EM-algorithm for motif discovery

Xiaohui Xie

University of California, Irvine

Position weight matrix

- Position weight matrix representation of a motif with width w:

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \cdots & \theta_{w1} \\ \theta_{12} & \theta_{22} & \cdots & \theta_{w2} \\ \theta_{13} & \theta_{23} & \cdots & \theta_{w3} \\ \theta_{14} & \theta_{24} & \cdots & \theta_{w4} \end{bmatrix} \tag{1}$$

where each column represents one position of the motif, and is normalized:

$$\sum_{j=1}^{4} \theta_{ij} = 1 \tag{2}$$

for all $i = 1, 2, \cdots, w$.

Likelihood

- Given the position weight matrix θ, the probability of generating a sequence $S = (S_1, S_2, \cdots, S_w)$ from θ is

$$P(S|\theta) = \prod_{i=1}^{w} P(S_i|\theta_i) \quad (3)$$

$$= \prod_{i=1}^{w} \theta_{i,S_i} \quad (4)$$

For convenience, we have converted S from a string of $\{A, C, G, T\}$ to a string of $\{1, 2, 3, 4\}$.

Likelihood

- Suppose we observe not just one, but a set of sequences $S_1, S_2, \cdots, S_n$. Assume each of them is generated independently from θ. Then, the likelihood for observing these n sequences is

$$P(S_1, S_2, \cdots, S_n|\theta) = \prod_{k=1}^{n} P(S_k|\theta) \quad (5)$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}} \quad (6)$$

Parameter estimation

- Now suppose we do not know θ. How to estimate it from the observed sequence data $S_1, S_2, \cdots, S_n$?
- One solution: calculate the likelihood of observing the provided n sequences for different values of θ,

$$L(\theta) = P(S_1, S_2, \cdots, S_n|\theta) = \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}} \tag{7}$$

Pick the one with the largest likelihood, that is, to find θ^* that

$$\max_\theta \; P(S_1, S_2, \cdots, S_n|\theta) \tag{8}$$

Estimating θ using maximum likelihood

- The optimal θ^* can be derived by setting

$$\frac{\partial \log L(\theta)}{\partial \theta_{ij}} = 0 \tag{9}$$

subject to the normalization constraint.

- The maximum likelihood estimate is

$$\theta_{ij} = \frac{n_{ij}}{n} \tag{10}$$

which is simply the frequency of different letters at each position. (n_{ij} is the number of letter j at position i).

Mixture of sequences

- Suppose we have a more difficult situation. Among the set of n given sequences, $S_1, S_2, \cdots, S_n$, some of them are generated by a weight matrix θ, but some of them are not. How to identify θ in this case?
- Let us first define the "*non-motif*" (also called *background*) sequence. Suppose they are generated from a single distribution

$$p^0 = (p^0_A, p^0_C, p^0_G, p^0_T) = (p^0_1, p^0_2, p^0_3, p^0_4) \tag{11}$$

Likelihood for mixture of sequences

- Now the problem is we do not know which sequence is generated from the motif (θ) and which one is generated from the background model (θ^0).
- Suppose we are provided with such label information:

$$z_i = \begin{cases} 1 & \text{if } S_i \text{ is generated by } \theta \\ 0 & \text{if } S_i \text{ is generated by } \theta^0 \end{cases} \tag{12}$$

  for all $i = 1, 2, \cdots, n$.

- Then, the likelihood of observing the n sequences conditioned on the label variables

$$P(S_1, S_2, \cdots, S_n | z, \theta, \theta^0) = \prod_{i=1}^{n} [z_i P(S_i|\theta) + (1 - z_i) P(S_i|\theta^0)]$$

Complete likelihood

- Suppose we have some prior knowledge on whether a sequence contains a motif or not, in terms of a prior distribution

$$P(z_i) = \begin{cases} \alpha & \text{if } z_i = 1 \\ 1-\alpha & \text{if } z_i = 0 \end{cases} \tag{13}$$

- Then, we can write down the *joint probability* of S and the label variable $z \equiv (z_1, z_2, \cdots, z_n)$

$$P(S_1, S_2, \cdots, S_n, z|\theta, \theta^0) = \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1-z_i)P(S_i|\theta^0)]$$

which is called the *complete likelihood*.

True likelihood

- However, the label variables are not directly observable (also called, hidden or latent). We will need to marginalize the joint distribution over z via summation:

$$\begin{aligned} P(S_1, S_2, \cdots, S_n|\theta, \theta^0) &= \sum_{z_1, z_2, \cdots, z_n} P(S_1, S_2, \cdots, S_n, z|\theta, \theta^0) \\ &= \prod_{i=1}^{n} [\alpha P(S_i|\theta) + (1-\alpha) P(S_i|\theta^0)] \end{aligned}$$

- Now, the parameter estimation can be formulated as maximize the true likelihood function

$$\max_{\theta, \theta_0} \; L(\theta, \theta_0) = P(S_1, S_2, \cdots, S_n|\theta, \theta_0) \tag{14}$$

Method I: Gradient Ascend

- As before, we use the log likelihood function

$$\log L(\theta, \theta_0) = \sum_{i=1}^{n} \log[\alpha P(S_i|\theta) + (1-\alpha) P(S_i|\theta^0)] \qquad (15)$$

- Gradient-based method: 1) calculate the gradient of $\log L(\theta, \theta_0)$ with respect to θ_{ij},

$$\frac{\partial \log L(\theta, \theta_0)}{\partial \theta_{ij}} \qquad (16)$$

Then, 2) update θ_{ij} with

$$\theta_{ij}^{t+1} = \theta_{ij}^{t} + \eta \left[\frac{\partial \log L(\theta, \theta_0)}{\partial \theta_{ij}} \right]_{\theta=\theta^t} \qquad (17)$$

where η is the step size.

Problems with gradient-based methods

- Gradient is hard to calculate
- Need to choose the correct step size
- Slow to converge
- Solution is only locally optimal

Method II: EM-algorithm

- Instead of optimizing the true likelihood function, we optimize an approximate likelihood

$$\log \tilde{L}(\theta, \theta_0) = \sum_{i=1}^{n} [q(z_i = 1) \log P(S_i|\theta) + q(z_i = 0) \log P(S_i|\theta^0)]$$

  where is also called *average log likelihood*. $q(z_i)$ is the posterior distribution of the label variable.

- The average log likelihood is a lower bound on the tree log likelihood function (Jensen's Inequality).

$$\begin{aligned} \log[\alpha P(S_i|\theta) + (1-\alpha) P(S_i|\theta^0)] \geq \\ q \log[P(S_i|\theta)\alpha/q] + (1-q) \log[P(S_i|\theta)(1-\alpha)/(1-q)] \end{aligned} \tag{18}$$

  for all $q \in [0, 1]$.

Expectation and Maximization

- The EM-algorithm iterates between two steps:
  - Expectation: calculate the posterior distribution of z_i,

$$q^{(t+1)}(z_i) \sim \begin{cases} P(z_i = 1)P(S_i|\theta^{(t)}) & \text{if } z_i = 1 \\ P(z_i = 0)P(S_i|\theta_0^{(t)}) & \text{if } z_i = 0 \end{cases} \tag{19}$$

  - Maximization: find optimal θ and θ_0,

$$\theta_{kl}^{(t+1)} \sim \sum_{i=1}^{n} q^{(t+1)}(z_i) I(S_{ik} = l) \tag{20}$$

- The two steps are guaranteed to converge to a locally optimal solution.

Pros and Cons of EM-algorithm

- Pros:
  - No need to choose step size
  - Guaranteed to converge
  - Fast
- Cons:
  - Locally optimal
  - Sensitive to the initialization of parameters

Method III: Gibbs Sampling

Motivation: the key problem is that the label variable z is unknown. Maybe we should try to generate a sample of these labels.

- **Initialization**: Randomly assign z_i to be 1 or 0 according to the prior probability $P(z_i)$.
- **Estimation step**: Traverse through S_i from $i = 1$ to n. Suppose we are considering S_i. Calculate the absolute frequency matrix of all other sequences (excluding S_i) with label 1. Let n_{ij} denote the number of letter j at position i. Set

$$\theta_{ij} = \frac{n_{ij} + \gamma}{n + 4\gamma} \tag{21}$$

where γ is a small number (called pseudocount). n is the total number of sequences with label 1, excluding S_i. Same for θ_0.

Gibbs Sampling

- **Sampling step**
  - Provided with current estimation of θ, θ_0. For sequence S_i, we can calculate the posterior probability of z_i:

$$q(z_i) \sim \begin{cases} P(z_i = 1)P(S_i|\theta) & \text{if } z_i = 1 \\ P(z_i = 0)P(S_i|\theta_0) & \text{if } z_i = 0 \end{cases} \tag{22}$$

    that is $q(z_i = 1) = \alpha P(S_i|\theta)/[\alpha P(S_i|\theta) + (1-\alpha)P(S_i|\theta_0)]$.
  - Randomly assign z_i to be 1 or 0 according to probability $q(z_i)$.
- Go to another sequence and repeat.

Pros and Cons of Gibbs sampling algorithm

- Pros:
  - Less susceptible to local optimal
  - Can naturally incorporate prior information
  - Guaranteed to converge
- Cons:
  - Can be slow
  - No good criterion on when to stop

Summary

We have discussed three algorithms for probabilistic motif discovery:

- Gradient-based method
- EM-algorithm
  - MEME:
    http://meme.sdsc.edu/meme/meme.html
- Gibbs sampling
  - BioProspector:
    http://ai.stanford.edu/∼xsliu/BioProspector/
  - AlignACE:
    http://atlas.med.harvard.edu/