SUNY Korea
The State University of New York
한국뉴욕주립대학교

Stony Brook University

Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 10:
GAUSSIAN MIXTURE MODEL APPLICATION EXAMPLE: NUCLEOSOME POSITIONING

Reference:

Polishko, A., Ponts, N., Le Roch, K. G., & Lonardi, S. (2012). NORMAL: Accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics*, *28*(12), i242–9.

BIOINFORMATICS

Vol. 28 ISMB 2012, pages i242–i249
doi:10.1093/bioinformatics/bts206

NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model

Anton Polishko[1,*], Nadia Ponts[2,3], Karine G. Le Roch[2] and Stefano Lonardi[1,*]

[1]Department of Computer Science and Engineering, [2]Department of Cell Biology and Neuroscience, University of California, Riverside, CA 92521, USA and [3]INRA, MycSA UR1264, 71 Avenue Edouard Bourlaux, F-33883 Villenave d'Ornon Cedex, France

ABSTRACT

"**Motivation:** Nucleosomes are the basic elements of chromatin structure. They control the packaging of DNA and play a critical role in gene regulation by allowing physical access to transcription factors. The advent of second-generation sequencing has enabled landmark genome-wide studies of nucleosome positions for several model organisms. Current methods to determine nucleosome positioning first compute an occupancy coverage profile by mapping nucleosome-enriched sequenced reads to a reference genome; then, nucleosomes are placed according to the peaks of the coverage profile. These methods are quite accurate on placing isolated nucleosomes, but they do not properly handle more complex configurations. Also, they can only provide the positions of nucleosomes and their occupancy level, whereas it is very beneficial to supply molecular biologists additional information about nucleosomes like **the probability of placement**, the **size of DNA fragments enriched for nucleosomes** and/or whether **nucleosomes are well positioned or 'fuzzy'** in the sequenced cell sample.

Results: We address these issues by providing a novel method based on a parametric probabilistic model. An expectation maximization algorithm is used to infer the parameters of the **mixture of distributions.** We compare the performance of our method on two real datasets against Template Filtering, which is considered the current state-of-the-art. On synthetic data, we show that our method can resolve more accurately complex configurations of nucleosomes, and it is more robust to user-defined parameters. On real data, we show that our method detects a significantly higher number of nucleosomes."

background info& importance

Current methods and their limitations

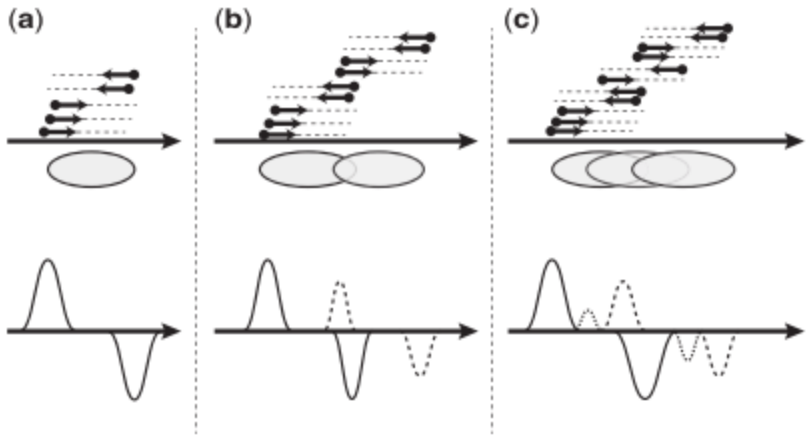
Problem Def.

Solution approach

Validation & Results

Background Info & Importance

Knowing the precise locations of nucleosomes in a genome is key to understanding how genes are regulated.



Nucleosome positions can tell us about

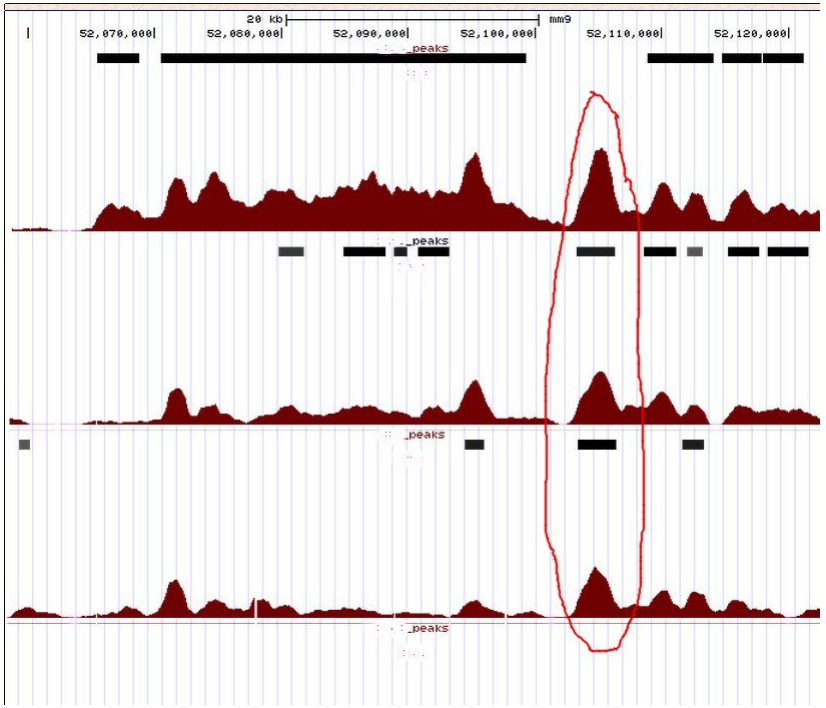
- How nucleosome positioning distinguish promoter regions and transcriptional start sites, and
- How the composition and structure of promoter nucleosomes facilitate or inhibit transcription.
- How diverse factors, including underlying DNA sequences and chromatin remodeling complexes, influence nucleosome positioning

Current Methods and Their Limitations

Peak calling

Problems of peak calling method

- The coverage profile function has to be cleaned of high-frequency noise,
  - typically via a kernel density estimation method
- peak finding algorithms have parameters (like the extension and the shift discussed above) that are difficult to optimize:
  - a set of parameter can work for a region of a chromosome but not for another
- peak calling do not properly resolve overlapping nucleosomes. F

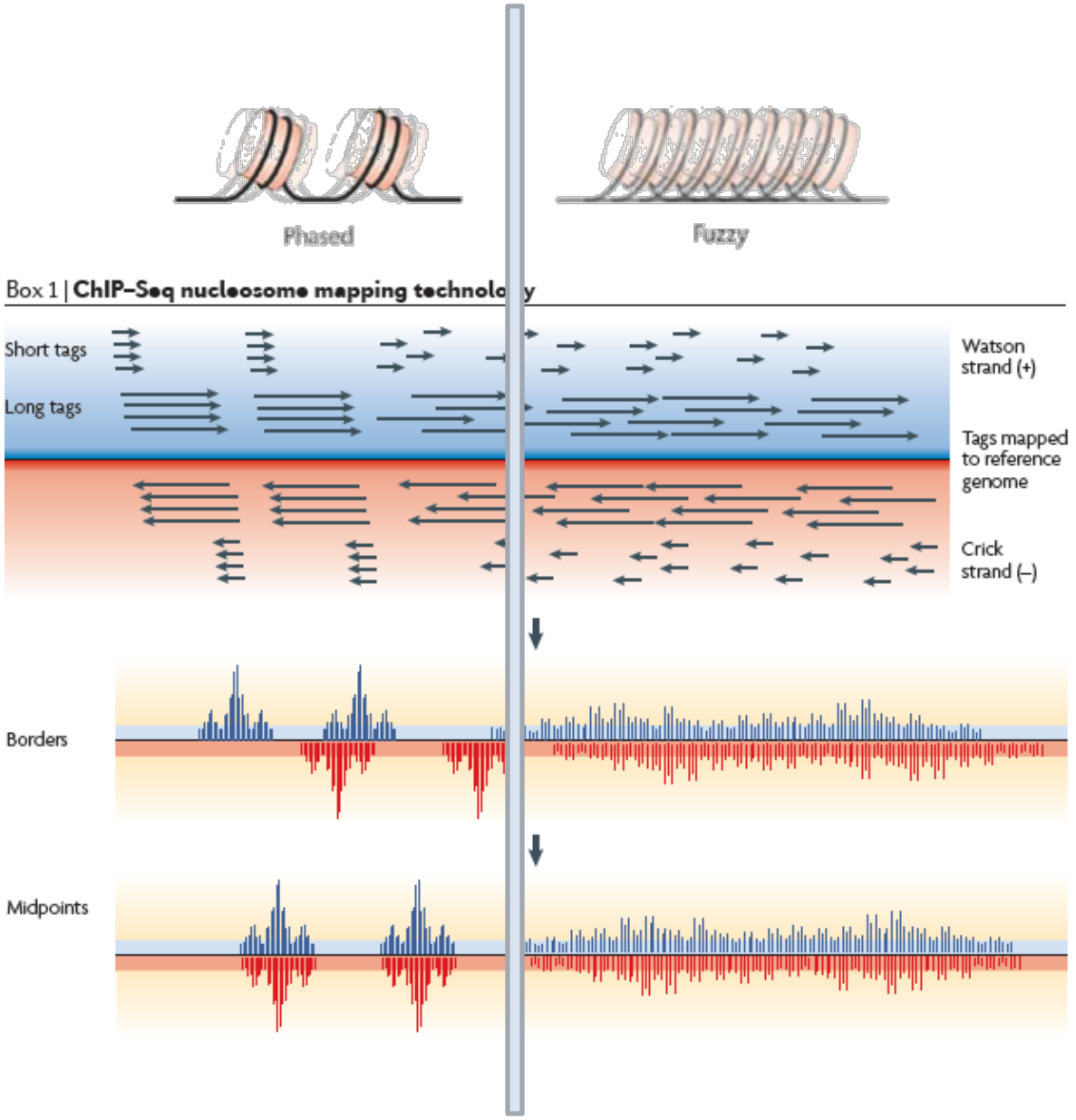


Problem Definition

Problem:

Determine nucleosome positioning and provide information about

- probability of position
- size of DNA fragments enriched for nucleosomes and/or
- whether nucleosomes are well positioned or 'fuzzy'



C. Jiang and B. F. Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* 10 161-172 (2009)

Solution Approach

NOrMAL: NucleOsome Mapping ALgorithm.

- Parametric probabilistic model for nucleosome positioning
- Uses Expectation Maximization (EM) to infer its parameters.

Conditions to consider

- Although DNA fragments, due to size of nucleosome, be around 146 bp, in reality they are shorter. The digestion process can either leave nucleosome-free DNA in the sample, or ‘over-digest’ the ends of nucleosome-bound DNA.
- The rate of digestion is sequence dependent so nucleosomes in different genomic locations will end up with different DNA fragment size.

Validation

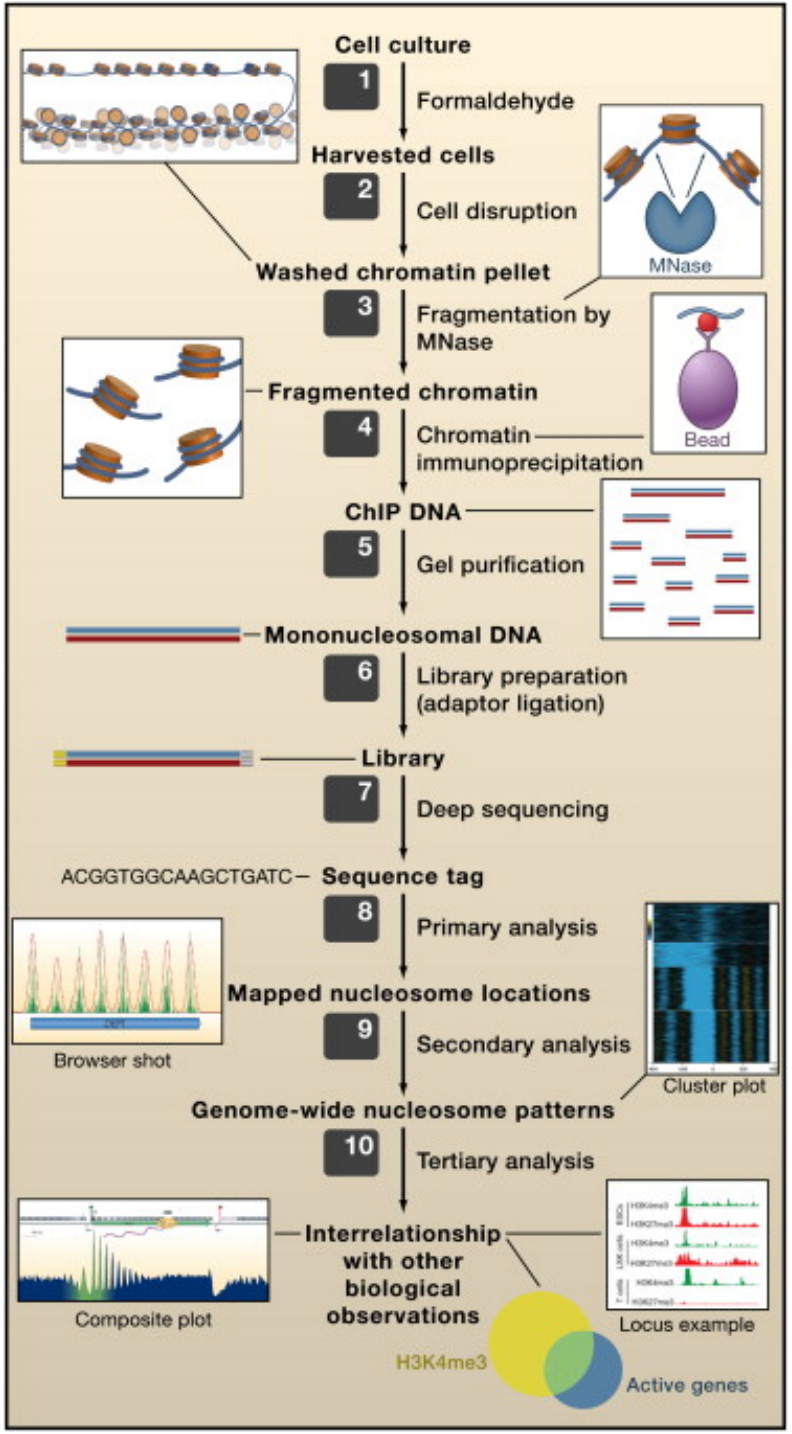
Compare the performance of NOrMAL against the state of the art method in 2011: Template Filtering (TF) algorithm

Dataset:

1. MAINE-seq data for Plasmodium falciparum (Ponts et al., 2010) and
2. MAINE-seq data for Saccharomyces cerevisiae (Weiner et al., 2010)

MAINE-seq data set: micrococcal nuclease digestion of mononucleosomes

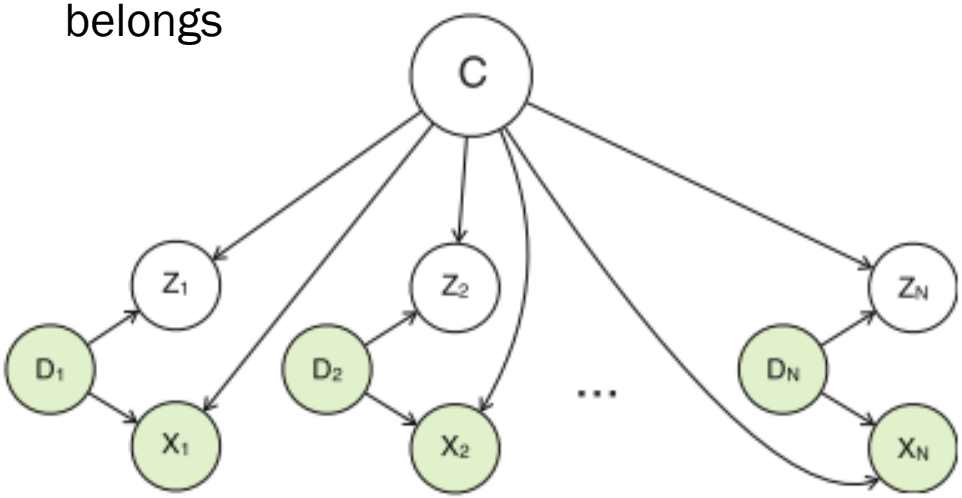
Isolates histone-associated DNA



Modified Gaussian mixture model framework: DATA

Data : i & Class: j

- $C_i \in [1, K]$: hidden variable representing the nucleosome to which it belongs



(Fig 2 of A. Polishk et al.) The proposed graphical mixture model: shaded nodes correspond to observed variables, white nodes correspond to hidden variables

- X_i, D_i, Z_i and M_i the random variables associated with variables x_i, d_i, z_i and m_i.

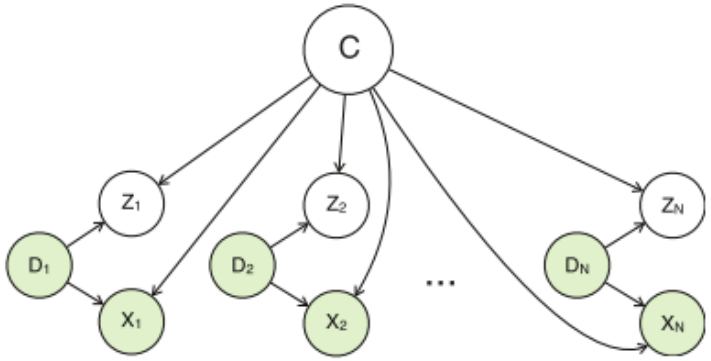
- i : any **DNA fragment** $i \in [1, N]$
- N: number of DNA fragments obtained after MNase digestion
- x_i: the starting position of the 5'-end of fragment i (obtained by mapping sequenced read)
- $d_i \in \{+1, -1\}$: the strand on which fragment i was mapped (+1 for the positive strand, and−1 for the negative strand).
- z_i: the length of fragment i. (can be observed directly only if sequencing produces paired-end reads)
- m_i: the position of the center of the fragment i,

$$m_i = x_i + \frac{d_i z_i}{2}$$

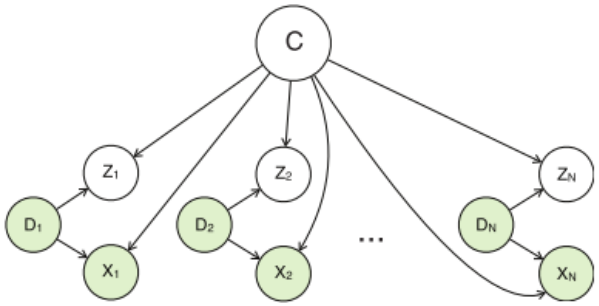
Modified Gaussian mixture model framework: CLASS

Each nucleosome $j \in [1, K]$ is described by set of **six variables**

$$\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$$
$$\theta_j = (\mu_j, \sigma_j, \Delta_j, \delta_j^{+1}, \delta_j^{-1}, \pi_j)$$



- μ_j : center position of the nucleosome j,
- σ_j : the fuzziness associated with the position of nucleosome j,
- Δ_j : describes the length of DNA fragments associated with nucleosome j, (z_i)
- $\delta_j^{+1}, \delta_j^{-1}$: variation on fragment site for positive and negative strands, respectively
- π_j : is the probability of nucleosome j.



* model **assumes** that our <u>random variables are distributed according to a **normal distribution**</u>

- assume that variable M_i associated with the center of the fragment i for a particular nucleosome j is distributed as follows

$$P(M_i|C_i = j, \boldsymbol{\theta}) \sim N(\mu_j, \sigma_j^2)$$

- assume that the length Z_i of fragment i for a particular nucleosome j is distributed as follows

$$P(Z_i|D_i = d_i, C_i = j, \boldsymbol{\theta}) \sim N(\Delta_j, \left(\delta_j^{d_i}\right)^2)$$

rule of linear combination of independent Gaussians

- Assuming the above Gaussians are independent and applying

$$x_i = m_i - \frac{d_i z_i}{2}$$

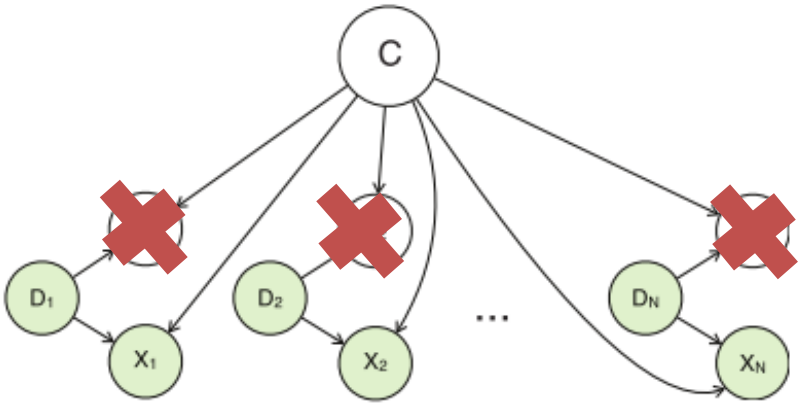
Probability of a given data point xi given the parameters of a nucleosome

$$P(X_i|D_i = d_i, C_i = j, \boldsymbol{\theta}) \sim N\left(\mu_j - \frac{d_i \Delta_j}{2}, \sigma_j^2 + \left(\frac{\delta_j^{d_i}}{2}\right)^2\right)$$

excluded variables zi from the equation.

Generative Mixture Model

$$P(X_i|D_i = d_i, C_i = j, \boldsymbol{\theta}) \sim N\left(\mu_j - \frac{d_i\Delta_j}{2}, \sigma_j^2 + \left(\frac{\delta_j^{d_i}}{2}\right)^2\right)$$



- Likelihood of point (x_i, d_i) given $\boldsymbol{\theta}$

$$P(X_i|D_i = d_i, \boldsymbol{\theta}) = \sum_{j=1}^{K} P(C_i = j|\boldsymbol{\theta}) P\left(X_i \middle| D_j = d_i, C_i = j, \boldsymbol{\theta}\right)$$

$$= \sum_{j=1}^{K} \pi_j f\left(x_i, \mu_j - \frac{d_i\Delta_j}{2}, \sigma_j^2 + \left(\frac{\delta_j^{d_i}}{2}\right)^2\right)$$

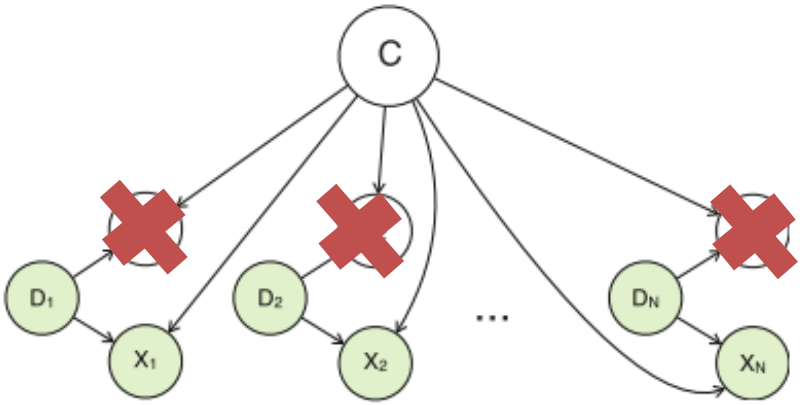
Where $f\ (x, a, b) = \frac{1}{\sqrt{2\pi b}} \exp\left(-\frac{(x-a)^2}{2b}\right)$

marginal log-likelihood of observed data

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}\,|\boldsymbol{\theta})\}$$

Expectation of the complete-data log likelihood

$$\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{z} p(\mathbf{Z}\,|\mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}}) \ln p(\mathbf{X}, \mathbf{Z}\,|\boldsymbol{\theta})$$



- log likelihood of observed data points X given parameters □

$$l(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} P(X_i|D_i = d_i, \boldsymbol{\theta})$$

$$= \sum_{i=1}^{N} ln \sum_{j=1}^{K} P(C_i = j|\boldsymbol{\theta}) P\left(X_i \middle| D_j = d_i, C_i = j, \boldsymbol{\theta}\right)$$

$$= \sum_{i=1}^{N} ln \sum_{j=1}^{K} \pi_j f\left(x_i, \mu_j - \frac{d_i \Delta_j}{2}, \sigma_j^2 + \left(\frac{\delta_j^{d_i}}{2}\right)^2\right)$$

Complete data log likelihood

- estimate the parameters of the model $\boldsymbol{\theta}$ via maximum likelihood□

$$\widehat{\boldsymbol{\theta}} = argmax_{\theta} l(X|\boldsymbol{\theta})$$

the posterior distributor of the latent variables

E step

Compute posterior probabilities $P(C_i = j|X_i = x_i; \boldsymbol{\theta}^t)$ of data points $x_i, i \in [1, N]$ w.r.t the distribution of C_i given the current estimate of parameters $\boldsymbol{\theta^t}$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta^t}) = E_{C|X,\boldsymbol{\theta^t}} l(X|\boldsymbol{\theta})$$

$$= \sum_{\boldsymbol{C}} p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}^{\mathbf{t}}) \ln p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})$$

marginal log-likelihood of observed data

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\}$$

Expectation of the complete-data log likelihood

$$\mathrm{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{z} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathbf{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

M step

Find new parameter estimation

$$\square \boldsymbol{\theta}^{t+1} = argmax_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta^t})$$

NOrMAL Algorithm

Initialize with **non-informative prior**

1: {Parameter initialization}
2: $\mu_0 \leftarrow (\mu_1, \mu_2, \ldots, \mu_K)$, where μ_i is uniformly distributed
3: $\pi_0 \leftarrow (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}) \in \mathbb{R}^K$
4: $\Theta^{(t)} \leftarrow (\mu_0, \sigma_0, \Delta_0, \delta_0^{+1}, \delta_0^{-1}, \pi_0)$
5: {Soft Learning}
6: $\Theta \leftarrow Learn(\Theta^{(t)})$
7: {Hard learning}
8: **for all** $i \in [1, N]$ **do**
9: $C_i \leftarrow argmax_j(T_{ij})$
10: **end for**
11: Recompute cluster parameters Θ
12: **return** Θ

1. Heuristics to find **number** of nucleosome positions (clusters)
2. Rough estimate of the nucleosome **positions**.

Learn($\boldsymbol{\theta}$)

E:
M:
Learn K

1: **repeat**
2: **while** not converged **do**
3: $\mathbb{Q}(\Theta|\Theta^{(t)}) \leftarrow E_{C|X,\Theta^{(t)}} l(X|\Theta)$
4: $\Theta^{(t+1)} \leftarrow argmax_\Theta \mathbb{Q}(\Theta|\Theta^{(t)})$
5: $t \leftarrow t + 1$
6: **end while**
7: **for all** $j \in [1, K-1]$ **do**
8: **if** $|\mu_j - \mu_{j+1}| \leq threshold$ **then**
9: {Merge clusters i and $i-1$}
10: **end if**
11: **end for**
12: **until** no clusters were merged
13: **return** $\Theta^{(t)}$

One iteration of **'hard learning'** assigning each data point x_i its maximum probable cluster

Determining the number of Gaussian components

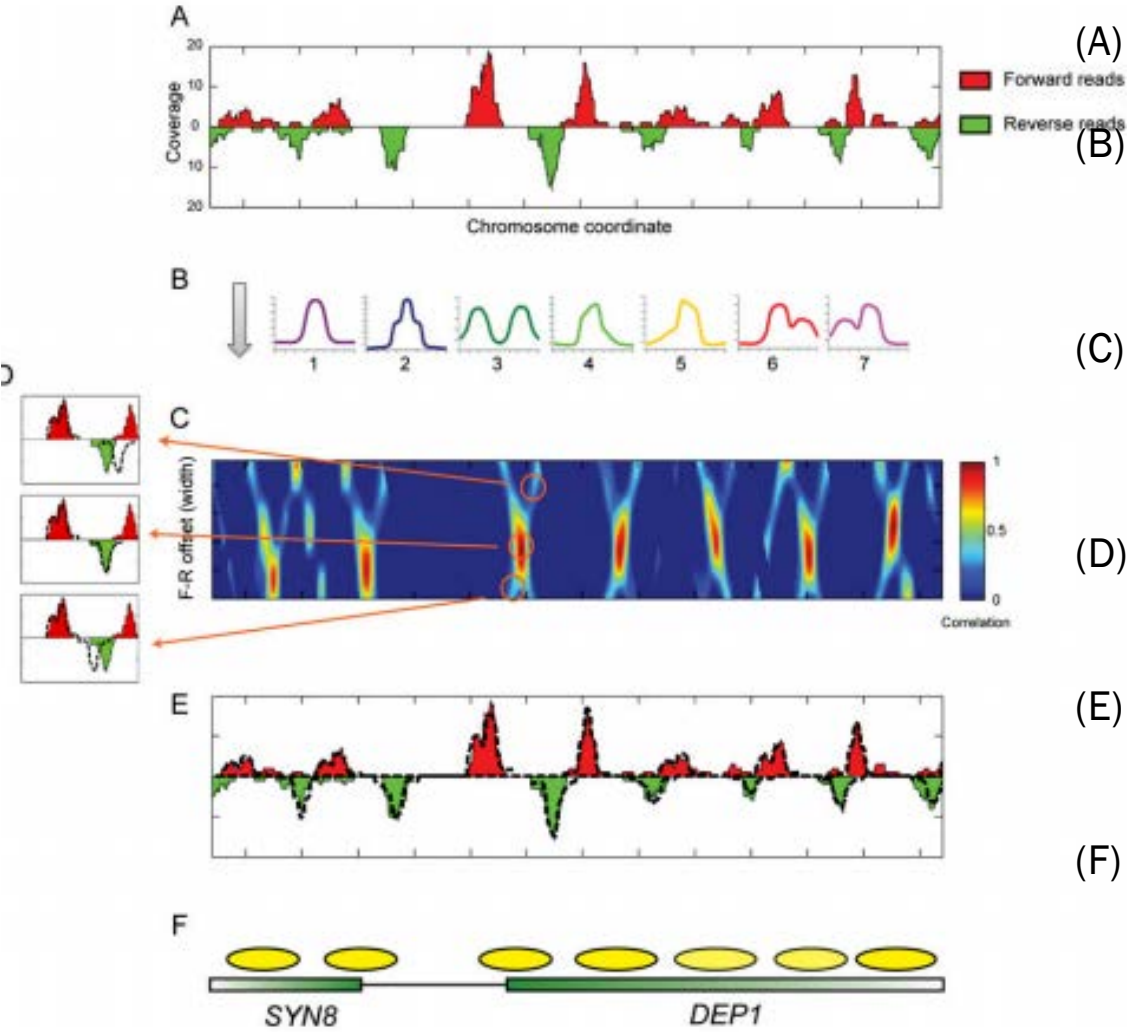
Heuristics to find K:

1. Initialize: Placing the maximum possible number of non-overlapping nucleosomes uniformly distributed on the chromosome,
   - $K = \frac{size\ of\ the\ chromosome(\approx ****bp)}{expected\ size\ of\ a\ nucleosome\ (\approx 146bp)}$

   where the expected size of nucleosomes is underestimated.
2. 'soft learning': we run our EM algorithm until convergence
3. merge those that have too much overlap (distance under a user-specified threshold).
   - In case of multiple overlaps for a nucleosome, we merge it with the closest one.
4. Repeat (2) and (3) until no additional clusters are merged.

- After a few cycles, we will obtain a set of non-overlapping clusters that best explain the given data points.
- Overlapping nucleosomes are merged into new ones and then the position of new nucleosomes are learned from the data.

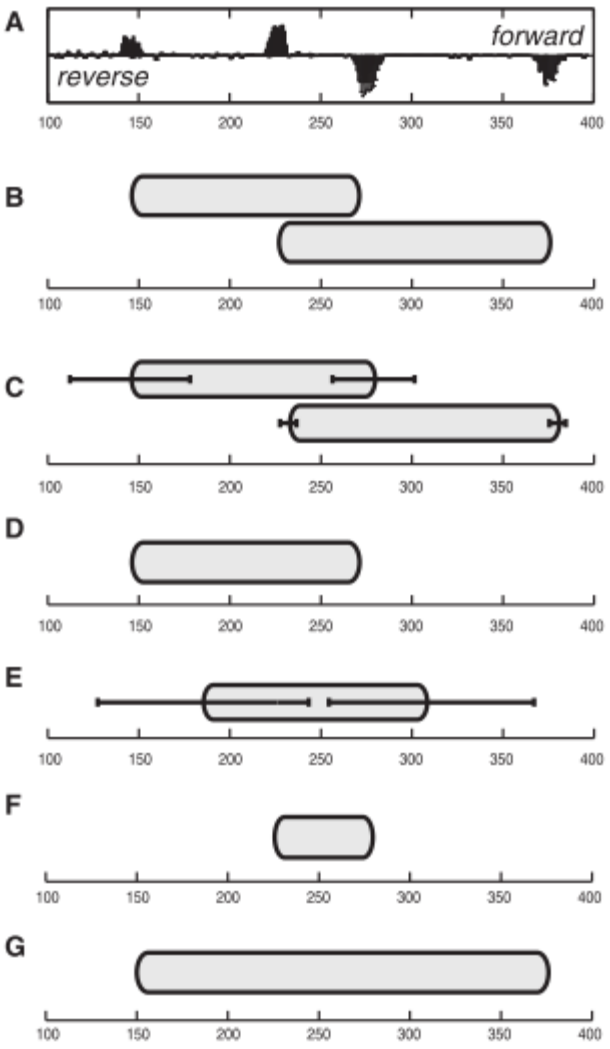
Template filtering overview:



(A) Deep sequencing data for a typical stretch of the yeast genome.
(B) Templates. Forward and reverse-strand read distributions are cross-correlated with each of the seven templates shown.
(C) Correlation coefficient heat map of template 1 for forward and reverse templates at varying center positions (x-axis) and distances (y-axis).
(D) Examples of templates spaced too far apart (top), at the optimal distance (middle), or too close together (bottom).
(E) Read distributions explained by the optimal template matches are shown as dotted lines for the region in A.
(F) Schematic of nucleosome calls and underlying gene annotations.

Weiner, A., Hughes, A., Yassour, M., Rando, O. J., & Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research*, *20*(1), 90–100.

Results: Synthetic data



Coverage profile from synthetic data

nucleosomes detected using TF allowing maximum 35% overlap [100,200]

NOrMAL allowing maximum 35% overlap (C)

TF allowing only 30% overlap, [100,200] (D)

NOrMAL allowing only 30% overlap (E);

nucleosome reported by TF with nucleosome size range [40,200] (F)

nucleosome reported by TF with nucleosome size range [100,300] (G)

Results: Real data

Like many bioinformatics problems, challenge for nucleosome position inference is that the true positions of the nucleosomes are unknown.

-> The author's have to do their best to convince that reader that their method is better.

number of nucleosome detected by TF and NOrMAL and corresponding execution time

Chromosome	No of mapped reads	TF	Time (s)	NOrMAL	Time (s)
1	16 688	1 033	1.38	**1 078**	6.86
2	78 543	4 284	7.37	**4 394**	84.56
3	30 589	1 583	4.43	**1 618**	8.49
4	138 801	7 975	16.36	**8 014**	369.11
5	55 601	2 986	4.02	**3 101**	38.80
6	26 141	1 403	1.63	**1 453**	4.45
7	101 981	5 727	9.84	**5 817**	126.34

1.Validation by agreement between exiting state of the art method

2. Validation by Empirical example