CSE 527
Lecture 14

The Gibbs Sampler

Talk Tomorrow

- UW Biostatistics Autumn Seminar Series

  Martin Tompa, Ph.D.
  Department of Computer Science and Engineering and
  Department of Genome Sciences

  "Discovery of Regulatory Elements by a Phylogenetic Footprinting Algorithm"

  Thursday, November 13, 2003, 3:30 pm

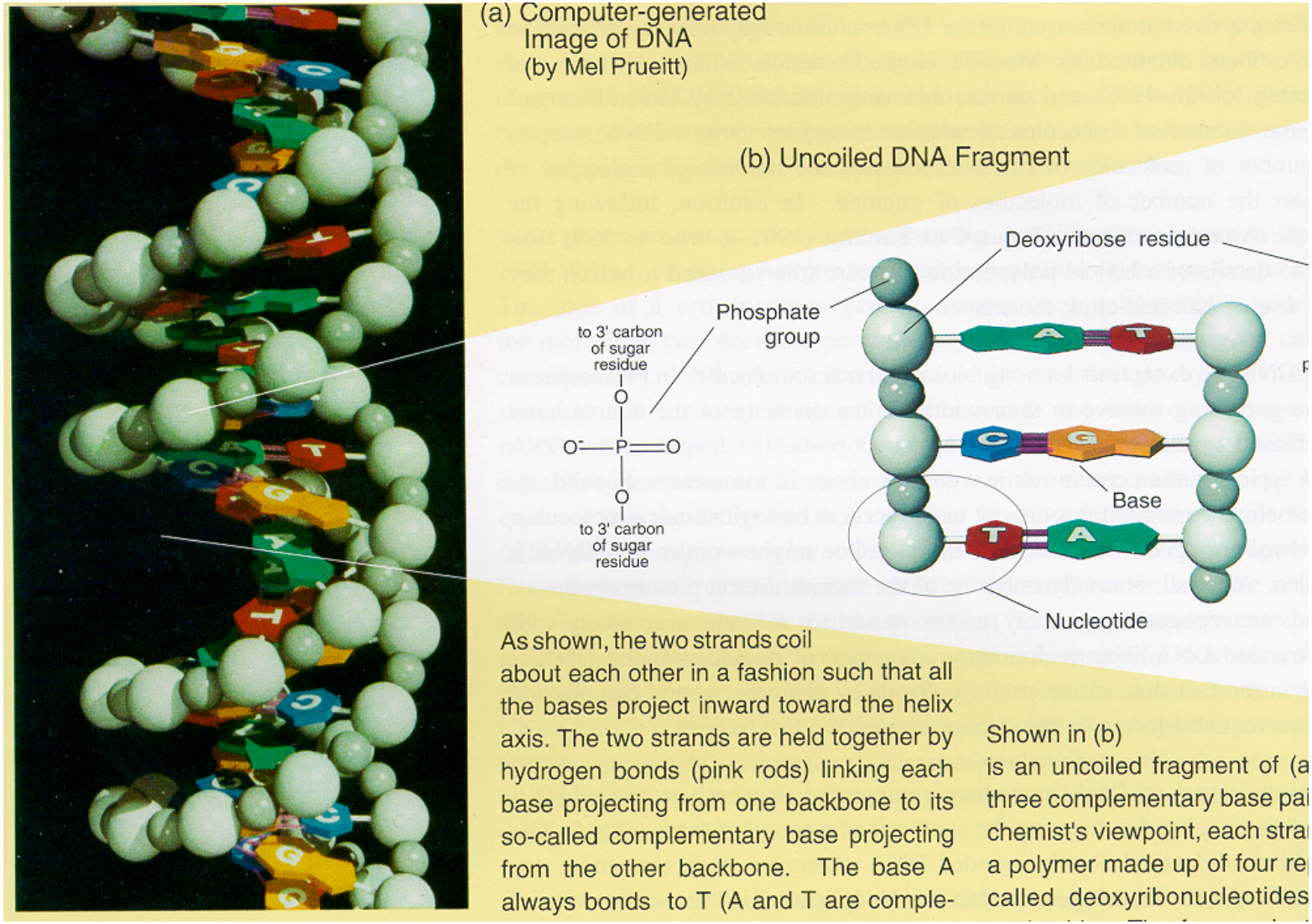
  T-639 Health Sciences Building

- Refreshments served outside the seminar room beginning at 3:15 pm.

The "Gibbs Sampler"

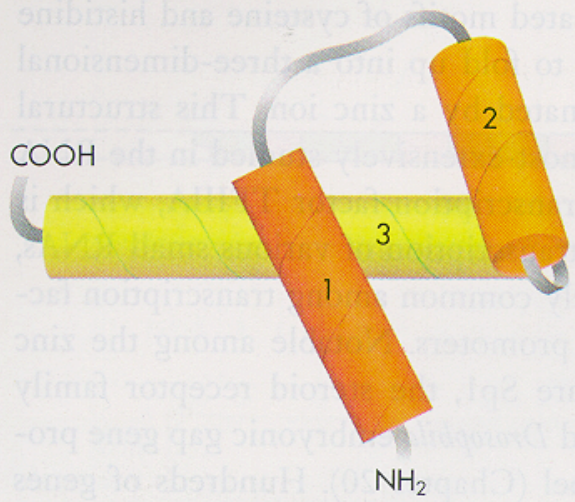
- Lawrence et al. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Sequence Alignment" Science 1993

The Double Helix

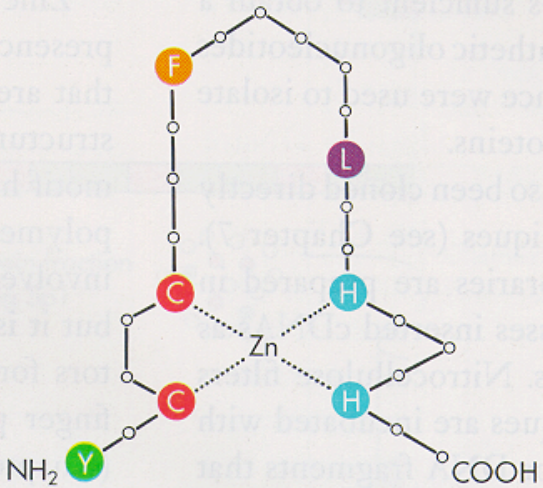


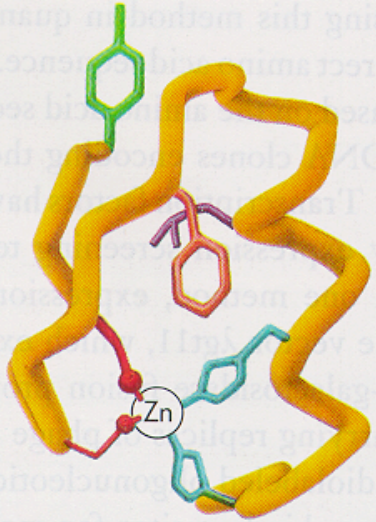
Los Alamos Science

Some DNA Binding Domains

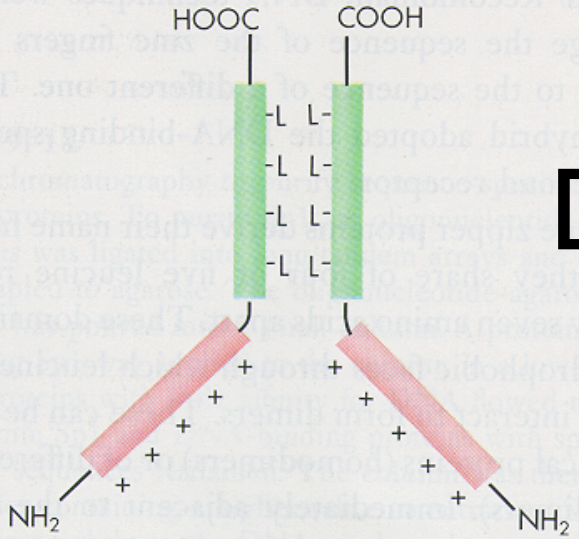


(a) Helix-turn-helix Homeodomain

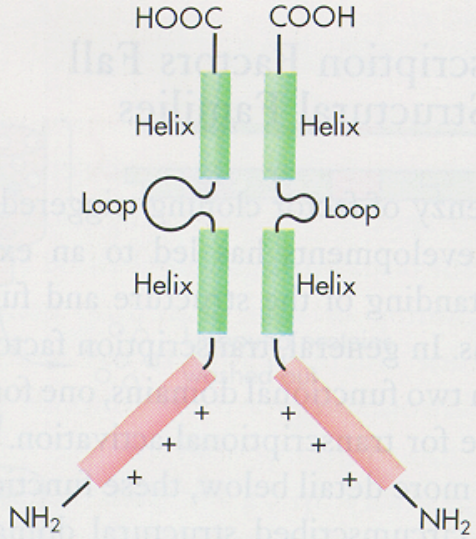




(b) C_2H_2 zinc finger



(c) Leucine zipper



(d) Helix-loop-helix

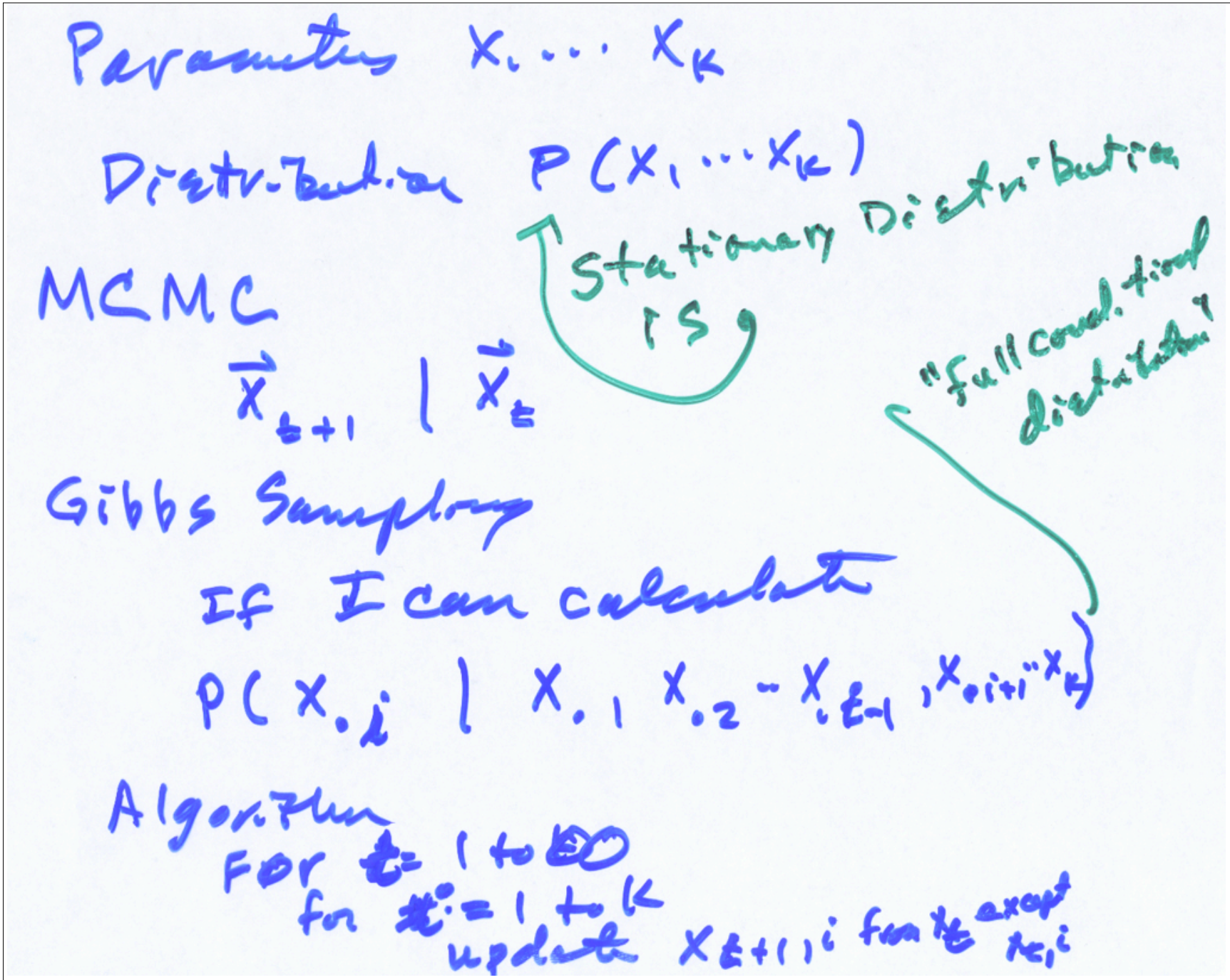
```
Sigma-37          223 IIDLTYIQNK SQKETGDILGISQMHVSR LQRKAVKKLR  240    A25944
SpoIIIC            94 RFGLDLKKEK TQREIAKELGISRSYVSR IEKRALMKMF  111    A28627
NahR               22 VVFNQLLVDR RVSITAENLGLTQPAVSN ALKRLRTSLQ   39    A32837
Antennapedia      326 FHFNRYLTRR RRIEIAHALCLTERQIKI WFQNRRMKWK  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 WKEVIPEKDA   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


Parameters $X_1 \dots X_K$
Distribution $P(X_1 \dots X_K)$
MCMC
$\vec{X}_{t+1} \mid \vec{X}_t$
Stationary Distribution
IS
"full conditional distribution"
Gibbs Sampling
If I can calculate
$P(X_{\cdot i} \mid X_{\cdot 1}\ X_{\cdot 2} \sim X_{\cdot t-1}, X_{\cdot i+1} \dots X_K)$
Algorithm
For t = 1 to ∞
for i = 1 to K
update $X_{t+1,i}$ from X_t except $X_{t,i}$

Again Sequences $S_1 \ldots S_K$

1 motif instance per sequence

each of length w

Motif model — WMM

↳ parameters for $1 \le i \le K$

$1 \le x_i \le |S_i| - w + 1$

"full joint"

$\text{Prob}(x_i = j \mid x_1 \ldots x_{i-1}, x_{i+1} \ldots x_K)$

build WMM from $x_1 \ldots x_K$ excl x_i

calc prob that i^{th} motif @ j

initial x_i's at random

for $t = 1$ to ...

for $i = 1$ to K

throw out motif instance from sequence i

calc WMM from rest

for $j = 1 \ldots |s_i| - w + 1$

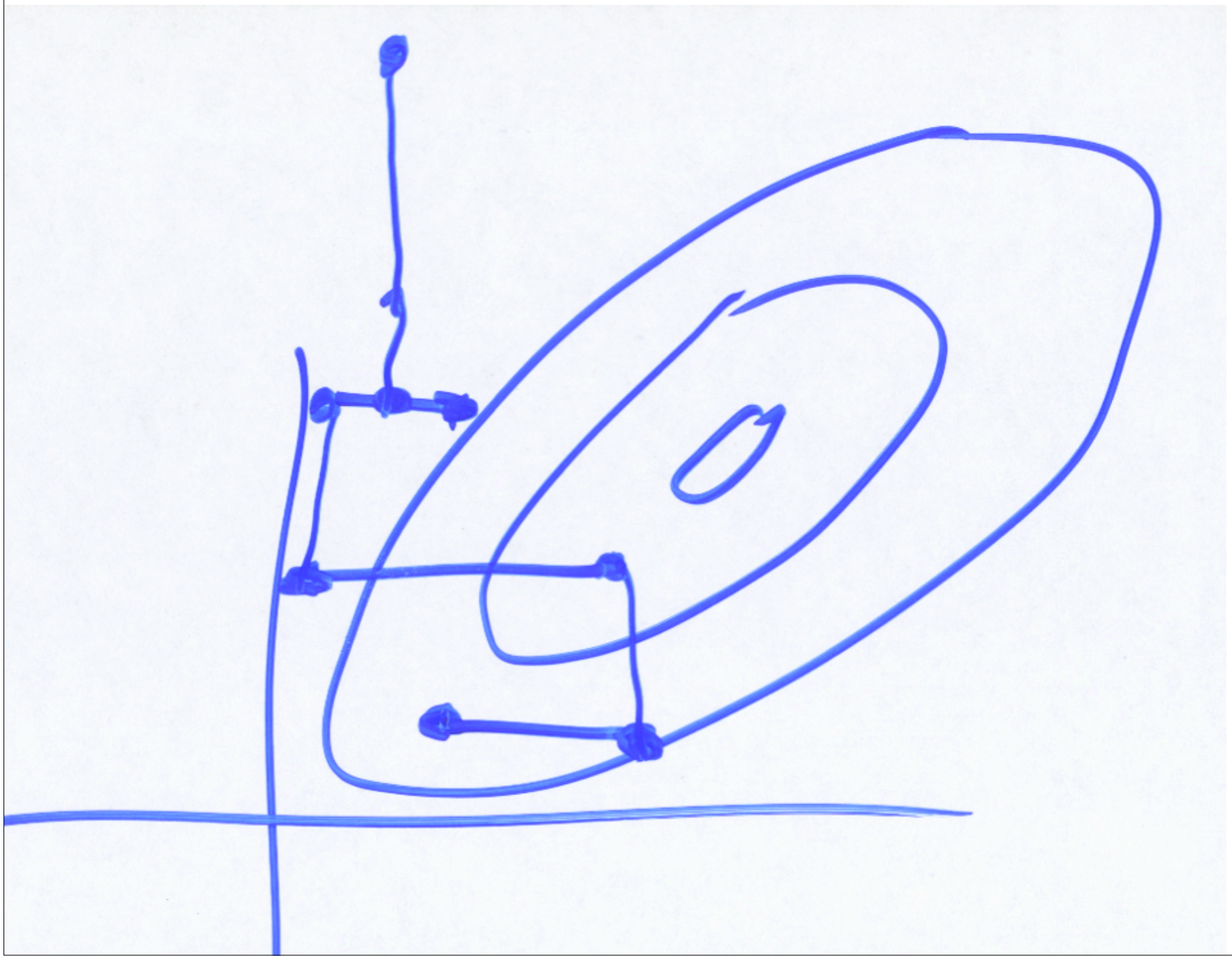
calc prob that i^{th} motif is @ j

pick x_i based on that probability distribution.

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

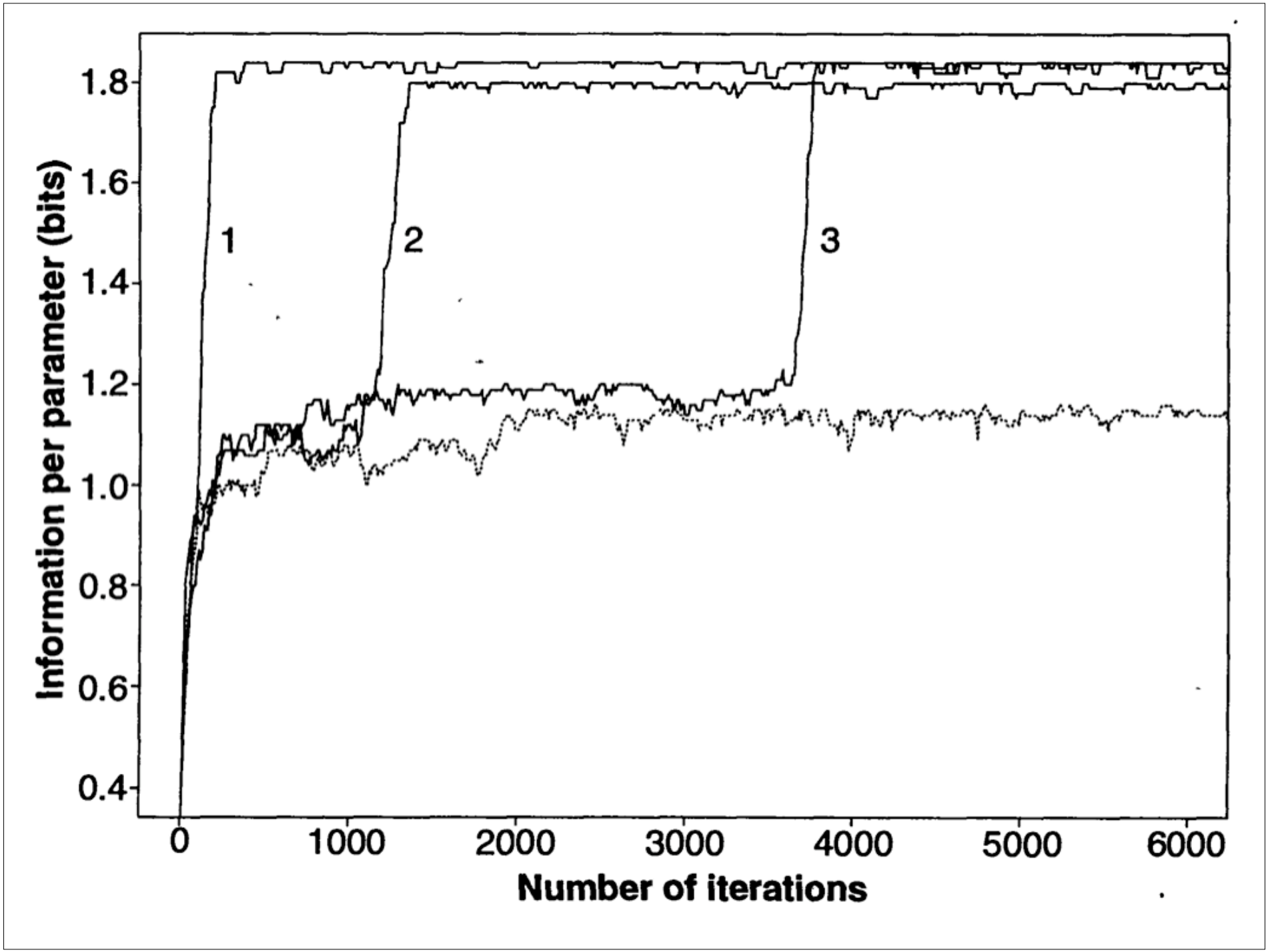
Issues

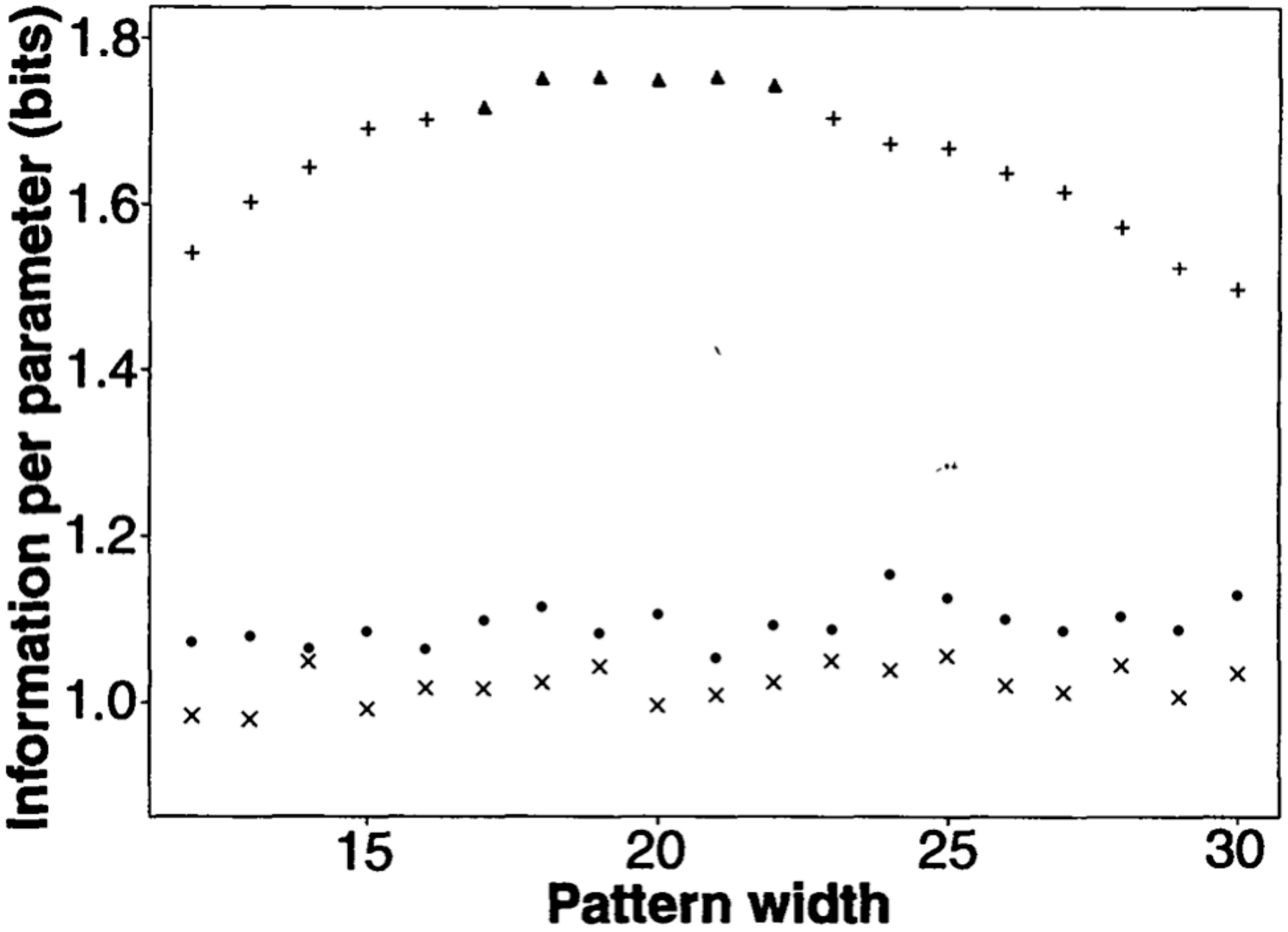
- Burnin - how long must we run the chain to reach stationarity?
- Mixing - how long a post-burnin sample must we take to get a good sample of the stationary distribution? (Recall that individual samples are not independent, and may not "move" freely through the sample space.)



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


Information per parameter (bits)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
1
2
3
0
1000
2000
3000
4000
5000
6000
Number of iterations


Information per parameter (bits)
1.8
1.6
1.4
1.2
1.0
15
20
25
30
Pattern width