Pacific Symposium on Biocomputing 13:465-476(2008)

A PARAMETRIC JOINT MODEL OF DNA-PROTEIN BINDING, GENE EXPRESSION AND DNA SEQUENCE DATA TO DETECT TARGET GENES OF A TRANSCRIPTION FACTOR

WEI PAN[1], PENG WEI[1], ARKADY KHODURSKY[2]

[1]*Division of Biostatistics, School of Public Health,* [2] *Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota*

This paper concerns with predicting the regulatory targets of a transcription factor (TF). We propose and study a joint model that combines the use of DNA-protein binding, gene expression and DNA sequence data simultaneously; a parametric mixture model is used to realize unsupervised learning, which however can be extended to semi-supervised learning too. We applied the methods to an E coli dataset to identify the target genes of LexA, which, along with applications to simulated data, demonstrated potential gains of jointly modeling multiple types of data over using only one type of data.

1. Introduction

This paper concerns with identifying the transcriptionally regulated target genes of a transcription factor (TF). The task is commonly approached based on one of the three data types: DNA-protein binding data (also called ChIP-chip data or genome-wide location analysis) surveying genome-wide DNA-TF interactions [11,12],microarray gene expression data comparing expression changes before and after perturbing the function of, e.g. by knocking-out, a TF-coding gene [5], and DNA sequence data which are aligned and scanned to find specific binding sites or motifs of a TF [1,13], Because of relatively high noise levels with high-throughput data, using only one data source may result in high false positives or false negatives. As a compensation, it is now widely recognized that an integrative analysis of multiple types of data should be more efficient in identifying the target genes of a TF [2,4,19,26]. With the ever-increasing availability of various types of high-throughput data, a main challenge is how to integrate them effectively. In the literature, there are several classes of the approaches. First, one can use one type of data to validate results from analyses of other types

Pacific Symposium on Biocomputing 13:465-476(2008)

of data [24]. Second, one first conducts a separate analysis on each type of data and then combine their results [27]. Third, regression analyses of one type of data (e.g. gene expression) on another type (e.g. DNA sequence) [3,4,23]. Fourth, one uses one type of data to generate priors or hypotheses for analyzing other types of data; e.g. Liu et al [14] used binding data to generate candidate binding regions, then used DNA sequence data to locate binding sites; Xie et al [29] used expression data to generate a prior list of potential binding targets, which was then utilized to analyze binding data. Finally, a joint model of multiple types of data can be employed to use all the data simultaneously to draw inference or make predictions, which is presumably more efficient than many other alternatives; our method belongs to this class, which also includes the following ones for detecting the targets of a TF. Wang et al [26] proposed a parametric mixture model for both DNA sequence data and binding (or expression) data; our method is similar to theirs except that we used three data sources and a different format of DNA sequence data. Pan et al [18] proposed a nonparametric mixture model; it requires duplicated arrays, not applicable to the E coli expression data to be analyzed here. Xie [28] proposed a fully parametric Bayesian approach using binding, expression and DNA data; because of analytically intractable posterior probability calculations, computationally intensive simulation methods (MCMC) were used to draw inference. Our work here shows that a simple parametric mixture model similar to that of Wang et al [26] works well, even when some parametric modeling assumptions are moderately violated, while accommodating more than two sources of data; furthermore, we extend the method from unsupervised learning to semi-supervised learning.

This paper is organized as follows. We first introduce our joint model as a parametric mixture model, then we outline an EM algorithm to estimate the parameters in the model and thus obtain posterior probabilities to draw inference. We present an application of the methods to an E coli dataset to identify the targets of LexA, comparing the results with the known and putative targets listed in regulonDB (v5.5) [21] and in Wade et al [25]. We also show results of simulation studies to demonstrate statistical efficiency gains from joint modeling over using only one data source. We end with a short discussion on some possible future work.

Pacific Symposium on Biocomputing 13:465-476(2008)

2. Methods

2.1. *A Joint Model*

Our goal is to identify which genes in a genome are the targets of a given TF. To be concrete, we consider three data sources corresponding to DNA-protein binding, gene expression and DNA sequence data, as to be used for an E coli example. We assume that the three data sources can be summarized as (X_i, Y_i, Z_i) for each gene i, $i = 1, ...G$: X_i is a summary or test statistic measuring the relative abundance of the TF binding to gene i, or the statistical significance of rejecting a null hypothesis that gene i is not bound by the TF; Y_i is a test statistic for differential expression of gene i when the TF-coding gene's function is perturbed; Z_i is a score measuring the degree to which one of its subsequences matches a known motif for the TF. Depending on whether gene i is a target or not, we have $T_i = 1$ or $T_i = 0$ respectively. To realize unsupervised learning, it is natural to assume that (X_i, Y_i, Z_i) comes from a mixture distribution: $f(x, y, z) = \pi f_1(x, y, z) + (1 - \pi) f_0(x, y, z)$, each component corresponding to the subpopulation of the genes with $T_i = 1$ or $T_i = 0$ respectively, and π is the prior proportion of the target genes. Further, we assume that conditional on T_i, the three data sources are independent; that is

$$\begin{aligned} f(x, y, z) = \; & \pi f_{11}(x; \theta_{11}) f_{12}(y; \theta_{12}) f_{13}(z; \theta_{13}) + \\ & (1 - \pi) f_{01}(x; \theta_{01}) f_{02}(y; \theta_{02}) f_{03}(z; \theta_{03}), \end{aligned}$$

where θ_{jk}'s are the (unknown) parameters for distributions f_{jk}. To infer whether gene i is a target, we use the posterior probability

$$Pr(T_i = 1 | X_i, Y_i, Z_i) = \frac{\pi f_{11}(X_i; \theta_{11}) f_{12}(Y_i; \theta_{12}) f_{13}(Z_i; \theta_{13})}{f(X_i, Y_i, Z_i)}.$$

Here we use $f_{jk} = \phi(.; \mu_{jk}, \sigma_{jk})$, a normal probability density function with mean μ_{jk} and variance σ^2_{jk}.

2.2. *Estimation via EM*

An EM algorithm [6] can be derived to estimate the unknown θ_{jk}'s. Given T_i, the complete data log-likelihood is

$$\log L_c = \sum_{i=1}^{G} T_i \log \pi f_1(X_i, Y_i, Z_i) + (1 - T_i) \log(1 - \pi) f_0(X_i, Y_i, Z_i).$$

Pacific Symposium on Biocomputing 13:465-476(2008)

The E-step is to calculate the conditional expectation

$$Q = E(\log L_c|Data) = \tau_i[\log\pi + \log f_1(X_i, Y_i, Z_i)] + (1-\tau_i)[\log(1-\pi) + \log f_0(X_i, Y_i, Z_i)],$$

where $\tau_i = Pr(T_i = 1|X_i, Y_i, Z_i)$. The M-step maximizes the above Q with respect to the unknown parameters. We use the generic notation $\theta^{(m)}$ to denote the updated estimate of θ in iteration m; it is easy to verify that, at iteration $m+1$,

$$\tau_i^{(m+1)} = \frac{\pi^{(m)} f_1^{(m)}(X_i, Y_i, Z_i)}{\pi^{(m)} f_1^{(m)}(X_i, Y_i, Z_i) + (1-\pi^{(m)}) f_0^{(m)}(X_i, Y_i, Z_i)}$$

where

$$f_j^{(m)}(X_i, Y_i, Z_i) = \phi(X_i; \mu_{j1}^{(m)}, \sigma_{j1}^{(m)})\phi(Y_i; \mu_{j2}^{(m)}, \sigma_{j2}^{(m)})\phi(Z_i; \mu_{j3}^{(m)}, \sigma_{j3}^{(m)}),$$

for $j = 1, 2$, and

$$\mu_{11}^{(m+1)} = \sum_{i=1}^{G} \tau_i^{(m)} X_i / \sum_{i=1}^{G} \tau_i^{(m)}, \quad \sigma_{11}^{2,(m+1)} = \sum_{i=1}^{G} \tau_i^{(m)} (X_i - \mu_{11}^{(m)})^2 / \sum_{i=1}^{G} \tau_i^{(m)},$$

$$\mu_{10}^{(m+1)} = \sum_{i=1}^{G} (1-\tau_i^{(m)}) X_i / \sum_{i=1}^{G} (1-\tau_i^{(m)}),$$

$$\sigma_{10}^{2,(m+1)} = \sum_{i=1}^{G} (1-\tau_i^{(m)}) (X_i - \mu_{10}^{(m)})^2 / \sum_{i=1}^{G} (1-\tau_i^{(m)}),$$

and $\pi^{(m+1)} = \sum_{i=1}^{G} \tau_i^{(m+1)}/G$, where the updates for other μ_{jk} and σ_{jk}'s are similar and omitted. The above iterations are continued until convergence. Because the EM may converge to a local maximum point, multiple starting values are needed, and the one with the maximum log-likelihood is chosen. The resulting estimates are maximum likelihood estimates (MLEs); the final τ_i are used to rank the genes for their likelihoods of being a target.

2.3. *Other Models*

The above joint model is for three data sources; it is straightforward to have a model for more or less than three data sources, and its corresponding EM updates for parameter estimation. For example, if we use only one source of data, say X_i's, we can have a corresponding mixture model

$$f(x) = \pi f_1(x; \theta_1) + (1-\pi) f_0(x; \theta_0),$$

Pacific Symposium on Biocomputing 13:465-476(2008)

and the posterior probability $Pr(T_i = 1|X_i) = \pi f_1(X_i;\theta_1)/f(X_i)$. The EM updates are

$$\tau_i^{(m+1)} = \frac{\pi^{(m)}\phi(X_i;\mu_{11}^{(m)},\sigma_{11}^{(m)})}{\pi^{(m)}\phi(X_i;\mu_{11}^{(m)},\sigma_{11}^{(m)}) + (1-\pi^{(m)})\phi(X_i;\mu_{01}^{(m)},\sigma_{01}^{(m)})},$$

and the updates for $\mu_{11}, \mu_{01}, \sigma_{11}, \sigma_{01}$ and π are exactly the same as before. Again at the convergence, we use the posterior probabilities τ_i to rank the genes.

2.4. *Extensions to Semi-supervised Learning*

The approaches taken so far are unsupervised learning, assuming that no known targets for the TF, which is not usually true. Supervised learning approaches have been proposed [32], which however may not work well if there are only few known targets for the TF, e.g., for LexA. We can extend our proposal to semi-supervised learning, combining the strengths of unsupervised and supervised learning, which is an advantage of the mixture model [15]. Suppose that the first G_1 genes are known targets while the remaining ones may or may not be. The models are the same as before. The parameter estimation procedures are also similar except that $\tau_i = 1$ for $i = 1, ..., G_1$.

Although in general semi-supervised learning improves over unsupervised or supervised learning, for our example, because there were only few known targets of LexA, the results of semi-supervised learning were similar to that of unsupervised learning. We will skip the discussion of semi-supervised learning. Nevertheless, we expect that this semi-supervised learning will be useful for other TFs and other types of data.

3. Results

3.1. *E coli data*

We extracted the DNA-protein binding data [25] and gene expression data [5] from the authors' supplied web sites respectively, and DNA sequence data from the NCBI and Affymetrix web sites.

The binding data contained two LexA samples (called LexA1 and LexA2 respectively) and two control samples (one Gal4 and one MelR (no Ab, no antibody) samples) hybridized on four Affymetrix Antisense Genome Arrays respectively. We downloaded the raw intensity data (i.e. CEL files) from the authors' supplied web page. Largely following Wade et al [25],

Pacific Symposium on Biocomputing 13:465-476(2008)

we processed the data in below steps. First, we used the Bioconductor R package `affy` to pre-process the data, including background correction with MAS 5 algorithm, and quantile normalization. Second, we calculated four $\log_2$ intensity ratios (LIRs), corresponding to the four combinations of any two arrays, for each probe: LexA_1/Gal4, LexA_1/no Ab, LexA_2/Gal4, LexA_2/no Ab; a large LIR indicated a locus containing enriched LexA. Third, we mapped each probe to a genome position based on the Affymetrix Ecoli_ASv2 annotation file. Fourth, for each of the four array combinations, we smoothed the LIRs over all probes with a sliding window of 1250 bp. Fifth, for each gene in each array combination, we identified its LIR peak among the probes belonging to the gene's coding and intergenic regions (if any) separately. Finally, each gene i's binding score or signal X_i was taken to be the average of its four LIR peaks from its coding region, or if there were probes from its intergenic region, X_i was the larger one of i) the average of its four LIR peaks from its coding region and ii) that from its intergenic region. The final step differed from that in Wade et al: they had an extra step to identify a candidate LexA-bound region/block containing ≥ 20 consecutive probes with all LIRs ≥ 0.17; they calculated the average of the four peaks only for the genes with such blocks, which were taken as candidate binding targets of LexA; they identified about 50 such binding targets. Because for our purpose, we would like to obtain a binding score for every gene, obviously we could not follow their route. This procedural difference contributed to some differences in X_i's between theirs and ours.

The expression data were drawn from four cDNA microarrays profiling gene expression levels for the wild type before and 20-minute after UV treatment, and for the lexA mutant before and 20-minute after UV treatment, respectively; a common control sample was used for each array. Two-channel intensities on each array were normalized using the loess local smoother to eliminate dye bias, as implemented in the R package `sma` [30] Suppose that normalized log-ratios of the two-channel intensities for gene i on the four arrays were M_{1i},...,M_{4i} respectively, then we used the summary statistics for gene expression data as $Y_i = (M_{1i} - M_{2i}) - (M_{3i} - M_{4i})$. Because LexA is known to be a repressor of some "SOS response" genes, it is expected that the transcriptional targets of LexA should have larger values of Y_i's (i.e. expression changes).

To extract DNA sequence data, on July 21, 2006, we downloaded ten known binding sites of LexA from regulonDB (v4.0), involving nine genes each with a binding site except two binding sites for gene lexA [20]. We input either these ten binding sites or five of them (in the order of #2, #4,...,#10

Pacific Symposium on Biocomputing 13:465-476(2008)

as ranked by MEME) into MEME [1] to find a top motif (Table 1). We then used scanACE [19] to scan the whole genome with a very low threshold such that at least one subsequence matching the motif could be obtained for most genes; we assigned the maximum of all the matching scores for gene i as Z_i, the summary statistic for the sequence data. Depending on whether the ten or five known binding sites were used to obtain the top motif, the resulting sequence data were denoted as Seq 1 (S1) or Seq 2 (S2).

After combining the three data sources and deleting genes with any missing values, we obtained $G = 3779$ genes in the combined data.

3.2. *Analysis*

Table 1. Ranks given by various methods for known (marked by *) and putative targets annotated in regulonDB (and in our data). Seq1 and Seq2 were the sequence scores obtained from the top motif using the nine and four known targets (marked by * and **) respectively.

Gene	Bind	Expr	B+E	Seq1	Seq2	B+E+S1	B+E+S2
polB	156	114	135	153	1593	127	146
phrB	1346	1826	2083	530*	81**	1516	452
uvrB	48	172	92	31*	6**	78	46
dinG	96	448	213	138	143	169	171
ftsK	75	3757	223	127	303	173	199
sulA	11	12	1	17*	728	1	1
umuD	31	29	1	19	8	1	1
umuC	192	12	1	3454	3652	34	37
ydjM	30	111	53	70	74	49	44
ruvB	2780	313	509	1471	2966	645	708
ruvA	127	147	141	10	38	94	108
uvrC	3015	3104	3646	3008	796	3377	2692
uvrY	3538	3473	3679	3008	796	3384	2685
recN	7	5	1	33	36	1	1
oraA	82	50	54	1220	871	61	59
recA	12	15	1	23*	4**	1	1
rpsU	464	1214	766	1097*	304	896	572
dnaG	2906	3621	3451	782	177	2620	954
rpoD	2906	3749	3455	782	177	2621	953
t150	2121	175	262	50	76	176	178
uvrD	263	245	274	4*	50	106	160
lexA	15	61	1	7*	1**	1	1
dinF	2549	217	323	7	1	118	77
uvrA	41	169	77	14*	114	58	72
ssb	41	143	74	14*	114	54	68

We considered using binding data alone, expression data alone, sequence

Pacific Symposium on Biocomputing 13:465-476(2008)

data alone (either using the motif found from the 10 or 5 known binding sites), both binding and expression data, and all the three data sources. For each type of data, a two-component mixture model was fitted, and the posterior probabilities were used to rank the genes, as discussed earlier. To motivate the mixture model for each data source, we fitted a two-component normal mixture model to each data source separately; it appeared that the mixture model fitted each data source well (not shown). The parameter estimates $(\hat{\pi}_j, \hat{\mu}_{j1}, \hat{\mu}_{j0}, \hat{\sigma}^2_{j1}, \hat{\sigma}^2_{j0})$ with $j = 1, ..., 4$ for the four data sources were $(0.063, 0.11, 0.89, 0.07, 0.64)$, $(0.174, 0.02, 0.20, 0.16, 2.81)$, $(0.278, 12.8, 15.3, 2.5, 7.3)$ and $(0.196, 17.1, 19.7, 2.4, 7.3)$. As a comparison, the joint model with the binding, expression and Seq 1 data resulted in estimates $\hat{\pi} = 0.122$, and $(\hat{\mu}_{j1}, \hat{\mu}_{j0}, \hat{\sigma}^2_{j1}, \hat{\sigma}^2_{j0})$ with $j = 1, ..., 3$ as $(0.11, 0.51, 0.07, 0.51)$, $(0.01, 0.29, 0.21, 3.53)$, and $(13.3, 14.9, 4.0, 11.2)$ for the three types of the data respectively.

Table 1 gave the results for all known/putative targets listed in regulonDB (v5.5) [21], downloaded in November 1, 2006, and in our combined data. In general, combining multiple types of data increased the chance of detecting the true targets as compared to using only binding data alone; for example, the ranks based on binding and expression data, or based on the three data sources, were higher, in some cases much higher, than those based on using binding data alone. This was due to the fact that our joint analysis combined the evidence from all the three data sources. For example, each of umuD, recA and lexA was ranked relatively high (but not highest) based on each of the three data sources alone, and combining any two or three sources of data led to a highest ranking (i.e. tied at the 1st with posterior probability equal to 1); umuC was ranked only at the 192nd based on the binding data alone, with the incorporation of the expression data its rank improved to a tied 1st.

We also obtained the results (not shown) for the putative targets with a common motif (Class II) and without any common motif (Class III) identified based on only the binding data by Wade et al. For the genes in class III, because no common motifs were found in the DNA sequences of the genes, it was not surprising that a separate or combined use of sequence data gave the genes lower ranks than those based on the binding data alone. More surprisingly, for most genes, a combined analysis using both the binding and expression data also gave lower ranks than that based on the binding data alone, due to low level expression changes.

Pacific Symposium on Biocomputing 13:465-476(2008)

3.3. *Simulation*

To further evaluate and compare the methods with various sources of data, we did a simulation study; the simulated data were generated from the fitted models for the real data to mimic real situations. Four simulation set-ups were considered. 1) Case I: we assumed that the joint model fitted to the three data sources (with Seq 1) was correct, and simulated data from the fitted joint model; this represented an ideal scenario for the joint analysis. 2) Case II: we assumed that the binding data came from its component from the fitted joint model as in 1), but each of the other two data sources came from a two-component normal mixture model as fitted to each data source separately (Figure 1); because there were a higher proportion of the genes in the first component for the expression data and sequence data, the joint model did not hold: in particular, the second component f_{02} and f_{03} were not a single normal distribution, but a mixture of two normals. This was a scenario for which a two-component mixture model for binding data alone was correct but a joint model was not. 3) Case III was similar to Case I except that some between-gene correlations were introduced for the binding data (which might arise when the probe intensities were smoothed as in the real data). Specifically, the genes were randomly divided into blocks with size about 10, then we added some noises drawn from $N(0, \sigma_{10})$ to binding data (as generated in Case I) such that the genes within each block had correlated X_i's. Hence all the methods had an incorrect independence assumption. 4) Case IV was a combination of Cases II and III: some between-gene correlations as in Case III were introduced to the binding data while other aspects were the same as in Case II. For each case, 100 independent datasets were generated; the realized false discovery rates (FDRs) were averaged over the 100 replicates for each method in each case.

Figure 1 summarized the results for using binding data alone, using both binding and expression data, and using all three types of data. It was clear that, compared to using only binding data, using more than one data source largely reduced the FDR; that is, at any given number of estimated positives (i.e. claimed targets), the joint model could identify a much larger number of true targets (and hence fewer false negatives). Although using three data sources improved over using two data sources, because of limited information available from sequence data (as measured by the small difference between the two component distributions for the sequence data), the improvement was not dramatic.

Pacific Symposium on Biocomputing 13:465-476(2008)

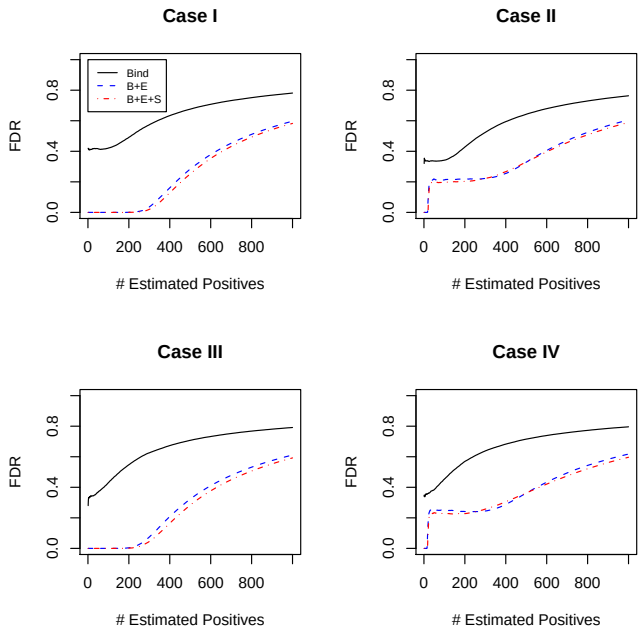

Figure 1. *Comparison of the FDRs from the three methods for simulated data.*

4. Discussion

We have demonstrated possible efficiency gains with a parametric mixture model to jointly combine multiple types of data for target detection. A key feature of our joint model is its simplicity, however, this does not exclude some possible modifications or extensions. First, rather than using a single normal distribution f_{jk} for each component for each data source, a more flexible choice is to use a mixture distribution for each f_{jk}; for the E coli data here, we considered this idea for binding data but it did not lead to much improvement, perhaps due to the goodness-of-fit of a single normal distribution to each component. We emphasize that, with some appropriate transformation, such as the Z-transformation [7], the normality of some component distributions is expected; furthermore, McLachlan et al [16] demonstrated that a two-component normal mixture model worked quite well for several typical expression datasets. Second, rather than using a sequence score for each gene, we may supply the sequence best matching the motif, and use a multinomial model for each component of the sequence data f_{3k}, as done in Wang et al [33]. This could possibly help refine the motif

Pacific Symposium on Biocomputing 13:465-476(2008)

model. We may also consider using multiple motifs for lexA and their multiple matching sequences for each gene. Third, an advantage of the mixture model is to use the estimated posterior probabilities to estimate FDR and false non-discovery rate (FNR) [16], However, such a use critically depends on the correctness of the assumed mixture model [18]. Because here we aim to use a simple parametric model which may or may not hold exactly, we did not pursue the task of estimating FDR or FNR, which however is important in practice. To relax the possibly too strong parametric assumption, we may consider the use of a more flexible mixture model approach as outlined above, alleviating the issue of FDR/FNR estimates' dependence on correct modeling. These are all interesting topics for future research.

Acknowledgments

This research was supported by NIH grants HL65462 (WP and PW) and GM066098 (AK), and a UM AHC Faculty development grant (WP, PW and AK).

References

1. Bailey T.L. and Elkan C. (1995). *Machine Learning*, **21**, 51-80.
2. Bar-Joseph, Z., gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. and Gifford, D.K. (2003). *Nature Biotechnology*, **21**, 1337-1342.
3. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001). *Nat. Genet.* **27**, 167-171.
4. Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003). *Proc. Natl. Acad. Sci. USA* **100**, 3339-3344.
5. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC (2001). *Genetics*, **158**, 41-64.
6. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). *J. R. Statist. Soc. B,* **39**, 1-38.
7. Efron, B. (2004). *J. Amer. Statist. Assoc.,* **99**, 96-104.
8. Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V.G. (2001). *J. Amer. Statist. Assoc.,* **96**, 1151-1160.
9. Holmes, I. and Bruno, W.J. (2000). *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 202-210.
10. Ihaka, R. and Gentleman, R. (1996). *Journal of Computational and Graphical Statistics* **5**, 299-314.
11. Lee, M.-L., Bulyk, M., Whitmore, G., and Church, G. (2002). *Biometrics*, **58**, 981-988.
12. Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., et al. (2002b). *Science*, **298**, 799-804.

Pacific Symposium on Biocomputing 13:465-476(2008)

13. Liu JS, Neuwald AF, and Lawrence CE (1999). *J. Amer. Statist. Assoc.* **94**, 1-15.
14. Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002). *Nat. Biotechnol.* **20**, 835-839.
15. McLachlan, G.J. and Peel, D. (2002) Finite mixture model. New York. John Wiley & Sons, Inc.
16. McLachlan, G.J., Bean, R.W., Jones, L.B.-T. (2006). *Bioinformatics* **22**, 1608-1615.
17. Newton, M.A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). *Biostatistics*, **5**, 155-176.
18. Pan W, Jeong KS, Xie Y, Khodursky A. (2006). To appear *Statistica Sinica*.
19. Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). *Nat. Biotech.*, **16**, 939-945.
20. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J. (2004). *Nucleic Acids Res.*, **32**, D303-306.
21. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. (2006). *Nucleic Acids Res.*, **34**, D394-D397.
22. Storey, J.D. and Tibshirani, R. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 9440-9445.
23. Sun N, Carroll RJ, Zhao H. (2006). *Proc Natl Acad Sci USA*, **103**, 7988-7993.
24. von Mering, C., Krause, R., Snel, B. et al. (2002). *Nature* **417**, 399-403.
25. Wade JT, Reppas NB, Church GM, Struhl K. (2005). *Genes and Development*, **19**, 2619-2630.
26. Wang, W., Cherry, J.M., Nochomovitz, Y., Jolly, E., Botstein, D. and Li, H. (2005). *Proc. Nat. Acad. Sci. USA,* **102**, 1998-2003.
27. Xiao G and Pan W. (2005). *Journal of Bioinformatics and Computational Biology* **3**, 1371-1389.
28. Xie, Y. (2006) Statistical analysis for microarray data: false discovery rate estimation, statistical testing and integrated analysis. PhD dissertation, University of Minnesota, Minneapolis, MN, USA.
29. Xie, Y., Pan, W., Jeong, K.S., Khodursky, A. (2007). *Statistics in Medicine* **26**, 2258-2275.
30. Yang, Y.H., Dudoit, S., Luu, P., and Speed, T. (2002). *Nucleic Acids Research* **30**, e15.
31. Zhao, H., Wu, B. and Sun, N. (2003). In Goldstein, D.R. (ed.), *Science and Statistics: A Festschrift for Terry Speed*, IMS Lecture Notes-Monograph Series, Vol. 40, p. 259-274.
32. Hong, P., Liu, X.S., Zhou, Q., Lu, X., Liu, J.S. and Wong, W.H. (2005). *Bioinformatics* **21**, 2636-2643.
33. Wang, W., Cherry, J.M., Botstein, D. and Li, H. (2002). *Proc. Nat. Acad. Sci. USA,* **99**, 16893-16898.