Motif representation using position weight matrix

Xiaohui Xie

University of California, Irvine

Position weight matrix

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

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

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

$$\sum_{j=1}^{4} \theta_{ij} = 1 \quad (2)$$

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 $S_1, S_2, \cdots, S_n$, each of which contains exactly w letters. Assume each of them is generated independently from the model θ. Then, the likelihood of observing these n sequences is

$$\begin{aligned} P(S_1, S_2, \cdots, S_n|\theta) &= \prod_{k=1}^{n} P(S_k|\theta) && (5) \\ &= \prod_{k=1}^{n} \prod_{i=1}^{w} \theta_{i,S_{ki}} = \prod_{i=1}^{w} \prod_{j=1}^{4} \theta_{ij}^{c_{ij}} && (6) \end{aligned}$$

where c_{ij} is the number of letter j at position i (Note that $\sum_{j=1}^{4} c_{ij} = n$ for all i).

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

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

Maximum likelihood estimation

- Maximum likelihood estimation of θ is

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_{\theta} \log L(\theta)) = \sum_{i=1}^{w} \sum_{j=1}^{4} c_{ij} \log \theta_{ij} \\ s.t. & \quad \sum_{j=1}^{4} \theta_{ij} = 1, \quad \forall i = 1, \cdots, w \end{aligned} \tag{9}$$

Optimization with equality constraints

- Construct a Lagrangian function taking the equality constraint into account:

$$g(\theta) = \log L(\theta) + \sum_{i=1}^{w} \lambda_i (1 - \sum_{j=1}^{4} \theta_{ij}) \tag{10}$$

- Solve the unconstrained optimization problem

$$\hat{\theta} = \arg\max_{\theta} g(\theta)) = \sum_{i=1}^{w} \sum_{j=1}^{4} c_{ij} \log \theta_{ij} + \sum_{i=1}^{w} \lambda_i (1 - \sum_{j=1}^{4} \theta_{ij}) \tag{11}$$

Optimization with equality constraints

- Take the derivative of $g(\theta)$ w.r.t. θ_{ij} and the Lagrange multiplier λ_i and set them to 0

$$\frac{\partial g(\theta)}{\theta_{ij}} = 0 \tag{12}$$

$$\frac{\partial g(\theta)}{\lambda_i} = 0 \tag{13}$$

which leads to:

$$\hat{\theta}_{ij} = \frac{c_{ij}}{n} \tag{14}$$

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

Bayes' Theorem

$$P(\theta|S) = \frac{P(S|\theta)P(\theta)}{P(S)} \tag{15}$$

Each term in Bayes' theorem has a conventional name:

1. $P(S|\theta)$ – the conditional probability of S given θ, also called the likelihood.
2. $P(\theta)$ – the prior probability or marginal probability of θ.
3. $P(\theta|S)$ – the conditional probability of θ given S, also called the posterior probability of θ
4. $P(S)$ – the marginal probability of S, and acts as a normalizing constant.

Maximum a posteriori (MAP) estmation

- MAP (or posterior mode) estimation of θ is

$$\hat{\theta}_{\mathrm{MAP}}(S) = \arg\max_{\theta} P(\theta|S_1, S_2, \cdots, S_n) \quad (16)$$

$$= \arg\max_{\theta} \log L(\theta) + \log P(\theta) \quad (17)$$

- Assume $P(\theta) = \prod_{i=1}^{w} P(\theta_i)$ (independence of θ_i at diffferent position i).
- Model $P(\theta_i)$ with a Dirichlet distribution

$$(\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4}) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3, \alpha_4). \quad (18)$$

Dirichlet Distribution

- Probability density function of Dirichlet distribution Dir(α) of order $K \geq 2$:

$$p(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_k^{\alpha_i - 1} \quad (19)$$

  for all $x_1, \cdots, x_K > 0$ and $\sum_{i=1}^{K} x_i = 1$. The density is zero outside this open $(K-1)$-dimensional simplex.
  $\alpha = (\alpha_1, \cdots, \alpha_K)$ are parameters with $\alpha_i > 0$ for all i.

- $B(\alpha)$, the normalizing constant, is the multinomial beta function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \quad (20)$$

Gamma function

- Gamma function for positive real z

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{21}$$

-

$$\Gamma(z+1) = z\Gamma(z) \tag{22}$$

- If n is a positive integer, then

$$\Gamma(n+1) = n! \tag{23}$$

Properties of Dirichlet distribution

- Dirichlet distribution

$$p(x_1, \cdots, x_K; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{24}$$

- Expectation, define $\alpha_0 = \sum_{i=1}^{K} \alpha_i$,

$$E[X_i] = \frac{\alpha_i}{\alpha_0} \tag{25}$$

- Variance

$$Var[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{26}$$

- Co-variance

$$Cov[X_i X_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{27}$$

Posterior Distribution

- Conditional probability:
  $P(S_1, S_2, \cdots, S_n|\theta) = \prod_{i=1}^{w} \prod_{j=1}^{4} \theta_{ij}^{c_{ij}}$
- Prior probability: $p(\theta_{i1}, \cdots, \theta_{i4}; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{4} \theta_i^{\alpha_i - 1}$
- Posterior probability:

$$P(\theta_i|S_1, \cdots, S_n) = \mathrm{Dir}(c_{i1} + \alpha_1, c_{i2} + \alpha_2, c_{i3} + \alpha_3, c_{i4} + \alpha_4)$$

- Maxmium a posteriori estimate:

$$\theta_i^{\mathrm{MAP}} = \frac{c_{ij} + \alpha_i - 1}{n + \alpha_0 - 4} \tag{28}$$

where $\alpha_0 \equiv \sum_i \alpha_i$.

Mixture of sequences

- Suppose we have a more difficult situation: Among the set of n given sequences, $S_1, S_2, \cdots, S_n$, only a subset of them are generated by a weight matrix model θ. 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

$$p^0 = (p_A^0, p_C^0, p_G^0, p_T^0) = (p_1^0, p_2^0, p_3^0, p_4^0) \tag{29}$$

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:

$$z_i = \begin{cases} 1 & \text{if } S_i \text{ is generated by } \theta \\ 0 & \text{if } S_i \text{ is generated by } \theta^0 \end{cases} \tag{30}$$

  for all $i = 1, 2, \cdots, n$.
- Then, the likelihood of observing the n sequences

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

Maximum Likelihood

- Find the joint probability of sequences and the labels

$$\begin{aligned} P(S, z|\theta, \theta^0) &= P(S|z, \theta, \theta^0)P(z) \\ &= \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1 - z_i)P(S_i|\theta^0)] \end{aligned}$$

where $z \equiv (z_1, \cdots, z_n)$ and $P(z) = \prod_i P(z_i)$.

- Marginalize over labels to derive the likelihood

$$L(\theta) = P(S|\theta, \theta^0) = \prod_{i=1}^{n} [P(z_i = 1)P(S_i|\theta) + P(z_i = 0)P(S_i|\theta^0)]$$

- Maximum likelihood estimate: $\hat{\theta}_{ML} = \arg\max_{\theta} \log L(\theta))$

Maximum Likelihood

- Find the joint probability of sequences and the labels

$$
\begin{aligned}
P(S, z|\theta, \theta^0) &= P(S|z, \theta, \theta^0)P(z) \\
&= \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1 - z_i)P(S_i|\theta^0)]
\end{aligned}
$$

  where $z \equiv (z_1, \cdots, z_n)$ and $P(z) = \prod_i P(z_i)$.

- Marginalize over labels to derive the likelihood

$$
L(\theta) = P(S|\theta, \theta^0) = \prod_{i=1}^{n} [P(z_i = 1)P(S_i|\theta) + P(z_i = 0)P(S_i|\theta^0)]
$$

- Maximum likelihood estimate: $\hat{\theta}_{ML} = \arg\max_{\theta} \log L(\theta))$

Lower bound on the $L(\theta)$

- Log likelihood function

$$\log L(\theta) = \sum_{i=1}^{n} \log \left[P(z_i = 1) P(S_i | z_i = 1) + P(z_i = 0) P(S_i | z_i = 0) \right]$$

where $P(S_i|z_i = 1) = P(S_i|\theta)$ and $P(S_i|z_i = 0) = P(S_i|\theta_0)$.

- Jensen's inequality:

$$\log(q_1 x + q_2 y) \geq q_1 \log(x) + q_2 \log(y)$$

for all $q_1, q_2 \geq 0$ and $q_1 + q_2 = 1$.

EM-algorithm

- Lower bound on $\log L(\theta)$.

$$\log L(\theta) \geq \sum_{i=1}^{N} \{q_i \log \frac{P(z_i = 1)P(S_i|z_i = 1)}{q_i} +$$

$$(1 - q_i) \log \frac{P(z_i = 0)P(S_i|z_i = 0)}{1 - q_i}\} \equiv \sum_{i=1}^{N} \phi(q_i, \theta)$$

- Expectation-Maximization: Alternate between two steps:
  - E-step

$$\hat{q}_i = \arg\max_{q_i} \phi(q_i, \hat{\theta})$$

  - M-step

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \phi(\hat{q}_i, \theta)$$

E-Step

- Auxiliary function

$$\phi(q_i, \theta) = q_i \log \frac{P(z_i = 1)P(S_i|z_i = 1)}{q_i} + (1-q_i) \log \frac{P(z_i = 0)P(S_i|z_i = 0)}{1 - q_i}$$

- E-step

$$\hat{q}_i = \arg\max_{q_i} \phi(q_i, \hat{\theta})$$

which leads to

$$\hat{q}_i = \frac{P(z_i = 1)P(S_i|z_i = 1)}{P(z_i = 1)P(S_i|z_i = 1) + P(z_i = 0)P(S_i|z_i = 0)} = P(z_i|S_i)$$

M-Step

- Auxiliary function

$$\phi(q_i, \theta) = q_i \log \frac{P(z_i = 1)P(S_i|z_i = 1)}{q_i} + (1-q_i) \log \frac{P(z_i = 0)P(S_i|z_i = 0)}{1 - q_i}$$

- M-step

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \sum_{i=1}^{N} \phi(\hat{q}_i, \theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \hat{q}_i [\log P(S_i|\theta) + (1 - \hat{q}_i) \log P(S_i|\theta_0)] \end{aligned}$$

which leads to

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} \hat{q}_k I(S_{ki} = j)}{\sum_{k=1}^{N} \hat{q}_k}$$

where $I(a)$ is an indicator function: $I(a) = 1$ if a is true and 0 $o.w.$

Summary of EM-algorithm

- Initialize parameters θ.
- Repeat until convergence
  - E-step: estimate the expected values of labels, given the current parameter estimate

$$\hat{q}_i = P(z_i|S_i)$$

  - M-step: re-estimate the parameters, given the expected estimates of the labels

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} \hat{q}_k I(S_{ki} = j)}{\sum_{k=1}^{N} \hat{q}_k}$$

The procedure is guaranteed to converge to a *local maximum* or *saddle point* solution.

What about MAP estimate?

Consider a Dirichlet prior distribution on θ_i:

$$\theta_i = \mathrm{Dir}(\alpha) \quad \forall i = 1, \cdots, w$$

- Initialize parameters θ.
- Repeat until convergence
  - E-step: estimate the expected values of labels, given the current parameter estimate

$$\hat{q}_i = P(z_i|S_i)$$

  - M-step: re-estimate the parameters, given the expected estimates of the labels

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} \hat{q}_k I(S_{ki} = j) + \alpha_j - 1}{\sum_{k=1}^{N} \hat{q}_k + \alpha_0 - 4}$$

Different methods for parameter estimation

So far, we have introduced two methods: ML and MAP

- Maximum likelihood (ML)

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} P(S|\theta)$$

- Maximum a posterior (MAP)

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} P(\theta|S)$$

- Bayes Estimator

$$\hat{\theta} = \arg\max_{\theta} E\left[\hat{\theta}(S) - \theta)^2\right]$$

that is the one minimizing MSE(mean square error).

$$\hat{\theta} = E[\theta|S] = \int \theta P(\theta|S) d\theta$$

Joint distribution

- Joint distribution of labels and sequences

$$\begin{aligned} P(S, z|\theta, \theta^0) &= P(S|z, \theta, \theta^0)P(z) \\ &= \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1 - z_i)P(S_i|\theta^0)] \end{aligned}$$

- Joint distribution of S, z, θ and θ_0

$$P(S, z, \theta, \theta_0) = \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1 - z_i)P(S_i|\theta^0)]P(\theta)P(\theta_0)$$

- Find the joint distribution of S and z

$$P(S, z) = \int_\theta \int_{\theta_0} \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1 - z_i)P(S_i|\theta^0)]P(\theta)P(\theta_0)d\theta_0$$

Posterior distribution of labels

- Posterior distribution of z:

$$
\begin{aligned}
P(z|S) &= \int_\theta \prod_{i=1}^{n} P(z_i)[z_i P(S_i|\theta) + (1-z_i)P(S_i|\theta^0)]P(\theta)d\theta / P(S) \\
&\sim q^m(1-q)^{n-m} \prod_{i=1}^{w} \left[\int_{\theta_i} \prod_{j=1}^{4} \theta_{ij}^{n_{ij}} P(\theta_i) d\theta_i \right] \frac{B(n_0+\alpha_0)}{B(\alpha_0)} \\
&\sim q^m(1-q)^{n-m} \prod_{i=1}^{w} \left[\frac{B(n_i+\alpha)}{B(\alpha)} \right] \frac{B(n_0+\alpha_0)}{B(\alpha_0)}
\end{aligned}
$$

where $m = \sum_{i=1}^n z_i$, $P(z_i = 1) = q$, $P(\theta_i) = \mathrm{Dir}(\alpha)$, $P(\theta_0) = \mathrm{Dir}(\alpha_0)$ are Dirichlet priors, and
$n_{ij} \equiv \sum_{k=1}^n z_k I(S_{ki} = j)$ is the number of letter j at position i among the sequences with label 1. $n_i \equiv (n_{i1}, \cdots, n_{i4})$.
$n_{0,j} \equiv \sum_{k=1}^n (1-z_k) \sum_{i=1}^w I(S_{ki} = j)$ and $n_0 \equiv (n_{0,1}, \cdots, n_{0,4})$.

Sampling

- Posterior distribution of z:

$$P(z|S) \sim q^m(1-q)^{n-m} \prod_{i=1}^{w} \frac{B(n_i+\alpha)}{B(\alpha)} \frac{B(n_0+\alpha_0)}{B(\alpha_0)}$$

- Posterior distribution of z_k conditioned on all other labels $z_{-k} \equiv \{z_i | i = 1, \cdots, n, i \neq k\}$:

$$P(z_k = 1 | z_{-k}, S) \sim q \prod_{i=1}^{w} \left[\frac{B(n_{-k,i} + \Delta(S_{ki}) + \alpha)}{B(n_{-k,i} + \alpha)} \right]$$

where $n_{-k,ij} \equiv \sum_{l=1, l\neq k}^{n} z_l I(S_{li} = j)$ is the number of letter j at position i among all sequences with label 1 excluding the k^{th} sequence. $n_{-k,i} \equiv (n_{-k,i1}, \cdots, n_{-k,i4})$.
$\Delta(l) = (b_1, \cdots, b_4)$ with $b_j = 1$ for $j = S_{ki}$ and otherwise 0.

Posterior distribution

- Posterior distribution of z_k conditioned on z_{-k}:

$$P(z_k = 1 | z_{-k}, S) \sim q \prod_{i=1}^{w} \frac{n_{-k,iS_{ki}} + \alpha_{S_{ki}} - 1}{\sum_j [n_{-k,ij} + \alpha_j - 1]} = q \prod_{i=1}^{w} \theta_{iS_{ki}}$$

  Note that $\theta_{iS_{ki}}$ is same as the MAP estimate of the frequency weight matrix using all sequences with label 1 excluding the k^{th} sequence.

- Similarly

$$P(z_k = 0 | z_{-k}, S) \sim (1-q) \prod_{i=1}^{w} \frac{n_{-k,0S_{ki}} + \alpha_{0,S_{ki}} - 1}{\sum_j [n_{-k,0j} + \alpha_{0,j} - 1]} = (1-q) \prod_{i=1}^{w} \theta_{0,S_{ki}}$$

  $\theta_{0,S_{ki}}$ is same as the MAP estimate of the background distribution.

Gibbs sampling

- Posterior probability

$$
\begin{aligned}
P(z_k = 1 | z_{-k}, S) &\sim \prod_{i=1}^{w} \theta_{iS_{ki}} \\
P(z_k = 0 | z_{-k}, S) &\sim (1-q) \prod_{i=1}^{w} \theta_{0,S_{ki}}
\end{aligned}
$$

- Gibbs sampling

$$
P(z_k = 1 | z_{-k}, S) = \frac{q \prod_{i=1}^{w} \theta_{iS_{ki}}}{q \prod_{i=1}^{w} \theta_{iS_{ki}} + (1-q) \prod_{i=1}^{w} \theta_{0,S_{ki}}}
$$

Gibbs sampling

- Initialize labels z: Assign the value of z_i randomly according to $P(z_i = 1) = q$ for all $i = 1, \cdots, n$.
- Repeat until converge
  - Repeat from $i = 1$ to n
    - Update θ matrix using the MAP estimate (excluding i^{th} sequence)
    - Sample the value of z_i