Microarray analysis

Overview

We used two methods to analyze our data: a modified Rosetta error model [1-4] and a maximum likelihood estimate of DNA concentration (MLEDC) method. To evaluate the accuracy of each method and to determine the optimal values of certain cutoffs, we analyzed the binding of the transcription factor GCN4 and compared our results with previously published lists of genes bound by GCN4 and genes that are not regulated by GCN4 [3]. We found that the MLDC method performed better than the Rosetta error model; however this result is due to factors that are specific to our transposon based method and cannot be generalized to ChIP-Chip data or microarray data. In this section we briefly describe the Rosetta error model and the MLDC method, we explain how we estimated the sensitivity and specificity of each method using the GCN4 as a test case, and we discuss why the MLDC method outperforms the Rosetta error model.

The Rosetta Model

The Rosetta error model for expression analysis has been described in detail [1, 2, 4], and a modified version has been used to analyze ChIP-Chip data [3]. We use the latter and briefly review it here. The intensity at spot j in microarray i is modeled by the following expression (using the framework presented by Dror [5]):

$$\tilde{y}_{ij} = v_i(g_{ij}(\zeta_{ij}t_{ij}) + f_j + e_{ij})$$

Here $\tilde{y}_{ij}$ is the intensity of the spot, v_i is the chip normalization factor, ζ_{ij} is a gene specific hybridization factor, t_{ij} is the absolute concentration of DNA complementary to feature j, f_j is spot specific additive noise, and e_{ij} is additive noise. g_{ij} represents the multiplicative noise. Different experiments are normalized by scaling all data on the chip so that the mean intensities are equal (alternatively, spiked in positive controls can be used to normalize [1]). The normalized intensity, y_{ij}, then becomes:

$$y_{ij} = (g_{ij}(x_{ij}) + f_j + e_{ij})$$

Here $x_{ij} = \zeta_{ij}t_{ij}$ and this represents the scaled concentration of DNA complementary to feature j. The scaled concentration can be used for this analysis because we are not interested in the absolute intensity at any given gene, but the difference in intensities between an experiment and a control at feature j. Our control condition is the Sir4 fragment expressed by itself (i.e. not fused to a transcription factor). We are interested in spots that display an increased intensity relative to the control, or the difference between the two measurements. If y_{2j} is the intensity measured at feature j in an experiment (e.g. with Sir4 fused to gcn4), and y_{1j} is the intensity measured at feature j in the control condition (e.g. with Sir4 alone), then we would like to model:

$$y_{2j} - y_{1j} = (g_{2j}(x_{2j}) - g_{1j}(x_{1j}) + e_{1j} + e_{2j})$$

Which we can write without loss of generality as

$$y_{2j} - y_{1j} = (1 - k_{12j}(x_{1j} - x_{2j}) + e_{ij} + e_{mj})$$

Here, k_{12j}, e_{ij}, and e_{mj} are

observations of random variables k and e. The Rosetta model assumes the probability distribution functions for these variables are normally distributed with mean zero. These assumptions yield good empirical results (a critical discussion of these assumptions can be found in [1]).

The variance of e can be estimated from the negative controls in each experiment, and the variance of k can be estimated using the method of Pokholok [3]. In this method, the random variable X is defined:

$$X = (y_{2j} - y_{1j}) / [2\sigma_e^2 + \sigma_k^2 (y_{2j}^2 + y_{1j}^2)]$$

Here σ_e^2 is the variance of e and σ_k^2 is the variance of k. If there was no transposition event at feature j, then X should be normally distributed with mean zero and variance 1. An estimate of σ_k^2 can be obtained by the following algorithm [the following text is taken from Pokholok[3] with minor modifications]:

(a) Select data that follow the noise distribution rather than the signal (e.g. X < 0)

(b) Calculate the X values for all spots given a starting value for σ_k^2 (e.g. 1)

(c) Compare the standard deviation of X for low intensities SD_{low} (e.g. top 10% of the list) to the standard deviation of X for higher intensities SD_{high} (e.g. 20-30% from top)

(d) Set $\sigma_k^2 = \sigma_k^2$ * SD_{high}/ SD_{low} and go back to b for calculation of X values.

Comment: If $SD_{low} > SD_{high}$, σ_k^2 is too large; if $SD_{low} < SD_{high}$, σ_k^2 is too small

(f) Loop (b-d) until the absolute difference between SD_{low} and SD_{high} is close to 0 with desired precision (e.g. 0.001)

The MLEDC method

The Rosetta error model works well when the distribution of intensities in the control channel is similar to the distribution of background intensities in the experimental channel. However, we observed a significant increase in integration "hot-spots" when no TF-SIR4 fusion protein is present, rendering the Rosetta error model inadequate. We developed a second way to analyze the calling card data. For a control we labeled genomic DNA and hybridized it to the microarray reading the green channel. The normalized intensity in the green channel can be written as,

$$Green_{ij} = (g_{ij}(\xi_{ij} D_{ij}) + f_j + e_{ij})$$

Here g_{ij}, ξ_{ij}, f_j, and e_{ij} have the same definitions as before, and D_{ij} is the concentration of genomic DNA. We use enough genomic DNA so that every spot yields a high intensity in the green channel. This means that the additive noise can effectively be ignored, so

$$Green_{ij} = (g_{ij}(\xi_{ij} D_{ij}))$$

The hybridization intensity in the red channel is as before:

$$\mathrm{Re}\, d_{ij} = (g_{ij}(\xi_{ij} t_{ij}) + f_j + e_{ij})$$

The ratio of red intensity to green intensity is,

$$R_{ij} = t_{ij} / D_{ij} + (f_j + e_{ij}) / g_{ij}(\zeta_{ij} D_{ij})$$

Since Dij is the same in each experiment and at each genomic location, we can set D_{ij} = 1 unit for all i,j. Then,

$$R_{ij} = t_{ij} + ((f_j + e_{ij}) / g_{ij}\zeta_{ij})$$

So, R_{ij}, the ratio of intensity in the red channel to the green channel, is given by the concentration of DNA in feature i in experiment j plus two additive noise terms. The means and standard deviations of both of these variables are very small when compared to the value of R_{ij} at features where a transposition event has occurred (data not shown). Thus, at these loci, R_{ij} is a good estimate of DNA concentration. Therefore, by ranking the probes by their average ratio across three experiments and applying a cutoff, we are able to identify the loci at which a transposition event occurred. To filter out random transposition events (i.e. those not directed by the fused transcription factor) we required that 2 of the 3 replicates displayed a R_{ij} value greater than an empirically derived cutoff. The appropriate cutoffs were determined by choosing the value of R that maximizes the sensitivity of the GCN4 experiments at a specificity of 97.5%.

Estimating the Sensitivity and Specificity of the Gcn4 calling cards.

To evaluate their CHIP-CHIP experiments, Young and colleagues compiled a list of 75 gcn4 targets as well as 935 genes whose regulatory regions are not likely to be bound by gcn4 [3]. To be included in the positive list, a gene must have displayed significant gcn4 binding as ascertained by previous ChIP-Chip experiments, displayed an expression change upon amino acid starvation, and contained a high-scoring gcn4 binding site in its upstream region. A gene on the negative list did not bind gcn4 in previous ChIP-Chip experiments, displayed <2 fold expression change upon amino acid starvation, and did not contain a high-scoring binding site in its upstream region. When the Rosetta analysis method was used, we obtained 45% sensitivity at 97.5% specificity. In contrast, the MLEDC method achieved 51% sensitivity at 97.5% specificity. We manually examined the intensities of the false negative genes in the MLEDC analysis – the majority of these features displayed little to no fluorescence in the red channel, suggesting that these features were categorized as negatives because no transposition event had occurred in these samples and were not due to inaccurate assumptions in our error model.

1. Dror, R.O., *Noise models in gene array analysis.* 2001, MIT Department of Electrical Engineering and Computer Science: Cambridge.
2. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.
3. Pokholok, D.K., et al., *Genome-wide map of nucleosome acetylation and methylation in yeast.* Cell, 2005. **122**(4): p. 517-27.
4. Weng, L., et al., *Rosetta error model for gene expression analysis.* Bioinformatics, 2006. **22**(9): p. 1111-21.
5. Dror, R.O., et al., *Bayesian estimation of transcript levels using a general model of array measurement noise.* J Comput Biol, 2003. **10**(3-4): p. 433-52.