Computational Biology CSE 527 Autumn 2004
Notes on Lecture 7, October 20th

Mathias Ganter
mganter@u.washington.edu

24th November 2004

Relative Entropy

Relative entropy, also known as the Kullback-Leibler distance or K-L divergence, between two probability mass functions $P(X)$ and $Q(X)$ is defined as

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

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 $(d(x,y) \leq d(x,z) + d(z,y))$. But, $H(P\|Q) > 0$ and $H(P\|Q) = 0$ iff $P = Q$. This means that the relative entropy is bounded at 0. This can be proved by

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

upper bound: $\log x \leq x - 1$
lower bound: $\log x \geq 1 - \frac{1}{x}$:

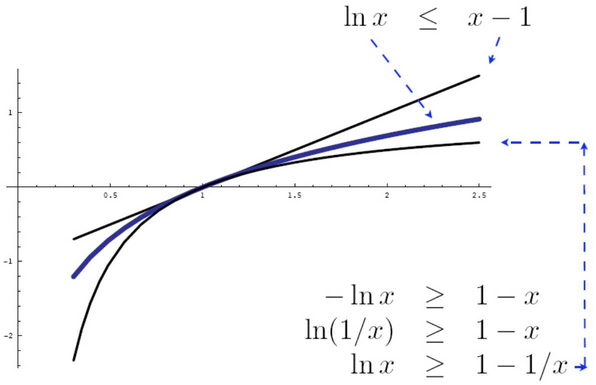

Figure 1: Relative entropy

Convergence of EM

What is convergence?
If X is a measure space, if $\Phi f_1, f_2, ...$ are measurable functions such that $\int_X \Phi < \infty$ and $f_n \leq \Phi$ for each n and if $f_n \rightarrow f$ almost everywhere, then f is integrable and

$$\lim_{n \rightarrow \infty} \int_X f_n = \int_X f.$$

Goal: Maximum likelihood estimate of θ i.e. find θ maximizing $Pr(x|\theta)$ (or $log(Pr(x|\theta))$.

- visible x: e.g., these are the points to be clustered
- hidden y: e.g., saying which point belongs to which cluster
- paramter θ: e.g., describes the various cluster distributions

The outline below follows the presentation in Durbin, et al.

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

Now, this can be rewritten as

$$\log P(X|\theta) = Q(\theta|\theta^t) - \sum_Y P(Y|X,\theta^t) \cdot \log P(Y|X,\theta)$$

and if you apply a little trick, i.e. optimizing Q rather than the whole equation, you'll receive:

$$\begin{aligned} \log P(X|\theta) - \log P(X|\theta^t) &= (1) \\ (2) &= Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \underbrace{\sum_Y P(Y|Y,\theta^t) \cdot \log \frac{P(Y|X,\theta)^t}{P(Y|X,\theta)}}_{\underbrace{H(P(Y|X,\theta^t) \| (P(Y|X,\theta)) \geq 0}_{\text{relative entropy}}} \end{aligned}$$

In addition, (1) ≥ 0 iff (2) ≥ 0. This reveals the difference in Qs! The aim is, that you end up at some local maximum of the objective function by finding a θ that maximizes $Q(\theta|\theta^t)$ by maximizing $Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$. But now, the question comes up, what all this has to do with the EM algorithm. The answer is easy and short:

- the E step gives $P(Y|Y,\theta^t)$, i.e. a probability distribution
- the M step finds a θ that maximizes $Q(\theta|\theta^t)$ by maximizing $Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$

Unfortunatelly, there is no guarantee that this works out perfectly. Among other things, the optimization process can get stuck in a very bad local maximum thus missing all better ones including the global maximum.

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 usually possible to find a DNA sequence (called the *consensus* sequence) to which all of them are very similar, like the TATA box, i.e. a consensus *5' TATAAT 3'* that is located about 10 bps upstream of the transcription start; involved in binding RNA polymerase via a TATA binding protein (TBP). Analogous to the Pribnow box in prokaryotes.

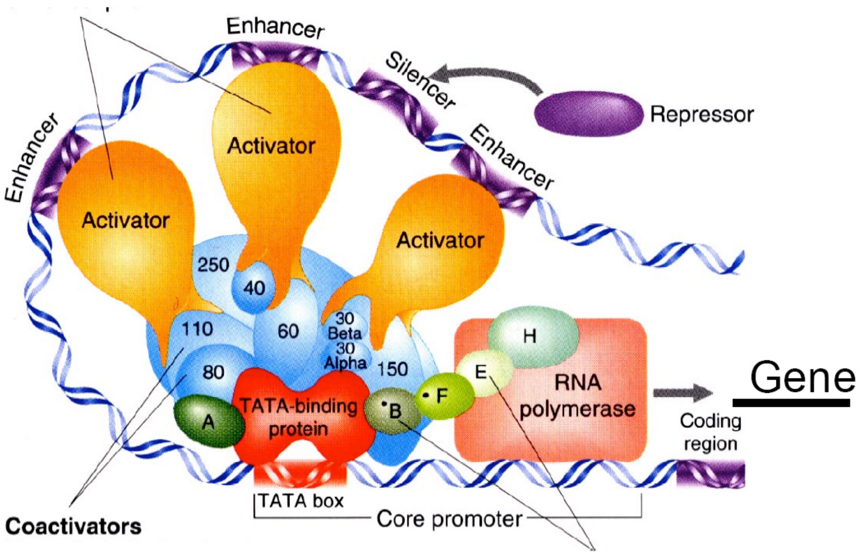

Figure 2: Illustrating the complexity of gene regulation - displaying the TATA box

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 can be constructed, $f_{b,i}$, 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. Let f_b denote the expected frequency of the base b in the genome. Thus, calculating the likelihood of a given sequence $S = B_1B_2B_3B_4B_5B_6$ being a TATA-box:

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

Similarly, the likelihood of observing it, given it is 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 contsructed, 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:

$$\begin{aligned} s_{b,i} &= \log\left(\frac{f_{b,i}}{f_b}\right) \\ s_{b,i} &< 0 \text{ means background probability} \end{aligned}$$

This attempt shows major drawbacks because it does not exploit all of the known information, e.g. CG rich regions, introns/exons, 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, you could say that the promoter region is very conserved.