Genomics and Bioinformatics: Week 12

3 December 2013

1 Relating protein copy number and binding site occupancies

We are interested in the interaction kinetics of a particular transcription factor (such as $PPAR\gamma$) with a mammalian-size genome. This genome contains a large number of potential binding sites. Each site will have an average occupancy (fraction of cells with the factor bound to the site) which is a function of the available concentration of the factor and the affinity (the dissociation constant K_d or the binding free energy $G(S)$) of the site for the factor.

The genome (of size $G[bp]$) can be divided into an accessible part (size M) and an inaccessible part (size $G-M$). The M accessible potential binding sites are classified by binding affinity (K_d) into $k+1$ categories ordered from 0 (non-specific background) to k (most specific). There are m_i sites in category no i to which n_i transcription factor proteins are bound:

$$\begin{aligned} N &= \sum_{i=0}^{k} n_i = \text{nb of nuclear TF proteins} \approx 10^3 - 10^4 \ , \\ M &= \sum_{i=0}^{k} m_i = \text{size of accessible genome} \approx 10^8 [bp] \ , \end{aligned}$$

The factor's DNA binding domain is described by a position-weight matrix $w(p,\alpha)$ where $p = 1, \ldots, L$ is the position within the domain and $\alpha \in \{A, C, G, T\}$ is the matching base in the DNA. We normalize the matrix w such that

$$w(p,\alpha) = e^{-g(p,\alpha)} \ , \qquad \max_{\alpha} e^{-g(p,\alpha)} = 1 \ ,$$

namely the most frequent base α_c at each position has a weight of 1 ($g(p,\alpha_c) = 0$). A given (accessible) site S has an energy $G(S)$ given by summing the contributions of each position:

$$G(S) = G_c + \sum_{p=1}^{L} g(p, S(p)) \ ,$$

where G_c is, by construction, the energy of the consensus site S_c. In terms of statistical weight, this is a product of independent variables:

$$W(S) \;=\; W_c e^{-\beta \sum_{p=1}^{L} g(p,S(p))} \;=\; \prod_{p=1}^{L} W_c e^{-\beta g(p,S(p))} \;,$$

where $W_c = W(S_c) = \exp(-\beta G_c)$. We can now express the average weight as

$$\overline{W} \;=\; \frac{1}{M} \sum_{i=0}^{k} m_i W_c e^{-\beta(G_i - G_c)} \;=\; \frac{1}{M} \sum_{i=0}^{k} m_i W_i \;,$$

where G_i is a representative energy for sequences in the category i.

Introducing the partition function

$$Z(N, \{m_i\}) \;=\; \sum_{\{n_i \le m_i \,\mid\, \sum n_i = N\}} \prod_{i=0}^{k} \binom{m_i}{n_i} W_i^{n_i} \;.$$

we can obtain the average number of proteins bound to each category of sites by

$$\overline{n}_i \;=\; -\frac{1}{\beta}\frac{\partial}{\partial G_i} \log Z(N, \{m_i\}) \;=\; W_i \frac{\partial}{\partial W_i} \log Z(N, \{m_i\}) \;.$$

1.1 Excess of proteins over binding sites

In this regime, $M \gg N \gg \sum_{i=1}^{k} m_i$, we can use an approximation for the contribution from the dominating unspecific sites (category 0):

$$\binom{m_0}{n_0} \approx \binom{M}{N} \left(\frac{N}{M}\right)^{N-n_0} \;,$$

which yields

$$\begin{aligned} Z(N, \{m_i\}) \;&\approx\; \sum_{\{n_i \le m_i \,\mid\, \sum n_i = N\}} \binom{M}{N} W_0^N \prod_{i=1}^{k} \binom{m_i}{n_i} \left(\frac{N W_i}{M W_0}\right)^{n_i} \\ &=\; \binom{M}{N} W_0^N \prod_{i=1}^{k} \left(1 + \frac{N W_i}{M W_0}\right)^{m_i} \;. \end{aligned}$$

therefore

$$\frac{\overline{n}_i}{m_i} \;=\; \frac{\frac{N W_i}{M W_0}}{1 + \frac{N W_i}{M W_0}} \;.$$

1.2 Excess of specific sites over proteins

In this regime, $M \gg m_i \gg N$ $(i > 0)$, and the following approximation using multinomial coefficients:

$$\prod_{i=0}^{k} \binom{m_i}{n_i} \approx \binom{N}{n_0, \ldots, n_k} \frac{\prod_{i=0}^{k} m_i^{n_i}}{N!} \ ,$$

leads to

$$\begin{aligned} Z(N, \{m_i\}) &\approx \frac{1}{N!} \sum_{\{n_i \mid \sum n_i = N\}} \binom{N}{n_0, \ldots, n_k} \prod_{i=0}^{k} (m_i W_i)^{n_i} \\ &= \frac{1}{N!} \left(\sum_{i=0}^{k} m_i W_i \right)^N \ . \end{aligned}$$

therefore

$$\frac{\overline{n}_i}{m_i} = \frac{N W_i}{M \overline{W}} \ . \tag{1}$$

1.3 Fitting quantitative binding data

Suppose next that we know the matrix $w(p, \alpha)$ (as well as $g(p, \alpha) = -\log w(p, \alpha)$) and we have a measure of genome-wide occupancy $\tau(S)$ (e.g. a ChIP-seq density profile). Then the average occupancy τ_i of sites in category i must be related to $\overline{n}_i$ by a simple calibration:

$$\begin{aligned} \tau_i &= \frac{1}{m_i} \sum_{j=1}^{m_i} \tau(S_j) \\ &= \lambda \frac{\overline{n}_i}{m_i} + \mu \\ &= \lambda \frac{N W_i}{M \overline{W}} + \mu \\ &= \lambda \frac{N e^{-\beta(G_i - G_c)}}{M \overline{W} / W_c} + \mu \ . \end{aligned}$$

In this equation we have applied the approximation (1). Unknown in this expression are the scaling factors β, λ, and μ, which can be optimized by least-square fit. Given β, the average $\overline{W}/W_c$ can be computed by sampling the matrix $w(p, \alpha)$ directly.

N and M must be determined experimentally: N by proteomics, M by ChIP-seq (H3K27ac histone modification) or by mapping DNAse hypersensitive sites.

1.3.1 Procedure

1. Determine the set M of accessible sites in the genome by using a threshold on H3K27ac ChIP-seq data.

2. Scan those regions with the matrix $w(p, \alpha)$ and record significant motif scores: $\{a_j\}$.

3. Measure the average ChIP-seq τ in a neighborhood of every motif to get $\{\tau_j\}$.

4. Fit λ_0, μ and β by least square optimization of $\tau_j = \lambda_0 a_j^\beta + \mu$.

5. Estimate the average motif weight by generating random sequences $\{S_j : j = 1, \ldots, J\}$ of length L: $\overline{W}/W_c = \frac{1}{J}\sum_{j=1}^{J} e^{-\beta \sum_{p=1}^{L} g(p, S_j(p))}$.

6. Infer λ from N, M and $\overline{W}/W_c$ as $\lambda = \lambda_0 \frac{M\overline{W}}{NW_c}$.