Motif finding with Gibbs sampling example

Colin Dewey

2007.02.13

The Data

- Hidden motif of width 7 in 4 sequences of length 10
- Each motif occurrence differs from consensus (GATTACA) in two positions

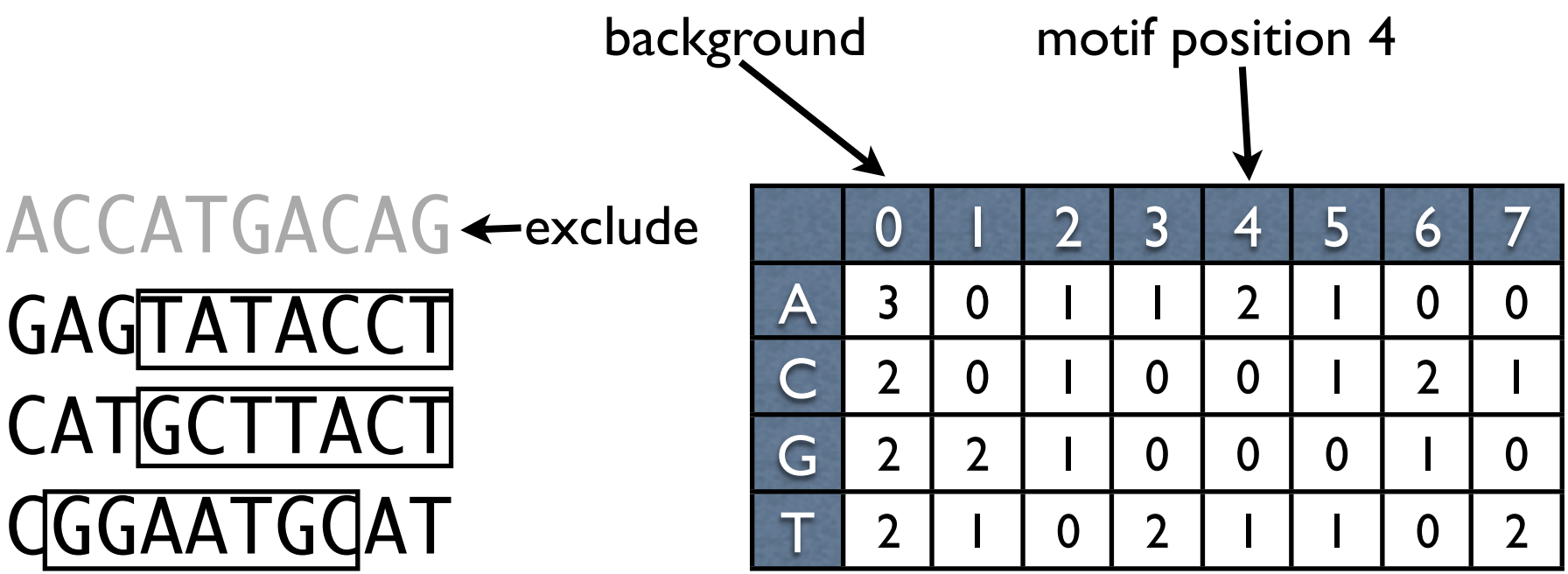
ACCATGACAG
GAGTATACCT
CATGCTTACT
CGGAATGCAT

Initialization

- Choose initial positions of motif at random

Predictive update step

- Update profile matrix based on motif and background frequencies and pseudocounts



	0	1	2	3	4	5	6	7
A	3	0	1	1	2	1	0	0
C	2	0	1	0	0	1	2	1
G	2	2	1	0	0	0	1	0
T	2	1	0	2	1	1	0	2

Predictive update step

- Calculate profile matrix from frequencies and pseudocounts

	0	1	2	3	4	5	6	7
A	3	0	1	1	2	1	0	0
C	2	0	1	0	0	1	2	1
G	2	2	1	0	0	0	1	0
T	2	1	0	2	1	1	0	2

$$p_{1,\mathtt{A}} = \frac{c_{1,\mathtt{A}} + b_{\mathtt{A}}}{N - 1 + B} = \frac{0 + 0.5}{4 - 1 + 2} = 0.1$$

$$p_{1,\mathtt{C}} = \frac{c_{1,\mathtt{C}} + b_{\mathtt{C}}}{N - 1 + B} = \frac{0 + 0.5}{4 - 1 + 2} = 0.1$$

$$p_{1,\mathtt{G}} = \frac{c_{1,\mathtt{G}} + b_{\mathtt{G}}}{N - 1 + B} = \frac{2 + 0.5}{4 - 1 + 2} = 0.5$$

$$p_{1,\mathtt{T}} = \frac{c_{1,\mathtt{T}} + b_{\mathtt{T}}}{N - 1 + B} = \frac{1 + 0.5}{4 - 1 + 2} = 0.3$$

$$p_{0,\mathtt{A}} = \frac{c_{0,\mathtt{A}} + b_{\mathtt{A}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{3 + 0.5}{3(3) + 2} = \frac{7}{22}$$

$$p_{0,\mathtt{C}} = \frac{c_{0,\mathtt{C}} + b_{\mathtt{C}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2 + 0.5}{3(3) + 2} = \frac{5}{22}$$

$$p_{0,\mathtt{G}} = \frac{c_{0,\mathtt{G}} + b_{\mathtt{G}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2 + 0.5}{3(3) + 2} = \frac{5}{22}$$

$$p_{0,\mathtt{T}} = \frac{c_{0,\mathtt{T}} + b_{\mathtt{T}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2 + 0.5}{3(3) + 2} = \frac{5}{22}$$

	0	1	2	3	4	5	6	7
A	0.31	0.1	0.3	0.3	0.5	0.3	0.1	0.1
C	0.23	0.1	0.3	0.1	0.1	0.3	0.5	0.3
G	0.23	0.5	0.3	0.1	0.1	0.1	0.3	0.1
T	0.23	0.3	0.1	0.5	0.3	0.3	0.1	0.5

Sampling step

- For each possible motif start position, calculate ratio of likelihood of next W positions from motif vs. background

	1	2	3	4
A_i	0.16	0.13	0.26	0.017

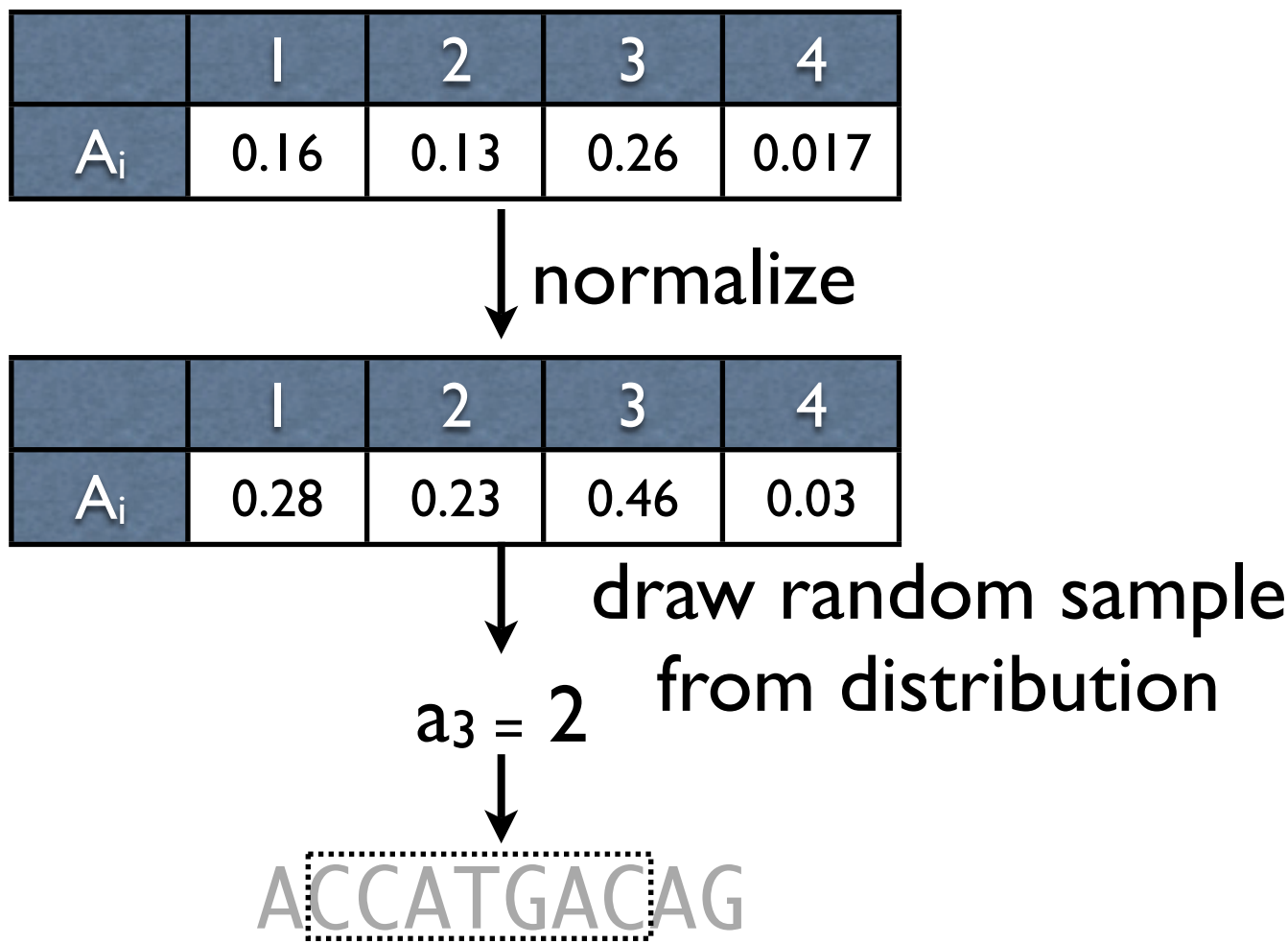
$$A_1 = \frac{p_{1,\mathtt{A}} \cdot p_{2,\mathtt{C}} \cdot p_{3,\mathtt{C}} \cdot p_{4,\mathtt{A}} \cdot p_{5,\mathtt{T}} \cdot p_{6,\mathtt{G}} \cdot p_{7,\mathtt{A}}}{p_{0,\mathtt{A}} \cdot p_{0,\mathtt{C}} \cdot p_{0,\mathtt{C}} \cdot p_{0,\mathtt{A}} \cdot p_{0,\mathtt{T}} \cdot p_{0,\mathtt{G}} \cdot p_{0,\mathtt{A}}} \approx \frac{0.1 \cdot 0.3 \cdot 0.1 \cdot 0.5 \cdot 0.3 \cdot 0.3 \cdot 0.1}{0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31} \approx 0.16$$

ACCATGACAG

	0	1	2	3	4	5	6	7
A	0.31	0.1	0.3	0.3	0.5	0.3	0.1	0.1
C	0.23	0.1	0.3	0.1	0.1	0.3	0.5	0.3
G	0.23	0.5	0.3	0.1	0.1	0.1	0.3	0.1
T	0.23	0.3	0.1	0.5	0.3	0.3	0.1	0.5

Sampling step

- Sample new position *i* in chosen sequence based on A_i



Calculate likelihood

- Calculate likelihood (or some related value) after each iteration
- Iterate:
  - choose sequence
  - predictive update
  - sample new motif position in sequence
- After many iterations, choose motif positions and corresponding profile matrix