Lecture 6 Mar 7 Characterizing Protein-DNA interactions using ChIP-seq and Motif Disc.

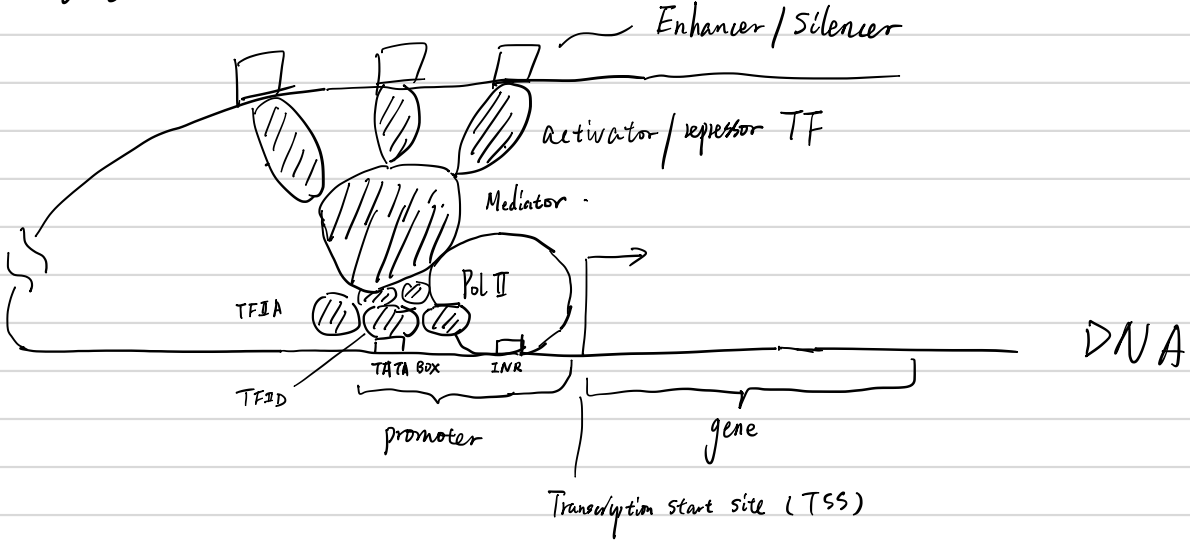
Agenda

- Landscape of gene regulation
- Assay in vivo TF binding via ChIP-seq
  - Binding events determination } EM.
  - Motif discovery

Gene Regulation

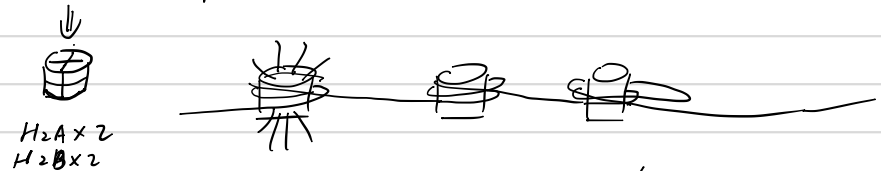
- Gene < 1% of the genome
- Different cells share the same DNA sequence
- Most of disease-associated mutations reside in non-coding region

} 99% of the genome regulate the expression of gene in a cell-type specific way.



- Transcription factors
  - Regulates gene regulation
  - Binding has sequence specificity

- Histone modification

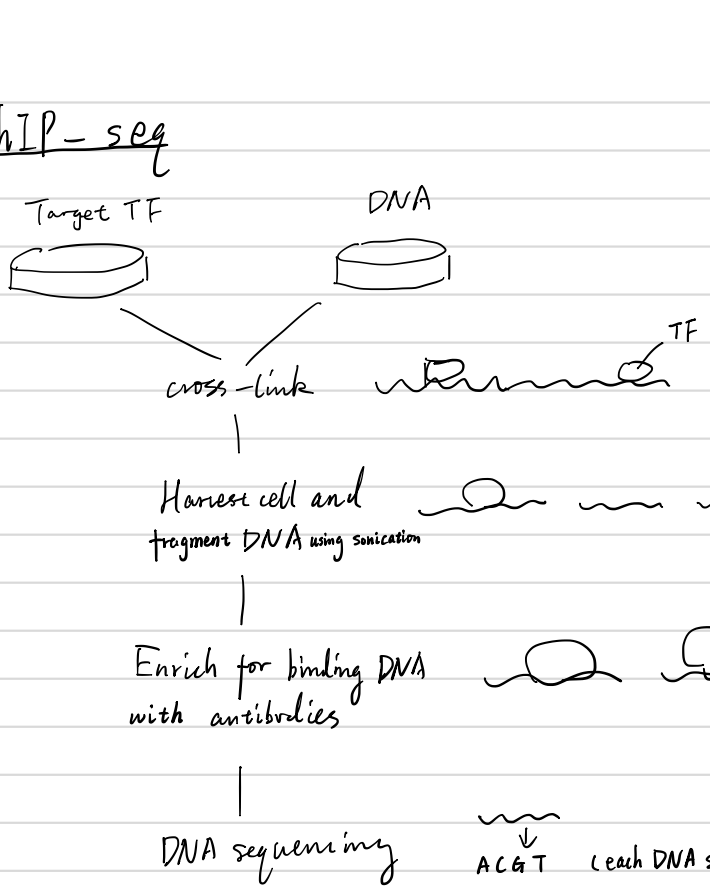


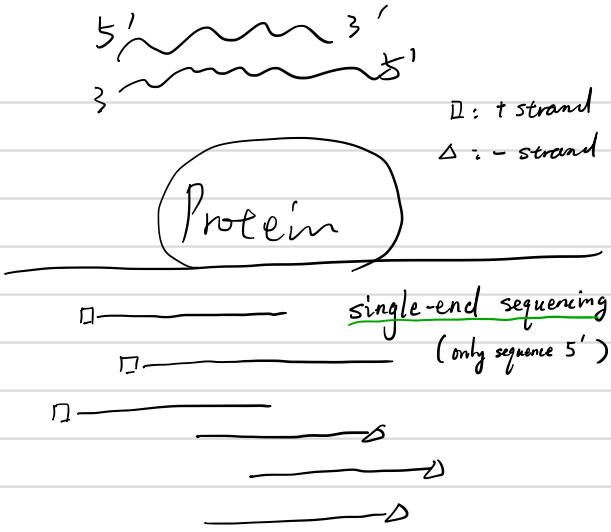
H_3K4me3: tri-methylation on the 4th lysine of H3 ⇒ promoters

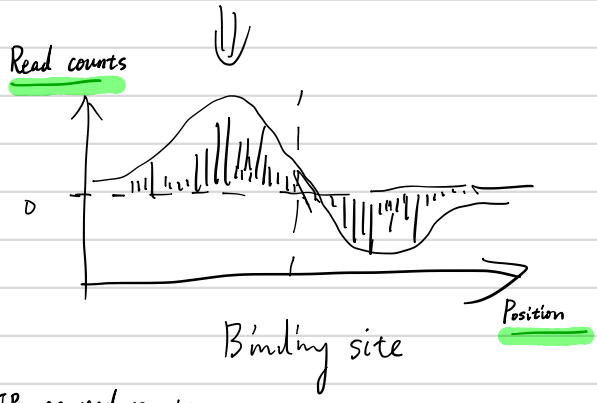
H_3K27ac: acetylation on the 27th lysine of H3 ⇒ active regulatory elements

- CTCF and cohesins

ChIP-seq



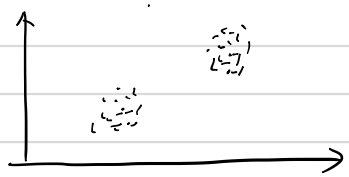




Computation problem

- Predict binding events : predict where the TF bind from the noisy ChIP-seq read counts
- Motif discovery : given a set of binding sequences (e.g 10 base pair around the predicted binding sites), model the sequence specificity (sequence "motif"

Expectation – Maximization (Mixture Model as example)



Goal: cluster the points into M groups

Assumption: (Gaussian Mixture) Points in the ith group distribute $\sim N(\mu_i, \Sigma_i) \triangleq N(\theta_i)$

Notation:
- $x_1, x_2 \cdots x_n$ samples $\in R^2$
- $z_1, z_2, \cdots z_n$ group assignments $\in \{1, 2, \cdots M\}$, M = # of groups
- θ_1, θ_2 gaussian parameters
- π_1, π_2 prior prob. of each group $\pi_k = p(z_i = k)$, $\forall i$

Likelihood of observ $p(x_i; \theta, \pi) = \sum_{k=1}^{M} p(z_i=k)\, p(x_i \mid z_i=k; \pi, \theta) = \sum_{k=1}^{M} \pi_k N(x_i; \theta_k)$

$$\Rightarrow \ln p(X; \theta, \pi) = \ln \prod_{i=1}^{n} p(x_i; \theta, \pi) = \sum_{i=1}^{n} \ln \sum_{k=1}^{M} \pi_k N(x_i; \theta_k)$$

$$\Rightarrow \theta^*, \pi^* = \underset{\theta, \pi}{\arg\max} \underbrace{\sum_{i=1}^{n} \ln \sum_{k=1}^{M} \pi_k N(x_i; \theta_k)}_{\text{hard to optimize}}$$

Instead we maximize $\ln p(X, Z; \theta, \pi)$

↓ unknown! → First estimate $p(Z|X;\theta,\pi)$ and then use $E_{p(Z|X;\theta,\pi)}(\ln p(X,Z;\theta,\pi))$ to approximate

Expection step: $r(Z) \triangleq p(Z|X;\theta,\pi) = \prod_{i=1}^{n} p(z_i|x_i;\theta,\pi)$

$r(z_i=k) \triangleq p(z_i=k|x_i;\theta,\pi) \propto p(x_i|z_i=k;\theta,\pi)\, p(z_i=k) = N(x_i,\theta_k)\,\pi_k$ (Bayes' Rule)

$\Rightarrow r(z_i=k) = \dfrac{N(x_i,\theta_k)\,\pi_k}{\sum_k N(x_i,\theta_k)\,\pi_k}$ $\Rightarrow$ (Soft assignment of point i to group k)

Maximization step:

$$\theta^*, \pi^* = \underset{\theta,\pi}{\operatorname{argmax}}\; E_{p(Z|X;\theta,\pi)}(\ln p(X,Z;\theta,\pi))$$

$$= \underset{\theta,\pi}{\operatorname{argmax}} \sum_Z r(Z) \ln p(X,Z;\theta,\pi) \Rightarrow \prod_{i=1}^{n} N(x_i,\theta_{z_i})\,\pi_{z_i}$$

$$= \underset{\theta,\pi}{\operatorname{argmax}} \sum_Z r(Z) \sum_{i=1}^{n} \left[\ln N(x_i,\theta_{z_i}) + \ln \pi_{z_i}\right]$$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{L}$

Set $\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \Sigma} = \frac{\partial L}{\partial \pi} = 0 \Rightarrow$

$$\mu_k^{new} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma(z_i=k)\, x_i$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_i=k)(x_i-\mu_k)(x_i-\mu_k)^T$$

$$\pi_k^{new} = \frac{N_k}{n}$$

$$N_k = \sum_{i=1}^{n} \gamma(z_i=k)$$ (total # of points assigned to kth group)

Binding event prediction

Genome Positioning System (GPS) Guo et al 2010 Bioinformatics

$$\begin{cases} X \rightarrow \text{the position of each read} \\ Z \rightarrow \text{the assignment of each read to bind events} \end{cases}$$

Special design:

- The property of ChIP-seq: ① Binding events generate reads in a range ⇒ segment the genome and perform EM separately
  ⇒ in each segment, there is one event on every base pair
  ② Shape of event is the same for any event ⇒ Instead of $N(\mu_i, \theta_i)$ for each event, use emp. distr estimated from genome-wide rough peak-calling (using other software)
  ⇒ After performing EM on each segment separately, re-estimate using the updated event position
- We want π_k to be sparse
  → negative Dirichlet prior on π_k $p(\pi) \sim \prod_k \frac{1}{(\pi_k)^\alpha}$, $\alpha > 0$
  → $\pi_k^{new} = \frac{N_k - \alpha}{\sum_{j=1}^{M}(N_j - \alpha)}$ $\xRightarrow{\text{component elimination}}$ Figueiredo and Jain 2002 $\pi_k^{new} = \frac{\max(0, N_k - \alpha)}{\sum_{j=1}^{M}\max(0, N_k - \alpha)}$

  $\alpha \Rightarrow$ the minimum # of reads an event needs to survive EM iterations

Motif discovery

- PWM (position weight matrices)

	1	2	3	4
A	0.1	0.4	0.1	0.1
C	0.3	0.1	0.1	0.4
G	0.4	0.2	0.7	0.4
T	0.2	0.3	0.1	0.1

- Aligned seqs → PWM is easy

AACT
AGCT ⇒
CTCA

	1	2	3	4
A	0.66	0.33	0	0.33
C	0.33	0	1	0
G	0	0.33	0	0
T	0	0.33	0	0.66

Goal:

Given a set of binding sequences (same length, not aligned), find a PWM (of length W) that describe the binding motif

$$\begin{cases} X \rightarrow \text{sequences} \\ Z \rightarrow \text{the starting position of each seq} \end{cases}$$

$p(X_i \mid Z_i = k)$ is parametered by motif PBM $M \in R^{4\times W}$ and background PBM $B \in R^{4\times 1}$

$$\Rightarrow p(X_i \mid Z_i = k; M, B) = \prod_{j=1}^{k-1} B(X_i^j) \prod_{j=k}^{k+W} M(j-k+1, X_i^j) \prod_{j=k+W}^{S(X_i)} B(X_i^j) \quad (1)$$

↓ The prob. that the motif sequence in X_i starts at the kth position

E step: $p(Z \mid X; M, B) = \prod_{i=1}^{n} p(Z_i \mid X_i; M, B) \triangleq \prod_{i=1}^{n} \gamma(Z_i)$

$$\Rightarrow \gamma(Z_i = k) \sim p(X_i \mid Z_i = k; M, B)\, p(Z_i = k)$$

⇒ If we assume $p(Z_i)$ is uniform (all the starting position are equally likely)

$$\gamma(Z_i = k) = \frac{\prod_{j=1}^{k-1} B \prod_{j=k}^{k+W-1} M \prod_{j=k+W}^{S(X_i)} B}{\sum_{k=1}^{S(X_i)-L+1} \pi B \; \pi M \; \pi B}$$

$$L$$

M step: $M^*, B^* = \underset{M,B}{\operatorname{argmax}}\ E_{p(Z|X;M,B)}\left(\ln p(X,Z;M,B)\right) + \sum_{l=1}^{W}\lambda_l\left(\sum_c M(l,c)-1\right)$

$\ln p(X,Z;M,B) \Rightarrow p(X|Z)\,p(Z) = \pi_B \pi_M \pi_B \times p(Z)$

$$\frac{\partial L}{\partial M} = \frac{\partial L}{\partial B} = 0$$

$$\Rightarrow M^{new}(k,c) = \frac{n_{k,c}+d}{\sum_c (n_{k,c}+d)}, \quad n_{k,c} = \sum_i \sum_{\{Z_i | X_i^{Z_i+k-1}=c\}} p(Z_i|X_i)$$

For all sequences and all the possible starting position, sum the prob. that the k-th nucleotide in the motif is c

d is a pseudo-count ↓ reflects how much we believe a priori that PWM should be uniform

$$B^{new}(c) = \frac{n_{0,c}+d}{\sum_c (n_{0,c}+d)}, \quad n_{0,c} = n_c - \sum_k n_{k,c}$$

Example of M step:

Assume $W=3$, after E step we have the following $p(Z|X)$ (prob. of motif starting position in each sequence)

X_1 =	A	C	A	G	C	A
$p(Z_1\|X_1)$	0.1	0.7	0.1	0.1		

X_2 =	A	G	G	C	A	G
$p(Z_2\|X_2)$	0.4	0.1	0.1	0.4		

X_3 =	T	C	A	G	T	C
$p(Z_3\|X_3)$	0.2	0.6	0.1	0.1		

⇒ M step

$n_{1,A} = 0.1 + 0.1 + 0.4 + 0.1 = 0.7$

$n_{1,C} = 0.7 + 0.4 + 0.6 = 1.7$

$n_{1,G} = 0.4$

$n_{1,T} = 0.2$

$\sum_c n_{k,c} = 3$

⇒ Let $d=1$, $M(1,A) = \frac{0.7+1}{3+4} = 0.24$

$M(1,C) = (1.7+1)/(3+4) = 0.39$

$M(1,G) = (0.4+1)/(3+4) = 0.2$

$M(1,T) = (0.2+1)/(3+4) = 0.17$

Similarly we can calculate the rest of M^{new}

$M^{new}(k,c)$ =

	1	2	3
A	0.24	0.39	0.21
C	0.39	0.21	0.18
G	0.2	0.24	0.44
T	0.17	0.16	0.16