Bayesian Model Based Approaches In The Analysis Of Chromatin Structure And Motif Discovery

by
Riten Mitra

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2010

Approved by:

Dr P.K.Sen, Advisor

Dr Mayetri Gupta, Committee Member

Dr Joseph Ibrahim, Committee Member

Dr Fred Wright, Committee Member

Dr Jason Lieb, Committee Member

ABSTRACT

RITEN MITRA: Bayesian Model Based Approaches In The Analysis Of Chromatin Structure And Motif Discovery.
(Under the direction of Dr P.K.Sen.)

Efficient detection of transcription factor (TF) binding sites is an important and unsolved problem in computational genomics. Recently, due to the poor predictive ability of motif finding algorithms, along with the recent proliferation of high-throughput genomic technologies, there has been a drive to utilize secondary information, such as the positioning of nucleosomes, for improving predictions. Nucleosomes prevent transcription factor binding at those sites by blocking the TF access to the DNA. We aimed to construct an accurate map of nucleosome-free regions (NFRs), based on data from high-throughput genomic tiling arrays in yeast. Direct use of Hidden Markov Models are not always applicable due to variable-sized gaps and missing data. So we have extended the hidden Markov model procedure to a continuous time version while efficiently incorporating DNA sequence features that are relevant to nucleosome formation. Simulation studies and an application to a yeast nucleosomal assay demonstrate the advantages of the new method. The established biological role of nucleosomes in relation to TF binding, led us to formulate a joint model in the fourth chapter. The algorithm was implemented on the FAIRE data set, and comparisons were made with existing motif search algorithms. The fifth chapter deals with HMM asymptotics. We obtained results on consistency asymptotic normality and contiguity of a hidden Markov model. These have helped our inference on the convergence properties of the posterior and the consistency of the Bayesian posterior estimates. This has led to the conclusion that the Bayesian inference of a HMM run on sufficiently large datasets (which is typical, in the case of genomic data) leads us very close to the underlying true parameters, as in the case of iid models. The result is fairly general in nature to provide the justification for HMM inference in a wide variety of datasets.

Contents

List of Tables

Chapter 1

Introduction

Research in computational biology over the past few decades has consistently supplemented and strengthened the field of genomics. It plays as much important a role in paving the the path to discoveries of great consequence, as the laboratory methods. The field has arisen out of a demand to address the new questions posed by the size and nature of the genomic data that has been in proliferation. These data arising out of new technological setup, has features and characteristics of its own. This, in turn, has led to framing innovative statistical methodologies and algorithms. Most of these algorithms are centered around a Bayesian paradigm, mostly in the field of motif discovery.
The growing acceptability of Bayesian methods in statistics applications can be partly attributed to the computational resources we have in hand today. With the help of these resources a Bayesian algorithm can help us make inferences from large complex models where the interrelationship between the components are often hierarchical. . However, the blind application of a Bayesian sampling scheme into a data model, without getting into the issues of the statistical features of the models, can be an issue of serious concern. The inference that we make from these models, is often subject to a set of assumptions that need to be fulfilled in order to guarantee issues of identifiability and proper posteriors.
The central problem that we deal with here is that of motif discovery in DNA. The DNA has short repetitive patterns present which are potential transcription factor binding sites and hence of a great biological significance. We have shown our methods to greatly improve on the shortcomings of earlier motif discovery models. This we have done by

trying to incorporate secondary information in the form of presence or absence nucleosomal structures. The secondary information has come up in . But our methods that we have implemented are just not based on merging of two data sources in coming up with more information . It goes beyond that in the sense that it implements the merger efficiently so that propagation biases from one prediction to another gets minimized. This issue has led us to come up with a suitable statistical model that jointly samples nucleosomes and motifs.

This modeling framework gives a biologically plausible structure to the data and forms one part of the picture. The other part comes from the set of issues originating from type of data that we are dealing with. The FAIRE technology that outputs the genome wide intensities has missing information inform of non contiguous probes. This must be dealt with first before we could possibly extend our nucleosome prediction models into a joint model. The first chapter is concerned with implementation and evaluation of such a model. We have been able to incorporate sequence features directly into the link function of a model to see if sequence features has an influence on the length of nucleosome and the propensity of nucleosome occurrence (measured by intensity data) Having satisfied ourselves that this framework produces remarkably favourable results on comparison with known data bases, and provides a significant improvement over the currently used HMM methods on modelling accurately the location features. we have been encouraged to build up on it and venture into the joint model described above. The joint model instead of one has two layer of hidden states, one for the nucleosome and one for the data structure. The data sources come in the form of nucleosomal intensity data and sequence data, obtained from Chip-Chip experiments and microarrays. The recursive forward algorithm that we implemented for the first chapter gets extended into a modified algorithm with two running indices for the two hidden states.

The inferences both from the joint model and the nucleosomal model (which consists of a single layer of hidden states) is aided by MCMC sampling from the posterior. The posterior mode here is actually the value attaining the maximum likelihood, when the prior is flat.The flatness of the prior distributions do not pose a problem since there arises no necessity of making the posterior proper. We get finite integrals when we integrate the likelihood kernel with respect to the parameters. On the other hand it helps us having a

non informative background on the parameters to be inferred, and make our inferences solely based on the data alone. The flat prior assumption essentially enables us to analyze the likelihood function and its properties through MCMC sampling techniques. So our inference on Maximum Likelihood Estimation of a a continuous time hidden markov model gets directly translated into making inferences on the posterior mode of the Bayesian model. By our conclusions in the third chapter, we have shown that the this mode is a consistent estimator. That is , under large sample size conditions the posterior mode is very close to the true parameter values. This is a very important and immensely relevant result considering the size of the genomic data that wee handle. In our application the sequence size runs in arange of 340000-500000 base pairs. Taking the mode of our posterior samples from MCMC would give us a very accurate prediction of the parametrs governing the underlying hidden markov process, and we would be able to predict the DNA structural positions to a very high degree of accuracy. This result has scopes beyobnd the applications that we have considered in the first two chapters and can find useful application in numerous classification and optimization problems in genomics which are based on likelihood approach .

We have taken the techniques of analyzing the HMM likelihood one step further and proved the asymptotic normality of the score function and the maximum likelihood estimates. This enabled us in building asymptotic confidence intervals of the sampled posterior mode of the parameters. By examining the position of the null with respect to the upper and lower bound of the interval we assessed the significance of the parameters. The fascinating thing about this result is that we no longer have to depend on the simulation sample sizes in order to get an estimate of the interval length, but we could directly obtain it from a single round of computations from the complete data likelihood. The MLE results and consistency are based on simple assumptions on the boundedness of transition parameters. By the biological constraints on the length of the nucleosome free regions and the nucleosomes, these conditions are trivially satisfied.

This methodology gives us a statistically rigorous estimate of the distance of our estimates from the null distribution However in some cases we need to assess the power of our test statistics under suitable alternatives For example we need to know if for organisms which share a very close geneological relationship have similar or different

nucleosomal lengths. Or in the case of yhe joint model, if a weak link joining the motif strength and signal is existent. Sometimes we are dealing with weakly pronounced motif signals that are not very different from the background and hence is difficult to distinguish or pick up. Here the decreasing genomic distance on the phylogenetic tree between related species, or the link value tending towards 0 or the weak motif signals, denote a sequence of alternatives tending towards the null. and testing procedures would necessitate deriving the distribution of power statistics under these conditions. This ahs been facilitated by the property of contiguity which we have been able to derive for the continuous time HMM. The contiguity property relied on asymptotic noramlity of the log likelihood and the asymptotic similarity of the observed and expected Fisher Information. The third chapter thus provides a repertoire of statistical tools to supplement our inferences in the first and the second. The document is organized as follows. The first chapter introduces the continuous hidden markov model and extends it to include sequence features. By implementing the model on a well known data set, we have been able get new insights into the relationships between nucleosomes and DNA characteristics. The second chapter of our document describes a nucleosome-motif model and quantifies the link between gene signal and motif strength. The third chapter deals with asymptotics of the models described above and provides the setup for testing hypothesis under biologically plausible local alternatives. The last chapter gives an overview of the body of work and points to the future directions in this area.
This document is organized into three main divisions. In Chapter 2, I have made a review of the research done in the area of motif search algorithms, nucleosome mapping,and asymptotic properties of Hidden Markov Models. I have depicted the importance of motif search algorithms in computational biology, its development since the beginning of the last decade, and its connection with nucleosome mapping. I have described in details, some of the main statistical methodologies that have dominated this field. In the course of the review, I have given a short overview of our proposed methods, and how they improve upon inadequacies of earlier models. The last section of the review deals with asymptotic results of Hidden Markov Models, which have an important role to play in connection with inferential procedures of our proposed models. For this section, I have described in details some of the main theoretical developments, giving weight to usage of

certain sophisticated theoretical tools in the proof of maximum likelihood estimation–which have been a recurrent theme in work done in this area. Within the same section, I have provided a short description of our proposed method, which attempt to extend the known results of discrete HMM to its continuous time version.
Chapter 3 deals with the first paper of our proposal. These sections give the motivation of our proposed method, describe in detail the biological background and data structure,provide an explanation of our estimation techniques, and show clearly the potential of our newly developed methodology to contribute significantly to current research.
The next major division, Chapter 4, constitutes the content of our second paper. It talks about a joint model that uniquely combines nucleosome finding, and motif search into a single Bayesian framework.
Chapter 5 talks about the asymptotic results and the contiguity arguments thereof.
The concluding chapter 6 deals with future directions of research.
The simulation tables and figures are provided in the Appendix.

Chapter 2

Literature Review

2.1 Introduction

The DNA sequence can be viewed as a series of letters taken from the set (A C G T), which correspond to the nucleotides Adenosine, Thymine, Cytosine and Guanin respectively. In reality, we do not have a single sequence, but a double helical structure, where one strand has the 'complementary' base, relative to another. Due to the chemical affinity formed from their structure, these nucleotide pairs form matching pairs. For example, if we take a subsection of one DNA strand and observe the sequence ACGT we can readily infer that the other strand has TGCA in the same positions. This complementarity allows us to focus our attention on a single DNA strand.
Since its discovery as the key constituent of cellular life in living organisms, attempts have been made to unearth the information contained in the genome. Advances in technology in the last two decades have made it possible to (a) make a map of the DNA sequence pattern and (b) reveal the physical structure of DNA. The DNA sequence, when considered alone, does not adhere to a regular deterministic pattern. Today,however, we know that there are subsequences in the DNA called genes, which are the key constituents in initiating the bio-chemical processes that lead to the formation of proteins in our cell, thus acting as the building block for all life processes in an organism. Transcription is the first step in the process. It is onset when the enzyme RNA polymerase binds to the upstream region of a gene. This enables the information from

the DNA sequence to be copied into mRNA. The information is then transferred from mRNA by the process of translation to form proteins. A special group of proteins, called the 'transcription factors' interacts with RNAp and regulate gene expression. They do so by attaching themselves to certain regions in the upstream section of a gene. The binding of transcription factors is a chemical process that depends on the binding energy of the interacting molecules,(the biology of which is not fully understood till date) and the chromatin structure of the region. Knowledge of the location of these binding sites would serve as a very useful guide in locating the functionally important regions of the genome, and hence ranks as one of the most important quests facing the field of bioinformatics today.

Experimental detection of the transcription factor binding sites remain an infeasible task. Hence, much of the work done in this direction, relies on statistical and numerical algorithms that exploit the quantitative differences in the base composition of the binding sites and the rest of the sequence. The binding sites are typically 8-20 base pairs long. One important property of these sites is that the sequence structure is more or less conserved across all binding sites corresponding to a particular transcription factor. Mismatches or some random deviations from a fixed pattern are, however, tolerated. The underlying fixed pattern is called a 'motif'.

Motifs are, thus, defined to be short repetitive (with some fuzzy mismatches) subsequences of unknown length within a DNA sequence. So now, the biological problem of finding the transcription factor binding sites translates to the computational and statistical problem of locating motifs in a sequence. It is expected that probabilistic models will have a big role to play in the motif search algorithms, since the patterns are repetitive with a degree of fuzziness or random error. The following section gives us an overview of the motif discovery models and methodologies.

2.2 Algorithms for motif discovery

One of the first approaches in motif search was a method called 'Consensus', developed by Stormo and Hartzell (1989). The main objective of the algorithm was to find motif sites across a set of unaligned sequences. A motif was designated by a 4 x k matrix,

where k is the motif length. Each column corresponded to a position in the binding site, and the frequencies in each column denoted the frequencies of occurences of different base pairs in that particular position. Initially all k length words (contagious subsequences) were chosen from one sequence . These words were equivalent to several 4 x k matrices where the first row is 1. In other words these initial choice of motifs were free of any random error components. Next, each of these matrices were compared to all k-length words from the next sequence. The comparison was quantitatively done using the Kullback- Liebler information score

$$I_{KL} = \sum_i \sum_j f_{ij} log \frac{f_{ij}}{f_{0j}}$$

where i ranged over the positions and j is the index of the base pair. For each matrix, the highest score was retained, and the k-word that gave rise to the score was added as a row in the matrix. The matrix was then normalized to the relative frequency scale. This was done progressively over all sequences and the highest scoring matrices were retained. Although computationally feasible, the above method lacked in a rigorous statistical framework that would allow us to test the significance of the motifs obtained. The scoring criterion, though having the intuitive appeal, was also chosen arbitrarily. This motivated researchers in the early 90s to come up with more rigorous stochastic models, that aimed in determining simultaneously the unknown matrix (the motif pattern) and the sites of their occurrence. Attempts to construct such a model was first taken by Lawrence and Reilly(1993). In their formulation, the sequences were allowed to have only one motif site . This was a restrictive assumption. Also, the model was based on the fact that there would be only one motif pattern prevalent in the entire set of sequences-not a valid assumption.

Both these conditions were relaxed, and a more complete version of the motif model emerged with the paper by Liu and Lawrence(1995) Liu et al. (1995). The model proposed by them, turned out to be the basic framework, on which further work was done. In the following sections we lay out the model setup and study in detail the inferential procedures.

2.2.1 Position Weight Matrix models

We have N different sequences with S_i of length L_i. For k different motif types, the k Position Weight matrices are denoted as $\Theta = \Theta_1, ...\Theta_d$. Each $\Theta_k = (\theta_{k1}, \theta_{k2}....\theta_{wk})$ where w_k is the width of the k^{th} motif type and $\theta_{ki} = (\theta_{ki1}, \theta_{ki2}\theta_{ki3}, \theta_{ki4})$ denotes the probabilities of the 4 bases to occur in the i^{th} binding position of motif type k. The corresponding probabilities for the distribution of sequences under the background (i.e in places not occupied by motifs) is denoted by $\theta_0 = (\theta_{01}, \theta_{02}\theta_{03}, \theta_{04})$.
The start position of the motif sites is indicated by the indicator variable
$A = ((A_{ijk})) = 1$ iff position j in sequence i, is the start position of the motif type k. A_k denotes the indicator variable for the start position of the motif type k, spanning all sequences. Let $S^i_{A_k}$ denote the set of letters occuring in position i of every instance of motif type k. If C(S) denotes the set of frequencies of all letters in a set S, then $(C(S^1_{A_k}), ...C(S^{w_k}_{A_k}))$ follows Product multinomial$MN(\theta_{k1}, ...\theta_{kw_k})$ i.e the i^{th} position in the motif type k, has column frequencies that follow a multinomial model with parameters θ_{ki}.
Using the notation $u^v = \prod_{i=1}^{p} u_i^{v_i}$ for vectors u and v in R^p the likelihood conditioned on the indicator matrix A is

$$P(S|\Theta, A, \theta_0) = \theta_0^{C(S^{A^c})} \prod_{k=1}^{K} \prod_{i=1}^{wk} \theta_{ki}^{C(S_i^{A_k})} \tag{2.2.1}$$

The prior assumed on θ_0 is Dirichlet(β_0) where $\beta_0 = (\beta_{01}.....\beta_{0K})$ and the corresponding Dirichlet prior is PD(B) for Θ_k where $B = (\beta_{k1}, ...\beta_{kw_k})$ is a 4 by w_k matrix. π denotes the probability that any position is a motif start site, i.e $\pi = P(A(ijk) = 1)$. The prior assumed on π is Beta(α, β) The posterior conditional probability of the data can now be written as

$$P(\Theta, A, \theta_0|S) = \theta_0^{C(S^{A^c})+\beta_0} \prod_{k=1}^{K} \prod_{i=1}^{wk} \theta_{ki}^{C(S_i^{A_k})+\beta_{ki}}, \tag{2.2.2}$$

where $L = \sum_{i=1}^{N}(L_i - w)$ and $|A|$ is the cardinality of the set A, or the number of positions covered by motifs.

2.2.2 Inference in Position Weight Matrix Models

Since the full conditionals are easily available, straight forward Gibbs sampling of the parameters is one possible approach. However for computational efficiency, Θ and π can be integrated out and sampling could be done directly from the posterior distribution P(A/S). The sampling of A would be then done using a Gibbs method, i.e. conditioned on the presence or absence of motif sites elsewhere in the set of sequences sequence, the probability of having a motif of type k, start at a particular position, is calculated. The position is then assigned to a possible motif start position with the calculated probability. Based on an approximation, Liu et al Liu et al. (1995) computed the following predictive formula

$$\frac{P(A(ijk)=1/S)}{P(A(ijk)=0/S)} = \frac{\pi}{1-\pi}\prod_{i=1}^{wk}\frac{\hat{\theta}_{kl}}{\hat{\theta}_0}^{C(S_{i,j+1,k})}$$

where $\hat{\theta}_{kl}$ and $\hat{\theta}_0$ are the estimated posterior means.
The above formula is intuitively appealing, since it shows directly the influence of the nucleotide counts of the motif models (the matrix column counts updated by the Bayesian parameters go into the posterior mean estimates). However the Gibbs approach becomes 'sticky', when the motif sites are abundant. (This is because we shall not be able to sample till after w positions of a selected motif site.)
One alternative to this would be to use a Data Augmentation technique. Under this approach we do not integrate out θ and π but conditioned on these parameters draw directly from the joint distribution of A. However, the conditional distribution of A does not have a closed form. To get rid of this problem, the algorithm of forward summation and backward sampling is implemented. This algorithm plays a key role in computational genomics and is widely used for models with unobserved hidden states.
The first step in the algorithm is constructing the forward sum- a sequence, which recursively computes the likelihood that the motif type k ends at a particular position j in the sequence. Let $F(j,k)$ denote the likelihood that the motif type k ends at position j, $k > 1$ and F(j,0) denote the likelihood till position j, when the j^{th} sequence comes from

background.

$$F(j,k) = \sum_{k\geq 0} F(j-w,l)\theta_k^{C(S_{j-w:j})}, k \geq 1 \quad (2.2.3)$$

$$F(j,0) = \sum_{k\geq 0} F(j-w,l)\theta_b^{C(S_{j-w:j})} \quad (2.2.4)$$

Backward sampling is then implemented to sample motif occurring sites from the end of the sequence. The rationale behind the backward sampling procedure is that the joint distribution of $P(A|S)$ can be written as the product of conditional distributions.
We initiate the backward sampling by first selecting the last position of the sequence as a motif ending site of type k with probability

$$\frac{F(N,k)}{\sum_{l\neq k} F(N,l)}$$

If the position is selected as a motif of type k, we move back w steps and sample at the N-w position.Else, we sample from the position N-1 with probability F(N-1, k). Generally for any position j, the probability of sampling a motif of type k is given as

$$\frac{F(j,k)}{\sum_{l\neq k} F(j,l)}$$

We continue the above chain of conditional sampling till we reach the beginning of the sequence. The beauty of the backward sampling is that the Markovian property makes the backward sampling probabilities only dependent on the forward probabilities, and not on the configuration of the previously sampled motifs.

2.2.3 Extensions

One extension to the above model is to have an unknown motif width w, with a suitable Poisson prior. However, in that case, we have a dimensionality change in θ. Reversible jump Monte Carlo can be implemented to circumvent this problem, or, as was discussed previously, θ can be integrated out. The width parameter,w, is then updated by a metropolis Hastings step. A more generalized form of the position weight matrix model

needed to be formulated for complex eukaryotes(e.g, humans, mouse) . This is because in such organisms, some potential problems can arise in the form of weak motif signals, sparseness of signals, (the binding site might be 2000 base pair away from gene), and the tendency of motifs to occur in clusters.
In order to accommodate these features, a more complete Bayesian model was proposed that incorporated the distance between the motifs(λ), a correlation transition matrix reflecting the probability of transition between adjacent motifs -(τ), and a variable list of putative motifs $D = (D_1, D_2, ...D_p)$.
D_0 denotes the background model.
The length,d, between the motifs is assumed to follow a geometric distribution with parameter λ truncated at w.
Let u denote a p-vector of binary random variables where $u_j = 1$ iff D_j is a part of the module. Thus number of non-zero entries in the vector u denote the number of motif types currently included in the model.
The prior distributions are as follows:

$$P(u_j = 1) = \pi \tag{2.2.5}$$

$$\lambda \approx Beta(a, b) \tag{2.2.6}$$

$$D_i = \prod_{j=1}^{k} Dirichlet(\beta_{ij}) \tag{2.2.7}$$

$$D_0 \approx Dirichlet(\beta_0) \tag{2.2.8}$$

The i^{th} row of τ follows $Dirichlet(\alpha_i)$. The multinomial parameters have closed form conditional distributions, as previously. The indicator matrix of motif positions, A, could be sampled through the data augmentation technique by first recursively computing the forward likelihood and then implementing backward sampling. The forward algorithm and backward sampling procedure needs to be adapted to the extended parameter settings.
However for sampling u, or in selecting the putative binding sites that go into the module, we encounter a problem, since the dimension of θ and τ vary according to the number of motif types. So θ and τ is integrated out analytically and the marginalized

posterior probability is given as

$$P(u/A,T,S) = \pi^{|u|}(1-\pi)^{p-|u|} \int P(S/A,T,D^u)P(D^u/u)d(D^u)$$

Evolutionary Monte Carlo is then applied to sample u. A set of temperatures $(t_1.....t)$ is chosen for each u_i . Now we define

$$\phi(u_i) = exp(log(P(u/A,T,S)/t_i)$$

A new configuration is now selected by the following steps:

1. Mutation: A new configuration v_k is chosen by randomly selecting u_k and changing. u_k is then replaced by v_k with the probability $min(\frac{\phi(v_k)}{\phi(u_k)}, 1)$

2. Crossover: Two configurations u_k and u_j are randomly selected and new units v_j and v_k are formed by a random exchange of their segments.

The two new units then replace the old ones with probability

$$min(\frac{\phi(v_k)\phi(v_j)}{\phi(u_k)\phi(u_j)}, 1).$$

Alternative models for motif discovery were being suggested since early 2000, that employed the concept of segmentation of the sequence data. Recursive formulae for computation of likelihood formed the basis of inference in such models. Below,we present a short overview of one such category of models.

2.2.4 Dictionary models

Bussmeker et al(2000) Bussemaker et al. (2000) introduced a dictionary model for motif discovery by making a novel formulation of the problem. The assumption was that Nature has dictionary of some words, which are sampled with replacement and are concatenated together to form the DNA sequence. The set of possible words is denoted by $(M_1, ...M_D)$ The sampling probability of the set of words is denoted by

$\rho = [\rho_{M_1}, ...\rho_{M_D}]$. H is the index for all possible segmentation of the sequence into words. $(H_1, ...H_k)$ represents the possible partition of the sequence into k words.
Let N(H) denote the number of words in H, $N_{Mj}(H)$ be the number of occurences of type Mj in the partition. The full likelihood of the data can be written as

$$P(S|\rho) = \sum_{H} \prod_{i=1}^{N_h} \rho(S_{H_i}) = \sum_{H} \prod_{j=1}^{d} [\rho(M_j)]^{N_{Mj}(H)}$$

The likelihood does not have a closed form, but its computation can be made feasible by using the forward recursive formula below. Let F_i denote the likelihood till position i. Then

$$F_i = \sum_{j=1}^{w} \rho(S_{i-j:i}) F_{i-j}$$

where w is the length of the longest word.
The maximum likelihood estimate of the parameters can be obtained through a Newton-Raphson algorithm. Alternatively, expectation maximization(EM) or Gibbs sampling can be used. The iterative procedure of estimation, introduced by Bussmeker et al,was to start off with an initial dictionary of only 4 single letter words (A, C, G,T) . At each stage of the dictionary, the maximum likelihood estimate of the word frequencies were computed. Based on these estimated probabilities, it was observed whether random concatenation of words were more probable than expected under the current model. These words were then added to the dictionary.
The above algorithm had two major drawbacks. One was that the algorithm would have the longer words consist of overrepresented segments, though that might not be true. The other was the fact that it overlooked the stochastic nature of motifs. An improved version of the above was presented in the paper by Keles et al(2003) [Keles et al. (2003)], where a probabilistic model was introduced to model the lengths of the first and last segments.
A substantial improvement to the dictionary model was done in Gupta and Liu(2003) Gupta and Liu (2003) where the concept of a stochastic dictionary was first introduced. A stochastic dictionary consisting of D words is equivalent to D Position Weight Matrices(introduced in the previous section). Each Position Weight Matrix Θ_k of width

w is represented as $(\theta_1, ...\theta_{wk})$. The occurrences of letters in motif sites corresponding to the same motif thus follow a product multinomial model. As in the notation defined in the earlier section, A_{ik} represents the indicator of the motif type k beginning at position i. Let $C = [C_{q+1}, ...C_D]$ represent the count matrices corresponding to the probability matrices. The complete data likelihood is

$$L(N, C, A|\Theta^D, \rho) = \prod_{l=1}^{D} \rho_l \prod_{k=5}^{D} \prod_{j=1}^{wk} \prod_{i=1}^{w} \theta_{ijk}^{c_{ijk}}$$

A Dirichlet distribution is assumed for

$$\rho \approx Dirichlet(\beta_0)$$

A corresponding Dirichlet prior PD(B)is assumed for $\Theta_k = [\theta_{1k}, ...\theta_{w_k k}]$ where B=$(\beta_1, \beta 2.....\beta_k)$ is a 4 x w_k matrix, with each $\beta_j = (\beta_{1j}, ..\beta_{4j})$ The posterior distribution of Θ_k is Product Dirichlet PD$(B + C_k)$ i,e the column counts of the k^{th} word are updated by the Beta prior coefficients.The posterior distribution of ρ is Dirichlet $(N + \beta_0)$. The complete likelihood till the i^{th} position, $L_i(\Theta)$ is computed from a recursive formula identical to the Bussemaker's model given above. Motif start sites are then conditionally sampled with the probability:

$$P(A_{ik} = 1|A_{i+wk}, \Theta) = \frac{P(S_{i:i+wk-1}, \Theta, \rho)L_{i-1}(\Theta)}{L_{i-1}(\Theta)}$$

Next, the stochastic matrix parameters and ρ is sampled from their conditional distributions, as given above.
The start sites and the motif parameters are sampled in this way till convergence. Then the number of words in the dictionary is incremented by 1 (D=D+1) and the algorithm starts from step 1.
A particular motif alignment is scored by the MAP criterion (Maximum A posteriori Probability). For a particular motif pattern A, the MAP is given by

$$\frac{P(A*|M_1)}{P(.|M_0)}$$

where M_0 denotes the model that the entire sequence is generated from background, while M_1 is the model that incorporates motifs. $P(A*|M_1)$ is the maximum of the posterior density of A. θ is integrated out in order to calculate the numerator and denominator. The MAP score was tracked along with the steps of the algorithm.

2.2.5 Relaxing the Product Multinomial assumption

The product multinomial likelihood obtained in the position weight matrix models was derived from the assumption that the nucleotide occurrences at motif positions were independent.
This constraint was first relaxed in Zhou and Liu(2003) Zhou and Liu (2004) . The motif binding positions were allowed to be correlated with each other, but only pairwise correlations were permitted. That is, a given position could be correlated with only one other position in the binding site.
The total model space, H is huge and contains H_m, the space of models having m correlations. A prior was put on the model spaces so as to penalize the models with larger number of correlations. Also, all models having the same number of correlations were assigned equal weight. That is

$$P(H_1) = P(H_2)$$

if H_1 and H_2 belong to H_m

$$\frac{P(H_1)}{P(H_2)} = Choose(w - 2m, 2)$$

i.e models that have larger correlations are given a weight inverse to the number of ways of inducing the additional correlations.
The prior probability for any $H \in H_m$ is calculated as proportional to $\frac{2^m}{(w-2m)!}w!$
The motif occurence probability is given as ρ. ρ is assigned a Beta (a, b) prior. The usual Dirichlet priors are imposed on the motif and background parameters.
The joint distribution of the data and the unknown parameters can be written as

$$P(S, A, |\theta_H, \pi) = P(H)P(\theta_H|H)P(\pi)P(A|\pi)P(S|A, \theta_H)$$

Sampling is done from the above distribution in the following steps.

1. Conditional on all other parameters, A is sampled through a Gibbs procedure.
2. ρ is sampled from its posterior distribution, Beta$(a + |A|, L - |A| + b)$
3. Motif parameters are sampled from their conditional distributions which are Dirichlet.
4. Updating of H directly is a problem, since the dimension of the motif parameters change with the cardinality of H.

So the motif parameters and ρ are integrated out from the joint posterior to give $P(S, A|H).P(H|S, A)$ is proportional to $P(SA|H)P(H)$.
H is now updated by a Metropolis-Hastings step. The new candidate H^* is obtained in the following two ways:
Addition: A pair of positions is randomly selected from $(w - 2m)$ positions of the motif and added to H.
Deletion: A pair of positions is selected from the m correlated pairs and removed.
We accept the new H^* with the probability min(1, r) where r is the ratio

$$\frac{P(H^*|S, A)T(H/H^*)}{P(H|SA)T(H^*|H)}$$

and T is the proposal density. A phase shifting step is implemented once in every 20 iterations of the algorithm to escape local modes.
Among the 95 sets of TF binding sites in which the algorithm was applied, 22 of them had the posterior probability of the most likely model about six times that of the PWM model . Simulation studies further showed that even with the same position weight matrices, the generalized weight model was able to pick up more true motifs than the independent model. This clearly showed that the independence assumption in motif positions will fail in detecting less conserved correlated sites. Barash et al in 2003 later extended the correlation structure to include all possible correlations by using a Bayesian Tree network.

Alongside the product multinomial hypothesis, we usually assume that the conservation is uniform over all regions in a motif. Again this is not biologically true. In reality, transcription factors act over certain segments of DNA bases at a time. A fragmentation model was proposed in Liu and Lawrence that allowed for a prior to be set up on the motif positions,reflecting the importance of each motif position. Kechris et al Kechris et al. (2004) used a prior distribution (normal or double exponential) that penalized deviations (absolute or squared)from the conservation profile. The parameters were updated using EM algorithm.
The methodologies discussed above were de novo, in the sense that they did not use any external information for finding motif sites. These methods suffer from the disadvantage of poor predictive ability and a high false positive rate. Recently, with the remarkable improvement of genome output technology, there has been a proliferation of genomic data that has motivated researchers to take advantage of external information and incorporate them in motif search. The most important sources of these external information arise in the form of

1. Evolutionary data. Genomic data for comparing different species are now available to us. The conservation pattern in the binding sites are directly dependent on the inter-species evolutionary distance.

2. Gene expression

3. Data on the underlying chromatin structure.

2.3 Using auxiliary information in motif discovery

2.3.1 Evolutionary conservation

When multiple sequence information is available it has been seen that multiply aligning sequences and using regions having a high sequence similarity increases the specificity of motif search. Wasserman et al Wasserman et al. (2000) generalized the species comparison to multi-cellular organisms. By this, they found that 98 percent (74/75) of experimentally defined sequence-specific binding sites of skeletal-muscle-specific

transcription factors are confined to the 19 percent of human sequences that are most conserved in the orthologous rodent sequences.

Based on the knowledge that regulatory regions are more conserved between human and mouse genomes than the background, Liu and Liu Liu et al. (2004) developed a method called Compare-Prospecter that extends Gibbs sampling by biasing the search in regions conserved across species. Recently, based on the interactions arising from clusters of TFBSs with known binding patterns, a variety of computational methods have been created for the discrimination of CRMs.

It is often the case that no prior information exists on binding patterns of any relevant transcription factors for sets of genes identified in large-scale expression studies. One approach is a method for identification of modules using known motifs, but it includes a preliminary step of motif identification using either a Gibbs sampling algorithm or an algorithm based on overrepresented oligonucleotide sequences One other approach uses suffix-trees and word consensus rather than a statistical model to locate ordered collections of motifs Marsan and Sagot (2000). In this method, sites of each motif type are assumed to occur exactly once in each module. An expectation-maximization algorithm based on a discriminant model with multiple iterative optimization steps has also been described Segal and Sharan (2005). Although these approaches are promising, computational identification of a TFBSs without prior knowledge of binding patterns remains elusive. Thompson et al Thompson et al. (2004) used the strategy of modelling neigboring interactions in synergy-model to improve predictions.

Phylogenetic analysis has also been used by Koch et al Koch et al. (2001) to determine conserved cis regulatory regions in certain plant species. Promoters were aligned by hand,phylogenetic distances were computed using a two-parameter model, and the resulting distance matrices were subjected to a neighbor-joining algorithm . However it remains challenge to efficiently incorporate phylogenetic trees in probabilistic motif models.

One major problem in using evolutionary approach is that cross-species comparison of sequences from orthologous genes, or phylogenetic footprinting, shortens the amount of sequence under consideration by focusing attention on conserved regions that are more likely to serve a biological function. Although such methods can increase binding-site

densities by fivefold, only the strongest sites are detected at this level.

2.3.2 Gene expression

Incorporating gene expression for finding regulatory modules has mostly relied on clustering genes based on their gene expression levels, and then finding motifs in the highly clustered genes. However, this technique might give rise to larger number of false positives due to spurious correlations. This happens because there can be genes in cluster without a motif, and all motifs in a gene might not respond. If the gene mechanism is multi factorial, then genes in separate clusters are not actually disjoint, and making them do so would lead to a loss of information. A filtering method was adopted to remove this inadequacy Hughes et al. (2000), however the algorithm sensitivity still remained low. Bussemaker et al Bussemaker et al. (2001) introduced a novel technique of modeling the relationship between gene expression and motifs . They came up with a linear model

$$A_g = c + \sum_{\mu \in M} F_\mu N_{\mu_g}$$

where A_g denotes the log ratio of the abundnace of gene expression between two cells, N_{μ_g} represents the number of occurences of motif μ in gene g. The F_μ coefficient measures the increment or decrement in the gene expression when a new motif is added. M denotes the set of motifs.

Both the gene expression and the number of motifs for each gene were centered around their mean and scaled by their variance. The model was then fitted to the data obtained from the chromosome sequence and ORF coordinates in the Sacchromyces genome data base.

Motifs were progressively included in the model based on their residual deviance. That is, based on a current set of motifs, the statistical significance of adding a new motif was judged by calculating the residual deviance of the model including the motif and the previous motif set. These residual deviances were ranked, and the largest one, which satisfied the significance criterion was added to the model. The significance criterion for the maximum residual deviance was found out from the extreme value distribution.

The model overall accounted for only 30 percent of the total signal present in gene

expression. However it performed well in the case of genes that had combinatorial effects in transcription regulation, i.e. where they co-varied in one circumstance but were different in another. It is in those cases that the previous clustering techniques failed.
Gene expression data was also used by Conlon et al Conlon et al. (2003a). First, MDscan was used to select motifs from a list of putative motif sites. All w mers in the top K sequences were used as seeds. For each seed, the w mers that shared at least m binding sites were used to construct motif matrices. These matrices were then scored using a Bayesian scoring function, and the 50 top scoring motifs were retained. W mers were then added from the top sequences to these motif matrices to improve the score . Motifs with average frequency of consensus bases less than .7 were removed. If several motifs of same width were similar upto m bases then the highest scoring one was retained. 30 motifs were reported at the final stage.
At the next step, the upstream sequence of gene is compared with a motif. The following scoring function was used to test how well a motif correlates with the gene expression

$$S_{mg} = log[\sum_{x_g} \frac{Pr(x|\theta_m)}{Pr(x|\theta_0)}]$$

where x_g denotes all w-mers in the upstream sequence, θ_m denotes the motif matrix and θ_0 is the background probability vector estimated from fitting a third order markov model to the intergenic sequence.
A simple linear regression is then performed to test the significance of the selected motifs. For each motif and gene, we have the following equation:

$$Y_g = \alpha + \beta_m S_{mg} + \epsilon_g$$

In the above equation, Y_g is log_2(gene expression data) and ϵ_g is the gene specific error term.
This model, though similar to the linear model used by Bussemaker et al, uses not just the number of motifs as the variable of interest, but the motif strength. (as quantified by the scoring function, obtained through position weight matrices.)
The significant motifs (pvalue for β_m less than .001) were then used for a multiple

stepwise regression.

$$Y_g = \alpha + \sum_m \beta_m S_{mg} + \epsilon_g$$

Initially only the intercept is added, and then motifs are added increasingly based on their residual error. The algorithm was designed to stop when no remaining motif met the criterion for entry.

Although the above methods are an improvement on the previous clustering techniques, the problem with them is that they unrealistically assume linearity, and are ill-suited to cope with problems dealing with multiple data sets and high dimensionality.

2.4 Statistical modeling of chromatin structure

2.4.1 The biology of Nucleosomes

The relationship between DNA flexibility, nucleosomes and the positioning of TFBS has recently been a hot area of research . Nucleosomes are compact chromatin structures which contain DNA sequence of 147 base pairs wrapped around a histon protein octamer. By this,they prevent RNA polymerase, regulatory proteins and other recombination complexes from acting on the genomic sequence at those positions. Hence the nucleosome-wrapped DNA would be expected to be free of transcription sites. So, the nucleosome positioning information would be a very useful guide in helping us locate the functionally important genomic regions.

The sharp bending of the nucleosomal sequence occurs at around 10 bp helical repeat of the DNA when the major groove faces outward and again at 5 bp away, when the major groove faces inward. Bends of each direction are facilitated by specific dinucleotides. Neighboring nucleosomes are separated by 10-50 bp long stretches of linker DNA, so about 75-90 percent of DNA is occupied by nucleosomes. High resolution images of nucleosomal DNA structure have suggested that the genome has sequence-specific conformational abilities, in the form of roll and twist angles, that enable its wrapping around the nucleosome. Typically AA TT signals have higher roll angles. Dinucleotides,overall, have very high correlation between roll and twist angles. The ability of the DNA sequence to bend around the histone octamer varies greatly, and this variance

is dependent on the sequence composition. The main question is whether the DNA sequence influences the positioning of nucleosomes across the genome in a way such that the ability of transcription factors to access binding sites is either inhibited or increased. Nucleosomes are being increasingly relied upon as the most important guiding tool in helping us determine a complete picture of gemome transcriptional activity. The biological role of nucleosomes is directly related to transcription, (in a way which we shall see, in the following section), and hence it would serve as a more powerful predictive tool than the information provided by comparative genomics and gene expression. The latter approaches definitely add to the knowledge provided by sequence information alone, but the prediction uncertainties associated with these methods place them in a rank lower to nucleosomes, in the context of motif search.

2.4.2 Nucleosome prediction algorithms

One of the first experiments to identify nucleosome positions was performed by the Yuan et al group Yuan et al. (2005). Based on the linker DNA's susceptibility to micrococcase nuclease, nucleosomal DNA was isolated and labeled with CY3 (green) fluorescent dye while the total genomic DNA was labeled CY5(red). These segments were then hybridized in microarrays printed with 50 mer oligonucleotide probes tiled at 20 base pairs across chromosome 3 of yeast. The log hybridization ratios(green to red) were considered as the signal of interest. A two state hidden Markov model was implemented on the data signal to classify the genomic region of interest into nucleosomal and non nucleosomal states. Regions classified as nucleosomes and having state length of 140-147 base pairs were labeled 'well positioned' whereas longer nucleosomal regions were labeled as 'delocalized' nucleosomes. It was observed that highly enriched genes were surrounded by regions of delocalized nucleosomes. This is supported by the biological reasoning that the action of RNAp causes nucleosomes to be disrupted and assemble behind the coding regions. Sequence conservation patterns were examined among related species of yeast. It was found that coding regions were always highly conserved. Intergenic sequences in nucleosomal regions were poorly conserved whereas, NFRs (Nucleosome Free Regions) were highly conserved.One interesting feature was that the conservation was not limited only to the coding regions. The conservation regions included stretches of poly A-dt,

along with TFBS. Also, the location of the NFRs had high correlation with the poly A-dt stretches indicating a possible role of these molecules in the formation of nucleosome free regions. The results of the prediction were compared with the existing databases of transcription factor motifs. 47 percent of unbound motifs were identified in linker sequences, whereas, 87 percent of motifs that are associated with transcription factors were found to be depleted of nucleosomes. Nucleosome free regions were identified as 51 percent of unbound motifs found on the array. This suggested that NFRs were transcriptional start sites, and a RNA hybridization experiment showed that indeed the 5' ends of the coding regions coincided with NFRs.
Segal et al(2007) Segal et al. (2006) also constructed a map of the nucleosomal positions in yeast genome, based on the statistical distribution of dinucleotides in nucleosomal DNA. Their approach was the first attempt to incorporate dinucleotides in nucleosome prediction. The statistical distribution was obtained from aligning a set of nucleosomal sequences from log phase yeast. Mononucleosomes were extracted by standard methods and protected fragments of length 147 bp were sequenced and cloned . These sequences and their reverse complements were then aligned about their centers. Next, dinucleotide frequency counts were obtained from each position. A moving average spanning three neighboring positions was then applied on the frequency counts to give a smoothed version of the frequencies for every position.From these frequencies, the empirical statistical distribution of dinucleotides for the nucleosomal sequence was computed from the given dinucleotide frequencies. The background distribution, i.e the distribution for a nucleosome free position was assumed to be a multinomial distribution, where each mononucleotide has equal frequency. Based on this empirical distribution, a two state constrained hidden Markov model was formulated. The nucleosome lengths were fixed at 147 bp, and at least 10 bp distance was allowed between adjacent nucleosomes. The set up resembled a motif model, with the only differences being

1. A dinucleotide distribution (obtained empirically) was used instead of a product multinomial distribution in the nucleosomal states

2. While both the parameters and the state structure is unknown in the case of motif models, in this case, the emission distribution is known.

The unconditional likelihood of the data is given as

$$L = \sum_{A} P_b[S_1, S_{c[1]-1}] \prod_{i=1}^{k} P[S_{c[i]}, S_{c[i]+146}] \prod_{i=1}^{k} P[S_{c[i]+147}, S_{c[i+1]-1}] P_b[S_{c[k]+147}, S_N]$$

where A represents all hidden state configurations, c[1],c[2]...c[k] represents nucleosome start positions, P_b is the distribution for background sequence, while P is the distribution in the nucleosomal state.
A forward algorithm is implemented as follows

$$F_0 = 1 \tag{2.4.1}$$

$$F_i = F_{i-1} P_b[S_i] \text{if} 1 \leq i \leq 146 \tag{2.4.2}$$

$$F_i = F_{i-1} P_b[S_i] + F_{i-147} P[S_{i-146}, S_i] \text{if} i > 147. \tag{2.4.3}$$

Similarly the the likelihood of S_N to S_i is calculated as follows

$$R_{N+1} = 1 \tag{2.4.4}$$

$$R_i = R_{i+1} P_b[S_i] \qquad i \geq N - 145 \tag{2.4.5}$$

$$R_i = R_{i+1} P_b[S_i] + R_{i-147} P[S_i, S_{i+146}] \qquad i \leq N - 146 \tag{2.4.6}$$

The probability of a nucleosomal position beginning at i is calculated as

$$P[i] = \frac{F_{i-1} P[S_i, S_{i+146}] R_{i+147}}{R_1}$$

The probability that a position i is covered by a nucleosome is given by

$$\sum_{k=0}^{146} P[i-k]$$

Stable nucleosomes were used to denote all those positions where $P[i] > .5$.
54 percent of predicted stable nucleosomes were found to be within 35 bp of the positions found in earlier literature. 68 percent of depleted coding regions, and 57 percent of

depleted intergenic regions had low predicted nucleosome occupancy. Together, the results indicate that approximately 50 percent of the nucleosome organization can be predicted from the nucleosomal sequence information.
Not taking into account the sequence information of the nucleosome free regions, does not make a strong case for the strength of the above classification approach, and is a potential reason why we do not get a high predictive value.

2.4.3 Relationship with sequence features

The relationship between DNA flexibility, nucleosome positioning and the sequence was more quantitatively investigated by Lee et-al(2007) Lee et al. (2007). It was found that the CTG trinucleotide correlates well with nucleosome occupancy while poly A-dT correlates negatively. A lasso model was constructed to bring out a linear relationship between the tip, tilt, twist angles, the sequence composition, and the transcription factor binding sites. Propeller twist capacity emerged as the most significant variable, (dinucleotides having highly negative propeller twist angles were found to be more rigid than those having lower negative values) The AAAA tetranucleotide was found to contribute to the rigidity of DNA conformation.
Nucleosomes were mostly found in coding regions and centromeres whereas nucleosome depleted regions corresponded with intergenic regions, most of which were promoter regions. In fact nucleosome free regions marked the boundary of transcription start sites, in that they were found just upstream of these start sites. A ladder of well positioned nucleosomes was found start from the beginning of transcription start sites. The nucleosome depleted regions also corresponded well with the occurence of transcription factor binding sites, which are clustered 100 bp upstream of the transcription start sites.126 known transcription factor binding sites and their position specific weight matrices were taken, and Wilcoxon mann whitney pvalues were computed for nucleosome occupancy. A strong relationship between nucleosome occupancy and the occurence of TFBs was observed. Four main clusters of coding regions were identified based on their nucleosome occupancy signature. Interestingly, these regions corresponded to four different gene regulatory functions.
Some recent nucleosome prediction methodologies were investigated by Peckham et al

(2007)Peckham et al. (2007) and Yassour-et-al(2008)Yassour et al. (2008). Yassour et al extended the HMM technique employed by Yuan-et-al. They implemented a hidden markov model with multiple states, in order to account for unstable nucleosome position and to adjust for global trends. Peckham-et-al employed a support vector machine for classification of nucleosomal and non nucleosomal states. They tested the efficiency of classification by using ROC curve. Their results were similar to that of studies, where G-C sequence features were found to be positively correlated with nucleosome formation, and A-T appeared to have inhibitory influence.

In 2008 Yuan and Liu Yuan and Liu (2008) used wavelet decomposition to represent the sequence signal and then used them as covariates in a logistic model. Based on the predictions from this model, they came up with the N -score, a statistical measure that quantifies the the tendency of a sequence to be occupied by a nucleosome. Below, we give a short sketch of the derivation of the N score.

Construction of the N score: 199 Nucleosomal DNA sequences were determined experimentally. They were then aligned by both forward and reverse strand as in Segal et al. Each nucleosomal DNA was about 145-153 base pair long. 296 NFR sequences were determined using a tiling micro array. Each NFR sequence had a 100 bp long linker region. From both the nucleosomal and NFR sequences, the central 131 bp was retained for statistical analysis. The linker sequences which were shorter than 131 bp was expanded symmetrically on both sides. The reverse strands were also added to the data set.

Next, for each sequence and each dinucleotide d, a 0-1 vector of length 130 is set up. The j^{th} component of the vector denotes the indicator of whether the dinucleotide is in j^{th} and the $(j+1)^{th}$ position. From this, a vector of length 128 is computed by averaging over the entire sequence. The i^{th} element in this vector is denoted by $f_{s,d}(i/128)$

The crucial step in the construction of N score comes in the wavelet transform of each of these dinucleotide frequency signals. Each of these signals is written as a linear combination of wavelets is the following manner.

$$f_{s,d}(i/128) = \sum_{jk} c_k^j(s,d)\psi_k^j(i/128) + c_0(s,d)$$

where ψ_k^j is a Haar wavelet function defined as follows:

$$\psi_k^j(x) = 1 \qquad 0 < x < 1/2 \tag{2.4.7}$$

$$= -1 \qquad 1/2 < x < 1 \tag{2.4.8}$$

$$= 0 \text{ otherwise} \tag{2.4.9}$$

The wavelet transform coefficients $c_k^j(s,d)$ can be obtained as in the coefficients for an orthonormal basis. (Note that $\psi_k^j(x)$ construct an orthonormal basis for a 0-1 valued function)

$$c_k^j(s,d) = \sum_i f_{s,d}(i/128)\psi_k^j(i/128)$$

The energy of a signal at level j is defined as

$$E = \sum_k [c_k^j]^2$$

It represents the variance of the signal at 2^{7-j} level. For 8 dinucleotides, we have 8 energy values for a given signal.
A logistic regression is now performed where the response is the indicator function of being a nucleosomal sequence, and the covariates are the 8 energy signals.

$$log\frac{P(s)}{1-P(s)} = \beta_0 + \sum_l \beta_l x_l(s)$$

where $x_l(s)$ is the energy of the l^{th} dinucleotide in sequence s. A stepwise procedure is then implemented to retain only the important covariates. The N score for each sequence is defined as the predicted logit from this model.
The N score model was validated by a two fold cross validation.The linker and nucleosomal sequences were divided into two groups, such that there are equal proportions of both types in each group. While one dataset was used to train, the other was used for testing. This was repeated five times. A ROC curve was obtained and the area under the ROC curve, i.e, the ROC score was used to tests its performance against the other models. One surprising result was that N scores derived from yeast data

matched quite well with the human genome data indicating that sequence specificity of nucleosomes may be conserved across eukaryotes.

Segal's model with a ROC score of .67 performed poorly against the N-score model. But the poor performance of the model was due to the fact that it did not use a discriminative approach for nucleosome prediction, as discussed earlier. It relied only on the properties of the nucleosomal sequence data for prediction. The ROC score for the support vector machine model was .82, insignificantly lower than the N score model. When the linker sequence data was incorporated into the Segal model, the ROC score shot up to .81, very close to that of the N score model. This showed that the superiority of the N score model was not derived from its use of wavelets, but in the fact it was the first nucleosomal prediction algorithm that use discriminative approach.

Poly Da-Dt tracks have been found to be correlated with the occurence of nucleosome free regions. The N scores helped to verify the result once again. Relationship between the N score at poly Da-Dt loci and the length of the ploy Da-Dt run was investigated. A negative correlation of .15 (pvalue less than .001) was obtained, confirming the experimental evidence, that poly Da-Dt tracks are depleted of nucleosomes. From the results of the logistic regression, it was seen that TT/AA/TA were the most important predictors, followed by TA/AC/GT, while GC had moderate predictive power. These results match with the results obtained from earlier studies. One interesting result was that the important predictors were related with nucleosome depletion rather than nucleosome formation. This suggested that the main role of the sequence information lay in determining boundaries of the nucleosome free regions.

Distribution of N score was compared at the transcription factor binding sites, unbound motif sites, and the genomic background. The average N score over the transcription factor binding sites is -1.03 which is less than that in the unbound motif sites(.70) and the genomic background (.77).

An interesting property of nucleosome occupancy investigated by Struhl et al Sekinger et al. (2005), was that deletion of sequence elements from promoter region increased DNA accessibility. In order to test this hypotheses, a computational experiment was performed. From every set of promoter regions, DNA sequences were progressively removed, starting from 10 bp. At each step 10 was further removed, and this was

continued till 200 bp. The change of n score was calculated for every deletion. A positive correlation of .48 ($pvalue < .001$) was obtained between the length of deleted sites and the N score. This result had motivated us to incorporate sequence features into our model for nucleosome prediction.
The nucleosome models discussed so far, however,leave sufficient room for significant improvements . In the following section, we discuss them and also suggest the approach that we took in this regard.

2.4.4 Mapping positions and its connection with motif search

From the previous approaches to nucleosome prediction, one pattern is clear. Most of the algorithms for nucleosome prediction aimed at classifying the nucleosomal states based on gene expression . After the prediction, the relationship between the predicted nucleosomal states, and other pertaining features such as presence of promoters, sequence features etc was investigated (by logistic regression as in Yuan (2008), or by Lasso regression in Lee(2007), or, mostly by just comparison of the statistical measures of these features in the two states). Also, the sequence features were limited to dinucleotide counts.
What has not been attempted so far, is the incorporation of the sequence features into a hidden markov model for improved classification. This is what we have implemented next,in our work. We have extended the above continuous time model in two ways:

1. Make the lengths of the states dependent on the sequence covariates. We did this by modeling the transition rates as a function of these covariates.

2. Make the emission means dependent on the covariates.

Selection of covariates is one issue of concern. Each probe has a number of non overlapping base pairs. We need to efficiently extract those features from the sequence set that contributes most to the signal. Earlier work in this area have mostly restricted themselves to di-tri-nucleotides. Here, we extend our feature set to include all oligomers upto length of 4. Principal components analysis is then performed on this set, and the top scoring variables retained for analysis.
Given an accurate prediction of the nucleosome states, one of its chief purposes, is to utilize this information in improving the sensitivity and specificity of the current motif

search algorithms. One way to achieve this would be to first obtain a prediction of the nucleosome positions, and then use it as a prior in motif search. We have attempted to formulate an unique joint model that predicts nucleosome positions and motifs simultaneously, based on the gene expression and sequence data. The joint model can be represented by the following components:

1. The indicator variable of motif start positions, similar to that in PWM
2. The hidden nucleosomal states.
3. Gene expression data.

Nucleosomal states are not allowed to have any instance of a motif. The probes in different nucleosomal states have different baseline means. Further in regions where a motif is present, the baseline mean gets incremented by a quantity, that reflects the strength of the motif (a function of the ratio of PWM probabilities to background probabilities) in that region. It is to be noted that due to the last assumption, the PWM parameters fail to have a Dirichlet prior. Metropolis hastings algorithm, needs to be thus suitably adapted for efficient estimation of these parameters.
Earlier,we have encountered models, where the relationship between gene expression data and sequence features, nucleosomal states and gene expression, gene expression and nucleosomal states were separately explored. In our improved algorithms for nucleosome expression, we have combined all three inter-relationships into a single hidden markov model.
The formulation of the hidden markov model is an important step with respect to setting up the combined approach, and also with respect to extending the model to continuous time frame. For this we need to have an overview of the asymptotic results, which we shall begin in the next section.

2.5 Statistical inference in hidden Markov models

From the previous sections, we have seen that hidden markov models (HMM, in short)have been the standard method of modeling sequence data. In the later sections we

shall see how we require extensions of such model to the continuous time case, in order to take into account the spatial and temporal lag between consecutive measurements. We also need a strong theoretical framework for the assessment of issues such as convergence rate, and identifiability in joint nucleosome-motif models, where there are two layers of hidden variables. The next subsection provides a brief overview of some of the theoretical work done in this context, where we discuss some of the major results in the area, and more importantly, some interesting techniques and properties used in the derivation of these results.

The main idea behind the hidden model framework, is that the observed sequence Y is in fact, dependent on an unobserved sequence X, which is a markov chain, (having discrete states typically). Conditional on X_k, Y_k are independent, and the density of Y depends upon the state of X. This density is generally referred to as 'emission density'. The markov chain transition probability q(x,x') is usually assumed to be ergodic. (This is a direct result of the irreducibility and aperiodicity of the transition matrix).

For example, we can have two unobserved states, 0 and 1. The transition probabilities are given as

$$\lambda = q(0,0) \tag{2.5.1}$$

$$\mu = q(1,1) \tag{2.5.2}$$

$$1-\lambda = q(0,1) \tag{2.5.3}$$

$$1-\mu = q(1,0) \tag{2.5.4}$$

If the state X is 0, Y follows ($N(\mu_1, \sigma_1)$. If the state X is 1,, Y follows $N(\mu_1, \sigma_1)$. The objective of inference is to produce estimates of the emission parameters,$\mu_1, \mu_0, \sigma_1, \sigma 0$, and the transition parameters, λ,μ. In the context of properties of the given sequence data, the transition probabilities can be thought of as modeling the lengths that the sequence spends in each state. We shall see, that in certain situations, such as the proposed joint nucleosome-motif model, it would me more convenient to model the lengths directly rather than in terms of transition probabilities.

2.5.1 Previous theoretical results: General assumptions and an overview

Theoretical results regarding the asymptotic properties of the MLE had started developing since early 1990s. The first major paper, in this direction, was by Leroux(1992) where consistency of MLE was established. Bickel and Ritov proved local asymptotic normality in 1996, and Ryden et al proved asymptotic normality under general conditions in 1998 . This result was later extended to autoregressive HMM models by Douc et al(2000). In most of these papers, an interesting technique of 'conditioning on the infinite past' and the principle of 'uniform forgetting' was used. We shall illustrate the methodologies in the following section. First, let us introduce some notations and assumptions.
Since the results have been extended to compact X, we will assume the most general form of notations: θ denotes the parameters indexing the transition probability and the emission density. q_θ is the transition probability function and g_θ is the emission density. X is the set of all hidden states, Y is the set of all possible observed values y. The assumptions are :

1. For all (x,x') $\in$ (X, X) $q_\theta(x, x'), g_\theta(x, y)$ is twice continuously differentiable.

2. $$sup_\theta sup_{x,x'} ||\nabla_\theta log(q_\theta(x, x'))|| < \infty$$
$$sup_\theta sup_{x,x'} ||\nabla^2_\theta log(q_\theta(x, x'))|| < \infty$$

3. $$E_\theta[sup_\theta sup_{x,x'} ||\nabla_\theta log(g_\theta(x, Y_1))||^2] < \infty$$
$$E_\theta[sup_\theta sup_{x,x'} ||\nabla^2_\theta log(g_\theta(x, Y_1))||^2] < \infty$$

4. The transition probability satisfies
$$0 < \sigma^- < q_\theta(x, x`) <= \sigma^+ \forall x, x' \in X, \forall \theta$$

5.

$$\forall y \in Y, 0 < \int_X g_\theta(x, Y_1) < \infty$$

In the following subsections, we shall try to give an outline of the proof of consistency and asymptotic normality established by the papers of Ryden et al in 1996 and 1998

2.5.2 MLE Results:Consistency

Here we shall see, how the ergodic properties have been made use of in establishing consistency results.
Since (x_k, y_k) is ergodic, it is possible to extend it to a doubly infinite sequence$(x_k, y_k)_{k=-\infty}^{\infty}$. This is the crucial step in the proof of the asymptotic results. Let us denote the log likelihood of the data till position n by $l_{x0,n}(\theta)$. This can be written as

$$\sum_{k=0}^{n} \int log[g_\theta(x_k, y_k) P_\theta(x_k \in dx_k | Y_{0:k-1})(dx_k; \theta)]$$

It would be convenient if the ergodic properties of $l_{x0,n}(\theta)$ could have been applied to obtain stationary martingale increments, and then the central limit theorems on martingales could be used However that is not the case. $l_{x0,n}(\theta)$ is not ergodic . So the trick is to approximate it by a function $l^s_{x0,n}(\theta)$ which is very similar to it, except that it is conditioned on past observations all the way till -∞, i.e

$$\sum_{k=0}^{n} \int log[g_\theta(x_k, y_k) P_\theta(x_k \in dx_k | Y_{-\infty:k-1})(dx_k; \theta)]$$

The results of the MLE are now obtained in the following three steps.

1. The ergodic properties of $l^s_{x0,n}(\theta)$ are established,leading to a unimodal function $l(\theta)$ which is the ergodic limit of $n^{-1}[l^s_{x0,n}(\theta)]$.

2. It is formally shown that $l^s_{x0,n}(\theta)$ and $l_{x0,n}(\theta)$ are very similar asymptotically and this was used to show that the latter converges to $l(\theta)$. We then show that the convergence is uniform.

3. The general argument for proving consistency in iid models, is used based on 2.

A finite approximation to $\int log[g_\theta(x_k, y_k)P_\theta(x_k \in dx_k|Y_{-\infty:k-1})(dx_k;\theta)]$ is obtained by conditioning till -m. We denote this by

$$h(k, m, x, \theta) = \int log[g_\theta(x_k, y_k)P_\theta(x_k \in dx_k|Y_{-m:k-1})(dx_k;\theta)]$$

It is then shown that $h(k, m, x, \theta)$ is a uniform Cauchy sequence, i.e

$$|h(k, m, x, \theta) - h(k, m', x', \theta| \leq \frac{\rho^{k+m-1}}{1-\rho}$$

where $\rho = \frac{1-\sigma^-}{\sigma^+}$. (This result is obtained from assumptions (4) and (5), whereby it can be shown that the total variation of $\int P_\theta(x_k \in dx_k|Y_{-\infty:k-1})(dx_k;\theta)]$ is less than ρ^k). The uniform Cauchy property of $h(k, m, x, \theta)$ leads to the result that this sequence has a limit independent of x, denoted by $h(k, \infty, \theta)$. Now since

$$l^s_{x0,n}(\theta) = \sum_{k=0}^{n} h(k, \infty, \theta)$$

and

$$l_{x0,n}(\theta) = \sum_{k=0}^{n} h(k, m, x0, \theta)$$

the Cauchy property produces a bound on $|l^s_{x0,n}(\theta) - l_{x0,n}(\theta)|$.
If $E_\theta[h(k, \infty, \theta)]$ is denoted by $L(\theta)$ (note that it doesn't depend on k, since $h(k, \infty, \theta)$ is ergodic) then by ergodic theorem $n^{-1}[l^s_{x0,n}(\theta)]$ converges to $l(\theta)$. From this it can be shown that $l_{x0,n}(\theta)$ also converges to $l(\theta)$. The fact that the convergence is uniform is obtained from the continuity properties of the functions stated in the assumptions.
The identifiability properties of HMM, which have been proved in 1967Teicher (1967),is next utilized to show that $l(\theta)$ has a global unique maximum. Let the maximum be

denoted by (θ_*). If $\widehat{\theta_n}$ denotes the MLE then

$$0 \leq l(\theta_*) - l(\widehat{\theta_n}) \tag{2.5.5}$$

$$\leq l(\theta_*) - n^{-1}l_n(\theta_*) + n^{-1}l_n(\theta_*) - n^{-1}l_n(\widehat{\theta_n}) + n^{-1}l_n(\widehat{\theta_n}) - l(\widehat{\theta_n}) \tag{2.5.6}$$

$$\leq 2sup_\theta|(l(\theta) - n^{-1}l_n(\widehat{\theta_n})| \tag{2.5.7}$$

This shows that for any compact subset

$$l(\theta_n) \to l(\theta_*)$$

as $n \to \infty$. Since l is continuous, this shows that the MLE converges to θ_* almost surely. The consistency result was proved by Leroux (1992). Asymptotic normality was established by Bickel Ritov Ryden(1998). Although they did not explicitly use the 'conditioning on infinite past' approach,the argument was built along those lines.

2.5.3 Alternative Estimation Methods

The consistency and asymptotic normality of the MLE was established , not before, 1998 . Before that, a number of alternative likelihood methods were suggested and their properties were studied under mild regularity conditions. The asymptotic normality of Maximum Split Data Likelihood methods was proved by Ryden (1996). A number of recursive estimation methods were suggested,by LeGland and Mevel (1997), e.g the RMLE (recursive maximum likelihood estimate).
A novel estimation method was suggested by LeGland and Mevel (2000) where by writing the log likelihood as the summation form discussed in the previous section, they were able to obtain recursive formulae for the score and observed information. (This result was later extended to compact spaces by Douc and Matias in 2002.) This estimation procedure was based on what is known as the 'geometric ergodicity property of an extended HMM, which we shall establish below.
Let p_n denote $P(X_n|Y_1, ...Y_n)$. An extended chain $[X_n, Y_n, p_n]$ is constructed.
p_{n+1} can be expressed in terms of p_n by the forward equation. Under ergodicity, there will be a solution to the equation, which would not depend on the initial conditions. So

the first step in establishing ergodicity is to get a bound on the difference between two solutions under two different initial conditions. The difference measure and its functions are defined below.

$$\delta(y) = \frac{min_x g(y,x)}{max_x g(y,x)} \tag{2.5.8}$$

$$\Delta_1 = min_x \int \delta(y) g(y,x) \lambda(dy) \tag{2.5.9}$$

$$\Delta = max_x \int \delta(y) g(y,x) \lambda(dy) \tag{2.5.10}$$

$f(y_n.....y_m, p) = p_n$ where the dependency is up to the m^{th} observation, where $m < n$,and the initial probability is p.
It can be then shown that

$$limsup_{n\to\infty}(1/n) log||f(y_n.....y_m, p) - f(y_n.....y_m, p')|| <= log(1-\epsilon)$$

where ϵ= minimum entry of the transition probability matrix, and

$$||f(y_n.....y_m, p) - f(y_n.....y_m, p')|| <= 2\epsilon^{-1}\delta y_m (1-\epsilon)^{n-m+1}||p - p'||$$

where p and p' are two different initial conditions. The above equation tells us that the markov chain becomes indifferent to the initial condition at an exponential rate. This is called the property of 'exponential forgetting'. Based on this, we have the following result. Under the condition that the transition matrix is irreducible and aperiodic, Δ is finite, and certain regularity conditions related to continuity and integrability of the emission densities, the markov chain $Z_n = (X_n, Y_n, p_n)$ has unique invariant measure μ .
This is a very useful result, and allows the log likelihood to be expressed as an additive functional of the extended markov chain.
For example, a recursive version of the conditional likelihood estimator can be obtained. This estimator minimizes the function

$$e_n(\theta) = \sum_{k=0}^{n} (1/2n)|Y_k - E_\theta(Y_k|Y_{k-1})|^2$$

Now this can be written as the additive functional

$$\sum_{k=0}^{n}(1/2n)|Y_k - \phi_\theta p_\theta^k|^2$$

It can be also proved that the law of large number holds i.e

$$e_n(\theta) \to e(\theta)$$

almost surely, where

$$e(\theta) = 1/2 \int_{R^d x} P(X)|y - \phi_\theta p_\theta|^2(dy, dp).$$

Then, the recursive algorithm defined by

$$\hat{\theta^k_{n+1}} = \pi(\hat{\theta^k_n} + \gamma_{n-1} H^k_{\hat{\theta^k_n}}(y_n, \hat{p_n}, \hat{w_n}))$$

where

$$H_\theta(y, p, w) = \phi_\theta(y - \phi_\theta p)w_k - \nabla(y - \phi_\theta p)p$$

is asymptotically normal.

2.5.4 L mixing processes

A new approach for maximum likelihood estimation in hidden markov models was investigated by Saska and Gerenser in 2003 , where they tried to use the HMM representation given by Borkar (1993). Links between hidden markov models and general stochastic system via L-mixing processes were established.
In the third chapter of our proposal, we shall talk about extending the asymptotic MLE results to a HMM, where the states are not equispaced. A continuous time markov chain has used to to model such a state space. The analysis of such models are more complicated, because due to a non-specific gap patterns, even the conditional likelihoods lack a closed form. We shall start from some basic assumptions on the boundedness of transition rates and emission densities, and use CLT for martingales to prove normality

for the emission and transition parameters.

2.5.5 Conclusions

Our review of the research done in this area provides us the necessary strength, in terms of tools developed by previous researchers, and motivation, in terms of the limitations and incompleteness of earlier approaches, to work towards a more efficient, accurate, and unified approach for nucleosome mapping and motif finding.
The subsequent sections describes in detail the proposed continuous time Hidden Markov Model and Bayesian estimation procedures that were performed for analysis. The data description, the peculiarities thereof, and the results related to the robustness and performance of our proposed model have been explicitly shown It constitutes the first paper of the proposal, and the second major division of this document. We begin by revisiting the definition and concept of nucleosomes, and the related experimental work done in this context.

Chapter 3

Determining chromatin features using continuous time hidden Markov models

3.1 Introduction

Genomic DNA in the cell nucleus exists as a DNA-protein complex called chromatin, in which the DNA is locally folded and compacted through a series of interactions with proteins called histones. At the first level of compaction, a stretch of DNA about 147 bp in length, is wrapped around a disc shaped octamer of histone protein, yielding a structure called a "nucleosome". Nucleosome structure has been determined by X-ray crystallography (Richmond and Davey, 2003), and steric constraints defining the separation of nucleosomes along the chromosome have been defined. This biological fact has an important implication in relation to gene regulation. The wrapping of DNA around the histone octamer prevents transcription factors (TFs) from binding to the sequence at nucleosomal positions (Widom, 1992). Thus, a well-defined map of the nucleosome positions could be used to improve the predictive power of current motif search algorithms. Most commonly used transcription factor motif search algorithms use the property of sequence conservation at binding sites as the only tool for analysis (Liu et al., 1995; Bussemaker et al., 2000; Liu et al., 2001; Gupta and Liu, 2003). However, in complex genomes, these methods often suffer from poor predictability and high false positive rates. More recent methods have used gene expression measurements through

regression-type models (Conlon et al., 2003b; Gupta and Ibrahim, 2007) or ChIP-chip (Chromatin Immunoprecipitation) experimental data (Buck and Lieb, 2004) to improve predictions, but weak relationships in the first approach often lead to missing important binding sites, and the second approach is limited in applicability to a small set of TFs. The connection of nucleosome positioning to TF-binding has motivated researchers to make use of information obtained from nucleosomal mapping for motif search. For example, Narlikar et al. (2007) have used nucleosome mapping information to create a prior for motif search. However, these methods will not work well unless the input data, that is, nucleosome positions, are based on high-quality predictions from experimental data. Before describing our methodology and its advantages over existing methods for nucleosome prediction, we discuss the various experimental assays for nucleosome detection and a number of existing computational approaches.

3.1.1 Genomic assays for nucleosome position detection

One of the first experiments to identify nucleosome positions was performed by Yuan et al. (2005), based on high-density genome tiling arrays. Tiling arrays are microarrays involving short overlapping probes, covering a genomic region of interest, or even the entire genome. These arrays have been used to experimentally measure positions of various genomic-level events, such as TF binding, histone methylation, and positioning of nucleosomes. In Yuan et al. (2005), based on the nucleosome-free DNA's susceptibility to micrococcal nuclease (*MNase*), nucleosomal DNA was isolated and labeled with green (cy3) fluorescent dye while the total genomic DNA was labeled with red (cy5) dye. These segments were then hybridized on to microarrays printed with 50-mer oligonucleotide probes tiled at 20 base pair overlaps across chromosome III in yeast. After pre-processing and normalizing of the data, the ratio of green-to-red measurement intensity values for spots along the chromosome indicate how likely that locus is to have been wrapped around a histone, constituting a nucleosome.
A more recent experimental technology– Formaldehyde assisted isolation of regulatory elements (FAIRE)– was developed by Hogan et al. (2006). The principle behind this procedure is that regions of the genome that are cross-linked with the histone protein are less susceptible to binding by transcription factors, and thus get retained at the

interphase of organic and aqueous phase, while the nucleosome-free (potentially regulatory) regions get enriched. The FAIRE procedure is initiated by first fixing whole yeast cells in a growth medium by formaldehyde, harvesting them by centrifugation, and finally sonicating the extracts, labeling the purified DNA and hybridizing them to microarrays. The microarray in Hogan et al. (2006) used 50-mer oligonucleotide probes that overlap every 20-mers to tile almost all of chromosome III and 1 kb of 223 additional regulatory regions. Four microarrays (three biological experiments and one technical replicate) were performed. Data of the median of ratios were extracted directly from arrays using Genpix. Acquired images were visually inspected and low quality spots were removed. The data were log transformed, block normalized, and the technical replicates were averaged and treated as one biological replicate, followed by averaging all three biological replicates. The resulting enrichment throughout the genome, in terms of logarithmic ratios of intensity measurements, represents the positioning of the regulatory regions which are free of nucleosomes. Yuan et al. (2005), in contrast, used the MNase (micrococcal nuclease) enzyme to extract histone proteins, the resultant microarray enrichment representing locations of the nucleosomal regions. Thus, as expected, the raw data from Yuan et al. (2005) roughly exhibits a negative correlation with the FAIRE enrichment values (Hogan et al., 2006). Both the experiments use identical probes in the same regions of the genome, thus can be used for comparison, however since FAIRE enriches nucleosome-free regions the exact model framework used in Yuan et al. (2005) is not appropriate for FAIRE data.

3.1.2 Computational approaches for nucleosome detection

We now discuss a number of existing computational approaches towards detection of nucleosome positioning in data from genomic assays discussed in the previous section. Most of these methods are adapted to the data from the first type of experiment (MNase digestion of Yuan et al. (2005)). In Yuan et al. (2005), a hidden Markov model (HMM) was implemented for classifying the genomic regions into nucleosomal and non-nucleosomal regions. This genome wide tiling array approach was repeated in other genomes, enabling the identification of many general features of the nucleosomal landscape shared across species (Pokholok et al., 2005). Ozsolak et al. (2007) performed a

similar genome wide DNA foot printing experiment on human fibroblasts, isolating mono-nucleosomal DNA using micrococcal nuclease. Their method used wavelet decomposition followed by outlier averaging and a peak trough analysis for classification. The HMM method could not be applied because of non contiguous DNA fragments. High-throughput tiling array data also has other complications that arise as a result of the data generation process, such as amplification bias, problematic or inaccurate reads, and varying degrees of non-random noise and measurement error. Relying solely on experimental signal intensities for nucleosome classification could lead to prediction errors that would be further propagated if the predictions are used as secondary input data for motif discovery methods.

Recent work has linked particular DNA sequence features with nucleosome formation (Sekinger et al., 2005), motivating researchers to exploit these features for improved classification. Segal et al. (2006) constructed a map of the nucleosomal positions in the yeast genome, based on the statistical distribution of dinucleotides in nucleosomal DNA. An empirical statistical distribution of dinucleotides was obtained by aligning a set of nucleosomal sequences from log phase yeast. However, neglecting the sequence information of the nucleosome free regions weakened the above classification approach. Some recent nucleosome prediction methodologies include Peckham et al. (2007), who employed a support vector machine for classification of nucleosomal and non nucleosomal states; and Yassour et al. (2008) who extended the HMM technique of Yuan et al. (2005) by implementing a hidden Markov model with multiple states, to account for unstable nucleosome positions and global trends. The relationship between the predicted nucleosomal states, and other features such as presence of promoters, and sequence features was investigated through prediction by Lasso regression in Lee et al. (2007). Yuan and Liu (2008) used wavelet decomposition to represent the sequence signal as covariates in a logistic model for the predicted states. However, in all these approaches, restricting the sequence features to dinucleotide distribution differences could be a potential reason for the overall poor classification rates, in spite of clear biological evidence that sequence composition strongly affects nucleosome positioning (Sekinger et al., 2005). Also, the sequence features were not used in the prediction models– the relationship was explored only on the basis of first obtaining nucleosome predictions (from

an HMM on intensity data) and then inferring potential influential sequence features. One important feature of the data encountered in the new technologies for nucleosome mapping are the gaps between adjacent probes. Employing a simple hidden Markov model, assuming equally spaced probes, would seriously misspecify the parameters and give us an erroneous state prediction in the case where there are unevenly spaced probes and/or missing data. In ChIP-chip assays for detecting potential transcription factor binding sites, where missing probe data may also be present, this limitation may not be as serious as the TF-bound enriched probes only constitute a tiny fraction of the total. Recent literature has also suggested that the conformability of DNA structure might strongly depend on the sequence combinations over an extended set of nucleotides. In this article we propose a novel general model framework for extending a hidden Markov model in two directions for this problem: (i) use of an underlying continuous-index Markov process that allows for gaps and missing data between probes, and (ii) incorporation of relevant sequence features, not limited to dinucleotide frequencies, into nucleosomal state prediction, through dimension reduction techniques. This unified model uses both the signal intensity and sequence features to predict the likelihood of nucleosome formation, lessening the dependence on only one data source to increase the accuracy of prediction. Continuous-index Markov processes have been successfully applied to certain other problems involving genomic data, such as copy-number variation detection (Stjernqvist et al., 2007) and genetic linkage analysis (Lander and Green, 1987), but the present problem presents some unique characteristics motivating our particular model formulation and fitting algorithm. The structure of the article is as follows. In Sections 3.2 and 3.3, we describe our proposed new model framework and methodology for model-fitting. In Section 3.5 we show that including sequence features of a higher order than di-nucleotides leads to a higher predictive power in this model framework, through applications to yeast nucleosomal assay data as well as a number of simulation studies.

3.2 Model framework

In the following sections, we describe our proposed model to extract positional estimates of nucleosome-free regions from the FAIRE enrichment data (Hogan et al., 2006), and

then illustrate its performance on real data as well as simulation studies. Although our model is developed with the FAIRE data structure as a template, it is a general model which can be used for a variety of high-throughput microarray settings. Also, from a biological standpoint, since the FAIRE technology is simpler to use, it is likely to be useful in nucleosomal studies for a wide variety of organisms.

3.2.1 Data structure

Due to experimental errors and image processing failures, the final data often has a lot of missing probe measurements. The actual intervals between probe measurements thus vary in length, especially for large data sets over entire chromosomes. For each position on the chromosome with fully observed measurements, the following data are available:

$$\begin{aligned} Y_{ij} &= \text{log-ratio of the intensity of the replicate } j \text{ of probe } i \\ t_i &= \text{ Nucleotide index of the beginning of probe } i \\ \boldsymbol{X}_i &= d\text{-dimensional vector of covariate measurements for probe } i, \end{aligned}$$

for $i = 1, \ldots, P$, and $j = 1, \ldots, R$. In addition, we also define $t_{a,b} = t_b - t_a$ as the gap between the center positions of probes indexed a and b, which is the genomic distance between two probes (measured on the base pair scale). If all probes are of the same length and equispaced, which is typically the case in these experiments, the gaps arise solely out of the existence of missing probes, leading to the distance $t(a, b)$ being a constant multiple of the number of missing probes in between. In this simplified case, it is equivalent to work with the metric of the number of missing probes as a distance measure.

3.2.2 Continuous index hidden Markov process

The observed probe measurements are subject to various sources of error. The actual underlying nucleosome occupancy status is thus assumed to be an unobserved stochastic process evolving over a continuous index, denoted by $[Z(t), t > 0]$. Further, we assume that $Z(t)$ represents a two-state Markov process, with $Z(t) = 1$ (0) representing a NFR

(non-NFR) state, such that

$$P[Z(t_i)|Z(t_{i-1}),\ldots,Z(t_1),\boldsymbol{Y}_{i-1},\ldots,\boldsymbol{Y}_1] = P[Z(t_i)|Z(t_{i-1})] = P_{z_{i-1}z_i}(t_i - t_{i-1}), \quad (3.2.1)$$

where $P_{z_{i-1}z_i}(\cdot)$ denotes the transition probability that state z_i was occupied at index t_i given that the process was in state z_{i-1} at index t_{i-1}. For the moment, let us assume that the transition probabilities of this process are stationary. We also assume that, conditional on the state of the process Z at index t_i, the observed measurement $\boldsymbol{Y}_i$ is independent of the hidden process and all observed measurements prior to index t_i, that is

$$P[\boldsymbol{Y}_i|Z(t_i),\ldots,Z(t_1),\boldsymbol{Y}_{i-1},\ldots,\boldsymbol{Y}_1] = P[\boldsymbol{Y}_i|Z(t_i)] = f(\boldsymbol{Y}_i|z_i). \quad (3.2.2)$$

Expressions (3.2.1) and (3.2.2) constitute a continuous index hidden Markov process. For notational simplicity, we shall henceforth refer to $Z(t_i)$ as simply Z_i.
We assume that the probability of staying in a state is linear with respect to index for an infinitesimal interval. The two-state hidden process can then be parameterized by the transition rate from a nucleosomal state to a NFR, (λ), and the reverse transition rate, (μ). The rate (or intensity) matrix Q is given by $Q = \left(\begin{smallmatrix} \lambda & -\lambda \\ -\mu & \mu \end{smallmatrix}\right)$. The matrix of transition probabilities $P(t)$ over an interval t is generated by the matrix exponential of a matrix of instantaneous transition rates Q, that is, $P(t) = \exp(Qt)$. Indexing the nucleosomal (non-NFR) state by 0 and NFR state by 1, we then have the following form for the transition probabilities:

$$\begin{aligned} P_{00}[t] &= \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \\ P_{11}[t] &= \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}. \end{aligned} \quad (3.2.3)$$

For sampling parameters under the MCMC framework, a parameterization in terms of the log-intensities is useful, i.e. we take $\log\lambda = \theta_{l0}$ and $\log\mu = \theta_{m0}$. The initial distributions of the hidden states are taken to be uniform across the states, that $\boldsymbol{\pi} = (\pi_0, \pi_1) = (.5, .5)$. In fact we can take any prior probability since a) the stationary distribution exists b) the markov chain irreducible, aperiodic and recurrent. So

convergence to the target distribution is guaranteed for any initial distribution . Also, no selective criteria are used to determine the starting point of the data sequence.
Fixing the transition structure, next we specify the emission densities. The data structure consists of three replicates within each probe. We can impose the same distribution for all replicates within all probes belonging to a particular state. However there might be a variation of the mean intensity between probes in the same state. In order to account for this variability, we implemented a hierarchical model where we have separate means for each probe, and these means are centered around a state-specific mean with a state-specific variance. The hierarchical Gaussian model is given by:

$$\begin{aligned} Y_{ij}|Z_i = k &\sim N(\nu_{ik}, \sigma_k^2); \quad k = 0,\ 1; \\ \nu_{ik}|Z_i = k &\sim N(\nu_{k0}, \tau_k^2), \end{aligned} \tag{3.2.4}$$

A more general form of the hierarchical model could involve probe-specific variances; but since we observed little evidence of variation among probe variances, we used the simpler form of the model to avoid an excessive computational burden. Though it is not possible to do an analysis of the variances belonging to different probes in different states, since the states themselves would be unknown before the estimation, luckily we found out that the variances within the probe were uniform throughout the data set, i.e, the variances of these probe specific variances were negligible (less than .01) So we settled for a less complex hierarchical model which took into account the diversity between the probes only in terms of th mean structure

3.2.3 Adding covariate effects to the model

Underlying characteristics of the DNA sequence that affect chromatin rigidity may affect positioning of nucleosomes. In previous studies, the prevalence of certain polynucleotide sequences such as poly-A (repetitions of the "A" nucleotide), poly-T and some others have been seen to affect nucleosome positioning. Hence such sequence characteristics may be assumed to affect the correct prediction of nucleosomal state. In order to model the association between the covariates and states, we used three levels of models, relating the

covariates to the transition rates, or emission probabilities, via link functions.

- Model M0: the original model (Eqns. 3.2.4 and 3.2.3) assuming that the covariates do not affect either the state transitions or emissions.

- Model M1 ("transition model"): Here we use a multiplicative intensity model for associating the covariates to the transition rates, by assuming that the transition rate during the interval between two points depends on the the value of the covariates at the end of the interval, i.e. the last probe. Let X_{ij}

  denote the value of covariate j ($j = 1, \ldots, d$) for probe i. Then the transition rates over the interval $[t_{i-1}, t_i]$ become:

$$\begin{aligned} \log \lambda_i(\boldsymbol{X}_i) &= \theta_{l0} + \sum_{j=1}^{d} \theta_{lj} X_{ij} \\ \log \mu_i(\boldsymbol{X}_i) &= \theta_{m0} + \sum_{j=1}^{d} \theta_{mj} X_{ij}. \end{aligned} \tag{3.2.5}$$

  In the later sections, to simplify notation, we shall denote $\lambda_i(\boldsymbol{X}_i)$ and $\mu_i(\boldsymbol{X}_i)$ simply as λ_i and μ_i.

- Model M2 ("emission model"): Here we assume that the probability distribution of the observations conditional on the hidden state can also be affected by the covariate measurements (i.e. probe sequence features). In this case, we assume that the state-specific probe measurement mean can be modeled as a linear function of the covariates, i.e.,

$$\nu_i | Z(t_i) = k \sim N\left(\nu_{k0} + \sum_{j=1}^{d} \beta_{kj} X_{ij}, \tau_0 \sigma_k^2\right) \tag{3.2.6}$$

  Note that the emission model (M2) has probe specific means already built into the model. So instead of modeling ν_{ij} of the hierarchical set up we directly model ν_i (We implemented the hierarchical setup for the base and transition models).

- Model M3 ("full model"): Here we assume both the transition intensities and state emission probabilities are affected by the covariates, that is, both expressions (3.2.5) and (3.2.6) hold.

None of these models described above lead to a closed-form analytical expression for the likelihood due to the hidden state variable Z, although this can be computed numerically through recursive techniques, as we discuss in Section 3.3. Before the model estimation procedure is described, we discuss the issue of identifiability in our proposed models, as it is highly relevant to the validity of our inference.

3.2.4 Identifiability

Here, we examine closely the identifiability conditions of the model described in Sections 3.2.2 and 3.2.3. Non-identifiability of a model implies that there are two sets of parameter settings that give rise to the same likelihood (and posterior distribution, with a suitably non-informative prior). Usually in a Bayesian framework imposing an informative prior may ensure identifiability. But since we want to impose minimal prior effects on inference and accordingly set flat priors, non-identifiability can lead to serious convergence problems in the Markov chain Monte Carlo (MCMC) sampling procedure and bias inference. There are two parts to the proof. Our approach is very similar to Leroux's in the sense that we both base our result on the fundamental theorem stated by Teicher in his 1967 paper 'On the identifiability of mixtures in product measures' This theorem has been the basis for proving identifiability conditions in several settings where the joint modeling of dependent data has come up. Hidden markov models form one such category of models. In Leroux's proof of identifiability in HMM, the part of identifiability of emission densities (that follow from the that the collection of emission densities should be same and the emission parameters have one to one correspondence with the densities) is the same as ours. However for the next part on equivalence of the mixture weights, we diverge. From this part it can be concluded in both the models that the laws of the processes are equal, which implies that the transition probabilities are equal. In Leroux's model, the proof is completed there itself since the parameters are the transition probabilities. However in our model, the transition parameters are related to the

transition probabilities by means of the transition equations, and it needs some work in this stage to translate the condition of equivalence of transition probabilities to that on transition rates.

To establish identifiability conditions for the proposed models, we extend the results of Teicher (1967). First, let us state the following definition and results.

Definition. Let $f_\phi(y)$ be a parametric family of densities of Y with respect to a common dominating measure μ and parameter ϕ in some set Φ. If π is a probability measure on Φ, then the density

$$f_\pi(y) = \int_\Phi f_\phi(y)\pi(d\phi)$$

is called a mixture density.

We say that the class of (all) mixtures of f_ϕ is identifiable if

$$f_\pi = f_{\pi'} \quad \mu - \text{a.e iff } \pi = \pi' \text{for all y}$$

Further, we say that the class of finite mixtures of f_ϕ is identifiable if for all measures π and π' with finite support, $f_\pi = f_{\pi'}\mu$ a.e iff $\pi = \pi'$.

Proposition 1. ***(Teicher, 1967)***. *The class of joint finite mixtures of the normal family is identifiable.*

Proposition 2. ***(Teicher, 1967)***. *Assume that the class of finite mixtures of the family f_ϕ of densities of Y with parameter $\phi \in \Phi$ is identifiable. Then the class of finite mixtures of n-fold product densities $f_\phi^{(n)}(y) = f_\phi(y_1)\dots f_\phi(y_n)$ with parameter $\phi \in \Phi^n$ is identifiable.*

Proposition 2 was proved by induction on n (Teicher, 1967).

Now, note that any hidden Markov model is a finite mixture of n-fold product densities, where the the weights of the mixture are functions of the transition probabilities.

Keeping this in mind, we applied the above results to prove the identifiability of models M0-M3, given in the following result.

Theorem 1. *Models M0, M1, M2 and M3 are identifiable. In other words, if $\boldsymbol{\eta}$ denotes the total set of all parameters in any of the four models, and $L(\boldsymbol{\eta}; \boldsymbol{y})$ denotes the likelihood of $\boldsymbol{\eta}$, there does not exist any $\boldsymbol{\eta}' \neq \boldsymbol{\eta}$ such that $L(\boldsymbol{\eta}; \boldsymbol{y}) = L(\boldsymbol{\eta}'; \boldsymbol{y})$ for all $\boldsymbol{y}$.*

The proof of Theorem 1 is given in the Appendix.

3.3 Model-fitting and Estimation procedure

First, we write down the likelihood under the proposed model framework, and then discuss estimation of parameters and hidden states. Let $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2)$ denote the set of all parameters in the model. (For model M0, $\boldsymbol{\theta} = (\lambda, \mu)\boldsymbol{\beta} = (\nu_0, \nu_1)$, and likewise for all models except M3, at least one parameter becomes void.) The likelihood of a sequence of N observed probe measurements, conditional on the covariates $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ and probe locations $\boldsymbol{T} = (t_1, \ldots, t_N)$, can be written as

$$\begin{aligned} L(\boldsymbol{\eta}) &= P(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N | \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\eta}) \\ &= [\pi_0 P(\boldsymbol{Y}_1 | z_1 = 0) + \pi_1 P(\boldsymbol{Y}_1 | z_1 = 1)] \sum_{z_1, \ldots, z_N} \prod_{i=2}^{N} P_{z_{i-1}, z_i | x_i}(t_i - t_{i-1}) f(\boldsymbol{Y}_i | \boldsymbol{X}_i, z_i) \end{aligned} \quad (3.3.1)$$

Direct evaluation of (3.3.1) would involve summing over all possible sequences of hidden states $z_1, \ldots, z_N$. However, under the Markovian assumption, this can be instead evaluated recursively, through a forward summation procedure.
One common approach to parameters and state estimation in HMMs is the expectation-maximization (EM) approach (Dempster et al., 1977). However, in fitting our model by EM, due to the complex nature of the transition probabilities, no closed form expression exists for the M-step, necessitating numerical optimization which would be very computationally intensive. The problem in the M step arises mainly due to the fact that the complete data likelihood is a product of the terms of the type coming in the transition equations 3.2.3. Although there have been analytically tractable M-steps for some continuous time hidden Markov models (Roberts and Ephrain, 2008), our model differs substantially from these in a very important aspect. In models of the former category, it is assumed that the state remains constant throughout the gap and there is a state jump only at the observation points. These models rely on the fundamental assumption that the probability of transition is linear with respect to the time spent in a time interval, and the proportionality constant is the rate of transition. Since the

transition structure is such that the observed value is obtained only after a jump over a period in a constant state, the transition equations are formulated simply in linear terms of the rates. On differentiating the log-likelihood we get a simple linear equation in terms of the rates. However, a basic aspect of our model is the assumption that there can be an arbitrary number of state changes within the gapped period, and the probability spanning the whole set of changes cannot be formulated in terms of the transition equations as described above. These make inferring a closed form M-step impossible. On the other hand, an MCMC sampling-based approach appears attractive due to the standard or log-concave forms of many of the conditional distributions. It is also more attractive to use a Bayesian MCMC approach in order to obtain the full posterior probability landscape of the parameters rather than a single maximum likelihood-based point estimate. In the next sections, we describe prior elicitation, and the steps of an MCMC algorithm that makes use of a recursive procedure to sample parameters from their posterior distributions.

3.3.1 Prior elicitation and sensitivity analysis

In an effort to keep the effect of the priors minimal, we assumed a flat prior for most parameters in the model. It can be deduced from the structure of the full model likelihood that the conditional posteriors assume proper densities with a flat prior. This can be observed from two facts. First, the emission densities are normal. Thus, marginalizing the joint density along the axes of emission parameters would be equivalent to integrating out normal densities. This would eventually yield a finite weighted summation of finite terms. Second, the expressions containing the transition parameters are a finite summation of terms of the type $f(a)\exp(-g(a))$ where f is a function bounded by 0 and 1 and g is a positive scalar multiple in M0 and is a monotonic function in the transition model M1. For all such functions of this type the integral is finite, and again leads to a finite summation of finite terms.
However there are two cases in the above likelihood corresponding to all probes being classified into either the 0 or 1 state, which when integrated out with a flat prior would yield infinity, as these parameters are not contained in such terms. In this case the posterior is improper and Gibbs sampling with such a model might lead to incorrect

results. However this problem can be circumvented by noting that the probabilities of these two extreme conditions is infinitesimally low (and typically zero in the real-life application here). However this problem can be circumvented by imposing a constrained hidden state model. This constrained state space will exclude these two states. Then we have a proper posterior. The adavantage of this technique derives from the observation that when we sample from these constrained space model, the conditional probabilities of the hidden states (excluding the all 0 or all 1 case) involved in the sampling procedures would differ from the conditional probabilities of our sampling scheme(which includes the extreme states) by a negligible margin. This is because the probabilities of these two extreme conditions is infinitesimally low (and typically zero in the real-life application here) and make practically no difference at all to the probabilities. Here we must note that that it is not because of the configuration of the states that we are able to get round this problem, since the likelihood of any state vector is very low for large N. It is because of the finiteness of such state vectors(2,in this case) that we can progress in the way mentioned above. In practical implementation,our sampling scheme can hit one of these two states. Apart from the assurance that such a situation can never occur in the current application on this genomic data, where the log intensities are so distributed that the sampler can never assign zeros to all the positions simultaneously(except when we initialize one transition rate to be a unrealistically large negative number such that the corresponding transition probability becomes 0), we can ignore the sample and re-sample. This is a valid step under the constrained sampling set-up. By this way of proceeding with a regular Gibbs sampler with a flat prior, we can approximate performing Gibbs sampling on the constrained space model.

Together with these observations and the fact that we would not like to bias the data with extraneous information, we initially assumed a flat uniform prior for the transition parameters and the emission mean parameters of our model. For the variance parameters, we used a non-informative prior for (τ^{-1}) where τ is the inverse of the variance σ^2. We fit the models again assuming a vague normal prior (with a variance of 5) on the transition parameters. Then we systematically reduced the variance by .25 units, to see the changes. For the vague prior there was no identifiable difference in the results. When the prior variance reduced beyond 1, there was an effect of pushing the log

of the transition parameters towards 0.
Thus for our analyses and simulation studies, we chose the emission parameters and transition parameters to be uniformly flat. Specifically, we chose, $p(log(\lambda)) \approx 1$; $p(log(\mu)) \approx 1$; $p(\nu_{k0}) \approx 1$ $(k = 0, 1)$; $p(\sigma_k^2) \approx \sigma_k^{-2}$ $(k = 0, 1)$; $p(\theta_{mj}) \approx 1, p(\theta_{lj}) \approx 1$ $(j = 1, \ldots, d)$; and $p(\beta_{kj}) \approx 1$ $(k = 0, 1;\ j = 1, \ldots, d)$.

3.3.2 The MCMC sampling algorithm

Let us denote $P_{jk}(t_{i+1}|t_i)$ as the transition probability from state j to state k, for adjacent probes situated at positions t_{i+1} and t_i, which is implicitly dependent on the transition rates λ and μ. The following steps constitute one iteration of the MCMC sampler designed to sample from the posterior distributions of all parameters of interest.

1. Use the forward algorithm to calculate the full likelihood given the set of parameters. This algorithm iteratively computes the likelihood of the data till a step in the sequence, with the last step being in a particular state. The first application of this algorithm in computing HMM likelihoods was shown by Baum et al. (1970) and later extended for a variety of models (Rabiner, 1989). Let F_j^i denote the "partial" likelihood until position i of the sequence, where position i is in state j, that is,
$$F_j^i = P(Z_i = j, Y_1, \ldots, Y_i | X_1, \ldots, X_i).$$
In our model, this is equivalent to:
$$F_j^i = \sum_{z_1, \ldots, z_i} P(\boldsymbol{Y}_1|z_1) \prod_{l=2}^{i} P_{z_{l-1}, z_l|x_l}(t_l|t_{l-1}) f(\boldsymbol{Y}_l|\boldsymbol{X}_l, z_l),$$
$j = 0, 1$ representing the nucleosomal and the NFR states respectively. The recursive procedure to calculate the likelihood is given by
$$F_j^i \ = \ [F_0^{i-1} P_{0j}(t_i|t_{i-1}) + F_1^{i-1} P_{1j}(t_i|t_{i-1})] f(\boldsymbol{Y}_i|Z_i = j, \boldsymbol{\eta}), \quad i = 2, \ldots, N. \quad (3.3.2)$$
We must note $\boldsymbol{Y}_i$ is a vector in our case consisting of one two or three replicates at the non missing positions. So its dimension varies with the number of non missing

replicates in the data. and
$f(\boldsymbol{Y}_i|Z_i = j, \boldsymbol{\eta}) = \int [\prod_{k=1}^{3} \phi(Y_{ik}, \nu_{ij}, \sigma_j)][\phi(\nu_{ij}, \nu_{0j}, \tau_j)]$
The above integral has a closed form expression. It can be computed by the usual method of completing squares of the exponent of the kernel of the Gaussian densities.

The initial conditions are: $F_j^1 = \pi_j f(\boldsymbol{Y}_1|Z_1 = j, \boldsymbol{\eta})$, $(j = 0, 1)$, and the full likelihood of the entire sequence is given by $F_0^N + F_1^N$.

2. Next, we employ a backward sampling procedure (Rabiner, 1989) to get a sample of the hidden states. Conditional on the sampled states at positions $t_{i+1}, \ldots, t_N$, the probability that position i is in state j is

$$P(Z_i = j|Z_{i+1}, \ldots, Z_N, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\eta}) \propto F_j^i P_{j,Z_{i+1}}(t_{i+1}|t_i)$$

The probability that the last position is in state j is proportional to F_j^N. In actual applications, the above procedures need to be reformulated in terms of the logarithms of the probabilities to avoid computational underflow.

3. Conditional on the other parameters, hidden state path, and observed data, update $\log \lambda$ and $\log \mu$ using a Metropolis-Hastings procedure. For models M0 and M2, this is done directly for θ_{l0} and θ_{m0}, while for models M1 and M3 this is done for each component in turn.

4. Conditional on the other parameters, hidden state path, and observed data, update the emission parameters ν_{k0} $(k = 0, 1)$ for model M0 and M1 and additionally, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kd})$, for models M2 and M3.

3.4 Details of sampling procedure

Our prior specification scheme is set as in Section 3.3.1. For the reader's convenience we present it again here
$p[log(\lambda)] \approx 1$; $p[log(\mu)] \propto 1$; $p(\nu_{ki}) \propto 1$ $(k = 0, 1)$; $p(\nu_{k0}) \propto N(\frac{\sum_{i=1}^{n_k} \nu_{ki}}{n_k}, \tau_k^2)$; $\quad k = 0,\ 1$;
$p(\sigma_k^2) \propto \sigma_k^{-2}$ $(k = 0, 1)$; $p(\tau_k^2) \propto \frac{1}{\tau_k^2}$ $p(\theta_{mj}) \propto 1, \pi(\theta_{lj}) \propto 1$ $(j = 1, \ldots, d)$; and $p(\beta_{kj}) \propto 1$

$(k = 0, 1;\ j = 1, \ldots, d)$.
Note that the log of the transition parameters (both in M0 and M1) has an uniform prior. This is because a uniform prior on any of these transition parameters, say, λ would result in an improper posterior for the logarithm of that parameter. The emission parameters in the model have been given a flat prior.
We adopt the following notations for convenience.
We denote the collection of emission parameters in the base model M0 and the emission model M2 as as η_b and η_e respectively. That is
$[\nu_{10}, \nu_{00}, \sigma_0, \sigma_1, \tau_0, \tau_1, \nu_{01}, \nu_{02} \ldots \nu_{0N}, \nu_{11}, \nu_{12} \ldots \nu_{1N}] = \eta_b$
$[\boldsymbol{\beta}, \sigma_0, \sigma_1] = \eta_e$
In a similar fashion we use a shortened notation for the collection of transition parameters. That is
$[\log(\lambda), log(\mu)] = \psi_b$
(This would denote the set of transition parameters for models M0 and M2.)
$[\theta_{l0}, \theta_{l1}\theta_{l2} \ldots \theta_{ld}, \theta_{m0}, \theta_{m1}, \theta_{m2} \ldots \theta_{md}] = \psi_t$
(This would denote the set of transition parameters for model M1.)
Here d=6. These correspond to the six covariates that we have included in the model for the nucleosomal and NFR states, including the intercepts.
Also we use n_k to denote the number of observations in state k. That is
$n_k = \sum_{i=1}^{N} I(Z_i = k)$.
We employ a Gibbs sampling method. The basic steps of the algorithm for model are enumerated as follows:

1. $P(Z|\boldsymbol{Y}, \eta_b, \psi_b)$.
   For this step we use the Data Augmentation technique described in subsection 3.3.2.

2. $P(\psi_b|\boldsymbol{Y}, \eta_b)$
   (As in a conventional Gibbs sampler, we should have sampled from $P(log(\lambda)|log(\mu)|\eta_b, Z, \boldsymbol{Y})$ and $P(log(\mu)|log(\lambda)|\eta_b, Z, \boldsymbol{Y})$. However,instead of that, we are integrating out Z from the full conditional. The reason we are doing that is that we want to replace the collection of these three full conditional steps $P(log(\lambda)|log(\mu)|\eta_b, Z, \boldsymbol{Y})$, $P(log(\mu)|log(\lambda)|\eta_b, Z, \boldsymbol{Y})$ and $P(Z|\boldsymbol{Y}, \eta_b, log(\lambda), log(\mu))$

by the collection of the two following block sampling steps.
1. $P(Z|\boldsymbol{Y}, \eta_b, log(\lambda), log(\mu)$
2. $P(Z|\boldsymbol{Y}, \eta_b, log(\lambda), log(\mu))$
In this way we would be drawing en block from the distribution of $P(Z, \psi_b|\boldsymbol{Y}, \eta_b)$
Note that
$P(Z, \psi_b|\boldsymbol{Y}, \eta_b) = P(log(\lambda), log(\mu)|\eta_b, \boldsymbol{Y})P(Z|\boldsymbol{Y}, \eta_b, log(\lambda), log(\mu))$
A Gibbs block sampler, if it can be implemented, is always a more efficient sampler than the one composed of full conditional distributions.)

- Note that in case of model M2 the above two steps are replaced by $P(\psi_b|\boldsymbol{Y}, \eta_e)$.
- In case of model M1 the above two steps are replaced by $P(\psi_t|\boldsymbol{Y}, \eta_b)$.

3. $P(\eta_b|\boldsymbol{Y}, Z, \psi_b)$
This step is replaced by $P(\eta_e|\boldsymbol{Y}, Z, \psi_b)$ in the case of model M2 and $P(\eta_b|\boldsymbol{Y}, Z, \psi_t)$ in case of Model M1.

We now elaborate on Steps (2) and (3).

3.4.1 Step I–Sampling the transition parameters

$P(\psi_b|\boldsymbol{Y}, Z, \eta_b)$

- For Models M0 and M2: The posterior distribution is proportional to the likelihood, since we have an uniform prior.

$$P(\psi_b|\boldsymbol{Y}, Z, \eta_b) = P(log(\lambda)|log(\mu), \eta_b, \boldsymbol{Y}) \propto L(log(\lambda), \mu, \eta_b|\boldsymbol{Y})p(log(\lambda))$$

$= L(log(\lambda), log(\mu), \eta_b|\boldsymbol{Y})$
The last expression is computed by the forward algorithm given in Subsection 3.3.2.
Here the forward algorithm essentially calculates the full likelihood expression
$\prod_{i=i}^{N} \sum_{z_1,\ldots,z_N} P(\boldsymbol{Y}_1|z_1)P_{z_{i-1},z_i|x_i}(t_i - t_{i-1}) \prod_{j=1}^{3} f(\boldsymbol{Y}_{ij}|\boldsymbol{X}_i, z_i)$.
(So essentially here we are integrating out Z from the conditional likelihood).

Since there is no closed form for the kernel of the posterior given by $L(log(\lambda), log(\mu)|\eta_b \boldsymbol{Y})$ we resort to a conditional Metropolis Hastings updating Scheme. The scheme is given below.

1. Generate a sample $\log(\lambda_{new}) \sim N(\log(\lambda_{old}), V)$.
2. Compute the ratio $R = \frac{L(log(\lambda_{new})),\mu,\eta_b|\boldsymbol{Y}}{L(log(\lambda),\mu,\eta_b,|\boldsymbol{Y})}$
3. Accept λ_{new} with probability min(1,R).
4. Repeat the above steps for $log(\mu)$.

(In our application on the Hogan dataset,we actually got a stable rejection rate of 35 percent by taking V=0.4.)

- For Model M1:
  Again as before, we do not have a closed form for the kernel of the posterior given by $P(\psi_t|\boldsymbol{Y}, \eta_b)$. Hence we implement a conditional Metropolis Hastings update as follows:

  1. Start from j=1.
  2. Generate a sample $\theta_{lj}^{new} \sim N(\theta_{lj}, V)$.
  3. Compute the ratio $R = \frac{L(\theta_{lj}^{new},\mu,\eta_b|\boldsymbol{Y})}{L(\theta_{lj},\mu,\eta_b,|\boldsymbol{Y})}$
  4. Accept θ_{lj}^{new} with probability min(1,R).
  5. Put j=j+1. Go back to step 2.
  6. Stop when j = d .
  7. Repeat all of the above steps for $log(\theta_{mj})$ $j = 1 \ldots d$.

3.4.2 Step II : Sampling of emission parameters

We first note that the posterior distribution
$P(\eta_b|\boldsymbol{Y}, Z, \psi_b) \propto L(\psi_b, \eta_b|, Z, \boldsymbol{Y})p(\eta_b)$
(The last expression is simply the complete data likelihood times the prior for η_b).
$= L(\psi_b, \eta_b|, Z, \boldsymbol{Y})[\prod_{i=1}^{N} \prod_{k=0}^{1} \phi(\nu_{ik}, \nu_{0k}, \tau_k^2)$

Now from the likelihood equation we see that

$$L(\psi_b, \eta_b|, Z, \boldsymbol{Y}) = \prod_{i=1}^{N} \sum_{z_1,\ldots,z_N} P(\boldsymbol{Y}_1|z_1) \prod_{i=2}^{N} P_{z_{i-1},z_i|x_i}(t_i - t_{i-1}) \prod_{j=1}^{3} f(\boldsymbol{Y}_{ij}|\boldsymbol{X}_i, z_i)$$

Note that in the right hand side of the equation the first part does not involve η_b and so the posterior is further proportional to
$\prod_{i=1}^{N} \prod_{k=0}^{1} \phi(\nu_{ik}, \nu_{0k}, \tau_k^2) \prod_{j=1}^{3} [f(\boldsymbol{Y}_{ij}|\boldsymbol{X}_i, \eta, z_i)]$.
(Here ϕ denotes the normal density function).
This is equal to the following expression
$\prod_{i=1}^{N} \prod_{k=0}^{1} \phi(\nu_{ik}, \nu_{0k}, \tau_k^2) \prod_{j=1}^{3} [\phi(\boldsymbol{Y}_{ij}, \nu_{ki}, \sigma_k^2)]$.
The above expression tells us that the components of η_b emerge as the mean and variance parameters in an expression of product of normal densities. Thus it is easy to deduce that their posterior distributions are normal and Inverse Gamma. The sampling steps are enumerated below

1. Sample $\nu_{k0} \approx N(\frac{\sum_{i=1}^{n_k} \nu_{ki}}{n_k}, \tau_k^2); \quad k = 0,\ 1;$
2. Sample $\nu_{ik} \approx N(\frac{\nu_{k0}/\tau_k^2 + y_{\bar{ik}}/\sigma_k^2}{1/\tau_k^2 + 1/\sigma_k^2}, \frac{1}{[1/\tau_k^2 + 1/\sigma_k^2]}); \quad k = 0,\ 1$
3. Sample $\tau_k^2 \approx IG(\frac{\sum_{i=1}^{n_k} (\nu_{ik} - \nu_{\bar{k}})^2}{2}, \frac{(n_k - 1)}{2})$
4. Sample $\sigma_k^2 \approx IG(\frac{\sum_{i=1}^{N} \sum_{j=1}^{3} (y_{ij} - \bar{y_i})^2 I(Z_i = k)}{2}, (n_k - 1))$

In the above expressions $y_{\bar{ik}}$ denotes the average of the intensity data over the replicates at the i^{th} probe where $Z_i = k$.
$\nu_{\bar{k}} = \frac{\sum_{i=1}^{n_k} (\nu_{ik})}{n_k}$ denotes the average of the probe specific means in all probes of the particular state.
For model M2, we do not have the hierarchical setup. However, following the exact same argument as in the base model M0, the posterior distribution of η_e can be found to be proportional to $\prod_{i=1}^{N} \prod_{k=0}^{1} \prod_{j=1}^{3} [f(\boldsymbol{Y}_{ij}|\boldsymbol{X}_i, \eta_e, z_i)]$.
Recall that the components of η_e are given by :
$\eta_e = [\boldsymbol{\beta}, \sigma_0, \sigma_1]$

Expanding the above expression of the kernel of the posterior in terms of these parameters we get

$$P(\eta_e|\boldsymbol{Y}, Z, \psi_b) \propto \prod_{i=1}^{N} \prod_{k=0}^{1} \prod_{j=1}^{3} [\phi(\boldsymbol{Y}_{ij}, [\boldsymbol{X}_i]'\beta^k, \sigma_k]$$

where $\beta^k = [\beta_{k1}, ...\beta_{k6}]$ Writing this in matrix notation we get

$$P(\eta_e|\boldsymbol{Y}, Z, \psi_b) \propto \prod_{k=0}^{1} \prod_{j=1}^{3} [\Phi(\boldsymbol{Y}_k, W_k'\beta^k, \sigma_k]$$

where Φ denotes the multivariate normal function. i.e
$\Phi(y, a, b) = exp[-\frac{(y-a)'(y-a)}{2b^2}]$
and $\boldsymbol{Y}_k$ represents a subset of the full data vector $\boldsymbol{Y}$ where $Z_i = k$. Z_k denotes a subset of the state vector where $Z_i = k$. W_k denotes a subset of the covariate matrix where $Z_i = k$. Thus we see that in the expression of the posterior, the $\boldsymbol{\beta}$ and the variance parameters emerge as the slope vector and the variance parameters of a simple linear regression model. Now we implement the following sampling steps:

1. Sample $\beta^k \approx P(\beta^k|Z_k, Y, \sigma_k) \quad k = 0,\ 1;$
   where
   $$P(\beta^k|Z_k, Y, \sigma_k) = N([W_k'W_k]^{-1}W_k'\boldsymbol{Y}_k, \sigma_k^2[W_k'W_k]^{-1})$$
   k=0,1.

2. Sample $\sigma_k \approx P(\sigma_k|Z_k, Y, \boldsymbol{\beta}) \quad k = 0,\ 1;$ where
   $$P(\sigma_k^2|\boldsymbol{Y}, Z, \boldsymbol{\beta}) = IG((n_k - d)/2, (1/2)s_k^2)$$
   k=0,1 and
   $s_k^2 = \sum_{i:Z_i=k}(Y_i - \boldsymbol{X}_i'\beta^k)^2/(n_k - 1)$.
   IG denoting the InverseGamma distribution.

3.5 Application to yeast nucleosome array data

The estimation procedure for the base, transition and emission models was applied to the FAIRE data set from Hogan et al. (2006), that comprised a total of 13947 probes. The data structure consists of two elements:

- Signal: This gives the log of the hybridization ratios obtained from the microarrays. For each probe we have three replicates each giving a measure of the intensity data at the particular probe. The number of probes is uniform throughout the data set The higher the value of these intensity measurement, the greater the chance for the probe to be in a nucleosome free region. The data was preprocessed by a z-score standardization to remove skewness. The average of the three replicates of the intensities was taken as the signal, denoted by Y_i. This signal was missing at certain probes. The missingness pattern was random, and the maximum length of a missing block was 13. The non-missing signal values numbered 12760, about 88% of the total data set. More details on the standardization used can be found in Hogan et al. (2006).

- Sequence: Each probe was of length 50 nucleotides, with an overlap of 20 nucleotides with the adjacent probe. As covariates which may potentially influence the nucleosomal signal, we first extracted an initial feature set consisting of counts of all oligomers upto length 5. For example a measurement of a covariate corresponding to the dinucleotide TA would be the number of occurrences of this dinucleotide throughout the length of the probe. Each probe thus corresponds to a point in a 1364-dimensional Euclidean space, the coordinate for each probe given by the corresponding oligomer word counts. These covariates, especially for longer size oligomers, have very sparse counts, and tend to be highly correlated with counts of words which are smaller segments of them. To avoid collinearity and also reduce dimensionality of the covariate space, we performed principal components analysis (PCA) on these 1364 covariates for all the probes in our data set. The first five principal components were seen to explain about 95% of the variability in the covariate space, and were retained as the final collection of covariates in the model.

3.5.1 Model-fitting using three models

Next, we compared the performance of the estimation methods for each of the models by applying them to the nucleosomal array data. All analyses were run in the statistical software R (http://www.r-project.org/). The MCMC procedure was initialized at values generated from a N(0,1) distribution for $\log(\lambda)$, $\log(\mu)$, $\log(\theta)$ $\log(\boldsymbol{\beta})$, ν_{00} and ν_{10}, while the variance parameters were initialized to 1. Multiple starting points were observed to make no significant difference in the results. Convergence was attained for all the models within 1000-5000 iterations. About 10000 iterations were used for each model after burn-in, for inference. The posterior estimates of the parameters were calculated from the mean of the lagged samples after burn-in, taken to be the first 10% of the iterations. We present below a subset of the analyses that indicate most strongly the power of the new approach. Table 3.1 gives the numerical summary of the parameter estimates obtained from all three models. The parameters corresponding to the second principal component variable was the only covariate that turned out to be significant at a 5% level in either state, in both models M1 and M2. On analyzing the data set with Model M3, MCMC convergence appeared to be very slow, hence we did not use this for further analysis. We used the same dataset consisting of 12760 probes for all the three methods. For this dataset,as we had mentioned earlier, the number of non missing probes varied. But the hierarchical model was able to take into account this discrepancy by being able to integrate out all the probe specific parameters at each probe level. We also ran the algorithm on a reduced data set where we had three non missing observations at each probe. The parameters obtained from there are very similar to the that obtained for the original data set, except for the values of σ_0 and σ_1. These two values are further separated out (.4 and .6 respectively) for the reduced data set. This indicates that the pattern of missig data is a little different between the two hidden states.
In the base model, M0, the parameter estimates for $\log(\lambda)$ and $\log(\mu)$ implied that if a particular probe was covered by a nucleosome free region, the probability of remaining in the state is very high, approximately about 0.98. These probability estimates support the belief that nucleosomes are separated by long nucleosome-free sequences. The nucleosomal states predicted also appear to be quite long; which is natural in this case as the low resolution of the data is motivated towards enrichment of nucleosome-free potential

regulatory regions, and do not allow for detection of short nucleosome free linker regions (< 10 bp) between adjacent nucleosomes, which are absorbed into the nucleosomal state.
In the transition model, M1, the estimates for the emission means and variances were exactly equal to that obtained in M0. The intercept estimates of the transition parameters matched quite well with the transition rates obtained from model M0. The second principal component was significant both in the the nucleosomal and NFR category (Table 3.1). The opposing signs of these estimates imply that the AT oligomers associated with this covariate help in continuation of the NFR subsequences and are detrimental to nucleosome formation. The estimates of other variables were not strictly negligible, but their absolute values were comparably lower than that of the second principal component variable. Recomputing the weights of the actual dinucleotide counts with respect to this variable, we see the following oligomers play an important role in the differentiation between nucleosomal and NFR regions. (All these oligomers have a weight of greater than .1): `A,T, AA, TT, AT,TA, AAT,ATA,ATT,TTA,TAT,TAA,TTTTA,TTTAT`.
The maximum weights were seen to be given to AT and TA simultaneously. The mono-nucleotides, the tetra-nucleotides and dinucleotides AA and TT had the lowest weights. The tri-nucleotides occupied the intermediate ranks. Interestingly, all statistically significant oligomers (of the AT combination) having the same number of nucleotides in their configuration, shared very similar weights.
In the emission model, M2, the estimates for the transition probabilities and variances were very similar to that obtained in the model M0. Again, the second principal component of the oligomer counts turned out to be the only significant variable. However unlike the transition model, this covariate was significant only in the NFR state.
We refitted the emission and transition models using only those 14 oligomer counts which were associated with the second principal component. The parameter estimates are given in tables 3.2 and 3.3
The starred values indicate that the estimates were less than .005.
AA and TT dinucleotides turned out to be the most significant variable for the emission model. As was in the case, when the principal components were fitted, none of the parameters corresponding to nucleosomes appeared to be significant. In fact many of them have starred values (less than .05) indicating that their effects were almost

negligible. This suggests that the intensity data within the nucleosomal segments is more or less uniformly distributed, however in the nucleosome free regions, the intensity mean is a variable function of the the AA and TT dinucleotide counts. It is interesting to observe that the effect of the other oligomer combinations, which were a part of the second principal component, vanishes when we refit the model with the oligomer counts. This implies that the effect of the second principal component was borne primarily by the counts of these two dinucleotide categories. The results from the transition model reiterates the sole importance of the AA and TT dinucleotides in influencing the state lengths. Here as in the PCA fitted transition model, these two features were important both for the nucleosomal and the NFR states. The other features though not individually significant, the combination may be significant in determining the states

Model comparison using the BIC. One important question is which of the models M0, M1 or M2 is most appropriate for the data. One possible criterion for model choice is the Bayes factor, which however would be difficult to compute analytically here due to the complexity of the model. A simpler alternative is the Bayesian Information Criterion (BIC) which can under many circumstances be considered an approximation to the Bayes Factor. It is straightforward to compute the BIC under the different models by the formula $BIC = -2\log(\hat{L}) + k\log(n)$, where $\hat{L}$ is the modal posterior likelihood, k is the number of parameters, and n is the number of data points (the log-likelihood is computed through a forward algorithm). The BIC for the three models are 17154.27 (M0) 14386.92 (M1) and 13337.19 (M2), showing that the incorporation of the sequence features was an essential part in determining the structural classification. The emission model, M2, turned out to have the best fit for this data set, indicating that local sequence features indeed influence nucleosome formation; however, the sequence does not exhibit as strong an effect in determining the lengths of the state of neighboring regions.

3.5.2 Comparison with known NFR regions from UCSC genome browser

As a biological validation step to check whether our algorithm correctly predicts known NFRs, we extracted the set of known regulatory regions in yeast from the UCSC data base (http://genome.ucsc.edu/). Although the set of all validated nucleosome-free regions

is not available, the presence of known active TF binding sites can serve as a useful indicator that the region is likely to be nucleosome-free. We looked at the overlap of our predicted nucleosome-free regions with the location of known TF binding sites
The results show an excellent performance for the continuous-index models as compared with the Yuan et al. (2005) method. Overall the base and transition model results perform very well (above 90% correct classification rate) against the database of known NFR regions. The predictive power of the emission model is also high (70%). Since we do not have a corresponding database of nucleosome regions, the false positive rate for all the models is underestimated. This might be one of the reasons why the emission model, M2, has a lower sensitivity than the base model, M0, in this comparison, although it fits the data better.

3.6 Simulation studies

In order to determine the power and robustness of our methodology, we next performed simulation studies to study the (i) consistency of model estimation under different parameter settings, (ii) detection of the correct model under model misspecification and (iii) importance of the continuous index assumption in the hidden Markov framework.

3.6.1 Consistency of model parameter estimation

Data sets were generated first under the three models M0, M1 and M2. Ten data sets of size 5000 probes were generated for each parameter setting under each of the models, and the models were fit using the MCMC procedure detailed in Section 3.3. In each case, the estimated bias and MSE of the estimators showed consistently accurate estimation of the model parameters. Below, as an example, we describe one such study.
For the base model, M0, the emission distributions were assumed to be normal with means -0.4 and 0.4, the transition parameters λ and μ were fixed at -3, and variances of the emission distributions were fixed at 0.35. For the emission and transition models, the intercept parameter corresponding to both states were fixed to 1. Next, each of the regression coefficients (θ for M1, and β for M2) corresponding to one of the five covariates were fixed to 1, while others were fixed to 0. This was done for both the nucleosomal and

nucleosome-free states. Thus a typical β (or θ) vector would be, for example, $(1, 0, 1, 0, 0, 0)$ for the nucleosomal state and $(1, 0, 0, 0, 0, 0)$ for NFR state. For all 3 models, 2500 iterations were done for each parameter setting, with a burn-in taken to be 20% of iterations. For purposes of evaluating the misclassification rates, the probes having an average state membership probability greater than 0.8 were rounded off to 1, while others were fixed at 0. For all data sets under the base model, M0, the MSE of all the emission parameters were less than 0.0001, the maximum MSE of the transition parameters was 0.03, and the maximum misclassification rate was 0, indicating the model was fit accurately in each case. For the transition model, M1, the maximum MSE for the corresponding parameters was also 0.03, with the maximum misclassification rate being 0.1. In addition, the MSE for the regression coefficient $\boldsymbol{\theta}$ ranged between 0.05 and 0.09. For the emission model M2, the maximum MSE for emission and transition parameters was slightly higher (0.08 for μ), the maximum misclassification rate was 0.1, and the range of MSE for $\boldsymbol{\beta}$ was .001 to .004. More details of these simulation studies are provided in the Supplementary materials.

3.6.2 Cross comparison under different models

Next, we simulated data under one specific model, and tried to estimate the parameters and states from a model different to the earlier model, in order to judge whether in each case the correct model was the one that most accurately fitted the generated data. For this set of simulations, the variances were fixed to 0.48 and 0.64 for the nucleosomal and NFR states respectively. The base model transition parameters were -3 and -3. For both the transition and emission models, M1 and M2, the parameter corresponding to the first principal component covariate in the NFR state was fixed to 1, while the other parameters were set to 0.

As expected, when the simulation and estimation models matched, the method overall performed very accurately, with the maximum MSE of the emission distribution parameters μ and σ being less than 0.01, and for λ and μ, less than 0.05. The classification rates are highest for the estimation models which match the simulation model. This pattern is also seen in the MSEs of the estimated parameters (more details in Supplementary materials).

We also computed the BIC for each analytical method under the condition that the data is simulated from one of these models (Table **??**). We used 5 data sets for computing the BIC under each simulation model. We then used the proportion of times BIC selected the model as a measure of model fit. to give a measure of the performance of an estimation method under a model. In general, we see that if we use the model for simulation in estimation, we tend to get the best results. Also given a simulation model, say M_A, if we try to estimate it by a model which contains it, say M_B, we get a very similar value of the maximal log likelihood. Although it is, in principle, erroneous to state that the same maximal log likelihood is achieved,however,in the ideal scenario of estimation, the extra components in the bigger model would turn out to be 0, and we would get the same likelihood This is not expected to be achieved in all data sets. However, in our case, we see that the log likelihood values are very close,(the error margin is less than .01) and the expected log-likelihood , obtained from the averages of the MCMC iterations are equal. This is so because of the large size of the data set. In another of our working papers' titled 'Asymptotics of continuous time Hidden Markov models' we have discussed how as in the iid cases, consistency results can be achieved for Bayesian estimates hidden markov models, as an extension of previous Bayesian consistency approaches. This implies that since our data set is sufficiently large, the errors in estimation have got smoothened out to a considerable degree, the parameters are actually equal to the true simulation parameters with a very high probability (a fact reflected in the table for mean squared errors and misclassification rates), and that the estimated log likelihood would be actually very close to the true log likelihood of the given model .
Five data sets were simulated under each model. The variances of the parameter estimates over the simulation sets were in the range of .01 to .04. The MSE s for each data set under a model was averaged to give an estimate of the estimation error of the parameters under a given model. The state classification percentages were similarly averaged over the five datasets. For all the five datasets, and for each simulation setting the BIC model chose the best estimation model to be the model under which the datasets were simulated. Thus the proportion of times BIC selected the model as a measure of model fit was always 100 percent for the simulation model. See Table 3.4
Due to issues with convergence, the fourth model proposed by us (Model M3) which had

both the transition and emission functions dependent on the covariates could not be tested. However we generated a total of 5 set from this model, and ran the other three estimation models on this simulated data to see how these models perform under this setup. For this set of simulations too, the variances were fixed to 0.48 and 0.64 for the nucleosomal and NFR states respectively. The base model transition parameters were -3 and -3. For both the transition and emission parameters, the parameter corresponding to the first principal component covariate in the NFR state was fixed to 1, while the other parameters were set to 0. The length of the simulated data sets were 5000. We took the sequence features of the first 5000 covariates of the Hogan data set to be the covariates. The MSEs and the BICs were averaged over all the five data sets to give a measure of the performance of the estimation models. We report the classification errors and the BICs in Table 3.5. In 2 of the five datasets, the transition model achieved the lowest BIC and MSE, while in the rest of the 3 the emission model scored the top position. In all five datasets , the base model M0 had the lowest rank.

3.6.3 Importance of the continuous index model

In order to show that the continuous index hidden Markov model was an essential improvement, required to fit the gapped probe data, we simulated a data set of length 5000, where the gap distribution was the same as in the FAIRE data hidden Markov model. It was then analyzed first by a continuous index HMM and then a discrete HMM. We then simulated a second data set of 5000 measurements where the gap structure was assigned randomly. This data set was analyzed using both discrete and continuous index models. Table 3.6 displays the results. The matching rate was calculated by comparing the number of prediction matches with the simulated state set. Clearly, we can see that the discrete HMM fails to capture the true picture in the gapped data framework.

3.7 Discussion

Researchers have been increasingly shifting focus to nucleosomes as a secondary source of information that will help in creating a more authentic map of the transcription factor binding site locations. Previous approaches to nucleosome positioning used gene signal

data to predict nucleosomes. Here, we propose a novel extension of the nucleosomal region prediction method with two major improvements: (i) use of a continuous-index Markov model to accommodate missing information or variably gapped probes; and (ii) use of underlying DNA sequence features that influence the transition rates and the emission densities. We have demonstrated the application of our proposed model on the data obtained from FAIRE, a method developed by Hogan et al. (2006), and now widely used for prediction of nucleosome free regions from the yeast genome. Our methods are reasonably efficient in terms of computer time, given their complexity– the time taken (in seconds) for 1000 iterations of the MCMC sampler for the models M0, M1, and M2 are 15343s, 47007s, and 18401s respectively.

Our first extension was in the form of a Bayesian continuous index hidden Markov model. Here we assumed that there is an underlying continuous index chain that generates the latent nucleosomal states, which we get to observe only in terms of emissions at the points where we have measured data. The transition probabilities are derived from the assumption that the probability of transition is linear with respect to index in an infinitesimal interval. We further extended the above model to create two new models M1 and M2 where the DNA sequence features influenced the transition rates and the emission densities respectively. For all three models, we proved the identifiability conditions, which allow us to construct an MCMC procedure that is simple to implement using minimally informative priors.

We employed Bayesian methodology in analyzing the model.Although the Bayesian setup allows us to incorporate prior information in addition what is provided by the data, we utilized minimal extraneous information. We employed powerful computational tools ,like MCMC and recursive likelihood computations which are now essential to any Bayesian analysis for their power in inference on complex posterior distributions.

In simulation studies, all models performed very well, with a correct classification rate of 1 in base and transition models, and above 0.9 for the emission model. The analyses were fairly insensitive to the assumptions of normality, as were seen by its performance under data sets simulated from the t-distribution.

The results of applying the models to the FAIRE data were compared with a set of known regulatory regions from the UCSC genome data base, as a surrogate for NFRs

(which have not all been marked with certainty). All three models performed fairly well, and most regulatory regions had a matching rate of 100 percent, with the corresponding predicted NFR regions. The performances of the transition and the base model were quite similar, and both were slightly better at predicting regulatory regions than the emission model, indicating that a non-homogeneous transition rate is probably most accurate for predicting nucleosome-free regions, unlike the simple HMMs that have been used in the past.

The non-zero parameter estimates from the emission and transition models clearly show that the DNA nucleotide combinations play a significant role in determining nucleosome positions. The covariates used in the model were obtained from principal component analysis of all nucleotide combination counts up to 4-mers (tetra-nucleotides). By computing the weights of the significant covariates with respect to the original collection of counts, we could determine the roles of specific nucleotide combinations. It turned out that the combinations of the nucleotides Adenosine and Thymine (A and T) are the most effective in determining nucleosome positioning. Of the different combinations, the dinucleotides (AT and TA) carry the maximum weight, followed by the tri-nucleotides. Our proposed models thus successfully extend the hidden Markov model methodology to gapped data sets, and by incorporating sequence information directly into the models, provide useful insights into the relationship between the DNA sequence and nucleosome positions. This also opens up the possibility for future work in synergistically modeling multiple complex data sets having differing intervals between probes.

Recently attempts have been made in the direction of combining data from multiple probe replicates into a hidden Markov framework (Johnson et al., 2009). Also, there have been efforts to develop more sophisticated technologies to overcome the major problems of ChIP-on-chip technology (large arrays and amplification bias) by performing massive parallel sequencing (Barski et al., 2007). However, such technologies are only in their preliminary stages of development and more critical refinement is necessary before their potential (and limitations) are fully realized.

Table 3.1: Panel 1: 95% Credible Intervals (CI) for parameter estimates for models M0 (base), M1 (transition), and M2 (emission). The indices l (or 0) and m (or 1) indicate the nucleosomal and NFR states; for instance, the term θ_{l0} refers to the intercept term for the nucleosomal state in the transition model M1, ν_{10} refers to the intercept term for the NFR state in the emission model M2. The CIs for parameters significant at a 95% level are given in bold fonts. Panel 2: Principal component weights of the significant covariate from the emission and transition models.
[Parameter Estimates]

	Parameter	95% CI	SE
	$\log(\lambda)$	**(-3.52,-3.75)**	.5
	$\log(\mu)$	**(-3.77,-4.19)**	.38
M0	ν_{00}	**(-.52,-.67)**	.08
	ν_{10}	**(.85,.99)**	.07
	τ_0	**(.45,.50)**	.02
	τ_1	**(.72,.79)**	.03
	σ_0	**(.25,.34)**	.04
	σ_1	**(.32,.38)**	.03
	θ_{l0}	**(-3.63,-3.75)**	.05
	θ_{m0}	**(-3.91,-.4.07)**	.05
	θ_{l1}	(.-.01,.02)	.01
	θ_{m1}	(.-.05,.05)	.04
	θ_{l2}	**(-.82,-.54)**	.68
	θ_{m2}	**(.78,.80)**	.1
M1	θ_{l3}	(-.04,.05)	.03
	θ_{m3}	(-.08,.07)	.07
	θ_{l4}	(-.03,.02)	.02
	θ_{m4}	(-.05,.04)	.04
	θ_{l5}	(-.02,.02)	.02
	θ_{m5}	(-.03,.03)	.03
	ν_{00}	**(-.54,-.63)**	.05
	ν_{10}	**(.88,.95)**	.03
	τ_0	**(.42,.51)**	.04
	τ_1	**(.72,.80)**	.04
	σ_0	**(.28,.33)**	.02
	σ_1	**(.31,.39)**	.02
	ν_{00}	**(-.54,-.56)**	.01
	ν_{10}	**(.92,.93)**	.01
	β_{01}	(-.05,.05)	.05
	β_{11}	(-.06,.05)	.05
	$\boldsymbol{\beta}_{02}$	(-.03,.03)	.02
M2	$\boldsymbol{\beta}_{12}$	**(.20,.22)**	.05
	β_{03}	(-.02,.01)	.01
	β_{13}	(-.03,.03)	.02
	β_{04}	(-.01,.01)	.01
	β_{14}	(-.04,.04)	.04
	β_{05}	(-.01,.01)	.01
	β_{15}	(-.02,.01)	.01
	$\log(\lambda)$	**(-3.53,-3.7)**	.25
	$\log(\mu)$	**(-3.85,-4.15)**	.15
	σ_0	**(.47,.49)**	.01
	σ_1	**(.74,.76)**	.01

Oligomer	Weight
A	.1
T	.1
AA	.12
TT	.13
AT	.17
TA	.18
ATA	.13
ATT	.14
TTA	.13
TAT	.13
TAA	.14
TTTTA	.1
TTTAT	.1

ht

Table 3.2: Significant oligomers for the emission model

Oligomer	NFR parameter (β_{1j})	Nucleosome parameter(β_{0j})
A	-.09	.02
T	-.08	.01
AA	.24	.02
TT	.35	.02
AT	.02	.01
TA	.06	.01
ATA	.04	-.03
ATT	.04	-.04
TTA	.03	.01*
TAT	.05	-.01*
TAA	..02	-.01
TTTTA	.02	-.01*
TTTAT	.02	-.01*

Table 3.3: Significant oligomers for the transition model

Oligomer	NFR parameter (θ_{1j})	Nucleosome parameter(θ_{0j})
A	.07	.05
T	.06	.05
AA	.25	-.18
TT	.17	-.10
AT	.09	-.04
TA	.10	-.07
ATA	.08	.01*
ATT	.06	.01*
TTA	.07	-.01
TAT	..02	-.02
TAA	.05	.04
TTTTA	.03	.01*
TTTAT	.04	.01*

Table 3.4: Simulation study:Proportion of datasets classified under the true model by the BIC criterion

		Simulation model		
		Base	Transition	Emission
Estimation Model	Base	**100**	.00	.00
	Transition	.00	**100**	.00
	Emission	.00	.00	**100**

Table 3.5: Tabulation of (a) correct state classification percentage (Match") and (b) Bayesian Information Criterion for model M3 under the three estimation models.

Model	Match	BIC
Base (M0)	81	21343.52
Transition (M1)	92	18279.65
Emission (M2)	94	18047.93

Table 3.6: Simulation study to compare the performance of discrete-index and continuous-index HMMs under two gap scenarios. The numbers "1" and "2" in the column "Set" refer to results obtained from data sets simulated under the FAIRE gap structure and the arbitrary gap structure respectively. "Match" refers to the percentage of probes classified into their correct state.

Model	Set	$MSE(\lambda)$	$MSE(\mu)$	$Ave_{MSE}(\nu_{00}, \nu_{10})$	$Ave_{MSE}(\sigma_0, \sigma_1)$	Match (%)
Continuous	1	.01	.01	.01	.01	99
Discrete	1	.03	.01	.01	.01	85
Continuous	2	.01	.01	.01	.01	99
Discrete	2	.08	.09	.05	.03	68

Chapter 4

A joint model approach to motif finding, using nucleosomal information

4.1 Introduction

As we have stated in the introductory chapters,finding functional DNA binding sites of transcription factors (TFs) throughout the genome is a necessary step in understanding transcriptional regulation. However, despite an explosion of TF binding data from high-throughput technologies like ChIP-chip Ren et al. (2000) Harbison et al. (2004), DIP-chip Liu et al. (2005), PBM Mukherjee et al. (2004), and gene expression arrays, finding functional occurrences of binding sites of TFs remains a difficult problem because the binding sites of most TFs are short, degenerate sequences that occur frequently in the genome by chance. In particular, matches to known TF motifs in the genome often do not appear to be bound by the respective TFs in vivo. This is because when the DNA is in the form of chromatin, not all parts of the DNA are equally accessible to TFs. In this state, DNA is wrapped around histone octamers, forming nucleosomes. In the first chapter ,our motivation for finding nucleosome positions was based on this specific role of nucleosomes to provide a mechanism for differential access to transcription factors . Indeed, it has been shown that functional binding sites of TFs at regulatory regions are typically depleted of nucleosomes in vivo Yuan et al. (2005)**?**Yuan et al. (2005). If we knew the precise positions of nucleosomes throughout the genome under various

conditions, we could increase the specificity of motif finders by restricting the search for functional binding sites to nucleosome-free areas. One way to achieve this would be to first obtain a prediction of the nucleosome positions, and then use it as a prior in motif search. Narlikar (Narlikar et al. (2007)) employed this approach to get an improved motif mapping of the yeast genome. Their main goal was to find transcription factor binding sites in sequences which are already known to bind to TFs For this they had to impose a prior which adequately reflected nucleosomal occupancy in w mers. This prior was constructed in an adhoc manner from nuclesomal occupancy scores. The scores themselves were obtained from a previous prediction algorithm by Segal et al. Each score gave the probability of a particular position being occupied by a nucleosome. These scores were then averaged over all w positions and normalized to give a positional prior probability for a w-mer starting a particular position. Next, a discriminative prior distribution was also used where a discriminative prior used. In this case, the probability of a w mer being accessible for transcription factor binding was calculated by taking a weighted ratio of the number of such w mers occurring at all positions in bound sequences over the total number total numbers of the occurrence of this particular w mer in all bound and unbound sequences. The weights were provided by the same positional cores used in the non-discriminative prior. It was shown that while nucleosome occupancy used as a simple positional prior only marginally improves the performance of a motif discovery algorithm, when it is used to compute a discriminative priortaking into account accessibility over the whole genome There were a number of problems with this method. First, the prior formulation was not based on likelihood of data of an external source or on the probability estimates of a w mer being occupied by a nucleosome. Instead it used a only the positional probability of a particular site to be occupied by a nucleosomes ,and used it as a proxy for the actual probability, by taking a simple average over the positions. This proxy can only work for short motif lengths, but for large w, the difference could be marked. Secondly the positional scores were dependent only on property of the distribution of dinucleotides within a nucleosome and hence missed other distinguishing sequence features. Thirdly and very importantly, the positional scores themselves were not obtained from a discriminative approach, but used the sequence distribution of nucleosomes only. So the power of the discriminative approach seem to

come only by the raw numbers of unbound sequences considered, along with the bound sequences, but the weights in itself do not emerge as an important factor because of the way the scores are built. The construction of a prior for nucleosomal positional estimates, though a useful technique, can suffer from one major drawback . The predictions for motif search lie would propagate the statistical uncertainties associated with prediction strategies directly into motif algorithms. This often yields results with low predictive power. The ideal would be to work with raw nucleosomal intensity data, and incorporate its likelihood into the broad Bayesian framework of a motif model. and Our approach in this paper is to use nucleosomal information, in a way such that these prediction biases do not arise. Also, we have addressed the goal of statistically quantifying the biological connection between motifs and signal strength . In order to achieve these,we have formulated an unique joint model that predicts nucleosome positions and motifs simultaneously, based on the gene expression and sequence data.

4.2 Methods

4.2.1 Data Structure and Assumptions

The joint model consists of four main components: the measured data: $\boldsymbol{Y}$, the log-intensities; $\boldsymbol{X}$, the sequence; and the unobserved (latent) components: $\boldsymbol{A}$, the indicator variable of all motif start positions, and $\boldsymbol{S}$, the set of latent nucleosomal states for each probe. The sequence data, $\boldsymbol{X}$, can be considered one long string of DNA, each element being a letter from A,C,G or T. The probe intensity data $\boldsymbol{Y}$, however, is at a resolution of individual probes, which consist typically of about 25-50 base pairs in our experimental data. However, in order to build a coherent joint model, we require data from both these sources to be at the same resolution. Hence we need to make a choice between two simplifying assumptions: (i) for all the base pairs within a single probe, $\boldsymbol{Y}$ has a constant probe specific value, or (ii) assign a value to each base pair of the probe (e.g. by numerical interpolation). If all probes are of equal length and equi-spaced the first option may work well, but since we typically have probes with gaps and of varying distance we choose option (ii), which also leads to including variability in the intensity over a probe.

4.2.2 Observed data and latent variables.

For the moment, consider the sequence as a single vector of length L, denoted by $\boldsymbol{X} = (X_1, \ldots, X_L)$. Each X_i $(i = 1, \ldots, L)$ takes values in the set $\{A, C, G, T\}$. Let $\boldsymbol{X}^{[a:b]}$ denote the subsequence $X_a, \ldots, X_b$. The signal intensity data, at the same resolution, are represented by $\boldsymbol{Y} = (Y_{11}, Y_{12}, \ldots, Y_{1j_1}, Y_{21}, \ldots, Y_{2j_2}, \ldots, Y_{Pj_P})$, where P denotes the total number of probes, and j_p $(p = 1, \ldots, P)$ is the number of basepairs corresponding to probe p (which is typically equal, but could vary in some cases, depending on experimental limitations), and $j_1 + j_2 + \cdots + j_P = L$. Along with $\boldsymbol{X}$ and $\boldsymbol{Y}$, we also define a vector $\boldsymbol{t}$ of length L, which contains the actual chromosomal location of the nucleotides corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$. For instance, if there is a gap of N nucleotides between probes $p1$ and $p2$, we will still denote the sequence corresponding to $Y_{p1,j_{p1}}$ as, say, some X_q, and the sequence corresponding to $Y_{p2,1}$ (the next probe measurement) as X_{q+1}; however, the value of $t_{q+1} = N$ would contain the information on the distance between the closest measured probes in this scenario. Next, $\boldsymbol{S} = (S_1, \ldots, S_P)$ denotes the indicator vector for nucleosomal state, with $S_p = 1$ if probe p belongs to a nucleosome-free region (NFR) and 0 if it belongs to a nucleosomal region. We can also write this in expanded form as an L-dimensional vector $(S_{11}, \ldots, S_{1j_1}, \ldots, S_{Pj_P})$, where S_{pj_p} takes the same value for $1, 2, \ldots, j_p$ $(p = 1, \ldots, P)$. Finally, let $\boldsymbol{A} = (A_1, \ldots, A_L)$ denote a latent vector of indicator variables with $A_i = 1$ if a motif site originates at position i of the sequence, and 0 otherwise.

4.2.3 Model formulation.

For simplicity, assume there is one motif type in the model, with w columns, characterized by a $4 \times w$ matrix of probabilities Θ. Column l $(l = 1, \ldots, w)$ of Θ is a vector $(\theta_{l1}, \ldots, \theta_{l4})'$, where θ_{lm} denotes the probability of observing the m-th base in the set $\{A, C, G, T\}$, and $\sum_{m=1}^{4} \theta_{lm} = 1$. We can generalize this to a model with, K motif types, characterized by weight matrices $\{\Theta_1, \ldots, \Theta_K\}$, of motif widths $w_1, \ldots, w_K$. Also, denote the *background* distribution of nucleotides by the parameter $\boldsymbol{\rho}$. In the simplest model, $\boldsymbol{\rho}$ is a vector of probabilities $(\rho_1, \ldots, \rho_4)$ of the 4 nucleotides; however, in complex genomes it may be necessary to use a Markov chain of moderate order (2 or 3) for

nucleotides due to the presence of long repeat sequences. We assume that each column l of a motif instance ($l = 1, \ldots, w$) is generated by a draw from the multinomial distribution characterized by parameter $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{l4})'$. Then, let $\Theta[\boldsymbol{X}^{[a:a+w-1]}]$, and $\boldsymbol{\rho}[\boldsymbol{X}^{[a:a+w-1]}]$ denote the probability of the segment $\boldsymbol{X}^{[a:a+w-1]}$ being generated from the motif and the background model respectively, that is, $\Theta[\boldsymbol{X}^{[a:a+w-1]}] = \prod_{l=1}^{w} \theta_{lX_{a+l-1}}$, and $\boldsymbol{\rho}[\boldsymbol{X}^{[a:a+w-1]}] = \prod_{l=1}^{w} \rho_{lX_{a+l-1}}$ (for an i.i.d. background). Additionally, π denotes the probability of occurrence of a motif at any position (only within the NFR regions). For the model of transition between the states of nucleosome (0) and NFR (1), we assume a continuous time Markov process characterized by a matrix of transition probabilities $P(t)$ for an interval t, where, we have, $P_{00}[t] = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)t}$, $P_{11}[t] = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)t}$. Next, the log-intensity data for probe p is assumed to be $N(\mu_{ps}, \sigma_s^2)$, where the index s denotes the nucleosomal state $s \in \{0, 1\}$. For the nucleosomal state, we assume a conjugate hierarchical model for μ_{p0}, such that $\mu_{p0} \sim N(\mu_0, \tau_0 \sigma_s^2)$. For the NFR state, the mean of the normal distribution at position a of the probe, μ_{pa} is modeled as a baseline value μ_1 linked to a parameter which associates the occurrence of motifs at that position of the probe to the observed intensity value. Specifically, we assume, conditionally on the latent variables $\boldsymbol{A}$ and $\boldsymbol{S}$, $\mu_{a1} = \mu_1 + \beta \log\left(\Theta[\boldsymbol{X}^{[a:a+w-1]}]/\boldsymbol{\rho}[\boldsymbol{X}^{[a:a+w-1]}]\right)$, where $1 \leq a \leq j_p - w + 1$. β is an unknown parameter measuring the direct association between signal data and the occurrence of motifs. The positive sign of the log-likelihood ratio translates into the scenario where the motif probabilities are different from the background; the higher the difference, the higher the strength of the motif. The priors for μ_0,μ_1 and β are assumed to be uniform (flat); and the priors for σ_s^2 are also taken to be flat $\frac{1}{\sigma_s^2}$ ($s = 0, 1$). For Θ and ρ we assume product Dirichlet and Dirichlet distributions with hyperparameters equal to 1. The only parameter for which we require an informative prior is the motif site prevalence π, since a non-informative prior for π typically leads to non-identifiability in practice (Gelfond et al., 2009). For π we thus use a relatively strong Beta$(\delta(1-\gamma), \delta\gamma)$ prior where δ is a large "pseudocount" and γ ($0 < \gamma < 1$) is a "prior expected value". Based on preliminary studies in a number of other studies including Gelfond et al. (2009), we propose to choose γ over a range of values between 10^{-5} and 10^{-3} which seem to be reasonable, and test the sensitivity of the final inference to such choice. Let us denote by η the set of all non-motif related parameters in the model, that

is, $\eta = (\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, \beta)$. We denote by ψ all the transition parameters. That is, $\psi = (log(\lambda), log(\mu))$ Now, the complete data likelihood can be expressed in the form: $L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{A}|\eta, \Theta, \boldsymbol{\rho}, \pi) = P(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{X}, \eta, \Theta, \boldsymbol{\rho}, \pi)P(\boldsymbol{X}|\boldsymbol{A}, \Theta, \boldsymbol{\rho}, \pi)P(\boldsymbol{A}|\boldsymbol{S}, \pi)P(\boldsymbol{S}|\lambda, \mu)$. The problem of interest is to estimate the latent variables $\boldsymbol{A}$ and $\boldsymbol{S}$, and the unknown parameters. Given the complex form of the likelihood, the estimation procedure we propose is a hybrid MCMC algorithm, that incorporates elements of recursive sampling-based data augmentation for efficient inference.

4.3 Estimation Procedure-The Setup

We implement Gibbs sampling by drawing from

1. $p(\boldsymbol{S}|\eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$
2. $P(\boldsymbol{A}|\boldsymbol{S}, \eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$
3. $P(\eta|\Theta, \pi, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{X})$
4. $P(\pi|\Theta, \pi, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{X})$
5. $P(\psi|\boldsymbol{S})$

4.4 Estimation Procedure

As with any Gibbs Bayesian procedure, computation of conditional probabilities play a major role. Since these probabilities are analytically intractable, a recursive algorithm is used. Below we present three forward algorithms forward algorithm for computing the unconditional likelihood. n this section we present three forward algorithms that would be useful in the sampling stage

1. Forward Algorithm I for computing the likelihood –$L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X})$. Note that in computing this quantity we are integrating out $\boldsymbol{A}$ from the posterior. This algorithm recursively computes the likelihood upto each position in the data and for each positions outputs a two component vector

$(L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 0), L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 1))$. The sum of these two terms give the complete likelihood for position i. The N by 2 matrix of likelihoods thus obtained is used in sampling from $p(\boldsymbol{S}|\eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$

2. Forward algorithm II for computing the Likelihood conditioned on states—$L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S})$. This is used in sampling from $P(\boldsymbol{A}|\boldsymbol{S}, \eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$

3. Forward algorithm III for joint sampling of $\boldsymbol{S}$ and $\boldsymbol{A}$. This also computes the unconditional likelihood but unlike (1) it outputs its results as a three component vector for each position i. $L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 0), L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 1\boldsymbol{A}[i] = 1)L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 1\boldsymbol{A}[i] = 0))$. The previous two algorithms can be implemented one after the other in order to sample from $p(\boldsymbol{S}, \boldsymbol{A}|\eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$. This is because the kernel of the joint posterior is proportional to $p(\boldsymbol{S}|\eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})P(\boldsymbol{A}|\boldsymbol{S}, \eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$. However, Forward algorithm III samples from this joint posterior directly with the help of N by 3 matrix of likelihoods. In this way it presents an alternative to the combined work of the first two algorithms.

4.4.1 Forward Algorithm I

Let F[i,k] denote the likelihood till position i of the data, with the position i being in the k^{th} state k= 0,1.

$$F[i, k] = L_i(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = k)$$

The following equations compute the likelihood for different values of i.

1. $1 < i < w$:

$$F[i, 1] = \prod_{j=1}^{i} (\phi(y_i, \mu_0, \sigma_0)\rho_{\boldsymbol{X}_j}) \tag{4.4.1}$$

$$F[i, 2] = \prod_{j=1}^{i} (\phi(y_i, \mu_1, \sigma_1)\rho_{\boldsymbol{X}_j}) \tag{4.4.2}$$

2. i=w:

$$F[w,1] = (1-\pi)\prod_{j=1}^{w}(\phi(\boldsymbol{Y}_w, \mu_0, \sigma_0)\boldsymbol{\rho}[\boldsymbol{X}^{1:w}])$$

$$F[w,2] = (1-\pi)\prod_{j=1}^{w}(\phi(\boldsymbol{Y}_j, \mu_1, \sigma_1)\rho_{X_i}) + \pi(\phi(\boldsymbol{Y}_{1:w}, \mu_{1a}, \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}])$$

$$F[i,2] = (1-\pi)[F[i-1,1](1-\lambda) + F[i-1,2](\mu)]\phi(\boldsymbol{Y}_i, \mu_1, \sigma_1)\rho_{X_i}$$
$$+(\pi)[F[i-w,1](1-\lambda) + F[i-w,2](\mu)](\phi(\boldsymbol{Y}_{i-w+1:i}, \mu_{21}, \theta_m, \theta_b), \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}])$$

4.4.2 Forward algorithm II

Let G[i] denote the likelihood of having a motif or a background at the position i, conditional on the states. That is $G[i] = L(\Theta, \boldsymbol{\rho}, \eta|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S})$

$$\boldsymbol{S}[1] = 1 \Rightarrow G[1] = \theta_b[seq[1]]$$

$$\boldsymbol{S}[1] = 0 \Rightarrow G[1] = 0$$

The computation is done differently for different values of i.

1. $2 < i < (w-1)$:

$$\boldsymbol{S}[i] = 0 \Rightarrow G[i] = G[i-1]$$

$$\boldsymbol{S}[i] = 1 \Rightarrow G[i] = G[i-1]\rho_{\boldsymbol{X}_i}\phi(\boldsymbol{Y}_i, \mu_0, \sigma_0)$$

2. i=w :

$$\boldsymbol{S}[i] = 0 \Rightarrow G[i] = G[i-1]$$

$$\boldsymbol{S}[1:w] = 1 \Rightarrow$$

$$G[i] = (1-\pi)G[i-1]\rho_{\boldsymbol{X}_i}\phi(\boldsymbol{Y}_i, \mu_0, \sigma_0)$$
$$+\pi(\phi(\boldsymbol{Y}_{i-w+1:i}, \mu_{a1}, \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}])$$

$$\boldsymbol{S}[i] = 1 \text{ and} (\boldsymbol{S}[1:w] \neq 1) \Rightarrow G[i] = (1-\pi)G[i-1]\phi(\boldsymbol{Y}_i, \mu_1, \sigma_1)\rho_{X_i}$$

3. $i > w$:

$$\boldsymbol{S}[i] = 0 \Rightarrow G[i] = G[i-1]$$

$$\boldsymbol{S}[i-w+1:i] = 1 \Rightarrow$$

$$\begin{aligned} G[i] = (1-\pi)G[i-1]\rho_{\boldsymbol{X}_i} * \phi(\boldsymbol{Y}_i, \mu_1, \sigma_1) \\ +\pi(\phi(\boldsymbol{Y}_{i-w+1:i}, \mu_{a1}, , \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}]) \end{aligned}$$

$$\boldsymbol{S}[i] = 1 \text{ and} (\boldsymbol{S}[1:w] \neq 1) \Rightarrow G[i] = (1-\pi)G[i-1]\phi(\boldsymbol{Y}_i, \mu_1, \sigma_1)\rho_{X_i}$$

4.4.3 Forward algorithm III to be used for joint sampling

Let H[i,1] $= L_i(\Theta, \boldsymbol{\rho}, \eta | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 0)$
H[i,2]$= L_i(\Theta, \boldsymbol{\rho}, \eta | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 1\boldsymbol{A}[i] = 1)$
H[i,3]$= L_i(\Theta, \boldsymbol{\rho}, \eta | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}[i] = 1\boldsymbol{A}[i] = 0))$.
The initial conditions are :

$$H[1,1] = h[1,2] = .5$$

The following equations compute the likelihood for different values of i.

1. $2 < i <= w$:

$$H[i,1] = [H[i-1,1]\lambda + H[i-1,2](1-\mu)]\phi(\boldsymbol{Y}_i, \mu_0, \sigma_0)\rho_{X_i}$$

$$H[i,2] = [H[i-1,1](1-\lambda) + H[i-1,2](\mu)]\phi(\boldsymbol{Y}_i, \mu_0, \sigma_0)\rho_{X_i}$$

$$H[1:w-1,3] = 0$$

$$H[w,3] = \pi(\phi(\boldsymbol{Y}_{1:w}, \mu_{a1}, \sigma_1)\Theta[\boldsymbol{X}^{1:w}])$$

2. $i > w$

$$H[i,1] = [H[i-1,1]\lambda + (H[i-1,2] + H[i-1,3])(1-\mu)]\phi(\boldsymbol{Y}_i, \mu_0, \sigma_0)\rho_{X_i}$$

$$H[i,2] = [H[i-1,1](1-\lambda) + (H[i-1,2] + H[i-1,3])(\mu)]\phi(\boldsymbol{Y}_i, \mu_1, \sigma_1)\rho_{X_i}$$

$$H[i,3] = [H[i-w,1](1-\lambda)^w + (H[i-w,2] + H[i-w,3])(\mu)^w]\pi(\phi(\boldsymbol{Y}_{i-w+1:i}, \mu_{a1}, \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}])$$

4.5 Sampling scheme and Conditional distributions

The main steps of the sampling procedure are:

1. $p(\boldsymbol{S}|\eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$
2. $P(\boldsymbol{A}|\boldsymbol{S}, \eta, \Theta, \boldsymbol{\rho}, \pi, \boldsymbol{Y}, \boldsymbol{X})$
3. $P(\eta|\Theta, \pi, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{X})$
4. $P(\pi|\Theta, \pi, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{X})$

The details for each of these steps are given below:

- $P(\boldsymbol{S}|\eta, \Theta, \boldsymbol{\rho}, \boldsymbol{Y}, \boldsymbol{X}, \pi)$
  Sample the nucleosomal states starting from the end of the sequence Use forward algorithm I. Recall that N is the entire length of the sequence Draw $\boldsymbol{S}$[N] with probability
  $$\frac{F[N,1]}{F[N,1] + F[N,2]}$$
  Conditioned on $\boldsymbol{S}$[i+1] Draw $\boldsymbol{S}$[i] with probability
  $$\frac{F[i,2]T[\boldsymbol{S}[i], \boldsymbol{S}[i+1]]}{F[i,1]T[2, \boldsymbol{S}[i+1]] + F[i,2]T[1, \boldsymbol{S}[i+1]]}$$

- $P(\boldsymbol{A}|\boldsymbol{S}, \eta, \Theta, \boldsymbol{\rho}, \boldsymbol{Y}, \boldsymbol{X}, \pi)$
  Sample the motifs conditional on on the states. In order to achieve this, we use the results of the forward algorithm II.

  Sample backward starting from position L.

At a given position i, compute

$$P = \pi(\phi(\boldsymbol{Y}_{i-w+1:i}, \mu_{a1}, \sigma_1)\Theta[\boldsymbol{X}^{i-w+1:i}])$$

The sampling probability that this position is a motif ending position is given by $\frac{P}{G[i]}$

If the motif position is selected, set a[i-w+1:i]=1. Or in other words declare the last w positions as a motif region. Else move back one position.

Set i=i-1.

Stop when $i < w$.

- $P(\eta|\boldsymbol{S}, \boldsymbol{A}, \Theta, \boldsymbol{\rho}, \pi)$

  Conditioned on the motifs and states, we sample the emission parameters as follows:

$$\sigma_0^2 \sim Gamma((n_0 - 1)/2, \sum_{i;\boldsymbol{S}[i]=0} (\boldsymbol{Y}_i - \boldsymbol{Y}_{m_0})^2/2)$$

  where n_0 is the number of positions in the nucleosomal state
  and $\boldsymbol{Y}_{m_0} = \sum_{i;\boldsymbol{S}[i]=0} \boldsymbol{Y}_i/n$ the average of the signal strength in the nucleosomal states.

  Sample

$$\mu_0 \sim N(a, b)$$

  where

$$a = \frac{\sum_{i;\boldsymbol{S}[i]=0} \boldsymbol{Y}_i}{n_0}$$

  and

$$b = \sigma_0^2$$

  Sample

$$(\mu_1, \beta) \sim MultivariateNormal(\hat{b}, \sigma_1^2)$$

  where $\hat{b}$ is the least squares estimate of the vector of regression from regressing $\boldsymbol{Y}_1$ on Z_1. Here, X_1 is a n_2 by 2 matrix . The first column of Z_1 is a one vector.

The second column of X_1 takes the value 0, when $\boldsymbol{A}[i]$=0.
Mathematically,

$$\hat{b} = (Z_1' Z_1)^{-1} Z_1' \boldsymbol{Y}$$

$\hat{\boldsymbol{Y}}_i$is the predicted value in position i, defined as $Z_1[i] * \hat{b}$

$$\sigma_1^2 \sim Gamma((n_1 - 1)/2, \sum_{i;\boldsymbol{S}[i]=1} (\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i)^2/2$$

- $P(\Theta|A, E)$
  The conditional posterior distributions are

$$\Theta \sim \prod_{i,j;\boldsymbol{A}[i:j]=1} [\Theta[\boldsymbol{X}^{i:j}] \prod_{k=i}^{k=j} \phi(\boldsymbol{Y}_k, \mu a1, \sigma 1) \tag{4.5.1}$$

$$\theta_b \sim \prod_{i,j;\boldsymbol{A}[i:j]=10} [\boldsymbol{\rho}[X^{i:j}]] \tag{4.5.2}$$

In order to sample Θ and $\boldsymbol{\rho}$ we needed to make a transformation of the 4*w probabilities of the position weight matrix . For each such probability, $\theta_{il}(l = 1 \ldots w, j = 1 \ldots 4)$ the logit function $p` = log((p)/(1 - p))$ will be used. That is, we define a new 4 by w matrix L and a new 1 by 4 vector R as follows

$$L_{il} = \frac{log(\Theta_{il})}{1 - log(\Theta_{il})} \tag{4.5.3}$$

$$R_{il} = \frac{log(\rho_i)}{1 - log(\rho_i)} \tag{4.5.4}$$

A Metropolis Hastings algorithm is then applied to sample the new parameters under the constraint

$$\sum_j \Theta_{jl} = 1 \forall \text{ l}$$

and,

$$\sum_j \rho_i = 1$$

where j is the nucleotide index .
The Metropolis Hastings algorithm is implemented in the following steps.

1. Start with l= 1.
2. Generate a 3 x 1 random vector N_l from Multivariate Normal $(R_l, 0.4I)$ where R_l is the vector (R_{1l}, R_{2l}, R_{3l})
3. Compute $J_{il} = \frac{exp(R_{il})}{1+exp(R_{il})}$
4. If $\sum_i J_{il} > 1$ then go to the previous step.
5. If $\sum_i J_{il} < 1$ then calculate $J_{4l} = 1 - \sum_{i=1}^{3} J_{il}$
6. Recompute N_l by appending $\frac{log(J_{4l})}{1-log(J_{4l})}$
7. By inverse transformation calculate the new motif matrix Θ^n where the l^{th} column of Θ^n is the inverse logit transform of N_l and all other columns remain the same.
8. Compute Ratio = $\frac{L(\Theta^n,\boldsymbol{\rho},\pi\eta|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S})}{L(\Theta,\boldsymbol{\rho},\pi\eta|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S})}$
9. Accept new matrix Θ^n with probability min(1,R)
10. Repeat the above steps for $l = 2, 3, \ldots w$
11. Update $\boldsymbol{\rho}$ in the same way. Note that, in place of w columns, we have to update only one vector.

- $P(\pi|\boldsymbol{A})$
  We sampled π from the Beta distribution, $Beta(n_1 - nsg + 1, n_1 + 1)$ where $nsg = \frac{|\boldsymbol{A}|}{w}$,which is equal to the number of motifs sampled in the previous step.

- $P(\psi|\boldsymbol{S})$
  Since there is no closed for for the kernel of the posterior given by $L(log(\lambda), log(\mu)|\boldsymbol{Y}, \boldsymbol{S})$ we resort to a Metropolis Hastings Scheme. The jump distribution was taken to be Gaussian,with a the variance of the jump denoted as V. The scheme is given below.

  1. Generate a sample $\log(\lambda_{new}) \sim N(\log(\lambda_{old}), V)$.
  2. Compute the ratio $R = \frac{L(log(\lambda_{new}),\mu,\eta_b|\boldsymbol{Y})}{L(log(\lambda),\mu,\eta_b,|\boldsymbol{Y})}$

3. Accept λ_{new} with probability min(1,R).

The same steps are repeated for μ. Note that in the above sampling procedure we have used a combination of Forward Algorithms I and II (i.e we have sampled out the states unconditionally first, and then the motifs) as an alternative to using the Forward Algorithm III. Initially the way the forward algorithm is formulated, forward algorithm III took more computational time than algorithms I and II combined. This is because a three by N matrix has to be defined and accessed throughout the algorithm. However, we see that essentially the time can be much reduced, because in the recursion equations, H[i,1] and H[1,2] occur together as a common summand. This motivated us to implement the same forward algorithm II without taking recourse to defining a new matrix, but using F[i,1] and F[i,2] of the forward algorithm I. That is, the idea is not use the three state representation in the computational stage. In computing the likelihood, we use only the two way representation of whether the state is in a nucleosomal position or not. But while sampling for say position i, we do not use F[i,1] or F[i,2] but go back a number of steps F[i-1,1] F[i-1,2] F[i-w,1] F[i-w,2] and use them in computing the three probabilities for the three possible scenarios in the position i. Here we are doing a joint sampling but without having a new matrix to define. This procedure cost us less computational time (44 seconds) instead of the combination of I and II(76 seconds). However we have seen that Forward Algorithm III , does not typically takes more time to converge than the combination of I and II. Because in both cases we are doing a block sampling of motifs and states together. In I and II , we are achieving this by by first doing $P(\boldsymbol{S}|.)P(\boldsymbol{A}|\boldsymbol{S}|.)$ and in the third we are doing $P(\boldsymbol{A}, \boldsymbol{S}|.)$ Both of them are improvements on trying to execute the full conditionals as in a conventional Gibbs sampler.

4.6 Simulations

A set of simulation studies were designed to test the efficacy of our sampling scheme. The main objective for this design of simulation studies is to ensure that our algorithm has the potential to work on a wide variety of data sets . Though, in essence, the data components would remain the same in the sense that it would comprise of the nucleosome intensity signal, and the sequence probes, there would differences in overall pattern across

different genomes This can be due to sparseness of signals, weak motifs , low motif occurrence probabilities, and varying strengths of the links between motifs and intensity. We would like to know whether in different settings as above, our algorithm is able to capture the true reality, and if so, to what extent. Is there a trend in the efficiency of this algorithm with an increase or decrease in the corresponding parameters? There is also the need to question how far the distributional assumptions that we have employed modeling the intensity is valid, and if we are doing a good job with large numbers. That is, though we have to face the computational burden of dealing with larger datasets, are we being able to estimate the underlying the parameters with greater precision? With these goals in mind we have designed the following simulation studies.

1. Checking the trend of convergence with increase and decrease in motif occurrence probability (π) [Section 4.6.2]
2. The influence of initial values on the MCMC chain, in particular, the issues with initializing the motif matrix.[Section 4.6.3]
3. Consistency: The errors of our predictions in increasingly large data sets.[Section 4.6.5]
4. Changing distributional assumptions[Section 4.6.4]
5. Varying β and π simultaneously.[Section 4.6.6]

4.6.1 Parameter Settings and inference methods for the simulation studies

For all these simulation studies,we restricted the length of the data set to 20000. We fixed $\boldsymbol{\rho}$ to (.25,.25,.25,.25) . The parameter values are taken to be

Table 4.1: Parameter values

(λ)	(μ)	(μ_1)	(μ_2)	(σ_1)	σ_2	(β)	(π)
.8	.57	-.6	.9	.48	.64	1	.01

The transition parameter values were chosen so as to get the NFR coverage region to be 30 percent, approximately. This basic parameter setting was used through out for all simulation studies. In the following subsections, where the values for parameter needed to be varied, the range has been specified. We used two simulation matrices for generating the motif matrix, which were obtained from the JASPAR data base. The matrices are the following

Table 4.2: Motif matrix

	1	2	3	4	5	6	7	8
1	0.40	0.10	0.75	0.80	0.15	0.04	0.04	0.10
2	0.40	0.06	0.05	0.08	0.02	0.04	0.04	0.10
3	0.10	0.80	0.10	0.08	0.82	0.04	0.04	0.10
4	0.10	0.04	0.10	0.04	0.01	0.88	0.88	0.70

and

	1	2	3	4	5	6	7	8
1	0.40	0.35	0.20	0.12	0.35	0.60	0.20	0.10
2	0.40	0.15	0.20	0.18	0.35	0.10	0.50	0.10
3	0.10	0.30	0.40	0.35	0.12	0.20	0.15	0.10
4	0.10	0.20	0.20	0.35	0.18	0.10	0.15	0.70

In the second table the weights are a little more uniformly distributed across the nucleotides than the first table.
In all the simulation settings in the current and the following subsections, the Auto Correlation function drop was taken as the measure of convergence. we marked the threshold iteration value from where the ACF of all the parameters dropped down to less than .1. We then took a burn in of 2000 samples and used the mean of the parameters as the posterior estimate.
The sensitivity rates are calculated by dividing the number of true positives by the total number of 'true' simulates sites. Specificity rates are calculated by dividing the number of true positives by the number of predicted positives. The average Mean squared errors of

the motif parameters were calculated on the basis of posterior sample mean and variance.

4.6.2 Checking convergence pattern with π

One important issue that could arise in the motif discovery algorithms is that we might be dealing with sequences where the motifs occur very rarely. In those extreme cases, many motif discovery algorithm could break down , and the resultant motif sensitivities would turn out to be very poor . By trying to check our convergence pattern by changing the π values we tried to determine the breaking point of our algorithm, the exact reasons and issues associated with the break-down , and methods to remedy the situation. We monitored the π values , reducing it step by step, and tried to see the convergence pattern. The parameter values were fixed at the numbers set in table 4.1. The motif matrix used was the first table from the JASPAR database, Table 4.2
The results are given in table 4.3:

Table 4.3: Convergence pattern

(π)	Num-iterations	Motif Sensitivity	Motif specificity.
.01	840	.88	.92
.009	930	.85	.91
.008	1150	.88	.93
.007	1200	.82	.87
.006	1350	.84	.86
.005	1500	.89	.93
.004	1700	.88	.92
.003	2000	.84	.90
.002	2200	.87	.1

When we further decrease the π value beyond .001, we see a breakdown of the algorithm. At a π value of .0008 we see that no motif sites are getting selected when the iteration number crosses 500. The algorithm breaks down because the the β parameter is sampled from a regression of the intensity data on a matrix which is a function of the motif covariates. In absence of motif covariates, the second column reduces to 0 and we get a less than full rank matrix, which cannot be inverted to give a estimate of β. Under a flat

prior the posterior sample would have generated from a normal distribution with this least squares estimate as the mean. However it is not possible under a circumstance where one column of the regression matrix is 0. This problem was circumvented by taking a vague normal prior , with mean 0 and a variance of (10^6) around β. Due to the vagueness of the prior used for β the difference in parameter estimates obtained under this modified posterior and that of the values obtained (where π was fixed at .01) was negligible. The differences were less than .01. We checked this pattern by repeating the simulation for values of $\pi = .004, .008, .005$. However as we see in Table 4.3, the lower the π value the more number of iterations needed to converge. This is because in the algorithm output there are quite a few runs where the β parameters gets only the random numbers generated the vague prior distribution in the absence of a motif, and the chain gets stuck in a certain pocket. And it needs a certain amount of time (as in the case of $\pi = .002$) when the state transitions parameters are sufficiently updated to steer the convergence in the correct direction.

4.6.3 Initial values for the MCMC chain

Initial values affect the convergence of a MCMC in a number of ways. It is well known that if we start close to the initial values, we would get convergence faster than if we would have done farther away. Sometime the convergence can take unusually long for an initial value which is at a large distance from the true parameters. Sometimes the chain would get stuck in a local mode due to particular set of initializations. Th motif matrix initialization is important and sometimes a weak motif initialization in a MCMC where the dataset is actually simulated under a strong motif can slow down the chain for a considerable period of time and lead to issues encountered in the previous subsection. Here we examine the different scenarios by plugging in different sets of initial values. The initial values for μ_1, μ_2, β were simulated from N(0,1) The variance parameters σ_1 and σ_2 were set at 1. The background parameters were generated uniformly from the unit simplex in R^4 This generation was done by generating three U(0,1) random numbers. Upon the condition that the sum is less than 1, the fourth number in θ_b vector was set by subtracting the sum of the three from 1.
A number of initial motif matrices were tried out. However not all of them led to

convergence Initially we tried to fix the 3 columns of the initial motif matrix to be the same as the generator motif matrix in the simulations. In all these cases we achieved convergence. Then we tried to generate a set of initial matrices by randomly permuting the numbers in each column of the generating motif matrix. When the initial motif configuration was such that the highest probability occupying nucleotide mismatched in more than 5 out of 8 positions, we ran into a situation where the number of motifs selected was 0. The second Jaspar motif matrix in Table 4.6.1 was used as a proxy for a motif matrix which distributed weights more uniformly. A set of initial matrices were generated from this second motif matrix by the same way of randomly permuting the row numbers within each column of the matrix This set did well with respect to convergence in all situations, when the motif matrix generator was the fixed second Jaspar motif matrix (Table 4.6.1). But it failed in some situations (i.e in some of the random settings) where the simulation motif matrix was the first motif matrix. Again the problem in convergence was mainly due to no motifs getting selected. And this was successfully taken care of by assuming a vague normal prior on β. A second technique employed in getting round this problem is to get a estimate of the transition parameters without assuming a motif in the estimation model. This would boil down to running a simple Hidden Markov Model(of the type used in Chapter 3. Taking estimates of these parameters after 500-1000 iterations, these values were then used with a randomly generated motif matrix model. This would always yield good convergence results . The convergence was achieved at less than 1500 iterations of running the full joint model with these updated estimates. However, when we applied the algorithm on real data, we employed the second method. The size of the data set acts hindrance which increases the computational time considerably because we had to run a precursor to the main algorithm just for generating the initial values. It has been observed (from one simulation setting with π fixed at .007) that this increase in computational time is less compared to the time spent in convergence for the MCMC with a non-flat β prior. However it is not possible to draw a general conclusion, regarding comparison of these two methods.

4.6.4 Changing distributional assumptions

Our model is based on the imposition of a Gaussian density on the intensity function. One of the motivations for this specification was that a Gaussian density would give us simple closed form for many full conditional posteriors in the sampling stage. Another motivation was that the log intensities are generally obtained from centering and scaling around several data points, and it is reasonable to assume that in the processed data, each data points corresponds to a large sample average, and hence a Gaussian approximation can hold true. There is also the evidence from the qq plots in Chapter 1, that the log intensities behave approximately normal. However we would liek to see how far the deviation from this assumption would affect our estimation results. We tried to change the distributional assumptions and check the robustness of our methods. For this, we generated data from a t distribution with 5 and 10 degrees of freedom . The results in 4.4 show that the model estimation procedure is pretty robust to the distributional assumptions. We see an improvement in motif specificities and sensitivities with an increase in the degrees of freedom of the T distribution. All other parameters were fixed at the values set in Table 4.1 and the motif matrix was fixed at the matrix given in Table 4.2

Table 4.4: Robustness

T degrees of freedom	Motif Sensitivity	Motif specificity.
5	.82	.89
10	.86	.91

4.6.5 Consistency

Increase in the size of the data set would lead us to more precision in estimation of our parameters. This is the basic notion of consistency and is known to be valid for frequentist estimates in iid and other dependent model extensions. Here we would like to test the precision of our Bayesian estimate with the increase in the size of our data set. It must be noted that there is no procedure to check on consistency from a Bayesian model

in itself. This is due to that fact that we do not have any closed form for the Bayesian estimates on which we can apply the limits. The only option is to check the precision of the MCMC estimates and compare it to the true parameter.(This would be known in a simulation setting) and see the variance of the estimate over a number of data sets simulated under the same setting. Thus we tried to see how close our estimates get to the real parameter values when the data size increases. In table 4.5 we see a strict pattern– the mean square errors decrease with increase in the lengths of the data sets. The other parameters were fixed at the values set in table 4.1 This clearly points to the consistency of the Bayesian estimates obtained through MCMC.

Table 4.5: Consistency of Parameter Estimates

Data length	Average-MSE-Transition	Average-MSE-Emission.
20000	.007	.004
30000	.006	.001
50000	.005	.001
70000	.003	.000
90000	.001	.000

4.6.6 Varying Beta and Motif occurrence Probability simultaneously

The simulation studies were designed with the purpose of investigating how the parameter estimates behave under different settings. One important object of finding was to to have a minimum bound or threshold on the motif probability π under which the algorithms converge. The other was to check the bias precision of parameter estimates obtained under decreasing or increasing levels of individual certain parameters. This was followed by a simulation setting where a combination of parameters were used . (Although the direction of biases and precision are expected to follow a set pattern in certain scenarios, the pattern obtained in all cases does not lend itself to an easy interpretation).
Along with this, we have tried to obtain the conditions under which the motif width is misspecified, and have obtained an estimate of its mis specifications on the the transition , emission and state estimates. We have also tried to see the effect of the conservation

pattern of a motif has on the variance of the estimates and the classification errors. This is done by considering a specific set of motif matrices.

1. Motif Category 0. We used the motif matrix of the Jaspar database that we used in earlier simulations

2. Motif Category A: This motif has a fixed width 7. We randomly chose 4 positions of the motif columns and to each column arbitrarily assigned a unique nucleotide to have the highest probability. This probability was taken to be .75 The other nucleotides in those columns were assigned a probability of .082. From the remaining 3 columns we assigned two columns to have nucleotide probabilities .5 and .3 in the first two rows (A and C) and equal probabilities in the remaining dinucleotides. We assigned a .8 to T in the last column and equal probabilities in the remaining rows.

3. Motif Category B: This motif also has fixed width 8. We choose any four columns and to each of these columns assign the probabilities (.5,.4,.05,.05) in random order. For the remaining 4 columns we assign .3 randomly to any of the three rows and .1 to the remaining row.

4.6.7 Motif Categories 0 and A

We first performed our simulations on motif category A and O. The transition and emission parameters were fixed at the values specified in 4.1
Keeping all other parameters constant, we changed the β value over the range (1.2-2.7). The β values were increased by an increment of 0.2 till 2 . We appended the values 2.5 and 2.7 later to the list to detect any changes over a greater range. There were four π values (.002,.003,.005 and .006) for each β setting.
We see that there is an increase in motif sensitivity and state classification rate when we increase β fixing a π value. The pattern of increase is similar over different π values till .006. Also fixing a β we see that the motif sensitivities and state classification rates increase over π. However at .006 we see break in the increasing pattern of sensitivities and state matching with β. We observe that at this π value there is an increasing trend

of motif sensitivity till $\beta = 2.5$ Although before this stage we have a considerable increase in motif sensitivity till $\beta = 2$. Similarly for the state classification rates, a slight decrease in motif sensitivity there we have an increased motif sensitivity till $\beta = 2.1$ and then a decreasing trend.
There was no trend observed for motif specificities, although we could see a somewhat increasing trend (with increasing β and π values) till $\pi = .004$.
The plots in the figure **??** gives us the pattern of change in motif specificities, sensitivities, and state misclassification rates over changes in β and π, where motif 0 is chosen as the motif under which the data is simulated.

4.6.8 Motif Category B

Next, we went on to investigate if the above trends of the mean square errors , obtained by varying β and π hold for a different type of motif matrix. For this we performed a set of simulations where the motif matrix in the simulation model was motif category B.
For this set of simulations, The transition and emission parameters were fixed at the values specified in 4.1. The motif width was fixed at 8.
Keeping all other parameters constant, we changed the β value over the range (1.2-2.7). We increase the β value by an increment of 0.2 till 2 . We appended the values 2.5 and 2.7 later to the list to detect any possible changes over a wider domain. Three π values (.002,.003, .006) were taken for each setting of the β parameter.
We observe that this pattern of change of β and π mentioned in the previous subsection is valid only for motif category A and motif category 0. In the B category of motif matrices, increase in β and π do not accompany a decrease in misclassification rates for the motifs. The motif specificities and sensitivities matches are obtained at a reduced level, and they stay more or less constant through out the increase in levels of β and π . But interestingly, there is a decrease in state classification percentage with π and β.
Figure **??** gives us a graphical representation of this scenario.
This can be accounted for the fact that since the conservation pattern in these category of motifs is relatively weak, the contribution to the mean of the intensity score from the motifs in the NFR region is equally distributed both on the positive and negative range . This makes differentiation between the NFR and nucleosomal states difficult,

contributing to an increase in the MSE s of the state transition parameters.

4.6.9 Mis-specifying the motif width

We wanted to see what effect the mis-specification of a motif width has on the estimation of parameter estimates. In all these simulation settings we worked with a category B motif matrix. The results were obtained for two sets of simulations.

- a)when the motif width was extended by 3 units
- b) when the motif width was extended by 5 units.

The β value was fixed to 1.4 and the motif probability π was set at .004. All other values were set at previous settings as in table 4.1.
After increasing the motif width by these amounts, we changed the value of β and π to see the effect of motif width misspecification, under these changed parameter settings. The effect is most pronounced in the estimation of β parameters and motif predictions. The motif specificities and sensitivities were reduced by .12 and .16 respectively. We observe that the MSEs of β increase by .09 and 1.27, when motif width is increased by 3 and 5 positions respectively. There is a slight but not very significant change in the transition parameters. The MSEs of λ and μ increase by .1 and .089.The state classification percentage also suffers and we see an increase of 1.13 in state classification error under a misspecification of 3 units. The motif specificities and sensitivities were reduced by .12 and .16 respectively.
However, when μ was increased from .57 to .7,keeping all other parameters fixed, we see that the mean square errors of the β estimates is reduced from that in the previous setting (when μ was set at .57) by an amount of .034 and .021 with a motif width increase of 3 units and 5 units respectively. However there is an absolute increase of 0.06 units and .07 units with the misspecification width, the mean squared errors margin only reduces when μ is increased. The increase in state specification error remains the same under these mis-specifications at this increased μ value.
On increasing π from .004 to .006, keeping all other parameters fixed we see a similar decrease (about .045 for 3 unit change and .057 for a change of 5 units) of the transition

parameters MSE's and a slight decrease in the state classification errors(.005 and .006) although there is no decrease in motif sensitivity and specificity.
Overall we see that the mis-specifying the motif width not only increases the mean squared errors in β parameters and the motif site predictions but also has an adverse effect on the state classification error. Increasing the mean of the intensity distribution in the NFR region and the motif occurrence probability slightly controls this error rate.

4.6.10 Two Step approach and its comparison with the joint model

We next tried to see how our method compares with the two step approach where we first estimate the nucleosomal positions and then try to estimate motifs based on nucleosome positions. We obtained the nucleosomal predictions by running a simple hidden markov model on the intensity data . The hidden states belonged to a nucleosomal or a Nucleosome Free Region. The emission and transition parameters for the nucleosomal prediction models were set to be the same as the previous simulations. We assigned a position to be nucleosomal if the posterior probability was found to be greater than .5. We then extracted the predicted nucleosome free regions and concatenated the NFR segments into one single data set where we ran the motif search algorithm– Bioprospector The motif width was selected to be 8. and the top motif was targetted for output. The β value was fixed at 1.2. For a fixed π we see that the two step approach fails to come close to the level of the joint modeling approaching ,as regards accurate motif detection. In table 4.6 we have denoted the two step approach by method 'B' and the joint modeling approach by 'A'. The β value was fixed at 1.2.

4.7 Identifiability

The identifiability results can be obtained by viewing the model as a hidden markov model where we have each position covered by one of three hidden states namely

1. Nucleosomal–S_0
2. NFR with no motifs–S_1
3. NFR with motif –S_2

Table 4.6: Comparison with the two step approach

Estimation method (π)	Motif Sensitivity	Motif specificity.	
A	.002	.88	.92
B	.002	.69	.87
A	.004	.91	.88
B	.004	.73	.82
A	.006	.90	.91
B	.006	.78	.76
A	.008	.94	.94
B	.008	.75	.82
A	.010	.95	.94
B	.010	.82	.84

Each of the states has its own unique emission density.
The transition probabilities between the different states are computed as the cross product of the motif occurrence probabilities, and the state transition probabilities. The following equations provide an example of the pattern in which the transition probabilities between the different states(27 equations in total) emerge.

$$T[S_0, S_0] = (1-\pi)(\lambda) \tag{4.7.1}$$

$$T[S_0, S_1] = (1-\pi)(1-\lambda) \tag{4.7.2}$$

$$T[S_0, S_2] = (\pi)(1-\lambda) \tag{4.7.3}$$

$$T[S_1, S_0] = (1-\pi)(1-\mu) \tag{4.7.4}$$

$$\tag{4.7.5}$$

and so on.
To establish identifiability conditions for the proposed models, we extend the results of Teicher,1960.

Proposition 2. . *Assume that the class of finite mixtures of the family f_ϕ of densities of Y with parameter $\phi \in \Phi$ is identifiable. Then the class of finite mixtures of n-fold product*

densities $f_{\phi}^{(n)}(y) = f_{\phi}(y_1) \dots f_{\phi}(y_n)$ *with parameter* $\phi \in \Phi^n$ *is identifiable.*

Proposition was proved by induction on n . Now, suppose there are two sets of parameters yielding the same likelihood, given by

$$\boldsymbol{\eta} = (\lambda, \mu, \nu_1, \nu_2, \sigma_1, \sigma_2, \theta_b, \theta_m, \beta, \pi) \quad \text{and} \quad \boldsymbol{\eta}' = (\lambda', \mu', \nu_1', \nu_2', \sigma_1', \sigma_2', \theta_b', \theta_m', \beta', \pi').$$

In order to prove identifiability, we need to show that $\boldsymbol{\eta} = \boldsymbol{\eta}'$.
From Proposition 1 and the observation that our model is a hidden markov model, and hence a model is a finite mixture of n-fold densities of normal likelihoods,a shown in chapter 1 we can directly invoke Proposition 2 to get the following. The finite collection of n-dimensional densities and the mixture weights attached to these densities shall be the same for the two parameter settings.
Since the finite collection of n-dimensional densities have a one to one correspondence with the emission parameters, and the equality of the collection is established from Proposition 1, we have from the equality of the emission densities in State 0

$$(\mu_0, \sigma_0), \theta_b) = (\mu_0', \sigma_0'), \boldsymbol{\rho}),$$

Similarly from the equivalence of densities in Stat1 we have

$$(\mu_1, \sigma_1)) = (\mu_1', \sigma_1')),$$

Lastly, the equivalence of emission densities yield the equation:

$$(\mu_{a1}, \Theta, \boldsymbol{\rho}), \sigma_2), \theta_m) = (\mu_{a1}, \Theta, \boldsymbol{\rho})', \sigma_2'), \theta_m'),$$

Now from the $\mu_2 = \mu_2'$ and $\Theta = \Theta'$. From this step, the equality of all emission parameters is proved.
Now we make use of the fact that the mixture weights are equal, and in order to translate this condition to an equality condition on the transition rates, we exploit the relation between the transition probabilities and the transition parameters. The equivalence of the mixture weights imply that law of the processes are equal and hence the transition probabilities between different given by the set of 27 equations are equal. Dividing the

equation $(1-\pi)(\lambda) = (1-\pi')(\lambda')$ and $(1-\pi)(1-\lambda) = (1-\pi')(1-\lambda')$ we get $\lambda = \lambda'$ Now from $(\pi)(1-\lambda) = (\pi')(1-\lambda')$ we get $\pi = \pi'$. Plugging this equality into $(1-\pi)(\mu) = (1-\pi')(\mu')$ we get $\mu = \mu'$. Hence the equality of transition parameters is proved.

4.8 Structure of the hidden Markov model

In order to prove identifiability results, we formulated the model as Hidden Markov Model with three hidden states. It is possible to extend the structure to a model with w+2 states, where the additional w states come from the w motif positions. There we can assume each position to add an an extra intercept in addition to the effect of the log likelihood score covariate. This would make the hidden Markov model with w+2 states identifiable. This formulation is essential to the deriving of the asymptotic properties of the Bayes estimators and the testing of hypothesis under contiguity. With the help of asymptotic properties and contiguity results proved in Chapter 5, we can test a range of hypothesis relating to the presence of a link connecting the motif strength and nucleosomal intensity signal.

4.9 Results

4.9.1 Applications on the yeast genome data set

The model was applied on the FAIRE data set. FAIRE is an acronym for Formaldehyde assisted isolation of regulatory elements (FAIRE) . This experimental technology is based on the principle that regions of the genome that are cross-linked with the histone protein are less susceptible to binding by transcription factors, and thus get retained at the interphase of organic and aqueous phase, while the nucleosome-free (potentially regulatory) regions get enriched. The FAIRE procedure is initiated by first fixing whole yeast cells in a growth medium by formaldehyde, harvesting them by centrifugation, and finally sonicating the extracts, labeling the purified DNA and hybridizing them to microarrays. The microarray in used 50-mer oligonucleotide probes that overlap every 20-mers to tile almost all of chromosome III and 1 kb of 223 additional regulatory

regions. Four microarrays (three biological experiments and one technical replicate) were performed. Data of the median of ratios were extracted directly from arrays using Genpix. Acquired images were visually inspected and low quality spots were removed. The data were log transformed, block normalized, and the technical replicates were averaged and treated as one biological replicate, followed by averaging all three biological replicates. The enrichment throughout the genome represents the positioning of the regulatory regions which are free of nucleosomes. The nucleosome intensity data and sequence features were simultaneously used for nucleosome and motif finding within the sequences. Due to the gaps between the probe positions in the FAIRE data, we had to incorporate continuous time equations for the transition intensities. This involved a set of Metropolis Hastings steps to sample λ and μ in addition to the sampling steps described in the estimation procedures in subsection 4.5.
The initial values for log(λ), $log(\mu)$, ,μ_0, μ_1, β were generated from N(0,1). π was assumed to be .05. The motif width was set at 8 and the initial motif matrix was fixed to Motif0, taken from the Jaspar database . $\boldsymbol{\rho}$ was initialized at (.25,.25,.25,.25). The parameter estimates were calculated over a sample of 2000 iterations after burn-in. The estimates are given in 4.7

Table 4.7: Parameter values

log(λ)	log(μ)	(μ_1)	(μ_2)	(σ_1)	σ_2	(β)
-4.8	-3.19	-.8	.56	.47	.66	.64

The motif matrix Θ parameter estimates are given in 4.8:

Table 4.8: Estimated Motif Matrix

	1	2	3	4	5	6	7	8
A	0.11	0.05	0.09	0.07	0.35	0.55	0.04	0..02
C	0.64	0.66	0.1	0.58	0.11	0.17	0.8	0.15
G	0.18	0.15	0.73	0.29	0.12	0.08	0.06	0.78
T	0.07	0.14	0.08	0..06	0.42	0.20	0.1	0.04

The background probability vector $[\theta_b]$ estimates are given in table 4.9:

Table 4.9: Estimated Background Probability vector

A	C	G	T
.34	.27	.21	.29

Both forward algorithms I and III were used in computing the log likelihood . Forward algorithm I incurred substantially lesser time than forward a algorithm III. Convergence was detected by monitoring the auto-correlation function.

4.9.2 Comparison with the denovo motif discovery and the two-step method

We compared our method with a motif search algorithm that did not take into account the nucleosome positional information. This de novo motif discovery method was implemented on the Hogan data set. The applied motif discovery algorithm is actually the classical motif discovery algorithm based on the Position Weight Matrix Model. For this implementation, the motif width was fixed to 8. A forward algorithm was used to compute the likelihood and motifs were sampled through backward sampling technique. The forward algorithm for computing the likelihood is given below:
T[i,1] denotes the likelihood till position i, with the position i being the end of a motif position. T[i,2] denotes the likelihood till position i, with the position i being in background

$$T[i,1] = (T[i-w,1] + T[i-w,2])\Theta[\boldsymbol{X}^{i-w+1:i}] \tag{4.9.1}$$

$$T[i,2] = (T[i-1,1] + T[i-1,2])\rho_{X_i} \tag{4.9.2}$$

$$\tag{4.9.3}$$

After getting the motif predictions from this algorithm, we implemented the two step approach on this real data set. As in the simulation we first got the nucleosomal

predictions by running a simple hidden markov model on the intensity data . We assigned a position to be NFR if the posterior probability of being NFR was found to be greater than .5. We then extracted the predicted nucleosome free regions and concatenated the NFR segments into one single data set where we ran the previous de-novo motif search algorithm. Table **??** shows the comparison between these approaches.

4.10 Conclusions

A joint motif nucleosome model, presented in this chapter, takes into account the nucleosome mapping information, and thereby, may have greater predictive power The increase in predictive power comes from efficiently incorporating nucleosomal mapping information. Although the main purpose of this model is to reduce the false positives in motif search, it simultaneously yields nucleosomal predictions. Also, for the first time, there has been an attempt to measure quantitatively the link between gene signal data and motifs, in spite of past biological evidence. Simulation studies have shown that the proposed nucleosome-adapted motif discovery algorithm beats the de-novo one both with respect to sensitivity and specificity of motif finding. It is also superior to the two-step approach of first getting nucleosomal predictions and then using it as a prior. This is because the prediction biases get propagated when we use predictions from a different model. The unified model takes this problem into account. The forward recursive algorithm and backward sampling procedure is a useful technique of Data Augmentation and likelihood computation in complex models which involve hidden variables. Our set up has two layers of hidden variables namely the state configuration and the motif positions. Also due to the link between the intensity data and the motif strength, the relationship between these components among themselves and with the data is complex and unique. Hence we needed to modify the existent recursive algorithms for motif discovery and come up with two new algorithms that are more general in nature. Gibbs sampling was implemented to sample from the posterior. Special attention was given so that block sampling replaces the full conditional steps wherever it was possible.This was done to speed up MCMC convergence to a considerable extent. There are scopes to extend our current model to accommodate length constrained states. The recursive

algorithms defined earlier, would definitely work in the new scenario, but at the cost of extra computational complexities. The link function can be suitably modified to accommodate a more general weighing scheme for the motif positions, whereby their positional preferences with respect to gene expression could be judged. The two layered hierarchical hidden structure of our model provides an interesting set up, that could find possible applications in other fields.

Chapter 5

Asymptotic results for continuous time hidden Markov models

5.1 Introduction

In recently introduced procedures like FAIRE (Formaldehyde Assisted Isolation of Regulatory elements), missing information due to experimental constraints in possibly non-contiguous probes may make hidden Markov models (HMM) intractable and lead to model misclassifications.The use of a continuous time Markov chain eliminates some of the impasses. In a setup of discrete state space, the transition probabilities are then derived from a continuous time Markov chain. This motives us to formulate a two-state exponential model of the hidden states. Along with the preliminary notion, this continuous time Markov chain model is introduced in Section 5.2. This chapter is primarily devoted to the development of general likelihood function based asymptotics providing the desired results of $\sqrt{n}$ consistency of the estimators as well as related asymptotic normality results. In this formulation a key tool is the incorporation of 'contiguity of probability measures' in a HMM setup. All these basic motivations are presented in Section 5.1.1. The derivations of the main results are considered in subsequent sections. The relevance of these asymptotic results in the two important problems considered in Chapters 3 and 4 is discussed in Section 5.6.2.

5.1.1 Preliminary Notion

Under the continuous time index framework, the tranistion between the hidden states occur in a continuous continuum .We get to observe and infer only in a discrete subset of that domain. We get the transition probability from one state to another from the properties of the continuous time markov chain. We assume that the chain leaves the current state at a rate proportional to the current duration. This gives us an two-state exponential model of the hidden states with transition rates λ and μ.
The transition probability from one state to itself in time interval t,is given as

$$P(0,0,t) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}$$

The basic assumptions and notations required for this section :
Let X_i denote the unobserved states.
X_i=1 or 0, i=1,2...N
The probabilities of transition between these states follow Equation 3.2.3
We assume that the transition rates are bounded from below and above . That is

$$0 < \sigma^- < \lambda, \mu < \sigma^+ < \infty$$

Let Y_i denote the observed data. Density of Y_i is f_{X_i} , i= 1,2,...N where f is a bounded density function indexed by the hidden states.
Let L_k denote the data likelihood up to position k.
Let a[k,0] denote $L(Y_1, Y_2, .Y_{k-1}, X_k = 0)$ Let a[k,1] denote $L(Y_1, Y_2, .Y_{k-1}, X_k = 1)$ where $L(Y_1, Y_2, .Y_{k-1}, X_k = i)$ is the likelihood up to position k with the k^{th} state=i.
We can see the motivation for these asymptotic results in view of their implications in the inferential procedures in Chapters 1 and 2. The goal of the set of all asymptotic results is to establish consistency of the posterior mode and the $\sqrt(n)$ convergence of the posterior.This discussion is important for the HMM model defined by the transition structure above and for any non-homogeneous HMM run on long chains, with bounded transition probabilities. Thus it would have huge and widespread applications in genomic data. In Chapter Three, we have applied the continuous time HMM on such a

dataset,where the length was 15000. Establishing the rate of this convergence to be $\sqrt{(n)}$ we would have a fair idea of the sample size needed to have such kind of asymptotic behavior. However the benefits of asymptotic results provided by increasing data size do not hold for arbitrary distribution of gap lengths. This we shall discuss in the section on posterior convergence and its implications. We can place a high degree of credibility of the results obtained through MCMC algorithm used for Bayesian inference due to the consistency results of the Bayesian parameters estimates and the log likelihood. It also validates the pattern of estimated log likelihood and BIC criterion which were observed in different simulation experiments performed on the model in Chapter 3.

5.2 Asymptotic inference for parameter estimates, score and information in continuous time HMMs

In this section we set about to prove the almost sure consistency of the maximum likelihood estimates for continuous time hidden index models. For homogeneous hidden Markov models, a proof of consistency by Leroux(1992) employs the technique of conditioning on the infinite past.We, on the other hand, utilized the martingale approach. It will be a recurrent theme in later subsections and play a major role in the proof of asymptotic normality.

5.2.1 The main results

- Result I. Consistency: Convergence of MLE s and the scaled log-likelihood.
- Result II: The Asymptotic Normality of the score function.
- Result III. Contiguity- We prove the asymptotic equivalence of observed information and Fisher's matrix. This, combined with Result II leads directly to the asymptotic distribution of parameter estimates.
- Result IV. Posterior Convergence:discusses the convergence of posterior under any continuous prior distribution.

5.3 Consistency

Here, we prove the consistency of HMM parameter estimates. We employ a martingale approach to centered ratios of log likelihood. In the course of this proof, we shall exploit the uniform forgetting property of the hidden markov models. The latter will follow from the results on the ignorability of the initial condition and equating moments of shifted markov processes. In a latter subsection, we prove identifiability. This will serve as a crucial condition for the final proof. Let

$$D_k = log(L_k) - log(L_{k-1})$$

denote the log likelihood increment. For notational convenience we will write $f_i(Y_k)$ as simply $f_{i,k}$ We can write this as
$log\frac{a[k,0]f_{0,k}+a[k,1]f_{1,k}}{a[k-1,0]f_{0,k-1}+a[k-1,1]f_{1,k-1}}$
This can be further written as
$log[[\frac{a[k,0]}{a[k-1,0]f_{0,k-1}+a[k-1,1]f_{1,k-1}}]f_{0,k} + [\frac{a[k,1]}{a[k-1,0]f_{0,k-1}+a[k-1,1]f_{1,k-1}}]f_{1,k}]$
Which is another form of the expression
$log(P(X_k = 0|Y_1....Y_{K-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k})$
We start by showing that $E(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_K = 1|Y_1....Y_{k-1})f_{1,k})$ is a Cauchy sequence in k.
For $m > k$ consider the expression

$$E((P(X_m = 0|Y_1....Y_{k-1})f_{0,m} + P(X_m = 1|Y_1....Y_{k-1})f_{1,m})$$

$$-(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k}))$$

We employ the trick of writing the above expression as

$$E((P(X_m = 0|Y_1....Y_{k-1})f_{0,m} + P(X_m = 1|Y_1....Y_{k-1})f_{1,m})$$

$$-(P(X_m = 0|Y_1....Y_{k-1}, X_{M-k} = 0)f_{0,m} + P(X_m = 1|Y_1....Y_{k-1}, X_{m-k} = 1)f_{1,m})$$

$$+(P(X_m = 0|Y_1....Y_{k-1}, X_{m-k} = 0)f_{0,m} + P(X_m = 1|Y_1....Y_{k-1}, X_{m-k} = 1)f_{1,m})$$

$$-(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k}))$$

Which can be written as the sum of two components:

$$E((P(X_m = 0|Y_1....Y_{k-1})f_{0,m} + P(X_m = 1|Y_1....Y_{k-1})f_{1,m})-$$

$$(P(X_m = 0|Y_1....Y_{k-1}, X_{M-k} = 0)f_{0,m} + P(X_m = 1|Y_1....Y_{k-1}, X_{m-k} = 1)f_{1,m}))$$

and

$$E(P(X_m = 0|Y_1....Y_{k-1}, X_{m-k} = 0)f_{0,m} + P(X_m = 1|Y_1....Y_{k-1}, X_{m-k} = 1)f_{1,m})$$

$$-(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k}))$$

Note that the second component is 0 since the two random variables whose first moments are taken have the same distribution. One is a shifted version of the other. Due to the Markov structure, the distribution of the processes depend only on the initial condition. For the first component we use the uniform forgetting property. This gives us the bound: $||(P(X_m = 0|Y_1....Y_{k-1}) - P(X_m = 0|Y_1....Y_{k-1}, X_{m-k} = 0)||_{TV} < (1 - \frac{\sigma^-}{\sigma^+})^k = \tau^k$ where τ is less than 1.
This shows that $E(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_K = 1|Y_1....Y_{k-1})f_{1,k}$ is a Cauchy sequence in k.
Now $|log(x) - log(y)| < \frac{|x-y|}{min(x,y)}$
Using the above and the fact that $(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k})$ is bounded it is easy to show that
$log[E(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k})]$ is a Cauchy sequence as well.
Jensen's inequality can now be used to show that:
$E[log(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k})]$ is a Cauchy sequence.
We now write the entire log likelihood

$$L_n = L_0 + \sum [L_k - L_{k-1}]$$

Let $E(L_k - L_{k-1})$ be denoted as U_k.
So $U_k = E(D_k)$

Now $(1/n)L_n = (1/n)[L_0 + \sum_{i=1}^{n}[Z_k + U_k]]$
We have shown that U_k is Cauchy and hence convergent.
The sequence Z_k is a mean zero martingale. Recall that
$Z_k = log(P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k})$
$P(X_k = 0|Y_1....Y_{k-1})f_{0,k} + P(X_k = 1|Y_1....Y_{k-1})f_{1,k} < 1$
and since the parameters lie in a bounded set. So any continuous function in this set is bounded . The boundedness of Z_k implies that Kolmogrov Strong Law of Large numbers for martingales holds and $\sum(Z_k/n)$ is convergent almost surely. From this, we get that L_n/n is convergent.

5.3.1 Identifiability

We go on to establish the identifiability conditions for the base and emission models by extending the results of Teicher (1967) Teicher (1967) . For this, we use the following main definition and result.
Definition: Let $f_\phi(y)$ be a parametric family of densities of Y with respect to a common dominating measure μ and parameter ϕ in some set Φ If π is a probability measure on Φ then the density

$$f_\pi(y) = \int_\Phi f_\phi(y)\pi d(\phi)$$

is called a mixture density .
We say that the class of (all) mixtures of (f_ϕ) is identifiable if

$$f_\pi = f_{\pi'}\mu \text{ a.e iff } \pi = \pi'$$

Furthermore we say that the class of a finite mixtures of f_ϕ is identifiable if for all measures π and π' with finite support $f_\pi = f_{\pi'}\mu$ a.e iff $\pi = \pi'$.
Result 1: The class of joint finite mixtures of the normal family is identifiable (Telcher,1960).
Result 2: Assume that the class of finite mixtures of the family f_ϕ of densities of Y with parameter $\phi \in \Phi$ is identifiable . Then the class of finite mixtures of n-fold product densities $f_\phi^{(n)}(y) = f_\phi(y_1).....f_\phi(y_n)$ with parameter $\phi \in \Phi^n$ is identifiable. The result was proved by induction on n. [Teicher (1967)].

We shall now apply the above results to prove the identifiability of the base model. (Note that any hidden markov model is a finite mixture of n-fold product densities, where the the weights of the mixture are functions of the transition probabilities). Now, suppose there are two sets of parameters

$$\Theta = (\lambda, \mu, \mu_1, \mu_2, \sigma_1, \sigma_2)$$

and

$$\Theta' = (\lambda', \mu', \mu_1', \mu_2', \sigma_1', \sigma_2')$$

yielding the same likelihood. We need to show that

$$\Theta = (\lambda, \mu, \mu_1, \mu_2, \sigma_1, \sigma_2) \tag{5.3.1}$$

$$= (\lambda', \mu', \mu_1', \mu_2', \sigma_1', \sigma_2') \tag{5.3.2}$$

Proof: From result 1 and the fact that and our model is a finite mixture of n-fold densities of normal likelihoods we can directly invoke Result 2 to get the following. The finite collection of n dimensional densities and the mixture weights attached to these densities shall be the same for the two parameter settings.
Now since the n-fold densities arise from the normal family, they have a one to one correspondence with the collection of n dimensional parameter sets formed by the n-fold convolution of the sets (μ_1, σ_1) and (μ_2, σ_2) . Since the finite collection of n dimensional densities have a one to one correspondence with the emission parameters, and the equality of the collection is established from result 1) we have

$$(\mu_1, \sigma_1), (\mu_2, \sigma_2) = (\mu_1'\sigma_1'), (\mu_2', \sigma_2')$$

and hence the individual parameters are equal up to a permutation or reordering.
Now we make use of the fact that the mixture weights are equal. In order to translate this condition to an equality condition on the transition rates, we exploit the relation between the transition rates and the transition probability functions. From the equality of the mixture weight probabilities, it can be inferred that the laws of these two processes are the same, i.e $P(X_i = a|X_{i-1} = b)$ for any sequence position i, and for any states a, b,

up to a permutation of the indices a and b. Let the permuted indices be denoted as a', b'. Thus, if we denote the transition probability under Θ' as $P'(X_i = a|X_{i-1} = b)$ we have $P = P'$

Adding up the equations for $P_{00}[t]$ and $P_{11}[t]$ for both Θ and $P'_{00}[t]$ and $P'_{11}[t]$ for both Θ' we get

$$e^{(-\lambda-\mu)t} = e^{(-\lambda'-\mu')t} \forall \text{ t} \tag{5.3.3}$$

$$\Rightarrow \lambda - \lambda' = \mu - \mu' \tag{5.3.4}$$

Assume $a' = a$ and $b' = b$.

Then, from the equation of $P_{00}[t]$ we have:

$$\frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} \exp -(\lambda+\mu)t = \frac{\mu'}{\lambda'+\mu'} + \frac{\lambda'}{\lambda'+\mu'} \exp -(\lambda'+\mu')t$$

Subtracting Right hand side from left hand side we get

$$[\frac{\mu}{\lambda+\mu} - \frac{\mu'}{\lambda'+\mu'}] + [\frac{\lambda}{\lambda+\mu} - \frac{\lambda'}{\lambda'+\mu'}] \exp -(\lambda'+\mu')t] = 0$$

Noting that $\lambda - \lambda' = \mu - \mu'$ and equivalently $\lambda + \mu = \lambda' + \mu'$ we get

$$(\lambda - \lambda')(1 - exp((-\lambda - \mu)t)) = 0 \forall \text{ t} \tag{5.3.5}$$

$$\Rightarrow \lambda = \lambda' = 0 \tag{5.3.6}$$

Again using

$$\lambda - \lambda' = \mu - \mu'$$

we get

$$\mu = \mu'$$

If we assume $a' = b$ and $b' = a$ we can repeat the above steps to see that $\lambda = \mu'$ and $\mu = \lambda'$ Thus the identifiability of the model is completely proved.

5.3.2 Identifiabilty and Asymptotics of log likelihood gives us consistency

We utilize the identifiability properties of HMM to show that $l(\theta)$ has a global unique maximum. Let the maximum be denoted by (θ_*). If $\widehat{\theta_n}$ denotes the MLE then

$$0 \leq l(\theta_*) - l(\widehat{\theta_n}) \tag{5.3.7}$$

$$\leq l(\theta_*) - n^{-1}l_n(\theta_*) + n^{-1}l_n(\theta_*) - n^{-1}l_n(\widehat{\theta_n}) + n^{-1}l_n(\widehat{\theta_n}) - l(\widehat{\theta_n}) \tag{5.3.8}$$

$$\leq 2sup_\theta|(l(\theta) - n^{-1}l_n(\widehat{\theta_n})| \tag{5.3.9}$$

This shows that for any compact subset

$$l(\theta_n) \to l(\theta_*)$$

as $n \to \infty$. Since l is continuous, this shows that the MLE converges to θ_* almost surely.

5.4 Asymptotic Normality of the score function

The main part of the proof of asymptotic normality lies in the application of the martingale central limit theorem. The proof of asymptotic normality for simple HMM is also dependent on the central limit theorem for martingales, but on the version of CLT where stationary processes are used. In order to facilitate CLT application, a large machinery was built by Ryden et al,which approximated the score function to a stationary process. We departed from this approach of proof for continuous index hidden markov models, as the algebraic calculations would be too cumbersome to handle.Instead,we used the non-stationary version of the martingale CLT that required the Lindberg-Feller condition and some properties of conditional squared expectations. For the latter, we again invoked martingale results. In this way, we could handle a more general scenario,involving any bounded non-homogeneous transition probability functions,and not just the specific structure we have from the transition equations. Here our function of interest is the derivative of the log likelihood, $l'_{x_0,n}(\theta)$. The double indices x_0 and n denote the initial state and the position up to which the likelihood is computed,

respectively. In keeping with the conventions of Douc, whose results we are going to use, we make a slight notational change. We denote the observed and unobserved processes (denoted hitherto by Y and X) as y and x respectively.

$$l'_{x_0,n}(\theta) = l'_{x_0,0}(\theta) + \sum_{k=1}^{n}[l'_{x_0,k}(\theta) - l'_{x_0,k-1}(\theta)] = \sum_{k=0}^{n} h'(k,0,x_0,\theta)$$

where

$$h'(k,0,x_0,\theta) = E_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0=x] - E_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0=x]$$

and

$$\phi_\theta(x,x',y) = \nabla_\theta log[q_\theta(x,x')g_\theta(x',y)]$$

$$h'(\infty,m,x,\theta) = E_\theta[\sum_{i=1}^{\infty}\phi_\theta(x_{i-1},x_i,y_i)|y_{1:k},x_0=x] - E_\theta[\sum_{i=1}^{\infty}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0=x]$$

We use the notation G_k denote the derivative of the log likelihood increment at stage k. Writing the score function as a sum of increments has important consequences.
We know that the derivative of log likelihood increments form a martingale.
The **Central Limit Theorem** for Martingales states that: If $[X_k]$ is a martingale increment sequence $T_n = \sum_{k=1}^{n} X_k$, $s_n^2 = \sum_{k=1}^{n} E(X_k^2)$, $v_k^2 = E(X_k^2|X_1...X_{k-1})$ and $w_n^2 = \sum_{k=1}^{n} v_k^2$. The sufficient conditions are:

1. $\frac{w_n^2}{s_n^2} \longrightarrow 1$ as $n \longrightarrow \infty$
2. For every $\epsilon > 0$ $s_n^{-2}\sum_{k=1}^{n} E[X_k^2 I_{[|X_k|>\epsilon s_n]}] \longrightarrow 0$ as $n \longrightarrow \infty$

We now try to get a L^2 bound on the difference between $h'(k,m,x,\theta)$ and $h'(\infty,0,x,\theta)$. For this, we take corresponding terms in $h'(k,m,x,\theta)$ and $h'(k,\infty,\theta)$ and match them one by one. Two types of terms come up

1. $[\phi_\theta(x_{i-1},x_i,y_i)|y_{1:k},x_0=x]$ and $[\phi_\theta(x_{i-1},x_i,y_i)|y_{-\infty:k}]$.
2. $[\phi_\theta(x_{i-1},x_i,y_i)|y_{1:k},x_0=x]$ and $[\phi_\theta(x_{i-1},x_i,y_i)|y_{1:k-1},x_0=x]$

We get a bound on the first group of terms by the Cauchy sequence property of $h'(k, m, x, \theta)$ proved earlier.
For the second type of difference terms, we use the 'uniform forgetting' and 'conditional mixing' properties. For this we invoke the backward decomposition theorem .
Conditional on $y_{0:n}$ the joint distribution of x_j downwards to x_m is a time reversed non homogeneous markov chain with the transition kernel $B[y_{m+1} : j](x;)$. This can be defined for each $j = -m + 1, ...k - 1$.
where these backward kernels satisfy the inequality

$$\frac{\sigma^-}{\sigma^+} v[y_{m+1:j}] <= B[y_{m+1:j}](x;) <= \frac{\sigma^+}{\sigma^-} v[y_{m+1:j}]$$

where

$$v[y_{m+1:j}] = \frac{\int .... \int \prod_{u=m+1}^{j} q_\theta(x_{u-1}, x_u) g_\theta(x_u, y_u) \lambda(dx_u) f(x_j)}{\int .... \int \prod_{u=m+1}^{j} q_\theta(x_{u-1}, x_u) g_\theta(x_u, y_u) \lambda(dx_u)}$$

The Derbushin coefficient of any kernel K is defined as

$$DC = \frac{1}{2} sup_{x,x'} ||K((x,) - K(x',)||_{TV}$$

The Transition kernel is bounded because the transition function is a continuous function on a compact bounded set, as we have a lower and upper bound on the transition rates. Using the result that for a kernel satisfying a condition as above, the Derbushin coefficient is bounded, we have the DC for the time reversed markov chain to be less than

$$1 - \frac{\sigma^-}{\sigma^+} = \rho$$

Now each of the terms of the type $P_\theta(x_i \in .|y_{m-1:k}, x_m)$ and $P_\theta(x_i \in .|y_{m-1:k-1}, x_m)$ can be viewed as time reversed distributions from k-1 to i with backward kernel being $< \rho$. Hence the total variation norm of the difference is ρ^{k-1-i}.
From this above property, we achieve bounds on the difference term . Next we combine these bounds and use Minkowski's inequality to obtain:-

$$||h'(k, m, x, \theta) - h'(k, \infty, \theta)||^2)^{1/2} \leq C \frac{\rho^{k+1/2-1}}{1 - \rho}$$

$\Rightarrow G_k^2$ is asymptotically bounded. $P(G_k^2 > k) < \rho^k$.
From the boundedness of G_k^2 the condition 1 of the Martingale CLT, the Lindberg-Feller condition is immediate While the sum in the numerator is convergent ($\sum \rho^k = \frac{1}{1-\rho}$) the sum in the denominator diverges to ∞. For the second condition we need to invoke the following theorem.
Theorem : Let X_n be a sequence of random variables and F_n be an increasing sequence of sigma fields with respect to F_n If X be a random variable such that $E|X| < \infty$ and there exists a constant c such that $P(|X_n| > x) <= cP(|X| > x)$for each $x > 0$ then
$n^{-1} \sum [E(X_i) - E(X_i|F_{i-1})] \Rightarrow 0$
We apply the result on random variables $[G_k^2]$.
From there we get $n^{-1}[\sum E[G_k^2|F_{k-1}] - \sigma_k^2] \longrightarrow 0$
We know that $n^{-1} \sum \sigma_k^2$ is bounded as σ_k^2 is bounded.
Dividing this by $n^{-1} \sum \sigma_k^2$ we get
$\frac{\sum_{i=1}^k E(G_k^2)/B_{k-1}}{\sum_{i=1}^k E(G_k^2)} - 1$ converges to 0.
$\Rightarrow$
$\frac{\sum_{i=1}^k E(G_k^2)/B_{k-1}}{\sum_{i=1}^k E(G_k^2)}$ converges to 1.
Hence the second condition of the Martingale Central Limit Theorem is verified. So we get

$$n^{-1/2} \sum_0^n h'(k, \infty, \theta_*) \rightarrow N(0, J(\theta_*))$$

weakly where

$$J = E_{\theta_*}[h'(1, \infty, \theta_*) h'^t(1, \infty, \theta_*)]$$

is the Fisher information matrix. This shows :
$n^{-1/2} \nabla_\theta l_n(\theta_*) \rightarrow N(0, J(\theta_*)$ weakly.

5.5 Contiguity

The concept of contiguity plays an essential role in developing the distribution of test statistics under a sequence of local alternatives. For this, we need to consider a sequence

of hypotheses, $P_n := P_n^{\theta_0}$ and $Q_n := Q_n^{\theta_n}$ where

$$\theta_n = \theta_0 + \delta/\sqrt{(}n)$$

P_n is said to be contiguous to Q_n if for any sequence of events A_n

$$[P_n(A_n) \longrightarrow 0] \Rightarrow [Q_n(A_n) \longrightarrow 0]$$

The relation of contiguity to asymptotic properties of test statistics under alternate hypotheses is established by the celebrated Le Cam's third lemma. In most cases,as in the continuous time HMM, the derivations of the distributions of these statistics are very difficult. The following lemma helps us surmount these obstacles:
Le Cam's Third Lemma: If Q_n is contiguous to P_n and the pair of statistics (S_n, L_n) is asymptotically normal with mean $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$ with $\mu_2 = -\frac{1}{2}\sigma_2^2$ and L_n being the likelihood ratio, then under Q_n
S_n is asymptotically normal $(\mu_1 + \sigma_{12}, \sigma_1^2)$
Thus we can translate the asymptotic normality under null distributions to that under alternatives.
Le Cam's first lemma gives us a very useful and widely used characterization of contiguity for a sequence of hypotheses.
Le Cam's first lemma: If under P_n the log-likelihood ratio L_n is asymptotically log-normal $(-\frac{1}{2}(\sigma^2), \sigma^2)$ then Q_n is contiguous to P_n
This lemma is used in establishing contiguity for a class of iid models by essentially expanding the log likelihood around the null by a Taylor's expansion. We extend that approach to the continuous time HMM model. For this,we approximate the behavior of log-likelihood as a ergodic sequence of stationary increments. Specifically we invoke the following propositions of Douc et al:

- **Proposition 1**: For all θ, there exists a stationary ergodic sequence $h_{k,\infty}$ such that $|l_n(\theta) - \sum_{k=0}^{n} h_{k,\infty}(\theta)| \to 0$ as $n \to \infty$
- **Proposition 2**: $h_{k,\infty}$ is obtained as a uniform limit of the continuous and differentiable sequnece $h_{k,m,x}(\theta)$ and hence is continuous and differentiable. We call

the gradient as $h'_{k,\infty}(\theta)$.

These propositions helps us mimic the Taylor's expansion argument as in the iid model. However our proof is not identical. We employ the trick of approximating the log likelihoods to a stationary sequence-a property that holds for our models. We then combine theorems of Douc et al that guarantee certain properties of the stationary processes and observed information. We then use the latter in our main proof to establish contiguity properties.

We write

$l_n(\theta_n) - l_n(\theta_0) = [l_n(\theta_n) - \sum_{k=0}^{n} h_{k,\infty}(\theta_n)] +$

$[-l_n(\theta_0) + \sum_{k=0}^{n} h_{k,\infty}(\theta_0)] + [\sum_{k=0}^{n} h_{k,\infty}(\theta_n) - \sum_{k=0}^{n} h_{k,\infty}(\theta_0)]$

The first two term tend to 0 by Proposition 1. For the third term we can write

$\sum_{k=0}^{n} h_{k,\infty}(\theta_n) - \sum_{k=0}^{n} h_{k,\infty}(\theta_0)$

$= \sum_{k=0}^{n} [h'_{k,\infty}(\theta_0) + 1/2(\theta_0 - \theta_n)^2 [h_{k,\infty}(\theta_0)'' + r(k, \theta_0)$

Now recall that $\theta_n = \theta_0 + \delta/\sqrt{(}n)$ and note that $r(k, \theta_0) \longrightarrow 0$ as $\theta_n \longrightarrow \theta_0$

The ergodic and stationarity property of $h_{k,\infty}(\theta)$ was proved by Leroux(1992). The ergodicity and stationarity of $h_{k,\infty}(\theta)'$ and $h_{k,\infty}(\theta)''$ is a consequence of theorems 4 and 5 of Douc et al.

So from Birkoff's ergodic theorem we have $\sum_{k=0}^{n} r(k, \theta_0)/n \to 0$ From section 5.4 on normality we know that the score function i.e $l_n(\theta_0)'/\sqrt{n} = \sum_{k=0}^{n} [h'_{k,\infty}(\theta_0)/\sqrt{n}]$ is asymptotically normal. Thus we have

$$\Delta log(L_n) = \delta N(0, J_0) - \frac{1}{2}\delta^2 J_1$$

For contiguity we need to show $J_0 == J_1$ asymptotically.

where $J_0 = E_{\theta_0}[h'(1, \infty, \theta_0) h'^t(1, \infty, \theta_0)]$ and $J_1 = lim_{n\to\infty} n^{-1} \nabla^2_\theta l_n(\theta_0)$ which is the limit of the observed information score .

The concluding part of this section establishes the asymptotic equivalence of the observed score information J_1 and the asymptotic covariance matrix J_0 (Note that in iid exponential model, they are equal for every n)

Recall that we had defined

$$h'(k,0,x_0,\theta) = E_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0=x] - E_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0=x]$$

where

$$\phi_\theta(x,x',y) = \nabla_\theta log[q_\theta(x,x')g_\theta(x',y)]$$

We drop the subscripts x_0 and θ for notational convenience. Note that

$$\nabla^2_\theta l_n(\theta_0) == E_\theta[\sum_{i=l}^{n}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0] + var_\theta[\sum_{i=1}^{n}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0]$$

As in the expression of the score function we can break up the above as
$E_\theta[\sum_{i=l}^{n}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0=x]$
$=\sum_{k=1}^{n}(E_\theta[\sum_{i-1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0]$
$-E_\theta[\sum_{i-1}^{k-1}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0]$
and
$var_\theta[\sum_{i=1}^{n}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0=x]$
$=\sum_{k=1}^{n}(var_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k},x_0]$
$-(var_\theta[\sum_{i=1}^{k-1}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0]$
Define

$$\tau_{1,k}(\theta) = var_\theta[\sum_{i=1}^{k}\phi_\theta(x_{i-1},x_i,y_i)|y_{0:k-1},x_0]$$

From Proposition 5 of Douc et al, under the assumptions $E_\theta[sup_x sup_\theta\phi(\theta,x_1,Y_1)] < \infty$ and $E_\theta[sup_x sup_\theta\phi(\theta,x_1,Y_1)]^2 < \infty$; $h'_{1,k}(\theta)$ and $\tau_{1,k}(\theta)$ both have limits as k $\longrightarrow \infty, P_\theta$ - a.s. The assumptions are trivially true from the boundedness of the transition function and the emission densities. Let $h_{k,\infty}(\theta)$ and $\tau_{k,\infty}(\theta)$ denote these limits. It follows from the definitions above that $[h_{k,\infty}]_{k=1}^{\infty}$ and $[\tau_{k,\infty}]_{k=1}^{\infty}]$ are $P_{\theta*}$-stationary and ergodic. Also, the limit of the observed Fisher information will be -$E_{\theta_0}[h_{0,\infty}(\theta_0) + \tau_{0,\infty}(\theta_0)]$. This is
$= E_{\theta_0}[h'(1,\infty,\theta_0)h'^t(1,\infty,\theta_0)]$ Thus we can conclude that
$\Delta log(L_n) = \delta N(0,J_0) - \frac{1}{2}\delta^2 J_0$
establishing the conditions for LeCam's first lemma. This proves that continuous time

Hidden Markov models admit a contiguous structure. The equivalence of the asymptotic information and Fisher's matrix now yields the asymptotic normality of the MLE.
To see this we recall the result of asymptotic normality of the score function:
$n^{-1/2}\nabla_\theta l_n(\theta_*) \to N(0, J(\theta_*)$ weakly.
Now we can apply a Taylor's expansion around θ_* to get

$$0 = \nabla_\theta l_n(\hat{\theta_n}) = \tag{5.5.1}$$

$$\nabla_\theta l_n(\hat{\theta_*}) + l_n[\theta_* + t(\hat{\theta_n} - \theta_*)](\hat{\theta_n} - \theta_*) \tag{5.5.2}$$

$$\Rightarrow n^{1/2}(\hat{\theta_n} - \theta_*) = -n^{-1}l_n[\theta_* + t(\hat{\theta_n} - \theta_*)]^{-1}n^{-1/2}\nabla_\theta l_n(\theta_*) \tag{5.5.3}$$

By strong consistency we have $\hat{\theta_n} \to \theta_*$ almost surely, so the first factor converges to $J(\theta_*)^{-1}$ a.s. The second factor by virtue of the equivalence converges to $N(0, J(\theta_*))$ .Hence the following result

$$n^{1/2}(\hat{\theta_n} - \theta_*) \to N(0, J^{-1}(\theta_*))$$

which is the standard asymptotic result of the MLE.

5.6 Posterior Convergence

In Bayesian analysis, one starts with a prior knowledge expressed as a distribution on the parameter space and updates the knowledge according to the posterior distribution given the data. It is therefore of importance to know whether the updated knowledge becomes more and more accurate and precise as data are collected indefinitely.This has motivated researchers to come up with frequentist results related to convergence and consistency of Bayesian estimates, mostly for iid models.
The proof of Walker(1969) for asymptotic normality for posterior density in iid models could be used as a guideline for similar results for Hidden Markov Models.The Bayesian Central limit theorem (The Bernstein Von Mises theorem) tells us that provided the prior is continuous and positive, the posterior distribution of a parameter of interest, will be asymptotically normal with mean as the maximum likelihood estimate, and variance as the inverse of the observed information. We can establish this result heuristically by

taking a Taylor's expansion around the MLE and then using the inverse formula of characteristic functions to deduce normality. However to rigorously prove the original theorem, restrictive conditions of an iid model are a must. In later extensions and modifications of this result by Walter and others, the iid structure was always retained. There have been attempts to extend it to dependent processes and stationary data (mostly by Walker(2003,2004) and Ghoshal(2006)). Till date, there has not been any extensions to HMM. In our work, we have combined HMM asymptotics and Bayesian consistency approaches to arrive at similar results for our model.
While we relax the iid condition in our proof for HMM, we mimic Walker's technique. What went into the heart of the previous proof is actually a collection of results on convergence of log likelihood and observed information (which are easily obtained for iid models). These have been obtained in the earlier sections of HMM. The complexity of Walker's proof is also simplified by the fact that we do not get infinite integrals(as expectations of log-likelihoods) anywhere in our calculations if we assume a positive continuous prior π .

5.6.1 The proof of asymptotic convergence of posterior density under a continuous prior

We first state and prove three conditions related to the behavior of the log posterior. It is noteworthy that Walker used the same three conditions as the building blocks of the proof for iid models. Our proof, however, does not depend on the surplus conditions about log likelihood and information, but only on the assumption of bounded transition rates. These are :

1. Let $N_0(\delta) = \theta : |\theta - \theta_0| < \delta$ be a neighborhood of θ_0 . Then there exists a positive number, $k(\delta)$ such that
   $lim_{n\to\infty} P[sup_{\theta \in N_0(\delta)'} n^{-1}[L_n(\theta) - L_n(\theta_0)] < -k(\delta)] = 1$

2. $lim_{n\to\infty} L_n(\theta_n)'' = -J(\theta_0)$
   But this is precisely the conclusion from the theorem on the asymptotics of the observed score information by Douc et al and our result of consistency on $\hat{\theta_n}$. To see this, we invoke Douc's theorem

$lim_{\delta \to 0} lim_{n \to \infty} sup_{|\theta - \theta*| < \delta} || n^{-1}[L_n(\theta*)'' - L_n(\theta)'']|| \to 0$
Our earlier proved result on consistency states $\hat{\theta_n} \to \theta$ a.s
Combining the two, we get $lim_{n \to \infty} L_n(\theta)'' = -J(\theta_0)$

3. $L_n(\hat{\theta^n}) - L_n(\theta_0) = O_p(1)$ For this we use Taylor's expansion once again:
$L_n(\theta_0) = L_n(\hat{\theta_n}) + 1/2(\theta_0 - \hat{\theta_n})^2 L_n(\theta*)''$
where $\theta* \in (\theta_0, \hat{\theta_n})$
Now we have already shown that the limiting distribution of $n^{1/2}(\theta - \hat{\theta_n})$ is $N(0, J(\theta_0)^{-1})$ and from (2) we have $lim_{n \to \infty} L_n(\hat{\theta^n})'' = -J(\theta_0)$ This implies $L_n(\hat{\theta^n}) - L_n(\theta_0) = O_p(1)$

Before going into the main part of the proof we prove condition(1).
For this we use the Douc method(used also in the proof by Ryden et al) of approximating the log likelihood to stationary ergodic sequence. Using this property we have two stationary sequences h_{1k} and h_{2k}, for which the following holds
$|L_n(\theta_0) - \sum_{k=1}^{n} h_{1k}| \to 0$
$|L_n(\hat{\theta_n}) - \sum_{k=1}^{n} h_{2k}| \to 0$
Therefore $n^{-1}[L_n(\hat{\theta_n}) - L_n(\theta_0)] - n^{-1}[h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)] \to 0$
Now since h_{1k} and h_{2k} is stationary, $[h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)]$ is a stationary sequence.
Now since $L_n(\hat{\theta_n}) - L_n(\theta_0) = \sum_k [L_{c,k}(\hat{\theta_n}) - L_{c,k-1}(\theta_0)]$
where $L_{c,k}$ denotes the conditional log likelihood till position k,(Pi denoting the conditional likelihood) we have
$E[L_{c,k}(\theta_0) - L_{c,k-1}(\hat{\theta_n})] < log E_{\theta_0}[Pi(\hat{\theta_n})/Pi(\theta_0)] = 0$ and hence
$E(n^{-1}[L_n(\hat{\theta_n}) - L_n(\theta_0)]) < 0$ for all n
Since $n^{-1}[L_n(\hat{\theta_n}) - L_n(\theta_0)] - n^{-1} \sum [h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)] \to 0$
we have $E(n^{-1}[L_n(\hat{\theta_n}) - L_n(\theta_0)]) - E(n^{-1} \sum [h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)]) \to 0$
Since $[h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)]$ is a stationary sequence this has a stationary mean, say q. The above equation implies $q < 0$
By the ergodic theorem we have $n^{-1}[h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)] \to q$.
And now by finally invoking once again the approximation equation
$n^{-1}[L_n(\hat{\theta_n}) - L_n(\theta_0)] - n^{-1}[h_{1k}(\hat{\theta_n}) - h_{2k}(\theta_0)] \to 0$
we have

$lim_{n\to\infty}P[sup_{\theta\in N_0(\delta)'}n^{-1}[L_n(\theta)-L_n(\theta_0)]<-k(\delta)]=1$ Hence the first condition is proved.

We now come to the main part of the proof.

We denote $p_n(\hat{\theta_n}|Y)$ as the likelihood under the MLE and we denote $\pi_n(\theta|Y)$ as the posterior likelihood.

Note that $\pi_n(\theta|Y)=\frac{p_n(\hat{\theta_n}|Y)\pi(\theta)}{\int p_n(\hat{\theta_n}|Y)\pi(\theta)d\theta}$

Now,

$p_n(Y|\theta)=p_n(Y|\theta_n)exp(L_n(\theta)-L_n(\theta_0))$

$=p_n(Y|\theta_n)exp[(-(\theta-\hat{\theta_n})^2/2)(1+R_n)]$ where $R_n=L_n''(\theta^*)-L_n''(\theta_0)$ with $\theta^*\in(\theta,\hat{\theta_n})$.

We now split the integral into two parts I_1 and I_2 taken over the sets $N_0(\delta)$ and $N_0(\delta)$‘(the complement set)

$$I_1=p_n(Y|\theta_n)exp(L_n(\hat{\theta_n})-L_n(\theta_0))\int_{N_0(\delta)^{'}}\pi(\theta)exp(L_n(\theta)-L_n(\theta_0))d\theta$$

Now from the first condition the integral is less than

$exp(-nk(\delta))\int_{N_0(\delta)^{'}}\pi(\theta)d\theta$

This proves that the posterior is finite. This does not hold true if we have a flat prior.

We must note here, that this property of the integrability of the posterior kernel is shown for all types of transition structures assumed. We just need to assume the boundedness of the transition probabilities for all time points within the continuous time index markov chain. This again is a direct outcome of the boundedness of the transition rates. In the discussion in the third chapter on continuous index chains, we proved the posterior integrability (upto exclusion of a finite number of extreme states) using the same assumptions, and for a particular case of the transition structure. The same general argument can be applied there too.

Now from the second condition,

$exp(L_n(\hat{\theta^n})-L_n(\theta_0))=O_p(1)$ and from the third condition

$Plim_{n\to\infty}\sigma_n^{-1}exp(-nk(\delta))=Plim[-n^{-1}L_n(\theta_n)'']^{1/2}exp(-nk(\delta))$

$=J(\theta_0)^{1/2}lim_{n\to\infty}n^{1/2}exp(-nk(\delta))=0$

Hence

$lim[[p_n(Y|\theta_n)\sigma_n]^{-1}I_1]=0$

We define I_2 as the second integral .

$$I_2 = p_n(Y|\theta_n) \int_{N_0(\delta)} \pi(\theta) exp[(-(\theta - \hat{\theta_n})^2/2)(1 + R_n)]d\theta$$

Since the prior is continuous and positive at $\theta = \theta_0$ for any $\epsilon > 0$ we can choose δ so that $|\pi(\theta) - \pi(\theta_0)| < \epsilon\pi(\theta_0)$ if $\theta \in N_0(\delta)$
Then if we define

$$I_3 = \int_{N_0(\delta)} exp[(-(\theta - \hat{\theta_n})^2/2)(1 + R_n)]d\theta$$

we have

$$(1 - \epsilon)I_3 < [\pi(\theta_0)[p_n(Y|\theta_n)\sigma_n]^{-1}I_2] < (1 + \epsilon)I_3$$

Now $|(n\sigma_n^2)^{-1}R_n| = |(n\sigma_n^2)^{-1}R_n| = n^{-1}[L_n''(\theta_*) - L_n''(\theta_0)]$ where $\theta_* \in (\hat{\theta_n}, \theta_0)$
From the second condition and Douc's theorem on asymptotic equivalance of observed information, this can be bounded by ϵ for any fixed δ.
When $\epsilon < 1$,
$\int_{N_0(\delta)} exp[(-(\theta - \hat{\theta_n})^2/2)(1 \pm \epsilon)]d\theta =$
$(2\pi)^{1/2}(1 \pm \epsilon)^{1/2}\sigma_n[\Phi(\sigma_n^{-1}(\theta_0 + \delta - \hat{\theta_n})(1 \pm \epsilon)^{1/2}) - \Phi(\sigma_n^{-1}(\theta_0 - \delta - \hat{\theta_n})(1 \pm \epsilon)^{1/2})]$
The factor in square brackets converges in probability to unity. So we have
$lim_{n\to\infty}P[(2\pi)^{1/2}(1 + \epsilon)^{1/2}\sigma_n < I_3 < (2\pi)^{1/2}(1 - \epsilon)^{1/2}\sigma_n] = 1$ Since δ can be chosen for arbitrary small ϵ we have
$lim_{n\to\infty}P[(2\pi)^{1/2}(1 - \eta)^{1/2}\sigma_n < [\pi(\theta_0)[p_n(Y|\theta_n)\sigma_n]^{-1}I_2] < (2\pi)^{1/2}(1 + \eta)^{1/2}\sigma_n] = 1$ This shows that
$[p_n(Y|\theta_n)\sigma_n]^{-1}p_n(Y|\theta_n) = (2\pi)^{1/2}\pi(\theta_0)$
Now let $I_4 = \int_{\hat{\theta_n}+b\sigma_n}^{\hat{\theta_n}-a\sigma_n} \pi(\theta_0)p_n(Y|\theta_n)$
Since $lim_{n\to\infty}P[(\hat{\theta_n} + b\sigma_n, \hat{\theta_n} - b\sigma_n) \subset N_0(\delta)] = 1$
we can make the replacements in the earlier equation with the condition that the probabilities tend to 1 as n tends to infinity. Thus we no longer requiring the constraint on the parameter space as a condition on these equations. Hence
$lim[[p_n(Y|\theta_n)\sigma_n]^{-1}I_4] = (2\pi)^{1/2}\pi(\theta_0)[\Phi(a) - \Phi(b)]$
Therefore $\int_{\hat{\theta_n}+b\sigma_n}^{\hat{\theta_n}-a\sigma_n} \pi_n(\theta|Y)d\theta = I_4/p_n(Y)$
converges in probability to $\Phi(a) - \Phi(b)$ which is the required result.

5.6.2 Consequences of posterior consistency and convergence on Bayesian inference in nucleosome positioning models

The results for asymptotics that has been developed all along this chapter are not all directly relevant to Bayesian inference. The main results that link our work to Bayesian inference is as follows:

1. Consistency of the MLE. For a flat prior this translates into the consistency of the posterior mode.

2. Asymptotic convergence of log-likelihood. This has been used to determine the behavior of BIC of estimation models in simulation settings

3. Asymptotic $\sqrt(n)$ convergence of the posterior. This gives us the rate of convergence of the posterior density, and since any continuous prior distribution is used, guarantees the asymptotic distribution of the posterior for a more general setup.

We must however note, that almost all of the subsidiary frequentist results proved along the way contribute to the above conclusions. While we have proved the first two results in the section of consistency, the third result uses asymptotic normality of MLE and a condition on observed information score-a powerful tool that went into proving contiguity Therefore, it is important to see that the frequentist results do not stand on their own, but is closely connected to its role in Bayesian inference. Instead of directly trying to prove only the results relevant to the Bayesian setup, we thought that is worth the effort to build a more complete and unified theory of frequentist asymptotics for continuous index HMM.
We now discuss the implication of these three results.
For result 1, the consistency results suggest that the estimated parameters will converge to the true parameter almost surely. In our simulation experiments we got MSEs as low as .001 for the Bayesian parameter estimates. See Tables **??**, **??** for base and emission models in Chapter 3. This pattern did not change with the different datasets. Also in Chapter 2, we have clearly shown how the parameter estimates MSE decrease towards zero with increase in the number of observations. See Table **??**.

This is indeed the Bayesian consistency result for HMM which is coming to play over here. In our sufficiently large datasets, the Bayesian estimates obtained are close to the true parameter with a very high probability. This implies we can have as much faith in our estimates from estimating a hidden markov model through MCMC on a large dataset, as we have when we are estimating an iid model on a large sample.
For result (2), we extend the consistency results of the parameter estimates to estimated log likelihood and hence to Bayesian Information Criterion. Note that
$BIC = -2log\hat{L} + klog(n)$ where n is the sample size and k is the number of parameters.
We know that:
$\hat{\theta_n} \to \theta_0$
Since L_n is continuous , there exists a sufficiently large N such that for all $n > N$
$\Rightarrow |L_n(\hat{\theta_n}) - L_n(\theta_0)| < \epsilon$ Now we invoke the result
$L_n(\theta_0)/n \to l(\theta_0), -P_{\theta_0}$ a.s. This is the main result proved in the section on consistency.
This implies that for sufficiently large n, $|n^{-1}(L_n(\hat{\theta_n}) - l(\theta_0))| < \epsilon$ for any ϵ.
The above result implies the asymptotic similarity of the actual and estimated data log likelihoods.
This combined with result $L_n(\theta_0)/n \to l(\theta_0), - - P_{\theta_0}$ a.s, implies that the estimated data log-likelihood converge to a single number dependent only on the parameter setting and the data-length. Also this number is equal to the actual data log likelihood for the model. This explains the pattern we see in the BIC tables for simulation settings in chapter 3.
The estimation models are superset of the simulation model. We must note that since the maximization is done over a bigger space, the MLE s could be potentially different. However we must also note that when the data is coming from the subset model, it is equivalent to assume that the data is generated from a superset model with extra parameters fixed to 0.The rest of them have the same values as the subset model. When we are estimating data coming from such a model with the same estimation model, our MLEs must converge to the true parameter set. This is the parameter space of the subset model, concatenated with 0s. Hence the estimated data log likelihood will be very close to the actual data log likelihood computed with this extended parameter set. In the same way when we use the simulation model in estimation, we see that the estimated data log-likelihood will be very close to the actual data log-likelihood. But these two data log

likelihoods computed (1) with the extended parameter set and (2) with the non extended parameter set, is the same, because we have the extra parameters in (1) fixed at null value. Hence, both the estimated log likelihoods are very close to a specific number (dependent on n and the parameter setting). For large sample size this is bound to occur with probability tending to one, for any dataset used. This suggests that if the simulation model is the minimal model, (in our case it is the base model) the BIC criterion will be driven solely by the number of parameters used. Hence the minimal model will emerge as the one with the lowest BIC criterion.

For simulation models which are not the minimal model (e.g transition and emission) the situation is little different.For convenience, let us fix the simulation model to be transition. When using a subset estimation model we are, in effect, maximizing our log likelihood within a constrained space. Hence the estimated log-likelihood of the base model will be always less than that under the simulation model. For the emission estimation model too, data coming from a transition model is equivalent to data coming from a combined transition-emission model where the emission parameters are fixed to 0. Estimating it by an estimation model is equal to estimation by a similar combined model, where the transition parameters are fixed to zero. This constraint on the estimation space once again makes the estimated data log-likelihood lesser.(unlike the earlier case on superset estimation models where we do not have constraints on the space of estimating models. But in both scenarios the parameter space is reduced for simulation models). So in all these settings we see that going solely by the criterion of estimated data log-likelihood only the simulation model will be chosen as the best model with probability 1. In the minimal simulation models, this principle extends to BIC. It might not be true for non minimal simulation models, as the increase in number of parameters might override the advantages of greater log likelihood. In our setting, however we see that BIC chooses the simulation model as the best model in 100 percent of the cases.

Result 3 implies that the rate of convergence of posterior is $\sqrt{(n)}$. This suggests that as the data size increases the shape of the posterior will be normal. Also, the distribution will grow tighter around the true parameter, with the length of a fixed alpha confidence interval rapidly shrinking at the rate of $\sqrt{(n)}$. In genome technology data like Chip-Chip and Faire, the size of the dataset is large. Thus we can have an accurate idea of exact

proximity of our parameter estimates to the true values for a particular dataset length. From the previous consistency result, we only could predict that the estimates will be close. This result gives us a direct relationship between the dataset length and the convergence.
However we must be clear about one restriction on the increase in data size.
We cannot assume that with the increase in data the discrete number of different gap lengths tend to infinity, or that the gap lengths themselves tend to infinity. This will result in an unboundedness of the 't'–the gap length parameter, in the transition function. This would go on to imply that the transition probabilities will not be bounded in a compact subset of [0,1]. All our inferences in this chapter are based on this boundedness assumption. So these results of convergence and consistency are good only under the condition that with the increase in data length size, the number of differing gap lengths will be bounded under a finite set.
Lastly, these results not only validate and strengthen the the inferences obtained in Chapters 3 and 4, but are fairly general in nature to provide the justification for HMM inference in a wide variety of datasets, specially in the fields of genomics and signal processing.

5.7 Conclusions

In this chapter of our thesis, we have extended the asymptotic MLE results to a continuous time index HMM. Such models are relevant in genomic data, where the probes are not equispaced. The analysis is more complicated since the conditional likelihoods lack a closed form. We started from some basic assumptions on the boundedness of transition rates and emission densities. The stationarity property of the underlying continuous markov chain and the ergodicity implied by the finite number of states in the model led us to consistency and normality results. One major step in the proof of asymptotic normality of the parameter estimates is the proof of asymptotic normality of the score function. We used this in establishing contiguity. The proof of contiguity is similar to the the iid case in the sense that we used Taylor's expansion to extract out the score function for which similar results hold. But it departs significantly from previous

approaches in our use of the asymptotic equivalence of the Information matrix and Fisher's matrix. Lastly, as in the case of i.i.d. observations, we obtained results on the long term behavior of Bayesian estimators and posterior for HMMs, directly from the asymptotic properties of the model log-likelihood, score function, and parameter estimates.

Chapter 6

Conclusion and Future directions

Nucleosomes are units of chromatin structure, consisting of DNA sequence wrapped around proteins called *histones*. Nucleosomes occur at variable intervals throughout genomic DNA, and prevent transcription factor (TF) binding by blocking TF access to the DNA. A map of nucleosomal locations would enable researchers to detect TF binding sites with greater efficiency. Our objective is to construct an accurate genomic map of nucleosome-free regions (NFRs), based on data from high-throughput genomic tiling arrays in yeast. These high volume data typically have a complex structure, in the form of dependence on neighboring probes as well as underlying DNA sequence, variable sized gaps and missing data. Previous methods often relied on the use of hidden Markov Models (HMMs), and ignored the role of sequence information in predicting nucleosomal positions. We propose a novel continuous time model appropriate for non-equispaced tiling array data, that simultaneously incorporates DNA sequence features relevant to nucleosome formation. Simulation studies and an application to a yeast nucleosomal assay demonstrate the advantages of using the new modeling framework, as well as its robustness to distributional misspecifications. Our results reinforce the previous biological hypothesis that higher order nucleotide combinations are important in distinguishing nucleosomal regions from NFRs.

Bayesian methodologies have a growing relevance in the field of genomics. In our work on the first paper, we have applied these methodologies and made extensions of the same, keeping in mind the new features of the data outputted from high throughput genome

technology. So in this way, our methodological approach, is driven by the current requirements that these new types of data demand. However, it is also a more general technique for inference from missing continuous data. This could have widespread applications for a model based classification approach in several areas of biomedical research . The asymptotic properties of the model, once established, would help us in using the distributional properties of the normal distribution in enriching our inferential procedures. We shall be able to test a set hypothesis related to length distributions of nucleosomal states and the gene signal features, directly from our proposed model.
The huge amount of genomic data is a blessing in terms of our ability to apply asymptotic results for inference. But it is also a curse, in term of computational time involved in analysis. This is also one area where we can concentrate on, for future work. Metropolis hastings algorithms applied conditionally on a set of parameters, utilizes the multiple computation of log likelihood in a model. For larger data sets, there is a need to plan efficient proposal distribution, for a multivariate Metropolis sampling, that can simultaneously cut down computational cost and give us a reasonable rejection rate.
The other technique that can be possibly applied in case of larger data sets is by combining our hidden markov model with the wavelets and peaks analysis– which have been used earlier to infer from gene signal data. The idea behind this approach would lie in the fact that the main differences between the two states are in the sharp rises and falls of the signal. Hence effectively, the data can be compressed to a size of reasonable amount for HMM application. For this, we need to formulate a model for the distribution between subsequent crests and troughs, (extreme value distributions can be used in this context) in order to statistically validate this compression.
The other feature of the data that can be a potential hurdle in future applications, is the differing gap structure between replicates. We need to see how this can be handled by the continuous time frame work.
The unified nucleosome model proposed in the second paper can be possibly extended in order to account for nucleosomal length constraints. We have formulated a combined model where the transition parameters are replaced by length distributions. However we need to put certain constraints on the assumptions of these distributions and the associated structure, in order to reduce computational complexities.

Lastly, there is a scope to extend the asymptotic work on continuous time HMM to areas where we have a specific form of gap structure., i.e where the gaps are not random, but follow a certain distribution. The extension of the state space to include a finite number of states , with bounded transition probability functions is obvious. The incorporation of these assumptions into our model would help us in applying our results to a wide variety of data sets.

Bibliography

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41:164–170.

Buck, M. and Lieb, J. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–60.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10096–10100.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.*, 27:167–171.

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003a). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 100:3339–3344.

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003b). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, 100(6):3339–3344.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38.

Gelfond, J., Gupta, M., and Ibrahim, J. G. (2009). A Bayesian hidden Markov model for jointly modeling probe sequences and ChIP-chip data for motif discovery. *Biometrics*, doi:10.1111/j.1541-0420.2008.01180.x.

Gupta, M. and Ibrahim, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Am. Stat. Assoc.*, 102(479):867–880.

Gupta, M. and Liu, J. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *J.Am.Stat.Ass*, 98:55–56.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104.

Hogan, G. J., Lee, C.-K., and Lieb, J. D. (2006). Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet*, 2(9):e158.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, 296:1205–1214.

Johnson, W., Liu, X., and Liu, J. (2009). Doubly-Stochastic Continuous-Time Hidden Markov Approach for Analyzing Genome Tiling Arrays. *Annals of Applied Statistics*.

Kechris, K. J., van Zwet, E., Bickel, P. J., and Eisen, M. B. (2004). Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biol.*, 5:R50.

Keles, S., van der Laan, M. J., Dudoit, S., Xing, B., and Eisen, M. B. (2003). Supervised detection of regulatory motifs in DNA sequences. *Stat Appl Genet Mol Biol*, 2:Article5.

Koch, M. A., Weisshaar, B., Kroymann, J., Haubold, B., and Mitchell-Olds, T. (2001). Comparative genomics and regulatory evolution: conservation and function of the Chs and Apetala3 promoters. *Mol. Biol. Evol.*, 18:1882–1891.

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 84:2363–2367.

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, 39:1235–1244.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, 90:1156–1170.

Liu, X., Brutlag, D. L., and Liu, J. S. (2001). Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, pages 127–138.

Liu, X., Noll, D. M., Lieb, J. D., and Clarke, N. D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.*, 15:421–427.

Liu, Y., Liu, X. S., Wei, L., Altman, R. B., and Batzoglou, S. (2004). Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, 14:451–458.

Marsan, L. and Sagot, M. F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.*, 7:345–362.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, 36:1331–1339.

Narlikar, L., Gordan, R., and Hartemink, A. J. (2007). A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, 3:e215.

Ozsolak, F., Song, J. S., Liu, X. S., and Fisher, D. E. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, 25:244–248.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Res.*, 17:1170–1177.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K., and Young, R. A. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122:517–527.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309.

Richmond, T. J. and Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, 423:145–150.

Roberts, W. and Ephrain, Y. (2008). An EM Algorithm for Ion-ChannelCurrent Estimation. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 56:Issue 1.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442:772–778.

Segal, E. and Sharan, R. (2005). A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol.*, 12:822–834.

Sekinger, E. A., Moqtaderi, Z., and Struhl, K. (2005). Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell*, 18:735–748.

Stjernqvist, S., Rydn, T., Skld, M., and Staaf, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, 23:1006–1014.

Teicher, H. (1967). Identifiability of mixtures of product measures . *Annals of Mathematical Statistics*, 38:1300–1302.

Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S., and Lawrence, C. E. (2004). Decoding human regulatory circuits. *Genome Res.*, 14:1967–1974.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, 26:225–228.

Widom, J. (1992). A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A.*, 89:1095–1099.

Yassour, M., Kaplan, T., Jaimovich, A., and Friedman, N. (2008). Nucleosome positioning from tiling microarray data. *Bioinformatics*, 24:i139–146.

Yuan, G. C. and Liu, J. S. (2008). Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, 4:e13.

Yuan, G. C., Liu, Y. J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., and Rando, O. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309:626–630.

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20:909–916.