Methods for the *discovery of* cis-regulatory modules, 3

Statistics 246
Week 14

Spring 2006
Lecture 2
From some ~6 yo slides of Jun Liu

Extensions to the basic Motif Alignment Model with a Gibbs sampler to multiple copies of multiple motifs were all outlined in the 1993 paper, ref on p.11. Full proofs came a bit later.
Phase-shift and fragmentation

- Sometimes get stuck in a local shift optimum
  - How to “escape” from this local optimum?
    - Simultaneous move: $A \rightarrow A + \delta$, $A + \delta = \{a_1 + \delta, \ldots, a_K + \delta\}$
    - Use a Metropolis step: accept the move with prob = $p$, 

$$p = \min\{1, \frac{\pi(A + \delta | R)}{\pi(A | R)}\}$$

Compare entropies between new columns and left-out ones.
The Fragmentation Model

Imagine we have \( w \) positions for the motif, only \( J \) of them are “important”:

Choose the best \( J \) columns among \( W \) possible choices via a Metropolis algorithm.

Prior distribution for fragmentation:

\[
\pi(\Delta) \propto \left( \frac{w - 2}{J - 2} \right)^{-1}
\]

Here \( w \) is regarded as the length of “span”. Using this model, no need to specify it exactly.
Generalizing the Single Block Motif Model
Repeated Motifs: the Gibbs Motif Sampler

- Sequences often have multiple repeats of several motifs in the set of sequences.
- **Idea 1**: along the line of predictive updating:

  Need a prior for the number of repeats in each sequence. Then we can conduct a similar PU step.
Idea 2: Mixture modeling

View the dataset as a long sequence with K motif types:

Idea: partition the input sequence into segments that correspond to different (unknown) motif models.

It is a mixture model (unsupervised learning).

Implement a predictive updating scheme.
Special Case: Bernoulli Sampler

- Sequence data: \( R = r_1 r_2 r_3 \ldots \ldots r_N \)
- Indicator variable: \( \Delta = \delta_1 \delta_2 \delta_3 \ldots \ldots \delta_N \)
  \[
  \delta_i = \begin{cases} 
  1 & \text{if it is the start of an element} \\
  0 & \text{if not.} 
  \end{cases}
  \]
- Likelihood: \( \pi(R, \Delta \mid \Theta, \varepsilon) \), \( \varepsilon \) is the prior prob for \( \delta_i=1 \)
- Predictive Update:
  \[
  \frac{\pi(\delta_k = 1 \mid \Delta_{[-k]}, R)}{\pi(\delta_k = 0 \mid \Delta_{[-k]}, R)} = \frac{\hat{\varepsilon}}{1 - \hat{\varepsilon}} \prod_{i=1}^{w} \left( \frac{\hat{p}_{i, r_{k+i-1}}}{\hat{p}_{0, r_{k+i-1}}} \right)
  \]
Example: Pore-like proteins

Membrane protein, channels for nutrients, waste products, and antibiotics.

- 25 proteins suspected of containing porins
- 70 segments found, with width=13
- 4 repeats detected for each of the 3 porins with known 3-D structures, corresponding to alternating membrane spanning $\beta$-strands.
- All on the outside of the structure.
References


• Liu, J.S. (1994). *JASA* 89, 986

Revisiting the Wu & Xie HMM

In the next few pages we revisit the Wu & Xie HMM for discovering cis-regulatory modules, as it mirrors quite closely what we have just done in a very natural way.

The major differences between what follows and what has just gone are a) here we permit more than one TFBS in the sequence with a modular structure for them; b) because of the HMM architecture, the collapsing step is not feasible, so that the probabilities need to be sampled, with the consequence that the sampling of the missing data indicators is carried out with explicit “known” probabilities; and c) these probabilities are not immediately accessible, but have to be calculated.

The structure of the HMM is repeated next, and then we turn to the Gibbs sampler.
Wu & Xie (2005)’s CRM HMM

Inter-CRM background state

Intra-CRM background state

3 motif model
Some details of the Gibbs sampler development of Wu & Xie (2005)

Consider a single observed sequence $X$. Let $M$ be the (unobserved) module indicators, and $A$ the (unobserved) TFBS indicators, $A = \{A_k, k=1,\ldots,K\}$, where $A_k$ is the indicator of the binding sites for the $k$th motif, and $A_{K+1}$ is the indicator of the non-site background inside the modules. We use the notation $X(M^c)$ to denote the background sequence outside the modules.

Write $\Theta = \{\theta_0, \Theta_1, \ldots, \Theta_K\}$, where $\theta_0$ is the parameter of a background Markov model and $\Theta_k$ is the $k$th PWM, all treated as known. Given $\Theta$, $q = \{q_0, q_1, \ldots, q_K\}$, and $r$, the complete data probability is

$$P(X, M, A \mid \Theta, q, r) = P(M, A \mid \Theta, q, r)P(X(M^c) \mid \theta_0, M)P(X(M) \mid M, A, \Theta).$$

Note that the parameters $q$ and $r$ are unknown, in addition to the hidden path defined by the unobserved $M$ and $A$. 
Combining the previous equation with priors for $q$ and $r$ gives the joint posterior distribution for the unknowns

$$P(M, A, q, r | X, \Theta) \propto P(X, M, A | \Theta, q, r)\pi(q)\pi(r),$$

where $\pi(q)$ is taken to be a Dirichlet prior distribution for $q$, and $\pi(r)$ is a Beta for $r$.

By sampling from this, we can obtain the values of the parameters $q$ and $r$ and of the hidden path $M$ and $A$ which maximize the right-hand side marginally and jointly. This will give us the optimal locations of the modules and of the TFBSs within the modules. Here are a few more details.

**Sampling $q$ and $r$ given $M$ and $A$ (and $X$, $\Theta$).**

First, we generate initial values of $q$ and $r$ using the Dirichlet and Beta distributions, respectively. (Clearly hyper-parameters for the priors need to be found before this step, but I skip this.)

Because of the conjugacy of the priors, $q \mid M, A$ and $r \mid M$ are again Dirichlet and Beta respectively, with easily computed parameters. Sampling from these is easy. Can you guess how one might do this?
Details, 3.

Sampling $M$ and $A$ given $q$ and $r$ (and $X, \Theta$).

Here we first use the HMM forward algorithm to calculate the marginal probability of $X = (x_1, \ldots, x_L)$ by summing over all possible paths. This gives us all the quantities of the form (notation as in the HMM literature)

$$\alpha_m(k) = pr(x_1, \ldots, x_m, s_m = k \mid \Theta, q, r), \; k = 0, \ldots, K+1,$$

where $s_m$ is the state of $x_m$. The sum of the $\alpha_L(k)$ over $k$ is just $P( X \mid \Theta, q, r)$. Now in our case (check the state diagram) if $k$ is a motif, then we take $x_m$ to be the last position of the motif. The forward recursions are then

$$\alpha_m(0) = (1-r) \times pr(x_m \mid x_{m-1}, \theta_0)\alpha_{m-1}(0) + q_0 \times pr(x_m \mid x_{m-1}, \theta_0)\alpha_{m-1}(K+1)$$

$$\alpha_m(K+1) = r \times pr(x_m \mid x_{m-1}, \theta_0)\alpha_{m-1}(0) + \sum_{k=1}^{K} \alpha_{m-1}(k) + q_{K+1} \times pr(x_m \mid x_{m-1}, \theta_0)\alpha_{m-1}(K+1)$$

$$\alpha_m(k) = q_k \times pr\{x_{m-w_k+1}, \ldots, x_m \mid \Theta_k\}\alpha_{m-w_k}(K+1), \; k=1, \ldots, K,$$

where $pr\{x_{m-w_k+1}, \ldots, x_m \mid \Theta_k\}$ is the probability of generating the segment of length $w_k$ from the $k$th motif model $PWM_k$.

The initial conditions here are $\alpha_0(0) = 1$, and $\alpha_m(k) = 0$ for $k=1, \ldots, K+1$, $m \geq 0$. 
Details, 4.

With the $\alpha_m$ all calculated, backward sampling can be used to sample $M$ and $A$. Here’s how it is done. Start from $m=L$. We first need to note that

$$
\text{pr}(s_L = k \mid X, \Theta, q, r) \propto \alpha_L(k).
$$

Thus we can normalize the $\alpha_L(.)$ and sample a value for $s_L$.

Moving backwards, once we generate the last position of a module, and are at $m$, we need to sample a state corresponding to a background letter observation ($s_m = K+1$), or to the sequence segment $\{x_{m-wk+1}, \ldots, x_m\}$ being one of the motifs ($s_m = k, 1 \leq k \leq K$), given the previously sampled states $s_{m+1}, \ldots s_L$. In their paper, Wu & Xie state that the probabilities of these events are again proportional to the $\alpha_m(.)$; for a proof, see next page.

Depending on the state sampled, we move to the first position of the motif, and repeat the sampling procedure until we reach $m=1$.

This ends one round of sampling of $A$ and $M$, and we switch to sampling $q$ and $r$ and repeat the cycle 1,000 times, say.
Details, 5.

When we reach position $m < L$ in the backward sampling, the next state should be sampled with probability

$$\text{pr}(s_m = k \mid s_{m+1}, \ldots s_L, X, \Theta, q, r).$$

Let’s drop the $\Theta, q$ and $r$, and simplify this quantity. Clearly

$$\text{pr}(s_m = k \mid s_{m+1}, \ldots s_L, X) = \text{pr}(s_m = k, s_{m+1}, \ldots s_L, X) / \text{pr}(s_{m+1}, \ldots s_L, X).$$

The numerator here can be written as

$$\alpha_m(k)\text{pr}(s_{m+1}, x_{m+1}, \ldots, s_L, x_L \mid s_m = k)$$

(check), while the denominator can obviously be written as

$$\sum_{k'} \alpha_m(k')\text{pr}(s_{m+1}, x_{m+1}, \ldots, s_L, x_L \mid s_m = k').$$

Since we are at the last position of a motif, all states go to $K+1$, and so the conditional probabilities given $s_m = k'$ are all the same, $1 \leq k' \leq K$, and our desired probability is indeed proportional to $\alpha_m(k)$. 