

## Gibbs Sampling in Motif Finding

## Gibbs Sampling

- Goal: Find most probable pattern by sampling from motif probabilities to maximize ratio of model to background probabilities [Lawrence 93]
- Given:
  - $X_1, \dots, X_N$
  - motif length  $k$
- Find:
  - Model  $M, B$
  - Locations  $a_1, \dots, a_N$  in  $X_1, \dots, X_N$

Motif finding

Bioinformatics, Fall 2011 2

## Algorithm

- Select random locations  $Z$  for motifs in sequences
- Repeat
  - Isolate a sequence  $y$  at random
  - Compute  $M, B$  based on  $Z$  and using all sequences except  $y$
  - Sample a new motif position for  $y$ 
    - Compute a weight for each starting position
    - Select at random using these weights
  - until enough samples have been obtained

Motif finding

Bioinformatics, Fall 2011 3

## Theoretical underpinnings

- Markov Chain Monte Carlo (MCMC)
- Suppose we wish to sample from a joint probability distribution  $p(X_1, X_2, \dots, X_n)$ .
- Set up a Markov Chain whose stationary distribution is  $p$ .
- Any random walk on the Markov Chain will eventually generate samples from  $p$ .
- Use these samples to understand the joint distribution.
  - Compute parameters such as mean, expected value of a function of the distribution, marginal distribution, or the state with the highest probability (for motif finding).
- Aperiodicity and irreducibility imply stationarity.
  - Were the mutation matrices aperiodic and irreducible?

Motif finding

Bioinformatics, Fall 2011 4

## Setting up the Markov Chain $Q$

- State of the distribution  $s(x_1, \dots, x_i, \dots, x_n)$ .
- Transitions between states that differ in only one position.
- For states  $s(x_1, \dots, x_i, \dots, x_n)$  and  $t(x_1, \dots, x_i', \dots, x_n)$ ,
  - $Q(t, s)$  = transition probability from  $s$  to  $t$  =  $(1/n) [p(x_1, \dots, x_i', \dots, x_n) / \sum p(x_1, \dots, *, \dots, x_n)]$
- Change one dimension at a time.
- Show that the above Markov Chain has  $p$  as its stationary distribution.

Motif finding

Bioinformatics, Fall 2011 5

## Detailed Balance

- Detailed balance requirement
  - $p(s) \cdot Q(t, s) = p(t) \cdot Q(s, t)$
- Implies stationarity of  $p$ 
  - $\sum_t p(s) \cdot Q(t, s) = \sum_t Q(s, t) \cdot p(t)$
  - $p(s) \cdot \sum_t Q(t, s) = \sum_t Q(s, t) \cdot p(t)$
  - $p(s) = [Qp](s)$
  - $p = Qp$
- Show that detailed balance is met.

Motif finding

Bioinformatics, Fall 2011 6

## Proof of Detailed Balance

- $$\begin{aligned}
 p(s) \cdot Q(t,s) &= \\
 &= p(x_1, \dots, x_i, \dots, x_n) \cdot (1/n) [p(x_1, \dots, x_i', \dots, x_n) / \sum p(x_1, \dots, *, \dots, x_n)] \\
 &= p(x_1, \dots, x_i', \dots, x_n) \cdot (1/n) [p(x_1, \dots, x_i, \dots, x_n) / \sum p(x_1, \dots, *, \dots, x_n)] \\
 &= p(t) \cdot Q(s,t)
 \end{aligned}$$
- Detailed balance implies reversibility of Markov chains.

Motif finding

Bioinformatics, Fall 2011 7

## Gibbs sampling is an instance of Metropolis-Hastings

- Markov chain defined on the space of outcomes
- Probability of outcome  $s$  is  $p(s)$
- $Q(s,t)$  = probability of transition from  $s$  to  $t$
- Postulate a function  $f(t,s)$  such that  $\sum_t f(t,s) = 1$
- At state  $s$ , choose a random neighbor  $t$  with probability  $f(t,s)$  and accept it with probability
  - $\min[1, p(t) \cdot f(s,t) / (p(s) \cdot f(t,s))]$
  - $Q(t,s) = f(t,s) \min[1, p(t) \cdot f(s,t) / (p(s) \cdot f(t,s))]$
  - $Q(s,t)$  is set so that column sums to 1
- Detailed balance is satisfied:
  - $$\begin{aligned}
 p(s)Q(t,s) &= p(s) \cdot \min[f(t,s), p(t) \cdot f(s,t) / p(s)] \\
 &= \min[p(s) \cdot f(t,s), p(t) \cdot f(s,t)] \\
 &= p(t) \cdot \min[f(s,t), p(s) \cdot f(t,s) / p(t)] \\
 &= p(t)Q(s,t)
 \end{aligned}$$
- Implies stationarity of  $p$  for Markov Chain  $Q$ .
- Free to choose an appropriate  $f$ 
  - For Gibbs sampling, for  $s$  and  $t$  differing in  $i$ th component,
    - $f(t,s) = p(i) / (n \sum_u p(i))$ , state  $u$  shares values of non- $i$  components with  $s,t$
    - probability of acceptance = 1

Motif finding

Bioinformatics, Fall 2011 8

## Advantages / Disadvantages

### Advantages:

- Easier to implement
- Less dependent on initial parameters
- More versatile, easier to enhance with heuristics

### Disadvantages:

- More dependent on all sequences to exhibit the motif
- Less systematic search of initial parameter space

### Tool:

- <http://bayesweb.wadsworth.org/gibbs/gibbs.html>

Motif finding

Bioinformatics, Fall 2011 9