

# Case Studies in Bayesian Computation

Michael Newton

IMA September 2003

## Five Cases, Four Lectures!

- 1*a.* Transcription Regulation: Lawrence/Liu Methods
- 1*b.* Cancer Genome Aberrations I: LOH, 1-gene model
2. Phylogenetics: evolutionary trees from aligned sequence
3. Cancer Genome Aberrations II: CGH, network model
4. Gene Expression: differential expression and microarrays

## Common Themes

Data

Model development (likelihood) (note on maturity)

Computation (MCMC, or not)

Inference

Lawrence/Liu et al. Model/Method

Data: multiple DNA or protein sequences, linked by common binding property

Example: upstream regions of 18 E-coli genes all regulated by CRP transcription factor

Good test: in this case the *binding elements* were determined experimentally. Generally not so. **NB!**

Problem: Predict binding elements from sequence data.

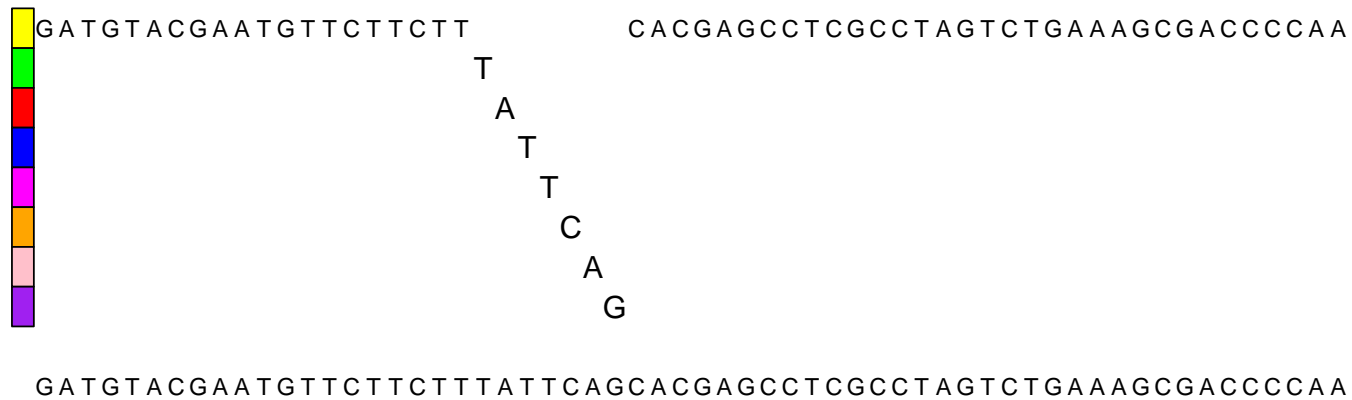
## CRP Data

cyclic AMP receptor protein  
18 sequences of length 105 bases

## Product Multinomial Motif Model

- Data  $D$ :  $K$  sequences, e.g. over  $m = 4$  bases or  $m = 20$  amino acids. Binding element length  $w$  treated as known.
- Alignment variables:  $A = \{A_1, A_2, \dots, A_K\}$  (missing data). These indicate the position of each binding element.
- Parameters  $\theta = (\theta_0, \theta_1, \dots, \theta_w)$ . Each  $\theta_i = (\theta_{i,1}, \dots, \theta_{i,m})$ .  $i = 0$  is background;  $i > 0$  is position within binding element. These parameters define a *motif*.
- Product Multinomial:  $P(D|\theta, A) = \prod_{i=0}^w \prod_{j=1}^m \theta_{i,j}^{c_{i,j}}$  where  $c_{i,j}$  counts the instances of base or a.a.  $j$  at 'position'  $i$ .

## Product Multinomial Motif Model



Repeat to generate each sequence, possibly changing binding element location, but maintaining motif.

## Prior

- Uniform for  $A$
- Product Dirichlet for  $\theta$ :

$$P(\theta) \propto \prod_{i=0}^w \prod_{j=1}^m \theta_{i,j}^{b_j}$$

for *pseudo-counts*  $b_j$

## MCMC

Generate states  $S_1, S_2, \dots$  aiming at target  $P(S|D)$

Propose  $S^* \sim q(S, \cdot)$  according to some move type.

Accept proposal w.p.  $\min(1, r)$  where

$$r = \frac{P(S^*|D) q(S^*, S)}{P(S|D) q(S, S^*)}.$$

Use simple move types for computational feasibility.

Use multiple move types for irreducibility.

Use well-chosen move types for statistical efficiency.

Subsample to simplify output analysis.

Run output analysis on saved states.

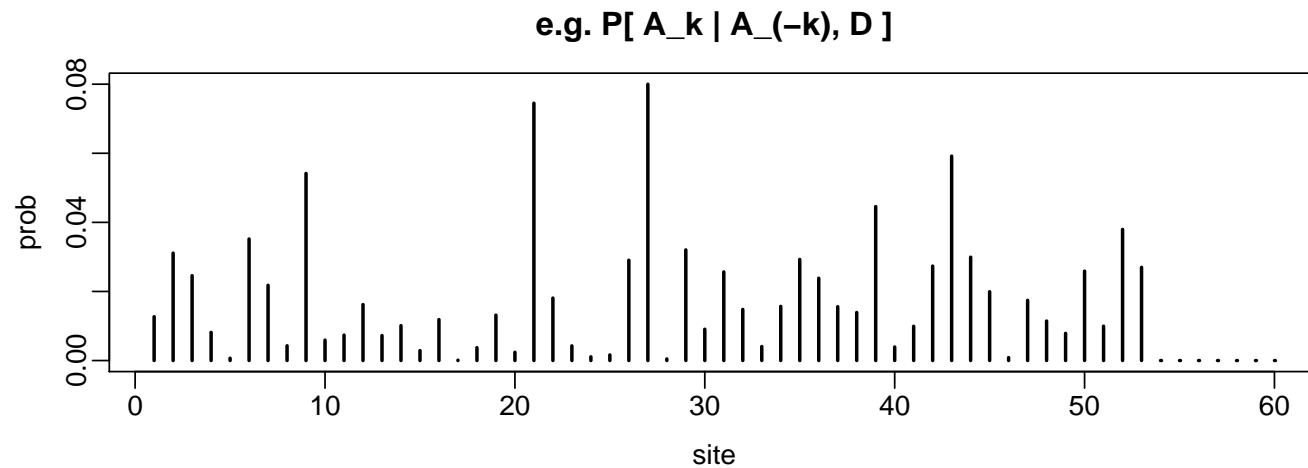
# MCMC

Method I:  $S = (\theta, A)$

- Update  $\theta$  using  $P(\theta|A, D)$ ; Update  $A$  using  $P(A|\theta, D)$

Method II:  $S = A$  (integrate the product Dirichlet prior)

- For each sequence  $k$ , update  $A_k$  using  $P(A_k|A_{(-k)}, D)$
- Shift proposal:  $A^* = A + / - 1$



## Results for CRP

Authors run MCMC as optimization, tracking  $P(D|A)$  over sampled  $A$ 's

## Some Issues

1. Many model extensions: unknown width; multiple motifs, ...
2. Web Interface
3. Homework: combine Dirichlet prior and Product multinomial likelihood to get  $P(A_k | A_{(-k)}, D)$  used in site sampler.