Recitation 3

Transcriptional regulation/ChIP-Seq, TF binding/EM

Ge (Saber) Liu

MIT - 6.802 / 6.874 / 20.390 / 20.490 / HST.506 - Spring 2019

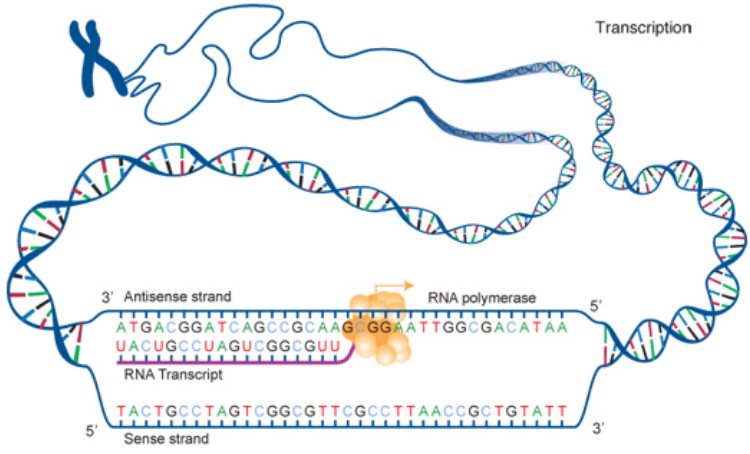
2019-02-21 / 2019-02-22

Outline

- Transcriptional regulation
- ChIP-Seq
- Motifs, Position weight matrix (PWM)
- TF binding prediction
- Expectation Maximization (EM)
- Mixture Model

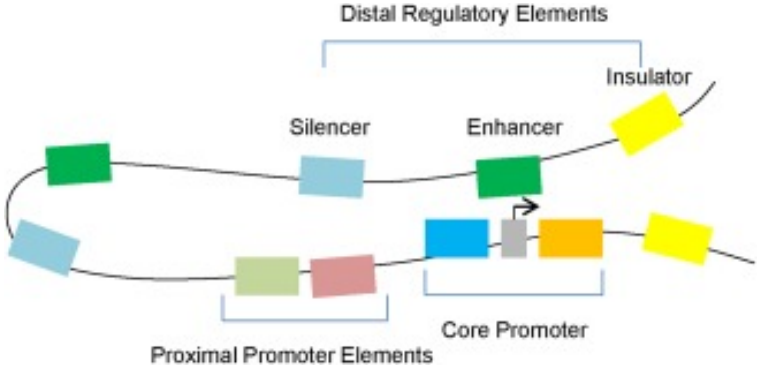
Transcriptional regulation

Transcription : first step of gene expression, in which a particular segment of DNA is copied into RNA (especially mRNA) by the enzyme RNA polymerase.



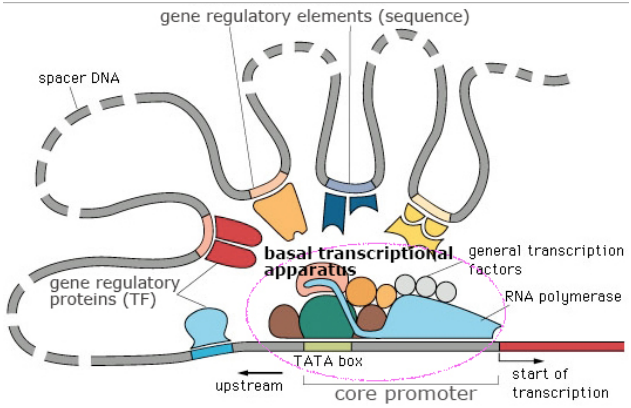
Transcriptional regulation

Transcription : first step of gene expression, in which a particular segment of DNA is copied into RNA (especially mRNA) by the enzyme RNA polymerase.
Transcriptional regulatory elements : nucleotide sequences of a gene that are involved in regulation of genetic transcription (e.g. enhancer/silencer, promoter, etc.), part of the non-coding DNA.



Transcription factors(TF): proteins that bind to specific DNA sequences in order to regulate the expression of a given gene, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.

- **General transcription factor(GTF, basal transcriptional factors):** a class of TF that is necessary for transcription to occur. Many of these GTFs do not bind DNA, but rather constitute a basal transcriptional apparatus with RNA polymerase and the mediator, which bind to the core promoter and start transcription.



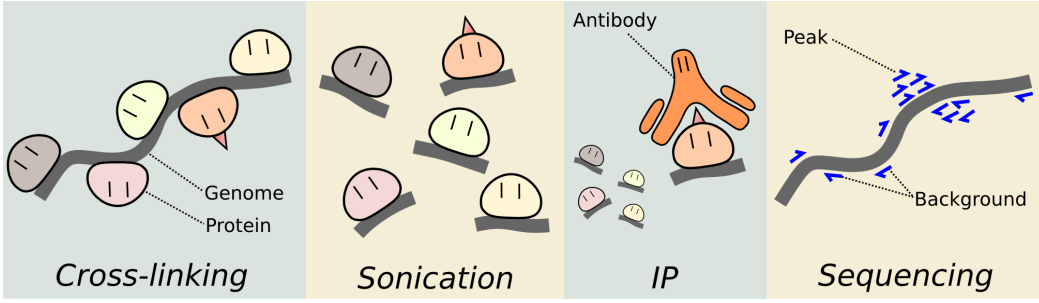
Transcriptional regulation

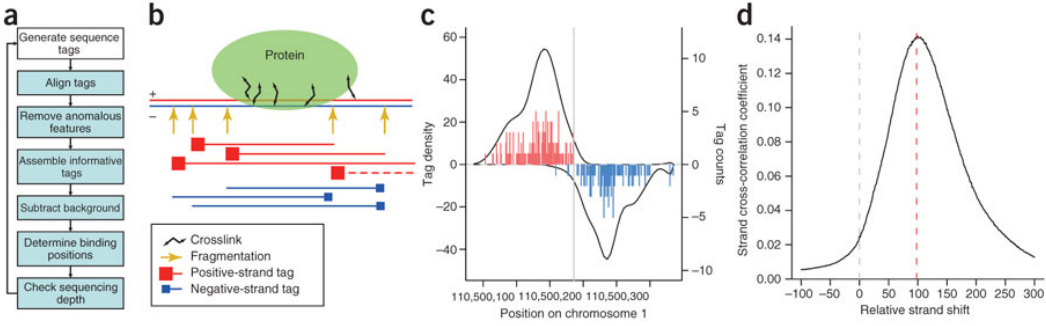
Transcription factors(TF): proteins that bind to specific DNA sequences in order to regulate the expression of a given gene, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.

- **General transcription factor (GTF, basal transcriptional factors):** a class of TF that is necessary for transcription to occur. Many of these GTFs do not bind DNA, but rather constitute a basal transcriptional apparatus with RNA polymerase and the mediator, which bind to the core promoter and start transcription.
- Other transcription factors: activators/repressors that bind to other regulatory elements.

Promoter: DNA region that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the same strand and upstream on the DNA. Eukaryotic promoters are diverse: Core promoter (minimal portion of the promoter required to properly initiate transcription), Proximal promoter, Distal promoter.
Enhancer: short DNA region that can be bound by activators to increase the likelihood of transcription of a particular gene.
Silencer: DNA sequence capable of binding repressors to inhibit the transcription.

ChIP-Seq


Genome
Protein
Cross-linking
Sonication
Antibody
IP
Peak
Background
Sequencing


a
Generate sequence tags
Align tags
Remove anomalous features
Assemble informative tags
Subtract background
Determine binding positions
Check sequencing depth
b
Protein
+
−
Crosslink
Fragmentation
Positive-strand tag
Negative-strand tag
c
Tag density
Tag counts
110,500,100 110,500,200 110,500,300
Position on chromosome 1
d
Strand cross-correlation coefficient
Relative strand shift

Motifs

Sequence motif is an amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. It is usually represented by a position weight matrix (PWM) and visualized using Seqeuence Logo.

position probability matrix: Nucleotide count at each position divided by the number of sequences in the alignment.

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

position frequency matrix

$$M_{k,j} = \sum_{i=1}^{N} I(X_{i,j} = k)$$

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

position probability matrix

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = k)$$

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

Both PPMs and PWMs assume statistical **independence** between positions in the pattern, as the probabilities for each position are calculated independently of other positions.

Position weight matrix (PWM): To get a PWM, we compare the PPM to a background frequency model b (usually uniform if not specified) and then take the log of the ratio to get the log-odds of observing nucleotide k at position j. And the log-odds of a sequence S is simply the sum of the log-odds of each nucleotide of the sequence at correponding location.

$$PWM_{k,j} = \textit{log-odds}_{(S_j=k)} = \log_2(PPM_{k,j}/b_k),$$
$$P(S|M) = \sum_j PWM_{S_j,j}$$

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

position weight matrix

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}$$

Information content: A measurement of how different a given PWM is from a uniform distribution.

$$IC = -\sum_{k,j} p_{k,j} * log(p_{k,j}/b_k)$$

Sequence logo is the most commonly used visualization of PWM, where the total height at position j is

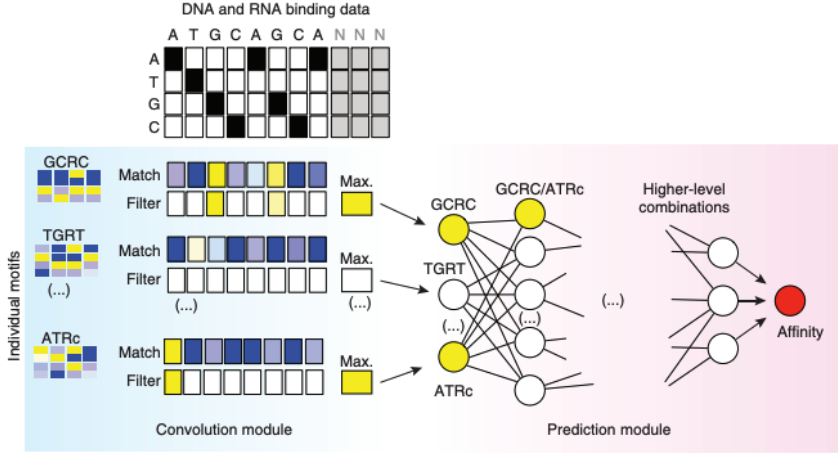
$$R_j = 2 - IC_j = 2 + \sum_{k} p_{k,j} * log(p_{k,j})$$

and the height of each base j is proportion to its frequency at position k.

TF binding prediction

Motif detector: Detect if a given sequence contains a given motif by calculating log-odds.

Convolutional filters in CNNs are narually equiavalent to motif detectors if one-hot encoding of the sequence is used as input. The higher layers learns combinatorial pattern of motifs and thus could learn complex logics.



Expectation Maximization (EM)

Given a set of ovserved variables $\mathbf{X}$, a set of unobserved latent variables $\mathbf{Z}$, and a parametric likelihood function $L(\Theta; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\Theta)$ with parameters Θ(also the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and $\mathbf{\Theta}$ is easily computable), the maximum likelihood estimate of the parameters is determined by the marginal likelihood of the observed data:

$$\Theta = arg \max_{\Theta} l(\Theta) \quad \text{where } l(\Theta) = p(\mathbf{X}|\Theta) = \sum_{i=1}^{N} \log p(x^{(i)}|\Theta) = \sum_{i=1}^{N} \log \sum_{z} p(x^{(i)}, z^{(i)}|\Theta)$$

However this is usually intractable since $z^{(i)}$s are unobserved, so EM algorithm tries to find the solution by iteratively conducting the 2-step procedures:

Expectation step (E-step): calculate the conditional distribution $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}, \mathbf{\Theta}^{(t)})$, then calculate the expected value of the joint likelihood function with respect to the conditional distribution. This gives a tight lower-bond of the $l(\Theta)$:

$$J(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) = \mathbf{E}_{\sim \mathbf{Z}|\mathbf{X}, \mathbf{\Theta}^{(t)}}[\frac{\log L(\Theta^{(t)}; \mathbf{X}, \mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \Theta^{(t)})}] = \sum_{i} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}|\Theta^{(t)})}{Q(z^{(i)})}$$

Maximization step (M-step): find the new parameter that maximizes this expectation.

$$\Theta^{(t+1)} = arg \max_{\Theta} J(\mathbf{\Theta}|\mathbf{\Theta}^{(t)})$$

Mixture model

Mixture model is a probabilistic model for representing the presence of sub-populations (characterized by latent variable **Z**) within an overall population (**X**), without requiring that an observed data set should identify the sub-population to which an individual observation belongs.

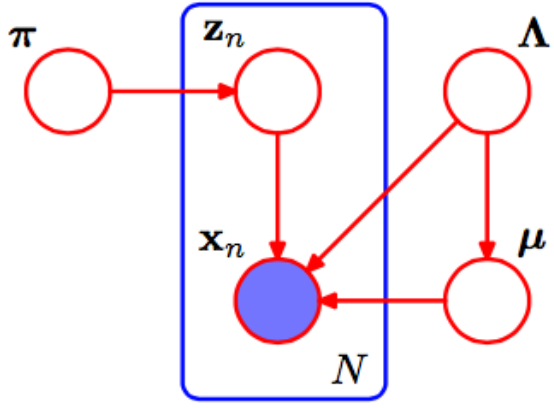


Figure: Gaussian Mixture Model

Gaussian Mixture model: the likelihood of observing each x_n is defined as a mixture of Gaussian given the sub-populations it is assigned to. Assuming there are K sub-populations ($z_n \in 1, 2, ..., K$)

$$p(x) = \sum_{i}^{K} \pi_i \mathbf{N}(x|\mu_i, \Sigma_i) = \sum_{z} p(z)p(x|z)$$

Responsibility of a mixture component takes for explaining an observation x is:

$$\begin{aligned}\gamma(z = k) = p(z = k|x) &= \frac{p(z = k)p(x|z = k)}{\sum_{j=1}^{K} p(z = j)p(x|z = j)} \\ &= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}\end{aligned}$$

EM algorithm for GMM given observable set $X = \{x_1, ..., x_N\}$

- **Randomly initialize** μ_k, Σ_k for each sub-population k.
- **E-step:** for each x_n estimate responsibility of the mixture components $\gamma(z_n)$.

$$\gamma(z = k) = p(z = k|x) = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

- **M-step:** find the MLE estimate for the parameters of the sub-population μ_k,Σ_k and π_k.

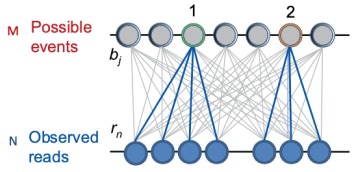
$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_n = k) x_n}{\sum_{n=1}^{N} \gamma(z_n = k)} \quad \text{weighted mean of the x in subpopulation k}$$

$$\Sigma_k = \frac{1}{\sum_{n=1}^{N} \gamma(z_n = k)} \sum_{n=1}^{N} \gamma(z_n = k)(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_n = k)}{N}$$

Binding event prediction: the latent variables are the possible events m, and the observable variables are the reads at each location r_n.

Mixture model



$$p(R \mid \pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n \mid j), \quad \sum_{m=1}^{M} \pi_m = 1$$

Position specific priors

- Events are sparse
- Events occurs more likely at motif positions

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_s + \alpha_m}$$

α_s : uniform sparse prior parameter governing the degree of sparseness, $\alpha_s > 0$;
α_m : position specific motif-based prior