March 9,10 Recitation

Prob ①: Cluster N points into M groups and determine the point distribution of each group

Prob ②: Cluster N ChIP-seq reads into M binding events and determine the read distribution of each event

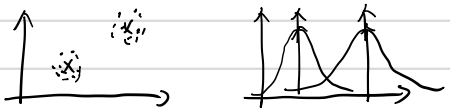
Prob ③: Cluster N binding sequences into M groups by where the motif starts and determine the PWM of motif and background

Z → groups / binding event / motif starting index

θ → μ_m, Σ_m, π_m $\quad \pi_m \quad$ M, B

X → points / pos of reads / binding sequence

} We don't know either Z or θ



seq1: BBMMMMMBB ($z=3$)

seq2: BBBMMMMMB ($z=4$)

⇒ We iteratively solve for (the distribution of) Z and the maximum-likelihood estimate of θ ($p(x,z)$)

E step: use θ^{t-1} to calculate $p(Z|X;\theta^{t-1})$

M step: Find $\theta^t = \underset{\theta}{\text{argmax}}\, E[p(x,z;\theta)]$, where E is over $p(Z|X;\theta^{t-1})$

⇒ Go to lecture notes to review motif discovery in detail