Pacific Symposium on Biocomputing 13:515-526(2008)

CMARRT: A TOOL FOR THE ANALYSIS OF CHIP-CHIP DATA FROM TILING ARRAYS BY INCORPORATING THE CORRELATION STRUCTURE

PEI FEN KUAN[1], HYONHO CHUN[1], SÜNDÜZ KELEŞ[1,2*]

[1] *Department of Statistics,*
[2] *Department of Biostatistics and Medical Informatics,*
1300 University Avenue, University of Wisconsin, Madison, WI 53706.
[*] *E-mail: keles@stat.wisc.edu*

Whole genome tiling arrays at a user specified resolution are becoming a versatile tool in genomics. Chromatin immunoprecipitation on microarrays (ChIP-chip) is a powerful application of these arrays. Although there is an increasing number of methods for analyzing ChIP-chip data, perhaps the most simple and commonly used one, due to its computational efficiency, is testing with a moving average statistic. Current moving average methods assume exchangeability of the measurements within an array. They are not tailored to deal with the issues due to array designs such as overlapping probes that result in correlated measurements. We investigate the correlation structure of data from such arrays and propose an extension of the moving average testing via a robust and rapid method called `CMARRT`. We illustrate the pitfalls of ignoring the correlation structure in simulations and a case study. Our approach is implemented as an `R` package called `CMARRT` and can be used with any tiling array platform.

Keywords: ChIP-chip, moving average, autocorrelation, false discovery rate.

1. Background

Whole genome tiling arrays utilize array-based hybridization to scan the entire genome of an organism at a user specified resolution. Among their applications are ChIP-chip experiments for studying protein-DNA interactions. These experiments produce massive amounts of data and require rapid and robust analysis methods. Some of the commonly used methods are ChIPOTle,[1] Mpeak,[2] TileMap,[3] HMMTiling,[4] MAT[5] and TileHGMM.[6] Although these algorithms have been shown to be useful, they don't address the issues due to array designs. The most obvious issue is the correlation of the measurements from probes mapping to consecutive genomic locations.[15] The basis for such a correlation structure is due to both overlapping probe

design and fragmentation of the DNA sample to be hybridized on the array. There are several hidden Markov model (HMM) approaches to address the dependence among probes but the current implementations are limited to first order Markov dependence.[4] Generalizations to higher orders increase the computational complexity immensely. We investigate the correlation structure of data from complex tiling array designs and propose an extension of the moving average approaches[1,7] that carefully addresses the correlation structure. Our approach is based on estimating the variance of the moving average statistic by a detailed examination of the correlation structure and is applicable with any array platform. We illustrate the pitfalls of ignoring the correlation structure and provide several simulations and a case study illustrating the power of our approach `CMARRT` (Correlation, Moving Average, Robust and Rapid method on Tiling array).

2. Methods

Let $Y_1, ..., Y_N$ denote measurements on the N probes of a tiling path. Y_i could be an average log base 2 ratio of the two channels or (regularized) paired t-statistic for arrays with two channels (e.g., Nimblegen) and a (regularized) two sample t-statistic for single channel arrays (Affymetrix) at the i-th probe. These wide range of definitions of Y make our approach suitable for experiments with both single and multiple replicates per probe.

A common test statistic for analyzing ChIP-chip data is a moving average of Y_i's over a fixed number of probes or fixed genomic distance.[1,3,7] The parameter w_i will be used to define a window size of $2w_i + 1$, i.e., w_i probes to the right and left of the i-th probe. In the case of moving average across a fixed number of probes for tiling arrays with constant probe length and resolution, the window size w_i is calculated by $L \times (2w_i + 1) - 2w_i \times O = FL$, where L is the probe length, O is the overlap between two probes and FL is the average fragment size. Our framework also covers tiling arrays with non-constant resolution. In this case, w_i will be different for each genomic interval and corresponds to the number of probes within a fixed genomic distance. For simplicity in presentation, we will utilize window size of fixed number of probes. We assume that the data has been properly normalized by potentially taking into account the sequence features,[8] and that $E[Y] = \mu$ and $\text{var}(Y) = \sigma^2$. Consider the following moving average statistic

$$T_i = \frac{1}{2w_i + 1} \sum_{j=i-w_i}^{i+w_i} Y_i. \quad (1)$$

Then, standard variance calculation leads to

$$\mathrm{var}(T_i) = \frac{1}{(2w_i+1)^2}\left((2w_i+1)\sigma^2 + \sum_{j=i-w_i}^{i+w_i}\sum_{k\neq j}\mathrm{cov}(Y_j, Y_k)\right). \quad (2)$$

The standardized moving average statistic is given by

$$S_i = \frac{T_i}{\sqrt{\mathrm{var}(T_i)}}. \quad (3)$$

Standard practice of using moving average statistics relies on (1) estimating σ^2 based on the observations that represent lower half of the unbound distribution; (2) ignoring the covariance term in equation (2); (3) and obtaining a null distribution under the hypothesis of no binding at probe i. In particular, ChIPOTle considers a permutation scheme where the probes are shuffled and the empirical distribution of the test statistic over several shufflings is used as an estimate of the null distribution. As an alternative, a Gaussian approximation is utilized assuming that Y_i's are independent and identically distributed as normal random variables under the null distribution. As discussed by the authors of ChIPOTle, both approaches assume the exchangeability of the probes under the null hypothesis. Exchangeability implies that the correlation within any subset of the probes is the same. However, empirical autocorrelation plots from tiling arrays often exhibit evidence against this (Fig. 1). In particular, in the case of overlapping designs, a correlation structure is expected by design. When the spacing among the probes is large, correlation diminishes as expected (the right panel of Fig. 1), and this was the case for the dataset on which ChIPOTle was developed.

We illustrate the problem with ignoring the correlation structure on a ChIP-chip dataset from an E-coli RNA Polymerase II experiment utilizing a Nimblegen isothermal array (Landick Lab, Department of Bacteriology, UW-Madison). The probe lengths vary between 45 and 71 bp, tiled at a 22 bp resolution. Approximately half of the probes are of length 45 bp. We compute the standardized moving average statistic S_i (assuming $\mathrm{cov}(Y_j, Y_k) \neq 0$) and S_i^* (assuming independence of Y_i's). A method of estimating $\mathrm{cov}(Y_j, Y_k)$ is described in the next section. The p-values for each S_i and S_i^* are obtained from the standard Gaussian distribution under the null hypothesis. We expect the quantiles of S_i and S_i^* for unbound probes to fall along a 45° reference line against the quantiles from the standard Gaussian distribution, whereas the quantiles for bound probes to deviate from this reference line. As evident in Fig. 2, if the correlation structure is ignored, the distribution of S_i^*'s for unbound probes deviates from the standard

Gaussian distribution. Since the data is obtained from a RNA Polymerase II experiment, we expect a larger number of points, corresponding to promoters, to deviate from the reference line. An additional diagnostic tool is the histogram of the p-values. If the underlying distributions for S_i and S_i^* are correctly specified, the p-values obtained should be a mixture of uniform distribution between 0 and 1 and a non-uniform distribution concentrated near 0. The histograms of the p-values (Fig. 2) again illustrate that the distribution for T_i is misspecified.

2.1. *Estimating the correlation structure*

Although it is desirable to develop a structured statistical model that captures the correlations, developing such a model is both theoretically and computationally challenging due to the complex, heterogeneous data generated by tiling array experiments. We propose a fast empirical method that estimates the correlation structure based on sample autocorrelation function. The covariance $\mathrm{cov}(Y_j, Y_{j+k})$ can be estimated from the sample autocorrelation $\hat{\rho}(k)$ and sample variance $\hat{\sigma}^2$,[10]

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}, \quad \widehat{\mathrm{cov}}(Y_j, Y_{j+k}) = \hat{\rho}(k)\hat{\sigma}^2. \qquad (4)$$

The following strategy is used in `CMARRT` for estimating the correlation structure. The top $M\%$ of outlying probes which roughly correspond to bound probes are excluded in the estimation of $\hat{\rho}(k)$. For the remaining probes, the sample autocorrelation at lag k ($\hat{\rho}_j(k)$) is computed for each segment j consisting of at least N consecutive probes. Genomic regions flanking a large gap or repeat masked regions will be considered as two separate segments. For any lag k, we let $\hat{\rho}(k)$ to be the average of $\hat{\rho}_j(k)$ over j. Here, N can be considered as a tuning parameter and our initial experiments with ENCODE datasets suggest that $N = 500$ works well in practice based on the diagnostic plots discussed in Section 1. M is an anti-conservative preliminary estimate of the percentage of bound probes which can be obtained under the assumption of independence among probes (usually $\sim 1 - 5\%$, depending on the type of ChIP-chip experiment).

3. Simulation studies

In this section, we investigate the performance of `CMARRT`, the conventional normal approximation approach under the independence assumption (Indep) and the HMM option in TileMap under various scenarios where we

know the true bound regions in terms of sensitivity and specificity while controlling FDR at various levels used in practice.

Simulation I: Autoregressive model. We consider the following model

$$Y_i = N_i + R_i, \quad N_i = \sum_{k=1}^{p} \alpha_{i-k} N_{i-k} + \epsilon_i, \tag{5}$$

where N_i is the autoregressive background component and R_i is the real signal. We generate 100,000 N_i from AR(p) to represent the background component under the assumption of $\text{cor}(N_i, N_{i+k}) = \rho^{0.4(k-1)+1}$ and randomly choose 500 peak start sites. We let the size of a peak to be 10 probes, so that $\sim 5\%$ of the probes belong to bound regions. To design scenarios similar to what we have observed in practice, we also allow for ~ 3 outliers within a bound region. The data is simulated from various p (AR order), ρ ($\text{cor}(N_i, N_{i+k})$) and σ ($\text{var}(N_i)$) for the background component, and strength c for the real signal.

Simulation II: Hidden Markov model. In this scenario, the data is simulated from hidden markov models (HMMs)[12] with explicit state duration distribution to introduce direct dependencies at the probe level observations. Let the duration HMM densities be $p_{S_i}(d_i)$ $\sim$Geometric(p_{S_i}). The transition probabilities (a_{ij}) and the parameters p_{S_i} in the duration HMM densities are chosen such that $\sim 5\%$ of the probes belong to bound regions. We consider the joint observation density $f_{N_i}(Y_1, Y_2, ...Y_{d_1}) \sim MVN(0, \Sigma_N)$ for the unbound regions and $f_{B_i}(Y_1, Y_2, ...Y_{d_1}) \sim MVN(\mu, \Sigma_B), \mu > 0$ for the bound regions, where MVN denotes the multivariate normal distribution. The parameters μ, Σ_N and Σ_B are chosen such that generated data resembles observed ChIP-chip data exhibiting correlations at the observation level.

Each simulation scenario is repeated 50 times. A probe is declared as bound if its adjusted p-value[11] is smaller than a pre-specified FDR level α when analyzing with `CMARRT` and Indep. For TileMap, we use the direct posterior probability approach[13] to control the FDR.

3.1. *Results of simulations I and II*

In Fig. 3, we summarize the sensitivity at the peak level and the specificity at various FDR thresholds from Simulation I for `CMARRT`, Indep, and TileMap. `CMARRT` is able to identify most of the bound regions at FDR of 0.05 and above while TileMap tends to be more conservative in declaring bound regions as shown in the sensitivity plots. Although Indep has

the highest sensitivity, it also has a high proportion of false positives. The specificity of Indep is significantly lower compared to CMARRT, even under the case of low correlation among the probes. Similar results are obtained in Simulation II under the duration HMM (Fig. 4). The left panels show the sensitivity and specificity for the case of smaller peaks with an average peak size of 10 probes while the right panels are for the case of larger peaks of size 20 probes on average. These results illustrate the superior performance of CMARRT in terms of both sensitivity and specificity even when the data is generated from a complex model. The heuristic way of estimating the correlation structure in CMARRT is able to reduce the false positives (specificity) significantly, but not at the expense of increasing false negatives (sensitivity). On the other hand, ignoring the correlation structure results in a higher proportion of false positives. Additionally, the HMM option in TileMap is more conservative than the moving average approach when the FDR is controlled at the same level.

4. Case study: ZNF217 ChIP-chip data

We provide an illustration of CMARRT with a ZNF217 ChIP-chip data tiling the ENCODE regions (available from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/)[14] with accession number GSE6624). The ENCODE regions were tiled at a density of one 50-mer every 38 bp, leading to $\sim$ 380,000 50-mer probes on the array. We analyze two different replicates of this dataset separately and compare the analysis on these single replicates. In Krig et al.,[14] the bound regions were identified with the Tamalpais Peaks program,[9] which requires a bound region to have at least 6 consecutive probes in the top 2% of the log base 2 ratios. This criteria tends to be too stringent and fails to identify bound regions which contain a few outlier probes with log base 2 ratios below the top 2% threshold and may result in a higher level of false negatives. In the top right panel of Fig. 5, we show one potential peak missed by the Tamalpais Peaks program. In such cases, the sliding window approach is more powerful for finding peaks. Moreover, this method also assumes the observations are independent. As evident in the left panel of Fig. 1, observations from nearby probes in this tiling array are correlated. As shown in Fig 5, the histograms of p-values for the unbound probes under the independence assumption deviates from the expected distribution in both replicates. Similar problem is present in the normal quantile-quantile plots (online supp. mat.) when the correlation structure is ignored.

As in Krig et al.,[14] we require the number of consecutive probes in each

bound region to be at least 6. A set of peaks is obtained for each replicate at a given FDR control. We assess the extent of overlaps between the set of peaks in these two replicates. The results are summarized in Table 1. All the methods identified more peaks in replicate 1 than replicate 2. Therefore, using the peaks from rep 1 as reference, the common peaks are defined as the percentage of overlapping peaks in replicate 2. For all FDR thresholds (except 0.01), `CMARRT` has the highest value of common peaks, followed by Indep and TileMap, which illustrates the consistency of the peaks identified by `CMARRT`.

As an independent validation, we determine the location of bound regions relative to the transcription start site (TSS) of the nearest gene using GENECODE genes from UCSC Genome Browser as in Krig et al.[14] (Table 1). For a given FDR control, the percentage of peaks located within $\pm 2kb$, $\pm 10kb$ and $\pm 100kb$ of the TSS is the highest in `CMARRT`, followed by Indep and TileMap. As expected, these numbers decrease as we increase the FDR threshold for all the three methods. These results illustrate the power of `CMARRT` in detecting biologically more plausible bound regions of ZNF217.

5. Discussion

We have investigated and illustrated the pitfalls of ignoring the correlation structure due to tiling array design in ChIP-chip data analysis. We proposed an extension of the moving average approaches in `CMARRT` to address this issue. `CMARRT` is a robust and fast algorithm that can be used with any tiling platform and any number of replicates. Both the simulation results and the case study illustrate that `CMARRT` is able to reduce false positives significantly but not at the expense of increasing false negatives, thereby giving a more confident set of peaks. We have recently became aware of the work of Bourgon[15] who carefully studies the correlation structure in ChIP-chip arrays and proposes a fixed order autoregressive moving average model (ARMA(1, 1)) and we are in the process of comparing `CMARRT` with this approach.

`CMARRT` is developed using the Gaussian approximation approach and the diagnostic plots illustrated can be utilized to detect whether a given dataset violates this assumption. One possible relaxation of this assumption is a constrained permutation approach that aims to conserve the correlation structure among the probes under the null distribution. Implementation of such an approach efficiently is a challenging future research direction.