An Oligonucleotide Microarray-Based Method for Determining Nucleosome Positioning in *Saccharomyces cerevisiae*

A Thesis presented

by

Yuen-Jong Liu

to

Computer Science and Biochemical Sciences

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

April 2, 2004

to my family

and to the yeast in our experiments

who gave their lives

for the advancement of science and the pursuit of knowledge

SUSCEPIT AUTEM VITA HOMINUM CONSUETUDOQUE COMMUNIS
UT BENEFICIIS EXCELLENTIS VIROS IN CAELUM
FAMA AC VOLUNTATE TOLLERENT.

It is a generally recognized custom among human beings
that those individuals who have excelled in services to humanity
should be raised to the heavens in fame and gratitude.

Marcus Tullius Cicero
De Natura Deorum, *On the Nature of the Gods*
2.62

Statement of Research

The following work was performed under the direction of Drs. Steven J. Altschuler and Lani F. Wu at the Bauer Center for Genomics Research. It was carried out in a period of time from June, 2003, through April, 2004.

Abstract

Eukaryotic DNA is packaged into nucleosomes, the basic repeating unit of chromatin. Consisting of approximately 146 base pairs wrapped around a core protein complex, nucleosomes have been shown to regulate gene expression both positively and negatively. Conventional methods to identify nucleosome positions are limited by the relatively small amount of DNA that can be assayed in one experiment. We have utilized a microarray-based technique that simultaneously measures the positions of thousands of nucleosomes across multiple genes. To facilitate the analysis of the microarray data, we developed a hidden Markov model (HMM) as an objective way to determine nucleosome positioning, thereby automatically and systematically processing DNA hybridization data on a genomic scale. I improved the HMM by incorporating the ability to simultaneously process data from multiple replicates, to consider the effects of cross-hybridization, and to ignore unusable data from poorly-hybridized microarray spots. After developing a graphical user interface for visualizing nucleosome positions, I found wide variations in nucleosome density. There was a significant difference in nucleosome density between gene-coding and intergenic regions. Though our preliminary results did not show significant correlation between transcription level and nucleosome density on promoter regions, exploration and quantification in this area is now possible.

1 Introduction

In eukaryotic cells, DNA needs to be packaged mainly to control the availability of genes for transcription and to prevent entanglement of its long strands. Nuclear DNA is packaged with various proteins to form chromatin. Chromatin structure is based on repeating units of the nucleosome, which consists of an eight-histone-molecule complex (containing two units each of H2A, H2B, H3, and H4) and approximately 146 DNA base pairs wrapped around it [1]. Whereas the length of the DNA in nucleosomes is believed to be fairly consistent, the linker DNA between consecutive nucleosomes can vary in length. Short linkers are typically associated with extremely compact chromatin structure, and long linkers give the appearance of "beads on a string" under electron microscopy (nucleosomes constitute the beads and linkers constitute the string) (Figure 1) [2]. Nucleosome-bound DNA can be further folded into higher-order compaction known as the 30-nm fiber, in which the nucleosomes and linkers are most likely folded in a zig-zag conformation [3, 4].

By affecting the positions of nucleosomes over regulatory elements in a promoter, chromatin structure plays an important role in regulating transcription. Nucleosome remodeling, the modification of nucleosome positions, appears to be a direct consequence of recruiting and assembling the holoenzyme required for transcription [5]. When bound to promoter regions, nucleosomes have been shown to prevent the initiation of transcription by bacterial and eukaryotic RNA polymerases through steric hindrance, *in vivo* [6] and *in vitro* [7, 8]. Nucleosomes are also capable of inducing gene expression by shifting their positions to allow the assembly of molecules necessary for transcription. For example, a change in nucleosome position is required for transcriptional activation of the human IFN-β gene in HeLa cells in response to viral infection. A nucleosome blocking the core promoter must slide to a downstream position, demonstrating the importance of temporal and spatial configuration of nucleosomes [9].

Traditional analysis of nucleosome positioning consists mainly of DNA footprinting with micrococcal nuclease (Figure 1), followed by Southern blots [10]. This technique is limited to measuring the positions of two or three nucleosomes at a time, usually in a single promoter region. Earlier global studies of chromatin structure were limited to low-resolution analysis of histone modification and histone subunit composition [11, 12, 13]. However, using an oligonucleotide microarray-based method [14, 15], we have recently developed a technique to identify the positions of several thousand nucleosomes in a single experiment (Figure 2) with a precision of about 10-20 base pairs. Our method allowed us to examine the structure and function of chromatin on a genomic scale.

The nucleosome microarray consists of short oligonucleotides designed to have overlapping sequences that corresponded to DNA regions of interest. DNA isolated from mononucleosomes was tagged with fluorophores and hybridized to the microarray. A scanned image identified fluorescent spots where mononucleosomal DNA hybridized to the oligonucleotides, which were then mapped to precise locations on the genome. Labeled genomic DNA was also hybridized to the microarray as a control to identify base-line levels of fluorescence.

We developed our method of nucleosome analysis using *Saccharomyces cerevisiae*, commonly known as baker's or budding yeast, as it is one of the best-characterized and simplest of eukaryotic organisms. In *S. cerevisiae*, chromatin remodeling occurs during the activation of the *MFA2* gene, specific to haploid mating type **a** cells [10], the *SUC2* gene encoding invertase, required for growth on sucrose [16, 17], and the *PHO5* gene, the structural gene for a highly regulated acid phosphatase [18, 19, 20, 21].

In our initial work using this microarray, we tested our method on the promoter regions of the *MFA2*, *SUC2*, and *PHO5* genes. DNA oligonucleotides with lengths of forty, fifty, and sixty nucleotides were tiled every twenty base pairs. We determined the optimal oligonucleotide length to be fifty nucleotides, which allowed us to

detect artificial linkers as short as seven base pairs. To distinguish between nucleosomes and linkers on the promoter regions of these genes, it was possible by visual inspection to set a threshold that separated the hybridization values of nucleosomes from those of linkers. On a genomic scale, however, the much larger amounts of data made it impractical to survey all the hybridization values manually. In order to automatically and systematically process the hybridization data, we needed a method to objectively determine the nucleosome/linker boundaries, thereby filtering out noisy artifacts. Variation between replicate experiments, within a single microarray chip, and in oligonucleotide-specific characteristics, such as cross-hybridization potential and GC content, could potentially introduce noise into the data.

To this end, I have developed a hidden Markov model (HMM) which accounts for the relatively static length of nucleosome-protected DNA, variability of linker lengths, the noisiness of the data, potential cross-hybridization artifacts, and data from biological and technical replicates. This model was first used to infer nucleosome positions on chromosome III and 233 genes of interest, including 100 genes from chromosomes II, XIV, and XVI, and 100 genes regulated by the cell cycle, the Swi/Snf chromatin remodeling complex, histone depletion, and histone tail modification.

I first determined that there exist very dense regions of nucleosomes, with extremely short linkers, as well as very sparse regions of nucleosomes. In addition, I also found that nucleosome density is higher in gene-coding regions than in intergenic regions, a result that Nagy *et al.* have recently discovered by a different technique [22].

2 Methods

2.1 Experimental

The methods described in this section were carried out by members of the Oliver Rando laboratory at the Bauer Center for Genomics Research.

2.1.1 High Throughput Analysis of Nucleosome Positions

Yeast cells were grown to mid-logarithmic phase. Their nuclei were treated with formaldehyde to cross-link DNA to nucleosomes [23]. The DNA was then digested with micrococcal nuclease, which destroyed linkers (Figure 1), and mononucleosomal DNA was isolated through gel electrophoresis [23]. DNA was released from nucleosomes and labeled with Cy5, a red fluorescent dye marker. Genomic (undigested) DNA was labeled with Cy3, a green fluorescent dye marker. These probes were mixed and hybridized to a microarray consisting of fifty-mer oligonucleotides. The oligonucleotides were designed to have overlapping sequences that corresponded to the DNA of interest: chromosome III and 233 genes from other chromosomes, including those that were regulated by the cell cycle, Swi/Snf chromatin remodeling complex, histone depletion, and histone tail modification. Chromosome III, the shortest chromosome in *S. cerevisiae*, was tiled in its entirety. For each of the 233 other genes, 900 base pairs of upstream, non-coding sequence and 100 base pairs of $5'$ coding sequence were tiled (Figure 2).

The overlap between consecutive oligonucleotides was constructed such that each oligonucleotide was offset from the next by twenty nucleotides. The hybridization value, or log ratio, for each oligonucleotide was calculated as the base-two logarithm of the Cy5 channel divided by the Cy3 channel,

$$\text{log ratio} = \log_2 \frac{\text{Cy5}}{\text{Cy3}} = \log_2 \frac{\text{nucleosomal channel}}{\text{genomic channel}} = \log_2 \frac{N}{G} \tag{1}$$

where N represents hybridization by nucleosomal DNA and G represents hybridization by genomic DNA. Thus, nucleosomal oligonucleotides produced higher log ratios, while linkers produced lower log ratios.

2.1.2 Artificial Nucleosomes

To aid in normalization, "artificial nucleosome" data was produced by hybridizing PCR-amplified sequences of DNA to our nucleosome microarray. Chromosome III was partitioned into non-overlapping contiguous sequences with an average length of about 6,000 base pairs per segment. These sequences were numbered and sorted into "even" and "odd" sets. An "even" experiment denotes one where only the even set of sequences were PCR-amplified, and an "odd" experiment denotes one where only the odd set of sequences were PCR-amplified. Several even and odd replicates were used in supervised learning to test our model.

The microarray method described in Section 2.1.1 was utilized, except that an equal amount of genomic DNA was also hybridized in the Cy5 channel. Similar to Equation 1, the hybridization value for each oligonucleotide in the "artificial nucleosome" experiments was calculated as

$$\text{log ratio} = \log_2 \frac{\text{Cy5}}{\text{Cy3}} = \log_2 \frac{\text{nucleosomal channel}}{\text{genomic channel}} = \log_2 \frac{N+G}{G} \tag{2}$$

where N represents hybridization by nucleosomal DNA and G represents hybridization by genomic DNA. We had found that the addition of genomic DNA in the nucleosomal channel was useful in reducing the dynamic range of the microarray output. Since N could theoretically vary from 0 in the absence of a nucleosome to $N_{\max}$ in the presence of one, the last expression of Equation 1 could vary from $\log_2 0 = -\infty$ to $\log_2 \frac{N_{\max}}{G}$, whereas the last expression of Equation 2 could vary from $\log_2 \frac{G}{G} = 0$ to $\log_2 \frac{N_{\max}+G}{G}$.

2.2 Computational

2.2.1 Normalization

Our microarray data was normalized to minimize variation in hybridization values among replicate experiments. The data was also normalized within the microarray chip of each experiment, as oligonucleotides near the edge of the chip might not be as protected by the coverslip, thereby becoming exposed to humidity and other factors in the environment. I performed preliminary normalization by linear regression, allowing individual experiments, blocks, and local regions to contribute a linear term to the hybridization values. Subsequently, Dr. Guocheng Yuan, a postdoctoral fellow in our laboratory, performed statistical analyses and normalized the data by rescaling followed by linear regression.

2.2.2 Cross Hybridization

While gene expression arrays are typically designed with oligonucleotides selected to minimize cross-hybridization with the rest of the genome [24, 25], cross-hybridization could not be avoided in our microarrays because the oligonucleotides were tiled according to of the actual DNA sequence of our regions of interest. A cross-hybridizing oligonucleotide will bind to many locations on the genome, leading to an increase in G in Equations 1 and 2. When a nucleosome is present, $N \gg G$, and an increase in G will cause $\left(\frac{N}{G}\right)$ to decrease and $\left(\frac{N+G}{G}\right)$ to approach 1 from above. Thus, cross-hybridization was expected to decrease the log ratio in both our real nucleosome and "artificial nucleosome" microarrays. I estimated the cross-hybridization potential of an oligonucleotide in several ways, using WU-BLAST 2.0, a search engine for nucleotide homologies, on a Linux 2.4 cluster.

First, I found all BLAST matches between each oligonucleotide and the yeast genome. I filtered for the hits with at least 60% sequence match. The percentages of

sequence match for each potential target were added.

Second, I identified sites on the genome that potentially cross-hybridize to each oligonucleotide. The ratios of predicted ΔG values for cross-hybridizing and non-cross-hybridizing matches were calculated from estimated nearest-neighbor thermodynamic parameters [26, 27, 28, 29, 30, 31]. Because the published thermodynamic parameters are limited to dinucleotide pairs with zero or one mismatch, the ΔG values of duplexes with gaps or more than one consecutive mismatch were estimated by the lowest ΔG of duplex subsequences with no more than one consecutive mismatch or by the ΔG of the whole duplex without considering contributions from gaps or consecutive mismatches.

Third, cross-hybridization potential was estimated by

$$\mathrm{XHYB} = \sum_{i=1}^{k} \exp\left(-\frac{\Delta G - \Delta G_i'}{RT}\right) \tag{3}$$

XHYB is an abbreviation for cross-hybridization potential, ΔG is the free energy change for non-cross-hybridizing duplex formation, $\Delta G_i'$ are the free energy changes for cross-hybridizing duplex formation, k is the number of potentially cross-hybridizing matches returned by BLAST, $R = 1.9872156 \times 10^3\,\mathrm{kcal}/(\mathrm{mol}\cdot\mathrm{K})$ is the gas constant, and $T = 310\,\mathrm{K}$ is the temperature at which the nearest-neighbor thermodynamic parameters were derived.

Fourth, for each oligonucleotide, I calculated the number of BLAST matches with at least n matching base pairs, using threshold levels $1 \leq n \leq 50$. Since I was trying to model two distinct output distributions, one for nucleosomes and one for linkers, I divided the oligonucleotides in our "artificial nucleosome" microarrays into the "artificial nucleosome" oligonucleotides, which hybridized to the PCR-amplified sequences of DNA, and the "artificial linker" oligonucleotides, which only hybridized to genomic DNA. I mainly considered the former category of oligonucleotides be-

cause cross-hybridization is most capable of distorting hybridization values when a nucleosome is present. Because the majority of these oligonucleotides did not have BLAST matches with cross-hybridizing sites on the genome, I partitioned them into two groups, (1) those which had no cross-hybridizing BLAST matches with at least n matching base pairs and (2) those which had at least one such match. (For example, at $n = 32$, of the 4998 oligonucleotides that were hybridized to "artificial nucleosomes," 3924 had no cross-hybridizing BLAST matches with at least n matching base pairs, 681 had exactly one BLAST match, and 393 had more than one.) For each oligonucleotide, I calculated the mean and standard deviation of its hybridization values across eight replicates of the "artificial nucleosome" data. Using the Wilcoxon rank-sum test, a non-parametric test that makes no assumptions of the shapes of the distributions, I tested the hypothesis that the means and standard deviations were drawn from the same distribution (Figure 3). I only tested the hypothesis for the means of the oligonucleotide hybridization values across replicates because we were more concerned with how cross-hybridization potential affects the mean output distribution. The Wilcoxon rank-sum test revealed a local optimum in confidence level, or p-value, at a threshold of $n = 32$ (Figure 4). The estimation of cross-hybridization was simplified to a Boolean value, such that any oligonucleotide with cross-hybridizing BLAST matches with at least 32 matching base pairs was considered to have cross-hybridization potential.

2.2.3 Hidden Markov Model

To objectively determine nucleosome/linker boundaries, we implemented a hidden Markov model (HMM). HMMs were originally developed for speech recognition [32] and have been applied to various problems in biology [33] such as finding genes in bacterial genomes [34] and nucleosome positioning signals in human intron and exon sequences [35]. An HMM makes three basic assumptions about the system at hand.

First, it assumes that the system passes through a sequence of hidden or unobservable states. However, at each point in the sequence, we have one or more observations that reflect upon the current state of the system in some way. The collection of observations and hidden states at each point in the sequence is called a *slice*. Second, the discrete or continuous observation at any node in a slice is conditionally independent of all other nodes given that the values of its parent nodes, the nodes that directly affect it, are known. Third, an HMM assumes that the state of the system is independent of history, in that the future states of the system are conditionally independent of all past states and observations given that the current hidden state is known. This is the Markov assumption and can be restated [36] as "The future is independent of the past given the present."

Our original implementation of the HMM sufficed for our preliminary work and was able to infer nucleosome positions that matched those published in the literature for regions such as the *MFA2* promoter (Figure 5). This HMM separated the hybridization values from the nucleosome microarrays into two categories, nucleosome-generated and linker-generated, represented by red and green, respectively (Figure 6a). The two categories were modeled as two separate Gaussian probability distributions (Figure 6b). The hidden state transition graph (Figure 6c) incorporated the fact that it takes approximately seven consecutively-tiled oligonucleotides to cover a nucleosome binding site. (The fifty-mer oligonucleotides are tiled every twenty base pairs, and each nucleosome binding site is 146 base pairs long on average [1].) The HMM optimized this categorization of oligonucleotides into nucleosomes and linkers through several passes of the expectation-maximization (E-M) algorithm and finally inferred the most likely sequence of nucleosome and linker states (Figure 6d) that matched the hybridization values on our microarray. From another point of view (Figure 7), the HMM represented the hidden states, nucleosome or linker, in the H nodes and the observed hybridization values in the O nodes. The arrows represent

the directed transitions for the hidden states according to the state transition graph, as well as the causal relationships between hidden states and the output distributions of the observed hybridization values.

However, our original HMM required further modifications in order to handle the larger output from genomic-scale nucleosome microarrays. In particular, we needed to normalize the microarray data prior to HMM analysis, thereby minimizing variation between different microarray experiments and between blocks on the same microarray chip. Oligonucleotide-specific biochemical characteristics, particularly cross-hybridization potential, were also likely to confound the HMM by altering the output distributions. Thus, I designed and implemented an improved HMM to be used in the same manner as described above. I augmented the topology of the HMM (Figure 8), incorporating information about potentially cross-hybridizing oligonucleotides into the X nodes and accounting for user-flagged, unusable microarray spots (Figure 9) in the F nodes. The calculation of values for the X nodes is described in Section 2.2.2. The value of the F node for an oligonucleotide is 1 if the corresponding microarray spot is unusable; normal spots have a value of 0.

Given an experimental data set, the HMM used the Baum-Welsh algorithm [33] to estimate the most likely values of seven parameters: μ_N (mean of nucleosomal output), μ_L (mean of linker output), σ_N (standard deviation of nucleosomal output), σ_L (standard deviation of linker output), w_N (contribution of cross-hybridization to nucleosomal output), w_L (contribution of cross-hybridization to linker output), and p (probability of remaining in the linker state). Though cross-hybridization potential was incorporated into the model, the HMM was free to assign as much weight to it as suggested by the data. The fitted model was then used to calculate the probability that a sequence of oligonucleotides on the nucleosome array corresponded to nucleosome or linker regions. In addition, the Viterbi algorithm allowed the model to compute the most likely explanation for the observed hybridization values, i.e. the

most likely sequence of hidden states.

2.2.4 Hardware and Software

Processing of the microarray data was conducted primarily in Matlab on a Mac OS X platform. The hidden Markov model was implemented using Kevin Murphy's Bayes Net Toolbox,[1] which includes algorithms such as Baum-Welch and Viterbi, to learn model parameters from the microarray data, to compute the likelihood of the data, and to estimate the most likely positions of nucleosomes and linkers to generate the data. Other software, such as Microsoft Excel and ad hoc Perl and shell scripts, were also used to handle the data.

3 Results

3.1 Graphical User Interface

I developed an automated method for simultaneously visualizing multiple replicates of our nucleosome microarray data (Figure 10), including the HMM-inferred nucleosome positions and gene-coding regions. The likelihood of a nucleosome's presence at any location is visualized by the thickness of a line that follows the mean of nucleosomal output and by a graph of the likelihood versus genomic position. In addition, nucleosome density is shown by a plot of p, the probability of remaining in the linker state. Higher values of p correspond to sparser nucleosome density and lower values correspond to denser nucleosome density. Finally, the cross-hybridization potentials are aligned with each oligonucleotide.

[1] The Bayes Net Toolbox is available from `http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html`.

3.2 Oligonucleotide-Specific Characteristics

As described in Section 2.1.2, we produced "artificial nucleosome" data as training sets from which we hoped to learn the effects of GC content and cross-hybridization potential on the hybridization values of the nucleosome microarray. For each oligonucleotide, I tabulated the mean and standard deviation of the log ratios across eight replicates of the "even" experiments. By plotting the means and standard deviations against the GC content and cross-hybridization potential of each oligonucleotide, I hoped to estimate the effects of these oligonucleotide-specific characteristics on the data.

The GC content of a DNA sequence affects its melting temperature and hence directly influences its binding affinity [25]. However, we were not sure how GC content would affect the hybridization values, and we hoped to find a correlation by analyzing the "artificial nucleosome" data. I attempted to correlate GC content with the mean and standard deviation of the hybridization signal using two common measurements: the total number of G's and C's in each oligonucleotide sequence or the length of the longest continuous stretch of G's and C's in the sequence. We were unable to detect any correlation through visual inspection or linear regression. Without clear evidence that GC content affects the hybridization outputs, I made the design choice to exclude this parameter from the HMM.

Another parameter that is specific to each oligonucleotide is its ability to cross-hybridize with alternate locations on the genome. Unlike gene expression arrays, tiled arrays must include all sequences in a certain region of the genome and therefore may include areas that share homology with other regions of the genome [37]. While gene expression arrays are designed to minimize cross-hybridization, this source of false positives is unavoidable in tiled arrays [24, 25]. After testing all the cross-hybridization estimation methods described in Section 2.2.2, I decided to utilize the simplified Boolean method as described below. X is set to 1 for all oligonucleotides

that have a BLAST match with at least 32 perfect base pairs; otherwise, X is set to 0. To learn the contribution of cross-hybridization potential in the real nucleosome data, I modeled the observed nodes O (Figure 8) as

$$O = \log_2\left(\frac{\text{Cy5}}{\text{Cy3}}\right) \sim \begin{cases} N(\mu_N + X \cdot w_N, \sigma_N) & \text{if } H = \text{N}_1, \ldots, \text{N}_7 \\ N(\mu_L + X \cdot w_L, \sigma_L) & \text{if } H = \text{L} \end{cases}$$

where $N(\mu, \sigma)$ is a normal distribution with a mean of μ and a standard deviation of σ. In addition to the parameters $\mu_N, \sigma_N, \mu_L, \sigma_L, p$ (Figure 7), my consideration of cross-hybridization potential introduced two more parameters, w_N and w_L, which were optimized in the learning process to produce the best-fit model. We found many cases where cross-hybridizing oligonucleotides caused decreases in the hybridization values and were therefore misleading, as they were supposed to indicate the presence of a nucleosome (Figure 11).

3.3 Hidden Markov Model

The HMM learned the following seven parameters each time it fitted to the data: μ_N (mean of nucleosomal output), μ_L (mean of linker output), σ_N (standard deviation of nucleosomal output), σ_L (standard deviation of linker output), w_N (contribution of cross-hybridization to nucleosomal output), w_L (contribution of cross-hybridization to linker output), and p (probability of remaining in the linker state). Using the HMM to learn a single model for entire regions of chromosome III proved to be inadequate because the hybridization values meander unpredictably (Figure 12). This is possibly due to preferential sites of micrococcal nuclease digestion or technical difficulties with the microarray. Thus, in order to allow the parameters of the HMM to vary over different regions of the chromosome, I ran the HMM on the leftmost window of 70 oligonucleotides on chromosome III, learned a set of parameters, and inferred the pos-

terior nucleosome probabilities (the probabilities of observing a nucleosome at a given oligonucleotide). The window was shifted to the right by one oligonucleotide, and the procedure was repeated until the end of chromosome III was reached. Subsequently, the learned parameters and posterior nucleosome probabilities from these windows were averaged. I varied the window size from from 20 to 100 oligonucleotides in increments of 10. In small window sizes (≤ 50), there were insufficient hybridization values from both nucleosome and linker output distributions, and the HMM could not distinguish between them. Large window sizes (≥ 80) defeated the purpose of removing the unpredictable trends. Thus, a window size of 70 oligonucleotides was suitable.

3.4 Biological Relevance

I identified regions on chromosome III where nucleosomes were spaced widely apart (Figure 13) and other regions where nucleosomes were tightly packed (Figure 14), showing both extremes of nucleosome density. I also found many intergenic regions that were devoid of nucleosomes and observed that gene-coding regions typically have higher nucleosome density than intergenic regions. Histograms of nucleosome densities on gene-coding and intergenic regions (Figure 15) show distributions that were nearly separable by a density threshold between 0.4 and 0.5 nucleosomes per 140 base pairs of DNA. Therefore, regions with higher nucleosome density were almost certain to be open reading frames, while regions with lower nucleosome density were almost certain to be intergenic regions.

4 Discussion

While searching for a correlation between cross-hybridization potential and hybridization values, I experimented with simple linear models, such as taking a count of num-

ber of BLAST matches between an oligonucleotide and the yeast genome, and more realistic models that incorporated the free-energy gain of duplex formation. Remarkably, I found that if I substituted a temperature that was a thousand times greater than normal in the free-energy model, the correlation was much more obvious. In fact, with such an exaggerated temperature, the free-energy contributions basically became a count of BLAST matches, since the amount of cross-hybridization can be approximated by Equation 3. An exaggerated temperature caused the argument of the exponential to approach zero, making each summand close to one, such that

$$\text{XHYB} = \sum_{i=1}^{k} \exp\left(-\frac{\Delta G - \Delta G_i'}{RT}\right) \approx \sum_{i=1}^{k} \exp(0) = \sum_{i=1}^{k} 1 = k$$

which is the number of BLAST matches above a certain threshold of significance. This surprising result was somewhat reflected in [37], where their best fit correlation between free-energy and cross-hybridization potential yielded a temperature that was seven times normal.

Initially, I had also set out to incorporate both GC content and cross-hybridization potential into my HMM. The lack of correlation between the former and trends in the "artificial nucleosome" data was disappointing, but it also simplified the model. The relationship between GC content and hybridization values may have been too complex to model from our data, for I was looking for simple, mostly linear relationships between these oligonucleotide-specific characteristics and the data.

Instead of using a single HMM to model entire chromosomes, I found that training the HMM on running windows yielded more accurate results. The data was affected by large-scale trends, evident by the vertical meandering of the hybridization values for each experiment (Figure 12), implying that the parameters of my HMM should vary over chromosome position. While a model can theoretically be formulated by specifying additional parameters to address the variation, this kind of variable HMM

was not supported by the Bayes Net Toolbox. Thus, to approximate the changes in parameters over chromosomal position, I trained my HMM on running windows as described in Section 3.3. I tested different window sizes to optimize the resolution of the running-windows HMM and found that a window size of 70 oligonucleotides showed the most reliable identification of nucleosome positions (Figure 12). In addition, the posterior probability of a nucleosome's presence on a given oligonucleotide was obtained by averaging the posterior probabilities for all HMM windows that contained the oligonucleotide, and this average was used as the final determination of nucleosome positioning.

Using the improved HMM, I examined nucleosome density on a chromosomal scale. First, I found that some genomic regions contained few nucleosomes (Figure 13) while others were highly packed with nucleosomes (Figure 14). The gene with relatively few nucleosomes may not be strongly regulated by nucleosome positioning. In contrast, the regular phasing of nucleosomes along a stretch of DNA may serve to sterically inhibit transcription. Figure 14 shows a gene-coding region that was almost completely protected by nucleosomes. This gene is likely to be silenced during logarithmic growth. Such a dense configuration of nucleosomes may also pack the DNA into regular chromatin structure. I also found many nucleosome-free intergenic regions (Figure 10) and verified the observation that gene-coding regions are more populated with nucleosomes than intergenic regions (Figure 15) [22]. Perhaps the aggregation of nucleosomes on gene-coding regions serves to better regulate gene expression, by sterically controlling the progression of transcription machinery.

Lastly, I hoped to correlate nucleosome density on promoter regions to the corresponding gene's transcription level. However, preliminary results have shown little correlation (Figure 16). Nevertheless, because nucleosomes both induce and repress genes [6, 7, 8, 9], their presence may be required for both highly- and moderately-expressed genes.

5 Future Directions

Our combined microarray-based method and HMM will be able to address questions regarding what fraction of the genome is found in strongly-positioned nucleosomes and what fraction lies in nucleosomes that are weakly-positioned in our yeast population. We expect that there will be some regions corresponding to each extreme. When a nucleosome is heterogeneously positioned, the next question that must be asked is the nature of the transitions between the distinct states. In other words, do delocalized nucleosomes shift back and forth on DNA on the time scale of seconds? Or do multiple distinct yeast states (e.g. prion states, subtelomeric silencing states, or cell cycle phases, each of which defines a stable set of nucleosome positions) exist simultaneously in our experimental population? We expect that there may be examples of the former [38], but examples of the latter probably exist as well.

With further analyses, we will correlate the movement and position of specific nucleosomes with changes in expression of the underlying gene. In particular, we are analyzing the DNA in the nucleosomes to identify binding sites for transcription factors. We are also looking towards comparing nucleosome positions of yeast in different cell cycle phases to identify the role of nucleosomes in controlling phase-specific events.

6 Conclusion

The identification of nucleosome positioning has been a considerable challenge over the years. While the traditional low-throughput methods have long been known, they are too tedious to easily expand beyond the scale of promoter regions of individual genes. My improved HMM circumvents the problem of scale by automatically and systematically inferring nucleosome positions from multiple replicates of genomic nucleosome microarrays.

Through this objective determination of nucleosome positioning, I have answered questions about nucleosome density. There were wide variations in nucleosome density, and there was a significant difference in nucleosome density between gene-coding and intergenic regions. Though preliminary results did not show significant correlation between transcription level and nucleosome density on promoter regions, exploration and quantification in this area is now possible.

7 Acknowledgments

I would like to thank Drs. Steven Altschuler, Lani Wu, and Oliver Rando for guidance on this project. I am grateful to Dr. June Oshiro for comments and feedback on the written report. Drs. George Church, Michael Dion, Eugenio Marco, Avrom Pfeffer, and Guocheng Yuan have helped me with useful discussion and advice. Finally, I would like to thank Chi-Ren Liu, Shu-Yuan Liu, and Yuen-Joyce Liu for their support and encouragement.

References

[1] Kornberg, R. D. and Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**(3), 285–94 (1999).

[2] Olins, A. L. and Olins, D. E. Spheroid chromatin units (v bodies). *Science* **183**(122), 330–2 (1974).

[3] Hansen, J. C. Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annu Rev Biophys Biomol Struct* **31**, 361–92 (2002).

[4] Rydberg, B., Holley, W. R., Mian, I. S., and Chatterjee, A. Chromatin conformation in living cells: support for a zig-zag model of the 30 nm chromatin fiber. *J Mol Biol* **284**(1), 71–84 (1998).

[5] Ptashne, M. and Gann, A. Transcriptional activation by recruitment. *Nature* **386**(6625), 569–77 (1997).

[6] Han, M. and Grunstein, M. Nucleosome loss activates yeast downstream promoters in vivo. *Cell* **55**(6), 1137–45 (1988).

[7] Knezetic, J. A. and Luse, D. S. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell* **45**(1), 95–104 (1986).

[8] Lorch, Y., LaPointe, J. W., and Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**(2), 203–10 (1987).

[9] Lomvardas, S. and Thanos, D. Modifying gene expression programs by altering core promoter chromatin architecture. *Cell* **110**(2), 261–71 (2002).

[10] Teng, Y., Yu, S., and Waters, R. The mapping of nucleosomes and regulatory protein binding sites at the *Saccharomyces cerevisiae MFA2* gene: a high resolution approach. *Nucleic Acids Res* **29**(13), E64–4 (2001).

[11] Alonso, A., Mahmood, R., Li, S., Cheung, F., Yoda, K., and Warburton, P. E. Genomic microarray analysis reveals distinct locations for the CENP-A binding domains in three human chromosome 13q32 neocentromeres. *Hum Mol Genet* **12**(20), 2711–21 (2003).

[12] Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* **99**(13), 8695–700 (2002).

[13] Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S. K., Wang, A., Suka, N., and Grunstein, M. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* **109**(4), 437–46 (2002).

[14] DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338), 680–6 (1997).

[15] Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**(1), 41–6 (1999).

[16] Wu, L. and Winston, F. Evidence that Snf-Swi controls chromatin structure over both the TATA and UAS regions of the *SUC2* promoter in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **25**(21), 4230–4 (1997).

[17] Gavin, I. M. and Simpson, R. T. Interplay of yeast global transcriptional regulators Ssn6p-Tup1p and Swi-Snf and their effect on chromatin structure. *Embo J* **16**(20), 6263–71 (1997).

[18] Almer, A. and Horz, W. Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the *PHO5/PHO3* locus in yeast. *Embo J* **5**(10), 2681–7 (1986).

[19] Almer, A., Rudolph, H., Hinnen, A., and Horz, W. Removal of positioned nucleosomes from the yeast *PHO5* promoter upon *PHO5* induction releases additional upstream activating DNA elements. *Embo J* **5**(10), 2689–96 (1986).

[20] Gregory, P. D., Barbaric, S., and Horz, W. Analyzing chromatin structure and transcription factor binding in yeast. *Methods* **15**(4), 295–302 (1998).

[21] Terrell, A. R., Wongwisansri, S., Pilon, J. L., and Laybourn, P. J. Reconstitution of nucleosome positioning, remodeling, histone acetylation, and transcriptional activation on the *PHO5* promoter. *J Biol Chem* **277**(34), 31038–47 (2002).

[22] Nagy, P. L., Cleary, M. L., Brown, P. O., and Lieb, J. D. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc Natl Acad Sci U S A* **100**(11), 6364–9 (2003).

[23] Gregory, P. D. and Horz, W. Mapping chromatin structure in yeast. *Methods Enzymol* **304**, 365–76 (1999).

[24] Rouillard, J. M., Herbert, C. J., and Zuker, M. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18**(3), 486–7 (2002).

[25] Tolstrup, N., Nielsen, P. S., Kolberg, J. G., Frankel, A. M., Vissing, H., and Kauppinen, S. OligoDesign: Optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling. *Nucleic Acids Res* **31**(13), 3758–62 (2003).

[26] Allawi, H. T. and SantaLucia, J., J. Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36**(34), 10581–94 (1997).

[27] Allawi, H. T. and SantaLucia, J., J. Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* **37**(8), 2170–9 (1998).

[28] Allawi, H. T. and SantaLucia, J., J. Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* **37**(26), 9435–44 (1998).

[29] Allawi, H. T. and SantaLucia, J., J. Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res* **26**(11), 2694–701 (1998).

[30] Peyret, N., Seneviratne, P. A., Allawi, H. T., and SantaLucia, J., J. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* **38**(12), 3468–77 (1999).

[31] SantaLucia, J., J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**(4), 1460–5 (1998).

[32] Rabiner, L. R. and Juang, B. H. *Fundamentals of speech recognition*. Prentice Hall signal processing series. PTR Prentice Hall, Englewood Cliffs, N.J., (1993).

[33] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis: Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, United Kingdom, (1998).

[34] Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**(2), 544–8 (1998).

[35] Baldi, P., Brunak, S., Chauvin, Y., and Krogh, A. Naturally occurring nucleosome positioning signals in human exons and introns. *J Mol Biol* **263**(4), 503–10 (1996).

[36] Pfeffer, A. J. CS181 lectures 19–20 — hidden Markov models, (2003).

[37] Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**(4), 342–7 (2001).

[38] Pazin, M. J., Bhargava, P., Geiduschek, E. P., and Kadonaga, J. T. Nucleosome mobility and the maintenance of nucleosome positioning. *Science* **276**(5313), 809–12 (1997).

A Figures

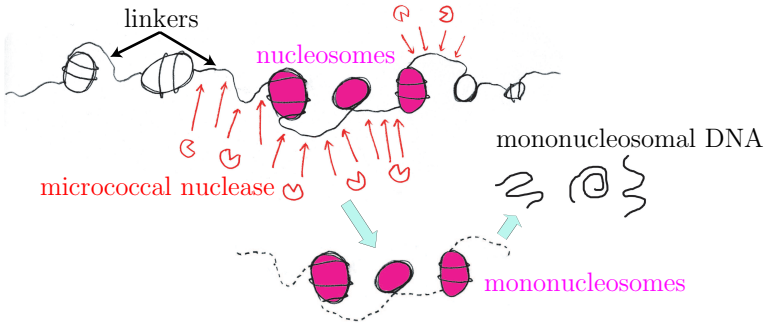

Figure 1: Preparation of Mononucleosomal DNA Formaldehyde was used to cross-link DNA to nucleosomes [23]. Linkers were destroyed by micrococcal nuclease digestion. DNA was dissociated from mononucleosomes and isolated through gel electrophoresis [23]. Mononucleosomal DNA was hybridized to a tiled microarray in our high-throughput method.

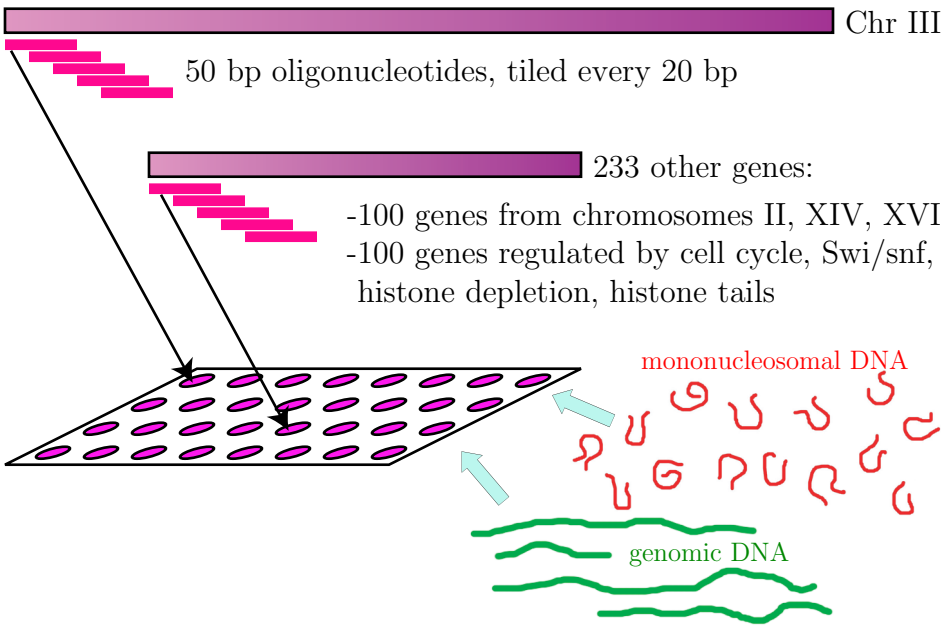

Figure 2: Nucleosome Microarray Each spot on our nucleosome microarray corresponded to a 50-mer oligonucleotide. The oligonucleotides were tiled with an offset of 20 base pairs on chromosome III in its entirety and the promoters of 233 other genes. Cy5-tagged mononucleosomal DNA and Cy3-tagged genomic DNA were hybridized to the oligonucleotides. The hybridization value, or log ratio, for each oligonucleotide was calculated as the base-two logarithm of the Cy5 channel divided by the Cy3 channel.

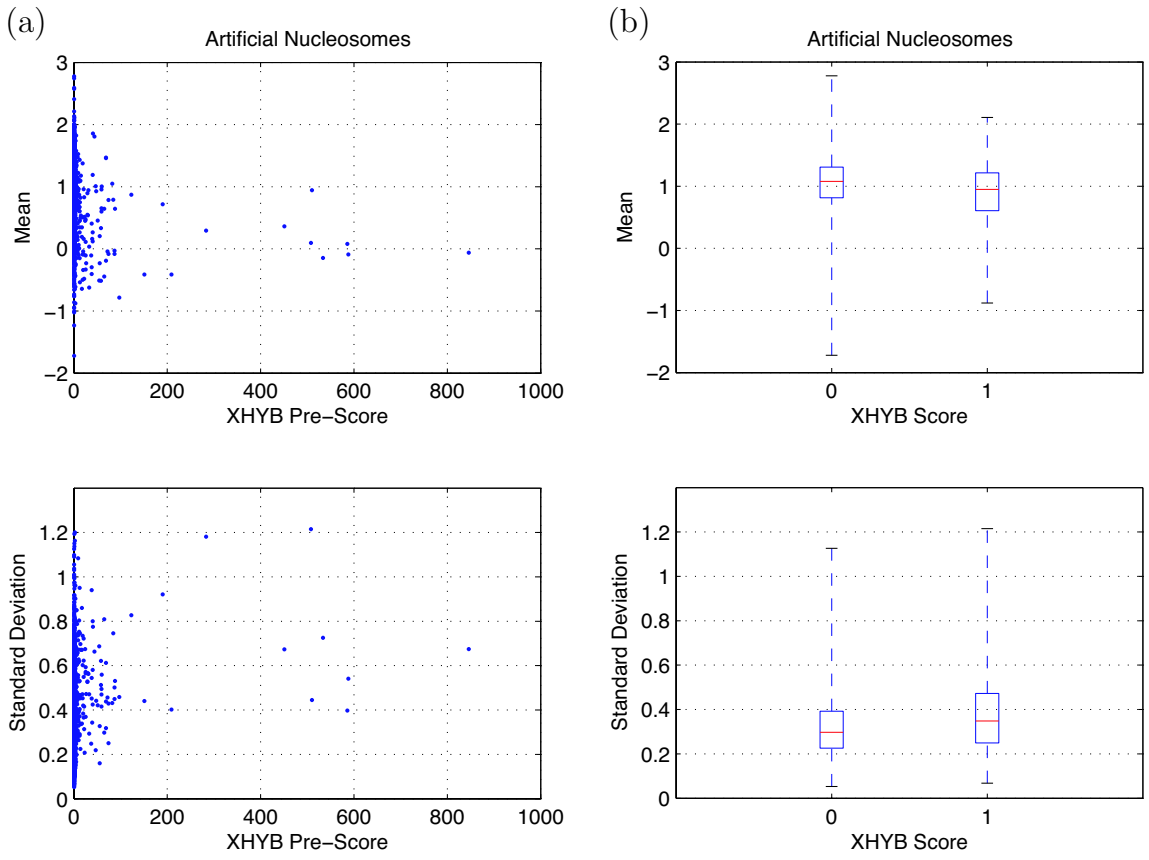

Figure 3: Means and Standard Deviations of Hybridization Values Across Replicates Each point represents an oligonucleotide that was hybridized by an "artificial nucleosome" on our "even" experiments. (a) For each oligonucleotide, the XHYB pre-score, an initial estimation of cross-hybridization potential, counted the number of cross-hybridizing BLAST matches with at least $n = 32$ matching base pairs and the mean and standard deviation were calculated on the hybridization values across eight replicates. (b) Because the majority of oligonucleotides did not have any cross-hybridizing BLAST matches with at least $n = 32$ matching base pairs, they were partitioned into two categories. Those which had no cross-hybridizing BLAST matches with at least n matching base pairs were assigned an XHYB score, or cross-hybridization potential, of 0, and those which had at least one such match were assigned an XHYB score of 1. The Wilcoxon rank-sum test was used as a quantitative test to differentiate the means of the oligonucleotides with an XHYB score of 0 from those with an XHYB score of 1.

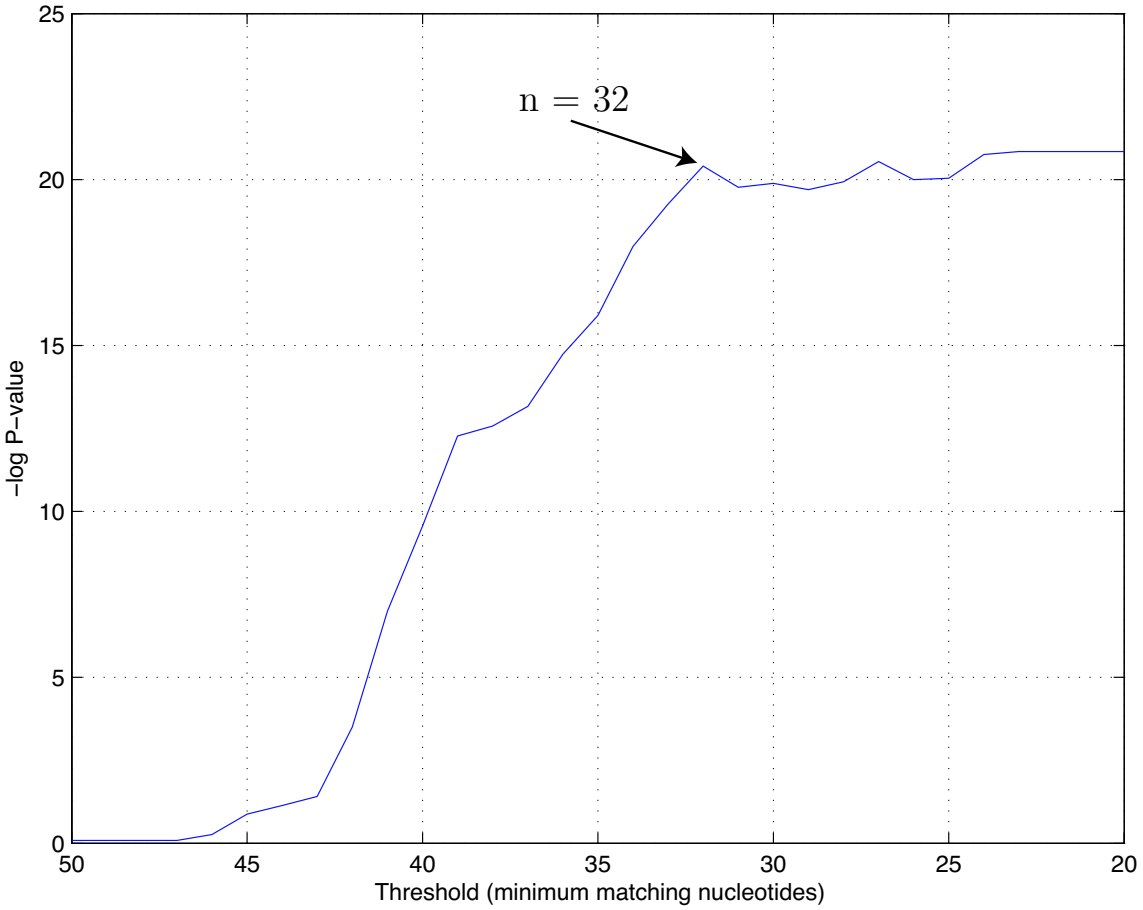

Figure 4: Estimation of Cross-Hybridization The means of the hybridization values of "artificial nucleosomes" were calculated to obtain a suitable threshold for the number of matching nucleotides in the BLAST matches of an oligonucleotide. I tested the hypothesis that the means of oligonucleotides with an XHYB score of 0 and the means of oligonucleotides with an XHYB score of 1 came from different distributions. Shown here are the negative base-10 logarithms of the p-values produced by the Wilcoxon rank-sum test (y-axis) for thresholds $20 \leq n \leq 50$ (x-axis). I chose the local optimum at $n = 32$ that minimized the p-value and maximized the confidence level in the hypothesis.

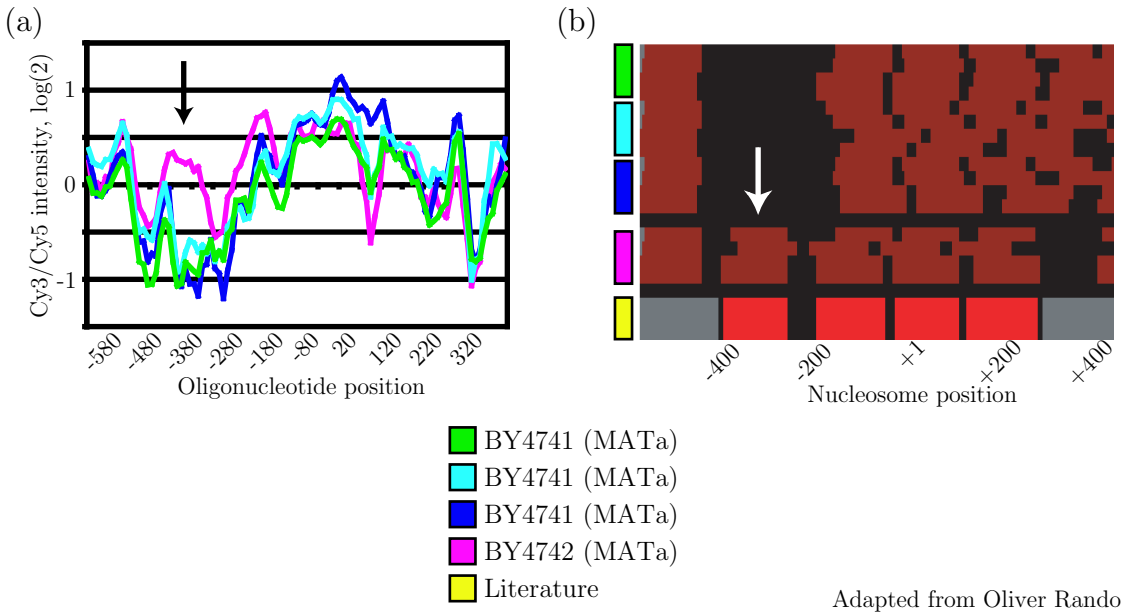

Figure 5: Nucleosome Positions on the *MFA2* Promoter A preliminary version of our nucleosome microarray was used to locate nucleosomes on the *MFA2* promoter in DNA isolated from BY4741 (*MATa*) and BY4742 (*MATα*) strains. The coordinates on the x-axes indicate positions relative to the *MFA2* transcriptional start codon. Three replicate data sets using strain BY4741 are included to show the low level of variation among multiple experiments. (a) The base-2 log ratios of Cy5-labeled mononucleosomal DNA (from haploid strain BY4741 or BY4742 as indicated) to Cy3-labeled yeast genomic DNA (from diploid strain BY4743) are plotted along nucleotide position. The arrow indicates the position of a nucleosome that is present only in BY4742 strains. (b) The HMM-inferred nucleosome positions are represented as dark red bars for nucleosomes and black areas for linker regions. Data from *MATα* nucleosome positions, published in [10], are depicted in bright red. (Gray areas indicate the unavailability of data.) The white arrow indicates the position of a nucleosome that is present only in BY4742 strains.

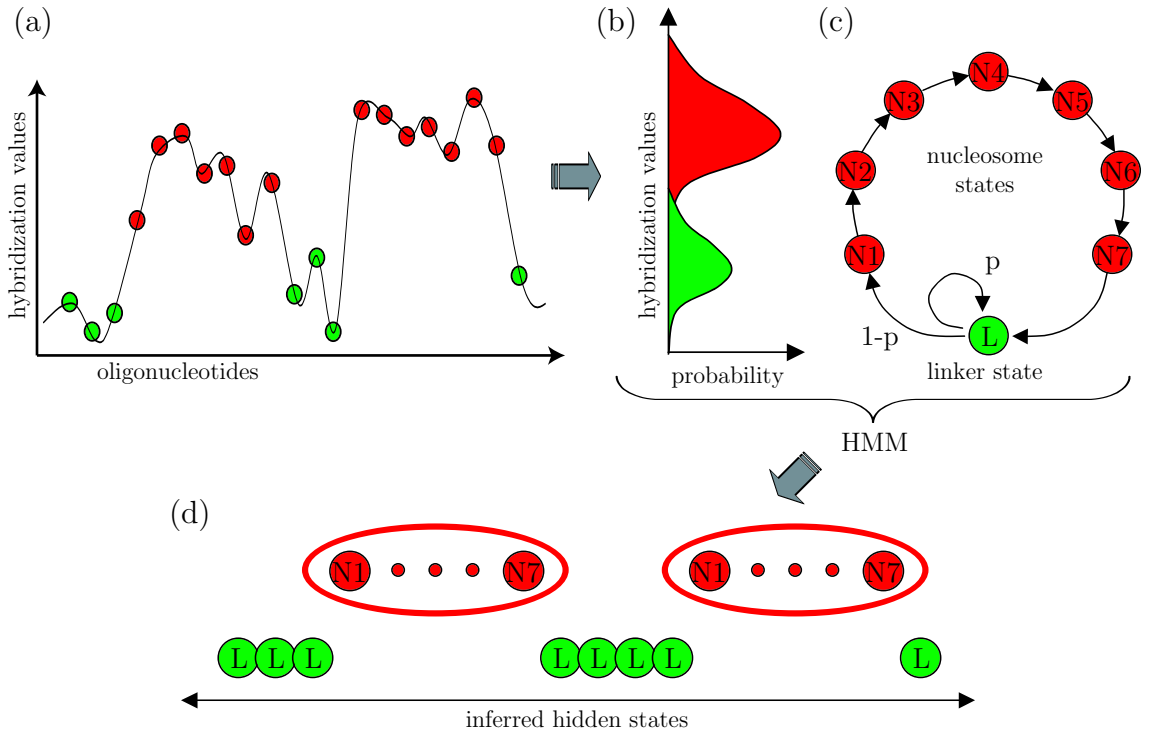

Figure 6: Hidden Markov Model A Hidden Markov Model (HMM) is used to infer nucleosome positions from hybridization data. The HMM models nucleosomes and linkers as two distinct probability distributions from which observed hybridization values are drawn. (a) The hybridization values are plotted along the y-axis, while the x-axis shows consecutive oligonucleotides mapped to specific locations on the genome. For illustrative purposes, hybridization values are colored red and green to indicate that they came from nucleosomes and linkers, respectively. (b) The observed probability distributions are modeled as Gaussian. The y-axis depicts log ratios while the x-axis represents probability density. (c) The transitions between all the states are deterministic, except from the linker state, where p is the probability of remaining in that state and $1 - p$ is the probability of entering a nucleosome. A nucleosome is represented by a sequence of seven nucleosomal hidden states because it takes approximately seven consecutive oligonucleotides, each offset from the previous by twenty nucleotides, to cover the ∼146 base pairs wrapped around the nucleosome [1]. (d) The Viterbi algorithm infers the most likely sequence of hidden states that could have generated the observed hybridization signals. Each occurrence of the sequence, $\mathrm{N1}, \ldots, \mathrm{N7}$, represents a nucleosome.

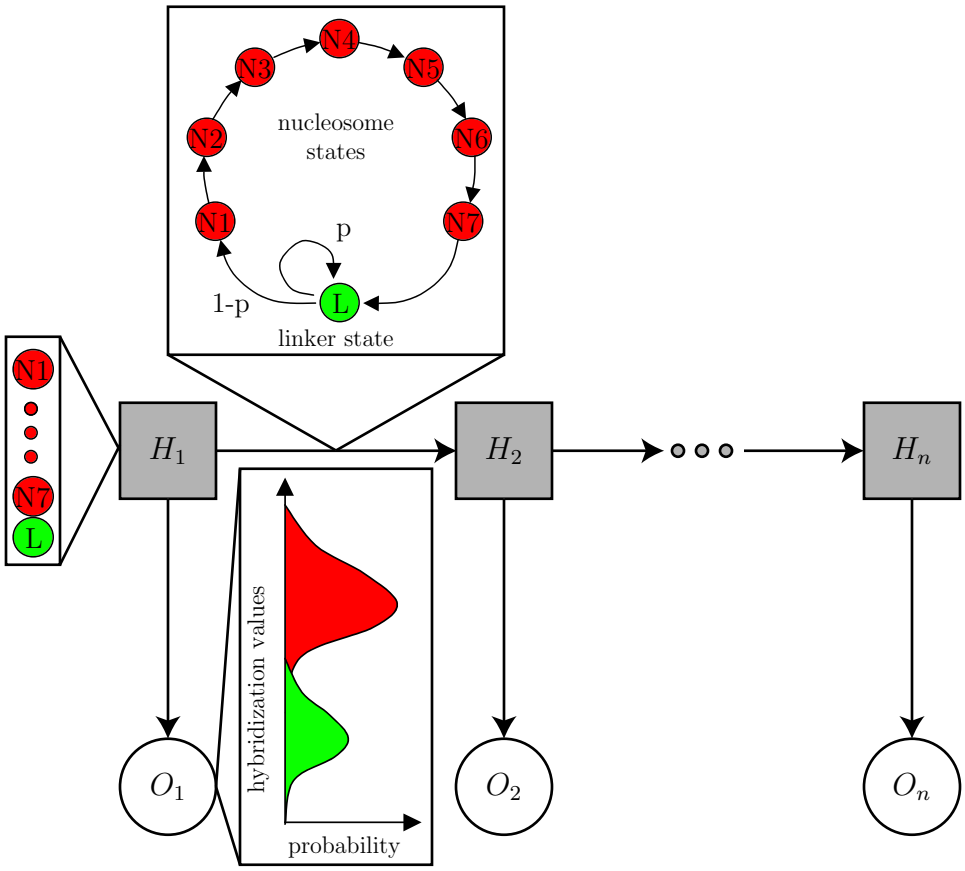

Figure 7: Original HMM Topology Circles denote continuous nodes, squares denote discrete nodes, clear means observed, and shaded means hidden. Each hidden node H assumes a hidden state that represents the type of DNA at the corresponding oligonucleotide, either nucleosome or linker. Each observed node O represents the normalized hybridization value and is modeled as a Gaussian probability distribution with mean and standard deviation of μ_N and σ_N for nucleosomes and μ_L and σ_L for linkers. The hidden states make transitions from one slice to the next according to the transition graph. The arrows represent causal relationships in which the probability distribution of a node is independent of all other nodes given the values of its parent nodes, i.e. the nodes with arrows pointing directly to it.

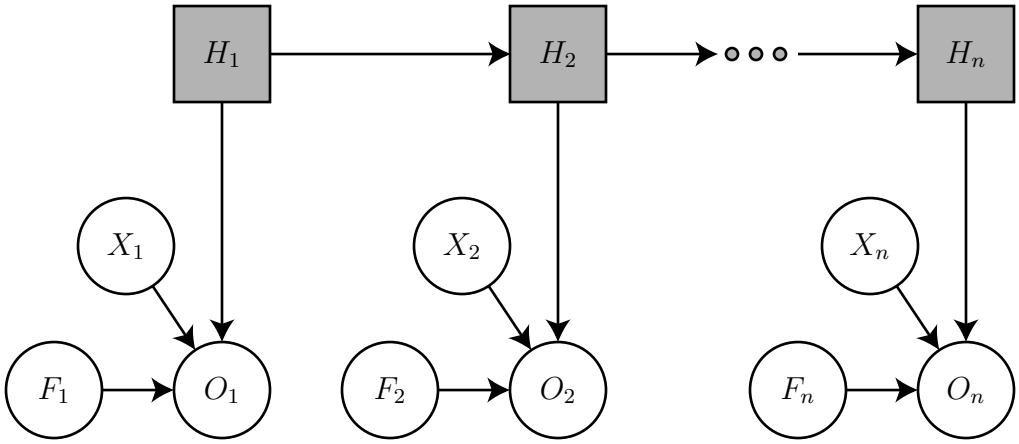

Figure 8: Improved HMM Topology Circles denote continuous nodes, squares denote discrete nodes, clear means observed, and shaded means hidden. This topology incorporates information about potential cross-hybridization and faulty microarray spots. Each hidden node H assumes a hidden state that represents the type of DNA at the corresponding oligonucleotide, either nucleosome or linker. Each observed node O represents the normalized hybridization value and is modeled as a Gaussian probability distribution with mean and standard deviation of μ_N and σ_N for nucleosomes and μ_L and σ_L for linkers. Each cross-hybridization node X is a Boolean value indicating whether the corresponding oligonucleotide is likely to cross-hybridize to alternate regions on the genome. If X is 1, then a weight, w_N for nucleosomes and w_L for linkers, is added to the mean of the log ratio distribution. Each flag node F is also a Boolean value indicating whether a particular observation has been flagged as an unusable measurement.

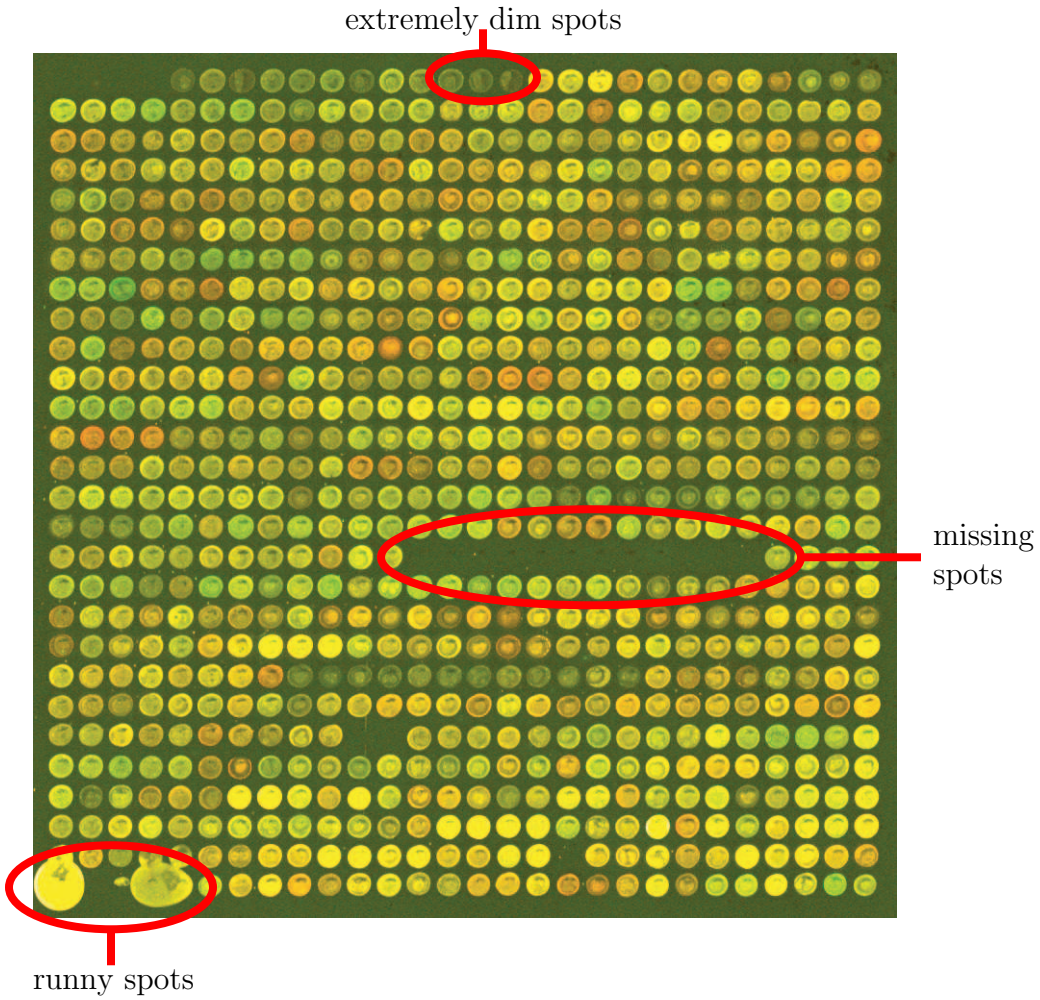

Figure 9: Flagged Spots on the Microarray This is a block from the nucleosome microarray. As indicated, some spots on our nucleosome microarray were not hybridized cleanly, and the corresponding oligonucleotides were flagged and ignored in the improved HMM, using the F nodes.

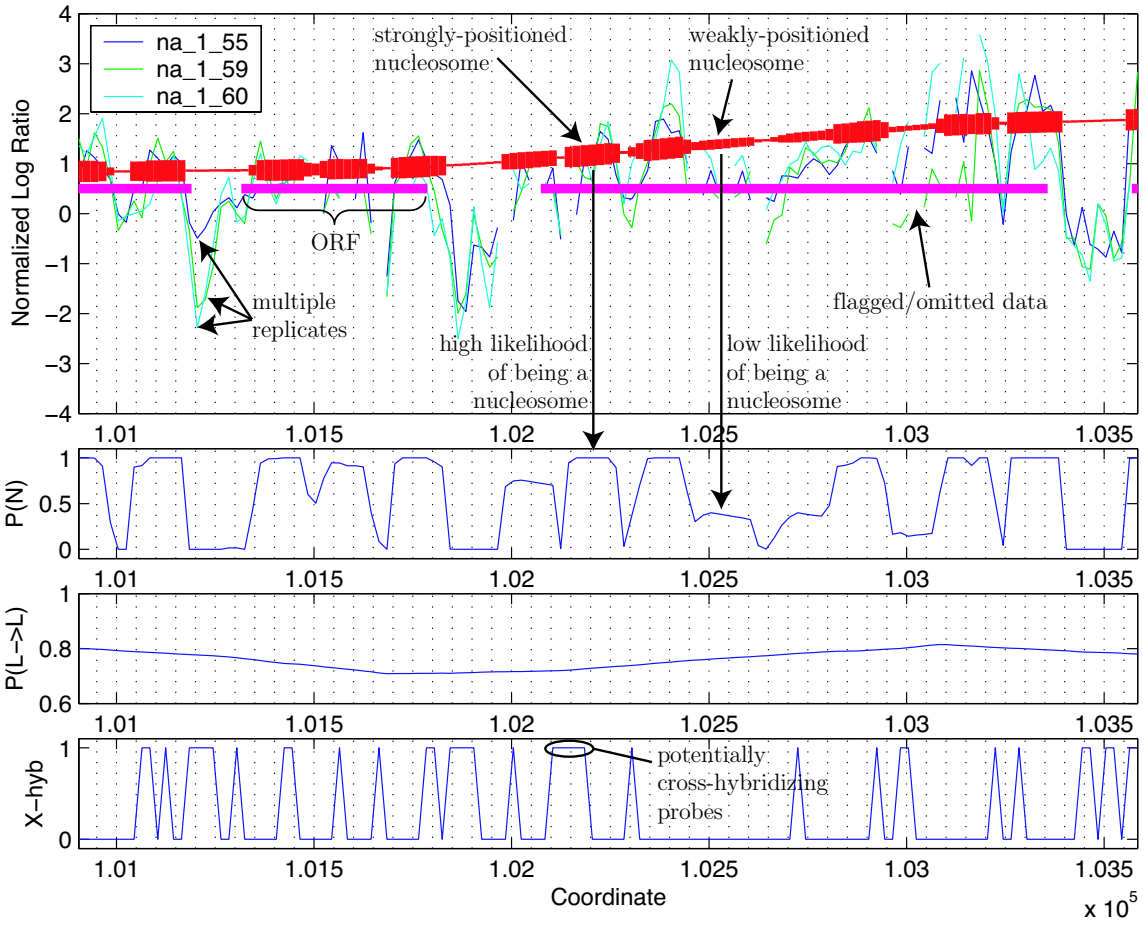

Figure 10: Graphical User Interface This figure shows the nucleosome positions inferred by the improved HMM on an internal region on chromosome III in mid-logarithmic phase *MATα* yeast cells. *Top panel*, the blue, green, and cyan lines represent data from three replicate experiments. The vertical thickness of the red line varies with the likelihood of a nucleosome's presence at each coordinate, and the vertical position of the red line varies with μ_N. Discontinuities in the data result from the omission of flagged oligonucleotides from the HMM's consideration. The magenta bar indicates the presence of gene-coding regions. *Second panel*, the likelihood of a nucleosome's presence is depicted. *Third panel*, the probability p of remaining in the linker state is depicted. *Bottom panel*, the Boolean value of the cross-hybridization node, X, is displayed. The x-axes show genomic coordinates in 10^5 base pairs.

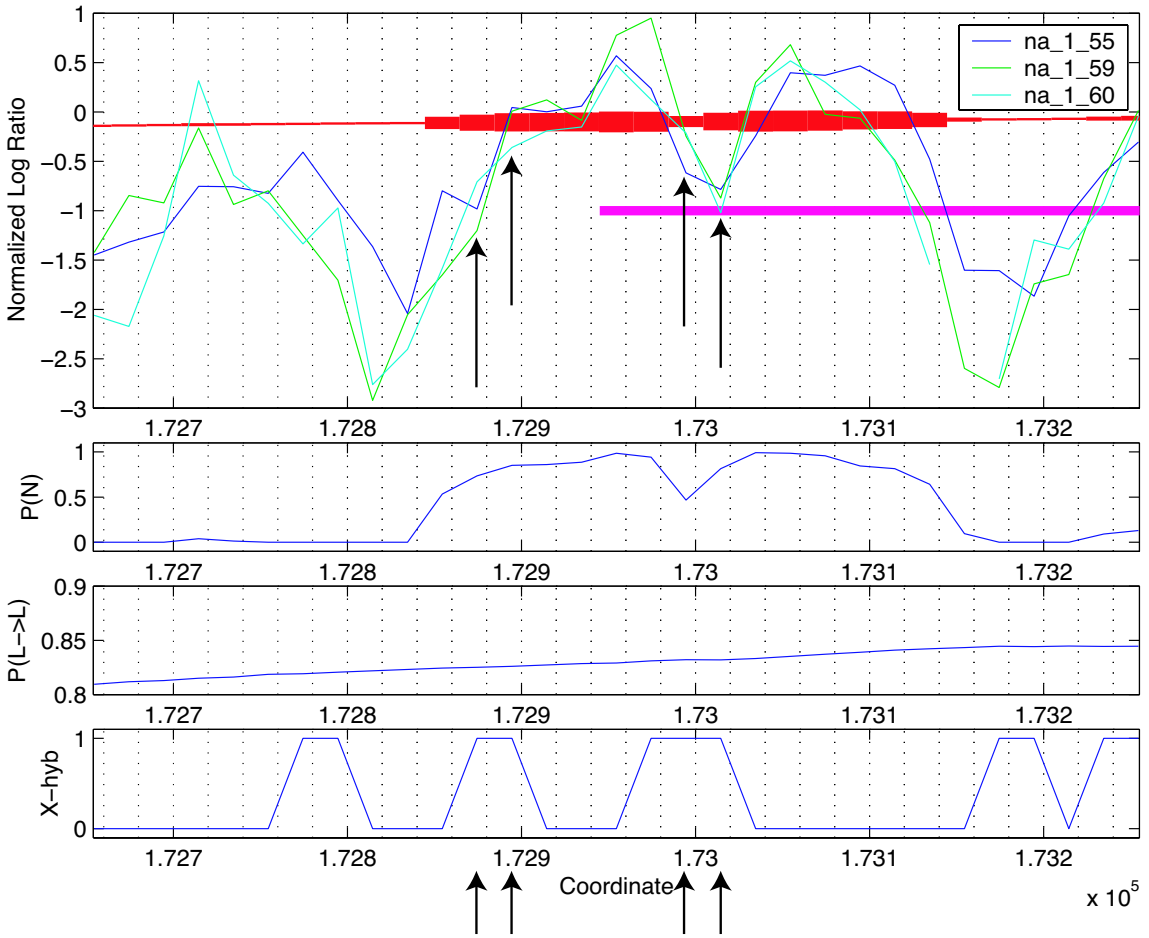

Figure 11: Inclusion of Cross-Hybridization Data Axes and plots are as described in Figure 10. Because we introduce genomic DNA into the genomic channel of the nucleosome microarrays, cross-hybridizing oligonucleotides are expected to cause decreases in the log ratio measurement. The inclusion of the cross-hybridization nodes X allows the improved HMM to discern a nucleosome even when cross-hybridization confounds the hybridization values, as shown by the arrows in the figure. In this region, cross-hybridization potential decreases the mean log ratio output by about 0.11 in the nucleosome state.

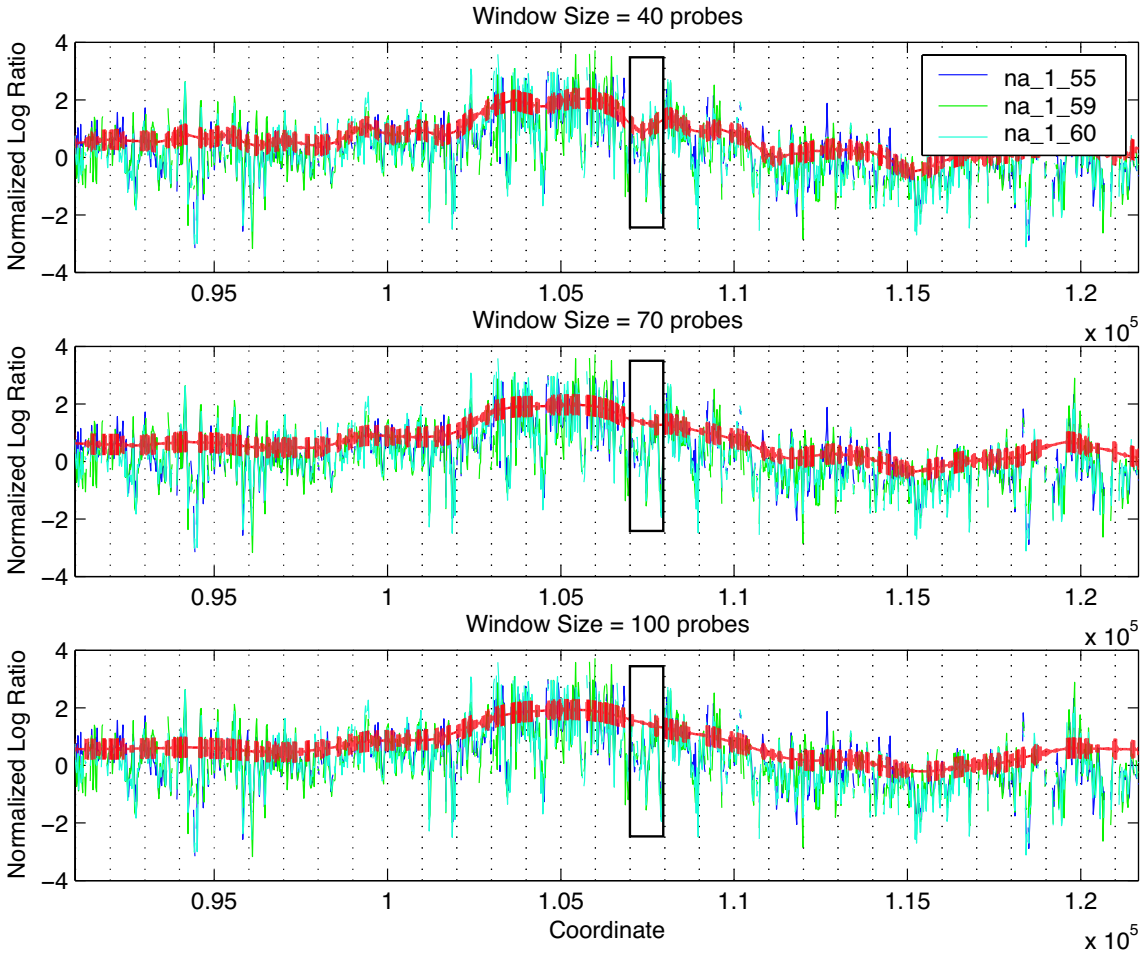

Figure 12: Windowed HMMs Axes and plots are as described in the top panel of Figure 10. There was an observable trend in the data, visible by following the vertical position of the red line, perhaps caused by preferential micrococcal nuclease digestion or technical difficulties with the microarray. The improved HMM was trained on running windows across this region to allow its parameters to vary, shown here with window sizes of 50, 70, and 100 oligonucleotides. Too small a window would deprive the HMM of enough data from both output distributions for it to distinguish them, while too large a window prevents it from following the overall trend. The differences in window sizes is most evident between coordinates 1.07×10^5 and 1.08×10^5, where μ_N seems to almost follow the mean of a linker region in the top panel. After looking at the trend-following capability of my HMM on window sizes from 20 to 100 oligonucleotides in increments of 10, I chose a window size of 70 oligonucleotides.

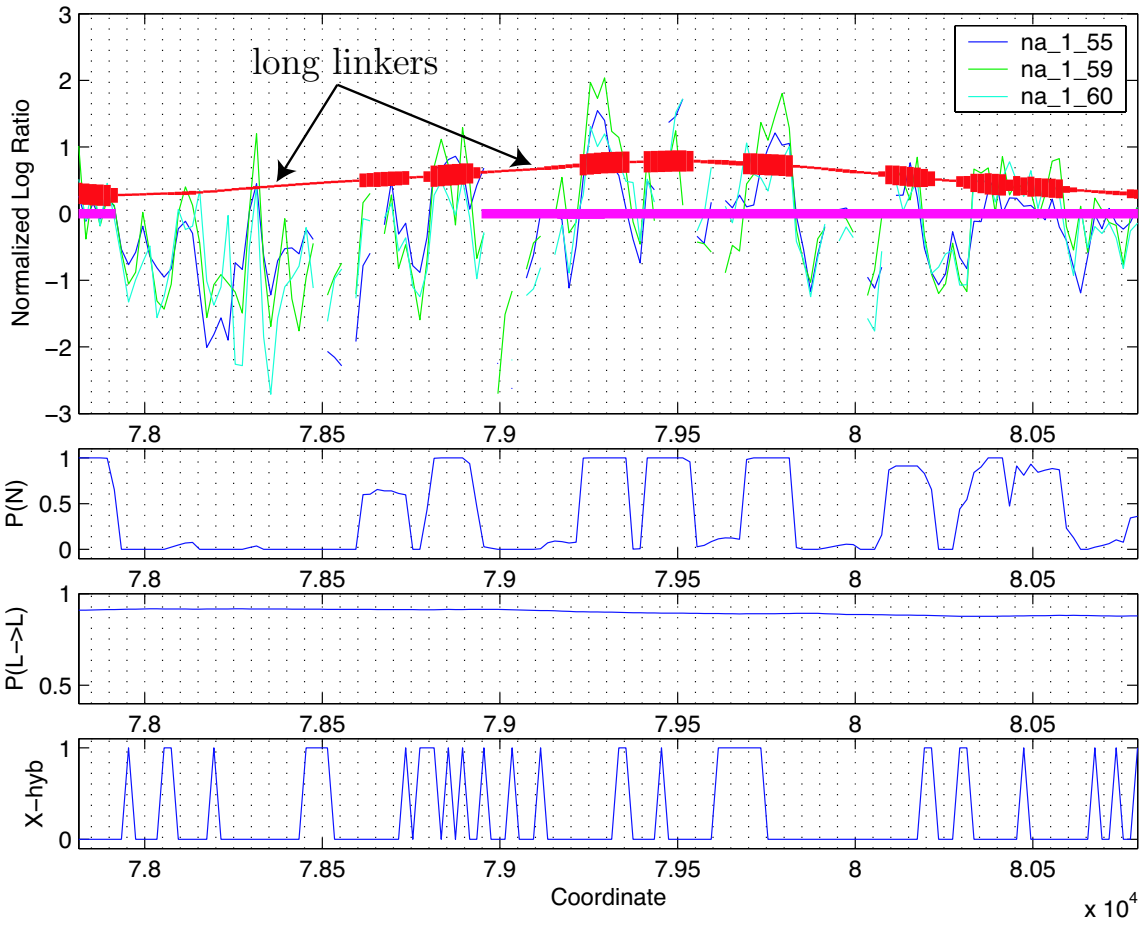

Figure 13: Sparsely-Populated Region Axes and plots are as described in Figure 10. Some regions of genomic DNA are sparsely-populated with nucleosomes and are characterized by long linkers and high, near-1 probabilities that the hidden state remains in the linker state, as shown in the third panel.

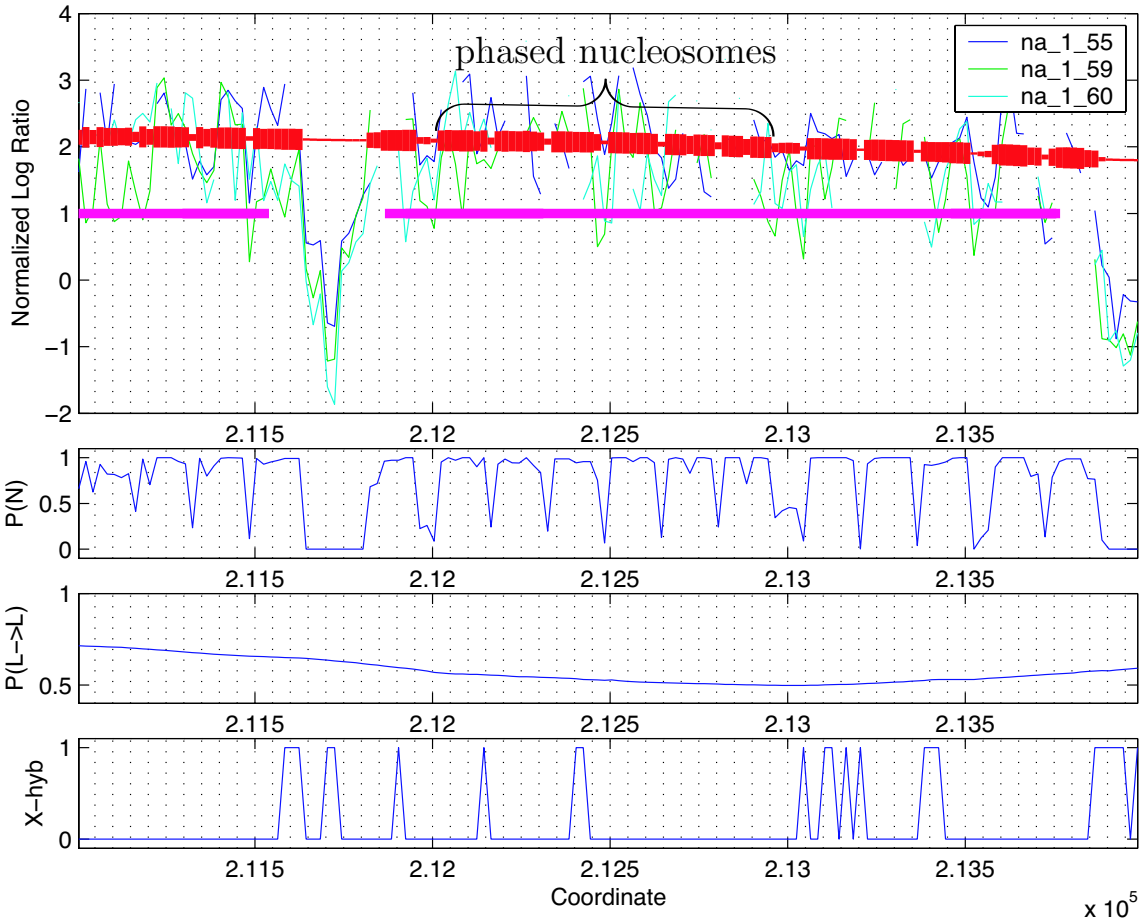

Figure 14: Densely-Populated Region Axes and plots are as described in Figure 10. Other regions of genomic DNA are densely-populated with nucleosomes and are characterized by phased, or tightly-packed, nucleosomes, short linkers, and low probabilities that the hidden state remains in the linker state, as shown in the third panel.

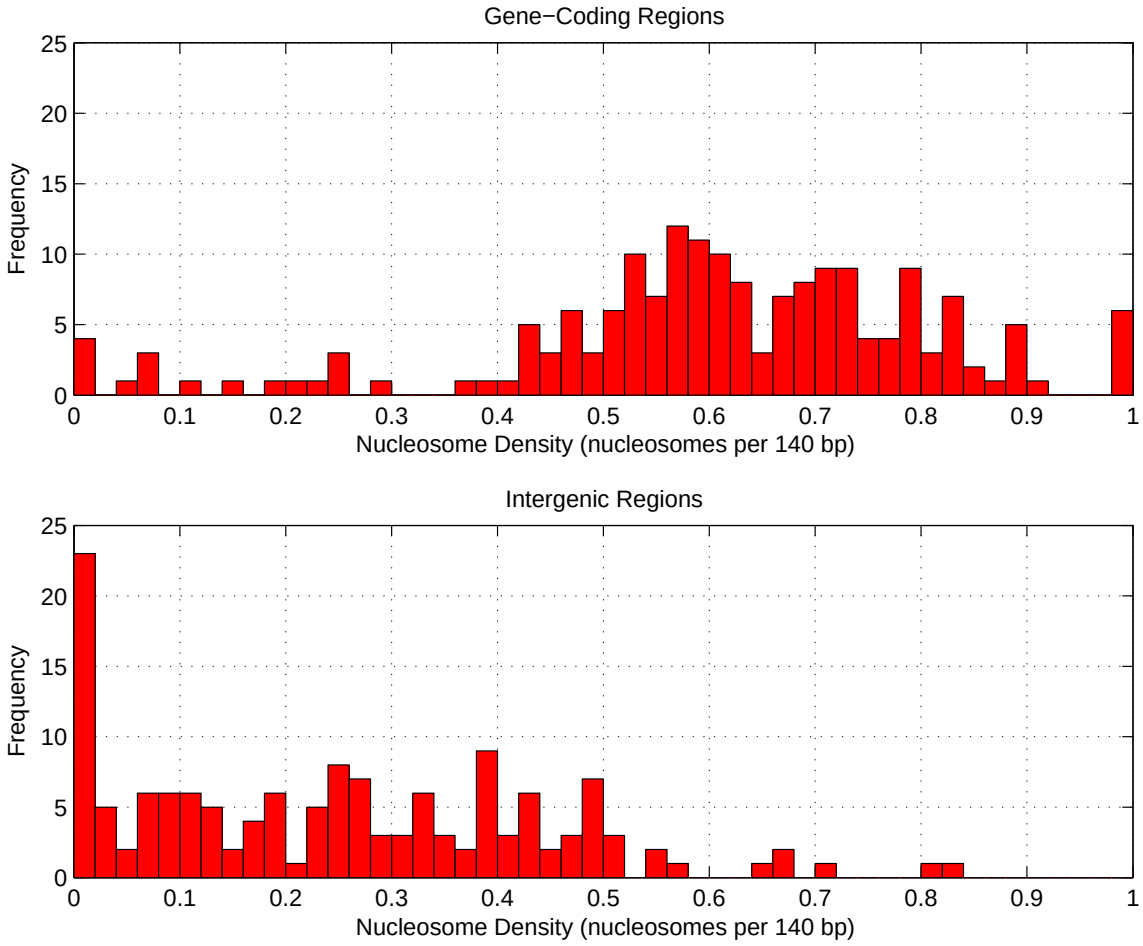

Figure 15: Nucleosome Density on Gene-Coding and Intergenic Regions Gene-coding regions tend to be densely populated by nucleosomes while intergenic regions tend to be sparsely populated, as recently shown by physical fractionation of chromatin, a completely different technique [22].

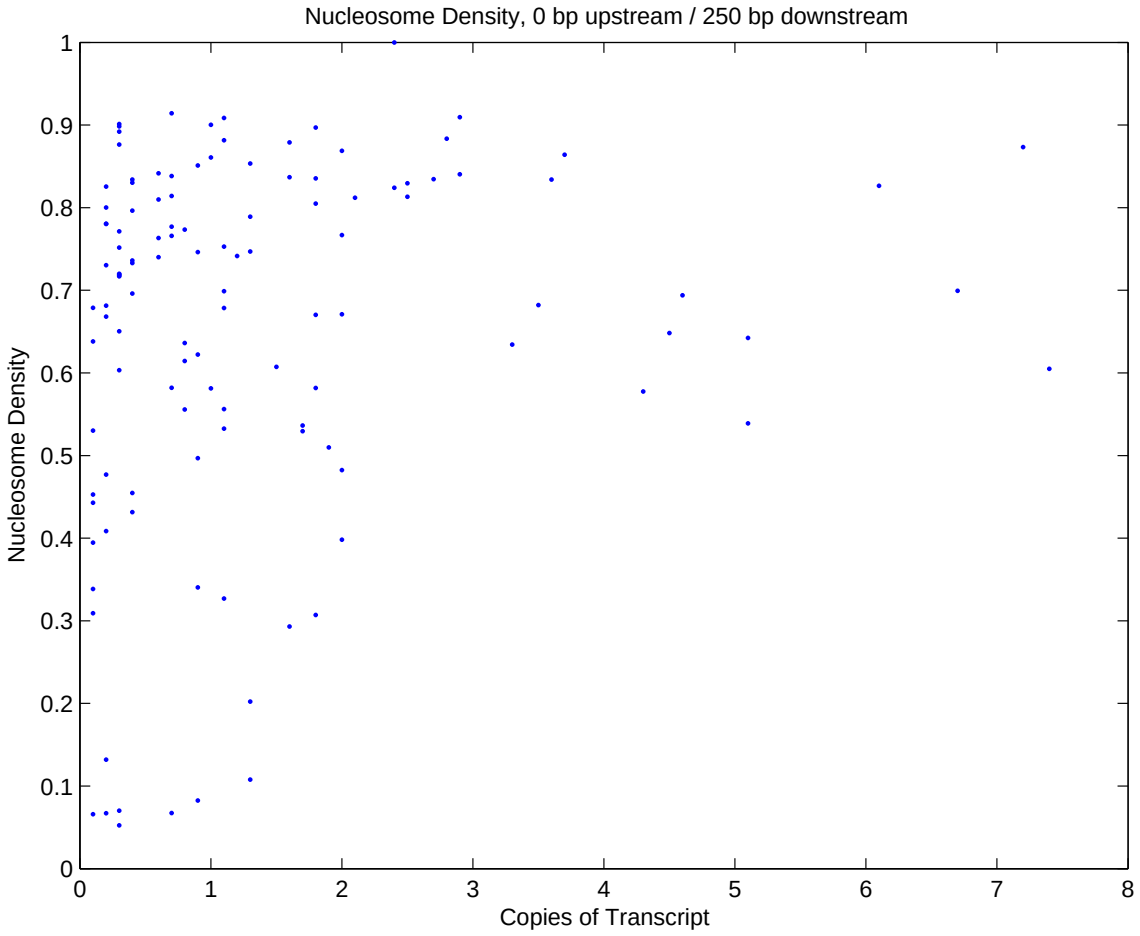

Figure 16: Nucleosome Density vs. Transcription Level Preliminary results do not show a significant correlation between the nucleosome density of a gene-coding region and its transcription level.