Published in final edited form as:

J Am Stat Assoc. 2014 January 1; 109(505): 48–62. doi:10.1080/01621459.2013.862169.

A locally convoluted cluster model for nucleosome positioning signals in chemical map

Liqun Xi^{1,4}, Kristin Brogaard^{2,4}, Qingyang Zhang^{1,4}, Bruce Lindsay³, Jonathan Widom^{2,5}, and Ji-Ping Wang^{1,2}

¹Department of Statistics, Northwestern University, IL 60208, USA

²Department of Molecular Biosciences, Northwestern University, IL 60208, USA

³Department of Statistics, The Pennsylvania State University, PA, 16802

Abstract

Nucleosome is the fundamental packing unit of DNA in eukaryotic cells, and its positioning plays a critical role in regulation of gene expression and chromosome functions. Using a recently developed chemical mapping method, nucleosomes can be potentially mapped with an unprecedented single-base-pair resolution. Existence of overlapping nucleosomes due to cell mixture or cell dynamics, however, causes convolution of nucleosome positioning signals. In this paper, we introduce a locally convoluted cluster model and a maximum likelihood deconvolution approach, and illustrate the effectiveness of this approach in quantification of the nucleosome positional signal in the chemical mapping data.

Keywords

Poisson cluster model; Deconvolution; Nucleosome positioning; EM algorithm; Chemical mapping; Negative Binomial cluster model

1 Introduction

Eukaryotic genomic DNA is compacted in different levels into a DNA-protein complex known as chromatin. In the first level of packing, the genomic DNA is packed into an array of nucleosome structures, each involving a stretch of DNA of ~147 bp, sharply bent and wrapped around eight protein particles called histones (Richmond and Davey (2003), Figure 1a). The free DNA between two consecutive nucleosomes is called linker DNA, which varies in length across different genomic regions and across different species. Any DNA sequence can potentially bind to histone particles, while the actual positions of nucleosomes *in vivo* are determined by various factors including histone binding affinity of the DNA, histone concentration, competition from other DNA-binding proteins and other chromatin remodelers (Segal and Widom (2009b)). Recent studies have shed light on the critical role of nucleosomes in chromosome functions and gene regulation (Segal *et al.* (2006), Schones *et al.* (2008), Lee *et al.* (2004, 2007), Whitehouse and Tsukiyama (2006), Ioshikhes *et al.* (2006)). Hence knowing the exact positions of nucleosomes *in vivo* will further advance understanding of the role of nucleosomes in various aspects.

Correspondence to: Ji-Ping Wang.

⁴These authors contributed equally to this work.

⁵Deceased.

The conventional method for mapping nucleosomes is to use a special enzyme called micrococcal nuclease (MNase) to digest chromatin fiber. MNase degrades the free DNA (linker DNA) (Horz and Altenburger (1981)) while the nucleosome DNA tends to survive the digestion because of the protection of the histones (Figure 1b). Sequencing the resulting DNA products thus yields a genome-wide map of nucleosome positions. The MNase approach, however, has its limitation in mapping accuracy due to several factors. Firstly DNA bound with other proteins rather than histones may be protected from MNase digestion, yielding over-long remaining linker DNAs in either end of the targeted nucleosome DNAs. Secondly the accuracy of MNase map is affected by the well-known MNase specificity (Dingwall *et al.* (1981), Chung *et al.* (2010)), by which the enzyme tends to cleave into a dinucleotide consisting of A/T, causing a systematic bias in the map. For example, if there is a dinucleotide "AT" in the linker region near the nucleosome edge (but not exactly the edge), a cleavage more likely occurs between the A and T rather than the exact nucleosome edge, leading to a DNA product that has extra base pairs of linker DNA in the end. Consequently the MNase map tends to be inaccurate to infer the exact centers of nucleosomes.

Recently a chemical approach to achieve a single-base-pair mapping accuracy has been developed by Widom's lab (Brogaard *et al.* (2012)) based on previous work by Flaus *et al.* (1996), Flaus and Richmond (1999). The inside chemistry and biology are referred to the papers above. Briefly, this approach targets to break the DNA backbone of the nucleosomes on both strands at specific positions near nucleosome centers (Figure 1c). The break points or cleavage sites are then mapped by high throughput sequencing technology. Figure 2 (upper part) shows the landscape of cleavage frequency on Watson and Crick strands in a typical genomic region. Some key facts of the chemical mapping can be summarized as follows:

- **Primary and secondary sites.** In theory the chemical method aims to break the nucleosome DNA backbone at the position -1 or $+6$ relative to the nucleosome center (defined as position 0) on each strand (Figure 1c, "+/-" refers to the 5'/3' direction respectively). These two sites are referred to as the *primary* and *secondary* sites respectively. For example, the highest cleavage peaks on Figure 2 likely (but not necessarily) correspond to either primary or secondary site (see details of identification of the two sites in Section 6).
- **Clustered cleavages.** The cleavages do not exclusively occur at primary and secondary sites, but rather in multiple positions around the two sites. In other words, cleavages occur in clusters around the primary and secondary sites. For example, in Figure 2, around genomic coordinates 31334 and 31478, we observe two major cleavage clusters around the peaked positions on each strand.
- **Cleavage convolution due to nucleosome overlap.** Ideally if there is only one positioned nucleosome in a 147 bp region centered at genomic location i , one should only observe cleavage clusters at $i - 1$ and $i + 6$ on the Watson strand and $i + 1$ and $i - 6$ on the Crick. However, the experimental data typically are collected from a heterogeneous population of cells. The nucleosome positioning can differ between cells by cell types or cell activities. Figure 2 shows that many peaked cleavage clusters may exist on the same strand in a 147 bp region, for example, in the regions of 31575–31650 and 31750–31850. This evidences the existence of multiple overlapping nucleosomes. Thus the observed cleavages at some positions can be convoluted if cleavage clusters arise from different but closely positioned nucleosomes.

The goal of the chemical map is to accurately identify the center positions of nucleosomes genome-wide and quantify their relative abundance. The chemical approach in nucleosome

mapping is new and will have far-reaching impact on this important topic. This paper is intended to provide a complete report on statistics aspects of this problem. In particular we propose a new maximum likelihood deconvolution approach that is far more efficient than the Bayesian approach that was originally proposed. In addition we improve the Poisson cluster model by extending it to negative binomial model to account for the over-dispersion issue of cleavage counts. A further discussion of the Bayesian deconvolution approach is presented in the supplementary materials accompanying this paper.

2 A generic convoluted cluster model

We consider a generic locally convoluted cluster model as follows. Let $\mathbf{I} = \{i : i = 1, 2, \dots, N\}$ define a one-dimensional lattice or a discrete time domain. We define a random cluster at position $i \in \mathbf{I}$ as $\mathbf{X}_i = \{X_i^j : j \in \mathbf{J}\}$ where $\mathbf{J}$ is a set of integers indexing the clustered positions relative to i , and each X_i^j is a random variable. For example a set $\mathbf{J} = \{-2, -1, 0, 1, 2\}$ will define a cluster of random variables that occur at positions $i-2, i-1, i, i+1$ and $i+2$. For simplicity of illustration, suppose X_i^j follows a Poisson distribution (but not limited to) with mean $k_i \lambda_j$, $j \in \mathbf{J}$, where λ_j is known, and k_i is a cluster specific but unknown parameter of interest, measuring the cluster signal intensity. A $k_i = 0$ implies there is no cluster positioned at i . We further assume that X_i^j 's are independent across i and j . As a cluster, $\mathbf{X}_i$ is a random unit governed by the cluster parameter k_i . It is only defined at position $i \in \mathbf{I}' := \{i : i \in \mathbf{I}, 1 \leq i+j \leq N, \forall j \in \mathbf{J}\}$, so is k_i . For example, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_{N-1}$ and $\mathbf{X}_N$ are undefined given $\mathbf{J} = \{-2, -1, 0, 1, 2\}$ above. Define $\Lambda := \{\lambda_j : j \in \mathbf{J}\}$, and $\mathbf{K} := \{k_i : i \in \mathbf{I}'\}$.

In real applications, X_i^j , as a counting statistic, often measures the number of individuals/ events observed at position $i+j$ derived from the i th cluster. Individuals or events from different Poisson clusters are undistinguishable regarding their cluster identity. For example, the total number of observed individuals/events at position i , denoted W_i , can be attributed to the count at the j th position within the $(i-j)$ th cluster (i.e., $i-j+j=i$):

$$W_i = \sum_{\{j: j \in \mathbf{J}, 1 \leq i-j \leq N\}} X_{i-j}^j.$$

In other words, the observed total is a local convolution of independent Poisson random variables from the neighboring Poisson clusters. We want to estimate the parameters $\mathbf{K}$.

To illustrate the complexity of this model, consider a simple cluster model where $N = 4$ and $\mathbf{J} = \{-1, 0, 1\}$ (This implies that k_i is only defined for $i = 2, 3$ due to constraint of $\mathbf{J}$). Then we have

$$\begin{aligned} W_1 &\sim f(k_2 \lambda_{-1}), \\ W_2 &\sim f(k_2 \lambda_0 + k_3 \lambda_{-1}), \\ W_3 &\sim f(k_2 \lambda_1 + k_3 \lambda_0), \\ W_4 &\sim f(k_3 \lambda_1), \end{aligned}$$

where f stands for a Poisson density. The likelihood function is

$$\begin{aligned}
L \propto & (k_2\lambda_{-1})^{W_1} \exp(-k_2\lambda_{-1}) \\
& \cdot (k_2\lambda_0 + k_3\lambda_{-1})^{W_2} \exp(-k_2\lambda_0 - k_3\lambda_{-1}) \cdot (k_2\lambda_1 + k_3\lambda_0)^{W_3} \exp(-k_2\lambda_1 - k_3\lambda_0) \\
& \cdot (k_3\lambda_1)^{W_4} \exp(-k_3\lambda_1),
\end{aligned} \quad (1)$$

or more generally, the log likelihood can be written as

$$\ell \propto \sum_{i=1}^N \left[W_i \log \sum_j k_{i-j} \lambda_j - \sum_j k_{i-j} \lambda_j \right]. \quad (2)$$

(note: for convenience of notation, $\sum_j$ in Eq. (2) stands for $j \in \mathbf{J}$ and $1 - i - j - N$).

3 Application in nucleosome mapping data

The convoluted cluster model has important genomic applications. For example, Qi and coauthors applied a similar convolution model to the ChIP-Chip data to detect the protein binding sites with a Bayesian deconvolution approach, where the convolution signal was continuous (Qi *et al.* (2006)). Here we applied this generic model to the chemical mapping data generated from budding yeast (*Saccharomyces cerevisiae*). This data set contains 105 million mapped cleavage sites from six independent mapping experiments using high throughput sequencing technology. Since each chromosome is separate, we can apply the deconvolution algorithm to each chromosome independently. For convenience of mathematical formularization, we pretend there is only one double-stranded DNA sequence of length N , i.e., $\mathbf{I} = \{1, \dots, N\}$. The observed cleavage frequency on the two strands are denoted as $\mathbf{W} = \{W_i : i \in \mathbf{I}\}$ and $\mathbf{C} = \{C_i : i \in \mathbf{I}\}$.

3.1 A single-template Poisson cluster model

We first consider a Poisson cluster model to the chemical mapping data by making the following assumptions:

1. Each genomic location can in principle be the center of a nucleosome except positions $1, \dots, 73$, and $N - 72, \dots, N$ because each nucleosome must have a fixed length of 147 bp. Let $\mathbf{I}' = \{74, \dots, N - 73\}$.
2. For each $i \in \mathbf{I}'$, the Poisson cluster is defined by $\mathbf{J} = \{-2, -1, 0, 1, 4, 5, 6, 7\}$, in which the first four indices define a sub-cluster around the primary site -1 , and the second four for the secondary site $+6$ (The detail of the template training is discussed in Section 6 below). Note: the cluster index set $\mathbf{J}$ is defined in the direction of $5' \rightarrow 3'$ for both Watson and Crick strands. In DNA sequence alignment the genome positions are indexed with reference to the Watson strand. Therefore if a nucleosome center is at position i , the Watson cluster is defined at genomic positions $i + \mathbf{J}$, and the Crick cluster is defined at $i - \mathbf{J}$.
3. Let $\mathbf{X}_i = \{X_i^j : j \in \mathbf{J}\}$ and $\mathbf{Y}_i = \{Y_i^j : j \in \mathbf{J}\}$ be the clusters on the Watson and Crick strands respectively governed by a nucleosome centered at i , in which X_i^j and Y_i^j are defined as cleavage frequency for genomic positions $i + j$ on the Watson and $i - j$ on the Crick respectively. All clusters ($\mathbf{X}_i$ or $\mathbf{Y}_i$) and all positions within the

cluster $(X_i^j \text{ or } Y_i^j)$ are assumed to be mutually independent. In addition, we assume X_i^j and Y_i^j follow a Poisson distribution with mean $k_{i1}\lambda_j$ and $k_{i2}\lambda_j$ respectively, where k_{i1} and k_{i2} are unknown parameters, and λ_j are known predefined template mean parameters (Section 6). The defined template $\Lambda = \{\lambda_j : j \in \mathbf{J}\}$ is plotted in Figure 3.

4. In addition to the nucleosome specific cleavages, we assume there are some cleavages non-specific to nucleosomes but due to noise at each position on each strand. These noise terms, denoted ε_{i1} and ε_{i2} for the Watson and Crick strands respectively, also follow a Poisson distribution with mean σ_{i1} and σ_{i2} .

Remark 1. We included the noise term in the model because we observe chemical cleavages virtually at every genomic location even in the well-known nucleosome free regions. This cannot be explained by the mechanism of the chemical method (i.e., cleavages only occur specifically around the center of nucleosomes). Later we shall use the signal/noise ratio to quantify the nucleosome positional signal.

Under these assumptions, we have the following convolution model:

$$W_i = \sum_{j \in J} X_{i-j}^j + \varepsilon_{i1},$$

$$C_i = \sum_{j \in J} Y_{i+j}^j + \varepsilon_{i2}, i \in \mathbf{I}'$$

The parameters k_{i1} and k_{i2} defined in Assumption 3 measure the nucleosome positioning signal strength relative to the Poisson cluster template $\Lambda := \{\lambda_j : j \in \mathbf{J}\}$ on Watson and Crick strand respectively. A larger k_{i1} or k_{i2} implies stronger evidence of a positioned nucleosome centered at i . Hence we term it the *nucleosome center positioning* (NCP) score at genomic location i . In theory these two scores should accord with each other as both measure the strength of nucleosome positioning signal at the same location. Nevertheless, we observed substantially unbalanced cleavages on the two strands in some regions. Hence we impose a constraint on k_{i1} and k_{i2} such that $k_{i2} = \delta_i k_{i1}$ where

$$\delta_i = \sum_{j=1-73}^{i+73} C_j / \sum_{j=i-73}^{i+73} W_j,$$

is the ratio of total number of cleavages on Watson and Crick strands in the region of $i \pm 73$. We denote these parameters of interest as $\mathbf{K} = \{k_{i1} : i \in \mathbf{I}'\}$.

For the non-specific cleavage or noise signal, we observe that it varies across regions dramatically. To simplify the calculation, we use the average of the lower half of the observed cleavage frequency in the region of $i \pm 73$ of each strand as the mean parameter of the Poisson noise, denoted by σ_{i1} and σ_{i2} henceforth. This simplification could over- or under-estimate the noise level, causing under- or over-estimation of the resulting k_{i1} or k_{i2} . But it is less of a concern than it seems to be since this rule is uniformly applied to every genomic location and we shall make nucleosome calls based on the NCP score/noise ratio at each genomic location (Section 3.5 and Section 5).

Under the independent Poisson assumptions,

$$W_i \sim f\left(\sum_{j \in J} k_{(i-j)1} \lambda_j + \sigma_{i1}\right),$$

$$C_i \sim f\left(\sum_{j \in J} k_{(i+j)2} \lambda_j + \sigma_{i2}\right).$$

The likelihood is given by

$$L(\mathbf{K}; \mathbf{W}, \mathbf{C}) = \prod_{i \in I'} f(W_i; \mathbf{K}) f(C_i; \mathbf{K}), \quad (3)$$

3.2 Maximum likelihood approach by generalized EM algorithm

The computing speed issue encountered in the Bayesian algorithm by Broggard et al. (2012) motivates us to seek a more efficient deconvolution algorithm. The log likelihood from Eq. (3) can be written as

$$\ell(\mathbf{K}) = \sum_{i \in I'} W_i \log\left(\sum_{j \in J} k_{(i-j)1} \lambda_j + \sigma_{i1}\right) - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i-j)1} \lambda_j + \sigma_{i1} \right] + \sum_{i \in I'} C_i \log\left(\sum_{j \in J} k_{(i+j)2} \lambda_j + \sigma_{i2}\right) - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i+j)2} \lambda_j + \sigma_{i2} \right]. \quad (4)$$

The summation sign inside of the log makes it difficult to find maximum likelihood estimate (MLE) of $\mathbf{K}$ by directly solving the likelihood equations based on (4). We use an elegant EM type of algorithm developed by Li et al. (2007) to find the MLE of $\mathbf{K}$. Suppose we have an initial estimate of $\mathbf{K}$, say $\mathbf{K}^0 \equiv \{k_{i1}^0, \dots\}$ (note: $k_{i2}^0 = \delta_i k_{i1}^0$). Define

$$t_{i1} \equiv t_{i1}(\mathbf{K}^0) := \sum_{j \in J} k_{(i-j)1}^0 \lambda_j + \sigma_{i1},$$

$$t_{i2} \equiv t_{i2}(\mathbf{K}^0) := \sum_{j \in J} k_{(i+j)2}^0 \lambda_j + \sigma_{i2}.$$

Then the log likelihood ratio under $\mathbf{K}$ and $\mathbf{K}^0$ can be written as

$$\begin{aligned} \ell(\mathbf{K}) - \ell(\mathbf{K}^0) &= \sum_{i \in I'} W_i \log \left[\frac{\sum_{j \in J} k_{(i-j)1} \lambda_j}{\sum_{j \in J} k_{(i-j)1}^0 \lambda_j} \frac{k_{(i-j)1}^0 \lambda_j}{t_{i1}} \right] \\ &+ \frac{\sigma_{i1}}{t_{i1}} - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i-j)1} \lambda_j \right] + \sigma_{i1} + \sum_{i \in I'} C_i \log \left[\frac{\sum_{j \in J} k_{(i+j)2} \lambda_j}{\sum_{j \in J} k_{(i+j)2}^0 \lambda_j} \frac{k_{(i+j)2}^0 \lambda_j}{t_{i2}} \right] \\ &+ \frac{\sigma_{i2}}{t_{i2}} - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i+j)2} \lambda_j \right] + \sigma_{i2} \\ &+ \sum_{i \in I'} t_{i1} + \sum_{i \in I'} t_{i2}. \end{aligned} \quad (5)$$

Note that the last two terms in Eq. (5) are constants given $\mathbf{K}^0$. In addition, the summation

inside of the log sign for the Watson strand terms, i.e., $\sum_{j \in J} \frac{k_{(i-j)1}^0}{k_{(i-j)1}^0} \frac{k_{(i-j)1}^0 \lambda_j}{t_{i1}} + \frac{\sigma_{i1}}{t_{i1}}$ can be regarded as the expectation of a discrete random variable that is defined on sample space $\{\{\frac{k_{(i-j)1}^0}{k_{(i-j)1}^0} : j \in J\}, 1\}$, with corresponding probabilities $\{\{\frac{k_{(i-j)1}^0 \lambda_j}{t_{i1}} : j \in J\}, \frac{\sigma_{i1}}{t_{i1}}\}$. The same applies to the Crick strand terms in Eq. (5). By Jensen's Inequality, we have

$$\begin{aligned} \ell(\mathbf{K}) - \ell(\mathbf{K}^0) &\geq \sum_{i \in I'} W_i \left[\sum_{j \in J} \frac{k_{(i-j)1}^0 \lambda_j}{t_{i1}} \log\left(\frac{k_{(i-j)1}}{k_{(i-j)1}^0}\right) \right] - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i-j)1} \lambda_j + \sigma_{i1} \right] + \sum_{i \in I'} C_i \left[\sum_{j \in J} \frac{k_{(i+j)2}^0 \lambda_j}{t_{i2}} \log\left(\frac{k_{(i+j)2}}{k_{(i+j)2}^0}\right) \right] - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i+j)2} \lambda_j + \sigma_{i2} \right] \\ &:= Q(\mathbf{K}, \mathbf{K}^0). \end{aligned} \quad (6)$$

Clearly the right hand side of inequality (6), defined as $Q(\mathbf{K}, \mathbf{K}^0)$, equals zero if $\mathbf{K} = \mathbf{K}^0$. If we find a $\mathbf{K}^*$ that maximizes $Q(\mathbf{K}, \mathbf{K}^0)$ given $\mathbf{K}^0$, then $Q(\mathbf{K}^*, \mathbf{K}^0) = 0$ and hence $\ell(\mathbf{K}^*) = \ell(\mathbf{K}^0)$. Note that the constant terms can be dropped from Q , which leads to a simplified objective function

$$\begin{aligned} Q^*(\mathbf{K}, \mathbf{K}^0) &= \sum_{i \in I'} W_i \left[\sum_{j \in J} \frac{k_{(i-j)1}^0 \lambda_j}{t_{i1}} \log(k_{(i-j)1}) \right] \\ &\quad - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i-j)1} \lambda_j \right] + \sum_{i \in I'} C_i \left[\sum_{j \in J} \frac{k_{(i+j)2}^0 \lambda_j}{t_{i2}} \log(k_{(i+j)2}) \right] \\ &\quad - \sum_{i \in I'} \left[\sum_{j \in J} k_{(i+j)2} \lambda_j \right]. \end{aligned} \quad (7)$$

A closed form solution to $Q^*(\mathbf{K}, \mathbf{K}^0)$ is given by (under the constraint $k_{i2} = \delta_i k_{i1}$, $\forall i \in I'$),

$$k_{m1} = \frac{k_{m1}^0 \left[\sum_{j \in J} W_{m+j} \lambda_j / t_{(m+j)1} + \delta_m \sum_{j \in J} C_{m-j} \lambda_j / t_{(m-j)2} \right]}{(1 + \delta_m) \sum_{j \in J} \lambda_j}, m \in I' \quad (8)$$

One can find MLE of $\mathbf{K}$ by iteratively updating k_{m1}^0 using Eq.(8). The algorithm was initialized with starting values of $\mathbf{K}^0$ and was run for each chromosome separately. In each run, every k_{m1} on the entire chromosome was updated sequentially and the algorithm was stopped when the relative difference of log likelihood values from two consecutive runs was smaller than 10^{-9} . We found this convergence criterion is sufficiently stringent and more stringent criterion does not result in noticeable difference in the resulting NCP scores. The convergence of algorithm is insensitive to the starting values of $\mathbf{K}$. We tried different starting values including fixed and randomly generated, but the algorithm always converged to the same solution.

3.3 Negative binomial cluster model

One limitation of the Poisson model is its lack of flexibility in modeling the variance component. Over-dispersion is a known issue in high throughput sequencing data due to various confounding factors including DNA sequence features. To account for this, we consider a negative binomial (NB) model for the cleavage count distribution. For a negative

binomial distribution $\text{NB}(r, p)$, the variance-mean ratio is $1/(1 - p)$, thus the success probability p tunes the over-dispersion extent. In addition, convolution of independent NB random variables with a common p results in a NB distribution with the same p . Thus for computational convenience, we shall assume that all NB distributions in the cluster model share a common but unknown p parameter.

For convenience of illustration, we define a single-template model by assuming that at each cluster position j , the cleavage frequency follows $\text{NB}(r_j, p)$. The data training step described in the Section 3.1 and Section 6 yielded an estimate of average cleavage level at each cluster positions, defined as λ_j for $j \in \mathbf{J}$. As the nucleosomes in the training set only represent a subset of possible nucleosomes genome-wide, the over-dispersion presented by this subset may be not optimal. In addition, it should be noted that the value of λ_j in the template only provides a measure of average cleavage level at cluster position j across the genome. For a given cluster at genomic location i , the actual distribution of the cleavages at any cluster position is scaled by the NCP score (i.e., k_{i1} and k_{i2}), which measures the nucleosome abundance at the given position:

$$\begin{aligned} X_i^j &\sim \text{NB}(k_{i1}r_j, p) = \text{NB}(k_{i1}\lambda_j(1-p)/p, p), \\ Y_i^j &\sim \text{NB}(k_{i2}r_j, p) = \text{NB}(k_{i2}\lambda_j(1-p)/p, p) \end{aligned}$$

Thus if we estimate p directly from the selected nucleosomes in the training set without accounting for k , this over-dispersion will be greatly inflated. Here we only keep the mean parameters in the template (i.e. $\lambda_j = pr_j/(1-p)$) fixed, and shall iteratively estimate the p parameter in the deconvolution algorithm.

Likewise, the noise terms are also assumed to follow the NB model with the same p and pre-determined mean level σ_{i1} and σ_{i2} as described in the Poisson case:

$$\begin{aligned} \varepsilon_{i1} &\sim \text{NB}(\sigma_{i1}(1-p)/p, p), \\ \varepsilon_{i2} &\sim \text{NB}(\sigma_{i2}(1-p)/p, p). \end{aligned}$$

Define $s_{i1} := k_{i1}(1-p)/p$, $s_{i2} := k_{i2}(1-p)/p$, $v_{i1} := \sigma_{i1}(1-p)/p$, $v_{i2} := \sigma_{i2}(1-p)/p$. As in the Poisson model, we impose the constraint $k_{i2} = \delta_i k_{i1}$, $\forall i \in \mathbf{I}'$ (and hence $s_{i2} = \delta_i s_{i1}$), where δ_i is the ratio of the total cleavages in the region of $i \pm 73$ on the Crick and Watson strands. Let $\mathbf{S} = \{s_{i1} : i \in \mathbf{I}'\}$. Then the convolution model becomes

$$\begin{aligned} W_i &\sim \text{NB}\left(\sum_{j \in \mathbf{J}} s_{(i-j)1} \lambda_j + v_{i1}, p\right), \\ C_i &\sim \text{NB}\left(\sum_{j \in \mathbf{J}} s_{(i+j)2} \lambda_j + v_{i2}, p\right). \end{aligned}$$

The likelihood function is given by:

$$\begin{aligned}
\ell(\mathbf{S}, p) = & \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{W_i-1} \log \left(\sum_{j \in J} s_{(i-j)1} \lambda_j \right. \right. \\
& \left. \left. + v_{i1} + l \right) + \left(\sum_{j \in J} s_{(i-j)1} \lambda_j \right. \right. \\
& \left. \left. + v_{i1} \right) \log(1-p) + W_i \log p + \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{C_i-1} \log \left(\sum_{j \in J} s_{(i+j)2} \lambda_j \right. \right. \right. \quad (9) \\
& \left. \left. + v_{i2} + l \right) + \left(\sum_{j \in J} s_{(i+j)2} \lambda_j \right. \right. \\
& \left. \left. + v_{i2} \right) \log(1-p) + C_i \log p \right].
\end{aligned}$$

Given p , MLE of $\mathbf{S}$ can be obtained using a similar strategy as used in the Poisson case. Suppose we have an initial estimate $\mathbf{S}^0$, we define

$$\begin{aligned}
t_{il1} &\equiv t_{il1}(\mathbf{S}^0) := \sum_{j \in J} s_{(i-j)1}^0 \lambda_j + v_{i1} + l, \\
t_{il2} &\equiv t_{il2}(\mathbf{S}^0) := \sum_{j \in J} s_{(i+j)2}^0 \lambda_j + v_{i2} + l,
\end{aligned}$$

where l is an integer. The log likelihood ratio under $\mathbf{S}$ and $\mathbf{S}^0$ can be written as:

$$\begin{aligned}
& \ell(\mathbf{S}) - \ell(\mathbf{S}^0) \\
&= \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{W_i-1} \log \left[\frac{\sum_{j \in J} s_{(i-j)1} \lambda_j + v_{i1} + l}{t_{il1}} \right] \right. \\
& \quad \left. + \left(\sum_{j \in J} (s_{(i-j)1} - s_{(i-j)1}^0) \lambda_j \log(1-p) \right) + \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{C_i-1} \log \left[\frac{\sum_{j \in J} s_{(i+j)2} \lambda_j + v_{i2} + l}{t_{il2}} \right] \right. \right. \\
& \quad \left. \left. + \left(\sum_{j \in J} (s_{(i+j)2} - s_{(i+j)2}^0) \lambda_j \log(1-p) \right) \right] \right].
\end{aligned}$$

Clearly the summation inside of the log sign for the Watson strand terms (same for the Crick

terms), i.e., $\log \left[\frac{\sum_{j \in J} s_{(i-j)1} \lambda_j + v_{i1} + l}{t_{il1}} \right]$ can be regarded as the expectation of a discrete random variable that is defined on sample space $\left\{ \left\{ \frac{s_{(i-j)1}}{s_{(i-j)1}^0} : j \in \mathbf{J} \right\}, 1 \right\}$, with corresponding probabilities $\left\{ \left\{ \frac{s_{(i-j)1}^0 \lambda_j}{t_{il1}} : j \in \mathbf{J} \right\}, \frac{v_{i1} + l}{t_{il1}} \right\}$. By Jensen's inequality,

$$\begin{aligned}
\ell(\mathbf{S}) - \ell(\mathbf{S}^0) &\geq \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{W_i-1} \sum_{j \in J} \frac{s_{(i-j)1}^0 \lambda_j}{t_{il1}} \log \frac{s_{(i-j)1}}{s_{(i-j)1}^0} \right. \\
&\quad + \sum_{j \in J} (s_{(i-j)1} - s_{(i-j)1}^0) \lambda_j \log(1 - p) \\
&\quad \left. + \sum_{i \in \mathbf{I}'} \left[\sum_{l=0}^{C_i-1} \sum_{j \in J} \frac{s_{(i+j)2}^0 \lambda_j}{t_{il2}} \log \frac{s_{(i+j)2}}{s_{(i+j)2}^0} \right. \right. \\
&\quad \left. + \sum_{j \in J} (s_{(i+j)2} - s_{(i+j)2}^0) \lambda_j \log(1 - p) \right] \right]. \quad (10)
\end{aligned}$$

Thus we can find maximizer of $\mathbf{S}$ given p based on right hand side of Eq. (10) iteratively as follows:

$$s_{m1} = \frac{s_{m1}^0 \left[\sum_{j \in J} \sum_{l=0}^{W_{m+j}-1} \frac{\lambda_j}{t_{(m+j)l1}} + \delta_m \sum_{j \in J} \sum_{l=0}^{C_{m-j}-1} \frac{\lambda_j}{t_{(m-j)l2}} \right]}{\left[\sum_{j \in J} \lambda_j + \delta_m \sum_{j \in J} \lambda_j \right] \log(1 - p)}, m \in \mathbf{I}'. \quad (11)$$

The entire algorithm started with an initial p and initial $\mathbf{S}^0$. Given p , $\mathbf{S}$ was updated for each chromosome separately based on Eq. (11) until the relative log likelihood values from two consecutive runs differed by less than 10^{-9} . Given the updated $\mathbf{S}$, the maximizer of p can be easily found by combining all chromosomes based on likelihood equation (9). The two steps were iteratively alternated until two consecutive p estimates differed by less than 10^{-3} . We found the results were insensitive to the choice of initial values just as in the Poisson model.

3.4 Four-template model

The nucleosome map from the single-template deconvolution approach allows us to examine how the cleavage pattern around the primary and secondary sites may depend on the base composition, particularly in the ± 10 bp of dyad. The most striking pattern is that about 60% of the unique nucleosomes have a nucleotide "A" at the -3 position of dyad (or "T" at $+3$ position by symmetry), which confirms the finding of a previously published *in vitro* MNase study (Thåström *et al.* (2004)) that the nucleosomes strongly prefers an "A" at the -3 position of dyad. The presence or absence of an "A" nucleotide at position -3 or a "T" at position $+3$ relative to the nucleosome center is associated with four distinct cleavage patterns (Figure 4a). The cleavage magnitude differs substantially across the four different situations, represented by A(-3)T(+3), A(-3)-T(+3), -A(-3)T(+3), -A(-3)-T(+3), where $-3/+3$ stands for position, A/-A and T/-T stand for presence/absence of A/T respectively. Interestingly the nucleosomes from the first three templates exhibit similar but stronger AA/TT/AT/TA pattern than the -A(-3)-T(+3) nucleosomes (the latter only represent 10% in the training set, Figure 4b). This could imply some difference in the structural requirements of dinucleotide motifs at these periodic positions when A/T is absent at -3 and $+3$ positions respectively; or it could indicate inaccuracy of alignment that degrades the signal due to absence of A/T at -3 and $+3$ positions or other reasons. Future studies are needed to investigate the cause of this degradation.

Under the single template model, this possible base-composition related bias will be carried over to the NCP score and hence to affect the downstream calculation of nucleosome occupancy. In addition, we note that the primary site tends to have higher cleavages than the secondary site except in the $-A(-3)T(+3)$ case where the opposite pattern is observed. The single template model (where the primary site has higher cleavages than the secondary site, Figure 3) may result in miscalls such that the true secondary site would be regarded as the primary site in the $-A(-3)T(+3)$ case. To account for this possible bias, we also consider a four-template deconvolution model. Denote the average cleavage frequency at the eight positions of the four templates as:

$$\Lambda^h = \{\lambda_j^h : j \in \mathbf{J}\}, h=1, 2, 3, 4.$$

For each genomic location on either strand, the applicable template is pre-determined based on whether there is an "A" or "T" at position -3 or $+3$ respectively. The templates on the two strands are mutually dependent because of the base pairing. For example, an "A" at position -3 on the Watson strand is paired with a "T" at $+3$ position of the Crick strand. The four-template model can be readily configured by replacing λ_j with the template-specific $\lambda_j^{w(i)}$ or $\lambda_j^{c(i)}$ in the Poisson and NB model, where $w(i)$ and $c(i)$ are the pre-determined template-specific configurations ($= 1, 2, 3, 4$) at position i on the Watson and Crick strands respectively. In Section 4 we will show the map accuracy based on the four-template model is better than that with single-template model.

3.5 Nucleosome maps

For both approaches, we followed the same method described in Brogaard *et al.* (2012)) to define the unique and redundant nucleosome maps respectively. We first defined the nucleosome center positioning (NCP) score at position i as the average of the NCP scores from both strands, i.e., $k_i = (k_{i1} + k_{i2})/2$ or $k_{i1}(1 + \delta_i)/2$. The NCP scores in the example genomic region in Figure 2 clearly suggest two dominant nucleosomes positioned at coordinates 31334 and 31478 (or 31477) respectively. In contrast in the region between 31750 and 31850, we observe many possible nucleosomes positions with comparable signal strength. To account for the noise, we further define the average noise at position i as $\sigma_i = (\sigma_{i1} + \sigma_{i2})/2$, and use the ratio of k_i/σ_i as the signal/noise ratio index for measuring the potential of a positioned nucleosome at location i . If $\sigma_i < 0.5$, we set it as 0.5 to avoid over-inflation of the ratio. The k_i/σ_i indices provide a comprehensive view of preferential positions genome-wide for nucleosomes. Nucleosome maps can be generated by applying different filtering criteria to the NCP score/noise ratio indices. For example, it is always appealing to define a unique map of the most prominent nucleosomes genome-wide (by uniqueness, neighboring nucleosomes are not allowed to overlap). Here we defined a "unique" map using a greedy algorithm by selecting positions based on the magnitude of k_i/σ_i in descending order but allowing up to 40 bp overlap between nucleosomes. The algorithm stopped when no gaps longer than 107 bp were left without nucleosome coverage. This small overlap helps reduce possible mis-calls due to mis-calls from previous rounds in the neighboring regions (though most of the called nucleosomes are identical in maps with or without allowing this small overlap). The exhaustive selection algorithm however may cause false definition of nucleosomes at sites with very small NCP/noise ratio, possibly in the long linker regions. Such long linkers exist in the genome, for example in the so-called nucleosome-depletion regions upstream of transcription start site (TSS) of each gene (in total about 6000 genes in yeast genome). Thus we defined the unique map as the subset that corresponded to the top 90% of NCP score/noise ratio, which excluded about 7400 defined positions from the greedy algorithm. Alternatively one may directly threshold the NCP/noise ratio. This was not pursued in this study as it is difficult to generate an objective and

convincing rule. Using the smallest ratio from the unique map as threshold value, we also defined a redundant map by including all positions whose NCP score/noise ratio was greater than the threshold.

4 Some results

In this section, we present some results that are related to justification of the proposed methods. In the supplementary materials accompanying this paper, we discussed the simple Bayesian approach originally by Brogaard *et al.* (2012) and an alternative Bayesian approach. In the latter, the simple Bayesian model was extended by including additional hierarchical layers to allow calculation of the posterior probability of whether any given position is a nucleosome center. Here we present a brief comparison between different approaches regarding computing efficiency and the resulting nucleosome maps. A thorough report of biological results (based on the simple Bayesian approach) is referred to the paper by Brogaard *et al.* (2012).

The Poisson MLE took about 30 minutes to finish genome-wide calculation with C++ codes running on a single 2.6GHz CPU core in Windows 7 Platform, which was about 6 fold faster than the NB MLE, 960 fold faster than the simple Bayesian, and 1350 fold faster than the alternative Bayesian algorithm under the current convergence criteria (Figure 5a). It is still practically feasible to apply the Poisson or NB MLE deconvolution to higher organisms including human (which is 275 times of the size of yeast) by utilizing the computing clusters, but not for the Bayesian approaches without algorithmic approximation for shortening computing time. The RAM usage was comparable in different approaches. For example, for yeast chromosome IV (of size 1.5 M, the largest among 16 chromosomes), the total RAM loading was about 90 M–100 M under the double precision mode (Figure 5a). The computing speed without accounting for effect of RAM usage is linear in chromosome length. For the MLE approaches, the demand of large RAM for running the algorithm on the entire chromosome of higher organisms may significantly deteriorate the computing speed even if the RAM is adequate. Thus we recommend to divide the entire chromosome into several pieces to run the algorithm separately by selecting some nucleosome free region (e.g., with extremely low cleavage counts) as the segmenting points.

Nucleosomes *in vivo* are enriched for particular DNA dinucleotide motifs at particular rotational positions along the nucleosome DNA, most notably the periodic occurrences of AA/TT/TA/AT and GC/GG/CG/CC dinucleotides with periodicity ~ 10 bp (Wang and Widom (2005)). These dinucleotide signals are collective features of nucleosomes when they are aligned at the nucleosome centers, thus the strength of periodicity is sensitive to the alignment quality and can be regarded as an indicator of the mapping accuracy of the nucleosome centers. Compared to the MNase map, the AA/TT/TA/AT signal from the chemical mapping (four-template Poisson model) was substantially stronger, proving that the chemical map is overwhelmingly more accurate than the MNase map (Figure 5b). The results from the two Bayesian approaches are extremely consistent with each other. Only 30 unique nucleosomes out of ~67,500 genome-wide were defined at different locations by the two approaches (see additional details in Supplementary Materials). The single-template and four-template models under the Poisson MLE yielded about 73% identically defined nucleosomes out of a total of ~ 67,500 genome-wide. While comparing the nucleosomes differentially defined by the two models (about 27%), we found the AA/TT/TA/AT signal from the four-template model was slightly but consistently stronger than the single-template model, manifesting that the four-template approach yields more accurate nucleosome calls (Figure 5c). Hence in the following comparisons, we shall only compare results from the four-template models including the simple Bayesian, Poisson MLE and NB MLE.

The overall consistency in terms of NCP score and the defined unique nucleosomes between different approaches are presented in Table 1. In summary most of the nucleosomes were consistently defined by different methods. The discrepancy typically corresponds to positions with relatively smaller NCP scores (Figures 6a,b), and thus the defined nucleosome positions in the discrepant sets could be more error-prone. We first examined the distance between the nucleosomes differentially defined in the Bayesian approach vs Poisson MLE and in Poisson MLE vs NB MLE. Nucleosomes *in vivo* are known to frequently prefer multiple positions spaced by $10n$ bp in the genome (Brogaard *et al.* (2012)). The distance frequency in both plots exhibits a similar dominant pattern peaked at $10n$ bp (Figure 6c,d), confirming that discrepant nucleosome positions in different maps mostly represent alternative bona fide nucleosome positions. Nevertheless the detailed difference may indicate a subtle difference in mapping accuracy. For example, the peaks at $\sim \pm 7$ bp in Figure 6c,d probably correspond to miscalls of nucleosomes in either set due to misclassification of the primary site as the secondary site or vice versa (note: primary and secondary sites are positioned at -1 and $+6$ relative to dyad respectively).

Compared to the Poisson MLE, the Bayesian prior tends to pull up small estimates and shrink large estimates (Figure 6a). The NCP scores from the two approaches are highly consistent with Pearson correlation of .99, and 91% of the resulting nucleosomes are identical in the two unique maps. The remaining 9% differentially defined nucleosomes show strong AA/TT/AT/TA signal (Figure 6e), comparable to the entire set of unique nucleosomes as shown in Figure 5b,c. The two sets of nucleosomes differ slightly in average AA/TT/AT/TA frequency (Figure 6e), while the amplitude of periodical signal (e.g. peak-valley height difference) is comparable, suggesting an overall close mapping accuracy.

The four-template NB model converged at $p = 0.668$, suggesting a substantial variance-mean over-dispersion ratio of 3. Interestingly accounting for over-dispersion in the NB model tends to dampen the resulting NCP scores compared to the Poisson MLE results (Figure 6b), resulting in a correlation of 0.978 in NCP scores and 13% differentially defined unique nucleosomes. In total there were 9152 unique nucleosomes selected by the NB model (out of $\sim 67,500$) but not by the Poisson model, and 9086 by the Poisson but not NB model. The NB-specific subset revealed a higher amplitude of periodicity of AA/TT/AT/TA signal than the Poisson set (Figure 6f), suggesting that accounting for over-dispersion may improve estimation of the NCP scores, and thus leading to a better definition of nucleosome map.

The improved performance in NB model is attributable to its resistance to violations of the model assumptions in the data. We conducted a systematic simulation study to compare the two approaches in various settings and results are presented in the Supplementary materials. When the cluster data are truly simulated from NB model without convolution, the Poisson method actually provides a consistent estimate of the NCP score even without considering the over-dispersion (Supplementary Figure 1). This is not surprising as the Poisson MLE of the NCP score is essentially a moment estimator. Nevertheless, the real data frequently contain cleavage outliers at some position(s) within the cluster that significantly deviate from the template models defined in Figures 3 and 4 (Supplementary Figure 2). The simulation study suggests that the both shrinkage and inflation of the mean parameter in any cluster position(s) tend to result in smaller NCP scores in the NB model than in the Poisson model, particularly when NCP score are larger, confirming the observation in Figure 6b for the real data. More importantly, the Poisson model is more susceptible to such perturbations (as expected) and yields slightly less accurate call of nucleosome positions in a simulated convolution data (Supplementary Figure 3).

5 Discussion

The convolution model developed in this paper only utilizes the cleavage information on the two strands. We did not use any DNA sequence information as model components to help predict the nucleosome positioning. Therefore the strong DNA sequence preferences including the AA/TT/TA/AT dinucleotide signal revealed by the chemical map in the entire nucleosome region are true facts about nucleosomes, not model "predicted" or "biased" features. Such dinucleotide sequence features are far stronger than ever were discovered in the MNase maps. Given this fact, we believe the definition of nucleosome map can be improved by combining sequence information, in particular in chemical mapping experiments where the cleavages are relative low due to inadequate sequencing.

We also explored several variants of the convolution model. The model specified in Eqs. (3) and (4) contains an adaptive noise term on top of the convoluted Poisson or NB clusters. This noise term was included for philosophical consideration rather than operational. Clearly cleavages observed at many genomic locations do not correspond to the true nucleosome center signals, but noise due to unknown reasons. The noise level also varies across regions for different reasons. First, difference in base composition of DNA sequence may react differently to chemical method, resulting in variation in noise level. Second, the nucleosome occupancy, which is essentially measured by the cleavage counts or the resulting NCP scores, is jointly determined by DNA sequence features and chromosome functions. For example, the nucleosome occupancy in active genes tend to have lower occupancy than inactive genes regardless of the DNA sequence features. This may cause variation of noise level accordingly. Thus normalizing the NCP score by noise could help account for these factors to recover the true positioning strength of the given DNA sequence.

In practice, it is difficult to distinguish the noise from weak nucleosome signals due to low occupancy. Noise in some regions could yield even higher cleavage counts than the true signal in regions where nucleosomes have lower occupancy. Nucleosome positioning *in vivo* has been frequently classified into two states, namely "stable" vs "fuzzy" (Segal and Widom (2009a)). A stable positioning often refers to a dominant nucleosome that is positioned in a genomic region of about 147 bp, forming a single strong nucleosome occupancy peak (Figure 7a); whereas a fuzzy positioning often implies the coexistence of many different nucleosome positions nearby causing a nucleosome-occupied region much wider than 147 bp (Figure 7b). In the latter case, the "noise" term as we defined above, regardless of whether corresponds to true noise or true signal, tends to be larger than that obtained from the stable nucleosome region. Using NCP score/noise ratio as a measure for nucleosome calls is hoped to provide insights into why the selected positions are more significant relative to other nearby positions in the fuzzy case. Figure 8a presents the AA/TT/TA/AT signal by quartiles of NCP score/noise ratio of the nucleosomes in the unique map, suggesting that stronger AA/TT/TA/AT signals are associated with more stably positioned nucleosomes under the NCP score/noise ratio criterion.

Alternatively one could use the NCP score itself as the measure to make nucleosome calls (with noise term in the model), or even drop the noise term from the model. We found about 87% of the unique nucleosomes defined based on the NCP score alone are identical to the original map based on NCP score/noise ratio (under NB model). The distance between the nucleosomes in the two discrepant sets again shows a pre-dominant pattern peaked at 10n bp (Figure 8b), illustrating that most of the nucleosomes in the two sets correspond to alternative positions. Interestingly the average AA/TT/TA/AT frequency from the NCP-defined nucleosomes is lower than NCP score/noise ratio-defined ones, while the amplitude of the periodical signals is comparable (Figure 8c), suggesting that these two methods do not yield pronounced difference in mapping accuracy of the most prominent nucleosomes. A

further plot of AA/TT/TA/AT frequency of the entire set by quartiles of NCP score shows that the amplitude of the signal is positively correlated with the NCP score, while the baseline level of AA/TT/TA/AT signal is reversely correlated with NCP score (Figure 8d). This suggests that the NCP score measure of nucleosome positioning signal might be confounded by the base composition of DNA sequences. In contrast, the NCP score/noise ratio seems to provide a clearer clue on the sequence features that facilitate nucleosome forming as illustrated in Figure 8a.

We defined the templates as the average cleavages at the eight positions around the primary and secondary sites. In principle we would like **J** to be concise and precise. If too few positions were defined in the templates, then a well positioned nucleosome could result in multiple NCP score peaks, yielding redundant nucleosome calls at neighboring positions. In the meanwhile one could consider to include more positions into the templates. But given the cleavage patterns presented in Figures 3 and 4 (see template training in Section 6), we believe the eight positions are adequate in capturing the cleavage information around the primary and secondary sites. Expansion of template size increases the computing time, but it does not results in noticeable improvement in mapping accuracy (assessed based on AA/TT/TA/AT signal, results not shown).

We can also relax the constraint imposed on the Watson and Crick NCP scores, i.e., $k_{i2} = \delta_i k_{i1}$ (where δ_i was defined as the ratio of total cleavages on the two strands in the region of $i \pm 73$) and estimate k_{i1} and k_{i2} separately. The computing time is doubled with independent k_{i1} and k_{i2} . The resulting NCP scores and nucleosome maps again are very consistent with the original maps.

6 Details of model training

The mechanism of chemical mapping dictates that there are two major cleavage sites around the nucleosome center on each strand, i.e., primary and secondary sites (Flaus *et al.* (1996)). The exact positions of the two sites in the *in vivo* mapping experiments are not clear to us. Our first effort is to identify the exact positions of the two major cleavage sites. One important fact of nucleosome is the structural mirror symmetry of the two strands around the nucleosome center. A major cleavage site at position x relative to the nucleosome center on one strand implies another site mirror-symmetrically positioned on the opposite strand. Thus the distance of the cross-strands major cleavage peaks can be used to deduce the positions of the two sites. We define local maximum peaks of the cleavage frequency on Watson and Crick strands separately using a greedy algorithm as follows:

1. Within each chromosome, the global maximum cleavage peak is first identified on each strand.
2. The next maximum peak is selected in the regions that is at least 147 bp away from any position(s) selected from the previous step(s).
3. Repeat step 2 until all local maximum peak positions are selected.

Figure 9 plots the frequency of distances between the Crick peaks and Watson peaks (Crick peak coordinate - Watson peak coordinate). Figure 9 shows that the major cut sites on Crick and Watson within the same nucleosome tend to be paired, with distance equal to 2 and -5 nucleotides (nt) respectively. If we regard that these cleavage peaks mainly arise from the primary and secondary sites, then there is only one possible configuration of the two sites that can give 2 and -5 nt distances. That is, on each strand the cleavage must occur at sites -1 and +6 relative to the nucleosome center (position 0). This configuration gives three different Crick-Watson peak-peak distances: 2 nt — a cleavage peak at position -1 on both Watson and Crick; -5 nt — a cleavage peak at position -1 on one strand and +6 on the other

strand, and -12 nt — a cleavage peak at position +6 on both strands. The last distance (-12 nt) was not obvious in the peak-peak distance plot. This does not mean that the secondary site is not one of the major cleavage sites, but rather it tends to have lower average cleavages than the primary site (Figure 3). As a result, there is only a small chance that the cleavages at the secondary sites from both strands are identified as the local maximum peaks simultaneously within the same nucleosome (as the peak-peak distance is defined above). The -1 and +6 sites are referred to as primary and secondary sites respectively.

The observed cleavage frequencies around the identified local maximum peaks suggest that neighboring positions immediately around the primary and secondary sites also tend to have substantially higher cleavages than the background signal. This motivates us to define an eight-position cluster index set as $\mathbf{J} = \{-2, -1, 0, 1, 4, 5, 6, 7\}$, in which the first four indices define a sub-cluster around the primary position -1, and the second four for the secondary site at position +6. Our second effort is to train a template that can represent the average cleavage frequency at each position of $\mathbf{J}$. We define a weight template $\mathbf{P} = \{p_j : j \in \mathbf{J}, \sum_j p_j = 1\}$ to measure the relative cleavage frequency at the eight positions. At each genomic location $i \in \mathbf{I}'$ (if assumed to be the center of a nucleosome), the template weighted score is calculated as

$$\sum_{j \in \mathbf{J}} (p_j W_{i+j} + p_j C_{i-j}),$$

where W_{i+j} and C_{i-j} are the observed cleavage frequencies at genomic locations $i+j$ on the Watson and $i-j$ on the Crick strands respectively. If $\mathbf{P}$ reflects the true weights, the most probable nucleosome center tends to achieve a maximum template score in a local region. With initial values set as follows:

$$\mathbf{P} = \{p_j : j \in \mathbf{J}\} = (1/8, 1/4, 1/8, 0, 0, 1/8, 1/4, 1/8),$$

we selected unique peaks of the template score as putative nucleosome centers using the same greedy algorithm described above. We updated the template weights by the relative average cut frequency at the eight positions around the selected putative nucleosomes. These two steps were repeated until the weight template converges. The converged template is plotted in Figure 3 in absolute scale. The template training is insensitive to the initial weights assigned in $\mathbf{P}$, as the scanning is performed simultaneously on the two strands and the relative positioning of the primary and secondary cleavage sites across strands (as shown in Figure 9) guarantees that the weights converge to the same values at the same cluster positions. It should be noted that different experimental protocol might yield different cleavage patterns in the chemical mapping. Thus we recommend to re-train the template for new data. In our data, the six experiments were performed under the same protocol and we found the templates generated from individual experiments were very consistent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We dedicate this paper to Dr. Jonathan Widom, who tragically passed away due to heart attack in 2011. Without his insights, passion and support, this work would not have been accomplished. The work was supported by National Institutes of Health under grant numbers U54CA143869, R01GM075313, R01GM058617, and T32GM00806.

References

- Brogaard K, Xi L, Wang J-P, Widom J. A map of nucleosome positions in yeast at base-pair resolution. *Nature*. 2012; 486:496–501. [PubMed: 22722846]
- Chung H, Dunkel I, Heise F, Linke C, Krobitsch S, Ehrenhofer-Murray A, Sperling S, Vingron M. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE*. 2010; 5(12):e15754. [PubMed: 21206756]
- Dingwall C, Lomonosoff G, Laskey R. High sequence specificity of micrococcal nuclease. *Nucleic Acids Research*. 1981; 9(12):1659–1673.
- Flaus A, Richmond T. Base-pair resolution mapping of nucleosome positions using site-directed hydroxy radicals. *Methods Enzymol*. 1999; 304:251–263. [PubMed: 10372364]
- Flaus A, Luger K, Tan S, Richmond T. Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proc Natl Acad Sci U S A*. 1996; 93(4):1370–1375. [PubMed: 8643638]
- Horz W, Altenburger W. Sequence specific cleavage of dna by micrococcal nuclease. *Nucleic Acids Research*. 1981; 9:2643–2658. [PubMed: 7279658]
- Ioshikhes IP, Albert I, Zanton SJ, Pugh B. Nucleosome positions predicted through comparative genomics. *Nature Genetics*. 2006; 38(10):1210–1215. [PubMed: 16964265]
- Lee C, Shibata Y, Rao B, Strahl B, Lieb J. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*. 2004; 36:900–905. [PubMed: 15247917]
- Lee W, Tillo D, Bray N, Morse R, Davis R, Hughes T, Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*. 2007; 39(10):1235–1244. [PubMed: 17873876]
- Li J, Ray S, Lindsay B. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*. 2007; 8:1687–1723.
- Qi Y, Rolfe A, MacIsaac K, Gerber G, Pokholok D, Zeitlinger J, Danford T, RD D, Fraenkel E, Jaakkola T, Young R, Gifford D. High-resolution computational models of genome binding events. *Nat Biotechnol*. 2006; 24(10):963–970. [PubMed: 16900145]
- Richmond TJ, Davey CA. The structure of dna in the nucleosome core. *Nature*. 2003; 423(6936):145–150. [PubMed: 12736678]
- Schones D, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 2008; 132(5):887–898. [PubMed: 18329373]
- Segal E, Widom J. From dna sequence to transcriptional behaviour: a quantitative approach. *Nature Reviews Genetics*. 2009a; 10:443–456.
- Segal E, Widom J. What controls nucleosome positions? *Trends Genet*. 2009b; 25(8):335–343. [PubMed: 19596482]
- Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore I, Wang J-PZ, Widom J. A genomic code for nucleosome positioning. *Nature*. 2006; 442:772–778. [PubMed: 16862119]
- Thåström A, Bingham L, Widom J. Nucleosomal locations of dominant dna sequence motifs for histone-dna interactions and nucleosome positioning. *J. Mol. Biol*. 2004; 338:695–709. [PubMed: 15099738]
- Wang J-PZ, Widom J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Research*. 2005; 22(22):6743–6755. [PubMed: 16339114]
- Whitehouse I, Tsukiyama T. Antagonistic forces that position nucleosomes in vivo. *Nat. Struct. Mol. Biol*. 2006; 13:633–640. [PubMed: 16819518]

Figure 1. Schematic diagrams of nucleosome structure, MNase mapping and chemical mapping of nucleosomes positions

a. Top: Diagrams showing two consecutive nucleosomes and the linker DNA. The right panel shows that the relative orientation of neighboring nucleosomes depends on the linker DNA length in crystal structure. Bottom: A diagram showing four nucleosomes positions along the genomic DNA. The last two corresponds to two overlapping nucleosomes which cannot exist simultaneously on the same DNA molecule, but can co-exist in the cell population due to heterogeneity. **b.** A diagram showing MNase mapping of nucleosome positions. The blue arrows mark the the ideal cleavage point, the edge of nucleosomes. In reality the MNase digestion is not perfect and the exact cleavage point may depend on the digestion time and is biased by DNA sequence features around the edge. **c.** A diagram showing chemical mapping of nucleosome positions. Blue arrows mark the exact position on the two strands where cleavages occur dominantly.

Figure 2. Raw cleavage frequency and nucleosome center positioning (NCP) scores from deconvolution in a typical genomic region on Chromosome XVI

The upper panel shows the raw cleavage frequency on both strands, where the cleavage peaks point to the approximate positions of nucleosomes. The lower panel plots the NCP score from the the Bayesian deconvolution algorithm, suggesting two well-positioned nucleosomes at coordinates 31334 and 31478 (or 31477) respectively, but many possible nucleosome positions in the region between 31750 and 31850 with comparable signal strength.

Figure 3. A template representing the average cleavage frequency in eight positions including $J = \{-2, -1, 0, 1, 4, 5, 6, 7\}$, defining the Poisson cluster used in the deconvolution algorithm

Figure 4. A four-template cleavage model

a. Four major distinct cleavage patterns around the primary and secondary sites. The four templates are specified by the presence/absence of a nucleotide "A" or "T" at position -3 and +3 respectively. For example, template A(-3)T(+3) stands for presence of an "A" at position -3 and a "T" at +3, and -A(-3)T(+3) stands for absence of "A" at -3 and presence of a "T" at +3 position etc. **b.** AA/TT/TA/AT frequency in nucleosomes that represent the four templates in the training set.

Figure 5. Overview of deconvolution approaches in computing efficiency and mapping accuracy
a. RAM usage on chromosome IV and genome-wide computing time of yeast. **b.** AA/TT/TA/AT frequency from nucleosomes defined in the chemical map shows substantially stronger signal compared with MNase map, suggesting significantly better mapping accuracy of chemical mapping technique. **c.** AA/TT/TA/AT frequency of unique nucleosomes differentially defined in the four-template model vs single-template model shows slightly but consistently improved signal in the former, suggesting improved mapping accuracy of the four-template model.

Figure 6. Comparison of unique maps defined by different deconvolution approaches
a, b. Scatter plots of NCP scores of all unique nucleosomes called by Poisson MLE vs Bayesian and Poisson MLE vs NB MLE respectively. **c, d.** The distance between the nearest nucleosomes differentially called by Poisson MLE vs Bayesian and Poisson MLE vs NB MLE respectively. **e, f.** Comparison of AA/TT/AT/TA signal of differentially called nucleosomes in Poisson MLE vs Bayesian and Poisson MLE vs NB MLE.

Figure 7. Stable and fuzzy nucleosome positioning

a. Shown are two nucleosomes of stable positioning with low noise. **b.** Shown is a region with fuzzy nucleosome positioning where NCP score tends to be scaled down more than the stable region because of larger noise. In both plots, the occupancy at a given position is defined the summation of NCP scores in the ± 73 bp region.

Figure 8. Comparison of nucleosome maps defined based on NCP score vs NCP score/noise ratio under NB MLE

a. AA/TT/TA/AT frequency by quartiles of NCP/noise ratio from the unique map based on Poisson MLE. **b.** Distance between differentially defined unique nucleosomes by NCP score vs NCP/noise ratio. **c.** AA/TT/TA/AT frequency of differentially defined unique nucleosomes by NCP score vs NCP/noise ratio. **d.** AA/TT/TA/AT frequency by quartiles of NCP score from the unique map based on Poisson MLE.

Figure 9. Plot of frequency of Crick-Watson peak-peak distance

The two peaks at 2 and -5 nucleotides imply two major cleavage sites at position -1 and +6 relative to the nucleosome center. The two peaks correspond to the combinations of primary-primary (+2 nt) and primary-secondary (-5 nt) pairs.

Table 1

Comparison of different deconvolution approaches

	Bayesian	Poisson MLE	NB MLE
Bayesian	-	0.990 (91%)	0.978 (87%)
Poisson MLE	-	-	0.976 (87%)

Presented in table are Pearson correlation of NCP scores genome-wide and the percentage (in parentheses) of consistent nucleosomes defined in unique maps from the simple Bayesian, Poisson MLE and NB MLE approaches.