The “Gibbs Sampler”

- Lawrence, et al.
  “Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Sequence Alignment,”
  Science 1993

Some DNA Binding Domains
Some History

- Geman & Geman, IEEE PAMI 1984
- Hastings, Biometrika, 1970
- Josiah Willard Gibbs, 1839-1903, American physicist, a pioneer of thermodynamics

How to Average

An old problem:
- n random variables:
- Joint distribution (p.d.f.): $P(x_1, x_2, \ldots, x_k)$
- Some function:
- Want Expected Value: $E(f(x_1, x_2, \ldots, x_k))$
**How to Average**

\[
E(f(x_1, x_2, \ldots, x_k)) = \int_{x_1} \int_{x_2} \cdots \int_{x_k} f(x_1, x_2, \ldots, x_k) \cdot P(x_1, x_2, \ldots, x_k) \, dx_1 \, dx_2 \ldots \, dx_k
\]

- Approach 1: direct integration
  (rarely solvable analytically, esp. in high dim)
- Approach 2: numerical integration
  (often difficult, e.g., unstable, esp. in high dim)
- Approach 3: Monte Carlo integration

\[
E(f(\vec{x})) \approx \frac{1}{n} \sum_{i=1}^{n} f(\vec{x}^{(i)})
\]

**Markov Chain Monte Carlo (MCMC)**

- Independent sampling also often hard, but not required for expectation
- MCMC \( \vec{X}_{t+1} \sim P(\vec{X}_{t+1} | \vec{X}_t) \) w/ stationary dist \( \equiv \)
- Simplest & most common: Gibbs Sampling

\[
P(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)
\]

- Algorithm
  for \( t = 1 \) to ∞
  for \( i = 1 \) to \( k \) do :
  \( x_{t+1,i} \sim P(x_{t+1,i} | x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_{t+1,i+1}, \ldots, x_{t+1,k}) \)

**Input:** again assume sequences \( s_1, s_2, \ldots, s_k \) with one length \( w \) motif per sequence
**Motif model:** WMM
**Parameters:** Where are the motifs?
  for \( 1 \leq i \leq k \), have \( 1 \leq x_i \leq |s_i| - w + 1 \)
**“Full conditional”: to calc**

\[
P(x_i = j | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)
\]

build WMM from motifs in all sequences except \( i \), then calc prob that motif in \( i^{th} \) seq occurs at \( j \) by usual “scanning” alg.
Randomly initialize $x_i$'s
for $t = 1$ to $\infty$
for $i = 1$ to $k$
   discard motif instance from $s_i$;
   recal WMM from rest
for $j = 1$ ... $|s_i| - w + 1$
   calculate prob that $i^{th}$ motif is at $j$:
   \[ P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k) \]
   pick new $x_i$ according to that distribution

Similar to MEME, but it would average over, rather than sample from $P(x_i = j \mid x_1,x_2,\ldots,x_{i-1},x_{i+1},\ldots,x_k)$

Issues

• Burnin - how long must we run the chain to reach stationarity?
• Mixing - how long a post-burnin sample must we take to get a good sample of the stationary distribution? (Recall that individual samples are not independent, and may not "move" freely through the sample space. Also, many isolated modes.)

Variants & Extensions

• “Phase Shift” - may settle on suboptimal solution that overlaps part of motif. Periodically try moving all motif instances a few spaces left or right.
• Algorithmic adjustment of pattern width: Periodically add/remove flanking positions to maximize (roughly) average relative entropy per position
• Multiple patterns per string

![Graph showing information per parameter over number of iterations]