

CS279, Autumn 2007 – Problem Set #1 Solutions

1. (a) K_{xz} determines the delay between the start of production of X and the start of production of Z . Increasing K_{xz} will move the beginning of the pulse to a later time, but will not affect how long it takes before repression begins. This will cause the pulse to become narrower and have a lower peak. If K_{xz} becomes sufficiently large, no pulse will take place at all because the repression of Z by Y will occur before Z is activated by X .
 K_{xy} and K_{yz} determine how long the rising edge of the pulse lasts. Increasing either one means it will take longer for Y to reach the critical level needed for repression of Z , leading to a longer rising edge of the pulse (and a higher peak level of Z).
 β determines how long it takes before X reaches the level needed to activate either Y or Z . Decreasing β results in a longer delay before the pulse begins. If $K_{xz} < K_{xy}$, decreasing β makes the pulse longer; if the reverse is true, decreasing β makes the pulse shorter.
- (b) To turn the genes on order, we can require

$$K_{xz}^1 < K_{xz}^2 < \dots < K_{xz}^n$$

To turn the genes off in order, we can require

$$K_{yz}^1 < K_{yz}^2 < \dots < K_{yz}^n$$

To turn the genes off in reverse order, we can require

$$K_{yz}^1 > K_{yz}^2 > \dots > K_{yz}^n$$

2. A few possible answers for each type of assay are listed below:
- (a) The regulation of g by t evolved recently and the motif is present in only one species.
The motif for t is highly tolerant to substitutions and is therefore free to evolve rapidly.
The motif for t is short, and so occurs often by chance, causing the sequence to not appear statistically overrepresented.
- (b) Under the experiment conditions (growth medium, tissue type, etc.), t does not regulate g , but does so under other circumstances.
The sequence to which t binds was not present on the ChIP-chip microarray.
The binding of t to g 's promoter region is highly transient and was not picked up by the experiment.
- (c) t binds to a site in g 's promoter, but requires a cofactor that was not present in the experiment to do so.
 t does not bind directly to the DNA upstream of g 's coding region, but instead acts through an intermediate protein.
- (d) Under the experiment conditions (growth medium, tissue type, etc.), t does not regulate g , but does so under other circumstances.
Redundancy in the transcription network compensates for the absence of t and maintains g 's normal expression level.
3. (a) The method of Beer and Tavazoie predicts which transcription factors regulate a particular gene. We could use this information to dramatically reduce the number of variables in the optimization problem described in the Bussemaker paper by eliminating all variables B_{fg} for which transcription factor f is not predicted to regulate gene g (that is, enforce $B_{fg} = 0$). This leads to a reduction in model complexity, and we would expect it to improve our generalization performance assuming the predictions from Beer and Tavazoie are sufficiently reliable.

- (b) The original Bussemaker formulation represents the expression of a gene g under condition t as

$$E_{gt} = \alpha_{0t} + \sum_f \alpha_{ft} B_{fg}$$

We could extend this to account for pairwise interactions among transcription factors as shown below:

$$E_{gt} = \alpha_{0t} + \sum_f \alpha_{ft} B_{fg} + \sum_{f_1, f_2} \alpha_{f_1 f_2 t} B_{f_1 g} B_{f_2 g}$$

Now we have a linear model over pairs of transcription factors. The disadvantage of formulating the learning problem this way is that the number of variables we have to learn is quadratic, rather than linear, in the number of transcription factors.

4. (a) We can use the objective function given by expression (5) from the Nguyen and D'haeseleer paper:

$$\sum_{g=1}^m \sum_{i=1}^n \left[E_{gi} - \sum_{j \in \Omega_g} M_{gj} A_{ji} \right]^2 + \lambda \sum_{j \in \Omega_g} [M_{gj} - M_{gj}^*]^2$$

Where the second term is a soft penalty for the learned binding strength M_{gj} deviating from the prior estimate for its value M_{gj}^* .

- (b) We assume that M_{gj} can be expressed as a PWM score:

$$M_{gj} = \sum_{k=1}^{L_j} w_j(k, c_{gjk})$$

where $w_j(k, c)$ is the score for having a character c at position k in the binding site of transcription factor j , and c_{gjk} is the character at position k of the binding site for transcription factor j in gene g 's promoter. We are given the c_{gjk} 's and wish to learn the w_j matrices (i.e, the PWMs weights). Our objective becomes

$$\sum_{g=1}^m \sum_{i=1}^n \left[E_{gi} - \sum_{j \in \Omega_g} \left(\sum_{k=1}^{L_j} w_j(k, c_{gjk}) \right) A_{ji} \right]^2 + \lambda \sum_j \sum_{k=1}^{L_j} \sum_c w_j(k, c)^2$$

- (c) We can modify the objective function to sum over species and add an additional term that penalizes the difference between the binding affinities of orthologous binding sites:

$$\sum_{s=1}^S \sum_{g=1}^{m_s} \sum_{i=1}^{n_s} \left[E_{sgi} - \sum_{j \in \Omega_{sg}} M_{sgj} A_{sji} \right]^2 + \lambda \sum_{s=1}^S \sum_{g=1}^{m_s} \sum_{j \in \Omega_{sg}} M_{sgj}^2 + \lambda_2 \sum_{s=1}^S \sum_{g=1}^m \sum_{j \in \Omega_{sg}} \sum_{k \in \alpha_{sgj}} [M_{sgj} - M_k]^2$$

where α_{sgj} is the set of binding sites orthologous to the binding site with affinity M_{sgj} . We could also encourage orthologous binding sites to have similar strengths by using a more realistic penalty based on a phylogenetic tree rather than the simple sum-of-pairs score shown above.