BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 14 2009, pages 1715–1721
doi:10.1093/bioinformatics/btp312

Genome analysis

Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data

Hyungwon Choi[1], Alexey I. Nesvizhskii[1,2], Debashis Ghosh[3,*] and Zhaohui S. Qin[2,4,*]

[1]Department of Pathology, [2]Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, [3]Departments of Statistics and Public Health Sciences, Penn State University, University Park, PA 16802 and [4]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Received on January 14, 2009; revised on April 15, 2009; accepted on May 6, 2009

Advance Access publication May 14, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Chromatin immunoprecipitation (ChIP) experiments followed by array hybridization, or ChIP-chip, is a powerful approach for identifying transcription factor binding sites (TFBS) and has been widely used. Recently, massively parallel sequencing coupled with ChIP experiments (ChIP-seq) has been increasingly used as an alternative to ChIP-chip, offering cost-effective genome-wide coverage and resolution up to a single base pair. For many well-studied TFs, both ChIP-seq and ChIP-chip experiments have been applied and their data are publicly available. Previous analyses have revealed substantial technology-specific binding signals despite strong correlation between the two sets of results. Therefore, it is of interest to see whether the two data sources can be combined to enhance the detection of TFBS.

Results: In this work, hierarchical hidden Markov model (HHMM) is proposed for combining data from ChIP-seq and ChIP-chip. In HHMM, inference results from individual HMMs in ChIP-seq and ChIP-chip experiments are summarized by a higher level HMM. Simulation studies show the advantage of HHMM when data from both technologies co-exist. Analysis of two well-studied TFs, NRSF and CCCTC-binding factor (CTCF), also suggests that HHMM yields improved TFBS identification in comparison to analyses using individual data sources or a simple merger of the two.

Availability: Source code for the software ChIPmeta is freely available for download at http://www.umich.edu/∼hwchoi/HHMMsoftware.zip, implemented in C and supported on linux.

Contact: ghoshd@psu.edu; qin@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Chromatin immunoprecipitation (ChIP) is a powerful method for isolating a transcription factor (TF) bound to DNA sequences *in vivo* (Orlando and Paro, 1993; Solomon *et al.*, 1988). In ChIP experiments, cells are first treated with reagents such as formaldehyde inducing protein–DNA crosslinks, and DNA is isolated and fragmented afterwards. An antibody specific to a TF is added to precipitate the interacting pairs, and their crosslinks are reversed. The resulting DNA fragments are direct evidence for physical interactions between the TF and its target genes. These DNA segments can be simultaneously mapped to the genome with array-based hybridization, known as ChIP-chip (Iyer *et al.*, 2001; Ren *et al.*, 2000). ChIP-chip has been widely used for identifying TF binding sites (TFBS). Recently, ChIP experiment coupled with massively parallel sequencing (Bentley *et al.*, 2008), or ChIP-seq, has been proposed as an alternative (Barski *et al.*, 2007; Chen *et al.*, 2008; Johnson *et al.*, 2007; Mikkelsen *et al.*, 2007; Robertson *et al.*, 2007; Schones *et al.*, 2008; Shivaswamy *et al.*, 2008). ChIP-seq offers genome-wide coverage in a single base pair resolution at low cost (Park, 2008).

Although a number of previous studies have demonstrated the power of ChIP-seq, it has also been shown that different mapping strategies may identify mutually exclusive peak regions as candidate binding sites. For instance, Robertson *et al.* (2007) reported that the overlap between the ChIP-enriched regions identified by ChIP-seq and ChIP-chip is around 60% in signal transducer and activator of transcription protein 1 (STAT1) data. Euskirchen *et al.* (2007) found that ChIP-chip and ChIP-PET (Loh *et al.*, 2006; Wei *et al.*, 2006), a sequencing-based method, are frequently complementary to each other in identifying validated targets when the signal is not sufficiently strong. The evidence by Robertson *et al.* (2007) suggests that massively parallel sequencing may not work well for all DNA fragments uniformly. For example, the sequencing can be biased toward certain parts of the genome due to the complex chromatin structure of DNA molecules in their native form. Also, sequence reads may also have reduced sensitivity in the genomic regions where repeat sequences appear frequently. For those DNA fragments, other mapping methods not relying on direct sequencing, e.g. ChIP-chip, can be a valuable source to complement the weakness of the sequencing technology.

For many of the existing ChIP-seq data, ChIP-chip experiments have also been conducted and the data are publicly available. It is desirable to take advantage of existing ChIP-chip datasets to assist TFBS identification using ChIP-seq. While such a joint analysis has a promise, it is a challenging task to account for the heterogeneity of data from the ChIP-chip and ChIP-seq platforms. This is because the two technologies show vastly different behavior in terms of sensitivity and specificity. Specifically, the peaks identified by ChIP-seq are expected to form regions that are much sharper than those in ChIP-chip due to its superior resolution, whereas ChIP-chip tends to report broader regions with moderate significance including

*To whom corresponding should be addressed.

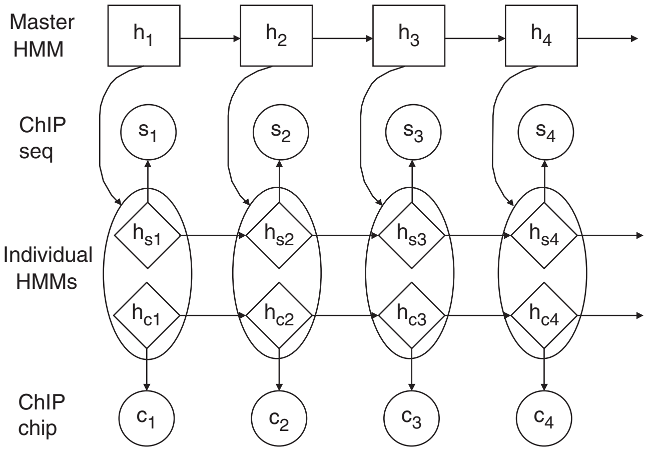

Fig. 1. HHMM framework with the master process in the top layer and the multiple individual processes in the bottom layer. The hidden states in ChIP-seq and ChIP-chip data are considered as emission from the master process.

potential false positives. Hence, the signals from the two data sources have to be appropriately weighted in order to keep the overall false positive rates low in the joint analysis.

To this end, hierarchical hidden Markov model (HHMM), a collection of multiple individual-level HMMs governed by a population or master-level HMM, is developed in this work. HMMs have been frequently used to analyze ChIP-chip data in the literature (Du *et al.*, 2006; Huber *et al.*, 2006; Humburg *et al.*, 2008; Ji and Wong, 2005; Li *et al.*, 2005; Munch *et al.*, 2006). The structure of the HHMM model is illustrated in Figure 1. In this method, individual-level HMMs function as de-noising filters that convert the raw data into inferred binary hidden states representing ChIP-enrichment and background noise, and the master-level HMM uses individual-level hidden states as a basis to infer the underlying *true* states. In this process, individual-level HMMs serve as a buffer to reduce the heterogeneity present in raw ChIP-chip and ChIP-seq data, and the master-level HMM summarizes their ChIP-enrichment status to produce the final probability score.

Development of HHMMs has been proposed previously in the literature (e.g. Bui *et al.* 2004; Fine *et al.* 1998). Recently, Shah *et al.* (2007) used this class of models for accurately detecting boundary points of copy number changes across multiple samples in genome-wide array-comparative genomic hybridization (aCGH) data. In their model, hidden states in the individual samples exchange mutual feedback with the hidden state in the master level. In contrast, for our problem, each data source is represented as an individual HMM, whose inferred hidden states are then modeled as the bivariate emission probabilities of the master-level HMM.

2 METHODS

2.1 Data

Data generated from ChIP-chip and ChIP-seq experiments are different. ChIP-chip data are fluorescent intensity levels from microarrays reflecting the amount of DNA fragments hybridized to the probes. Probes on tiling arrays are usually 36–50 nt long. Elevated intensity levels from multiple adjacent probes indicate ChIP-enrichment. In contrast, ChIP-seq data are sequencing reads that map to the reference genome. Reads piled up at a tight neighborhood indicate ChIP-enrichment.

Because a HMM framework was adopted, the data are first summarized into fragment counts in units of windows of fixed size (25 nt in this study and adjustable) along the genome. Dissecting chromosomes into windows of equal length has been used previously in the ChIP-seq literature (Mikkelsen *et al.*, 2007). Since the start and end positions of ChIP-chip probes do not exactly overlap with these windows, ChIP-chip probe boundaries were redefined so as to match the ChIP-seq windows (later in the master-level HMM). With typical probes having length >25 nt, one ChIP-chip probe can be mapped to multiple windows. The impact of varying window sizes is further discussed in Section 3.

2.2 HHMM

Let the ChIP-seq count data and the ChIP-chip intensity data be denoted by $S=(s_1,...,s_T)$ and $C=(c_1,...,c_T)$, respectively, for a chromosome that has been divided into T windows. We assume that the number of windows is identical in the two data. It is assumed that each data source follows its own independent HMM. Their respective hidden states are denoted by $h_s=(h_{s1},...,h_{sT})$ and $h_c=(h_{c1},...,h_{cT})$. As shown in Figure 1, these hidden states are modeled as bivariate random variables in the emission of master HMM, whose hidden states are denoted by $h=(h_1,...,h_T)$. Both the individual-level states (h_s,h_c) and the master-level states h consist of either ChIP enriched (denoted 1) or background (denoted 0) states. Note that the ultimate goal of HHMM is to infer the master-level hidden states h.

The model parameters are now specified in the individual level first. The three main components of HMM—initial state distribution, transition probability matrix and emission (Rabiner, 1989)—are defined. The initial state distribution $\pi(h_{s1})$ and $\pi(h_{c1})$ and the transition probabilities A^s and A^c can either be fixed or estimated from the data. Each matrix has two rows and two columns, with probability moving from one state in the row to another state in the column. In the latter case, one can assume that each row of A^s and A^c follows multinomial distribution and estimate the probabilities from the frequency of relevant moves in the inference of h_s and h_c, respectively. Parametric models are used to describe emission probabilities in the ChIP-enriched states and the background states.

In the following, we briefly describe the two individual-level HMMs proposed to model ChIP-chip and ChIP-seq data, respectively. More details of the two models can be found in the Supplementary Material. We will then explain in details about the master-level HMM which is the main focus of this study. Note that the individual-level HMMs proposed here can be replaced by alternative methods to infer ChIP-enrichment in the individual data sources.

2.2.1 HMM in ChIP-chip In ChIP-chip data, we use uniform and normal distributions to model the hybridization intensities in the ChIP-enriched states and the background states, respectively. The uniform–normal mixture model has been previously used to model differential gene expression in microarray data analysis (Parmigiani *et al.*, 2002). While the assumption of uniform distribution for signals in the ChIP-enriched states may be an over-simplification due to possible variation in TF binding affinity across the genome, this assumption alleviates the computational loading of HHMM when no prior knowledge is given as to the preferential binding site enrichment. A possible replacement for the distribution in the ChIP-enriched states is another normal distribution (Li *et al.*, 2005).

In the uniform–normal model, $C_t|h_{ct}=1\sim\mathcal{U}_{\theta_{c1}}(\cdot)$ and $C_t|h_{ct}=0\sim\mathcal{N}_{\theta_{c0}}(\cdot)$, where $\mathcal{U}$ and $\mathcal{N}$ denote the uniform and normal distributions in the ChIP-enriched states and the background states, respectively. The uniform distribution parameters θ_{c1} are fixed as the minimum and maximum of intensities $\{C_t\}_{t=1}^T$, and the mean and variance parameters of the normal distribution $\theta_{c0}=(\mu_c,\sigma_c^2)$ will be estimated. Bayesian inference for HMM was implemented with conjugate priors and efficient sampling algorithm was presented in Scott (2002) (See Supplementary Material Section 1.1 for details).

2.2.2 HMM in ChIP-seq In ChIP-seq data, overdispersion and higher proportion of zero counts must be accounted for in the model. We assume single sample analysis for now. We use generalized Poisson (GP) distribution and zero-inflated Poisson (ZIP) distribution to model read counts in the ChIP-enriched states and the background states, respectively (Consul, 1989; Johnson *et al.*, 1992). For inference purposes, a latent variable Z_t is defined for sequence count s_t at location t such that $Z_t = 0$ if $s_t = 0$ and s_t is generated from the point mass at zero, and $Z_t = 1$ otherwise (thus, it is always the case $Z_t = 1$ if $s_t > 0$). As in the model for ChIP-chip data, Bayesian inference was implemented (See Supplementary Material Section 1.2 for details).

The same model can be extended to two sample analysis (ChIP-treated sample and untreated control sample). In the paired design, we used a bivariate GP/ZIP distribution to model the read counts between the two types of samples. A HMM is then designed to perform inference on the ChIP-enriched/background states (See the Supplementary Material Section 1.2 for details).

2.2.3 Master HMM In the master level, the initial state distribution $\pi(h_1)$ and transition probabilities A are defined the same way as in the individual level. For the emission, the data (h_s, h_c) are modeled with two multinomial distributions, i.e. $(h_{st}, h_{ct})|h_t = 1 \sim \mathcal{M}_{\theta_1}(\cdot)$ and $(h_{st}, h_{ct})|h_t = 0 \sim \mathcal{M}_{\theta_0}(\cdot)$, where the distribution for the enriched state $\mathcal{M}$ denotes multinomial distribution, and $\theta_1 = (p^1_{00}, p^1_{01}, p^1_{10}, p^1_{11})$ and $\theta_0 = (p^0_{00}, p^0_{01}, p^0_{10}, p^0_{11})$ are their parameters for the ChIP-enriched states and the background states, respectively. These parameters are given a conjugate Dirichlet prior with parameters $(\gamma^1_{00}, \gamma^1_{01}, \gamma^1_{10}, \gamma^1_{11})$ and $(\gamma^0_{00}, \gamma^0_{01}, \gamma^0_{10}, \gamma^0_{11})$, respectively.

Given the posterior probability pairs (q_{st}, q_{ct}) at all positions $t = 1, \ldots, T$ estimated in the individual-level HMMs, hidden states in the master level are inferred as follows. Had (h_{st}, h_{ct}) been observed directly, the likelihood for the master-level HMM would be

$$\pi(h_1)\prod_{t=2}^{T}\pi(h_{st}, h_{ct}|h_t, \theta_{h_t})\pi(h_t, h_{t-1}, A)$$

From the inference of individual HMM, $\{(q_{st}, q_{ct})\}_{t=1}^{T}$ are computed, but the actual hidden states $\{(h_{st}, h_{ct})\}_{t=1}^{T}$ remain still unknown. Treating this as a missing data problem, the likelihood is integrated over all four possibilities of (h_{st}, h_{ct}) based on the marginal weights (q_{st}, q_{ct}), i.e.

$$\pi(h_1)\prod_{t=2}^{T}\left[\sum_{(h_{st}, h_{ct})} g_t \cdot \pi(h_{st}, h_{ct}|h_t, \theta_{h_t})\right]\pi(h_t, h_{t-1}, A)$$

where $g_t = (q_{st})^{h_{st}}(1-q_{st})^{(1-h_{st})}(q_{ct})^{h_{ct}}(1-q_{ct})^{(1-h_{ct})}$. This multiplicative factor g_t weights the four possible cases of (h_{st}, h_{ct}) based on the product of their corresponding marginal posterior probabilities in ChIP-seq and ChIP-chip at position t, as an approximate solution to the missing data problem.

With this likelihood, imputation and posterior sampling steps are iterated as in the ChIP-chip case: (i) Imputation: draw $h^{(i+1)} \sim \pi(h|h_s, h_c, \theta_0, \theta_1, A)$ using the forward–backward algorithm, and (ii) Posterior sampling: draw $\theta_j^{(i+1)} \sim \pi(\theta_j|h_s, h_c, h^{(i+1)}, A)$ for $j = 0, 1$. With the multinomial likelihood and the Dirichlet prior, the posterior is again Dirichlet distribution, thus $\theta_j = (p^j_{00}, p^j_{01}, p^j_{10}, p^j_{11})$ are drawn from $\mathcal{D}(\gamma^j_{00} + H^j_{00}, \gamma^j_{01} + H^j_{01}, \gamma^j_{10} + H^j_{10}, \gamma^j_{11} + H^j_{11})$ where $H^j_{kl} = \sum_t 1\{h_{st} = k, h_{ct} = l, h_t = j\}$ for $k, l = 0, 1$ and $j = 0, 1$.

Prior was elicited to reflect the known technological difference between ChIP-seq versus ChIP-chip in terms of precision and sensitivity. Since ChIP-seq signals tend to be more sparse but more precise than ChIP-chip signals, elicitation of informative prior that elevates the confidence for ChIP-seq signals more than ChIP-chip signals was preferred for the master-level HMM. In fact, there are ways to conjecture the optimal posterior distribution in real data. For example, if one is aware of the false positive rates in ChIP-seq and ChIP-chip, then the posterior can be set so that the ratio p^1_{10}/p^1_{01} is inversely proportional to the ratio of false positives. One can also learn this knowledge from preliminary motif search in TFBS identified in ChIP-seq and ChIP-chip and reflect the sensitivity ratios in p^1_{10}/p^1_{01} and p^1_{11}/p^1_{10}. Through multiple simulations and real data analysis, it was found that the following prior works well: $\gamma^1_{11} = M/2$, $\gamma^1_{10} = M/5$ and $\gamma^1_{01} = M/10$ in the ChIP-enriched windows, and $\gamma^0_{kl} = 1$ in the background windows.

The elicited prior results in the posterior probability ratios $1 < r_{01} < r_{10} < r_{11}$ where $r_{kl} = p^1_{kl}/p^0_{kl}$. This requirement is important since the noise in the ChIP-chip data will substantially increase the number of windows with unique ChIP-chip signals H^1_{01} assigned to the ChIP-enriched state, and as a result the posterior probability of ChIP-enrichment for these windows will be overestimated unless informative prior is specified. Admission of ChIP-chip unique signals with higher frequency than ChIP-seq unique signals is likely to result in elevated false positive rates.

2.2.4 Regions with missing data Due to the limitations in technology and the presence of repetitive regions, neither ChIP-seq nor ChIP-chip is able to completely survey all bases of the human genome. Regions that are inaccessible from both are marked and skipped. There are also regions on the genome that are uniquely accessible by either technology. When data from one source is missing, the inference of the hidden states at the upper level in HHMM will rely on the other data source alone. That is, using the marginal distribution (Bernoulli) of the joint distribution to model the observed (non-missing) data.

3 RESULTS

3.1 Simulation study

A simulation study was conducted in order to assess the performance of HHMM. The posterior probabilities were generated instead of the raw signals, as the focus of this simulation study is the assessment of master-level HMM, where the information from both data sources are combined.

First, the master-level hidden states h in a chromosome containing a 100 000 probes (25 Mb chromosome) were simulated from a stationary Markov chain with a transition probability matrix

$$A = \begin{pmatrix} 0.99 & 0.01 \\ 0.15 & 0.85 \end{pmatrix}$$

Hidden state 1 denotes ChIP-enrichment. ChIP-enriched states have been accepted only when the probes formed a contiguous block, i.e. all 'singletons' in the ChIP-enriched state have been converted to the background state. This generates the baseline 'truth' where the true ChIP-enriched sites are 150 bp long on average (range from 75 bp to 1375 bp and IQR of 100–250 bp).

Given a value of hidden state $h_t = 1$ at each locus t, posterior probabilities $P(h_{st} = 1|S)$ and $P(h_{ct} = 1|C)$ have been generated from Beta distributions with mean 0.9 and 0.8, respectively. In order to reflect higher resolution in ChIP-seq over ChIP-chip, data were generated so that each true ChIP-enriched region is almost exactly covered by a ChIP-seq peak region with ChIP-chip signals surrounding it. Negative signals ($h_t = 0$) have been placed as follows. Reflecting the actual false positive rates of <5% in ChIP-seq and 25% in ChIP-chip previously reported in analyses of real datasets, e.g. Robertson *et al.* (2007), these false positive signals were planted in blocks of 3–8 windows with probability 0.05 and 0.25 in the two datasets, respectively.

Datasets with four possible sampling behaviors (p_s, p_c) have been simulated. Sampling behavior here refers to the sensitivity of each data source producing signal within the true ChIP-enriched regions. Case I $(p_s = 0.75, p_c = 0.9)$ and Case II $(p_s = 0.6, p_c = 0.8)$ represent scenarios where ChIP-chip signals appear with a greater frequency (with a greater error rate) than ChIP-seq, which may represent the

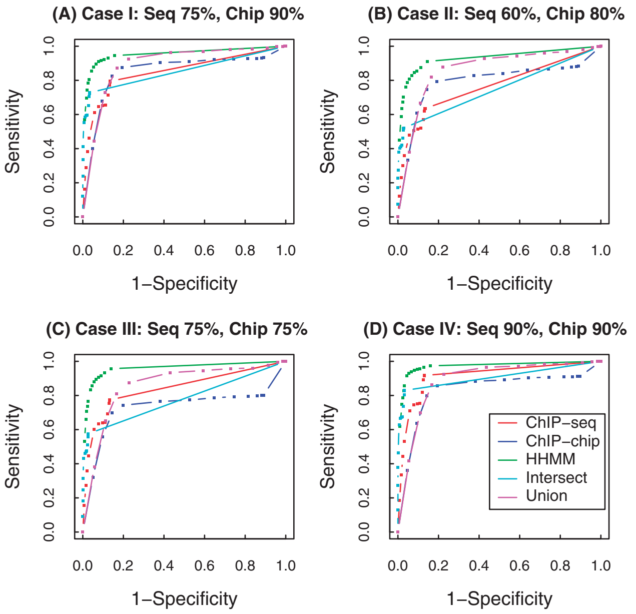

Fig. 2. Plots of the ROC in the four simulation datasets, comparing ChIP-seq only, ChIP-chip only, HHMM, Intersection and Union. Four different settings of ChIP-seq and ChIP-chip data were generated. Signal present in 75% and 90% (**A**); 60% and 80% (**B**);75% and 75% (**C**); and 90% and 90% (**D**) of the ChIP-enriched regions detected by ChIP-seq and ChIP-chip, respectively.

cases where the sequencing depth is low and hence a number of real ChIP-enriched regions are missed by ChIP-seq. Case III ($p_s = 0.75, p_c = 0.75$) and Case IV ($p_s = 0.9, p_c = 0.9$) represent scenarios where both data sources cover real binding site motif regions with a good sensitivity and some of the platform-specific regions host a good number of real motifs. Other scenarios of a varying range of combinations with the fixed p_s/p_c ratio have also been simulated, and the results were consistent.

HHMM was compared with four other ways to identify ChIP-enriched regions with high probability: (i) ChIP-seq only: peak regions from ChIP-seq HMM; (ii) ChIP-chip only: peak regions from ChIP-chip HMM; (iii) Intersection: common peak regions in both sources; and (iv) Union: peak regions from either source. Figure 2 shows the receiver operating characteristic (ROC). In all examples, HHMM is the best performing method in terms of sensitivity followed by Union, outperforming both single-source analyses. More importantly, HHMM keeps the specificity higher than Union for nearly all decision points (square dots). The fact that the ROC curve bent to the right significantly for high specificity decision points in the Union indicates that blind picking of all signals would result in high false positive rates at a fixed specificity, mostly due to ChIP-chip data. HHMM removes most of the low-key negative signals, which can be seen in the upper left corner of the ROC curves.

In sum, the results in Cases I through IV indicate that the area under the curve of the ROC is the highest in HHMM followed by Union and the number of positive calls is almost always highest in HHMM at fixed specificity. In all scenarios where either the more advanced mapping platform misses some of the true signals or both platforms complement the identification for each other, HHMM has the potential to collect the highest number of binding sites and, at the same time, to keep the false discovery rates lower than blind picking of all signals.

Meanwhile, another dataset was simulated with $(p_s, p_c) = (0.8, 0.6)$, where the better performing platform ChIP-seq covers most of the signals picked up by ChIP-chip. Examination of ROC curve shows that HHMM, ChIP-seq only and Union methods perform equivalent to one another, indicating that there is no additional benefit earned by HHMM as expected. Also, this is consistent with the fact that the ROC improved the least in Case IV out of the four scenarios, where the number of overlapping signals in ChIP-seq and ChIP-chip is the largest among all.

3.2 Application to NRSF data

3.2.1 Data HHMM was applied to a real dataset for the TF NRSF (Johnson *et al.*, 2007). In the study, ChIP-seq was used to study genome-wide mapping of binding sites of NRSF, a neuron-restrictive silencer factor known for its negative regulation of many neuronal genes in non-neuronal cells (Schoenherr *et al.*, 1996), were mapped to ∼2000 locations in the human genome using ChIP-seq. ChIP-seq data for the treated and control cell lines were available from the Illumina web site and an unpublished ChIP-chip data was also available in Gene Expression Omnibus (GSE7372). Since the array platform used in the ChIP-chip data (Nimblegen ENCODE tiling arrays) does not cover the whole genome, this section focuses on the 10 ENCODE regions each spanning 5 million bp, i.e. around 1% of the human genome.

High probability signals (0.9 and above) appeared in 26.5 and 272.9 kb in ChIP-seq and ChIP-chip data, respectively, indicating significant differences between the two platforms. Among these, 422 windows were overlapping, which accounts for 37% of ChIP-seq. The posterior probabilities were then combined into a single data as mentioned previously, and master-level HMM was fit.

3.2.2 Motif enrichment For the peak regions identified by each method, we used two motif search engines to find TFBS. First, we used MatInspector (Cartharius *et al.*, 2005) of Genomatix (http:// www.genomatix.de) using the position weight matrix (PWM) reported in Schoenherr *et al.* (1996) (See Supplementary Material Section 2 for motif logo plot). Since the sequence alignment in MatInspector does not offer confidence assessment, we generated permuted sequences by randomly shuffling the nucleotides within each sequence in the peak regions and the motif search was reiterated, providing a reference for assessing the significance of the hits. The enrichment of TFBS motifs was tested by χ^2 test in a contingency table setting. The rows of the 2×2 table indicate whether the motif search was done in the original sequence or in the permuted sequence, and the columns indicate whether the sequences contain motifs or not Frith *et al*. (2004). Second, we applied Clover to construct more realistic background sequences from a specific composition of real sequences (human chromosome 20). We used the same PWM used above for sequence alignment. Clover reports the motifs that are statistically significant matches compared with matches to background sequences at a certain significance threshold (e.g. $P = 0.05$).

Table 1 shows the result of the analysis using the probability threshold 0.9. The total number of motifs is the highest at 67 in the Union method, but the HHMM picks up 46 motifs while keeping

Table 1. Motif summary for the five methods in NRSF data

Method	#Match[a] (#Permute[b])	#Peaks	Coverage (kb)	OR[c]	χ^2	Match rate[d]	#Clover[e]	Rate in Clover[f]
HHMM	46 (11)	424	179.2	4.56	21.74	0.19	51	0.28
Union	67 (24)	860	293.0	2.94	20.47	0.15	64	0.22
ChIP-seq	25 (4)	61	26.5	9.89	18.09	0.79	37	1.39
ChIP-chip	52 (17)	830	272.9	3.20	17.48	0.13	40	0.15
Intersect	10 (1)	25	6.6	16.00	7.46	1.36	13	1.97

Regions containing at least one peak with probability 0.9 were selected and motifs with PWM of NRSF motif were searched.

[a] #Match is the number of peaks including a motif match in the original sequence.

[b] #Permute is the number of peaks including a motif match in the permuted sequence.

[c] OR stands for odds ratio.

[d] Match rate is (#Match − #Permute)/1 kb.

[e] # Clover corresponds to the number of statistically significant motif hits in the selected peak regions at *P*-value threshold 0.05.

[f] Rate in Clover refers to the # motif hits divided by the sequence coverage.

the false positives less than half of the Union, indicating improved control of false positive rates, at the expense of fewer low ranking signals. This is congruent with the Clover analysis shown in the last two columns of Table 1, where the absolute number of motifs is the highest in the Union but the identification rate (per 1 kb) is the highest in HHMM (using the default *P*-value threshold 0.05).

3.2.3 Comparison of individual HMMs Since the performance of HHMM depends on the success of individual HMMs, we evaluated the individual-level HMMs in ChIP-seq and ChIP-chip data in comparison to cisGenome (Ji *et al.*, 2008) in the NRSF data. CisGenome is an integrative system for detecting ChIP-enriched regions from ChIP-seq or ChIP-chip data. The same datasets used in the HHMM analysis were fed into CisGenome. All default arguments were used in CisGenome. The significance criterion was set at the posterior probability threshold 0.9 (default in HHMM) for HMMs in HHMM and FDR 0.1 for CisGenome (default in CisGenome).

The comparison in the ChIP-chip data reveals that a considerable overlap exists between the two methods. Our uniform–normal model and CisGenome identified 697 and 830 peaks, respectively, with 530 peaks (76%) identified in CisGenome in the peaks identified by our uniform–normal HMM. The peaks unique to each method are likely due to the distinct approach taken in TileMap (Ji and Wong, 2005) implemented in CisGenome. In the ChIP-seq data, GP-ZIP HMM identified 61 peaks with probability threshold 0.9 whereas CisGenome identified 37 peaks with FDR threshold 0.1. Thirty-five peaks from CisGenome (95%) have overlap with at least one region in the selected peak regions in our GP-ZIP HMM.

We also found that, for the ChIP-chip data, the peaks identified by CisGenome tend to be longer than the peaks identified by our uniform–normal HMM (589 bp versus 328 bp) on average, while, for the ChIP-seq data, peaks identified by CisGenome tend to be narrower than peaks identfied by our GP-ZIP HMM (136 bp versus 435 bp).

Despite the differences in the width of peak regions, both comparisons suggest that the individual-level HMMs in HHMM perform reasonably in concordance with other implementations such as CisGenome.

3.2.4 Impact of window size We evaluated the impact of the window size in HHMM using 10 and 50 nt long windows in addition to the 25 nt window used above. In terms of sequence overlap, 25 and 50 nt windows gave consistent result (see Table 2). Peaks identified using 10 nt windows but missed using the other two larger windows from 3 to 8 windows (30 to 80 nt), indicating that they are short peak regions.

Table 2. Impact of window sizes in HHMM analysis

Window size (nt)	Peaks	Coverage (kb)	Overlap with		
			10 nt	25 nt	50 nt
10	1195	350.9	–	–	–
25	424	179.2	97%	–	–
50	405	194.4	98%	91%	–

Finer window size helps to identify more motifs at the expense of extended computational loading.

This inconsistent result with 10 nt window suggests that using microscopic windows may exaggerate short peaks, let alone increased computational loading. Based on the consistency in the analyses using 25 and 50 nt windows, windows of comparable size to sequence reads (20–30 nt) or array probes (36–50 nt) is deemed optimal for HHMM analysis.

3.3 Application to CTCF data

3.3.1 Data and model fit For an example of genome-wide mapping, binding sites of CTCF were mapped using the data from Kim *et al.* (2007) and Barski *et al.* (2007). CTCF is a zinc finger protein that has a multivalent character as a TF (Dunn and Davie, 2003; Ohlsson *et al.*, 2001) capable of participating in both repression and activation due to the combinatorial use of its 11 zinc fingers. CTCF zinc fingers can be selectively utilized based on the different needs of target genes, and thus the binding sites are likely to be more variable than other transcription factors. For instance, Kim *et al.* (2007) has reported 62 genes for which multiple CTCF binding sites were identified. We used the PWM reported in that study for motif search (See Supplementary Material Section 2 for the motif logo plot).

Individual HMM fits in this data showed that 419 457 windows in ChIP-seq and 3.4 million probes (7.1 million windows worth) in ChIP-chip had positive posterior probabilities, where 152 025

Table 3. Motif summary for the five methods in CTCF data

Method	#Match[a] (#Permute[b])	#Peaks	Coverage (Mb)	OR[c]	χ^2	Match rate[d]
HHMM	23 772 (4815)	65 808	30.31	7.16	16 057.36	0.63
Union	26 788 (6200)	83 325	40.08	5.89	16 018.71	0.51
ChIP-seq	16 771 (1836)	25 372	9.33	25.00	18 926.85	1.60
ChIP-chip	16 599 (5134)	69 246	33.83	3.94	7172.77	0.34
Intersect	6310 (719)	9576	3.06	23.80	7023.18	1.83

Peak regions that contains a signal with probability 0.9 were selected and motifs with PWM of CTCF motif were searched.
[a]#Match is the number of peaks including a motif match in the original sequence.
[b]#Permute is the number of peaks including a motif match in the permuted sequence.
[c]OR stands for odds ratio.
[d]Match rate is (#Match − #Permute)/1 kb.

windows overlapped with each other (37% of ChIP-seq). Among these, 1.5 million windows had posterior probabilities 0.9 in at least one data source, and 1.2 million of these had 0.9 and above probability in the master level.

3.3.2 Motif enrichment Table 3 presents the motif enrichment test results in MatInspector based on the data filtered at the probability threshold 0.9. For background sequences, we randomly shuffled the original sequences within each peak region as in the NRSF data. It is easy to see that HHMM and Union are the two methods that collect the highest number of TFBS motifs, but the number of hits in the permuted sequences shows almost a 3:4 ratio, indicating that the relative significance of motif search results is improved in HHMM.

Since the number of hits in a genome-wide data is extremely large, all χ^2 tests reported extremely small *P*-values. However, the odds ratio of observing motifs in the selected regions was higher in HHMM (7.16) than in Union (5.89), and the match rate was also higher in HHMM (0.98/1 kb) than in Union (0.84/1 kb). This improvement is an obvious consequence of the fact that the regions picked by HHMM (30 Mb) is far narrower than those picked by Union (40 Mb) on average.

On the other hand, ChIP-seq data from Barski *et al.* (2007) seem to demonstrate the ultra-performance of ChIP-seq, where 62% of the motifs found in Union were identified, but the search regions are so specific that the number of hits in the permuted sequences is low (1836 in ChIP-seq, 6200 in Union) and therefore the odds ratio and the match rates are high. Nonetheless, it is the goal of HHMM to find a compromise between Union and ChIP-seq only analysis, in which extra 7000 motifs were saved by allowing some of the most significant ChIP-chip-specific regions at the expense of a reduced overall statistical significance of motif enrichment.

4 DISCUSSION

The availability of multiple experimental datasets profiling the activity of a specific TF is an important asset for delineating regulatory mechanisms. The proposed HHMM method not only identifies more binding sites with increased specificity, but also serves as an assessment of agreement and discrepancy between both technologies. It is noted that HHMM may not be optimal when the best performing experimental platform (ChIP-seq in this case) identifies most of the true ChIP-enriched regions, since additional information with a decreased precision will do nothing but dilute the signal with little contribution to finding extra binding site motifs (See the example of STAT1 data in the Supplementary Material Section 3). Nevertheless, it is difficult to expect that the new sequencing technology will always be able to provide perfect coverage of the genome in practice, and thus the previously deposited ChIP-chip datasets may be of significant value in improving TFBS identification in most cases.

Although our method is designed for combining ChIP-chip and ChIP-seq data, the HHMM framework is rather general and can be applied to other scenarios where information collected from multiple sources may be integrated. The opportunities for this type of joint analyses frequently arise in biomedical research. With the rapid development of new technologies, there are often multiple assays co-existing, measuring the same or closely related quantities of interest. Also for measuring protein–DNA binding, a series of assays have been developed, e.g. ChIP-PCR, ChIP-chip, ChIP-PET and ChIP-seq. Since these assays often have different sensitivity and specificity, straightforward combinations such as Union and Intersection do not work well. HHMM, on the other hand, is built under a coherent probability framework that is able to handle heterogeneity in sensitivity and specificity from the individual data sources, and therefore allows for easy incorporation of data from multiple experimental platforms. Because technologies are constantly changing, the two-stage HHMM estimation framework allows straightforward incorporation of data from new platforms.

ACKNOWLEDGEMENTS

Authors are grateful to the three reviewers for their helpful comments that significantly improved the article.

Funding: National Institute of General Medical Sciences (grant R01GM72007 to D.G. in part); the Huck Institute for Life Sciences at Penn State University (to D.G. in part); Career development award from University of Michigan Prostate SPORE (to Z.S.Q.'s research, in part).

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Bentley,D. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bui,H. *et al.* (2004) Hierarchical hidden Markov models with general state hierarchy. In *Proceedings of AAAI*. San Jose, CA.

Cartharius,K. *et al.* (2005) Matinspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

Chen,X. *et al*. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Consul,P. (1989) *Generalized Poisson Distributions*. Marcel Dekker, New York.

Dunn,K. and Davie,J. (2003) The many roles of the transcriptional regulator CTCF. *Biochem. Cell Biol.*, **81**, 161–167.

Du,J. *et al.* (2006) A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, **22**, 3016–3024.

Euskirchen,G. *et al.* (2007) Mapping of transcription factor binding regions in mammalian cells by chip: Comparison of array- and sequencing-based technologies. *Genome Res.*, **17**, 898–909.

Fine,S. *et al.* (1998) The hierarchical hidden Markov model: analysis and applications. *Mach. Learn.*, **32**, 41–62.

Huber,W. *et al.* (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**, 1963–1970.

Humburg,P. *et al.* (2008) Parameter estimation for robust HMM analysis of chIP-chip data. *BMC Bioinformatics*, **9**, 343.

Iyer,V. *et al.* (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Ji,H. and Wong,W. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.

Ji,H. *et al.* (2008) An integrated software system for analyzing chip-chip and chip-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Johnson,D. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Johnson,N. *et al.* (1992) *Univariate Discrete Distributions*. John Wiley & Sons, New York.

Kim,T. *et al.* (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.

Li,W. *et al.* (2005) A hidden Markov model for analyzing chIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**, i274–i282.

Loh,Y. *et al*. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.

Mikkelsen,T. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Munch,K. *et al.* (2006) A hidden Markov model approach for determining expression from genomic tiling microarrays. *BMC Bioinformatics*, **7**, 239.

Ohlsson,R. *et al.* (2001) CTCF is uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.*, **17**, 520–527.

Orlando,V. and Paro,R. (1993) Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell*, **75**, 1187–1198.

Park,P. (2008) Epigenetics meets next-generation sequencing. *Epigenetics*, **3**, 318–321.

Parmigiani,G. *et al.* (2002) A statistical framework for expression-based molecular classification in cancer. *J. R. Stat. Soc. B*, **64**, 717–736.

Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA-associated proteins. *Science*, **290**, 2306–2309.

Robertson,G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Schoenherr,C. *et al.* (1996) Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl Acad. Sci. USA*, **93**, 9881–9886.

Schones,D. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Scott,S. (2002) Bayesian methods for hidden Markov models: recursive computing in the 21st century. *J. Am. Stat. Assoc.*, **97**, 337–351.

Shah,S. *et al.* (2007) Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, **23**, i450–i458.

Shivaswamy,S. *et al.* (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, e65.

Solomon,M. *et al.* (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.

Wei,C. *et al.* (2006) A global map of p53 transcription factor binding sites in the human genome. *Cell*, **124**, 207–219.