

---

# Computational functional genomics

(Spring 2002: Lecture 16)

David K. Gifford

(Adapted from a lecture by Tommi S. Jaakkola)

MIT LCS and AI Lab

*gifford@mit.edu*

---

## Discovering DNA motifs

- We are given:
  - $N$  sequences
  - Information on background sequence that does not contain the motif
  - We may have priors on the likelihood that a sequence contains a motif (location)
  - We may have priors on motif location (consensus)
  - We may have priors on motif length
- We seek to discover:
  - One or more patterns called **motifs**
  - A motif is sequence of conserved bases that is described by multinomial model (block model)
- Today
  - Review of Dirichlet priors
  - Multinomial models
  - EM and Gibbs based multiple local alignment

---

## Types of alignment

- Local vs. Global Sequence Alignment
  - Local seeks to discover common subsequences
  - Global seeks to align entire sequences
- Pair-wise vs. Multiple Alignment
  - Pair-wise considers two sequences at once
  - Multiple considers more than two sequences at once
- Local multiple alignment

---

## Maximum likelihood principle: Binomial

- **Maximum likelihood principle**: we find the parameter  $\hat{\theta}$  that maximize the likelihood of the observed data

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(x^{(1)}, \dots, x^{(n)} | \theta) \quad (1)$$

The **Maximum likelihood estimate** (MLE) for the Binomial PMF is

$$P(k_N | \theta) = \binom{N}{k} \theta^k (1 - \theta)^{(N-k)} \quad (2)$$

$$\log P(k_N | \theta) = \log \binom{N}{k} + k \log \theta + (N - k) \log(1 - \theta) \quad (3)$$

$$\frac{d P(k_N | \theta)}{d \theta} = \frac{k}{\theta} - \frac{N - k}{1 - \theta} \quad (4)$$

$$0 = \frac{k}{\theta} - \frac{N - k}{1 - \theta} \quad (5)$$

$$\hat{\theta} = k/N \quad (6)$$

---

## Maximum a Posterior Estimators (MAP)

- Assume that we know something about a coin before we observe  $N$  trials
- Prior knowledge can take on many forms
  - Assumptions (mRNA levels are never negative)
  - Data (other experiments suggests that protein A regulates gene B)
  - Estimates (our best estimate of the parameters so far)
- How do we express this knowledge so that it can be used in a principled way?
- Represent this knowledge as a **distribution over model parameters**
  - In the case of a coin, as a distribution over  $\theta$

---

## Maximum a Posterior Estimators (MAP)

- Bayesians use prior knowledge when analyzing data
  - This can lead to different conclusions from the same data, depending on your prior
- Frequentists believe that conclusions from data should always be the same
- Using **Bayes' Rule** in our Binomial example:

$$P(\theta|k_N) = \frac{P(k_N|\theta)P(\theta)}{P(k_N)} \quad (7)$$

- Let's represent  $P(\theta)$  as:

$$P(\theta) = C(\alpha)\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \quad (8)$$

$$\alpha_1 = pS + 1 \quad (9)$$

$$\alpha_2 = (1-p)S + 1 \quad (10)$$

---

## Dirichlet Distributions

- $P(\theta)$  is a Dirichlet distribution, and is a conjugate distribution to the Binomial distribution:

$$P(\theta) = C(\alpha)\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \quad (11)$$

$$\alpha_1 = pS + 1 \quad (12)$$

$$\alpha_2 = (1-p)S + 1 \quad (13)$$

- This binomial form of the Dirichlet distribution is called the Beta distribution.
- Now:

$$P(\theta|k_N) = \frac{\binom{N}{k} C(\alpha) \theta^{k+pS} (1-\theta)^{(N-k)+(1-p)S}}{P(k_N)} \quad (14)$$

$$\frac{d P(\theta|k_N)}{d\theta} = \frac{k+pS}{\theta} - \frac{(N-k)+(1-p)S}{1-\theta} \quad (15)$$

$$\theta_{\hat{MAP}} = \frac{k+pS}{N+S} \quad (16)$$

- $pS$  can be viewed as a **pseudo-count**

---

## A simple motif problem

- Imagine we have 10 sequences of 100 coin flips each
- Each sequence is made by:
  - Flip 90 fair coins
  - Flip 10 unfair coins, each may have a different bias
  - Now line the coins up, with the ten unfair coins next to one another and always in the same order in every sequence
- We seek to discover the probability of heads of each unfair coin  $\theta = [p_1, p_2, \dots, p_{10}]$  and the alignment  $A = [a_1, a_2, \dots, a_{10}]$
- $\theta$  is a model of the motif
- A potential score is  $\log \frac{P(D|A, \theta)}{P(D)}$
- However, we may wish to compute a score that is regularized by the amount of information required to determine the location of the pattern in each input sequence
- Dividing by the number of pattern parameters (10) gives us **information per parameter**
- Can use this to pick motif length if it is unknown