18.417 Introduction to Computational Molecular Biology	
Lecture 19: November 16, 2004	Scribe: Tushara C. Karunaratna
Lecturer: Ross Lippert	Editor: Tushara C. Karunaratna

Gibbs Sampling

Introduction

Let's first recall the *Motif Finding Problem*: given a set of n DNA sequences each of length t, find the *profile* (a set of l-mers, one from each sequence) that maximizes the consensus score.

We have already seen various naive brute-force approaches for solving this problem.

In this lecture, we will apply a probabilistic method known as *Gibbs Sampling* to solve this problem.

A probabilistic approach to Motif Finding

We can generalize the Motif Finding Problem as follows: given a multivariable scoring function $f(y_1, y_2, \ldots, y_n)$, find the vector $\vec{\mathbf{y}}$ that maximizes f.

Consider a probability distribution p where $p \sim f$. Intuitively, if f is relatively large at the optimum, then if we repeatedly sample from the probability distribution p, then we are likely to quickly encounter the optimum.

Gibbs Sampling provides us a method of sampling from a probability distribution over a large set.

We will use a technique known as *simulated annealing* to transform a probability distribution into one that has a relatively tall peak at the optimum, to ensure that Gibbs sampling is likely to quickly encounter the optimum. In particular, we will observe visually that the probability distribution $p \sim f^{1/T}$, for a sufficiently small T, is a good choice.

Gibbs Sampling

Gibbs Sampling solves the following problem.

- Input: a probability distribution $p(y_1, y_2, \ldots, y_n)$, where each $y_i \in S$. $|S|^n$ may be big, but $|S|$ is assumed to be manageable.
- Output: a random $\vec{\mathbf{y}}$ chosen from the probability distribution p.

Gibbs Sampling uses the technique of Monte Carlo Markov Chain simulation. The idea is to set up a Markov Chain having p as its steady-state distribution, and then simulate this Markov Chain for long enough to be confident that an approximation of the steady-state has been attained. The final state of the simulation approximately represents a sample from the steady-state distribution.

Let's now define our Markov Chain. The set of states of our Markov Chain is S^n. Transitions exist only between states differing in at most one coordinate. For states $\vec{\mathbf{y}} = (y_1, \ldots, y_m, \ldots, y_n)$ and $\vec{\mathbf{y}}' = (y_1, \ldots, y'_m, \ldots, y_n)$, we define the transition probability $T(\vec{\mathbf{y}} \to \vec{\mathbf{y}}') = \frac{1}{n} \frac{p(y_1,\ldots,y'_m,\ldots,y_n)}{\sum_{y_m} p(y_1,\ldots,y_m,\ldots,y_n)}$.

We now show that the distribution p is a steady-state distribution of our Markov Chain.

Recall that the definining property of a steady-state distribution π is

$$\pi T = \pi$$

This property is known as *global balance*.
The stronger property

$$\pi(\vec{\mathbf{y}})T(\vec{\mathbf{y}} \to \vec{\mathbf{y}}') = \pi(\vec{\mathbf{y}}')T(\vec{\mathbf{y}}' \to \vec{\mathbf{y}})$$

is known as *detailed balance*. We can see that detailed balance implies global balance by summing both sides of the detailed balance condition over $\vec{\mathbf{y}}'$:

$$\begin{aligned}
\sum_{\vec{\mathbf{y}}'} \pi(\vec{\mathbf{y}})T(\vec{\mathbf{y}} \to \vec{\mathbf{y}}') &= \sum_{\vec{\mathbf{y}}'} \pi(\vec{\mathbf{y}}')T(\vec{\mathbf{y}}' \to \vec{\mathbf{y}}) \\
\pi(\vec{\mathbf{y}}) \sum_{\vec{\mathbf{y}}'} T(\vec{\mathbf{y}} \to \vec{\mathbf{y}}') &= \sum_{\vec{\mathbf{y}}'} \pi(\vec{\mathbf{y}}')T(\vec{\mathbf{y}}' \to \vec{\mathbf{y}}) \\
\pi(\vec{\mathbf{y}}) &= (\pi T)(\vec{\mathbf{y}})
\end{aligned}$$

Therefore, let's just check whether p satisfies detailed balance. If $\vec{\mathbf{y}}'$ differs from $\vec{\mathbf{y}}$ in zero or more than one place, then detailed balance trivially holds (in the latter case,

both sides of the detailed balance condition evaluate to zero). So, suppose that $\vec{\mathbf{y}}'$ differs from $\vec{\mathbf{y}}$ in only one place, say coordinate m. The left-hand-side of the detailed balance condition evaluates to $p(\vec{\mathbf{y}})\frac{1}{n}\frac{p(\vec{\mathbf{y}}')}{\sum_{y_m} p(y_1,\ldots,y_m,\ldots,y_n)}$. The right-hand-side evaluates to $p(\vec{\mathbf{y}}')\frac{1}{n}\frac{p(\vec{\mathbf{y}})}{\sum_{y_m} p(y_1,\ldots,y_m,\ldots,y_n)}$. The two sides are equal, as desired.

Therefore, p is indeed the steady-state distribution of our Markov Chain.

Scoring profiles

Let's investigate a probabilistic approach to scoring profiles, as an alternative to simply using the consensus score.

We assume a background frequency P_x for character x.

Let $C_{x,i}$ denote the number of occurences of character x in the i^{th} column of the profile. We call this the *profile matrix*.

Then, in the background, the probability that a profile has profile matrix C is given by

$$\begin{aligned} prob(C) &= \prod_{i=0}^{l-1} \binom{n}{C_{a,i}\ C_{c,i}\ C_{g,i}\ C_{t,i}} P_a^{C_{a,i}} P_c^{C_{c,i}} P_g^{C_{g,i}} P_t^{C_{t,i}} \\ &\sim \prod_{x,i} \frac{1}{C_{x,i}!} P_x^{C_{x,i}} \end{aligned}$$

Since the profile corresponding to the actual motif locations should have small background probability, we assign

$$\begin{aligned} score(C) &\sim 1/prob(C) \\ &\sim \prod_{x,i} C_{x,i}! P_x^{-C_{x,i}} \end{aligned}$$

Now, $\log(n!) = \Theta(n \log n)$. Therefore,

$$score(C) \sim \exp\left(\sum_{x,i} C_{x,i} \log \frac{C_{x,i}}{P_x}\right)$$

The exponent is known as the *entropy* of the profile.

In summary, maximizing the entropy, rather than the consensus score, is a statistically more adequate approach of finding motifs.

Motif finding via Gibbs Sampling

Here is pseudocode for Motif Finding using the Gibbs Sampling technique.

1. Randomly generate a start state $y_1, \ldots, y_n$.
2. Pick m uniformly at random from $1, \ldots, n$.
3. Replace y_m with y'_m picked randomly from the distribution that assigns relative weight $1/prob(C(y_1, \ldots, y'_m, \ldots, y_n))$ to y'_m.
4. <do whatever with the sample>
5. Goto step 2.

Note that we are just doing a simulation of the Markov Chain defined by the Gibbs Sampling technique.

Simulated Annealing

Annealing is a process by which glass is put into a highly durable state by a process of slow cooling.

We can use the same idea here: to amplify the probability of sampling at the optimum of a probability distribution p, we instead sample from $p^{1/T}$ where $T \to 0$.

Figure 19.1 shows us a graph of a probability distribution p. The optimum occurs at state 4, but there are other peaks that have significantly large height.

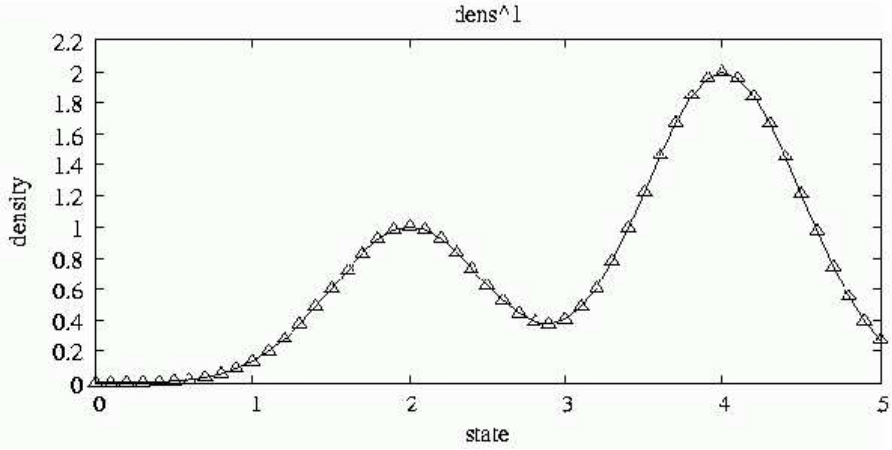

Figure 19.1: Graph of a probability distribution p.

Figures 19.2 and 19.3 show the graphs of the probability distributions p^5 and p^{50} respectively. The height of the peak at state 4 has increased considerably with respect to the heights of the other peaks.

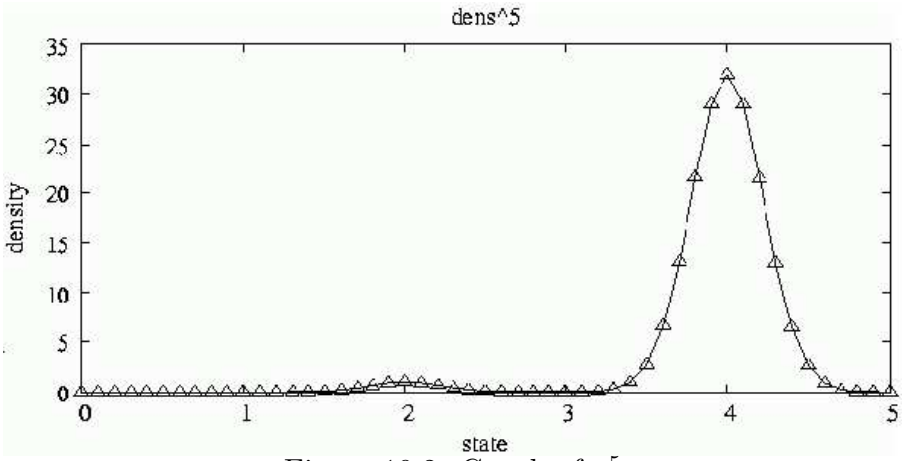

Figure 19.2: Graph of p^5.

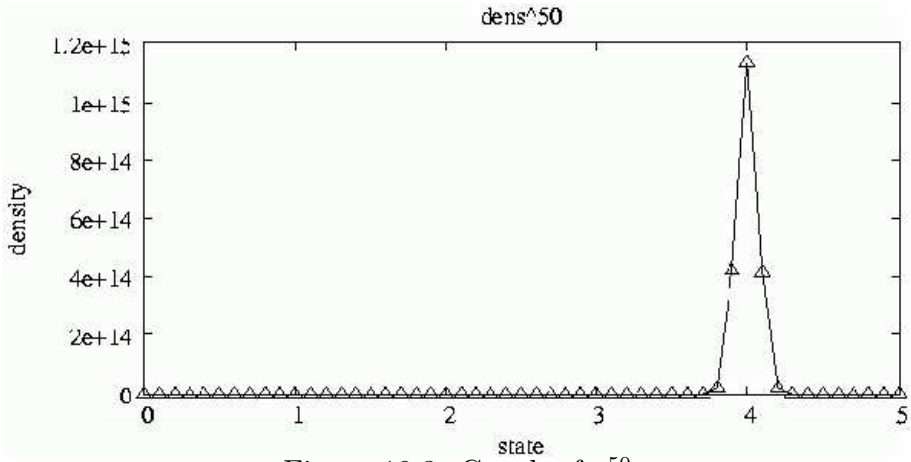

Figure 19.3: Graph of p^{50}.

How do we find the right T? Here are two possible approaches: we can either drop T by a small amount after reaching steady-state, or we can drop T by a small amount at each step.

Some questions that we didn't answer

- For how long should we run the Markov Chain?
- How often can we sample?