Methods for the *discovery of* cis-regulatory modules, 2

Statistics 246
Week 14

Spring 2006
Lecture 1
Finding motifs/modules using gene expression data

• Clustering methods
  – Assume that co-expression implies co-regulation.
  – Group genes into disjoint clusters based on similarity in expression profile
  – Run an algorithm such as MEME, AlignAce, MDScan, etc to find shared motifs in upstream regions of the groups of genes
  – Then seek to cluster your motifs. We start on this today.

• Linear regression methods
  – Model the gene expression (ratio) as a linear function of one or more TFBS variables (motif counts, PWM scores or more), and select the most relevant subsets. In a way these methods are variants on that in Wasserman and Fickett (1998). We come to this later.
Using sequence data from the organism of interest after clustering genes

The sequences are typically from promoter regions of genes believed to be co-regulated, and hence to share regulatory elements. If the organism is yeast, people typically take ~600 bp upstream of the transcription start site, whereas if it is mouse or human, 3kb or more might be used.

There are several methods in use for identifying shared binding sites of sets of sites in unaligned sequences. Because of its importance, I’ll give a quick historical overview of the Gibbs sampler approach to this problem. Programs implementing this include AlignAce and BioProspector. In general these methods work brilliantly with bacteria, reasonably well with yeast, less well with the fly, and not very well at all with mouse and human. The issue here is the frequency of occurrence, length, and degree of conservation, and proximity to the TSS of TFBSs.

For a recent review, see Tompa et al, Nat. Biotech. 23 (1) Jan 2005 p.137.
The *E. coli* CRP dataset
Lawrence & Reilly, 1990

- 18 unaligned DNA sequences, each of length 105 bp.
- There is at least one CRP binding site, known experimentally, in each sequence. (We met these sites last week in our discussion of the regulation of *lacZ*.)
- The binding sites are about 16-19 base pairs long, with considerable variability in their contents.
- We are interested in seeing if we can find and characterize these sites computationally.
- This paper started a stream of research leading to many of the methods reviewed in Tompa *et al*, 2005.

The figures in the next few slides are from slides or papers by Jun Liu.
The CRP dataset: actual sites indicated
A first approach to this problem

- Was made by Lawrence and Reilly, *Proteins: Structure, Function and Genetics*, 1990 who introduced what we call the motif alignment model (next slide). They used the EM algorithm.
- The best contemporary implementation of their method is probably in the program MEME. I’ll omit further details just now, and turn to
- An alternative approach is via the Gibbs sampler, whose approach I’ll discuss in some detail.
Motif alignment model

The observed data: \( K \) sequences \( b_k = (b_{k,1}, \ldots, b_{k,N_k}) \)

The missing data: the alignment variable \( A = \{a_1, a_2, \ldots, a_K\} \) of motif start positions

- Bases in background (= non-motif) positions assumed to follow a common multinomial with \( p_0 = (p_{0,a}, p_{0,c}, p_{0,g}, p_{0,t}) \)
- Bases in position \( j \) in the motif assumed to follow a multinomial distribution with \( p_j = (p_{j,a}, p_{j,c}, p_{j,g}, p_{j,t}), j=1,\ldots,w. \)
- All bases are assumed mutually independent
Suppose that the multinomial probabilities \( \{p_0, p_1, \ldots, p_w\} \), are all known, and that we have one observed nucleotide sequence \( \mathbf{b} = (b_1, b_2, \ldots, b_N) \) known to contain one instance of the motif. Where is it? Denote its start index by \( a \).

Let’s suppose that our prior distribution \((\pi_n)\) for the start site \( a \) is uniform on the indices \( n=1, \ldots, N-w+1 \), and let’s update that given the observed sequence \( \mathbf{b} = (b_i) \). That is, we calculate the posterior probability

\[
pr(a=m | \mathbf{b}) = \frac{pr(b | a=m)\pi_m}{\sum n pr(b | a=n)\pi_n}
\]

\[= pr(b | a=m) / \sum n pr(b | a=n), \quad \text{since } \pi \text{ is uniform.}
\]

Now each term in the numerator and denominator is a product of \( p \)'s with first subscript \( 0 \), for background, or \( j = 1, \ldots, w \), for position \( j \) in the motif.

Divide both by

\[
\prod_{i=1}^{N} p_{0,b_i}.
\]

We find the following, which you should interpret and remember:

\[
pr(a = m | \mathbf{b}) = \prod_{j=1}^{w} \frac{p_{j,b_{m+j-1}}}{p_{0,b_{m+j-1}}} / \sum_{n=1}^{N-w+1} \prod_{j=1}^{w} \frac{p_{j,b_{n+j-1}}}{p_{0,b_{n+j-1}}} \propto \prod_{j=1}^{w} \frac{p_{j,b_{m+j-1}}}{p_{0,b_{m+j-1}}}.
\]
The Liu-Neuwald-Lawrence Gibbs sampler

We are going to suppose that the $p_j$ have independent Dirichlet priors $\text{Dir}(p_j, \beta_j)$ with parameters $\beta_j$, $j = 0, \ldots, w$.

Our data will be a set of $K$ sequences $b_k = (b_{k,1}, \ldots, b_{k,N_k})$ of bases assumed all to have just one instance of the motif. This instance is supposed \textit{a priori} to be equally likely to begin at any feasible index, i.e. we have independent uniform priors for the missing motif start indicators $A = (a_k)$.

Call the sequences \textit{Data}, and the probabilities $\Theta$. The basic Gibbs sampler would involve sampling from the distribution $pr(A, \Theta \mid \text{Data})$, and typically this will be done by alternating between sampling $A \mid \Theta, \text{Data}$, and sampling $\Theta \mid A, \text{Data}$. In the present case we skip the sampling of $\Theta$, utilizing what is known in the MCMC trade as \textit{collapsing}, simply sampling from $A \mid \text{Data}$. This can done sequence by sequence, and we now illustrate one step, denoting the indicator set $A$ without $a_k$ by $A_{[-k]}$. 
The predictive update (PU) step, 1

First, note that $\text{pr}(\text{Data}, A \mid \Theta)$ which we call (*) is simply a product of $p$’s for bases $b=a, c, g$ and $t$, some with first subscript $0$ for background, and others with first subscripts $j$ for position $j$ in the motif, raised to the powers which correspond to their frequencies of occurrence. We are going to calculate $\text{pr}(a_k = m \mid A_{[-k]}, \text{Data})$, which is just the integral of the product of the Dirichlet priors by $\text{pr}(a_k = m, A_{[-k]}, \text{Data}, \Theta)$, divided by the sum from $m = 1$ to $N_k - w + 1$ of such terms. Now this product has exactly the same form as (*), but with different exponents, these now including the parameters $\beta_{j,b}$ from the priors. However, these integrals all end up giving ratios of products of gamma functions, so our calculation boils down to finding a simple representation of these products. We now turn to this, following Liu et al, JASA 1995.
The PU step, 2

The product of (*) by the Dirichlet priors has the form

\[
\prod_{j=0}^{w} \prod_{b \in \{a,c,g,t\}} p_{j,b}^{\beta_{j,b} - 1} \times p_{j,b}^{C_{j,b}^{[-k]}} \times p_{j,b}^{C_{j,b}^{k,m}}
\]

where \(C_{j,b}^{[-k]}\) and \(C_{j,b}^{k,m}\) are the counts of base \(b\) in position \(j = 0\) (for background) and \(j=1,\ldots,w\) for the motif, in the sequences other than the \(k\)th, or in the \(k\)th when its motif starts at \(m\), respectively. Integrating out the \(\rho\)'s, we get a quantity proportional to

\[
\prod_{b \in \{a,c,g,t\}} \Gamma(\beta_{0,b} + C_{0,b}^{[-k]} + C_{0,b}^{k,m}) \prod_{j=1}^{w} \Gamma(\beta_{j,b} + C_{j,b}^{[-k]} + \delta(b, b_{k,m+j-1}))
\]

where \(\delta(b,b') = 1\) if \(b=b'\), and = 0 otherwise.
The PU step, 3

Now we try to simplify. Just as we divided by a quantity on slide 8 to simplify, so we also do here. This time the quantity with which to divide is

\[
\prod_{b \in \{a,c,g,t\}} \{ \Gamma(\beta_{0,b} + C_{0,b}^{[-k]} + C_{b}^{k}) \prod_{j=1}^{w} \Gamma(\beta_{j,b} + C_{j,b}^{[-k]}) \}
\]

where \( C_{b}^{k} \) is the count of base \( b \) in sequence \( k \).

After dividing, we first collect the terms to the right of the product from 1 to \( w \) which have \( \delta=1 \) (the rest cancel). Now use a familiar property of the gamma function, and obtain

\[
\prod_{j=1}^{w} (\beta_{j,b_{k,m+j-1}} + C_{j,b_{k,m+j-1}}^{[-k]}).
\]
Apart from the missing denominators, which do not depend on \( m \), this is a \textbf{product} of the \textbf{posterior probabilities} of the \( w \)-tuple \( b_{k,m}, \ldots, b_{k,m+w-1} \) being a motif, given \textbf{Data} and \( A_{[-k]} \), the motif locations in the sequences other than \( k \).

Next we turn to the other component of the ratio. It is

\[
\prod_{b \in \{a,c,g,t\}} \frac{\Gamma(\beta_{0,b} + C_{0,b}^{[-k]} + C_{0,b}^{k,m})}{\Gamma(\beta_{0,b} + C_{0,b}^{[-k]} + C_{0,b}^{k})}. 
\]

Here we notice that for each \( b \), the gamma function in the denominator has more terms than the corresponding one in the numerator, specifically, \( n(k,m,b) \) more terms, where \( n(k,m,b) \) is the number of bases \( b \) in the \( w \)-tuple \( b_{k,m}, \ldots, b_{k,m+w-1} \). The \textbf{reciprocal} of the product above thus simplifies to

\[
\prod_{b \in \{a,c,t,g\}} \prod_{r=1}^{n(k,m,b)} (\beta_{0,b} + C_{0,b}^{[-k]} + C_{b}^{k} - r) \approx \prod_{b \in \{a,c,t,g\}} (\beta_{0,b} + C_{0,b}^{[-k]})^{n(k,m,b)} = 13
\]
The PU-Step, completed

(from the previous page) \[ = \prod_{j=1}^{w} \left( \beta_{0,b_{k,m+j-1}} + C_{0,b_{k,m+j-1}}^{[-k]} \right). \]

Again these are the numerators of posterior probabilities, given Data and $A_{[-k]}$. If we denote these posterior probabilities by $q_{j,b}$, then completely in line with the formula on p.8 we have

\[ P(a_k = m \mid A_{[-k]}, Data) \propto \prod_{j=1}^{w} \frac{q_{j,b_{m+j-1}}}{q_{0,b_{m+j-1}}}. \]
The PU algorithm (idea only)

- **Initialize** by choosing random starting positions
- **Iterate** the following steps many times:
  - Randomly or systematically choose a sequence, say, sequence $k$, to exclude.
  - Carry out the P-U step as just described to update $a_k$
- **Stop** when not much change observed, or when some criterion is met.

There are several extra features now, particularly to deal with column shift, >1 motif/sequence, to use better background models, to estimate the motif width $w$, etc. Again I omit details, as my aim was to present the essentials.
Results: comparing Gibbs method (GM) with the earlier EM on the CRP dataset

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Footprint sites</th>
<th>Two most likely sites</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EM</td>
<td>GM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>First</td>
<td>Second</td>
<td>First</td>
</tr>
<tr>
<td>cole 1</td>
<td>17, 61</td>
<td>61</td>
<td>45</td>
<td>61</td>
</tr>
<tr>
<td>eco arabop</td>
<td>17, 55</td>
<td>55</td>
<td>76</td>
<td>55</td>
</tr>
<tr>
<td>eco bgirl</td>
<td>78</td>
<td>78</td>
<td>40</td>
<td>76</td>
</tr>
<tr>
<td>eco crp</td>
<td>63</td>
<td>63</td>
<td>73</td>
<td>63</td>
</tr>
<tr>
<td>eco cya</td>
<td>50</td>
<td>50</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>eco deop</td>
<td>7, 60</td>
<td>7</td>
<td>39</td>
<td>7</td>
</tr>
<tr>
<td>eco galE</td>
<td>42</td>
<td>24</td>
<td>76</td>
<td>42</td>
</tr>
<tr>
<td>eco ilvBpr</td>
<td>39</td>
<td>39</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td>eco lac</td>
<td>9, 80</td>
<td>9</td>
<td>73</td>
<td>9</td>
</tr>
<tr>
<td>eco malE</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>eco malK</td>
<td>29, 61</td>
<td>61</td>
<td>29</td>
<td>61</td>
</tr>
<tr>
<td>eco malt</td>
<td>41</td>
<td>41</td>
<td>11</td>
<td>41</td>
</tr>
<tr>
<td>eco ompA</td>
<td>48</td>
<td>48</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>eco trnaA</td>
<td>71</td>
<td>71</td>
<td>34</td>
<td>71</td>
</tr>
<tr>
<td>eco uXul</td>
<td>17</td>
<td>17</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>pbr-P4</td>
<td>53</td>
<td>53</td>
<td>84</td>
<td>53</td>
</tr>
<tr>
<td>tm9Bcat</td>
<td>1, 84</td>
<td>5</td>
<td>66</td>
<td>5</td>
</tr>
<tr>
<td>(tdc)</td>
<td>78</td>
<td>78</td>
<td>76</td>
<td>78</td>
</tr>
</tbody>
</table>