Computational Biology CSE 527 Autumn 2004
Notes on Lecture 7, October 19th

Jonathan Claridge (& Mathias Ganter)

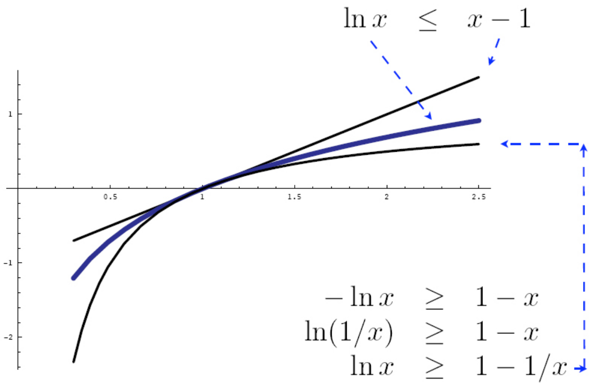
November 10, 2005

Relative Entropy

Consider distributions P and Q on a sample space Ω. The relative entropy (also known as the Kullback-Leibler distance/divergence) between P and Q is defined as

$$H(P\|Q) = \sum_{x\in\Omega} P(x)\log\frac{P(x)}{Q(x)}.$$

When dealing with relative entropy, it is useful to recall the following bounds on natural log:



$$\ln x \quad \leq \quad x-1$$

$$\begin{aligned} -\ln x &\geq 1-x \\ \ln(1/x) &\geq 1-x \\ \ln x &\geq 1-1/x \end{aligned}$$

Figure 1: Bounds on natural log

It is important to note that the relative entropy is non-negative. Although the above sum involves both positive and negative terms, the factor of $P(x)$ gives the positive terms (those with $P(x) > Q(x)$) greater weight than the negative terms (those with $P(x) < Q(x)$). This may be proved rigorously as follows:

$$\begin{aligned} H(P\|Q) &= \sum_{x\in\Omega} P(x)\log\frac{P(x)}{Q(x)} \\ &\geq \sum_{x\in\Omega} P(x)\left(1-\frac{Q(x)}{P(x)}\right) \\ &= \sum_{x\in\Omega}(P(x)-Q(x)) \\ &= \sum_{x\in\Omega} P(x) - \sum_{x\in\Omega} Q(x) \\ &= 1-1 \\ &= 0. \end{aligned}$$

While $H(P\|Q)$ is often called a distance, the relative entropy is not a metric because it is asymmetric and does not satisfy the triangle inequality ($d(x,y) \leq d(x,z) + d(z,y)$). However, it is true that $H(P\|Q) \geq 0$, and $H(P\|Q) = 0$ iff $P = Q$.

Convergence of EM

Recall the quantities we consider for the EM algorithm:

- Visible data x, e.g. the points to be clustered;
- Hidden data y, e.g. which point belongs to which cluster;
- Parameter θ, e.g. description of the various cluster distributions.

The visible data x is fixed. The steps of the algorithm are:

- The E (expectation) step. For fixed parameter θ, we estimate the expected values of the hidden data y.
- The M (maximization) step. Given expected values of the hidden data y, we find parameter θ to maximize $P(x|\theta)$.

We would like to show a way of executing the M step so that our θ estimates will converge. The outline below follows the presentation in Durbin, et al. For any y,

$$\begin{aligned}\log P(x|\theta) &= \log P(x,y|\theta) - \log P(y|x,\theta) \quad \text{(while } x \text{ is fixed)} \\ &= \underbrace{\sum_y P(y|x,\theta^t)\cdot\log P(x,y|\theta)}_{Q(\theta|\theta^t)} - \sum_y P(y|x,\theta^t)\cdot\log P(y|x,\theta)\end{aligned}$$

Letting $Q(\theta|\theta^t) = \sum_y P(y|x,\theta^t)\cdot\log P(x,y|\theta)$ (the first term on the right), we can rewrite this as

$$\log P(x|\theta) = Q(\theta|\theta^t) - \sum_y P(y|x,\theta^t)\cdot\log P(y|x,\theta).$$

In general, this equation is extremely difficult to optimize. However, we can simplify our task by attempting to optimize $Q(\theta|\theta^t)$. Subtracting $\log P(x|\theta^t)$ from the above equation, we obtain

$$\log P(x|\theta) - \log P(x|\theta^t) = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \underbrace{\sum_y P(y|y,\theta^t)\cdot\log\frac{P(y|x,\theta)^t}{P(y|x,\theta)}}_{H(P(y|x,\theta^t)\|P(y|x,\theta))}.$$

The last term here is a relative entropy, and thus it is nonnegative. Consequently, we find that

$$\log P(x|\theta) - \log P(x|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t).$$

So if we find θ that yields a higher value of $Q(\theta|\theta^t)$, this will yield a higher value of $P(x|\theta)$ as well.

There is no guarantee that this will work out perfectly. For instance, the optimization process may get stuck at a bad local maximum that is very far from the global maximum. However, we have at least shown that we obtain some sort of maximum, and issues such as these may be addressed by other means.

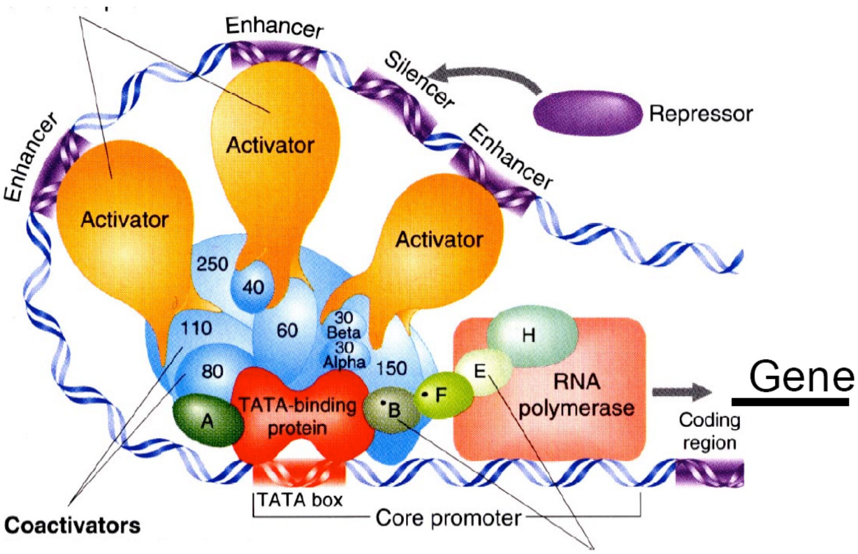


Figure 2: Illustrating the complexity of gene regulation — displaying the TATA box

Sequence Motifs and Weight Matrices

Promoter regions in DNA sequences do not follow a strict pattern. This makes the identification of promoter regions very difficult. Although promoter regions vary, it is often possible to find a DNA sequence (called the *consensus* sequence) to which they are very similar. One such example is the “TATA box,” i.e. a consensus *5’ TATAAT 3’* that is located about 10 bps upstream of the transcription start in E. coli, which is involved in binding RNA polymerase via a TATA binding protein (TBP). This is analogous to the Pribnow box in prokaryotes.

Due to the high variability, exact methods cannot be used for identifying promoter regions by the TATA box. Instead, a pattern search method based on frequencies is used. A table of statistics $f_{b,i}$ can be constructed, where $f_{b,i}$ is the frequency of the base b in position i of the known promoter region suffixes, assuming that positions are independent. Furthermore, we let f_b denote the expected frequency of the base b in the genome, i.e. the background probabilities.

Given a sequence $S = B_1B_2B_3B_4B_5B_6$, the likelihood that it occurs as a TATA-box is given by

$$P(S|S \text{ is a TATA-box}) = \prod_{i=1}^{6} f_{B_i,i}.$$

On the other hand, the likelihood of sequence S occurring as a “non-promoter” is

$$P(S|S \text{ is not a TATA-box}) = \prod_{i=1}^{6} f_{B_i}.$$

Thus, the log-likelihood ratio is

$$\log\left(\frac{P(S|\text{promoter})}{P(S|\text{non-promoter})}\right) = \log\left(\frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}}\right) = \sum_{i=1}^{6} \log\left(\frac{f_{B_i,i}}{f_{B_i}}\right)$$

From the table $f_{B_i,i}$ a scoring matrix can be constructed, with each entry $s_{b,i}$ denoting the score that a sequence should be given for having the base b in the *ith* position. The score $s_{b,i}$ is computed by the following formula:

$$s_{b,i} = \log\left(\frac{f_{b,i}}{f_b}\right).$$

Note in particular that $s_{b,i} < 0$ means that base b has a greater chance of occurring in position i according to the background probabilities.

This attempt has some major drawbacks because it does not exploit all of the known information, such as CG-rich regions, introns/exons, and relations between adjacent bases. But on the other hand, these sequence variations can be considered as a controlling mechanism of expression levels of various genes.

Finally, it can be noted that experiments show ∼80% correlation of log likelihood weight matrix scores to measured binding energy of RNA polymerase to variations on TATAAT consensus. Thus, one could say that the promoter region is very conserved.

Which WMM is the best?

Suppose we have a set of sequences assumed to be generated by a WMM. How do we determine which WMM best describes our data? This is the WMM whose entries are the frequencies-per-position of the bases in our sample sequences.

Neyman-Pearson Theorem

Suppose we are given a sample $x_1, x_2, \ldots, x_n$ from a distribution $f(\cdot|\theta)$, and we wish to test hypothesis $\theta = \theta_1$ versus $\theta = \theta_2$. The Neyman-Pearson Theorem states that we lose no information by looking at the likelihood ratio

$$\frac{f(x_1, x_2, \ldots, x_n|\theta_1)}{f(x_1, x_2, \ldots, x_n|\theta_2)}.$$

Equivalently, we may use the log-likelihood ratio,

$$\log \frac{f(x_1, x_2, \ldots, x_n|\theta_1)}{f(x_1, x_2, \ldots, x_n|\theta_2)} = \log f(x_1, x_2, \ldots, x_n|\theta_1) - \log f(x_1, x_2, \ldots, x_n|\theta_2).$$

This theorem motivates our use of likelihood ratios in weight matrix models, as well as in many subsequent discussions.