Predicting Nucleosome Positioning Using Multiple Evidence Tracks

Sheila M. Reynolds[1], Zhiping Weng[2], Jeff A. Bilmes[1], and William Stafford Noble[1]

[1] University of Washington, Seattle, Washington, USA
[2] Boston University, Boston, Massachusetts, USA

Abstract. We describe a probabilistic model, implemented as a dynamic Bayesian network, that can be used to predict nucleosome positioning along a chromosome based on one or more genomic input tracks containing position-specific information (evidence). Previous models have either made predictions based on primary DNA sequence alone, or have been used to infer nucleosome positions from experimental data. Our framework permits the combination of these two distinct types of information. We show how this flexible framework can be used to make predictions based on either sequence-model scores or experimental data alone, or by using the two in combination to interpret the experimental data and fill in gaps. The model output represents the posterior probability, at each position along the chromosome, that a nucleosome core overlaps that position, given the evidence. This posterior probability is computed by integrating the information contained in the input evidence tracks along the entire input sequence, and fitting the evidence to a simple grammar of alternating nucleosome cores and linkers. In addition to providing a novel mechanism for the prediction of nucleosome positioning from arbitrary heterogeneous data sources, this framework is also applicable to other genomic segmentation tasks in which local scores are available from models or from data that can be interpreted as defining a probability assignment over labels at that position. The ability to combine sequence-based predictions and data from experimental assays is a significant and novel contribution to the ongoing research regarding the primary structure of chromatin and its effects upon gene regulation.

Keywords: Nucleosome prediction, dynamic Bayesian network, chromatin structure.

1 Introduction

DNA in eukaryotes is packaged with histone and other proteins into a chromatin complex. The most basic element of chromatin is the nucleosome "core", which consists of a bundle of eight histone proteins around which is wound approximately 147 base pairs (bp) of double-stranded DNA. Between adjacent cores exists a variable-length stretch of DNA commonly called the "linker" which is

B. Berger (Ed.): RECOMB 2010, LNBI 6044, pp. 441–455, 2010.

generally more accessible to elements such as transcription factors than the compacted DNA in the core. The precise positioning of the nucleosome cores and the inter-nucleosomal linker regions allows for selective access to the DNA by the cellular machinery; understanding the mechanisms that control this positioning is therefore crucial to our understanding of gene regulation and expression.

Numerous computational approaches to inferring nucleosome positions either from experimental data or from the primary DNA sequence have been published in recent years. These methods generally use a hidden Markov model (HMM) or similar framework (*e.g.* Boltzmann chain) in which a sequence of hidden states, representing the nucleosome core and the linker, form a Markov chain, and the observations "emitted" by each state are derived either from DNA-sequence models or experimental assays. Common model assumptions include the requirement that adjacent nucleosomes may not overlap, as well as constraints on the length of a nucleosome and a model of the linker lengths. The model of linker lengths generally specifies a minimum linker length due to steric hindrance between adjacent nucleosomes, and may also define a geometric or other distribution over longer linker lengths [1] or an upper limit on linker length [2]. Although very similar in implementation, models based on DNA-sequence scores and models based on experimental data are solving two different problems. When the inputs to the HMM are sequence-model scores [1,2,3,4,5], the HMM framework predicts the most probable nucleosome positions based on the DNA sequence alone. In contrast, when the inputs originate from experimental data such as tiling microarrays [6,7,8,13], the goal is data analysis and interpretation.

In this work, we exploit the power of dynamic Bayesian networks (DBNs) to create a general framework for predicting nucleosome positions using one or more input tracks of arbitrary position-specific genomic scores. A DBN is a generalization of the widely used HMM [9], and generalized versions of the standard inference algorithms commonly applied to HMMs exist for the broader class of DBNs. The typical HMM falls into the broad class of generative models in that, in addition to being used in the standard way, the model can also be (although rarely is) used to generate instances of evidence sequences according to the model parameters. The model that we present here is more discriminative in nature, and uses the input evidence to directly inform the probabilities at each state in the Markov chain. Furthermore, our model allows multiple evidence tracks to be combined to jointly influence the current state, while the Markov chain simultaneously enforces the sequential grammar that is described by the state transition matrix. Specifically, we show how we can use either sequence-model scores or experimental data independently, or both together, with the result that the sequence scores can be used to fill in gaps in the experimental data and provide a more complete picture of the nucleosome landscape. Alternatively, sequence-model scores can be used in conjunction with transcription factor (TF) binding probabilities, resulting in a competitive model similar to the one described by Wasson and Hartemink [5] with the assumption that a TF can only bind to the DNA between nucleosome cores. The ability to combine sequence-based predictions and data from experimental assays is a significant and novel contribution to the ongoing research regarding the primary structure of chromatin and its effects upon gene regulation.

2 Results

2.1 Predicting Nucleosome Positioning from Arbitrary Sequence-Preference Scores

In recent years, numerous methods have been proposed for scoring a DNA segment for the purposes of distinguishing nucleosome-inhibiting vs nucleosome-forming regions. The DBN presented in this work can integrate the information contained in these types of local sequence scores, regardless of the method used to produce them, to infer nucleosome positioning along a chromosome. In this section we illustrate this application of our method with three specific examples. First we show that we can recapitulate the average occupancy predicted by the Segal model [3,12,14] using the Segal raw binding scores as inputs, and then we show predictions based on our recently developed nucleosome dyad scores using two different linker-length models. Our probabilistic framework permits two types of linker models: a geometric length-distribution which prefers shorter linkers, or a uniform distribution which gives the same probability to all possible linker lengths (see *Methods* for details). These two different linker-length models can be thought of as describing two variations on the statistical positioning idea [10] in regions where sequence-directed positioning is weak.

Our nucleosome dyad score, *dScore*, is based on a discriminative pattern-correlation method [11] which computes a score for the central position of an input sequence of length 301 bp, based on sequence information alone, by weighting and combining information from all k-mers for $k \in \{1, 2, 3\}$. This score is the continuous-valued output of a binary classifier and can be interpreted in a manner similar to a log-ratio. The Segal raw binding score is the log-ratio of two model components: one captures the periodic positioning of dinucleotides along the nucleosome core, while the other encodes the relative linker-region preferences for all 5-mers.

Figure 1 shows the two different sequence-preference scores in the top panel: the Segal raw binding score and our dyad score (dScore), plus a GC-content track for reference (computed using a sliding window of width 71 bp). In the bottom panel, each trace corresponds to the posterior probability that a position is covered by a nucleosome core, inferred by the model from the input local sequence scores. The output based on the Segal raw binding score and using the uniform linker-length model closely recapitulates the average occupancy probability predicted by the full Segal model [12] (Pearson correlation $r = 0.96$). Two separate output traces are shown based on the dScores: the first uses the uniform linker-length model, and the second uses the geometric linker-length model.

There are significant qualitative similarities as well as differences both between the Segal and dScore sequence-scores and the posterior probabilities shown in Figure 1. These differences are due to the differences in the input scores as well as

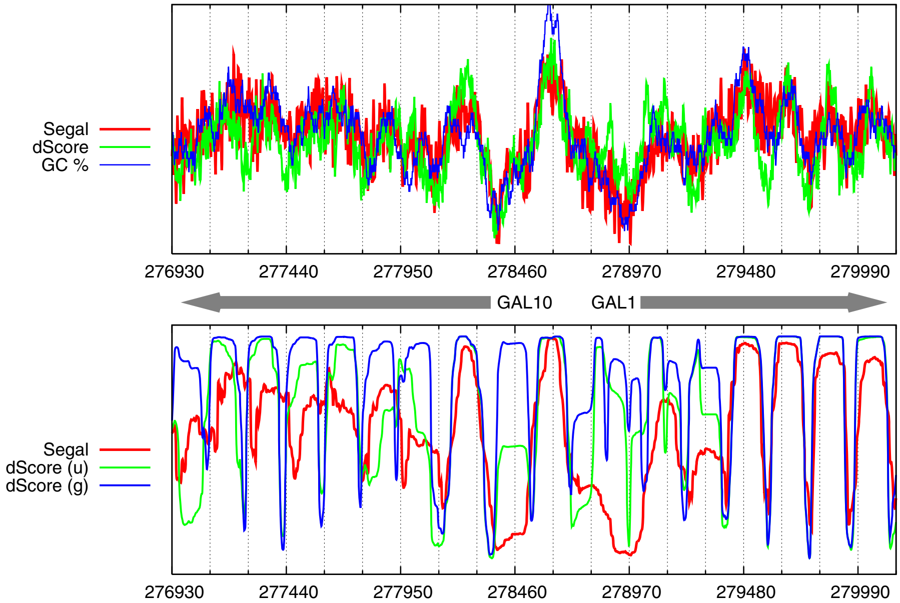


Fig. 1. *S. cerevisiae* chromosome II: raw sequence-model scores and local GC % (bottom) and nucleosome core posterior probabilities (top) for the Segal model and our pattern-classification model with a uniform linker model (green) and a geometric linker model (blue)

to differences in the linker length models. The most striking difference can be seen immediately upstream of the GAL10 transcription start site, in an AT-rich region wide enough for one nucleosome core, where both sequence-models produce low scores. The Segal model predicts a very long nucleosome-free region (NFR), while the two dScore models predict a weakly-positioned nucleosome—the model that prefers shorter linker lengths places a nucleosome with high probability while the uniform linker-length model places one with lower probability.

2.2 Interpretation of Experimental Data Alone or in Conjunction with Sequence Scores

Another application of our model is to interpret experimental data, similar to what has been done previously with microarray data [6,8,13]. By incorporating additional information in the form of sequence-based scores or even just a model of linker lengths, the model can fill in gaps in the experimental data. Experimental data is frequently also expressed as a log-ratio, so the same mapping to probabilities described above can be used here.

Figure 2 shows a region on yeast chromosome II for which there is a gap in one of the *in vivo* experimental data sets from Kaplan *et al.* [12]. The gap is 1340 bp wide and corresponds to the ribosomal protein RPL4A. Using the experimental data as an evidence track, the probabilistic model was run twice—once using the geometric linker model, and once using the uniform linker model (top panel of Figure 2). When the model includes a preference for shorter linker lengths, it places 8 nucleosomes, evenly distributed across the 1340 bp gap in the data. With the uniform linker model, we observe two interesting changes in the predictions: first, they track the input data much more closely because, aside from the grammar, the data is the only source of information; and second, the model is much less certain about how many nucleosomes fill the gap—without the preference for short linkers, the model is considering all possible placements of between one and eight nucleosomes. In both cases the uncertainty grows with the distance from the nearest data, as indicated by the decreasing local maxima and the increasing local minima.

2.3 Evaluation of Predicted Nucleosome Position Accuracy

We have previously created a set of 50,814 estimated nucleosome dyad positions in yeast based on the experimental data of Field *et al.* [14]. The genomic positions of these dyads were estimated by applying a simple peak-detection algorithm to a nucleosome occupancy map, and a confidence score derived from the number

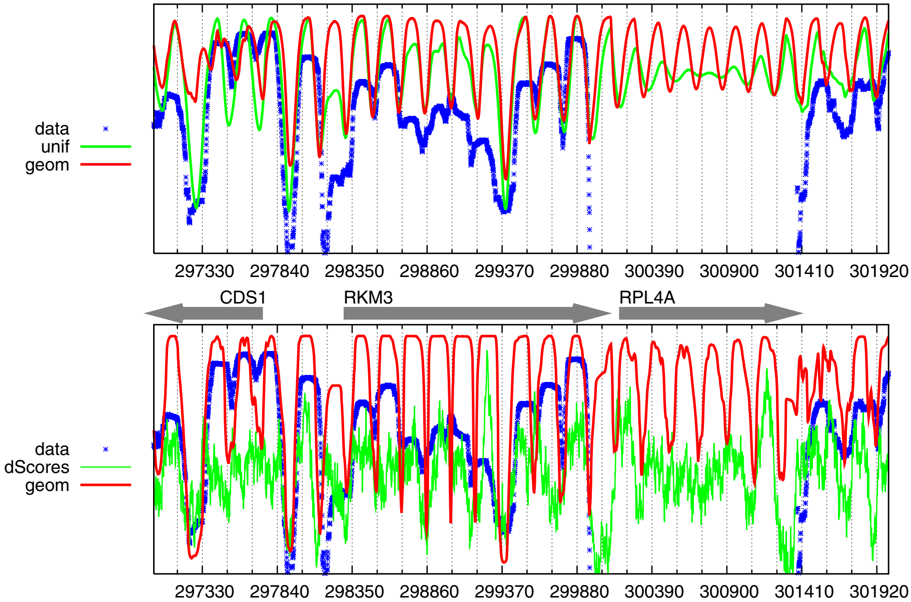


Fig. 2. *S. cerevisiae* chromosome II: Experimental data (blue stars) with a gap spanning the coding sequence for ribosomal protein RPL4A (approximately 300,000-301,400) Top: nucleosome core posterior probabilities inferred from experimental data using geometric linker (red) or uniform linker (green). Bottom: sequence-model scores (green) are added as additional evidence and nucleosome positions recomputed.

of overlapping reads was associated with each dyad [11]. In order to evaluate the positional accuracy of the predictions based on the two different sequence-model scores described above, we compare the predicted dyad positions (local maxima in the posterior probability of being in the dyad state) to the experimental benchmark set and compute the fraction of the positions in the experimental set that are within X nucleotides of a predicted dyad.

Posterior probabilities of nucleosome positions were computed using three different input tracks (one at a time): (a) the experimental Field occupancy map, (b) the dScores, and (c) the Segal raw binding scores. Predicted dyad positions were then compared to the entire benchmark set and to a small subset of the highest scoring positions (Fig 3). Because the estimated positions being used as the benchmark were derived from the same data used in (a), one would expect a near perfect concordance, and in fact the majority of the 50,814 dyads have corresponding predictions within 3 bp. The fact that the predictions based on the experimental dataset do not match up more exactly to the positions estimated using a simple peak-detection approach highlights the strengths of using a sequence model which simultaneously integrates all available information along the entire sequence. For example, if the experimental data indicates a sharply demarcated NFR, the edges of the NFR will affect the positioning of adjacent nucleosomes. These effects are automatically considered by the DBN but not by a simplistic

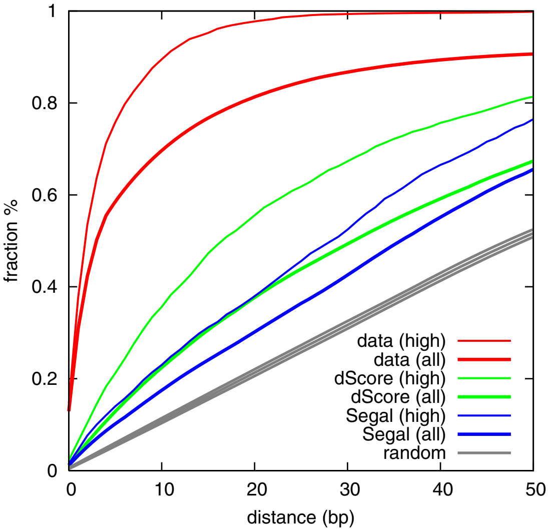


Fig. 3. Dyad positions inferred by the DBN using experimental data (red), dScores (green), or Segal raw binding scores (blue), compared to previously estimated dyad positions. Each pair of curves represents an evaluation over the entire set of 50,814 estimated dyad positions (all) and the top-scoring 3,180 (high). Each curve represents the fraction y of the estimated dyad positions for which a dyad was predicted by the DBN to within x nucleotides. The grey curves represents the performance that would be expected by chance (mean, and mean ± one standard deviation, from simulations).

peak-detection approach. For the purposes of comparing to predictions based on sequence scores, this comparison to predictions based directly on the data provides an upper bound on the performance of any other method.

The dyad positions predicted using either type of sequence-based scores are both much less similar to the benchmark positions, although for both models the high-scoring benchmark dyads are more likely to be predicted accurately. At a maximum distance of 15 nucleotides between a benchmark dyad and a predicted dyad, corresponding to a 90% overlap between the reference nucleosome core and the prediction, the dScore-based predictions match 47% of the high-scoring subset and 31% of the entire set, compared to 31% and 24% respectively for the Segal-based predictions, and the 16% that would be expected by chance. All three sets of predictions contained very similar numbers of predicted dyads (∼62,500), so these accuracy figures are directly comparable.

2.4 Competition with Transcription Factors

Histone proteins do not interact with the DNA to form nucleosomes in isolation, but rather compete dynamically with other DNA binding factors. To illustrate how this notion of competition can be incorporated into our model, we show an example of combining nucleosome-sequence scores with a landscape of transcription factor binding probabilities. We scanned the yeast genome using the

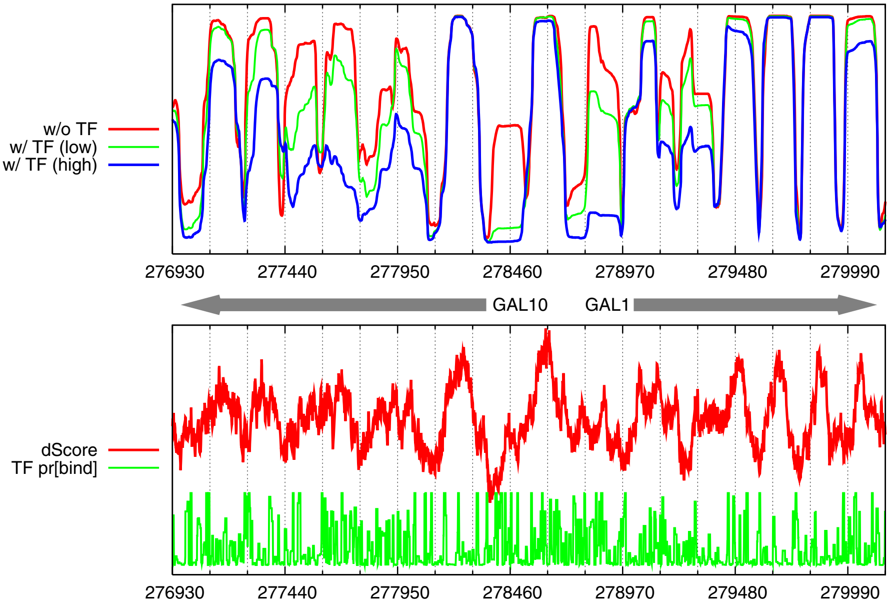


Fig. 4. *S. cerevisiae* chromosome II: competition with transcription factors destabilizes weakly positioned nucleosomes first. Top: nucleosome positions inferred from dScores without (red) and with low (green) and high (blue) levels of TF competition. Bottom: dScores (red) and TF binding probability landscape (green).

112 DNA-binding protein sequence specificities described by Badis *et al.* [15], and created an overall TF landscape by taking the maximum resulting binding probability at each position (see *Methods* for details). This information was then used in parallel with the dScores described earlier, and results are shown in Figure 4. This region of yeast chromosome II has two genes transcribed in opposite directions, with transcription start sites separated by approximately 600 bp. Immediately upstream of each TSS is a region of very high AT-content which includes strong matches for several TFs including SIG1 and PHO2. The figure shows that including the TF binding landscape almost completely eliminates the formerly weakly predicted nucleosome upstream of the GAL10 TSS while not significantly affecting the most strongly predicted nucleosomes.

3 Discussion

We have developed a novel solution to the problem of predicting nucleosome positions along a chromosome by incorporating arbitrary sources of information within a coherent probabilistic framework. Previous approaches have solved only part of this problem, using either sequence information alone or experimental data alone. Using sequence-based evidence in combination with experimental data provides a mechanism for interpreting the experimental data while filling in gaps using sequence predictions. Gaps in experimental data can be a significant problem in organisms with much larger (and more highly repetitive) genomes than yeast, where even genome-wide assays of nucleosome positioning produce relatively sparse data sets [16,17]. Combining multiple input tracks also permits us to investigate the relative impacts of different factors on the nucleosome landscape. Two different sequence-models could even be combined to see whether, jointly, they can make more accurate predictions than either one individually.

While we acknowledge the ongoing debate as to the impact *in vivo* of sequence-directed nucleosome positioning, we believe that predictive models that can incorporate the mechanisms that affect nucleosome positioning will increase our understanding of the chromatin structure and the impact it has on gene regulation and expression. Based on our genome-wide comparison of nucleosome positions estimated from an *in vivo* dataset to those predicted using dScore, we find that roughly 15% more of the nucleosome cores are predicted with at least a 90% overlap than would be expected by chance. The remaining nucleosomes are likely to follow a statistical positioning pattern, which this DBN naturally models. It may be interesting to explicitly compare a nucleosome-occupancy probability computed using purely local information to the probability computed by a full sequential model in order to understand which nucleosomes are predicted to be well-positioned due to a locally strong sequence signal and which might be predicted to be well-positioned as a result of a nearby, strongly-positioned "barrier" [10].

In this study, we opted not to evaluate our methods by computing a correlation between the posteriors produced by our model and an experimentally determined nucleosome occupancy profile [1]. Empirically, such profiles generally exhibit a strong dependence on local GC-content; consequently, a simple sliding window of GC-content yields a pseudo nucleosome positioning signal that correlates at 0.70 with an empirical *in vitro* profile and between 0.56 and 0.63 for three *in vivo* sets from [12]. Although the inherent GC-richness of the nucleosome cores and AT-richness of the linkers will naturally produce this type of correlation, our concern is that the known GC-bias of the Illumina high-throughput sequencing will further enhance this effect. In contrast, a separate *in vivo* data set [14], from the same lab but based on the Roche 454 sequencing platform, has a lower correlation with local GC-content (r=0.42), which is consistent with the lower GC-bias previously observed with these longer reads [18]. A recent study investigating the impact of chromatin structures on laboratory DNA manipulation [19] also noted that the sequencing bias toward higher read-density in GC-rich regions of Illumina-based deep sequencing [20] can result in a misleading overrepresentation of sequence reads in GC-rich DNA that will correlate strongly with GC-rich genomic features. The dScore was explicitly designed to be insensitive to GC-content across its analysis window (301 bp), and is less correlated ($r = 0.46$) with GC-content computed on a smaller scale (71 bp) than the Segal raw binding score ($r = 0.74$). Rather than trying to reproduce the wandering baseline seen in experimental nucleosome occupancy maps, we choose to focus on trying to accurately predict the most likely positions of linkers *vs* cores. In the posterior probabilities produced by our model, a deep null indicates a highly confident linker position and in turn a highly confident adjacent nucleosome, while regions of greater uncertainty are characterized by smaller differences between adjacent local maxima and local minima.

We believe that our discriminative framework for incorporating arbitrary heterogeneous scores directly into a sequential model will also prove useful in other segmentation applications in which a score can be interpreted directly as a label probability and may not lend itself well to being modeled using Gaussian mixtures in a generative framework—one possible example being inferring copy number variation from experimental data [21]. This framework can also be extended by using indicator variables [22] to explicitly allow for missing data or to specify, for example, that when two input tracks are both present only one of the two should be used.

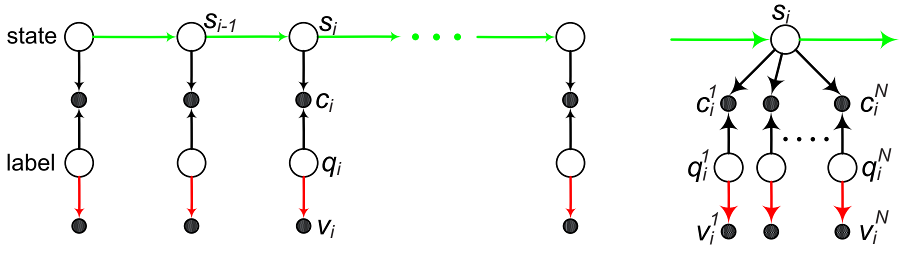
4 Methods

4.1 A Dynamic Bayesian Network for Nucleosome Prediction

The DBN that we use in this work is similar to a previous DBN-based method we developed to predict transmembrane protein topology from sequence [23], and is implemented using the Graphical Models Toolkit (GMTK) [24]. The task addressed by *Philius*, the topology prediction DBN, is the segmentation of an input protein into a series of non-overlapping regions belonging to one of three classes: *membrane*, *inside*, or *outside*. In this nucleosome prediction task, our goal is even simpler because there are only two classes of interest: *nucleosome core* and *linker*. Philius introduced a novel approach to using partially labeled data during training which we will further generalize here. Typically, when labeled data is used to train an HMM (*i.e.* supervised training), the label accompanying each observation (*e.g.* nucleotide or amino acid) specifies the value of the associated "hidden state". Philius allows for a more flexible relationship between the label and the state during training: a one-to-many relationship is defined between the labels and the states, and a special "wildcard" label allows the state variable to take on *any* value that is otherwise consistent with the topology of the model. In the case of Philius, the wildcard label is used to address the uncertainty inherent in the segment boundaries—at each segment boundary, some labels were replaced by the wildcard in order to allow the model to make small adjustments to the boundary locations during training. For the purposes of nucleosome prediction, we exploit this idea to define a similarly flexible relationship between labels and states, although in the model presented here, the labels are not observed in the traditional sense—instead they are constrained by the evidence.

Philius uses a two-pass decoding process that makes use of so-called "soft" labels to find the protein topology that maximizes the posterior probabilities at each position while obeying the grammar constraints required by the membrane topology. In this work, we show that a similar mechanism can be used to incorporate a variety of information sources to predict nucleosome positioning while obeying the grammar constraints required by the chromatin "topology".

Figure 5a shows the graphical model of our DBN, in which a single track of virtual evidence is incorporated as a soft constraint on the value of the *label* node. For simplicity, this graphical model omits the portion of the graph which takes care of the counting for the fixed-duration states. This counting mechanism is implemented exactly as in Philius [23]. To fully define the nucleosome positioning DBN, in addition to the graphical model shown in Figure 5a, the precise form of the relationship between each node and its parent(s) must be defined. We will proceed by describing each of the DBN components in turn, starting with the Markov chain over states, then the relationship between each connected state and label pair (s_i,q_i) joined by the observed child c_i, and finally how the input



(a) Single track DBN **(b)** Multiple tracks

Fig. 5. Graphical models for (a) the DBN with a single track of evidence and (b) a single frame showing the incorporation of an arbitrary number N of evidence tracks. The small black nodes represent the virtual evidence, the white nodes represent hidden variables, the subscript i refers to the genomic position and the superscripts in (b) index the evidence tracks.

model scores, experimental data, or other types of information are injected into the DBN via the relationship between the label node q_i and the virtual evidence node v_i.

A Markov Chain over Hidden States. Our model consists of five states: three to model the fixed-length nucleosome and two to model the variable-length linker. The three states that are used to describe the nucleosome and their associated lengths are the *dyad* (5 bp), and the *5' and 3' turns* (71 bp each), where the dyad refers to the central position of the nucleosomal sequence, at the axis of symmetry of the histone core. The linker is described using two states: a fixed-length state (9 bp), and a state with a geometrically-distributed length (implemented as a simple self-looping state, with minimum length 1 bp). Together these two states capture the steric hindrance constraint between adjacent nucleosomes, enforcing a minimum linker length of 10 bp, while also allowing for arbitrarily long linkers. The state transition diagram is shown in Figure 6 and consists of a simple cycle in which each state has only one possible *next* state, meaning that when a *change* in state is to occur, there is only one possible new, different state given the current state. This simple sequence of states defines the nucleosome “grammar”. For simplicity, the initial state is always defined to be the geometric-length linker state. This hard constraint greatly reduces the complexity of the inference while having relatively little effect on the predictions. For all subsequent states, the conditional relationship between each state and the previous state $Pr[s_i|s_{i-1}]$ is defined according to the deterministic grammar described above, with the exception of the self-looping linker state which transitions to the next state (the 5' turn) with probability p or remains in the linker state with probability $1 - p$. The duration model realized by this self-looping state is a geometric distribution, $Pr[k] = (1 - p)^{k-1}p$ for $k > 0$, with mean $1/p$. By using a feature in GMTK that allows for exponential weights to be applied to any edge in the DBN, we can also run our model with a completely unbiased linker model. We do this by setting a weight of 0 on the state-transition edge: this exponential weight is applied to any non-zero probability in the state-transition matrix, causing all non-zero values in the matrix to become 1. In this mode, the $Pr[k]$ defined above is equal to 1 for all values of k. The effect of this exponential weight is similar to that of the temperature constant in a Boltzmann model,

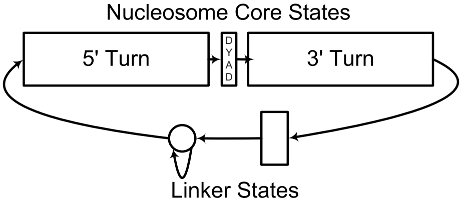


Fig. 6. State transition diagram. The width of each rectangular state is proportional to the duration specified for that state. The circular state represents the self-looping linker state which follows a geometrical duration distribution.

albeit inverted: a weight of 0 corresponds to an infinite temperature at which all possible outcomes become equally likely.

Virtual Evidence Constraints. While the backbone of our model is the same Markov chain over hidden states that exists in the traditional HMM, the relationship between the hidden state and the "observation" is quite different. While each state in an HMM is traditionally thought of as "emitting" a particular discrete or continuous observed random variable, and the probability distribution over the observed variable is conditioned on the hidden state, our model has a more discriminative flavor in which the information available at each genomic position is used to directly influence the local probability distribution over possible state assignments. The result is that the probability of a particular sequence of state assignments is weighted according to the information available at each position. This direct influence on the local probability over the possible assignments to the state variable is accomplished using the concept of "virtual evidence" [23,25,26], as will be described in more detail below. Below each state in the graphical model, a typical HMM would have a single observed node o_i, dependent on the parent state s_i according to some distribution $Pr[o_i|s_i]$. In this DBN, we have instead two distinct relationships: the first is a deterministic relationship between the state s_i, the label q_i, and the virtual evidence node c_i: $Pr[c_i|s_i, q_i]$. This construct, in which c_i is called an *observed child* because it induces a relationship between its parents, is used to define which states are consistent with a particular label: c_i is observed to be equal to 1, and the table $Pr[c_i = 1|s_i, q_i]$ implements an indicator function $I(s_i, q_i)$, which is equal to 1 if s_i and q_i are consistent with one another, and otherwise is equal to 0.

The second probabilistic relationship shown in the graphical model is between the label q_i and a second virtual evidence node v_i, and is defined as $Pr_i[v_i = 1|q_i]$. We add the subscript i to this conditional relationship to indicate that it depends on the current position, i, unlike the relationship between the *state* and the *label*, and unlike the observation distribution in a typical time-homogeneous HMM. Finally, we assign a uniform marginal distribution over the possible values of q_i: $Pr[q_i = Q] = 1/|Q|$ where Q represents a specific label, and $|Q|$ is the cardinality of the discrete label variable.

Joint Probability Distribution. We can now give the equation for the probability of a particular assignment to all of the hidden nodes, in other words to a particular sequence of states $\mathbf{s}$, and a particular sequence of labels $\mathbf{q}$:

$$Pr[\mathbf{s}, \mathbf{q}] \propto \left(Pr[s_1] \prod_{i=2}^{N} Pr[s_i|s_{i-1}] \right) \left(\prod_{i=1}^{N} \mathbf{I}[s_i, q_i]\, Pr[q_i]\, Pr_i[v_i|q_i] \right)$$

in which we use the indicator function $I(s_i, q_i)$ in place of $Pr[c_i = 1|s_i, q_i]$. The indicator function $\mathbf{I}[s_i, q_i]$ will cause all inconsistent pairs of sequences $\mathbf{s}$ and $\mathbf{q}$ to have probability zero. Considering only the subset of sequence pairs that are self-consistent $\{\bar{\mathbf{s}}, \bar{\mathbf{q}}\}$, this probability can be restated as:

$$Pr[\bar{\mathbf{s}}, \bar{\mathbf{q}}] \propto \left(Pr[\bar{s}_1] \prod_{i=2}^{N} Pr[\bar{s}_i | \bar{s}_{i-1}] \right) \left(\prod_{i=1}^{N} Pr[\bar{q}_i] \, Pr_i[v_i | \bar{q}_i] \right)$$

in which the first term in parentheses scores the sequence of states and enforces the grammar defined by the state-transition matrix, while the second term incorporates the virtual evidence at each position. Finally, we sum over all consistent label sequences $\bar{\mathbf{q}}$, to find the probability of a particular sequence of states:

$$Pr[\bar{\mathbf{s}}] \propto \left(Pr[\bar{s}_1] \prod_{i=2}^{N} Pr[\bar{s}_i | \bar{s}_{i-1}] \right) \left(\sum_{\bar{\mathbf{q}}} \prod_{i=1}^{N} Pr[\bar{q}_i] \, Pr_i[v_i | \bar{q}_i] \right)$$

This probability can be computed efficiently using the junction tree algorithm, which is a generalization of the forward-backward algorithm for HMMs, because of the underlying tree structure of the graph. We can similarly compute the posterior probabilities for the state variable at each position, and this will be the standard output of our model—specifically we plot the posterior $Pr_i[core]$ computed by summing the posterior probabilities of the three nucleosome states (the dyad and the 5' and 3' turns). Furthermore, multiple tracks of evidence can be incorporated into the model simply by replicating the evidence portion of the model as shown in Figure 5b. All of the information available at each genomic position will be used to infer the probabilities of the possible assignments to the state variable at that position.

Evidence Track Definition. We have defined the state space of our model but we have not yet precisely defined either the labels or the virtual evidence that we intend to use to define the function $Pr_i[v_i|q_i]$. We describe three possible sources of information to be used as inputs to our model, although our intent here is to describe a framework in which arbitrary sources of information can be combined in a principled manner to predict nucleosome positioning along a chromosome. The three types of nucleosome-positioning information that we describe are: a) scores from a DNA-sequence model of nucleosome positioning; b) nucleosome-occupancy data from a high-throughput sequencing experiment; and c) a transcription factor “landscape”. The first two types of information can each be used as the sole source information, while the TF landscape is shown used in conjunction with scores from a sequence model. The one-to-many relationship between each label variable and the associated state variable is customized for each type of input data.

Sequence model scores. Assuming that a sequence model score z_i can be interpreted as a log-ratio, in other words a choice between two hypotheses, we define q_i to be a binary label such that $q_i = 1$ corresponds to the dyad state, and $q_i = 0$ corresponds to any non-dyad state. The virtual evidence node, v_i is also a binary random variable, although we always observe $v_i = 1$ for all i. We assign uniform marginal probability distributions to both of these binary variables, and then use the law of total probability to find that the sum of the conditional probabilities $Pr[v_i = 1|q_i = 1]$ and $Pr[v_i = 1|q_i = 0]$ is equal to 1. Furthermore, we define the log-ratio of these two conditional probabilities to equal the aforementioned score, z_i, and therefore:

$$Pr[v_i = 1|q_i = 1] = \frac{1}{1 + e^{-z_i}} \quad \text{and} \quad Pr[v_i = 1|q_i = 0] = \frac{1}{1 + e^{z_i}}$$

Experimental data. Experimental data derived from a microarray or sequencing assay can similarly be interpreted as a log-ratio and supplied as an evidence track exactly as described for the sequence scores above.

Transcription factor binding probabilities. The third type of input information that we consider is a binding probability track representing one or more TFs. We model the relative affinity of a binding site to a particular transcription factor X using a position weight matrix (PWM) as described in [27]. Assuming that a TF can only bind in the absence of a nucleosome, *i.e.* in a linker region, we define q_i such that $q_i = 1$ corresponds to either linker state, and $q_i = 0$ corresponds to *any* state. A high TF-binding probability (high probability that $q_i = 1$) will therefore result in a higher probability of being in a linker state, while a low TF-binding probability (high probability that $q_i = 0$) will have little to no effect.

References

1. Lubliner, S., Segal, E.: Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy. Bioinformatics 25, 1348–1355 (2009)
2. Yuan, G.C., Liu, J.S.: Genomic Sequence is Highly Predictive of Local Nucleosome Depletion. PLoS Comp. Biol. 4, e13 (2008)
3. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thøaström, A., Field, Y., Moore, I.K., Wang, J.Z., Widom, J.: A genomic code for nucleosome positioning. Nature 44, 772–778 (2006)
4. Peckham, H.E., Thurman, R.E., Fu, Y., Stamatoyannopoulos, J.A., Noble, W.S., Struhl, K., Weng, Z.: Nucleosome positioning signals in genomic DNA. Genome Research 17, 1170–1177 (2007)
5. Wasson, T., Hartemink, A.J.: An ensemble model of competitive multi-factor binding of the genome. Genome Research 19, 2101–2112 (2009)
6. Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., Rando, O.J.: Genome-scale identification of nucleosome positions in *S. cerevisiae.* Science 309, 626–630 (2005)
7. Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., Nislow, C.: A high-resolution atlas of nucleosome occupancy in yeast. Nature Genetics 39, 1235–1244 (2007)
8. Yassour, M., Kaplan, T., Jaimovich, A., Friedman, N.: Nucleosome positioning from tiling microarray data. Bioinformatics 24, i139–i146 (2008)
9. Bilmes, J., Bartels, C.: Graphical Model Architectures for Speech Recognition. IEEE Signal Processing Magazine 22, 89–100 (2005)
10. Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., Pugh, B.F.: A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research 18, 1073–1083 (2008)

11. Reynolds, S.M., Bilmes, J.A., Noble, W.S.: Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in *Saccharomyces cerevisiae* and *Homo sapiens* (in submission)
12. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., Segal, E.: The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 548, 362–366 (2009)
13. Sun, W., Xie, W., Xu, F., Grunstein, M., Li, K.-C.: Dissecting Nucleosome Free Regions by a Segmental Semi-Markov Model. PLoS One 4, e4721 (2009)
14. Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J., Segal, E.: Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. PLoS Comp. Biol. 4, e1000216 (2008)
15. Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L., Gebbia, M., Talukder, S., Yang, A., Mnaimneh, S., Terterov, D., Coburn, D., Yeo, A.L., Yeo, Z.X., Clarke, N.D., Lieb, J.D., Ansari, A.Z., Nislow, C., Hughes, T.R.: A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell 32, 878–887 (2008)
16. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K.: High-resolution profiling of histone methylations in the human genome. Cell 129, 823–837 (2007)
17. Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., Zhao, K.: Dynamic regulation of nucleosome positioning in the human genome. Cell 132, 887–898 (2008)
18. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., Frazer, K.A.: Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10, R32 (2009)
19. Teytelman, L., Özaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., Eisen, M.B.: Impact of Chromatin Structures on DNA Processing for Genomic Analyses. PLoS One 4, e6700 (2009)
20. Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H.: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36, e105 (2008)
21. Marioni, J.C., Thorne, N.P., Tavaré, S.: BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22, 1144–1146 (2006)
22. Hoffman, M.M., Buske, O.J., Bilmes, J.A., Noble, W.S.: Segway: a dynamic Bayesian network method for segmenting genomic data (in preparation)
23. Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A., Noble, W.S.: Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. PLoS Comp. Biol. 4, e1000213 (2008)
24. Bilmes, J., Zweig, G.: The Graphical Models Toolkit: An Open Source Software System for Speech and Time-Series Processing. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE Press, New York (2002)
25. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
26. Reynolds, S.M., Bilmes, J.A.: Part-of-speech tagging using virtual evidence and negative training. In: Proc. HLT and EMNLP, pp. 459–466. IEEE Press, New York (2005)
27. Granek, J.A., Clarke, N.D.: Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol. 6, R87 (2005)