Motif finding, EM algorithm, Gibbs sampling

Askar Gafurov

November 30, 2023

Motifs

- Motivation: DNA binding sites for a certain protein
  - E.g. RNA polymerase (in gene expression)
- The protein prefers some locations on DNA, but not a unique sequence
  - E.g. AATATACC, but also AGTATACG or CATATCTC
  - The probability of binding is not constant, some sequences are more likely to bind
- Motif = a table of probabilities for each position of a binding site

$$W = \begin{pmatrix} A: & 0.7 & 0.7 & 0.05 & 0.89 & 0.05 & 0.82 & 0.1 & 0.01 \\ C: & 0.2 & 0.05 & 0.05 & 0.01 & 0.05 & 0.1 & 0.8 & 0.8 \\ G: & 0.05 & 0.2 & 0.05 & 0.05 & 0.10 & 0.03 & 0.05 & 0.1 \\ T: & 0.05 & 0.05 & 0.85 & 0.05 & 0.80 & 0.05 & 0.05 & 0.09 \end{pmatrix}$$

- $\Pr[\text{AGTATACG is binding} \mid W] = 0.7 \cdot 0.2 \cdot 0.85 \cdot 0.89 \cdot 0.80 \cdot 0.82 \cdot 0.8 \cdot 0.1 \approx 0.006$
- $\Pr[\text{AATATACC is binding} \mid W] = 0.7 \cdot 0.7 \cdot 0.85 \cdot 0.89 \cdot 0.80 \cdot 0.82 \cdot 0.8 \cdot 0.8 \approx 0.156$

Generative model of a sequence with a motif

- Goal: define $\Pr[S \mid O, W]$ and $\Pr[O \mid W]$
- $\Pr[O \mid W]$ is easy: binding is equally likely to occur at every position (if we don't know the sequence)
  - $\Pr[O \mid W] := \dfrac{1}{m - L + 1}$, where $m = |S|$
- $\Pr[S \mid O, W]$ is a bit tricky.
  - We already know the prob. of letters at binding positions $(O, O+1, \ldots, O+L-1)$
  - Assign *background frequency* $q(\cdot)$ for letters outside the binding site
    - ⋆ e.g. $q(A) = q(T) = 0.3, q(C) = q(G) = 0.2$
  - Now, the prob. of observing S is a product of probs. for each letter:

$$\begin{aligned}\Pr[S = CCTATTGTATACCTATACC \mid O = 6, W] = \\ = q(C)q(C)q(T)q(A)q(T)\cdot \\ \cdot W[T,1]W[G,2]W[T,3]W[A,4]W[T,5]W[A,6]W[C,7]W[C,8]\cdot \\ \cdot q(T)q(A)q(T)q(A)q(C)q(C) \approx \\ \approx 1.11 \times 10^{-9}\end{aligned}$$

Motif in a larger sequence

- Question: Given a motif $W \in \mathbb{R}^{4 \times L}$ and sequence $S = CCTATTGTATACCTATACC$, where does the binding site starts?
  - Assume there is exactly one binding site
  - Let's denote the binding site start as O.
- We can compute $\Pr[O \mid S, W]$ for each possible value of O.

$$\Pr[O \mid S, W] \overset{\text{Bayes}}{=} \frac{\Pr[S \mid O, W] \cdot \Pr[O \mid W]}{\sum_{O'} \Pr[S \mid O', W] \cdot \Pr[O' \mid W]}$$

- In human words: *compute prob. of observing S given start O and motif W for each value of O, and then normalize them to sum up to* 1.
- Notation for the (eventual) renormalization:

$$\Pr[O \mid S, W] \sim \Pr[S \mid O, W] \cdot \Pr[O \mid W]$$

Motif in a larger sequence

- Now let's compute $\Pr[S = CCTATTGTATACCTATACC \mid O, W]$ for each value of O and then renormalize it:

$$Pr[S|O=1,W] \approx 4.76 \times 10^{-13}$$
$$Pr[S|O=2,W] \approx 7.00 \times 10^{-17}$$
$$Pr[S|O=3,W] \approx 1.04 \times 10^{-14}$$
$$Pr[S|O=4,W] \approx 7.59 \times 10^{-14}$$
$$Pr[S|O=5,W] \approx 2.92 \times 10^{-16}$$
$$Pr[S|O=6,W] \approx 1.11 \times 10^{-09}$$
$$Pr[S|O=7,W] \approx 7.87 \times 10^{-16}$$
$$Pr[S|O=8,W] \approx 1.54 \times 10^{-14}$$
$$Pr[S|O=9,W] \approx 9.19 \times 10^{-17}$$
$$Pr[S|O=10,W] \approx 1.34 \times 10^{-15}$$
$$Pr[S|O=11,W] \approx 6.12 \times 10^{-15}$$
$$Pr[S|O=12,W] \approx 1.67 \times 10^{-09}$$
$$\Sigma \approx 2.78 \times 10^{-09}$$

$$Pr[O=1 \mid S,W] \approx 4.76 \times 10^{-13}/(2.78 \times 10^{-09}) \approx 0.00017$$
$$Pr[O=2 \mid S,W] \approx 7.00 \times 10^{-17}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=3 \mid S,W] \approx 1.04 \times 10^{-14}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=4 \mid S,W] \approx 7.59 \times 10^{-14}/(2.78 \times 10^{-09}) \approx 0.00003$$
$$Pr[O=5 \mid S,W] \approx 2.92 \times 10^{-16}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=6 \mid S,W] \approx 1.11 \times 10^{-09}/(2.78 \times 10^{-09}) \approx 0.39992$$
$$Pr[O=7 \mid S,W] \approx 7.87 \times 10^{-16}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=8 \mid S,W] \approx 1.54 \times 10^{-14}/(2.78 \times 10^{-09}) \approx 0.00001$$
$$Pr[O=9 \mid S,W] \approx 9.19 \times 10^{-17}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=10 \mid S,W] \approx 1.34 \times 10^{-15}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=11 \mid S,W] \approx 6.12 \times 10^{-15}/(2.78 \times 10^{-09}) \approx 0.00000$$
$$Pr[O=12 \mid S,W] \approx 1.67 \times 10^{-09}/(2.78 \times 10^{-09}) \approx 0.59987$$

Quick summary so far

- Motif W = a table of letter probabilities at each position of the site
  - $W[a, j] :=$ Pr[j-th letter of a site is letter a]
- Probability of sequence S with a site at position O, motif $W \in \mathbb{R}^{4 \times L}$ and background letter frequency q is computed as a product:
  - $\Pr[S \mid O, W] = \prod_{j=1}^{O-1} q(S[j]) \cdot \prod_{j=1}^{L} W[S[O+j-1], j] \cdot \prod_{j=O+L}^{m} q(S[j])$
- Probability of site being at position O of sequence S, motif W and b.f. q is computed by renormalizing $\Pr[S \mid O, W]$ (assuming a unique occurrence):
  - $\Pr[O \mid S, W] \sim \Pr[S \mid O, W]$

Motif finding with known O (Two hands)

- Task: Given a vector of sequences $\mathbf{S} = (S_1, \ldots, S_n)$ of length m each, a vector of site starts $\mathbf{O} = (O_1, \ldots, O_n)$ and b.f. q, find the *best* motif W of length L!
  - Assuming that the motif **occurs exactly once** in each sequence
- Example of input data (the sites are shown as red text):

CGACTAAACCACGGA
AGATATAACAAAAAG
AAGTCACCATAAACT
AGTATTCCTATAGCA
TGACACATACCATGG
TAATATACCGCTTAC
TGCTAATAGTCCATA
TAATATACCGTATCT

Motif finding with known O (Two hands)

- *best* = most likely = with the maximum (log-)likelihood
  - Likelihood of W is $\mathcal{L}(W; \mathbf{O}, \mathbf{S}) \stackrel{\text{def.}}{=} \Pr[\mathbf{S}, \mathbf{O} \mid W]$
  - $W^* = \arg\max_{W \in \mathcal{W}} \mathcal{L}(W; \mathbf{O}, \mathbf{S}) = \arg\max_{W \in \mathcal{W}} \ln \mathcal{L}(W; \mathbf{O}, \mathbf{S})$
- Intuition: best W is obtained by counting letter frequencies at the sites
  - $W^*[a, j] \stackrel{??}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{S_i[O_i+j-1]=a} =: \frac{\#_{a,j}(\mathbf{O})}{n}$

CGACTAAACCACGGA
AGATATAACAAAAAG
AAGTCACCATAAACT
AGTATTCCTATAGCA
TGACACATACCATGG
TAATATACCGCTTAC
TGCTAATAGTCCATA
TAATATACCGTATCT

$$W_{\text{counting}} = \begin{pmatrix} 5/8 & 6/8 & 0/8 & 8/8 & 2/8 & 6/8 & 1/8 & 0/8 \\ 2/8 & 1/8 & 1/8 & 0/8 & 0/8 & 0/8 & 7/8 & 7/8 \\ 1/8 & 1/8 & 0/8 & 0/8 & 1/8 & 0/8 & 0/8 & 0/8 \\ 0/8 & 0/8 & 7/8 & 0/8 & 5/8 & 2/8 & 0/8 & 1/8 \end{pmatrix}$$

Let's check the intuition

$$W^* := \arg\max_{W \in \mathcal{W}} \ln \mathcal{L}(W; \mathbf{O}, \mathbf{S}) \overset{\text{def.}}{=} \arg\max_{W \in \mathcal{W}} \ln \Pr[\mathbf{S}, \mathbf{O} \mid W] =$$

$$= \arg\max_{W \in \mathcal{W}} \ln \Pr[\mathbf{S} \mid \mathbf{O}, W] + \ln \Pr[\mathbf{O} \mid W] =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{i=1}^{n} \ln \Pr[S_i \mid O_i, W] =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{i=1}^{n} \left(\sum_{j=1}^{O_i - 1} \ln q(S_i[j]) + \sum_{j=1}^{L} \ln W[S_i[O_i + j - 1], j] + \sum_{j=O_i+L}^{m} \ln q(S_i[j]) \right) =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{j=1}^{L} \sum_{a \in \{A,C,G,T\}} \ln W[a, j] \cdot \sum_{i=1}^{n} \mathbf{1}_{S_i[O_i + j - 1] = a} =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{j=1}^{L} \sum_{a \in \{A,C,G,T\}} \ln W[a, j] \cdot \#_{a,j}(\mathbf{O})$$

Let's check the intuition, p.2

- $W^* = \arg\max_{W \in \mathcal{W}} \sum_{j=1}^{L} \sum_{a \in \{A,C,G,T\}} \ln W[a,j] \cdot \#_{a,j}(\mathbf{O})$
- Each column of W can be optimised independently:
  - $W_j^* = \arg\max_{\substack{x_A,x_C,x_G,x_T \geq 0 \\ \sum x_a = 1}} \sum_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O}) \ln x_a$
- Using the method of Lagrange multipliers, we obtain $x_a^* = \frac{\#_{a,j}(\mathbf{O})}{\sum_a \#_{a,j}(\mathbf{O})} = \frac{\#_{a,j}(\mathbf{O})}{n}$
- Thus indeed $W^*[a,j] = \frac{\#_{a,j}(\mathbf{O})}{n}$

Method of Lagrange multipliers

- Optimisation task: $\underset{\substack{x_1,\ldots,x_4\geq 0\\ \sum_i x_i=1}}{\arg\max} \sum_{i=1}^4 a_i \ln x_i$
- Define new function $T(x_1,\ldots,x_4,\lambda) := \sum_{i=1}^4 a_i \ln x_i + \lambda \cdot \left(-1 + \sum_{i=1}^4 x_i\right)$
- Solve (unconstrained) optimisation task $\underset{x_1,\ldots,x_4,\lambda\in\mathbf{R}}{\arg\max} T$ e.g. by setting the gradient to zero:

$$\nabla T = \left(\frac{a_1}{x_1} + \lambda, \ldots, \frac{a_4}{x_4} + \lambda, -1 + \sum_{i=1}^4 x_i\right)$$

$$\nabla T = 0 \Longrightarrow x_i = \frac{-a_i}{\lambda}, \sum_{i=1}^4 x_i = 1$$

$$\Longrightarrow \sum_{i=1}^4 \frac{-a_i}{\lambda} = 1 \Longrightarrow \lambda = -\sum_{i=1}^4 a_i$$

$$\Longrightarrow x_i = \frac{a_i}{\sum_{i=1}^4 a_i}$$

Quick summary so far

- Previous:
  - $\Pr[S \mid O, W] = \prod_{j=1}^{O-1} q(S[j]) \cdot \prod_{j=1}^{L} W[S[O+j-1], j] \cdot \prod_{j=O+L}^{m} q(S[j])$
  - $\Pr[O \mid S, W] \sim \Pr[S \mid O, W]$
- (new!) Given sequences $\mathbf{S}$ and motif starts $\mathbf{O}$, we can find the most likely motif W of length L using letter frequency counting
  - $W^* = \arg\max_{W \in \mathcal{W}} \ln \Pr[\mathbf{S} \mid \mathbf{O}, W] = \left(\frac{\#_{a,j}(\mathbf{O})}{n} \right)_{a \in \{A,C,G,T\}, 1 \le j \le L}$

Motif finding with distribution of O (One hand)

- Task: Given a vector of sequences $\mathbf{S} = (S_1, \ldots, S_n)$ of length m each, a distribution of site starts $g(\mathbf{O}) = \prod_{i=1}^{n} g_i(O_i)$ and b.f. q, find the best motif W of length L!
  - Assuming that the motif **occurs exactly once** in each sequence
  - In human words: we don't know exactly where the motif starts occur, but we have a guess g_i for each sequence.
  - The values $g_i(1), \ldots, g_i(m - L + 1)$ are "weights" for each position in sequence S_i
- Example of input data:

```
CGACTAAACCACGGA
AGATATAACAAAAAG
AAGTCACCATAAACT
AGTATTCCTATAGCA
TGACACATACCATGG
TAATATACCGCTTAC
TGCTAATAGTCCATA
TAATATACCGTATCT
```

$$(g_i(o))_{i,o} = \begin{pmatrix} 0.07 & 0.07 & 0.51 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.51 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\ 0.12 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.40 \\ 0.46 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.09 & 0.08 \\ 0.07 & 0.07 & 0.07 & 0.51 & 0.07 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.51 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.07 & 0.07 & 0.07 & 0.51 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.51 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 & 0.07 \end{pmatrix}$$

Motif finding with distribution of O (One hand)

- *best* = with the maximum log-likelihood across all possible $\mathbf{O}$ with prob. g
  - $W^* = \arg\max_{W \in \mathcal{W}} E_{\mathbf{O}\sim g}\left[\ln \mathcal{L}(W; \mathbf{O}, \mathbf{S})\right] \stackrel{\text{def.}}{=} \sum_{\mathbf{O}\in\mathcal{O}} \ln \mathcal{L}(W; \mathbf{O}, \mathbf{S}) \cdot g(\mathbf{O})$
- Intuition: best W is obtained by counting letter frequencies at all possible sites weighted by g
  - $W^*[a, j] \stackrel{??}{=} \frac{1}{n} \sum_{i=1}^{n} \sum_{O_i=1}^{m-L+1} g_i(O_i) \cdot \mathbf{1}_{S_i[O_i+j-1]=a} =: \frac{\#_{a,j}(g)}{n}$

Example of weighted frequency counting

- Looking for a motif of length $L = 3$
- Input: $\mathbf{S} = \begin{matrix}\texttt{AAACCT}\\ \texttt{ACGACA}\\ \texttt{TTACCG}\end{matrix}$, distrubution of starts $g = \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.1 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.1 & 0.1 & 0.6 & 0.2 \end{pmatrix}$

$$\#: \begin{pmatrix} A: & 0.2+0.3+0.4+0.4+0.3+0.6 & 0.2+0.3+0.2+0.1 & 0.2+0.1+0.3+0.1 \\ C: & 0.1+0.1+0.2 & 0.4+0.1+0.4+0.3+0.6+0.2 & 0.3+0.4+0.2+0.1+0.6 \\ G: & 0.2 & 0.1 & 0.4+0.2 \\ T: & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$W_{counting} = \begin{pmatrix} A: & 2.2/3 & 0.8/3 & 0.7/3 \\ C: & 0.5/3 & 2.0/3 & 1.6/3 \\ G: & 0.2/3 & 0.1/3 & 0.6/3 \\ T: & 0.1/3 & 0.1/3 & 0.2/3 \end{pmatrix}$$

Let's check the intuition

$$W^* := \arg\max_{W \in \mathcal{W}} E_{\mathbf{O} \sim g}[\ln \mathcal{L}(W; \mathbf{O}, \mathbf{S})] = \arg\max_{W \in \mathcal{W}} E_{\mathbf{O} \sim g}[\ln \prod_{i=1}^{n} \mathcal{L}(W; O_i, S_i)] =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{i=1}^{n} E_{\mathbf{O}_i \sim g_i}[\ln \mathcal{L}(W; O_i, S_i)] =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{i=1}^{n} \sum_{O_i=1}^{m-L+1} (\ln \Pr[S_i \mid O_i, W] + \ln \Pr[O_i \mid W]) g_i(O_i) =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{i=1}^{n} \sum_{O_i=1}^{m-L+1} \left(\sum_{j=1}^{O_i - 1} \ln q(S_i[j]) + \sum_{j=1}^{L} \ln W[S_i[O_i + j - 1], j] + \sum_{j=O_i+L}^{m} \ln q(S_i[j]) \right) g_i(O_i) =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{j=1}^{L} \sum_{a \in \{A,C,G,T\}} \ln W[a, j] \left(\sum_{i=1}^{n} \sum_{O_i=1}^{m-L+1} \mathbf{1}_{S_i[O_i+j-1]=a} \cdot g_i(O_i) \right) =$$

$$= \arg\max_{W \in \mathcal{W}} \sum_{j=1}^{L} \sum_{a \in \{A,C,G,T\}} \ln W[a, j] \cdot \#_{a,j}(g).$$

Quick summary so far

- Previous:
  - $\Pr[S \mid O, W] = \prod_{j=1}^{O-1} q(S[j]) \cdot \prod_{j=1}^{L} W[S[O+j-1], j] \cdot \prod_{j=O+L}^{m} q(S[j])$
  - $\Pr[O \mid S, W] \sim \Pr[S \mid O, W]$
  - For known $\mathbf{S}$ and starts $\mathbf{O}$:
    $W^* = \arg\max_{W \in \mathcal{W}} \ln \mathcal{L}(W; \mathbf{S}, \mathbf{O}) = \left(n^{-1} \cdot \#_{a,j}(\mathbf{O})\right)_{a \in \{A,C,G,T\}, 1 \le j \le L}$
    - ⋆ $\#_{a,j}(\mathbf{O}) := \sum_{i=1}^{n} \mathbf{1}_{S_i[O_i+j-1]=a}$
- (new!) Given sequences $\mathbf{S}$ and motif starts distribution $g(\mathbf{O})$, we can find the most likely motif W of length L using weighted letter frequency counting:
  - $W^* = \arg\max_{W \in \mathcal{W}} E_{\mathbf{O} \sim g}[\ln \mathcal{L}(W; \mathbf{S}, \mathbf{O})] = \left(n^{-1} \cdot \#_{a,j}(g)\right)_{a \in \{A,C,G,T\}, 1 \le j \le L}$
  - $\#_{a,j}(g) := \sum_{i=1}^{n} \sum_{O_i=1}^{m-L+1} g_i(O_i) \cdot \mathbf{1}_{S_i[O_i+j-1]=a}$

Motif finding without O (No hands)

- Task: Given a vector of sequences $\mathbf{S} = (S_1, \ldots, S_n)$ of length m each and b.f. q, find the *best* motif W of length L!
  - Assuming that the motif **occurs exactly once** in each sequence
  - No information about the motif starts nor motif itself...
- Example of input data:

```
CGACTAAACCACGGA
AGATATAACAAAAAG
AAGTCACCATAAACT
AGTATTCCTATAGCA
TGACACATACCATGG
TAATATACCGCTTAC
TGCTAATAGTCCATA
TAATATACCGTATCT
```

Expectation-Maximisation algorithm

- The algorithm:
  - Start with random motif $W^{(0)}$
  - Repeat:
    - ★ (E-step) infer $g^{(t+1)}(\cdot)$ from $W^{(t)}$:

$$g^{(t+1)}(\mathbf{O}) := \Pr[\mathbf{O} \mid \mathbf{S}, W^{(t)}] \sim \prod_{i=1}^{n} \Pr[S_i \mid O_i, W^{(t)}]$$

    - ★ (M-step) infer $W^{(t+1)}$ from $g^{(t+1)}(\cdot)$:

$$W^{(t+1)} := \arg\max_{W \in \mathcal{W}} E_{\mathbf{O} \sim g^{(t+1)}}[\ln \mathcal{L}(W; \mathbf{S}, \mathbf{O})] = \left(n^{-1} \cdot \#_{a,j}\left(g^{(t+1)}\right)\right)_{a \in \{A,C,G,T\}, 1 \le j \le L}$$

- Random $\Rightarrow W^{(0)} \overset{\text{E}}{\Rightarrow} g^{(1)} \overset{\text{M}}{\Rightarrow} W^{(1)} \overset{\text{E}}{\Rightarrow} g^{(2)} \overset{\text{M}}{\Rightarrow} W^{(2)} \overset{\text{E}}{\Rightarrow} g^{(3)} \overset{\text{M}}{\Rightarrow} W^{(3)} \overset{\text{E}}{\Rightarrow} \dots$
- Each next $W^{(t)}$ is better than the previous one:

$$\Pr[S \mid W^{(t+1)}] \ge \Pr[S \mid W^{(t)}]$$

Time for a demo

Reconstruction of a missing motif start

- Task: Given sequences $\mathbf{S}$, motif length L and motif starts with missing i-th coordinate $\mathbf{O}_{-i} = (O_1, \ldots, O_{i-1}, ?, O_{i+1}, \ldots, O_n)$, reconstruct the missing coordinate O_i.
  - We don't know the motif W
- Let's compute prob. for each possible value of O_i:

$$\Pr[O_i = k \mid \mathbf{O}_{-i}, \mathbf{S}] \overset{\text{cond.}}{=} \frac{\Pr[O_i = k, \mathbf{O}_{-i} \mid \mathbf{S}]}{\sum_\ell \Pr[O_i = \ell, \mathbf{O}_{-i} \mid \mathbf{S}]} \sim$$

$$\sim \Pr[\mathbf{O} = (O_1, \ldots, O_{i-1}, k, O_{i+1}, \ldots, O_n) \mid \mathbf{S}] =$$

$$= \frac{\Pr[\mathbf{S} \mid \mathbf{O}] \cdot \Pr[\mathbf{O}]}{\sum_{\mathbf{O}'} \Pr[\mathbf{S} \mid \mathbf{O}] \cdot \Pr[\mathbf{O}]} \sim$$

$$\sim \Pr[\mathbf{S} \mid \mathbf{O}]$$

- So, we only need to be able to compute $\Pr[\mathbf{S} \mid \mathbf{O}]$ for arbitrary $\mathbf{O}$...

How to compute Pr[**S** | **O**]

- Since we don't know the true motif, we have to average over all possible motifs. It's called marginalization:

$$\Pr[\mathbf{S} \mid \mathbf{O}] = E_W[\Pr[\mathbf{S}, W \mid \mathbf{O}]] = \int_{\mathcal{W}} \Pr[\mathbf{S} \mid W, \mathbf{O}] \cdot p(W)\ dW$$

- We need to define the "probability" $p(W)$ of motif W.
- We want all combinations of frequencies to be "equally likely":

$$W_j \sim \text{Dirichlet}(1, 1, 1, 1)$$
$$p(W_j) = \frac{\Gamma(4)}{\Gamma(1)^4} = \frac{3!}{0!^4} = 6$$
$$p(W) = 6^n$$

Let's compute some integrals

$$\Pr[\mathbf{S} \mid \mathbf{O}] = \int_{\mathcal{W}} \Pr[\mathbf{S} \mid W, \mathbf{O}] \cdot 6^n \, dW \sim$$

$$\sim \int_{\mathcal{W}} \Pr[\mathbf{S} \mid W, \mathbf{O}] \, dW = \int_{\mathcal{W}} \prod_{i=1}^{n} \Pr[S_i \mid W, O_i] \, dW =$$

$$= \int_{\mathcal{W}} \prod_{i=1}^{n} \prod_{j=1}^{O_i - 1} q(S[j]) \cdot \prod_{j=1}^{L} W[S[O_i + j - 1], j] \cdot \prod_{j=O_i+L}^{m} q(S[j]) \, dW =$$

$$= \int_{\mathcal{W}} \prod_{i=1}^{n} \prod_{j=1}^{L} W[S[O_i + j - 1], j] \, dW =$$

$$= \int_{\mathcal{W}} \prod_{j=1}^{L} \prod_{a \in \{A,C,G,T\}} W[a, j]^{\#_{a,j}(\mathbf{O})} \, dW =$$

$$= \prod_{j=1}^{L} \int_{\mathcal{S}^4} \prod_{a \in \{A,C,G,T\}} W[a, j]^{\#_{a,j}(\mathbf{O})} \, dW_j \ldots$$

Computation of integral over unit 4-simplex

- We need to compute a definite integral of form $\int_{\mathcal{S}^4} x_1^{a_1} \cdot x_2^{a_2} \cdot x_3^{a_3} \cdot x_4^{a_4} \; d\mathbb{X}$, where $\mathcal{S}^4 = \left\{(x_1, \ldots, x_4) \in [0,1]^4 : \sum_i x_i = 1\right\}$.
  - A magic formula: $\mathcal{B}(z_1, z_2) \overset{\text{def.}}{=} \int_0^1 t^{z_1-1} \cdot (1-t)^{z_2-1} \; dt = \dfrac{\Gamma(z_1) \cdot \Gamma(z_2)}{\Gamma(z_1+z_2)}$

$$\int_{\mathcal{S}^4} x_1^{a_1} \cdot x_2^{a_2} \cdot x_3^{a_3} \cdot x_4^{a_4} \; d\mathbb{X} =$$

$$= \int_0^1 x_1^{a_1} \int_0^{1-x_1} x_2^{a_2} \int_0^{1-x_1-x_2} x_3^{a_3} \cdot (1-x_1-x_2-x_3)^{a_4} \; dx_3 \; dx_2 \; dx_1 =$$

$$= \int_0^1 x_1^{a_1} \int_0^{1-x_1} x_2^{a_2} \int_0^{\xi} x_3^{a_3} \cdot (\xi - x_3)^{a_4} \; dx_3 \; dx_2 \; dx_1 =$$

$$= \int_0^1 x_1^{a_1} \int_0^{1-x_1} x_2^{a_2} \xi^{a_3+a_4} \int_0^{\xi} \left(\frac{x_3}{\xi}\right)^{a_3} \cdot \left(1 - \frac{x_3}{\xi}\right)^{a_4} \; dx_3 \; dx_2 \; dx_1 =$$

$$= \frac{\Gamma(a_3+1)\Gamma(a_4+1)}{\Gamma(a_3+a_4+2)} \int_0^1 x_1^{a_1} \int_0^{1-x_1} x_2^{a_2} (1-x_1-x_2)^{a_3+a_4+1} \; dx_2 \; dx_1 =$$

$$= \frac{\Gamma(a_3+1)\Gamma(a_4+1)}{\Gamma(a_3+a_4+2)} \frac{\Gamma(a_2+1)\Gamma(a_3+a_4+2)}{\Gamma(a_2+a_3+a_4+3)} \int_0^1 x_1^{a_1} \cdot (1-x_1)^{a_2+a_3+a_4+2} \; dx_1 =$$

$$= \frac{\Gamma(a_3+1)\Gamma(a_4+1)}{\Gamma(a_3+a_4+2)} \frac{\Gamma(a_2+1)\Gamma(a_3+a_4+2)}{\Gamma(a_2+a_3+a_4+3)} \frac{\Gamma(a_1+1)\Gamma(a_2+a_3+a_4+3)}{\Gamma(a_1+a_2+a_3+a_4+4)} =$$

$$= \frac{a_1! a_2! a_3! a_4!}{(3+a_1+a_2+a_3+a_4)!}.$$

Back to the main integral

$$\Pr[\mathbf{S} \mid \mathbf{O}] \sim \ldots \sim \prod_{j=1}^{L} \int_{\mathcal{S}^4} \prod_{a \in \{A,C,G,T\}} W[a,j]^{\#_{a,j}(\mathbf{O})}\, dW_j =$$

$$= \prod_{j=1}^{L} \frac{\prod_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O})!}{(3+n)!} \sim$$

$$\sim \prod_{j=1}^{L} \prod_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O})!$$

Finally, we can compute the probabilities for the missing motif start:

$$\Pr[O_i = k \mid \mathbf{O}_{-i}, \mathbf{S}] \sim \prod_{j=1}^{L} \prod_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O})!$$

Hooray! Time to take a breath and regain posture.

Example of computing the prob. of missing start

- Formula: $\Pr[O_i = k \mid \mathbf{O}_{-i}, \mathbf{S}] \sim \prod_{j=1}^{L} \prod_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O})!$

  AAACCT
- Input: $\mathbf{S} = \texttt{ACGTCA}$, $\mathbf{O}_{-2} = (2, ?, 3)$

  TTACCG

$$\Pr[O_2 = 1 \mid \mathbf{O}_{-2}, \mathbf{S}] \sim 3!0!0!0! \cdot 1!2!0!0! \cdot 0!2!1!0! = 24 \quad : \text{sites AAC, ACG, ACC}$$
$$\Pr[O_2 = 2 \mid \mathbf{O}_{-2}, \mathbf{S}] \sim 2!1!0!0! \cdot 1!1!1!0! \cdot 0!2!0!1! = 4 \quad : \text{sites AAC, CGT, ACC}$$
$$\Pr[O_2 = 3 \mid \mathbf{O}_{-2}, \mathbf{S}] \sim 2!0!1!0! \cdot 1!1!0!1! \cdot 0!3!0!0! = 12 \quad : \text{sites AAC, GTC, ACC}$$
$$\Pr[O_2 = 4 \mid \mathbf{O}_{-2}, \mathbf{S}] \sim 2!0!0!1! \cdot 1!2!0!0! \cdot 1!2!0!0! = 8 \quad : \text{sites AAC, TCA, ACC}$$

$$\Sigma = 24 + 4 + 12 + 8 = 48$$

$$\Pr[O_2 = 1 \mid \mathbf{O}_{-2}, \mathbf{S}] = 24/48 = 0.50$$
$$\Pr[O_2 = 2 \mid \mathbf{O}_{-2}, \mathbf{S}] = 4/48 \doteq 0.08$$
$$\Pr[O_2 = 3 \mid \mathbf{O}_{-2}, \mathbf{S}] = 12/48 = 0.25$$
$$\Pr[O_2 = 4 \mid \mathbf{O}_{-2}, \mathbf{S}] = 8/48 \doteq 0.17$$

Sampling from Pr[**O** | **S**] via Gibbs sampling algorithm

- A bigger goal: To sample motif starts **O** from Pr[**O** | **S**]
- Gibbs sampling algorithm:
  - Start with a random $\mathbf{O}^{(0)}$
  - Repeat:
    - Select a random coordinate $i \in_R \{1, \ldots, n\}$
    - Erase i-th coordinate from $\mathbf{O}^{(\mathbf{t})}$
    - Sample a replacement O' for it from $\Pr[O_i = k \mid \mathbf{O}_{-i}, \mathbf{S}] \sim \prod_{j=1}^{L} \prod_{a \in \{A,C,G,T\}} \#_{a,j}(\mathbf{O})!$
    - new sample $\mathbf{O}^{(t+1)}$ is the same as $\mathbf{O}^{(t)}$, but with i-th coordinate replaced by O'
- This algorithm produces samples $O^{(0)}, O^{(1)}, O^{(2)}, O^{(3)}, \ldots$ from Pr[**O** | **S**]
- Example:

$$
\begin{array}{lcccccccc}
O^{(0)} = (& 1 & 3 & 2 & 1 & 10 & 6 &) \\
O^{(1)} = (& 1 & 3 & 2 & 3 & 10 & 6 &) \\
O^{(2)} = (& 1 & 3 & 2 & 3 & 7 & 6 &) \\
O^{(3)} = (& 1 & 3 & 2 & 3 & 16 & 6 &) \\
O^{(4)} = (& 1 & 5 & 2 & 3 & 16 & 6 &) \\
O^{(5)} = (& 1 & 5 & 2 & 3 & 16 & 9 &) \\
\ldots
\end{array}
$$

Back to Motif finding without O (No hands)

- Task: Given a vector of sequences $\mathbf{S} = (S_1, \ldots, S_n)$ of length m each and b.f. q, find the *best* motif W of length L!
- Algorithm using Gibbs sampling:
  - Sample many motif starts vectors $\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(B)}$ from $\Pr[\mathbf{O} \mid \mathbf{S}]$
  - For each sampled motif starts vector $\mathbf{O}^{(t)}$, compute the optimal motif $W^{(t)} := \underset{W \in \mathcal{W}}{\arg\max} \ln \mathcal{L}(W; \mathbf{S}, \mathbf{O}^{(t)}) = \left(n^{-1} \cdot \#_{a,j}\left(\mathbf{O}^{(\mathbf{t})}\right)\right)_{a \in \{A,C,G,T\}, 1 \le j \le L}$
  - Return the pair $(W^{(t)}, \mathbf{O}^{(t)})$ with the highest log-likelihood $\ln \mathcal{L}(W^{(t)}; \mathbf{S}, \mathbf{O}^{(t)})$

Time for a demo

Summary

- Motifs are used to represent e.g. DNA bind sites of proteins
  - Motif W = a table of letter probabilities at each position of the site
    - $W[a, j] :=$ Pr[j-th letter of a site is letter a]
- Motif finding = given sequences, where motif occurs, find the best motif W
- Motif finding can be solved by Expectation-Maximisation algorithm
  - Alternating improvement of $g(\mathbf{O})$ and W:
    - Random $\Rightarrow W^{(0)} \overset{\text{E}}{\Rightarrow} g^{(1)} \overset{\text{M}}{\Rightarrow} W^{(1)} \overset{\text{E}}{\Rightarrow} g^{(2)} \overset{\text{M}}{\Rightarrow} W^{(2)} \overset{\text{E}}{\Rightarrow} g^{(3)} \overset{\text{M}}{\Rightarrow} W^{(3)} \overset{\text{E}}{\Rightarrow} \ldots$
- Motif finding can be solved by Gibbs sampling
  - Random sampling of $\mathbf{O}$ from Pr[$\mathbf{O} \mid \mathbf{S}$], selecting the best one
  - Gibbs sampling works by altering one coordinate of a previous sample, sampling its value from a conditional distribution