Bioinformatics Advance Access published October 12, 2006

BIOINFORMATICS

A Supervised Hidden Markov Model Framework for Efficiently Segmenting Tiling Array Data in Transcriptional and ChIP-chip Experiments: Systematically Incorporating Validated Biological Knowledge

Jiang Du [1], Joel S. Rozowsky [2], Jan O. Korbel [2], Zhengdong D. Zhang [2], Thomas E. Royce [2,3], Martin H. Schultz [1], Michael Snyder [2,4], Mark Gerstein [1,2,3*]

[1]Department of Computer Science, Yale University, New Haven, CT 06520, USA, and [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT 06520, USA, [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven CT 06520, USA, [4]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven CT 06520, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Large-scale tiling array experiments are becoming increasingly common in genomics. In particular, the ENCODE project requires the consistent segmentation of many different tiling array data sets into "active regions" (e.g. finding transfrags from transcriptional data and putative binding sites from ChIP-chip experiments). Previously, such segmentation was done in an unsupervised fashion mainly based on characteristics of the signal distribution in the tiling array data itself. Here we propose a supervised framework for doing this. It has the advantage of explicitly incorporating *validated biological knowledge* into the model and allowing for formal training and testing.

Methodology: In particular, we use a hidden Markov model (HMM) framework, which is capable of explicitly modeling the dependency between neighboring probes and whose extended version (the generalized HMM) also allows explicit description of state duration density. We introduce a formal definition of the tiling-array analysis problem, and explain how we can use this to describe sampling small genomic regions for experimental validation to build up a gold-standard set for training and testing. We then describe various ideal and practical sampling strategies (e.g. maximizing signal entropy within a selected region versus using gene annotation or known promoters as positives for transcription or ChIP-chip data, respectively).

Results: For the practical sampling and training strategies, we show how the size and noise in the validated training data affects the performance of an HMM applied to the ENCODE transcriptional and ChIP-chip experiments. In particular, we show that the HMM framework is able to efficiently process tiling array data as well as or better than previous approaches. For the idealized sampling strategies, we show how we can assess their performance in a simulation framework and how a maximum entropy approach, which samples sub-regions with very different signal intensities, gives the maximally performing gold-standard. This latter result has strong implications for the optimum way medium-scale validation experiments should be carried out to verify the results of the genome-scale tiling array experiments.

Supplementary information: The supplementary materials are available at http://tiling.gersteinlab.org/hmm/.

Contact: mark.gerstein@yale.edu

1 INTRODUCTION

1.1 Motivation

Tiling arrays are used to survey genomic transcriptional activity (Bertone *et al.*, 2004; Cheng *et al.*, 2005; Kapranov *et al.*, 2002; Rinn *et al.*, 2003; Schadt *et al.*, 2004) and transcription factor binding sites (Buck and Lieb, 2004; Cawley *et al.*, 2004; Iyer *et al.*, 2001) at high resolution. The raw/preprocessed data from tiling array experiments are first processed by certain analysis methods, which produce a list of predicted genomic "active regions". These are either transcriptionally active regions (TARs)/transcribed fragments (transfrags) (Bertone *et al.*, 2004; Cheng *et al.*, 2005; Kampa *et al.*, 2004; Rinn *et al.*, 2003) or transcription factor binding sites. Usually a subset of these regions is further studied by experimental validation, which answers the question of whether these regions are actually active or not.

With the beginning of projects such as ENCODE (ENCODE Project Consortium, 2004), which aims to annotate the genome sequence with the function of specific elements (e.g. whether they are regulatory sites, exons or introns), the large scale tiling array experiments that are carried out present a number of new challenges. One of these is how to build up an existing *knowledge base of validated biological information* about genomic elements such as the location of exons and introns or of transcription factor binding sites, and how to use this knowledge base in combination with the tiling array data on a limited region of the genome to construct a predictive model that we can extrapolate to the rest genome in order to best segment it into functional elements.

*to whom correspondence should be addressed

We also have the related problem of how to grow this knowledge base of validated biological information systematically so as to do the extrapolation most efficiently. We envision that it will not be possible to validate every single ChIP-chip experiment binding site, or every single exon in the human genome using RT-PCR. However, we can imagine that following the large scale tiling array experiments there will be medium-scale validation experiments done on thousands of predicted binding sites and gene structures to try to verify them. The question is: how should these binding sites and gene structures for validation be picked? They could, of course, be selected in terms of having the best scores, but one would like to pick them so as to derive a model that would be best able to analyze the remainder of the data accurately.

Here we tackle both of these challenges by proposing a hidden Markov model (HMM) (Rabiner, 1989) framework which integrates the existing *validated biological knowledge* about gene structures and transcription factor binding sites, and then uses this encapsulated biological knowledge to segment tiling array data. In particular, we also show how one can systematically pick un-annotated or unlabeled regions from the tiling array data for further validation to grow the validated biological knowledge base of labeled examples most optimally in order to get a maximally predictive model.

We do our analysis side by side on both transcriptional data and ChIP-chip binding site data. We have two reasons for this. First of all, it shows the general utility of the approach, that we can apply the same formalism to tiling array data from both types of experiments. Second, since data from the two different experiments have different levels of validated biological knowledge, it allows us to see how our formalism performs in two areas with different amounts of knowledge. Finally, because we can get a better handle of how things work on the better studied transcriptional data, we can have great confidence that we are applying a correct approach when segmenting the ChIP-chip data.

1.2 Previous work

In tiling array data analysis, the goal is to identify genomic active regions with high signal intensities. This procedure can not be implemented in a naïve fashion, due to the noise in the background and the possible low signal intensities in some active regions (Hoyle *et al.*, 2002; Royce *et al.*, 2005). Different statistical algorithms have been developed to process the tiling array data. Earlier examples include pseudo-median threshold with maxgap/minrun (Karplus *et al.*, 1999), p-value cutoff with maxgap/minrun (Bertone *et al.*, 2004), sliding-window PCA with MD (Schadt *et al.*, 2004), and variance stabilization (Gibbons *et al.*, 2005). More recently, several HMM approaches and HMM variants have been developed (Ji and Wong, 2005; Li *et al.*, 2005; Marioni *et al.*, 2006). Flicek *et al.* (personal communication) have also applied an HMM to ChIP-chip data resulting from tiling array signals characteristic of histone modifications.

Some of these existing methods, such as maxgap/minrun (Bertone *et al.*, 2004; Kampa *et al.*, 2004), involve parameters that have to be decided manually. HMM approaches, formerly introduced in the field of sequence analysis (Eddy, 1998; Karplus *et al.*, 1999; Krogh *et al.*, 1994), have the advantage of not using any additional parameters other than the model itself. Li *et al.* (2005) proposed the construction of a two-state HMM for ChIP-chip data partially based on the results of Affymetrix SNP arrays (Lieberfarb *et al.*, 2003). Ji and Wong (2005), more recently, proposed a more general Unbalanced Mixture Subtraction (UMS) approach to recover different emission distributions in a HMM from a mixture distribution. However, in some cases, there may exist neither corresponding experimental results that can be utilized to build the HMM, nor validated biological knowledge comprehensive enough for an unbiased evaluation in the UMS analysis.

On the other hand, the use of partially validated knowledge about the array data, such as gene annotation or experimental validation results on small genomic regions, has not been specifically considered by existing methods; and there does not exist a systematic framework to optimally obtain and utilize this kind of knowledge in tiling array analysis. Such a framework will have the potential to better assist the analysis of tiling array data, as the related validated knowledge becomes more abundant and accurate via experimental validations.

1.3 Methodology

In this paper we propose a new supervised scoring framework based on HMM that will consistently score different types of tiling array data by incorporating validated biological knowledge. As our framework will be based on both transcriptional and regulatory data, we can demonstrate its efficiency on the better described transcriptional data so that we have greater confidence when applying it to the ChIP-chip data.

An integral part of our strategy is developing a scheme to intelligently select sub-regions for validation, in order to better build up gold standard sets to incorporate into our statistical model. We investigate the performances of different sample selection schemes described in section 2 on a simulated dataset in section 4, and propose to employ the *MaxEntropy* scheme as a measure for sample selection: we want to select sub-regions that have the highest entropies for experimental validation first, so as to effectively build up the validated biological knowledge for our HMM approach.

After the sample sub-regions are selected and their corresponding state sequences are obtained via further validation experiments or according to existing validated biological knowledge, a frequency-based supervised learning algorithm is applied to build the HMM and then the Viterbi algorithm is utilized to compute the most likely state sequence for the whole sequence of array signals. Since current experimental validation data are insufficient to apply our *MaxEntropy* sampling scheme, we also propose alternative methods for choosing sample sub-regions. As described in section 3, for transcriptional tiling array experiments, a four-state HMM can be constructed by learning from the sequences of probes which fall into regions of the corresponding gene annotation. For ChIP-chip data, the knowledge of gene annotation is again relevant to the identification of binding sites, because transcription factor binding sites (TFBS) are usually considered to be enriched in upstream regions of genes and unlikely to occur in inner regions of genes. By incorporating this knowledge, a two-state HMM can be constructed for further analysis. Empirical results in section 4 show that our methods effectively handle large datasets, even with relatively noisy training data.

2 METHODS

2.1 Idealized definitions of the problem

In this section we give two idealized definitions of the tiling-array analysis problem, which will form the basis of our core algorithms on both sample sub-region selection and HMM analysis based on the selected samples.

Definition 1. Idealized HMM Tiling Problem (*HTP*). An idealized HMM tiling problem is a tuple $\langle D, C_{sample}, O \rangle$, where D is the emission sequence corresponding to a hidden state sequence S generated by an unknown HMM M, C_{sample} is the constraint on how sample sub-regions can be selected in D (e.g. the maximum length of each sample sub-sequence), and O is a labeling oracle (an imaginary black box which is able to answer certain questions) that can discover the corresponding hidden state sequence of any sample sub-region in D. A solution to the problem first selects a set of sample sub-regions in D according to the constraint C_{sample}, asks the labeling oracle O about the corresponding state sequences of these sample sub-regions, then efficiently computes a model M' for D and outputs the corresponding state sequence S' for D.

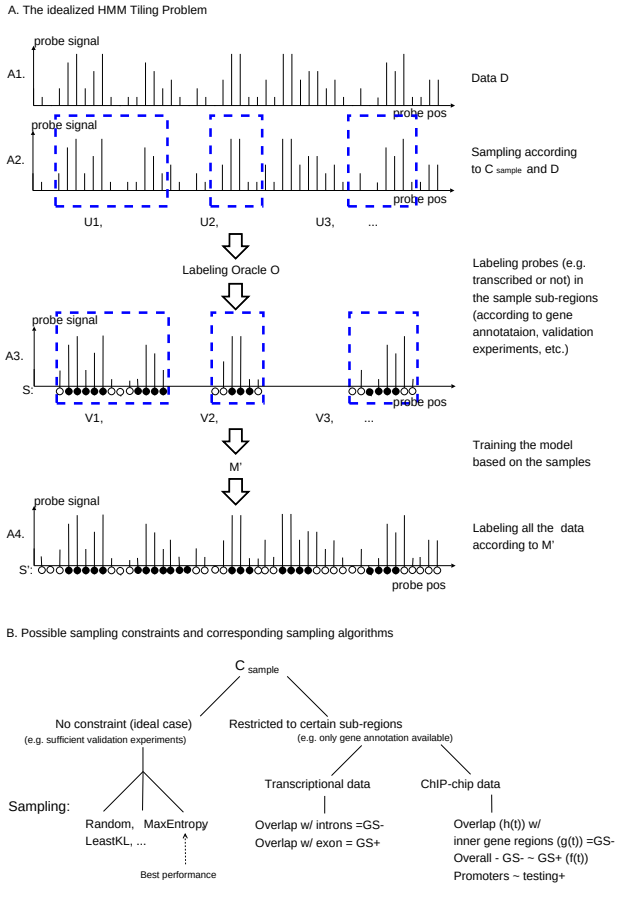

Fig. 1. Idealized HMM tiling-array analysis problem. (A) Idealized HMM tiling problem. (B) Sampling constraints and corresponding strategies.

As shown in Figure 1A, S and D, generated by M in the problem's assumption, corresponds to the biological state (for instance, transcribed or not transcribed) sequence and signal intensity sequence of the probes in the array, preferably after necessary preprocessing such as normalization. The length of the sequence, L, corresponds to the size of the tiling array. The solution to the problem, which is also the framework we propose, first selects m sample sub-regions $\{U_1, U_2, ..., U_m\}$ in D according to the sampling constraint C_{sample}, and passes them to the labeling oracle O, which corresponds to an experimenter who refers to *validated biological knowledge* (existing annotation, validation experiments, etc.) and then discovers the hidden state (label) sequences $\{V_1, V_2, ..., V_m\}$ for these small subsets of neighboring probes in the array. These sub-sequences of U_is and V_is form the samples/training set of our analysis methods. A model M' is then learned based on this training set, and processed by a decoding algorithm on D, which outputs the predicted corresponding state sequence S' for D.

The sampling constraint C_{sample} corresponds to the possible limitations in selecting sample sub-regions in real tiling array problems. As shown in Figure 1B, when experimental validations can be done on any set of genomic sub-regions, there will be no constraint on sampling at all and C_{sample} will be equal to null/empty. In the other extreme, if no further validation experiments can be done and the only available validated knowledge is the gene annotation related to the transcriptional tiling experiment, C_{sample} will only allow those sub-regions inside the gene annotation to be selected (otherwise the labeling oracle will fail to label all the sample sub-regions). One can imagine intermediate situations between these extremes.

HTP differs from the real problem of tiling array data analysis in two main aspects. On one hand, the actual state sequence S of the array data is not necessarily generated by a certain HMM. Such an HMM assumption is stated in *HTP* not only because that it is a reasonable approximation to the real problem, whose data fits the continuing nature of a HMM, but also because it is necessary for further performance analysis of the solutions to this problem. On the other hand, the labeling oracle O (e.g. experimental validation) in real problems is not always perfect and can make mistakes, from which we can give a generalization of *HTP* in the following definition:

Definition 2. Idealized HMM Tiling Problem with an Imperfect Oracle (*HTPIO*). An idealized HMM tiling problem with an imperfect labeling oracle is a tuple $\langle D, C_{sample}, O^I \rangle$, which has the same definition as *HTP*, except that the labeling oracle O^I is not perfect and may make mistakes when discovering the underlying state sequences $\{V_1, V_2, ..., V_m\}$ for sample sub-sequences $\{U_1, U_2, ..., U_m\}$. Obviously, *HTPIO* is a generalization of *HTP*.

Here we also define an intuitive metric for the solution S' to both problems:

Definition 3. Error rate of a solution S' for *HTPIO* ($Error(S', S)$).

$$Error(S', S) = \frac{Difference(S', S)}{L} \quad (1)$$

where the difference of two state sequences is computed as the number of corresponding elements that do not agree with each other, and S' and S are of the same length L.

The smaller the error rate, the better is the solution. However, in real problems it is hard to apply this metric, since the actual hidden sequence is unknown. This definition only serves as a performance measurement in section 4 about results on simulated datasets. Other possible performance measures for real experimental datasets are also discussed in section 4.

A similar problem to *HTPIO* has been studied by Abe and Warmuth (1992) in the context of Probabilistic Automata (PA). Our work differs from theirs in several aspects. First of all, we investigate the problem of sample sub-region selection whereas they do not. Second, we take errors in the labeling oracle into consideration. Third, we introduce a more intuitive measurement of error, compared to the *Kullback-Leibler divergence* of different PAs in their paper. Last but not least, we seek a time-efficient solution, whereas their work focuses on obtaining sample complexity bounds for learning the model while ignoring computational efficiency.

As described above, *HTPIO* asks for solutions to two different kinds of sub-problems simultaneously: one solution on an effective sub-region sampling scheme and one corresponding solution on an efficient algorithm to output a good approximation of S. These two solutions form our HMM framework, which systematically incorporates validated knowledge into tiling

array data analysis. In the following two sub-sections, we present efficient solutions to both sub-problems separately.

2.2 Selection of sample sub-regions

When deciding which sample sub-regions in D should be selected as inputs to the labeling oracle, we investigate a set of sample selection schemes besides random selection. To simplify discussion, we assume that C_{sample} is equal to null/empty and that we are selecting m non-overlapping sample sub-sequences $\{U_1, U_2, ..., U_m\}$, each of length k.

Some of these sampling schemes employ entropy as a measure. The first one of these, *MaxEntropy*, selects m non-overlapping sub-regions with the highest entropies. The second one, *UnbiasedEntropy*, divides all the sub-regions into m groups according to their entropy values, and randomly selects one sub-region out of each group. The third one, *MaxMinEntropy*, selects $m/2$ sub-regions with the highest entropies and $m/2$ sub-regions with the lowest entropies. *MaxEntropy* tends to pick up those sub-regions that contain both active and inactive probes in the same region (e.g. the transcribed gene regions in transcriptional tiling arrays), while the other two methods will pick up totally inactive sub-regions as well.

Another sampling scheme, *LeastKL*, employs a well-known measure in information theory called *"Kullback-Leibler divergence"* (Kullback and Leibler, 1951), between D of length L and its sub-sequence U_i of length k.
Definition 4. Kullback-Leibler Divergence (*K-L divergence*). Let D and U_i be probability distributions over a countable domain Z. The *Kullback-Leibler Divergence* of D with respect to U_i, $d_{KL}(D, U_i)$ is defined as follows:

$$d_{KL}(D, U_i) = \sum_{z \in Z} P_D(z) \log_2 \frac{P_D(z)}{P_{U_i}(z)} \quad (2)$$

By convention we let $0 \log 0 = 0$, and $0/0 = 1$.

Normally we think the smaller $d_{KL}(D, U_i)$, the more similar U_i is to D in terms of their probability distributions over Z. When selecting sample sub-sequences for *HTPIO* using *LeastKL*, we want to select m sub-sequences U_i with the smallest $d_{KL}(D, U_i)$ values. The underlying idea is to obtain information from those most representative regions for future learning algorithms.

For tiling array data, D is usually a sequence of uncountable real numbers, so the elements in D need to be discretized to integers (either by direct rounding, or rounding after $\log$ transformation, depending on the nature of the data), which requires $O(L)$ operations. When m, k are constants and $m, k << L$, an approximate result of the m non-overlapping sub-sequences can be obtained in $O(L)$ for all these schemes.

Empirical results in section 4 show that when the labeling oracle is perfect, the *MaxEntropy* and *LeastKL* sample selection algorithm are superior to other schemes; when the oracle makes relatively small mistakes, *MaxEntropy* always outperforms other schemes.

2.3 An efficient HMM approach for *HTPIO*

After the sample sub-sequences and their corresponding state sequences have been obtained, a frequency-based supervised learning algorithm is applied to build the HMM and then a Viterbi algorithm (Rabiner, 1989; Viterbi, 1967) is utilized to compute the most likely state sequence S' for the whole sequence D, which is an approximate answer to *HTPIO*. The forward-backward algorithm (Rabiner, 1989) can also be used to generate detailed scores for each element in D, although it will be more time consuming than the Viterbi algorithm.

The supervised learning algorithm takes as input the sample sub-sequences $\{U_1, U_2, ..., Um\}$ and corresponding state sequences $\{V_1, V_2, ..., V_m\}$, each of length k, and outputs the following matrices:

$$A_{ij} = \frac{\sum_{V \in \{V_1, V_2, ..., V_m\}} \xi_V^S(i, j)}{\sum_{V \in \{V_1, V_2, ..., V_m\}} \gamma_V^S(i)} \quad (3)$$

$$B_{ik} = \frac{\sum_{(V,U) \in \{(V_1,U_1),(V_2,U_2),...,(V_m,U_m)\}} \xi_{V,U}^O(i, k)}{\sum_{V \in \{V_1, V_2, ..., V_m\}} \gamma_V^S(i)} \quad (4)$$

where $\xi_V^S(i, j)$ is the number of transitions from state i to j in state sequence V, $\gamma_V^S(i)$ is the number of occurrences of state i in V, $\xi_{V,U}^O(i, k)$ is the number of times state i in V emits k in U. We can then build a discrete HMM with A as the transition matrix, and B as the emission matrix. We set the initial state distribution of the HMM to uniform to avoid biased estimation for this parameter. As long as the initial state distribution is set to a reasonable distribution, it should not have a great impact on the final result when L is sufficiently large. When the sample size is relatively small, the discrete emission matrix B may be ill-formed if estimated directly, in which case we build a continuous HMM and use kernel density estimation (Parzen, 1962) to construct smoother emission distributions for different states: if $x_1, x_2, ..., x_N$ are the observed emissions for a certain state, then its corresponding emission distribution is computed as $P(x) = \frac{1}{N} \sum_{i=1}^{N} W(x - x_i)$, where in this case W is a Gaussian function with mean 0 and predefined variance σ^2.

The supervised learning algorithm runs in $O(mk)$ time, and the Viterbi algorithm requires $O(n^2 L)$ time, where n is the number of states (which is 2 or 4 in examples in section 3) in the HMM and L is the length of D. Since $mk < L$, the total time cost of our solution (sampling, learning, and decoding) to *HTPIO* is thus $O(n^2 L)$, which is comparable to most of the existing tiling array analysis methods. Results in section 4 show that our methods handle large datasets effectively.

3 IMPLEMENTATIONS

In this section, we will show that even though at present there may exist too little experimentally validated data to be incorporated in our HMM approach described above, other kinds of validated knowledge such as gene annotation already provide a good basis for our methods in both transcriptional and ChIP-chip data analysis.

3.1 Incorporating gene annotation in transcriptional data analysis

In transcriptional tiling array experiments, TARs or transfrags form the subject of interest. Here the gene annotation of the organism in study is obviously the validated biological knowledge we should consider to incorporate into our HMM approach.

Despite its inaccuracy, the knowledge of gene annotation usually involves a large amount of information. This allows the construction of a four-state HMM instead of a two-state HMM. The structure of the HMM is illustrated in Figure 1A in Supplementary figures. Each probe in the tiling array can be in one of the four HMM states (TAR, NONTAR, and two other intermediate transition states), emitting the assigned intensity/score. As shown in Figure 1B, the parameters of the HMM can be estimated by learning from both positive and negative samples in the sequences of probes which fall into regions with known transcription characteristics, in this case, the knowledge of corresponding gene annotation.

What is more, the choice of annotated genes as the training set conforms to our MaxEntropy sample selection scheme, since these regions usually contain both high and low signals, thus having relatively high entropy values.

3.2 Incorporating gene annotation in ChIP-chip data analysis

For ChIP-chip data, we should first identify the possible knowledge to incorporate into our HMM approach, since this is not as obvious as for transcriptional data, where gene annotation is an intuitive choice. One option is the dataset of those experimentally verified regions, which at present is

usually limited in size and cannot form a valid training set for HMM construction. On the other hand, the knowledge of gene annotations is somewhat related to the identification of binding sites, since transcription factor binding sites (TFBS) are usually considered to be enriched in upstream regions of genes, and unlikely to occur in inner regions of genes. By incorporating this knowledge, a two-state HMM can be constructed in the following way:

As shown in Figure 1B in Supplementary figures, the HMM contains a TFBS state 0 and a non-TFBS state 1. The overall emission distribution $h(t)$ is computed based on the ChIP-chip data. As shown in Figure 1B, the emission distribution of the non-TFBS state, $g(t)$, according to the above discussion, can be estimated based on the knowledge of inner regions in genes. The emission distribution of the TFBS state, $f(t)$, can then be obtained by subtracting $g(t)$ from $h(t)$, using canonical FDR procedures. The transition parameters of the HMM can be estimated based on empirical knowledge. Actually, if $f(t)$ and $g(t)$ are significantly different from each other, a small variance in transition parameters should not affect the result of HMM approach very much.

However, the HMM constructed in this way may not be as effective as in the case of transcriptional data, since the knowledge involved in the construction does not relate to the TFBS very closely. Further scoring on the initial analysis results can be done by computing the posterior probabilities $P(S_i = k|D)$ for the predicted states on probes, where S_i is the state of the ith probe, k is the predicted state, and D is the emitted sequences of the probes involved. These scores indicate the confidence in every single prediction and can be used to refine the prediction results obtained by HMM analysis. The identified active probes can then be ranked according to the overall confidence levels in their regions and a threshold confidence level may either be set manually or be learned automatically to refine the original results.

3.3 Incorporating other validated knowledge in tiling array data analysis

Since our HMM framework defined in section 2 provides a general interface for incorporating validated knowledge about the dataset in question, virtually any such knowledge can be utilized by this approach. For example, our framework can take the data from a tiling array experiment, and select a medium-sized set of sub-regions by using some appropriate analysis method (e.g. the *MaxEntropy* sampling scheme in section 2.2). These sub-regions can be further studied by experimental validations, which identifies the underlying state (e.g. transcribed or not, in a transcriptional tiling array experiment) of every single probe inside these sub-regions. These knowledge form a well-established training set and can then be incorporated into our HMM approach in the framework, which will lead to more accurate analysis results than that obtained using only information from the array data. Since all these can be done systematically within our framework, it actually provides a way to consistently analyze tiling array data across a number of experiments and also across different types of experiments.

4 RESULTS

4.1 Performance measurement

We use $Error(S', S)$ defined in section 2.1 as an intuitive measure to analyze the results on a simulated dataset, where we have access to the actual hidden state sequence S. We also investigate some key issues in our HMM approach, including sample selection, size of the training set, and error in the training data.

When we analyze the results on real experimental data, it is hard to get a good estimation of S, which makes it difficult to compute the overall error rate. One the other hand, for a rigorous performance evaluation like cross-validation, a gold-standard dataset with exact information is required. Unfortunately, in many cases no such dataset exists, especially over large genomic regions. In the absence of such a gold standard, we evaluate the performance of different methods by comparing their results against the imperfect training set used in the approach, and also against previous segmentation results of other non-HMM methods on the same dataset. Furthermore, we investigate how the size and noise of the training set affects the performance of our HMM approach.

4.2 Results on simulated dataset

A simulation on our framework of the solution to *HTPIO* proposed in section 2 was done to investigate its performance. We performed ~ 17000 trials, each of which solved a randomly generated *HTPIO* of $\langle D, C_{sample}, O^I \rangle$, where the length L of D is 1M, constraint C_{sample} specifies that $m = 2^i$ ($i = 1, 2, ..., 8$) sub-regions, each of length $k = 50$, should be selected as samples, and O^I makes mistakes randomly with probability $e = 0, 0.05, 0.1$; $Error(S', S)$ was computed in each trial for different sample selection schemes described in section 2.2. The results in Figure 2 (and Figure 2 in Supplementary figures) show that *MaxEntropy* and *K-L divergence* based sample selections are superior to other selection schemes when the labeling oracle O^I is perfect. When O^I makes mistakes with a relatively low probability, *MaxEntropy* outperforms all other sampling schemes. We also observe that as the sample size mk increases, the overall performances of all methods improve, and become stable when the sample size is larger than ~ 0.013M. This observation leads to a hypothesis that an intelligently selected medium-sized training set is sufficient for our HMM approach on real experimental datasets, which is supported by the results in section 4.3 as well.

4.3 Results on transcriptional dataset

We tested our method on a transcriptional tiling array dataset which has 25mer oligonucleotide probes tiled approximately every 21bp covering all the non-repetitive DNA sequence of the ENCODE regions ($\sim$ 30Mb) (ENCODE Project Consortium, 2004). This dataset is sufficiently large for our performance test, and the corresponding prediction result of a minrun/maxgap method (Bertone *et al.*, 2004) is available as well, which provides a good estimation of the TARs.

We formed the training set (~ 7.5Mb) from the normalized dataset by using the method in section 3.1 with the RefSeq annotation (Pruitt *et al.*, 2005). In order to investigate the performances of our methods with different-sized training sets, we also randomly selected a certain portion of the whole training set, and then built a basic discrete four-state HMM (Figure 1A in Supplementary figures) and a continuous HMM (by using kernel density estimation) based on that portion. The portions we selected were $1/2$, $1/4$, $1/8$ and $1/16$ of the whole training set, and every selection was repeated 16 times so that the variance of the corresponding performances could be estimated empirically. We also built a generalized HMM (GHMM) (Mohamed and Gader, 2000; Rabiner, 1989) based on the whole training set to test the possible gain of using a more sophisticated model which captures length characteristics.

Figure 3 uses *Youden's J* (Youden, 1950), which is $Sensitivity + 1 - Specificity$, as a measure of the overall performances of different methods with different-sized training sets. The sensitivity and specificity of the HMM prediction results are computed based on both the whole training set and the previous prediction results of maxgap/minrun. Figure 3 shows that even when $1/4$ (~ 1.9Mb) of the whole training set is used, our HMM approach gives a performance comparable to or better than existing methods, with either gene annotation or previous prediction results as performance criteria. Another important fact shown in Figure 3 is that the continuous

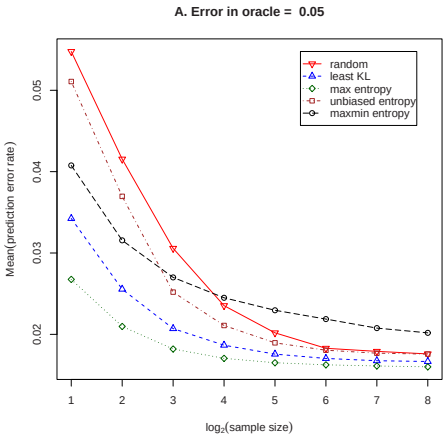

Fig. 2. Results on simulated dataset. "Error in Oracle" is the probability with which O^I makes mistakes. (A) Mean of the prediction error rates. (B) Standard deviation of prediction error rates.

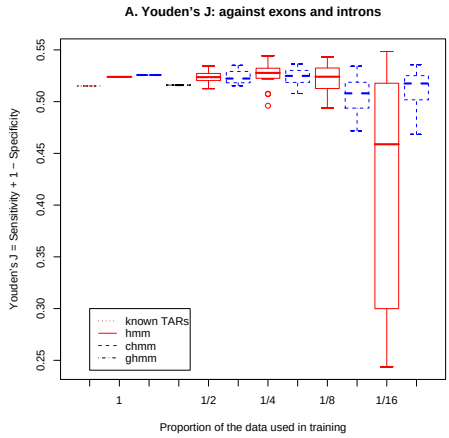

Fig. 3. Results on transcriptional dataset: Youden's J. (A) RefSeq exon regions are used as positives, and intron regions as negatives. (B) Known TARs predicted by maxgap/minrun method are used as positives.

HMM has much more stable performance than the discrete model, especially when the training set is small (less than $1/4$ of the whole training set). This is because the continuous HMM has smoother emission distribution estimations than the discrete one, and its performance is thus less likely to be affected by a small set of biased samples. We can also observe that GHMM does not seem to give significantly better performance than simpler models.

We further computed the posterior probabilities for the predicted states on probes, and set different thresholds to identify TARs. Figure 4 shows the ROC curves of different models with different training sets. Again the continuous HMM outperforms the discrete one, and has good performance even with a relatively small ($\sim$ 1.9Mb) training set. The similarity of A and B diagrams in Figure 3 and Figure 4 also shows that gene annotation is a good criterion for performance measurement, if we do not have any existing prediction results to utilize.

The minimum training set guaranteeing good performance for our approach on this dataset is $\sim$ 1.9Mb, which includes $\sim$ 0.1M probes. Since the size of the training set needed for satisfying performance of our method does not increase with the size of the dataset, it seems that if $\sim$ 0.1M probes in this type of tiling array experiment can be labeled and put into the training set, our method becomes immediately applicable to identify TARs for the whole dataset. We also want to point out that the labeling process does not have to be perfect: in this case, Figure 3A shows that less than 60% of the training set is actually correct, while Figure 3B shows that our method has satisfying performance with this training set.

4.4 Results on ChIP-chip dataset

We tested our method on a STAT1 ChIP-chip tiling array dataset which has 50mer oligonucleotide probes tiled approximately every 38bp covering most of the non-repetitive DNA sequence of the ENCODE regions ($\sim$ 30Mb). This dataset, as in the case of section 4.3, is sufficiently large for our performance test, and the corresponding prediction result of a maxgap/minrun method is available as well, which provides a good estimation of the TFBSs.

Due to the lack of available validated biological knowledge, we built a simple two-state continuous HMM (Figure 1B in Supplementary figures) based on the negative training set ($\sim$ 8Kb) from the normalized dataset by using the method described in section 3.1

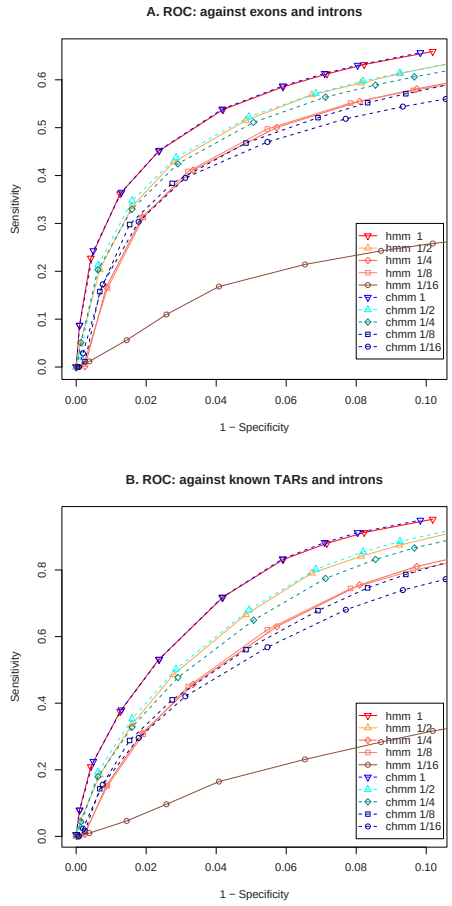

Fig. 4. Results on transcriptional dataset: ROC curves. "hmm 1/2" stands for the discrete HMM built with $1/2$ of the whole training set, and so on. (A) RefSeq exon regions are used as positives, and intron regions as negatives. (B) TARs predicted by the maxgap/minrun method are used as positives.

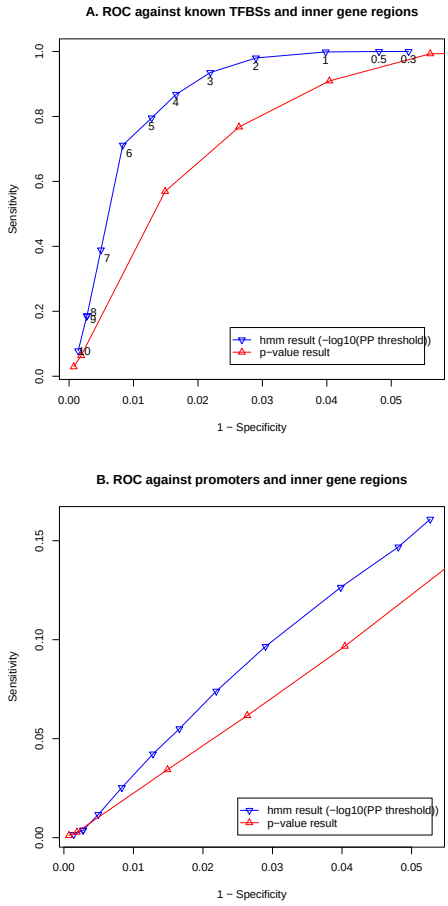

Fig. 5. Results on ChIP-chip dataset: ROC curves. (A) Previously predicted TFBSs are used as positives, and the inner gene regions as negatives. The numbers along the ROC curve of HMM result are the $-\log_{10}(PP\ threshold)$, where PP is the posterior probability of a probe being in NON-TFBS state. (B) The promoter regions in the array are used as positives.

with RefSeq annotation, computed the posterior probabilities for the probes being in NON-TFBS state, and set different thresholds to get different sets of TFBSs. Figure 5 shows the ROC curves of predictions by using our HMM approach and a p-value cutoff method. The inner gene regions are used as negatives, while both previously predicted TFBSs and the promoter regions in the array are used as positives. We can observe that the HMM approach has better performance than the p-value cutoff approach in both criterions.

The near-linear ROC curves in Figure 5B also show that the promoter regions may not be as good a criterion as the previous TFBS results. Analogous to the case with transcriptional data, when experimental validation results become sufficient to form a medium-sized (covering ~ 0.1M probes) knowledgebase about the dataset in question, this knowledgebase can be utilized as a performance measure as well as the training set for our HMM approach.

5 DISCUSSION AND CONCLUSIONS

We present an efficient HMM framework which systematically incorporates *validated biological knowledge* into tiling array data analysis. This framework, which consists of a *MaxEntropy* sample selection algorithm and HMM learning and decoding approaches, is proposed based on *HTPIO*, an idealized definition of the tiling array analysis problem. Empirical results of our methods in the framework on a simulated dataset, a transcriptional dataset and a ChIP-chip dataset show that our framework effectively handles large datasets, even with a relatively noisy training set.

Our work differs from previous studies in tiling array data analysis by specifically taking *validated biological knowledge* into consideration and systematically incorporating it using an empirically tested *MaxEntropy* sample selection scheme for optimal analysis. These features ensure the good performance of our framework with even a relatively small gold standard training set, which has not been specifically considered by previous methods. In this way our framework can consistently analyze tiling array data across a number of experiments, and can process different types of array data automatically, without the need to manually set additional parameters. This

feature will become an advantage for analyzing very large datasets (e.g. for the $\sim$ 3Gb human genome): when sufficient experimental validations are done afterwards, a *medium-sized* (covering ~ 0.1M probes, according to the empirical results in section 4.3) *validated biological knowledgebase* can be formed for the array data in question. Our framework can then improve its performance with the guidance of this medium-sized knowledgebase, and its refined analysis results can in turn assist further experimental studies. What is more, in section 4.3 our framework gives good performance by incorporating some relatively inaccurate biological knowledge (with approximately 60% correctness), and the sub-regions in the training set are not specifically chosen according to our proposed sampling scheme. We can expect that for real problems which use validated biological knowledge from highly accurate experimental validations, the necessary minimum size of the biological knowledgebase could be even smaller than ~ 0.1M probes for our framework to achieve satisfying performance.

Another feature of our method is that given a set of regions with similar signal intensities, it can identify all the regions in the whole dataset with similar signal distributions. This feature is potentially useful for identifying regions with different transcription levels. For instance, our HMM method can take as the training set all the known highly expressed genes in the tissue, and then identify all the regions in the corresponding transcriptional tiling array that have the similar transcription level.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their advices and comments. We acknowledge support from the NIH (1U01HG003156-01). J.O.K. is supported by a European Molecular Biology Organization Long-Term Fellowship.

REFERENCES

Abe,N. and Warmuth,M. (1992) On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning,* **9**, 205–260.

Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S., Gerstein,M. and Snyder,M. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science,* **306** (5705), 2242–2246.

Buck,M.J. and Lieb,J.D. (2004) Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics,* **83** (3), 349–360.

Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J., Wheeler,R., Wong,B., Drenkow,J., Yamanaka,M., Patel,S., Brubaker,S., Tammana,H., Helt,G., Struhl,K. and Gingeras,T.R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell,* **116** (4), 499–509.

Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G., Sementchenko,V., Piccolboni,A., Bekiranov,S., Bailey,D.K., Ganesh,M., Ghosh,S., Bell,I., Gerhard,D.S. and Gingeras,T.R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science,* **308** (5725), 1149–1154.

Eddy,S.R. (1998) Profile hidden markov models. *Bioinformatics,* **14** (9), 755–763.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science,* **306** (5696), 636–640.

Gibbons,F.D., Proft,M., Struhl,K. and Roth,F.P. (2005) Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biol,* **6** (11), R96.

Hoyle,D.C., Rattray,M., Jupp,R. and Brass,A. (2002) Making sense of microarray data distributions. *Bioinformatics,* **18** (4), 576–584.

Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature,* **409** (6819), 533–538.

Ji,H. and Wong,W.H. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics,* **21** (18), 3629–3636.

Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G., Tammana,H. and Gingeras,T.R. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res,* **14** (3), 331–342.

Kapranov,P., Cawley,S.E., Drenkow,J., Bekiranov,S., Strausberg,R.L., Fodor,S.P.A. and Gingeras,T.R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science,* **296** (5569), 916–919.

Karplus,K., Barrett,C., Cline,M., Diekhans,M., Grate,L. and Hughey,R. (1999) Predicting protein structure using only sequence information. *Proteins,* **Suppl 3**, 121–125.

Krogh,A., Brown,M., Mian,I.S., Sjlander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol,* **235** (5), 1501–1531.

Kullback,S. and Leibler,R. (1951) On information and sufficiency. *Annals of Mathematical Statistics,* **22(1)**, 7986.

Li,W., Meyer,C.A. and Liu,X.S. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics,* **21 Suppl 1**, i274–i282.

Lieberfarb,M.E., Lin,M., Lechpammer,M., Li,C., Tanenbaum,D.M., Febbo,P.G., Wright,R.L., Shim,J., Kantoff,P.W., Loda,M., Meyerson,M. and Sellers,W.R. (2003) Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res,* **63** (16), 4781–4785.

Marioni,J.C., Thorne,N.P. and Tavar,S. (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics,* **22** (9), 1144–1146.

Mohamed,M. and Gader,P. (2000) Generalized hidden markov models part i: theoretical frameworks. *IEEE Transcations on Fuzzy Systems,* **8**, 67–81.

Parzen,E. (1962) On estimation of a probability density function and mode. *Ann. Math. Stat.,* **33**, 1065–1076.

Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* **33** (Database issue), D501–D504.

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

Rinn,J.L., Euskirchen,G., Bertone,P., Martone,R., Luscombe,N.M., Hartman,S., Harrison,P.M., Nelson,F.K., Miller,P., Gerstein,M., Weissman,S. and Snyder,M. (2003) The transcriptional activity of human Chromosome 22. *Genes Dev,* **17** (4), 529–540.

Royce,T.E., Rozowsky,J.S., Bertone,P., Samanta,M., Stolc,V., Weissman,S., Snyder,M. and Gerstein,M. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet,* **21** (8), 466–475.

Schadt,E.E., Edwards,S.W., GuhaThakurta,D., Holder,D., Ying,L., Svetnik,V., Leonardson,A., Hart,K.W., Russell,A., Li,G., Cavet,G., Castle,J., McDonagh,P., Kan,Z., Chen,R., Kasarskis,A., Margarint,M., Caceres,R.M., Johnson,J.M., Armour,C.D., Garrett-Engele,P.W., Tsinoremas,N.F. and Shoemaker,D.D. (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol,* **5** (10), R73.

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory,* **13(2)**, 260–267.

Youden,W.J. (1950) Index for rating diagnostic tests. *Cancer,* **3** (1), 32–35.