after correct normalization their M values will be close to zero. The normalized M values of the enriched targets will lie either significantly above or below zero, but not necessarily with any known distributions, or even symmetry.

This paper describes a fast and general multiscale method for normalization of microarray data and identification of enriched targets without assuming any prior distribution. Its utility is greatest when the data are difficult to model statistically, for instance, when they contain unavoidable distortion and asymmetry. Such problems are most acute for ChIP-on-chip experimental data.

ChIP-on-chip experiments combine microarrays ("chips") with Chromatin immunoprecipitation (ChIP) assays to identify the genomic loci bound by any given transcription factor (Ren *et al*., 2000; Blais and Dynlacht, 2005; van Steensel *et al*., 2005). The immunoprecipitated sample represents gene fragments attached to the transcription factor, and is compared with a sample not subjected to immunoprecipitation and thus representing all genes equally ("input" sample). The two samples are labeled with different fluorescent dyes and are co-hybridized onto a DNA microarray representing all gene promoters of the particular species examined. Those spots that show a significant increase in fluorescence in IP sample relative to the input sample are termed enriched spots, and are considered to represent the target genes bound by the transcription factor in the cell nucleus.

There are currently well-established methods for normalization and identification that work well for many cDNA array data

*to whom correspondence should be addressed. E-mail: lerman@umn.edu

Published by Oxford University Press 2006

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

sets (Quackenbush, 2002; Smyth *et al*., 2003). However, in ChIP-on-chip experiments, enriched target DNA segments deviate in only one direction. That is, the M values (log ratio of input to IP signals) of enriched sites are mostly negative. In this case, the statistical distribution of the corresponding M values is asymmetric, hard to model and thus difficult to estimate. Moreover, in such data the whole M values are frequently skewed, their observed distribution varies locally, and the dependence of their local means on the A values is nonlinear. Consequently, common probabilistic techniques have been difficult to adapt to data arising from ChIP-on-chip experiments (Buck *et al*., 2004).

Occasionally, one also encounters cDNA array data with asymmetric distribution of expression values. For example, the sex-biased genes of *D. melanogaster* constitute a subpopulation whose expression values deviate asymmetrically from the bulk of expression values (Parisi *et al*., 2003).

Three algorithms for ChIP-on-chip experimental data have been developed very recently. Two of them: ChIPOTle (Buck *et al*., 2005) and Mpeak (Zheng *et al*., 2005) take into account the "neighbor effect", observed in experiments using locus tiling arrays, to improve the identification of targets. However, those methods cannot be used with microarrays, where genes are represented by only one spot. Another method Chipper (Gibbons *et al*., 2005) applies to microarrays. We show here better performance of our algorithm over the latter method for a specific data set.

We model the microarray data as arising from a mixture of two distributions. The first one (the "stable" part) represents unenriched targets. In the M vs. A plots introduced earlier, this part is concentrated along a graph of an unknown function f mapping A (mean log-intensity) to M (difference in log-intensities): $M = f(A) +$ noise. In the ideal case of perfect correlation between the two intensities, the function f is zero, and the noise is symmetric and homoscedastic (its local variance is independent of A). In reality, the graph is frequently curved due to the systematic sources of variation discussed above, and the local variances around the normalizing curve are not necessarily constant (namely, the data is heteroscedastic). The second component of the mixture distribution represents outliers (enriched targets). The goal of our algorithm is to estimate the conditional mean and variance of the "stable" distribution, ignoring the presence of outliers.

We refer to our algorithm as Multiscale Strip Construction (MSC), as it constructs a normalizing curve (the estimated conditional mean) with an enveloping strip around it (the estimated conditional variance) in a multiscale fashion. The algorithm zooms in adaptively on the "stable" part of the data by constructing parallelograms of decreasing sizes, centered and oriented along the underlying curve. Higher dimensional generalizations of the algorithm for different kinds of data appear elsewhere (Lerman *et al*., 2006).

Our MSC algorithm performs well on various ChIP-on-chip and cDNA array data, even in problematic cases of significant outliers and strong asymmetries, skewness and curvature of main curve.

2 ALGORITHM AND METHODS

We provide a simplified description of the algorithm and leave its more detailed elaborations and analyses to a mathematical paper (Lerman *et al*., 2006).

The main input to our algorithm is a planar data set E of N points. The algorithm identifies a "stable" set within that data and estimates its local mean and standard deviations. It also uses those estimates in order to assess "outliers" coming from a different model.

In order to simplify the technical description, we assume that the data is normalized along the x-axis so that the M values of the data remain constant at 0. We also assume that the "unenriched" ("stable") part of the data is distributed symmetrically around the x-axis. In this case, the algorithm only estimates the local standard deviations of the "stable" set. We refer to this ideal case as the linear-symmetric case, or LS-Case, and explain its generalization later.

In some cases (e.g. the artificial data exemplifying the algorithm in the supplementary material) the task of the algorithm is well-studied. Indeed, it can be addressed by robust estimation of local means (robust regression) and local standard deviations (Rousseeuw *et al*., 1987). However, in many cases, in particular, ChIP-on-chip data, the local percentages of outliers are significantly high, in particular larger than 50% in some regions, and their distribution is asymmetric and hard to model. In order to overcome this obstacle, we suggest a stopping procedure which separates significant "outliers" in a multiscale fashion and forms local regions that cover the "stable" set (excluding those "outliers"). In each local region standard techniques could then be applied to estimate local statistics and then use it to reassess the significance of outliers coming from a different distribution.

The algorithm depends on the following parameters: l_0, c_0, n_0, n_{sh} and α_0. We explain later how to choose their values and elaborate further in Lerman *et al*., 2006.

We sketch our algorithm below (Algorithm 1) for the linear-symmetric case with $n_{sh} = 1$. Later subsections contain the details of the different steps of this scheme and its generalization to other instances. Lerman *et al*., 2006, Figure 2 illustrates its various steps when applied to an artificial data set.

Dyadic Intervals and Rectangles

We fix an interval Q_0 of nearly minimal length containing the projection of the data set onto the x-axis. A dyadic interval with respect to Q_0 is an interval that occurs when dividing recursively Q_0 to halves. It is of level l, if it has been obtained by l consecutive partitions. We denote all dyadic intervals with respect to Q_0 by $\mathcal{D}(Q_0)$. If Q is a dyadic interval, we denote its length by $\ell(Q)$. 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 $P_{Q_0} := Q_0$.

In order to describe the stopping constructions more formally, we will need to define properties of several regions, which extend dyadic intervals to the plane. Figure S1 (supplemental material) illustrates those regions. For an interval $Q \in \mathcal{D}(Q_0)$, its extension $Str(Q)$ to an infinite strip is

$$Str(Q) = Q \times \mathbb{R}.$$

Its extension $Rec(Q)$ to a rectangle (centered around Q) is

$$Rec(Q) = Q \times [-c_0 \cdot \ell(Q), c_0 \cdot \ell(Q)].$$

The "outer" or "putative-enriched" part of $Rec(Q)$ is defined, by removing a subrectangle of appropriate height, as follows:

$$Out(Q) = Rec(Q) \setminus \quad Q \times \left[-\frac{c_0}{2} \cdot \ell(Q), \frac{c_0}{2} \cdot \ell(Q)\right] \quad .$$

Algorithm 1 MSC Algorithm (LS-Case)

- Transform E, so that x-axis = best approximating line
- Set $Q_0 :=$ interval $\supseteq$ projection of E on x-axis
- Set $\ell = 0$, $Stop\_Int = \emptyset$, $Good\_Int = \{Q_0\}$, $N_{\text{stop}} = 0$

while $l \leq l_0$ and $N_{\text{stop}} < N$ **do**

- For each interval Q in $Good\_Int$ form a fixed-ratio rectangle $Rec(Q)$ symmetric around Q
- Compute f_Q which describes a local fraction of putative "outliers" inside $Rec(Q)$
- Compute F_Q which combines f_P's for intervals P's from current and previous levels containing Q
- Compute σ_Q: standard deviation in $Rec(Q)$
- $New\_Stop :=$ all intervals in $Good\_Int$ satisfying: $F_Q > \alpha_0$ or $|Rec(Q) \cap E| < n_0$
- $Stop\_Int := Stop\_Int \cup New\_Stop$
- $Good\_int :=$ dyadic children of intervals in $Good\_int$
- $Good\_int := Good\_int \setminus Stop\_int$
- $N_{\text{stop}} :=$ number of points in stopping intervals
- $l := l + 1$

end while

- Record local standard deviations for stopping rectangles
- Obtain the score R via local standard deviations
- Identify outliers according to R

The Stopping Criterion

We describe the formal steps of the stopping construction and then explain their motivation. Figures S2 and S3 illustrate the stopping construction for an artificial data set. More properties implied by the stopping construction are formulated and proved in Lerman *et al*., 2006.

The algorithm proceeds in a top-down procedure and computes f_Q and F_Q at any dyadic interval Q it visits. The fraction f_Q has the form

$$f_Q = \frac{|Out(Q) \cap E|}{|Str(Q) \cap E|}.$$

The cumulative sum of fractions, F_Q, is computed as follows: First, the algorithm initializes $F_{P_{Q_0}} = 0$, 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 any one of the following two conditions is satisfied:

$$\begin{aligned} &1. \quad F_Q > \alpha_0. \qquad (1)\\ &2. \quad |Rec(Q) \cap E| < n_0. \end{aligned}$$

The main stopping criterion (equation (1)) implies a global estimate on the percentages of initially detected outliers (points outside the rectangular regions) as a function of the parameter α_0 (Lerman *et al*., 2006, Proposition 5.1). The second stopping criterion is necessary for having valid local estimates in each interval.

The stopping construction results in local rectangles which aim to cover most of the "stable" set and to separate away "significant" outliers. The heuristic justification for the success of this separation can be given as follows. The local quantity f_Q measures the local fraction of "putative outliers" in $Rec(Q)$. High values of f_Q, occurring in combination with sufficiently farther local distance from the core of the "stable" distribution, imply presence of locally significant outliers. In order to identify outliers that are also globally significant, we follow several strategies commonly used in harmonic analysis, which combine local quantities at different scales to identify global structure. We use an additive function F_Q, whose analogs have appeared in similar formulations (Jones, 1990; Lerman, 2003).

The Output Functions

The main output function $\tilde{S}$ for the LS-Case estimates the local "standard deviations" of the stable distribution. That is,

$$\tilde{S}(x) = \begin{cases} \sigma_{Q(x)}, & \text{if } |Q| \geq n_0; \\ \sigma_{P_Q(x)}, & \text{otherwise.} \end{cases}$$

We create a smoother version of the above function by generating n_{sh} instances of the corresponding piecewise constant function according to different grids and averaging those piecewise constant functions (Lerman *et al*., 2006, Section 4.6).

The function $\tilde{S}$ estimates "standard deviations" in the rectangles associated with stopping intervals, and requires a small correction to extend it to the region outside those rectangles. We thus alter it by assuming that for each stopping interval Q, the points in $Rec(Q)$ were sampled from the restriction of a Gaussian random variable to that region. The function $\hat{S}$ estimates the standard deviations of the underlying local Gaussian distributions (Lerman *et al*., 2006, Section 4.5). Note that except in this last stage, the algorithm need not make any assumptions about the exact nature of the statistical distributions of data.

Generalization of the Algorithm

Our approach permits two generalizations of the LS-Case algorithm that are necessary for our applications. The first generalization constructs the normalizing curve (which we denote by C) instead of assuming an underlying line. The generalized algorithm approximates the data by lines at different scales and shear the regions $\{Rec(Q)\}_{Q \in \mathcal{D}(Q_0)}$ around those lines. That is, it uses appropriately chosen parallelograms at different scales and locations instead of the rectangles used in the LS-Case. Once the sheared regions $Rec(Q)$ and $Out(Q)$ are defined appropriately for any $Q \in \mathcal{D}(Q_0)$, the algorithm then proceeds mutatis mutandis. The curve C is obtained as the union of line segments. The averaging process described above results in a smooth curve. Figures S4-S7 illustrate this process. We initialize the process by shifting and rotating the data on its principal axis (Lerman *et al*., 2006).

The second generalization allows the algorithm to adapt to asymmetric data, in particular ChIP-on-chip experimental data. It uses asymmetric regions $\{Rec(Q)\}_{Q \in \mathcal{D}(Q_0)}$ (illustrated by Figures S8 and S9). Details of both generalizations appear in Lerman *et al*., 2006.

We also modify slightly the function F_Q following Lerman *et al*., 2006, Section 4.9.1.

Ranking and Identification of Outliers

In order to rank and identify enriched targets, we define a scoring function R for a point (A,M) as follows:

$$R = \begin{cases} \frac{|M - C(A)|}{\hat{S}(A)} & \text{for cDNA arrays} \\ \max(\frac{-(M - C(A))}{\hat{S}(A)}, 0) & \text{for ChIP-on-chip .} \end{cases}$$

Initial p-values are obtained from those scores by assuming that the stable distribution is normal. That is,

$$p\text{-val}(A, M) = 1 - \text{ erf}\,(\frac{R}{\sqrt{2}})\,.$$

Following Reiner *et al*., 2003, we have adjusted the p-values in order to control the false discovery rate of the multiple testing procedure. That is, given a false discovery rate level q, we order the computed p-values: $p_{(1)} \leq \ldots \leq p_{(N)}$ and set

$$p^* = p\text{-value}(\max\{i :\ p_{(i)} \leq q \cdot \frac{i}{N}\}). \qquad (2)$$

We identify the points with p-values less than or equal p^* as enriched.

Choice of Parameters

We fix the values of the following parameters: $l_0 := 10$, $n_0 := 30$, $n_{sh} := 30$ and c_0 is the minimal constant (or almost minimal) for which $E \subseteq Rec(Q)$.

The parameter α_0 is important for good performance of the algorithm. It describes the global expected percentage of outliers. We have developed an algorithm for estimating this parameter (Lerman *et al*., 2006, Section 4.8). The main idea is to apply the MSC algorithm with different values of α_0 and identify outliers at different fixed levels of FDR. For each value of α_0, we draw the curve of the number of outliers detected by the algorithm as a function of the FDR level. We then observe the jumps between the curves. We choose the value of α_0 according to the first significant jump in the profile curves, so that it corresponds to separating the first significant subgroup of outliers (Lerman *et al*., 2006, Section 4.8). In cases of ambiguity of first significant jump of a given replicate, we choose the one closest to the median jump (among all replicates). We show later that the output of our algorithm is not too sensitive to the choice of α_0, but is optimal when fixing α_0 according to our method.

Complexity

The speed of the algorithm for a data set of N points, when using ℓ_0 levels and n_{sh} shifts is of order $O(N \cdot \ell_0 \cdot n_{sh})$ and the required storage is $O(N)$ (Lerman *et al*., 2006, Section 5.3).

Note that in the ideal situation (homoscedastic variance), the algorithm never recurs beyond the highest level Q_0 and outputs essentially the same constant width strip in same time complexity as would the binding-ratio-algorithm, the most popular alternative algorithm currently used to isolate outliers in ChIP-on-chip array data.

In practice, the CPU time of our algorithm (written in a Matlab code which was not optimized) was 1.11 seconds when computing C and the strip $\tilde{S}$ and using a data with $N = 5823$ points and a laptop with Intel Pentium processor of 1.60 GHZ and 1 Gigabyte of RAM (the data is replicate A of Myogenin ChIP-on-chip described in Subsection 3). When also computing the strip $\hat{S}$, the CPU time was 7.96 seconds. For comparison, the CPU times of LOESS using the same data and pc with the bandwidths parameters 0.1, 0.3 and 0.7 are 8.54, 15.28 and 28.05 seconds respectively. While LOESS only normalizes the data, our algorithm also estimates the local standard deviations of the stable distribution and the significance of enriched points.

Clearly, the use of $\tilde{S}$ instead of $\hat{S}$ reduces considerably the computational time. Our experience shows that for values of α_0 less than 0.2, the differences between the two functions are not significant.

3 CASE STUDIES

We demonstrate results of our algorithm for both gene expression and ChIP-on-chip data with emphasis on the latter.

C. acetobutylicum Gene Expression Data

Using our algorithm, we have analyzed cDNA array data comparing gene expression of megaplasmid pSOL1 deficient *C. acetobutylicum* strain M5 relative to its wild type (WT) strain (Yang *et al*., 2003). The pSOL1 genes are postulated to have expression with a broad range of levels in WT, but unexpressed in M5. Therefore, these genes were expected to be characterized as enriched in the WT strain versus the M5 strain.

To measure the statistical power of our algorithm, we focused on the following quantities: the false positive rate (FPR), the true positive rate (TPR) and the identification error (E_r), all described in Yang *et al*., 2003.

Yang *et al*., 2003 have used the same data in order to compare various algorithms for identification of differentially expressed genes, including their own algorithm: SNN-LERM (segmental nearest neighbor method of logarithmic expression ratios). They concluded that their algorithm performed better than the other algorithms.

We have compared FPR, TPR and E_r of both MSC and SNN-LERM for the six glass arrays of M5 vs. WT in the supplemental material of Yang *et al*., 2003. We maintained a similar ratio of outliers and summarized our findings in Table 1. Figures S10 displays the separating strips of the two algorithms for slide 804 and Figure S11 demonstrate the corresponding ROC curves.

Numerical Results	Slide 422	Slide 424	Slide 783	Slide 784	Slide 786	Slide 805
SNN						
TPR	0.093	0.087	0.059	0.257	0.202	0.176
FPR	0.089	0.073	0.069	0.046	0.057	0.058
E_r	0.498	0.493	0.505	0.394	0.427	0.441
MSC						
TPR	0.11	0.10	0.059	0.236	0.21	0.191
FPR	0.085	0.07	0.069	0.05	0.055	0.055
E_r	0.488	0.484	0.505	0.407	0.423	0.432

Table 1. Comparison of SNN-LERM and MSC for identification of C. acetobutylicum pSOL1 genes in six slides of M5 vs. WT experiments (using data where SNN-LERM was shown to be superior to other methods (Yang *et al*., 2003)).

The results indicate better identification by MSC in four out of the six experiments, though the magnitude of improvement is arguably small. In view of the superiority of SNN-LERM over other existing algorithms for this particular data (as claimed by Yang *et al*., 2003), we find our results noteworthy.

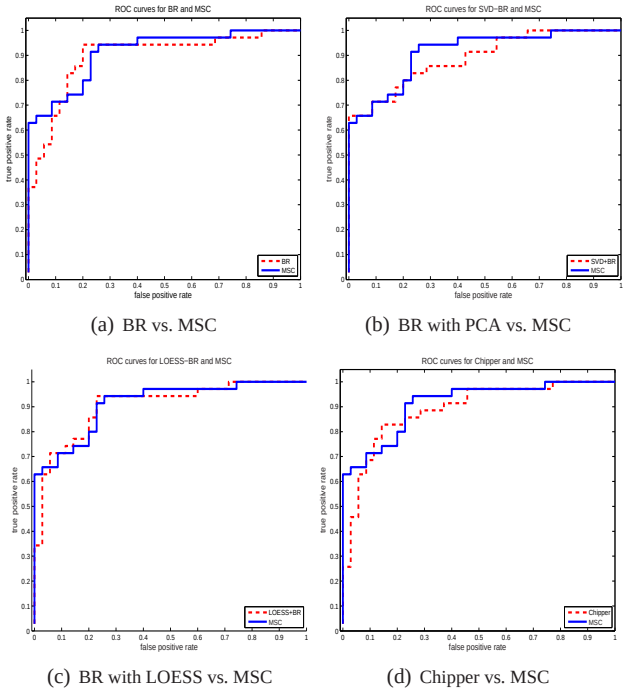
Mouse DNA microarray from ChIP-on-chip experiment

We have performed ChIP-on-chip experiments using the Mm4.7k mouse promoter DNA microarray, with highly specific antibodies

against well-characterized transcription factors. The experiments have been replicated three times. A detailed biological analysis of these experiments is published elsewhere (Blais *et al*., 2005).

In the main data described here, the antibodies recognized Myogenin and the experiment was performed in myotubes. In the supplemental material we have also analyzed ChIP-on-chip experiments where antibodies recognized MyoD in both growing cells and myotubes. Following Blais and Dynlacht, 2005, we have excluded any experiment with more than one replicate with dust speckles on the glass slide or with low spot fluorescence intensity (65% with respect to background).

With an aim to independently validate observed binding of a transcription factor to a given genomic locus, we have performed confirmatory, gene-specific PCR on immunoprecipitated chromatin in the special case of the Myogenin data. This is a method that does not involve DNA amplification, DNA labeling and microarray hybridization, which are the most prominent sources of error and noise in the ChIP-on-chip procedure. We have chosen microarray spots from different levels of binding ratios. Thirty-five tested genes were determined as unambiguously enriched (and thus considered true targets of the transcription factor under study), while thirty-five were considered unenriched. Original data for this comparison is described in Blais *et al*., 2005 and also provided in the supplemental material of this paper.



(a) BR vs. MSC (b) BR with PCA vs. MSC

(c) BR with LOESS vs. MSC (d) Chipper vs. MSC

Fig. 1. ROC Curves comparing each one of the methods: BR, BR with PCA, BR with LOESS and Chipper with MSC. The MSC curve is described by a solid line and the other curves by dashed lines.

In order to determine an optimal value of the parameter α_0, we have applied our method of detecting first jumps in the number of outliers found by MSC. Figure S12 describes those jumps. Accordingly we have chosen $\alpha_0 = 0.2$ for replicates A and C and $\alpha_0 = 0.21$ for replicate B.

We have compared the MSC with the binding ratio method (BR) as applied to this data in Blais *et al*., 2005, binding ratios with respect to the principal axis of the data, binding ratios together with LOESS normalization and the recent Chipper algorithm (Gibbons *et al*., 2005). The binding ratio method identifies enriched sites by selecting (according to a subjective threshold) the points with highest ratios of IP signal to input signal (equivalently, lowest M values). BR can be combined with LOESS by initial application of LOESS normalization and then identifying enriched sites by selecting points with lowest second coordinates. Similarly, BR can be applied with respect to the principal axis by shifting the data so its center of mass is zero, rotating it so the x-axis coincides with the main principal axis and then applying BR. Other approaches for normalization and identification of cDNA arrays yielded even less compelling results than the four methods and are thus not presented here.

Method	MSC	BR	LOESS+BR	PCA+BR	Chipper
Area	0.913	0.895	0.905	.0.891	0.888

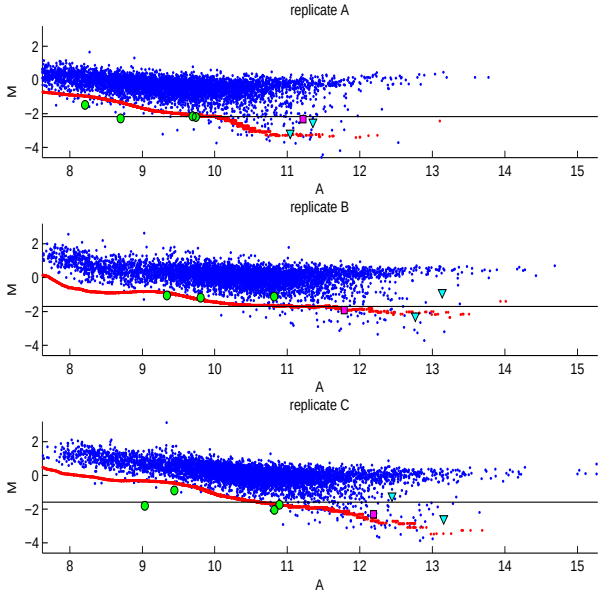
Table 2. Areas below ROC curves for the different methods. The LOESS span parameter, 0.3, has been chosen to maximize its area. The MSC parameter α_0 has been chosen according to first significant jumps (see Figure S12).

α_0	0.1	0.21	0.23	0.26	0.3	0.4	0.5
Area	0.900	0.913	0.915	0.913	0.909	0.907	0.902

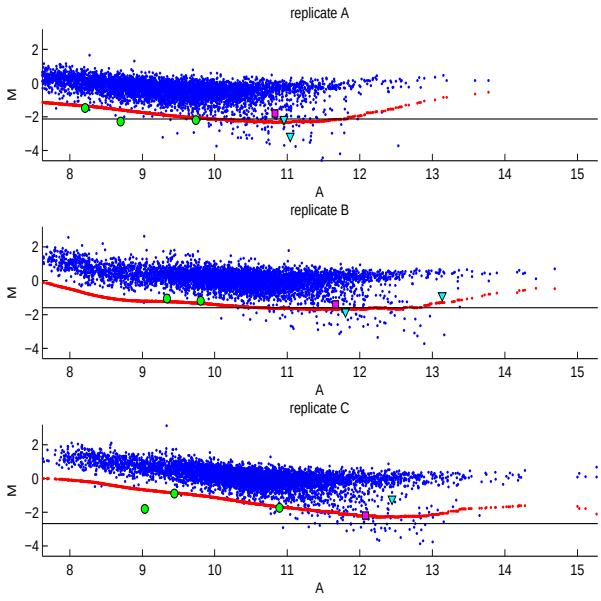
Table 3. Areas below ROC curves for MSC with different values of α_0s.

A comparison along a full range of true positive and negative rates is described by a ROC curve. Figure 1 presents ROC curves comparing MSC with the other four Methods. The ranks of the various methods are averaged among unexcluded replicates and their sorted values are used for identification. The areas below the curves for the different methods are recorded in Table 2. We have chosen carefully the LOESS span parameter to maximize its area below the ROC curve. In both instances, MSC performs slightly better than the four other algorithms over a full range of false positive rates. However, only 1.19% of the data has been verified to be either enriched or unenriched and therefore the differences between the methods (in particular LOESS and MSC) are not statistically significant. We nevertheless find those results important as we are not aware of ChIP-on-chip experimental data with larger percentage of confirmatory PCRs (in practice, not commonly used with large data, since it is both a time-consuming and an expensive process).

While the ROC curve describes a full range of true positive and negative rates, in practice, there is a specific range which is important for identification. We identify such a range by controlling the



(a) BR vs. regular MSC (initial transformation)



(b) BR vs. MSC without initial transformation

Fig. 2. In (a) regular MSC (with initial transformation) is compared with BR for the three replicates of our data. In (b), we have applied MSC without initial shift and rotation onto principal axis. BR threshold is indicated by a straight line, while MSC strip is represented by the thicker curve We have identified the enriched points of MSC in at least two replicates while applying FDR level of 0.1 for each replicate; We identified the same number of enriched points with a weighted BR score. Circles reveal enriched spots that the MSC algorithm distinguished and the binding ratio method failed to distinguish over all 3 replicates, while squares reveal enriched points identified by BR and not by MSC. Triangles reveal points which are not enriched and were identified by BR as enriched, unlike MSC. There were no spots that MSC failed to identify as not enriched, while BR did not. Figure S14 describes the comparison of regular MSC and BR with respect to the coordinates obtained after applying the initial transformation. Figure S17 compares the normalizing curves of MSC with and without initial transformation.

FDR level. We have chosen a level of 0.1. In the absence of clear underlying models in some of the other competing methods, we have balanced them with the same number of identified enriched points (combining all 3 replicates by a weighted score) for the purpose of fairly comparing identification rates. Figure 2(a) illustrates such a comparison between BR and MSC. MSC has identified correctly Chrnb1, Chrng, Cited2 and Myc as enriched, whereas BR misidentified them. However, BR has identified correctly Sema6c as enriched whereas MSC missed it. BR has falsely identified Hist1h2bc and Cacng1 as enriched, unlike MSC. MSC true positive rate is 0.629, whereas that of BR is 0.542. MSC false positive rate is 0, whereas that of BR is 0.057.

Optimal performance of MSC depends on a correct choice of the parameter α_0 and our algorithm for detecting such a choice is a distinctive advantage of our method. Nevertheless, MSC is not highly sensitive to variations in the parameter α_0. Table 3 illustrates this point, by recording areas below ROC curve for different values of α_0 (more details appear in Table S1). The variations in areas is not significant and the optimal area is near our choice of the optimal parameter. Similarly, in Table S4 we record identification of true and false positives and negatives for MSC with different values of α_0 and the corresponding identification values for the other methods with same percentage of detected enriched points, while maintaining an FDR level of 0.1 for MSC.

Our application of MSC includes an initial rotation on principal axis. We have compared it to BR with respect to this axis in order to show that the initial transformation is not enough for good identification (it is worse or at least comparable to regular BR). When applying our method without this initial transformation the area below the ROC is 0.904 (Figure S13 explains the choice of α_0 and Figure S18 presents the corresponding ROC curves). However, the area decreases with higher values of α_0 (see Table S2). Nevertheless, those differences are not statistically significant. Indeed, they are mainly due to a single spot: Cacng1. This spot is falsely identified by MSC without initial rotation as enriched with very low false positive rate. The other methods have identified it as enriched with higher false positive rates (see Figures S15, S16, S19 and Table S3).

The identification for fixed FDR (we have chosen 0.1 but other values worked as well) of the MSC without the initial transformation is good when compared with the other methods, even for a large range of α_0 (see Table S5). Figure 2(b) illustrates such a comparison between BR and MSC. Figure S17 demonstrates the normalizing curves obtained by MSC with and without the initial transformation as well as that of LOESS. The differences are noticeable only in a very sparse region.

Our conclusion is that applying MSC without initial rotation can also work well in identifying outliers. It is more convenient to plot the resulted curve and strip that way, as there is no need to rotate backward. Differences of the two implementations have also been compared for the MyoD data (see Figures S26-S31). The good performance of our method irrespective of the initial transformation is a strong indicator of its robustness. On the other hand, LOESS did not perform as well when rotated on the principal axis (e.g. its area under ROC is $.896$).

We believe that our experimental results provide clear indication of the attractive performance of the MSC method, as it allows investigators to extract more meaningful and reliable information from their data sets. Figures S20 and S21 show instances of failures of some standard techniques to the latter data. We have also applied

our algorithm to ChIP-on-chip experiments where antibodies recognized MyoD in both growing cells and myotubes, but with no confirmatory PCRs and report the results in Figures S22-S31. There are several marked advantages enjoyed by our algorithm: its adaptability to regions of lower or higher variance and to areas of the data which exhibit significant nonlinearity between channels as well as asymmetry; its model for identifying enriched points under a fixed false discovery rate; its fast implementation; its robustness to transformation of the data and to change of parameters and its ability to choose the main parameter α_0 to improve the identification results. In the ideal case when there is no nonlinearity between channels and the variance is relatively constant throughout the data, we expect MSC to perform similarly to the binding ratio method. However, MSC proves its effectiveness in analyzing many important data sets, because one is frequently confronted with experimental results that stray far from the ideal, as numerous types of artifacts (unequal dye incorporation, unequal background in the two channels, different quantum yield of the dyes, etc) remain difficult to control and confound the analysis in the presence of the inherent asymmetry of ChIP-on-chip experiments.

4 BRIEF DISCUSSIONS

The approach described here fills in a substantial void in the analysis of general DNA arrays, in particular arrays from ChIP-on-chip experiments. Namely, it represents an effective method for identifying enriched targets while handling logarithmic ratios of intensities with asymmetric and heteroscedastic characteristics. Currently, most standard techniques fail to analyze a large fraction of these data and many investigators resort to the simple binding ratio method in order to rank "outliers" (e.g. IP-enriched sites in ChIP-on-chip experiments).

The MSC will prove most advantageous for ChIP-on-chip data sets that display mild or pronounced non-linearity, as well as for data sets where the proportion of enriched spots is very large. However, when working on data sets that are close to ideality, it still performs as well as other existing methods.

ACKNOWLEDGEMENT

We are grateful to the NYU Cancer Institute Genomics Facility for providing necessary instrumentation and expertise. We thank Mary Tsikitis and Diego Acosta for their help in performing confirmatory PCR, and Rick Young and Duncan Odom of the Whitehead Institute for advice in the design of a mouse promoter Microarray. We also thank Fang Cheng, Ronald R. Coifman, Peter Jones and Yi (Joey) Zhou for helpful discussions; E. Terry Papoutsakis and Carles Paredes for their help in interpreting the data appearing in their original paper on pSOL1 genes; James Glimm and Jacob Schwartz for commenting on earlier versions of this paper and finally, Mark Green and IPAM (UCLA) for inviting GL and JM to take part in their bioinformatics as well as multiscale geometry meetings, where discussions of similar topics stimulated our research. Special thanks for the careful anonymous reviewers and their constructive suggestions. JM, GL and BM are supported by grants from NSF's ITR program, Defense Advanced Research Projects Agency (DARPA), and New York State Office of Science, Technology & Academic Research (NYSTAR). GL is supported by NSF grant #0612608. BD is supported by an NIH grant #5R01 GM067132-02. AB is supported by a postdoctoral training fellowship from the Fonds de la Recherche en Sante du Quebec.

REFERENCES

Blais,A., Dynlacht,B.D. (2005) Devising transcriptional regulatory networks operating during the cell cycle and differentiation using ChIP-on-chip, *Chromosome Research*, **19**(13), 1499-511.

Blais,A. et al. (2005). An initial blueprint for myogenic differentiation, *Genes Dev.*, **1**:19(5), 553-69.

Buck,M.J., Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics*, **83**, 349-360.

Buck,M.J., Nobel,A.B., Lieb,J.D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data, *Genome Biol.*, **6**(11):R97.

Gibbons,F.D., Proft,M., Struhl,K., Roth,F.P. (2005) Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization, *Genome Biology*, **6**(R96).

Jones,P.W. (1990) Rectifiable sets and the traveling salesman problem, *Invent. Math.*, **102**(1), 1-15.

Lerman,G. (2003) Quantifying curvelike structures of measures by using L_2 Jones quantities, *Comm. Pure App. Math.*, **56**(9), 1294-1365.

Lerman,G., McQuown,J., Mishra,B. (2006) Multiscale Curve and Strip Constructions, Preprint; attached in supplemental material.

Parisi,M., et al. (2003) Paucity of Genes on the *Drosophila* X Chromosome Showing Male-Biased Expression, *Science*, **299**(5607), 697-700.

Pollack,J.R., et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nat Genet*, **23**, 41-46.

Quackenbush,J. (2002) Microarray data normalization and transformation, *Nat Genet*, **32**, 496-501.

Reiner,A., Yekutieli,D., Benjamini,Y. (2003) Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures, *Bioinformatics*, **19**(3), 368-375.

Ren,B., et al. (2000) Genome-wide Location and Function of DNA Binding Proteins, *Science*, **290**, 2306-2309.

Rousseeuw,P.J., Leroy,A.M. (1987) *Robust regression and outlier detection*, Wiley, New York.

Smyth,G.K., Yang,Y.H., Speed,T.P. (2003) Statistical issues in microarray data analysis, In: Functional Genomics: Methods and Protocols, M.J. Brownstein and A.B. Khodursky (eds.), Methods in Molecular Biology, **224**, 111-136.

van Steensel,B. (2005) Mapping of genetic and epigenetic regulatory networks using microarrays, *Nature Genetics*, **37**, S18-S24.

Yang,H., Haddad,H., Tomas,C., Alsaker,K., Papoutsakis,E.T. (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.

Zheng,M., Barrera,L.O., Wu,Y.N., Ren,B. (2005) A probability theory of ChIP-chip data, *Proceedings of Joint Statistical Meetings* (extended preprint can be found at http://preprints.stat.ucla.edu/443/CHIP_zmdl_1036.pdf).