

Supplementary Materials to “A Non-Homogeneous
Hidden Markov Model on First Order Differences for
Automatic Detection of Nucleosome Positions”

Pei Fen Kuan¹, Dana Huebert², Audrey Gasch³, Sündüz Keleş^{1,4*}

¹Department of Statistics, University of Wisconsin,
Madison, WI 53706.

²Department of Cellular and Molecular Biology, University of Wisconsin,
Madison, WI 53706.

³Department of Genetics, University of Wisconsin,
Madison, WI 53706.

⁴Department of Biostatistics and Medical Informatics, University of Wisconsin,
Madison, WI 53706.

*E-mail: keles@stat.wisc.edu

May 12, 2008

1 Model fitting for the non-homogeneous HMM with the Expectation-Maximization Algorithm

Let q_t be the hidden state latent variable for probe t and $\lambda = (\pi, A, B)$, where A is the transition probability matrix and B is the emission distribution. The complete log likelihood is given by:

$$\begin{aligned} \log P(O, Q | X, \lambda) &= \log P(O | Q, X, \lambda) P(Q | X, \lambda) \\ &= \log \left[\prod_{t=1}^T \prod_{i=1}^N b_i(O_t)^{I(q_t=i)} \right] \left[\prod_{t=1}^{T-1} \prod_{i=1}^N \prod_{j=1}^N a_{i,j}(X_{t+1})^{I(q_t=i, q_{t+1}=j | X_{t+1})} \right] \\ &\quad \left[\prod_{i=1}^N \pi_i^{I(q_1=i)} \right]. \end{aligned}$$

We assume that

$$\begin{aligned} b_i(O_t) &= N(\mu_i, \sigma_i^2), \\ a_{i,j}(X_{t+1}) &= \frac{\exp(\gamma_{i,j} + \beta_j X_{t+1})}{\sum_{k=1}^N \exp(\gamma_{j,k} + \beta_k X_{t+1})}. \end{aligned}$$

Expected complete log likelihood is given by

$$\begin{aligned} E[\log P(O, Q | X, \lambda)] &= \sum_{i=1}^N P(q_1 = i | O, \lambda) \log \pi_i \\ &\quad + \sum_{t=1}^T \sum_{i=1}^N P(q_t = i | O, \lambda) \log \left[\frac{1}{\sqrt{2\pi(X_1)\sigma_i^2}} \exp\left(-\frac{(O_t - \mu_i)^2}{2\sigma_i^2}\right) \right] \\ &\quad + \sum_{t=1}^{T-1} \sum_{i=1}^N \sum_{j=1}^N P(q_t = i, q_{t+1} = j | O, X_{t+1}, \lambda) \log a_{i,j}(X_{t+1}). \end{aligned}$$

E-step: Define two variables $\gamma_t(i)$ and $\xi_{g,t+1}(i, j)$:

$$\gamma_t(i) = P(q_t = i | O, \lambda)$$

$$\begin{aligned}
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_T(i)} \\
&= \sum_{j=1}^N P(q_t = j, q_{t+1} = j \mid O, X_{t+1}, \lambda),
\end{aligned}$$

$$\begin{aligned}
\xi_{t+1}(i, j) &= P(q_t = i, q_{t+1} = j \mid O, X_{t+1}, \lambda) \\
&= \frac{\alpha_t(i)a_{i,j}(X_{t+t})b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i)a_{i,j}(X_{t+t})b_j(O_{t+1})\beta_{t+1}(j)} \\
&= \frac{\alpha_t(i)a_{i,j}(X_{t+t})b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \alpha_T(i)},
\end{aligned}$$

where $\alpha_t(i) = P(O_1, \dots, O_t, q_t = i \mid \lambda)$ and $\beta_t(i) = P(O_{t+1}, \dots, O_T \mid q_t = i, \lambda)$.

M-step:

$$\begin{aligned}
&\max E[\log P(O, Q \mid X, \lambda)] \\
&\text{s.t. } \sum_{i=1}^N \pi_i = 1, \\
&\quad \sum_{j=1}^N a_{i,j}(X_{t+1}) = 1, \quad t = 1, \dots, T-1
\end{aligned}$$

yields

$$\begin{aligned}
&\hat{\pi}_i = \gamma_1(i), \\
&\left\{ \begin{aligned} \hat{\mu}_1 &= \frac{\sum_{t=1}^T \gamma_t(B_N)O_t - \sum_{t=1}^T \gamma_t(B_L)O_t}{\sum_{t=1}^T \sum_{i \in \{B_N, B_L\}} \gamma_t(i)} \\ \hat{\mu}_2 &= \frac{\sum_{t=1}^T \sum_{i \in \{N_1, N_{2a}, L_3\}} \gamma_t(i)O_t - \sum_{t=1}^T \sum_{i \in \{N_{2c}, N_3, L_1\}} \gamma_t(i)O_t}{\sum_{t=1}^T \sum_{i \in \{N_1, N_{2a}, N_{2c}, N_3, L_1, L_3\}} \gamma_t(i)}, \quad \text{if } \hat{\mu}_1 \geq \hat{\mu}_2 \end{aligned} \right. \\
&\hat{\mu}_1 = \hat{\mu}_2 = \frac{\sum_{t=1}^T \sum_{i \in \{B_N, N_1, N_{2a}, L_3\}} \gamma_t(i)O_t - \sum_{t=1}^T \sum_{i \in \{B_L, N_{2c}, N_3, L_1\}} \gamma_t(i)O_t}{\sum_{t=1}^T \sum_{i \in \{B_N, B_L, N_1, N_{2a}, N_{2c}, N_3, L_1, L_3\}} \gamma_t(i)}, \quad \text{if } \hat{\mu}_1 < \hat{\mu}_2.
\end{aligned}$$

The non-parametric transition probabilities for the case study is updated as follows:

$$\begin{aligned}
\hat{a}_{i,j} &= \frac{\sum_{t=1}^{T-1} \xi_{t+1}(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \text{ for } i \neq N_3, B_L, L_3, \\
\hat{a}_{N_3, B_L}^p &= \frac{\sum_{t=1}^{T-1} \xi_{t+1}(N_3, B_L) I(X_{t+1} \geq 0)}{\sum_{t=1}^{T-1} \gamma_t(N_3) I(X_{t+1} \geq 0)}, \\
\hat{a}_{B_L, B_N}^n &= \frac{\sum_{t=1}^{T-1} \xi_{t+1}(B_L, B_N) I(X_{t+1} < 0)}{\sum_{t=1}^{T-1} \gamma_t(B_L) I(X_{t+1} < 0)}, \\
\hat{a}_{L_3, B_N}^n &= \frac{\sum_{t=1}^{T-1} \xi_{t+1}(L_3, B_N) I(X_{t+1} < 0)}{\sum_{t=1}^{T-1} \gamma_t(L_3) I(X_{t+1} < 0)}.
\end{aligned}$$

Note that the parameters in the logistic regression model cannot be solved analytically. The parameters can be optimized via a conjugate gradient algorithm. The details can be found in Robertson et al. (2004). The computation of $\alpha_t(i)$ and $\beta_t(i)$ is based on the forward and backward procedure (Rabiner; 1989).

1.1 Forward procedure

Let $\alpha_t = P(O_1, O_2, \dots, O_t, q_t = i \mid \lambda)$.

- Initialization:

$$\alpha_1(i) = P(O_1, q_1 = i \mid \lambda) = \pi_i b_i(O_1), \text{ for } 1 \leq i \leq N.$$

- Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^N \alpha_t(i) a_{i,j} X_{(t+1)}] b_j(O_{t+1}), \text{ for } 1 \leq t \leq T-1, 1 \leq j \leq N.$$

- Termination:

$$P(O \mid \lambda) = \sum_{i=1}^N \alpha_T(i).$$

1.2 Backward procedure

Let $\beta_t(i) = P(O_{t+1}, \dots, O_T \mid q_t = i, \lambda)$.

- Initialization:

$$\beta_T(i) = 1, \text{ for } 1 \leq i \leq N.$$

- Induction:

$$\beta_t(i) = \sum_{j=1}^N a_{i,j}(X_{t+1})b_j(O_{t+1})\beta_{t+1}(j), \text{ for } t = T - 1, T - 2, \dots, 1, 1 \leq j \leq N.$$

- Termination:

$$P(O | \lambda) = \sum_{i=1}^N \beta_1(i)b_i(O_1)\pi_i.$$

The optimal hidden state sequence is obtained via Viterbi algorithm (Rabiner; 1989).

1.3 Viterbi algorithm

Define $\delta_t(i) = \max_{q_1, \dots, q_{t-1}} P(q_1, \dots, q_t = i, O_1, \dots, O_t | \lambda)$.

- Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \text{ for } 1 \leq i \leq N,$$

$$\psi_1(i) = 0, \text{ for } 1 \leq i \leq N.$$

- Recursion:

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{i,j}^{(t)}] b_j(O_t), \text{ for } 2 \leq t \leq T, 1 \leq j \leq N,$$

$$\psi_t(j) = \operatorname{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i) a_{i,j}^{(t)}], \text{ for } 2 \leq t \leq T, 1 \leq j \leq N.$$

- Termination:

$$P(O | \lambda)^* = \max_{1 \leq i \leq N} [\delta_T(i)],$$

$$q_T^* = \operatorname{argmax}_{1 \leq i \leq N} [\delta_T(i)].$$

- Path backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \text{ for } t = T - 1, T - 2, \dots, 1.$$

2 HMM architecture in simulation studies

Figures 1(a) and 1(b) show the HMM hidden states architecture used in Simulation I and II, respectively. We simulate data for tiling arrays consisting of 50-mer oligonucleotides probes tiled at 20 base pairs resolution. In Simulation I, we consider sinusoidal curves with different periodicity as shown in Figure 2. An example of simulated data from Simulation II with mixture emission distributions is given in Figure 3.

3 Example of “hand picked” annotation

An example of “hand picked” annotation from Yuan et al. (2005) is given in Figure 4, which illustrates that some uncertainties in mapping nucleosome-linker boundaries still exist despite careful visual inspection. For instance, the linker between the nucleosomes 2 and 3 in the top left panel could possibly be shifted by one probe.

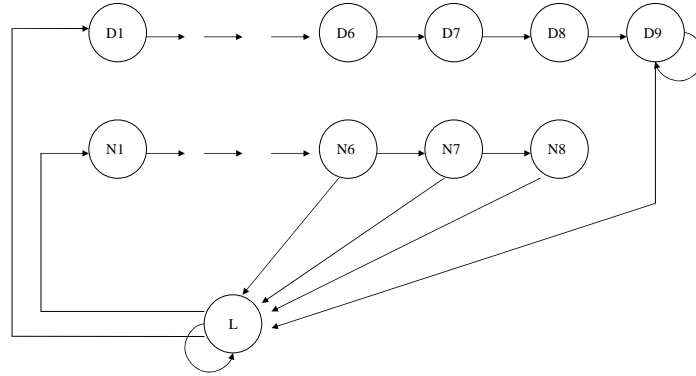
4 Summarizing reads across genomic position in ChIP-Seq data

Figure 5 illustrates the procedure in obtaining total reads for each genomic position in the ChIP-Seq data. In particular, we first extend these reads to 150 base pairs according to the sequence orientation for both the plus and minus strands. The total reads for each genomic position is then taken to be the sum of all extended reads at the position.

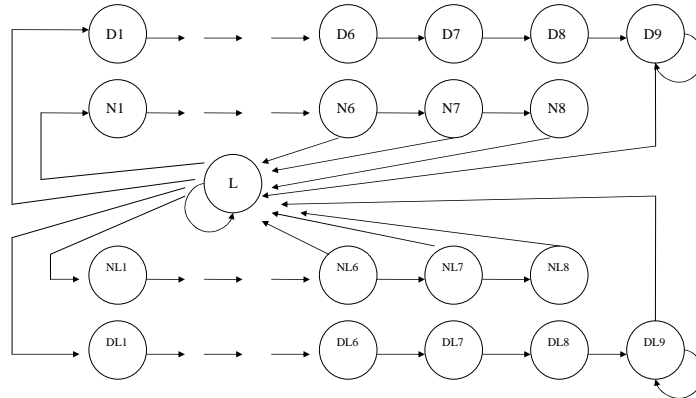
References

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286.

- Robertson, A., Kirshner, S. and Smyth, P. (2004). Downscaling of daily rainfall occurrence over northeast brazil using a hidden markov model, *Journal of Climate* **17**(22): 4407–4424.
- Yuan, G., Liu, Y., Dion, M., Slack, M., Wu, L., Altschuler, S. and Rando, O. (2005). Genome-scale identification of nucleosome positions in *s.cerevisiae*, *Science* **309**: 626–630.



(a)



(b)

Figure 1: *HMM architecture in Yuan et al. (2005)*. D1-D9 represent delocalized high nucleosomes, N1-N8 represent well-positioned high nucleosomes, DL1-DL9 represent delocalized low-signal nucleosomes, NL1-NL8 represent well-positioned low-signal nucleosomes and L represents a linker probe.

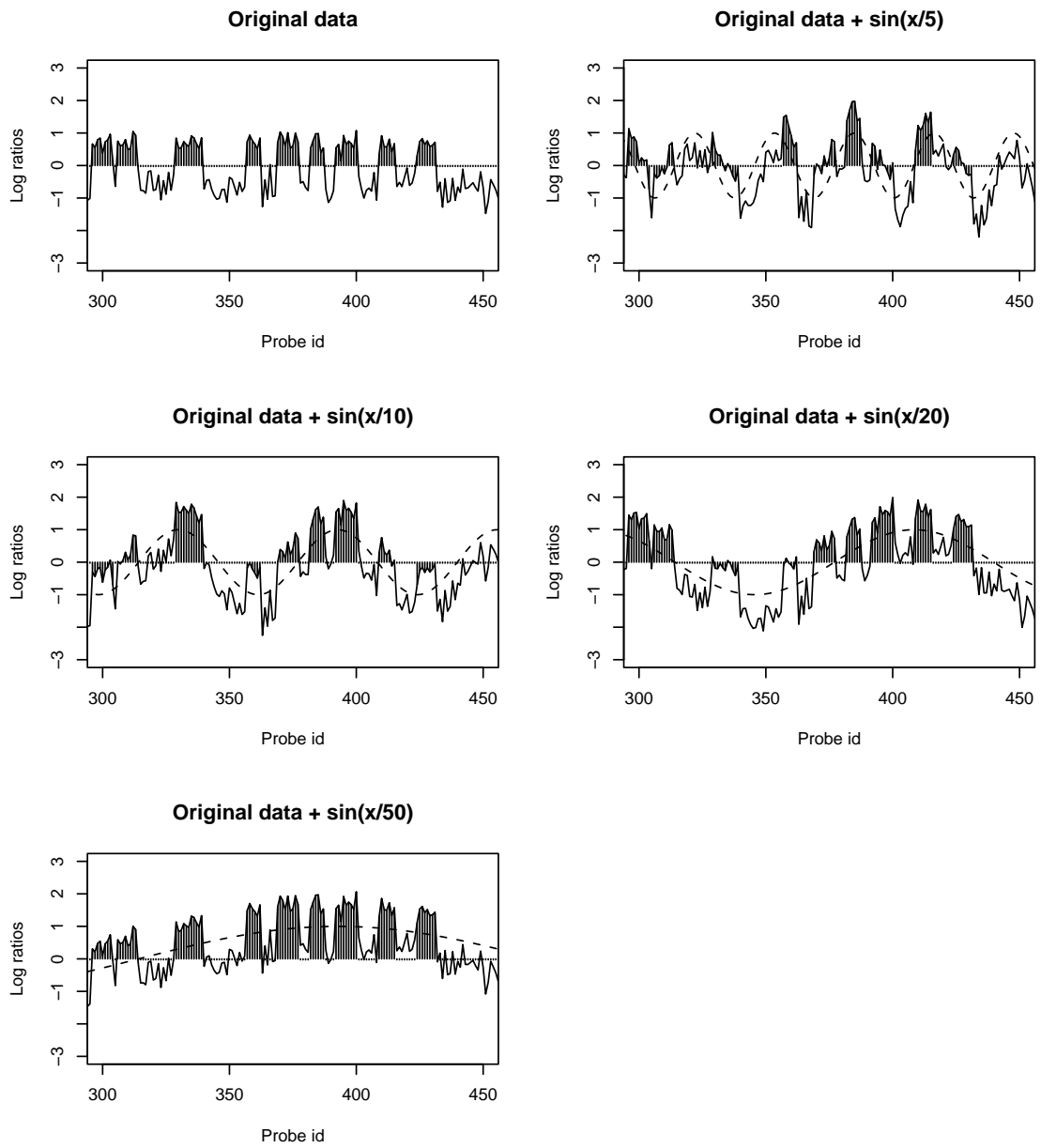


Figure 2: *Trend lines in the simulated data.* The periodicity of the sinusoidal trend lines is varied in each simulation scenario.

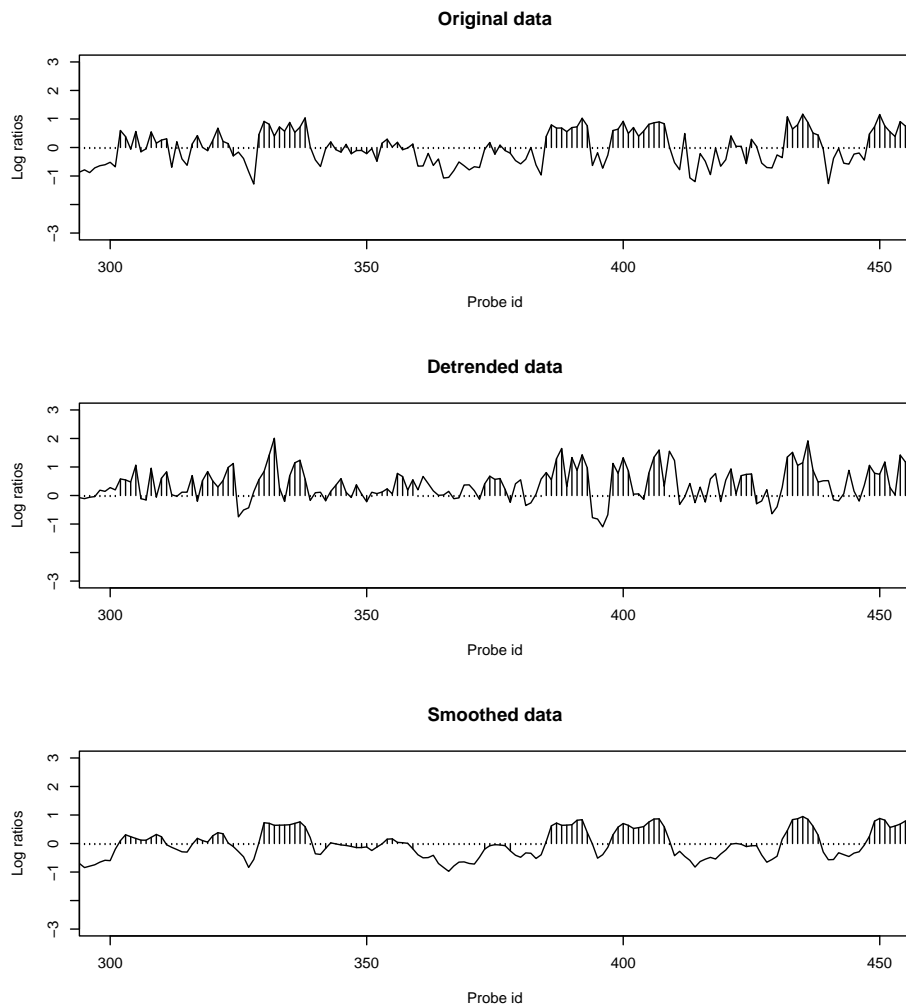


Figure 3: *An example of simulated data from Simulation II.* Middle panel is the data detrended by comparing peak and trough within a window size of 7 probes. Bottom panel is the smoothed data. Black vertical lines represent nucleosome probes.

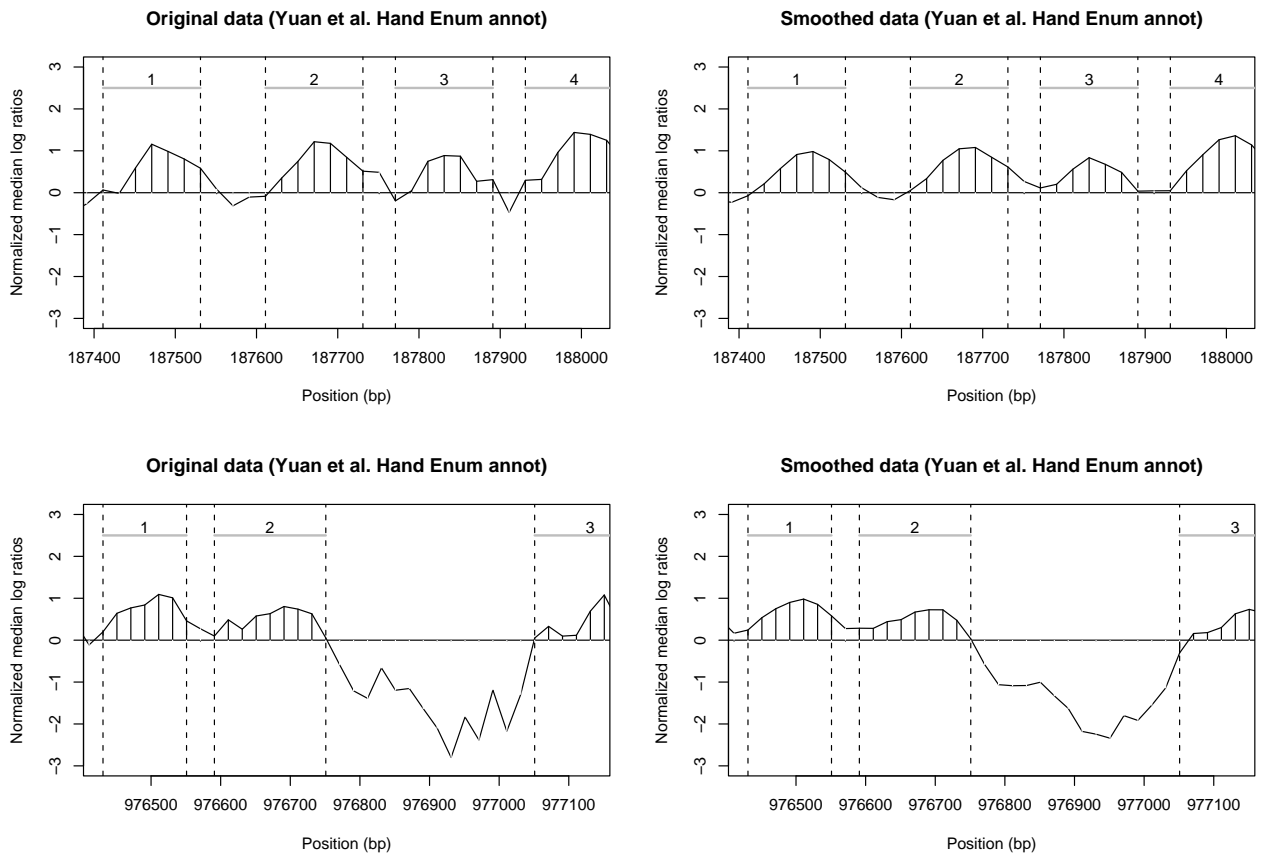


Figure 4: *Examples of “hand picked” annotations in Yuan et al. (2005).* The left panels are original data based on the “hand picked” annotations in Yuan et al. (2005) for two regions in chromosomes 5 and 7, respectively. The right panels are the smoothed data for similar regions. Although the “hand picked” nucleosomes are reliable, there are still some uncertainties in picking the boundaries of nucleosome-linker, for instance between nucleosomes 2 and 3 in the top panels and between nucleosomes 1 and 2 in the bottom panels.

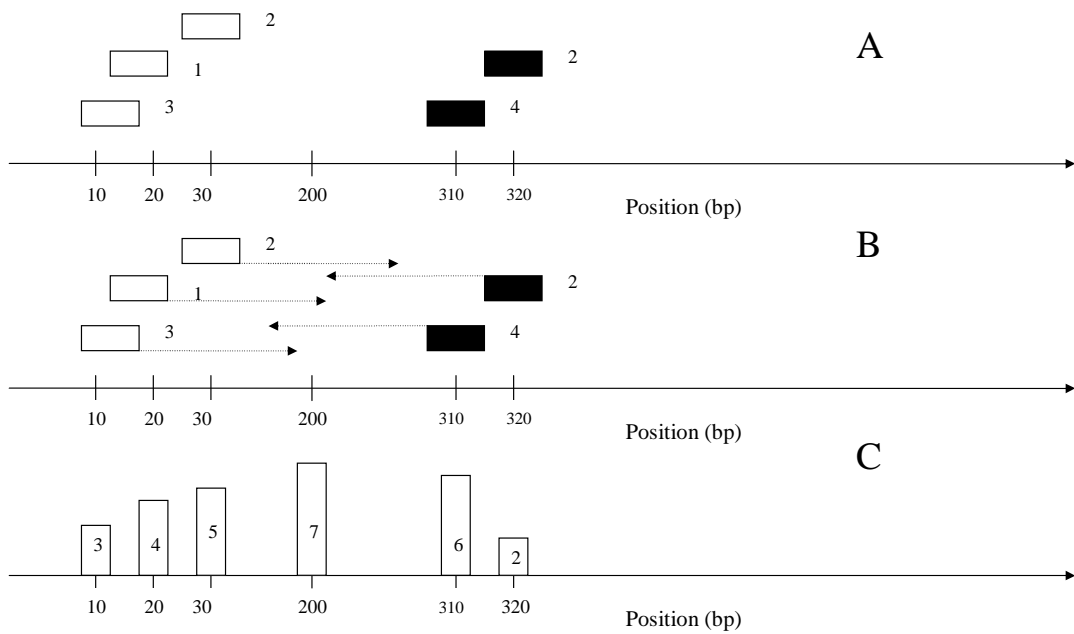


Figure 5: *Illustration of obtaining reads for each genomic position in ChIP-Seq data.* The white rectangles are reads mapped to the plus strand and the black rectangles are reads mapped to the minus strand. Panel B shows the extended reads (150 base pairs). Panel C shows the total read for each genomic position.