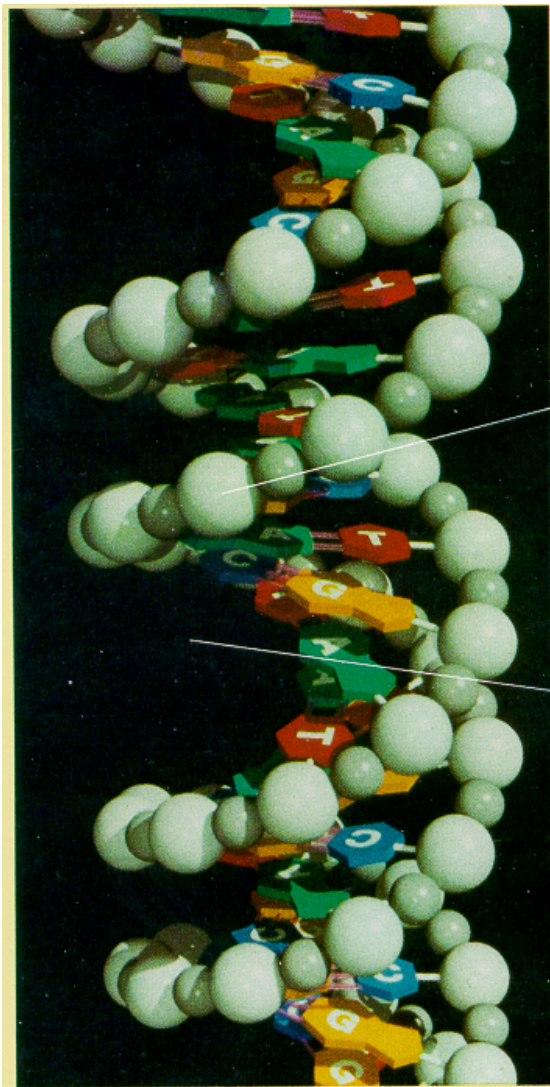# CSE 527
Lecture 9

The Gibbs Sampler

Talk Today

- Zasha Weinberg
  Combi
  HSB K-069, 1:30
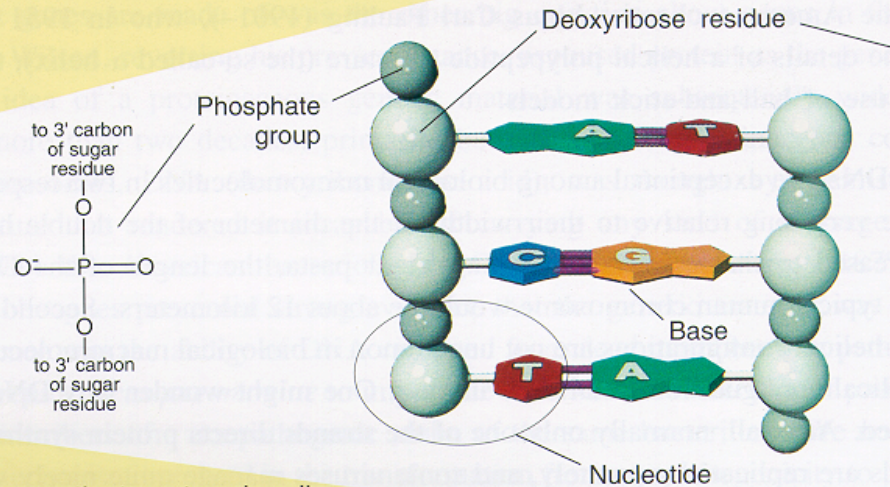 "Fast, accurate annotation of non-coding RNAs"

The "Gibbs Sampler"

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

The Double Helix

(a) Computer-generated Image of DNA (by Mel Prueitt)

(b) Uncoiled DNA Fragment

Deoxyribose residue

Phosphate group

to 3' carbon of sugar residue

O^-—P—O

to 3' carbon of sugar residue

Base

Nucleotide

As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-
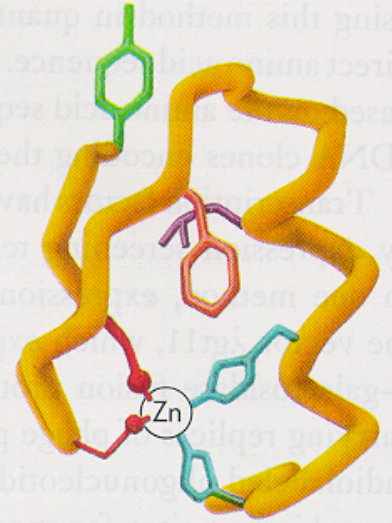
Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

Los Alamos Science
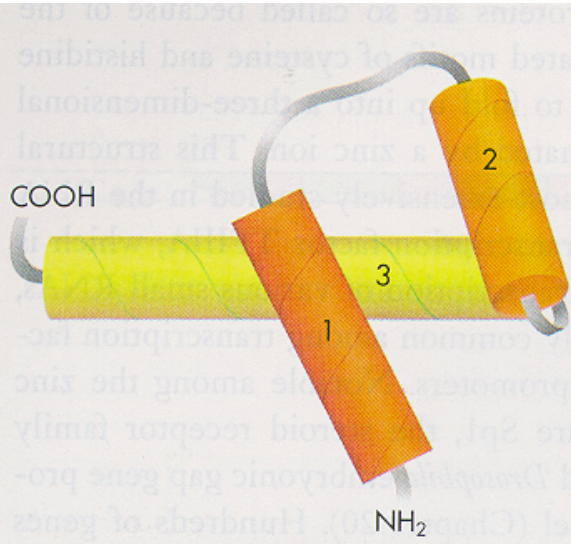
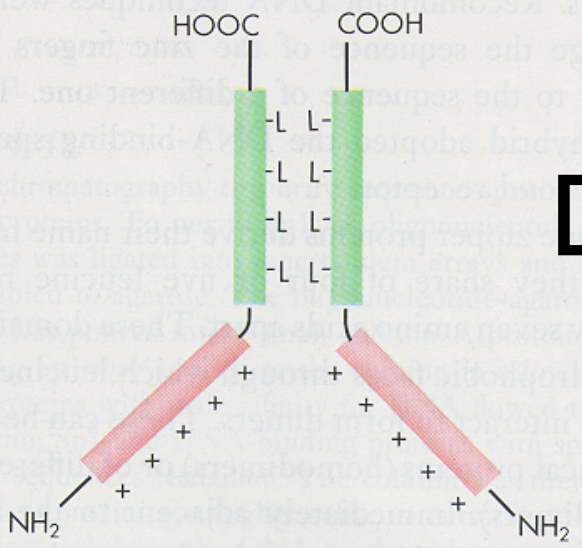(a) Helix-turn-helix Homeodomain

(b) C₂H₂ zinc finger
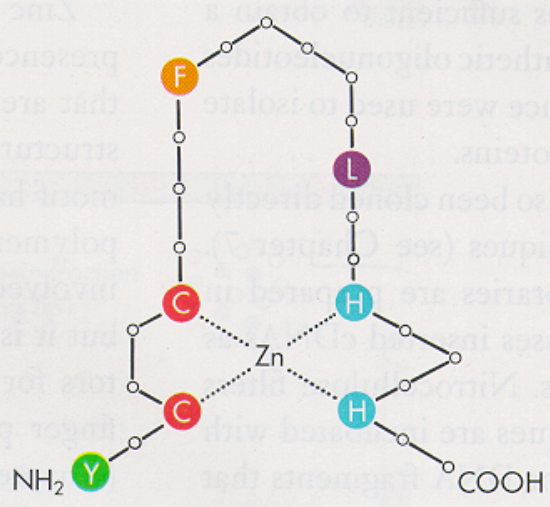
(c) Leucine zipper

(d) Helix-loop-helix

Some DNA Binding Domains
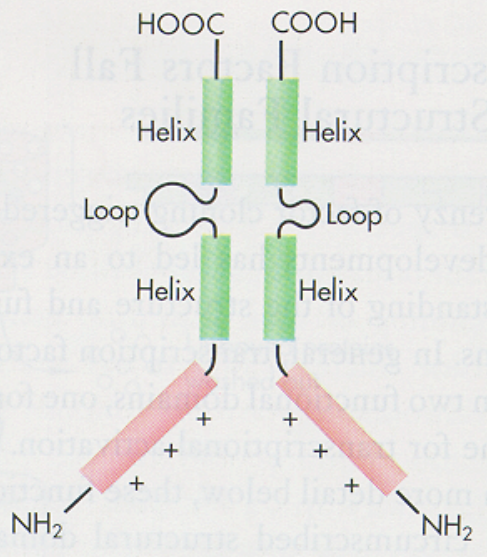
```
Sigma-37        223 IIDLTYIQNK SQKETGDILGISQMHVSR LQRKAVKKLR 240   A25944
SpoIIIC          94 RFGLDLKKEK TQREIAKELGISRSYVSR IEKRALMKMF 111   A28627
NahR             22 VVFNQLLVDR RVSITAENLGLTQPAVSN ALKRLRTSLQ  39   A32837
Antennapedia    326 FHFNRYLTRR RRIEIAHALCLTERQIKI WFQNRRMKWK 343   A23450
NtrC (Brady.)   449 LTAALAATRG NQIRAADLLGLNRNTLRK KIRDLDIQVY 466   B26499
DicA             22 IRYRRKNLKH TQRSLAKALKISHVSVSQ WERGDSEPTG  39   B24328 (BVECDA)
MerD              5       MNAY TVSRLALDAGVSVHIVRD YLLRGLLRPV  22   C29010
Fis              73 LDMVMQYTRG NQTRAALMMGINRGTLRK KLKKYGMN   90   A32142 (DNECFS)
MAT a1           99 FRRKQSLNSK EKEEVAKKCGITPLQVRV WFINKRMRSK 116   A90983 (JEBY1)
Lambda cII       25 SALLNKIAML GTEKTAEAVGVDKSQISR WKRDWIPKFS  42   A03579 (QCBP2L)
Crp (CAP)       169 THPDGMQIKI TRQEIGQIVGCSRETVGR ILKMLEDQNL 186   A03553 (QRECC)
Lambda Cro       15 ITLKDYAMRF GQTKTAKDLGVYQSAINK AIHAGRKIFL  32   A03577 (RCBPL)
P22 Cro          12 YKKDVIDHFG TQRAVAKALGISDAAVSQ WKÉVIPEKDA  29   A25867 (RGBP22)
AraC            196 ISDHLADSNF DIASVAQHVCLSPSRLSH LFRQQLGISV 213   A03554 (RGECA)
Fnr             196 FSPREFRLTM TRGDIGNYLGLTVETISR LLGRFQKSGM 213   A03552 (RGECF)
HtpR            252 ARWLDEDNKS TLQELADRYGVSAERVRQ LEKNAMKKLR 269   A00700 (RGECH)
NtrC (K.a.)     444 LTTALRHTQG HKQEAARLLGWGRNTLTR KLKELGME   461   A03564 (RGKBCP)
CytR             11 MKAKKQETAA TMKDVALKAKVSTATVSR ALMNPDKVSQ  28   A24963 (RPECCT)
DeoR             23 LQELKRSDKL HLKDAAALLGVSEMTIRR DLNNHSAPVV  40   A24076 (RPECDO)
GalR              3         MA TIKDVARLAGVSVATVSR VINNSPKASE  20   A03559 (RPECG)
LacI              5       MKPV TLYDVAEYAGVSYQTVSR VVNQASHVSA  22   A03558 (RPECL)
TetR             26 LLNEVGIEGL TTRKLAQKLGVEQPTLYW HVKNKRALLD  43   A03576 (RPECTN)
TrpR             67 IVEELLRGEM SQRELKNELGAGIATITR GSNSLKAAPV  84   A03568 (RPECW)
NifA            495 LIAALEKAGW VQAKAARLLGMTPRQVAY RIQIMDITMP 512   S02513
SpoIIG          205 RFGLVGEEEK TQKDVADMMGISQSYISR LEKRIIKRLR 222   S07337
Pin             160 QAGRLIAAGT PRQKVAIIYDVGVSTLYK TFPAGDK    177   S07958
PurR              3         MA TIKDVAKRANVSTTTVSH VINKTRFVAE  20   S08477
EbgR              3         MA TLKDIAIEAGVSLATVSR VLNDDPTLNV  20   S09205
LexA             27 DHISQTGMPP TRAEIAQRLGFRSPNAAE EHLKALARKG  44   S11945
P22 cI           25 SSILNRIAIR GQRKVADALGINESQISR WKGDFIPKMG  42   B25867 (Z1BPC2)
                          ***************** ***
```

B

Position in site

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Arg	94	222	265	137	9	9	137	137	9	9	9	52	222	94	94	9	265	606
Lys	9	133	442	380	9	71	380	194	9	133	9	9	71	9	9	9	71	256
Glu	53	9	96	401	9	9	140	140	9	9	9	53	140	140	9	9	9	53
Asp	67	9	9	473	9	9	299	125	9	67	9	67	67	9	9	9	9	67
Gln	9	600	224	9	9	9	224	9	9	9	9	9	278	63	278	9	9	170
His	240	9	9	9	9	9	125	125	9	9	9	9	125	125	125	9	9	240
Asn	168	9	9	9	9	9	168	89	9	89	9	248	9	168	89	9	89	89
Ser	117	9	117	117	9	9	9	9	9	9	9	819	63	387	63	9	819	9
Gly	151	9	56	9	9	151	9	9	9	1141	9	151	9	56	9	9	56	9
Ala	9	9	112	43	181	901	43	181	215	9	43	9	43	181	112	43	78	9
Thr	915	130	130	9	251	9	9	9	9	9	9	311	130	70	855	9	130	9
Pro	76	9	9	9	9	9	9	9	9	9	9	9	210	210	9	9	9	9
Cys	9	9	9	9	9	9	9	9	295	581	295	9	9	9	9	9	9	9
Val	58	107	9	9	500	9	9	9	156	9	598	9	205	58	9	746	9	58
Leu	9	121	9	9	149	9	93	149	458	9	149	9	37	37	9	177	9	9
Ile	9	166	114	61	323	9	114	166	9	9	427	9	61	9	61	427	9	61
Met	9	104	9	9	9	9	9	198	198	9	104	9	9	198	9	9	9	9
Tyr	9	9	136	9	9	9	9	262	262	9	9	136	136	9	262	9	262	136
Phe	9	9	9	9	9	9	9	9	9	9	108	9	9	9	9	9	9	9
Trp	9	9	9	9	9	9	9	9	9	9	366	9	9	9	9	9	9	366

Some History

- Geman & Geman, IEEE PAMI 1984

- Hastings, Biometrika, 1970

- Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, "Equations of State Calculations by Fast Computing Machines," J. Chem. Phys. 1953

- Josiah Williard Gibbs, 1839-1903, American physicist, a pioneer of thermodynamics

How to Average

- An old problem:
  - n random variables: $x_1, x_2, \ldots, x_k$
  - Joint distribution (p.d.f.): $P(x_1, x_2, \ldots, x_k)$
  - Some function: $f(x_1, x_2, \ldots, x_k)$
  - <u>Want</u> Expected Value: $E(f(x_1, x_2, \ldots, x_k))$

How to Average

$$E(f(x_1, x_2, \ldots, x_k)) \quad =$$

$$\int_{x_1} \int_{x_2} \cdots \int_{x_k} f(x_1, x_2, \ldots, x_k) \cdot P(x_1, x_2, \ldots, x_k) dx_1 dx_2 \ldots dx_k$$

- 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
  sample $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots \vec{x}^{(n)} \sim p(\vec{x})$ and average:
  $$E(f(\vec{x})) \approx \frac{1}{n} \sum_{i=1}^{n} f(\vec{x}^{(i)})$$

Markov Chain Monte Carlo (MCMC)

- Independent sampling also often hard, but not required for expectation
- MCMC $\quad \vec{X}_{t+1} \mid \vec{X}_t$
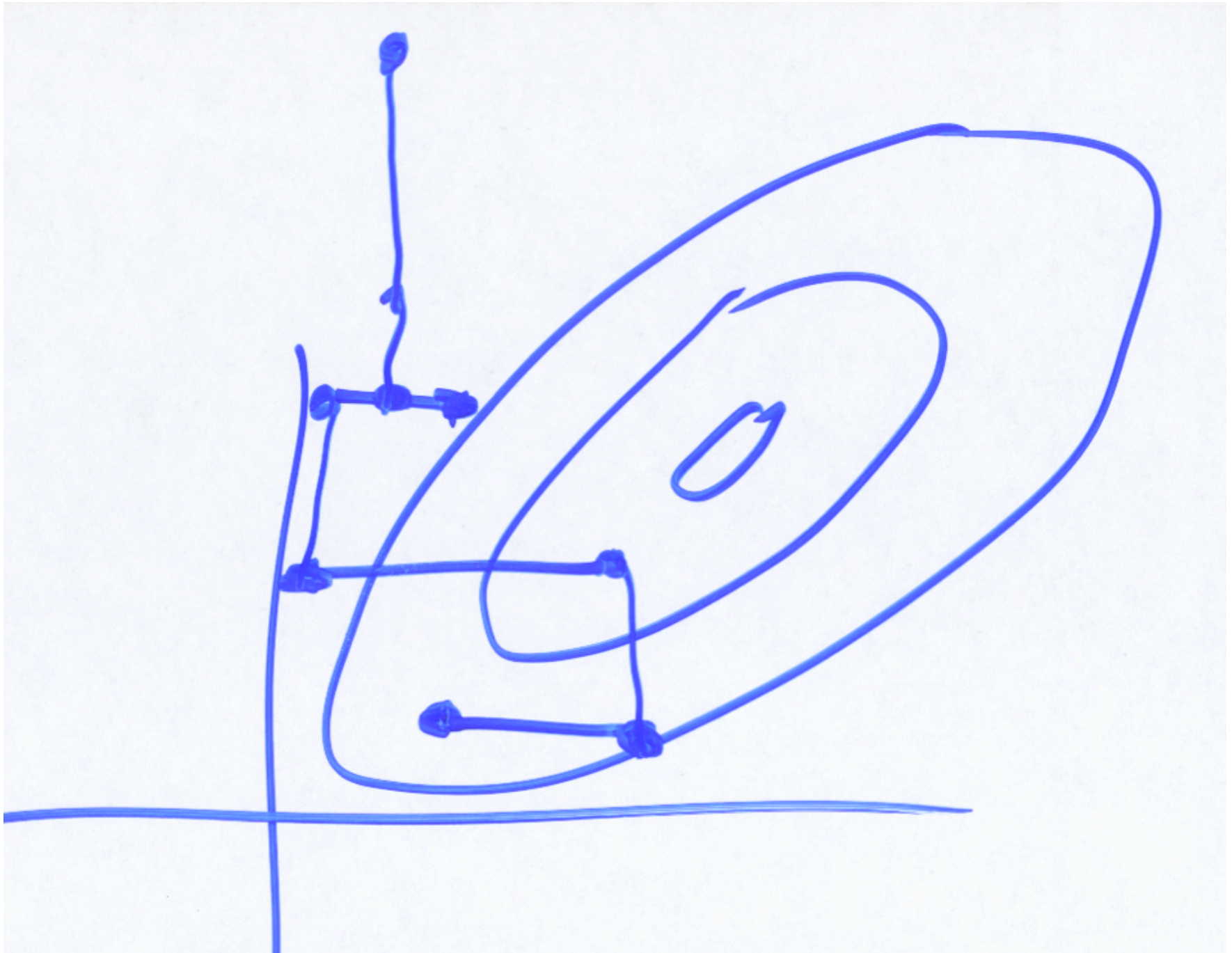- Simplest & most common: Gibbs Sampling

$$P(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$

- Algorithm

  for t = 1 to ∞

     for i = 1 to k do :

$$x_{t+1,i} \sim P(x_{t+1,i} \mid \underline{x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}}, \underline{x_{t,i+1}, \ldots, x_{t,k}})$$

$$t+1 \qquad t$$

- Input: again assume sequences s1, ..., sk with one length w motif per sequence
- Motif model: WMM
- Parameters: Where are the motifs?
  for 1 <= i <= k, have 1 <= xi <= |si|-w+1
- "Full conditional": to calc

$$P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$

  build WMM from motifs in all sequences except i, then calc prob that motif in ith seq occurs at j by usual "scanning" alg.

Randomly initialize xi's
for t = 1 to ∞
   for i = 1 to k
      discard motif instance from si;
      recalc WMM from rest
      for j = 1 ... |si|-w+1
         calculate prob that ith motif is at j:

$$P(x_i = j \mid x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_k)$$

      pick new xi according to that distribution

Similar to MEME, but it would average over, rather than sample from

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.)

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