

Computational Genomics

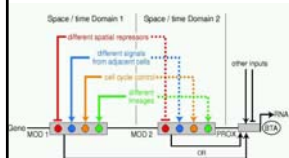
10-810/02-710, Spring 2009

Advanced models and algorithms for motif detection

Eric Xing

Lecture 9, February 11, 2009

Reading: Durbin Chap 9,
class assignment



© Eric Xing @ CMU, 2005-2009

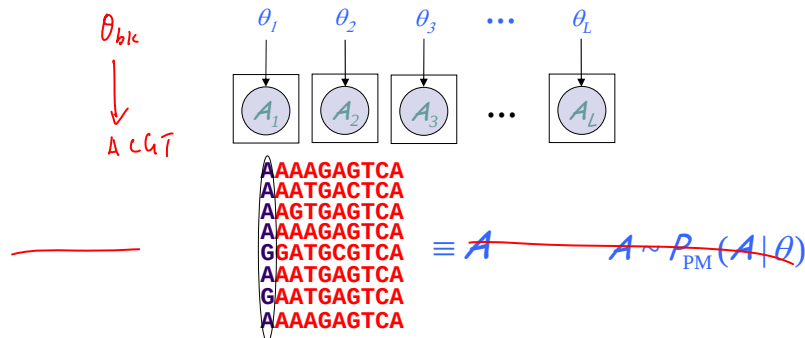
1

The product multinomial model

[Lawrence *et al.* Science 1993]

- Positional specific multinomial distribution:

$$\theta_l = [\theta_{lA}, \dots, \theta_{lC}]^T$$

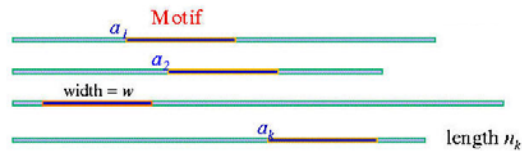


- Position weight matrix (PWM): θ
 - The nucleotide distributions at different positions are independent

© Eric Xing @ CMU, 2005-2009

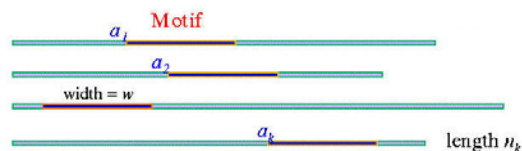
2

The oops model:



Data: sequence: $Y = \text{motif} + bk.$
 model para: $\theta \quad \theta_{bk}$
 unknown: motif location: $z_k = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} a_k$

The oops model:

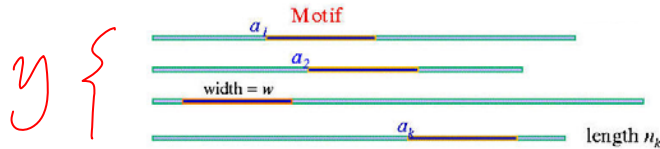


Data: sequence: $Y = \text{motif} + bk.$
 model para: $\theta \quad \theta_{bk}$
 unknown: motif location: $z_k = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} a_k$

given z_k, n_k , we can form alignment matrix A .
 having A , we can compute θ, θ_{bk} .

The oops model:

$$A = \begin{bmatrix} y(a_1) & \dots & y(a_1+L-1) \\ y(a_2) & \dots & y(a_2+L-1) \\ \vdots & & \vdots \\ y(a_k) & \dots & y(a_k+L-1) \end{bmatrix}$$



Data: sequence: $Y = \text{motif} + bk$.

model para: $\theta \quad \theta_{bk}$

unknown: motif location: $z_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} a_k \\ n_k \end{matrix}$

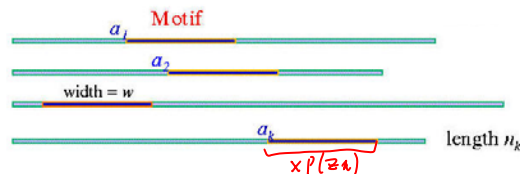
given z_k, v_k , we can form alignment matrix A .
having A , we can compute θ, θ_{bk} .

© Eric Xing @ CMU, 2005-2009

5

The oops model:

$$A = \begin{bmatrix} y(a_1) & \dots & y(a_1+L-1) \\ y(a_2) & \dots & y(a_2+L-1) \\ \vdots & & \vdots \\ y(a_k) & \dots & y(a_k+L-1) \end{bmatrix}$$



Data: sequence: $Y = \text{motif} + bk$.

model para: $\theta \quad \theta_{bk}$

unknown: motif location: $z_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} a_k \\ n_k \end{matrix}$

given z_k, v_k , we can form alignment matrix A .
having A , we can compute θ, θ_{bk} .

EM: initial: guess a_k, v_k .
compute $\theta' = f(A')$
or directly guess θ'

iterate:

① compute $P(z_k^i | Y_k, \theta^i), \forall i, k$
(how?)

form $\langle A \rangle$:

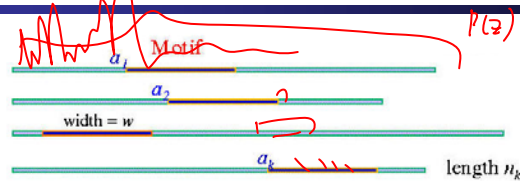
(how?)

② compute θ^{t+1} from $\langle A \rangle$

© Eric Xing @ CMU, 2005-2009

6

The pops model:



EM: initial: guess $a_k, \forall k$.
 compute $\theta' = f(A')$
 or directly guess θ'
 iterate:
 ① compute $P(Z_k^i | Y_k, \theta^i), \forall i, k$
 Form $\langle A \rangle$?
 ② compute θ^{i+1} from $\langle A \rangle$

variants:

Max: $A_{max} = \left[y(a_k^*) \right]$

sampling:

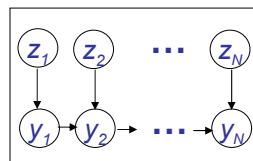
Random (10⁴)

© Eric Xing @ CMU, 2005-2009

7

The Zoops model:

Z_k



One more parameter: " ϵ "
 which is the frequency of motif
 $P(z=1) = \epsilon$.

$Z \in \{0, 1\}^N$

EM: initial: guess $a_k, \forall k$.
 compute $\theta' = f(A')$
 or directly guess θ'

iterate:

① compute $P(Z_k^i | Y_k, \theta^i), \forall i, k$ $\Rightarrow P(Z_{k,i}=1 | Y_k, \theta^i, \theta_{bk})$
 $\hookrightarrow [0, 1]$

Form $\langle A \rangle$

② compute θ^{i+1} from $\langle A \rangle$

$y(z_{k,i})$
 $\frac{y(z_{k,i})}{i}$ k

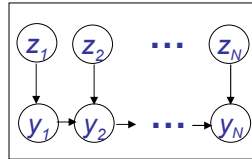
$= \frac{P(y(z_{k,i}) | \theta)}{P(y(z_{k,i}) | \theta) + P(y(z_{k,i}) | \theta_{bk})}$

$P(z_{k,i}=0 | \text{---}) = \frac{P(y(z_{k,i}) | \theta_{bk})}{P(y(z_{k,i}) | \theta) + P(y(z_{k,i}) | \theta_{bk})}$

© Eric Xing @ CMU, 2005-2009

8

Bayesian models:



One more parameter: " ϵ "
which is the frequency of motif
 $P(z=1) = \epsilon$.
 $Z \in \{0, 1\}^N$

EM: initial: guess $\theta_k, \forall k$.
compute $\theta' = f(\theta)$
or directly guess θ'
iterate:
① compute $P(z_k^i | y_k, \theta^i), \forall i, k$

Form $\langle A \rangle$

② compute θ^{i+1} from $\langle A \rangle$

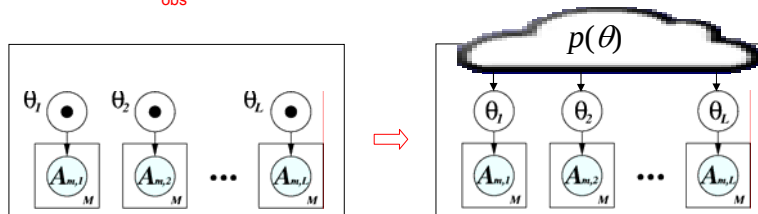
$p(\theta | z)$
 $p(\theta | y)$, γ
 $p(\theta | y, z, \gamma)$

© Eric Xing @ CMU, 2005-2009

9

Bayesian approach

- For model $P(A|\theta)$:
 - Treat parameters θ as unobserved random variable(s)
 - probabilistic statements of θ is conditioned on the values of the observed variables A_{obs}



- Bayes' rule:

$$p(\theta | A) \propto p(A | \theta) p(\theta)$$

posterior likelihood prior

- Bayesian estimation:

$$\theta_{\text{Bayes}} = \int \theta p(\theta | A) d\theta$$

© Eric Xing @ CMU, 2005-2009

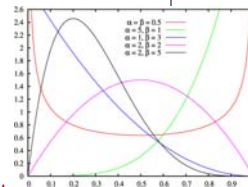
10

Bayesian estimation for Bernoulli

- Beta distribution:

$$P(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} = B(\alpha, \beta) \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$x_1, \dots, x_N \quad p(x_1, \dots, x_N) = \theta^{n_h} (1-\theta)^{n_t}$$



- Posterior distribution of θ :

$$P(\theta | x_1, \dots, x_N) = \frac{p(x_1, \dots, x_N | \theta) p(\theta)}{p(x_1, \dots, x_N)} \propto \theta^{n_h} (1-\theta)^{n_t} \times \theta^{\alpha-1} (1-\theta)^{\beta-1} = \theta^{n_h + \alpha - 1} (1-\theta)^{n_t + \beta - 1}$$

- Notice the isomorphism of the posterior to the prior,
- such a prior is called a **conjugate prior**

© Eric Xing @ CMU, 2005-2009

11

Bayesian missing data problem

- θ : parameter of interest
- $\mathbf{X} = \{x_1, \dots, x_N\}$: a set of complete i.i.d. **observations** from a density that depends upon θ : $p(\mathbf{X} | \theta)$

$$p(\theta | \mathbf{X}) = \prod_{i=1, \dots, n} p(x_i | \theta) p(\theta) / p(\mathbf{X})$$

- In **practical** situations, x_i may **not** be completely observed.
 - Assuming the **unobserved values** are missing completely at random,
 - let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, $x_i = (y_i, z_i)$, $i = 1, \dots, n$
 - y_i : observed part, z_i : missing part

$$p(\theta | \mathbf{Y}) = \int p(\theta | \mathbf{Y}, \mathbf{Z}) p(\mathbf{Z} | \mathbf{Y}) d\mathbf{Z}$$

- This integration is usually hard to obtain in close-form $\rightarrow$ **Imputation**

© Eric Xing @ CMU, 2005-2009

12

Example

```

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...HIS7
5' - ATGGCAGAATCACTTTAAACGTGGCCCCACCCGCTGCACCCGTGTCATTTTGTACGTTACTGCGAAATGACTCAACG ...AR04
5' - CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTCGCATCGCCGAAGTGCCATAAAAAATATTTTT ...ILV6
5' - TGGGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAAATGAAAAATTTTCGACAAAATGTATAGTCATTCTATC ...THR4
5' - ACAAGGTACCTTCTG6CCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCGAACATGAAA ...AR01
5' - ATTGATTGACTCATTTTCCTGACTACTACGAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAATAAATAA ...HOM2
5' - GGCGCCACAGTCGCGTTTG6TTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAGTG6GCATGTGCTTCACACA ...PRO3
  
```

Data Imputation

- Multiple values, $z^{(1)}, \dots, z^{(m)}$ are drawn from $p(Z | Y)$ to form m complete data sets.

$$p(\theta | Y) = \int \text{---} p(z | Y) dz.$$

- With these **imputed** data sets and the ergodicity theorem,

$$z \sim p(z | Y)$$

$$p(\theta | Y) \approx 1/m \cdot \{p(\theta | Y, z^{(1)}) + \dots + p(\theta | Y, z^{(m)})\}$$

- But in most applied problems it is **impossible** to draw Z from $(Z | Y)$ directly.

$$p(z | Y, \theta)$$

Data Augmentation



- Tanner and Wong's data augmentation (DA)
 - apply Gibbs Sampler to draw multiples of θ 's and multiples of Z 's jointly from $p(\theta, Z | Y)$
- DA algorithm
 - Notice that: $p(Z | Y) = \int p(\theta, Z | Y) d\theta = \int p(Z | \theta, Y) p(\theta | Y) d\theta \approx 1/m \cdot \{p(Z | \theta^{(1)}, Y) + \dots + p(Z | \theta^{(m)}, Y)\}$
 - I-step $z^{(m)} \sim p(Z | \theta^{(m)}, Y)$ $\leftarrow$ gr the EM like
 - Recall that: $p(\theta | Y) \approx 1/m \cdot \{p(\theta | Y, z^{(1)}) + \dots + p(\theta | Y, z^{(m)})\}$
 - P-step $\theta^{(m)} \sim p(\theta | Y, z^{(m)})$ $\leftarrow$ Bayes rule
 - By iterating between drawing θ from $p(\theta | Y, z)$ and drawing Z from $p(Z | \theta, Y)$, DA constructs a Markov chain whose equilibrium distribution is $p(\theta, Z | Y)$.

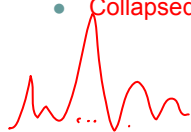
© Eric Xing @ CMU, 2005-2009

15

Collapsed Gibbs Sampler (J. Liu)



- Consider Sampling from $p(\theta | D)$, $\theta = (\theta_1, \theta_2, \theta_3)$
 - Original Gibbs Sampler
 - (i) $\theta_1 \sim p(\theta_1 | \theta_2, \theta_3, D)$ $\mathcal{Z}$
 - (ii) $\theta_2 \sim p(\theta_2 | \theta_1, \theta_3, D)$ $\rightarrow$ PMM
 - (iii) $\theta_3 \sim p(\theta_3 | \theta_1, \theta_2, D)$
 - Collapsed Gibbs Sampler (J. Liu):
 - (i) $(\theta_1, \theta_2) \sim p(\theta_1, \theta_2 | D)$ ✓
 - (ii) $\theta_3 \sim p(\theta_3 | \theta_1, \theta_2, D)$ ✗



© Eric Xing @ CMU, 2005-2009

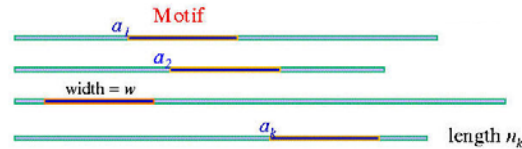
16

Back to *de novo* Motif Detection

— Bayesian missing data problem



- The oops model (one occurrence per sequence)



- Parameter of Interest

$$\theta_{4 \times W} = \begin{pmatrix} \theta_{1A} & \theta_{2A} & \cdots & \theta_{WA} \\ \theta_{1T} & \theta_{2T} & \cdots & \theta_{WT} \\ \theta_{1G} & \theta_{2G} & \cdots & \theta_{WG} \\ \theta_{1C} & \theta_{2C} & \cdots & \theta_{WC} \end{pmatrix} \quad \theta_{0,4 \times 1} = \begin{pmatrix} \theta_{0A} \\ \theta_{0T} \\ \theta_{0G} \\ \theta_{0C} \end{pmatrix}$$

- Missing Data $A = \{a_1, a_2, \dots, a_k\}$
- Observed Data $Y = \text{Given Sequences}$

© Eric Xing @ CMU, 2005-2009

17

The Gibbs Motif Sampler

$$p(a_k | b) \quad \theta$$



- Standard Gibbs Sampler (Iterative sampling):** $p(\theta | A, Y); p(A | \theta, Y)$
 - Draw from $[\theta | A, \text{Data}]$, then draw from $[A | \theta, \text{Data}]$

- Collapsed Gibbs Sampler (Predictive Updating):** $A \sim p(A | Y)$
 - pretend that $K-1$ motif instances have been found. We stochastically predict for the K -th instance!!

- Step0. choose an arbitrary starting point

$$A^{(0)} = (a_1^{(0)}, a_2^{(0)}, \dots, a_K^{(0)});$$

- Step1. Generate $A^{(t+1)} = (a_1^{(t+1)}, a_2^{(t+1)}, \dots, a_K^{(t+1)})$ as follows:

$$\text{Generate } a_1^{(t+1)} \sim p(a_1 | a_2^{(t)}, \dots, a_K^{(t)}, Y);$$

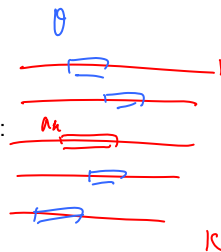
$$\text{Generate } a_2^{(t+1)} \sim p(a_2 | a_1^{(t+1)}, a_3^{(t)}, \dots, a_K^{(t)}, Y);$$

...

$$\text{Generate } a_K^{(t+1)} \sim p(a_K | a_1^{(t+1)}, a_2^{(t+1)}, \dots, a_{K-1}^{(t+1)}, Y);$$

$$\text{Generate } \theta^{(t+1)} \sim p(\theta | a_1^{(t+1)}, a_2^{(t+1)}, \dots, a_K^{(t+1)}, Y);$$

- Step2. Set $t=t+1$, and go to step 1



© Eric Xing @ CMU, 2005-2009

18

The Predictive Update Version

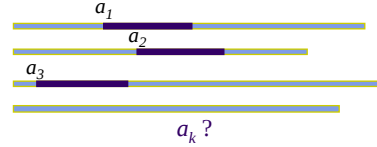


- Initialized by choosing random starting positions

$$a_1^{(0)}, a_2^{(0)}, \dots, a_K^{(0)}$$

- Iterate the following steps many times:

- Randomly or systematically choose a sequence, say, **sequence k**, to exclude.



- Carry out the **predictive-updating** step to update a_k
(no need to sample $\theta^{(t)}$ at each t, we can compute it in close-form, see next slide)

- Notations:

$$\mathcal{A}_{[-k]} \equiv \{a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_K\} \Rightarrow c_{l,j,l-k}, l=1 \dots W, j=A, T, G, C$$

$$a_k \Rightarrow c_{l,j}(a_k)$$

- Stop when not much change observed, or some criterion met.

© Eric Xing @ CMU, 2005-2009

19

Predictive Distribution



- The predictive distribution for a_k :

$$p(a_k = i | \mathcal{A}_{[-k]}, \mathcal{Y}) \approx \text{Const} \cdot \prod_{l=1}^W \prod_j \left(\frac{\hat{\theta}_{l,j,l-k}}{\hat{\theta}_{0,j,l-k}} \right)^{c_{l,j}(i)} \quad (*)$$

- Predictive update:

$$\hat{\theta}_{l,j,l-k} = \frac{c_{l,j,l-k} + \beta_{l,j}}{K-1 + \sum_j \beta_{l,j}}, \quad \hat{\theta}_{0,j,l-k} \text{ similarly}$$

assuming: $\theta_0 \sim \text{Dirichlet}(\alpha)$, $\theta \sim \text{Product Dirichlet}(B)$, $B = (\beta_{l,j})$, $\alpha \sim \text{Uniform}$

- Sampling: every segment of width W in Y_k has probability (*). Choose one at random according to these probabilities, and this becomes the new a_k .

© Eric Xing @ CMU, 2005-2009

20

Derivation



$$p(a_k | a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_K, Y) = \frac{p(Y | a_k, A_{-k}) p(A)}{\sum_i p(Y | a_k = i, A_{-k}) p(A)} = \frac{p(Y | a_k, A_{-k})}{\sum_i p(Y | a_k = i, A_{-k})} \quad (1)$$

$$p(Y | A) = \int p(Y | \theta, \theta_0, A) p(\theta) p(\theta_0) d\theta d\theta_0 \quad (2)$$

$$\begin{aligned} p(Y | \theta, \theta_0, A) p(\theta | \beta) p(\theta_0 | \beta) &= \left(\prod_j \theta_{0,j}^{c_{0,j}(A)} \times \prod_{l,j} \theta_{l,j}^{c_{l,j}(A)} \right) \times \left(\frac{\Gamma(\beta_0)}{\prod_j \Gamma(\beta_{0,j})} \prod_j \theta_{0,j}^{\beta_{0,j}-1} \times \prod_l \frac{\Gamma(\beta_l)}{\prod_j \Gamma(\beta_{l,j})} \prod_j \theta_{l,j}^{\beta_{l,j}-1} \right) \\ &= \frac{\Gamma(\beta)}{\prod_j \Gamma(\beta_j)} \prod_j \theta_{0,j}^{c_{0,j}(A)+\beta_{0,j}-1} \times \prod_l \frac{\Gamma(\beta_l)}{\prod_j \Gamma(\beta_{l,j})} \prod_j \theta_{l,j}^{c_{l,j}(A)+\beta_{l,j}-1} \end{aligned} \quad (3)$$

$$\begin{aligned} p(Y | A) &= \int p(Y | \theta, \theta_0, A) p(\theta) p(\theta_0) d\theta d\theta_0 \\ &= \left(\frac{\Gamma(\sum_j \beta_{0,j})}{\prod_j \Gamma(\beta_{0,j})} \times \frac{\Gamma(\sum_j \beta_j + c_{0,j}(A))}{\Gamma(\sum_j \beta_{0,j} + c_{0,j}(A))} \right) \times \left(\prod_l \frac{\Gamma(\sum_j \beta_{l,j})}{\prod_j \Gamma(\beta_{l,j})} \times \frac{\Gamma(\sum_j \beta_{l,j} + c_{l,j}(A))}{\Gamma(\sum_j \beta_{l,j} + c_{l,j}(A))} \right) \\ &= \left(\frac{\Gamma(\sum_j \beta_{0,j})}{\prod_j \Gamma(\beta_{0,j})} \times \frac{\Gamma(\sum_j \beta_j + c_{0,j}(A_{-k}) + c_{0,j}(a_k))}{\Gamma(\sum_j \beta_{0,j} + c_{0,j}(A_{-k}) + c_{0,j}(a_k))} \right) \times \left(\prod_l \frac{\Gamma(\sum_j \beta_{l,j})}{\prod_j \Gamma(\beta_{l,j})} \times \frac{\Gamma(\sum_j \beta_{l,j} + c_{l,j}(A_{-k}) + \delta(c_{l,j}(a_k)))}{\Gamma(\sum_j \beta_{l,j} + c_{l,j}(A_{-k}) + \delta(c_{l,j}(a_k)))} \right) \end{aligned} \quad (4)$$

...

$$p(a_k | a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_K, Y) = \frac{p(Y | a_k, A_{-k})}{\sum_i p(Y | a_k = i, A_{-k})} \approx \prod_l \left(\frac{\beta_{l,j} + c_{l,j}(A_{-k})}{\sum_j \beta_{l,j} + K - 1} \right)^{c_{l,j}(a_k)} = \prod_l \left(\frac{\hat{\theta}_{l,j}}{\hat{\theta}_{0,j}} \right)^{c_{l,j}(a_k)} \quad (5)$$

© Eric Xing @ CMU, 2005-2009

21

Phase-shift and fragmentation



- Sometimes get stuck in a local shift optimum



- How to "escape" from this local optimum?

$V(a_k | -)$

- Simultaneous move: $A \rightarrow A + \delta$, $A + \delta = \{a_1 + \delta, \dots, a_K + \delta\}$
- Use a Metropolis step: accept the move with prob = p ,

$$r = \min \left\{ 1, \frac{p(A + \delta | Y)}{p(A | Y)} \right\}$$

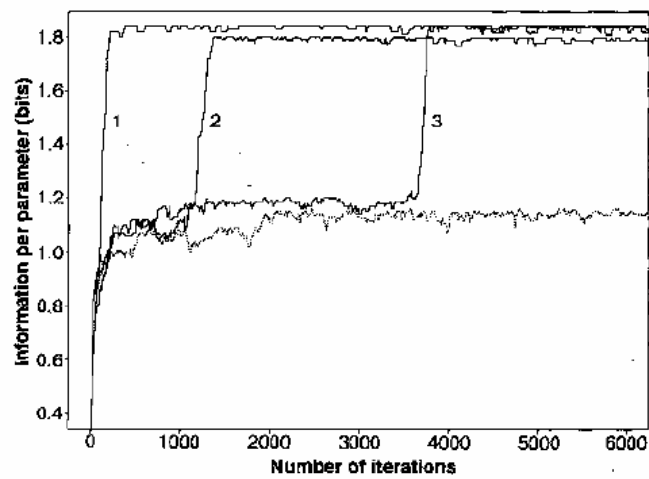


Compare entropies between new columns and left-out ones.

© Eric Xing @ CMU, 2005-2009

22

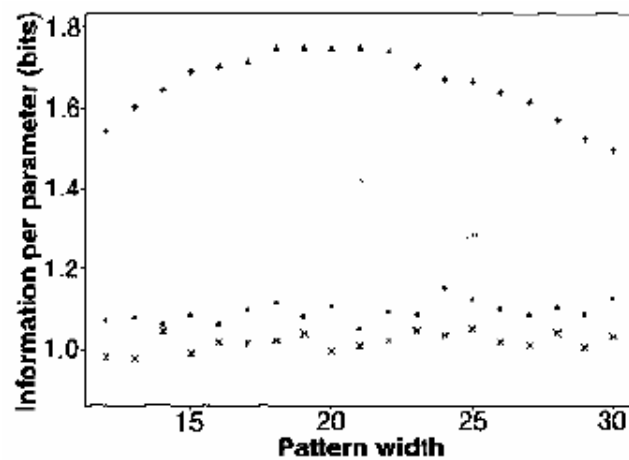
Convergence



© Eric Xing @ CMU, 2005-2009

23

Model Selection



© Eric Xing @ CMU, 2005-2009

24

Natural Extensions to Basic Model I



- Multiple Pattern Occurrences in the same sequences:

Liu, J., "The collapsed Gibbs sampler with applications to a gene regulation problem," *Journal of the American Statistical Association* 89 958-966.

- Prior:** any position i has a small probability ε to start a binding site:

$$A = (a_1, \dots, a_k) \quad P(A) \approx \varepsilon^k (1 - \varepsilon)^{N-k} \quad (\text{with nonoverlapping constraints})$$



- Recall**

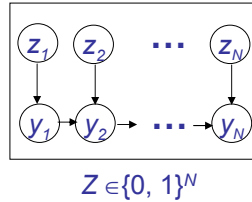
$$p(a_k | a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_k, Y) = \frac{p(Y | a_k, A_{-k}) p(A)}{\sum_i p(Y | a_k = i, A_{-k}) p(A)}$$

- augmented proposal distribution for the Gibbs motif sampler

© Eric Xing @ CMU, 2005-2009

25

Back to the binary indicator model



Although sampling $\mathbf{z}$ directly is prohibitive, a very simple form of the conditional distribution of any z_n given all the rest $\mathbf{z}_{[-n]}$ is available

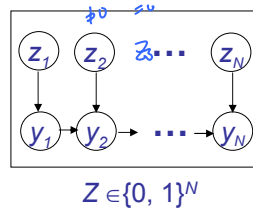
$$\frac{p(z_n = 1 | \mathbf{z}_{[-n]}, \mathbf{y})}{p(z_n = 0 | \mathbf{z}_{[-n]}, \mathbf{y})} = \frac{\hat{\varepsilon}}{1 - \hat{\varepsilon}} \prod_{l=1}^L \prod_{j=1}^4 \left(\frac{\hat{\theta}_{l,j}}{\hat{\theta}_{0,j}} \right)^{\delta(y_{n+l-1}, j)}$$

where $\hat{\varepsilon}$ and $\hat{\theta}$ are estimated from $\mathbf{y}$ and $\mathbf{z}_{[-n]}$.

© Eric Xing @ CMU, 2005-2009

26

Any problem with the model?



- How about multiple types of motifs?
This model can be easily extended to solving K different types motif simultaneously by letting $Z \in \{0, \dots, K\}^{NT}$
- Can motif overlap?
A Markov chain for Z .
- Are motif sites independent?
See next ...

© Eric Xing @ CMU, 2005-2009

27

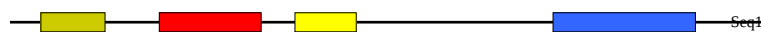
Natural Extensions to Basic Model II

- **Composite Patterns:**

BioOptimizer: the Bayesian Scoring Function Approach to Motif Discovery *Bioinformatics*



- Multiply the motif p -values



$$\begin{aligned} \text{Combined } p\text{-value} &= 0.00001 * 0.0035 * 0.007 * 0.00000005 \\ &= 1.225 * 10^{-17} \end{aligned}$$

© Eric Xing @ CMU, 2005-2009

28

The product of p -values



- Theorem: The probability $F_n(p)$ that the product of n independent, uniform $[0, 1]$ random variables

$$Z_n = \prod_{i=1}^n p_i$$

- has an observed value less than or equal to p , is given by

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!}$$

- for $0 < p \leq 1$, and is zero when p is zero.

Timothy L. Bailey and Michael Gribskov, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, 14:48-54, 1998.

© Eric Xing @ CMU, 2005-2009

29

Natural Extensions to Basic Model III



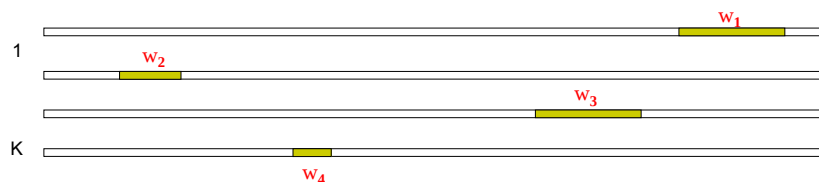
- Correlated in Nucleotide Occurrence in Motif:**

Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 6, 909-916.



- Insertion-Deletion**

BALSA: Bayesian algorithm for local sequence alignment *Nucl. Acids Res.*, 30 1268-77.

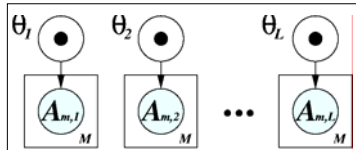


© Eric Xing @ CMU, 2005-2009

30

Recall the PM Model

- The PM parameter, $\theta_l = [\theta_{lA}, \dots, \theta_{lC}]^T$, corresponds exactly to the PWM of a motif



The nucleotide distributions at different sites are independent !

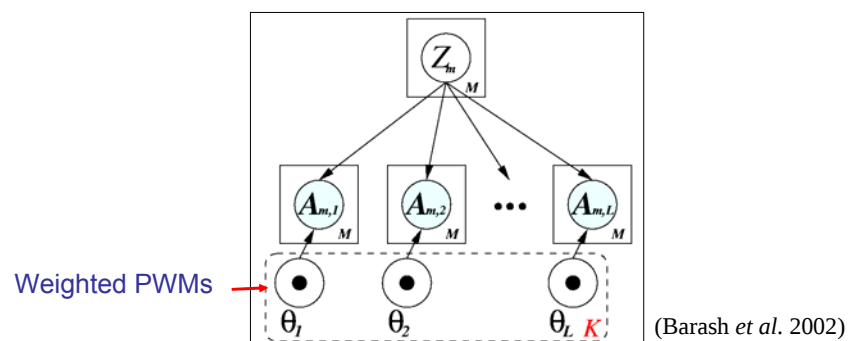
- The **score** (likelihood-ratio) of a candidate substring: **AAAAGAGTCA**

$$R = \frac{p(x = \{\text{AAAAGAGTCA}\} | \text{PWM})}{p(x = \{\text{AAAAGAGTCA}\} | \text{bk})} = \prod_{l=1}^{10} \frac{p(y_l | \text{PWM})}{p(y_l | \text{bk})} = \prod_{l=1}^{10} \frac{\theta_{l,y_l}}{\theta_{0,y_l}}$$

© Eric Xing @ CMU, 2005-2009

31

Mixture of PM models



The nt-distributions at different sites are conditionally independent but marginally dependent !

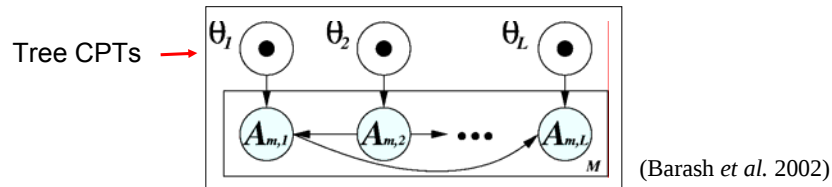
The **likelihood** of a candidate substring: **AAAAGAGTCA**

$$P(x = \{\text{AAAAGAGTCA}\}) = \pi_1 p(\cdot | \text{PWM}_1) + \pi_2 p(\cdot | \text{PWM}_2)$$

© Eric Xing @ CMU, 2005-2009

32

Tree models



The nt-distributions at different sites are pairwise dependent !

The **likelihood** of a candidate substring: **AAAAGAGTCA**

$$P(x = \{\text{AAAAGAGTCA}\}) = \prod_i p(x_i | \text{pi}(x_i)) = p(x_1 | x_2) p(x_1 | x_2) \cdots p(x_m | x_1)$$

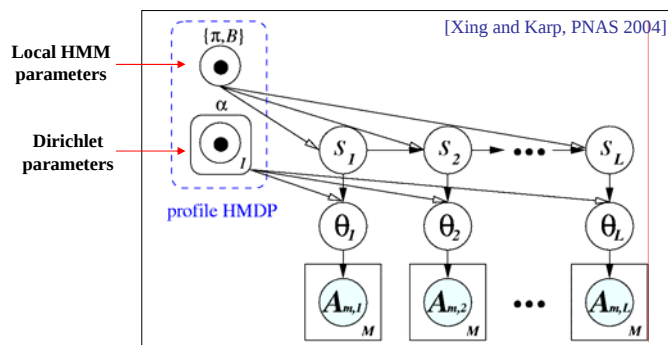
Mixture of trees

...

© Eric Xing @ CMU, 2005-2009

33

MotifPrototyper: A Bayesian Markovian model



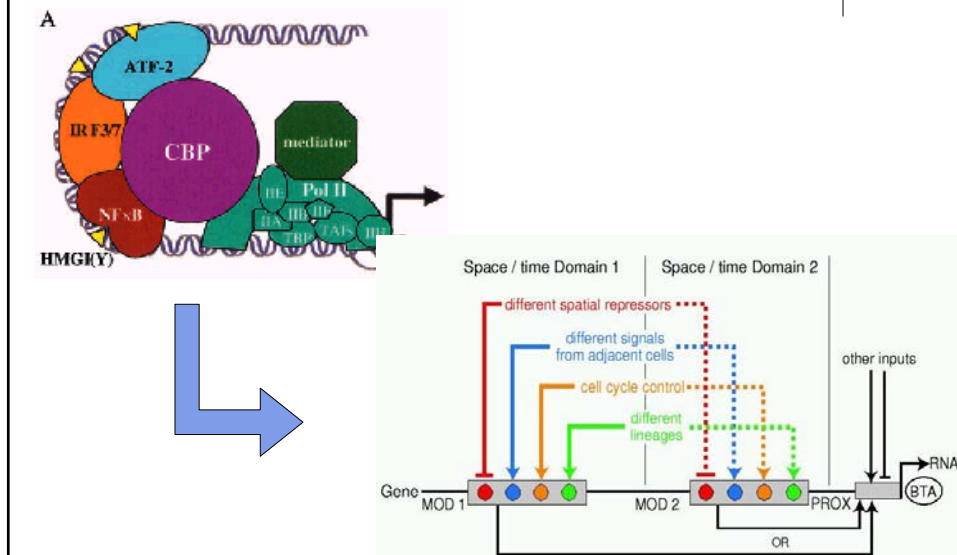
Learning: empirical Bayes estimation:

a family of training motifs $\{A\}_k \Rightarrow$ hyperparameters $\{\alpha, \pi, B\}_k$

© Eric Xing @ CMU, 2005-2009

34

Enhanceosome: Cis-Regulatory Modules



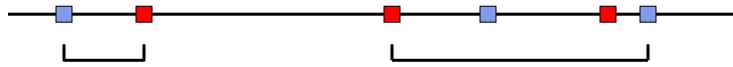
Cluster Finding Methods



- Poisson model *not a model*
 - A. Wagner
- Cister (CIS-element clusTER finder)
 - M. Frith *et al.*
- Comet (Cluster Of Motifs E-value Tool)
 - M. Frith *et al.*
- cis-regulatory module finder
 - Gupta M, Liu JS.
- LOGOS
 - Xing *et al.*

Finding Motif Clusters

- Poisson Method



Exponential distribution:

$$pdf(x) = ae^{-ax}$$

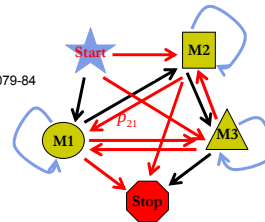
“Pearson type III distribution”:

$$pdf(x) = \frac{a}{(k-2)!} (ax)^{k-2} e^{-ax}$$

- Regulatory Modules:

Cister & Comet

De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci USA*, 102, 7079-84
LOGOS



© Eric Xing @ CMU, 2005-2009

37

Cister & Comet: Introduction

DNA sequence

segment



$$\text{score}(\text{segment}) = \ln \left[\frac{\text{Prob}(\text{segment} | \text{cluster model})}{\text{Prob}(\text{segment} | \text{random model})} \right]$$

Cluster model:

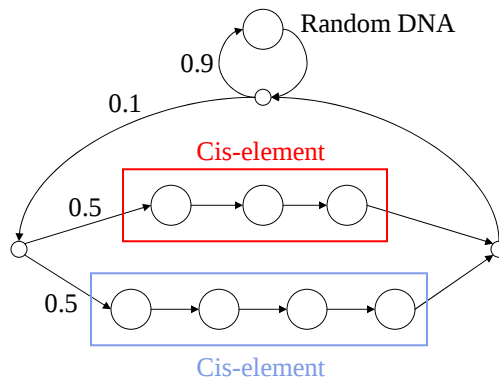


Poisson-distributed cis-elements, embedded in random DNA

© Eric Xing @ CMU, 2005-2009

38

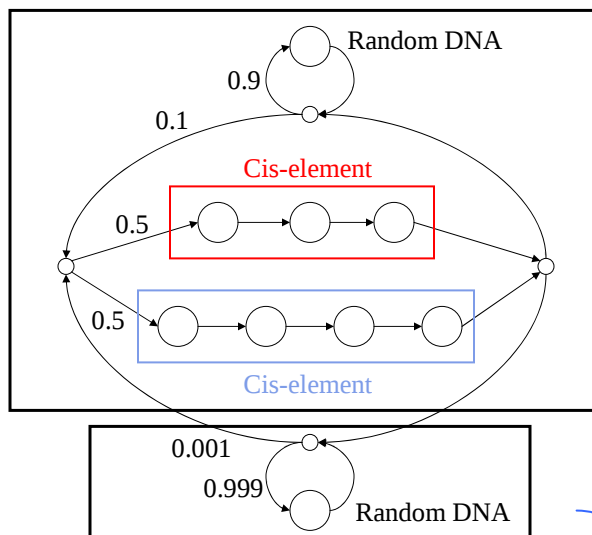
Hidden Markov Model



© Eric Xing @ CMU, 2005-2009

39

Cister



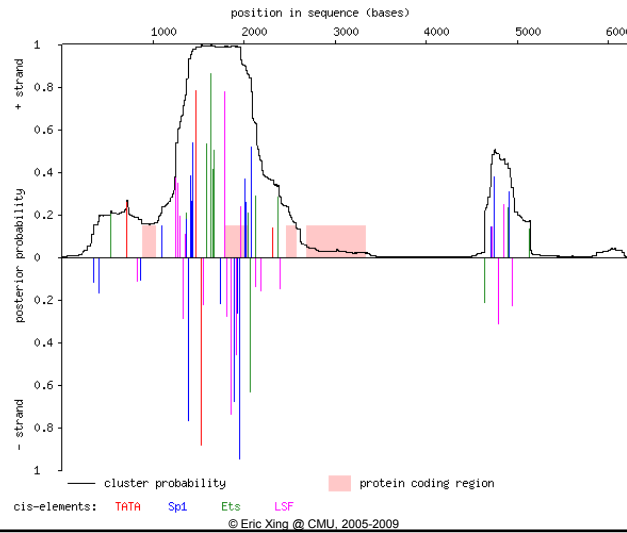
Cluster model

Random model

© Eric Xing @ CMU, 2005-2009

40

Cister applied to Human c-fos



41

Comet

DNA sequence

segment

$$\text{score}(\text{segment}) = \ln \left[\frac{\text{Prob}(\text{segment} \wedge \text{optimal parse} \mid \text{cluster model})}{\text{Prob}(\text{segment} \mid \text{random model})} \right]$$

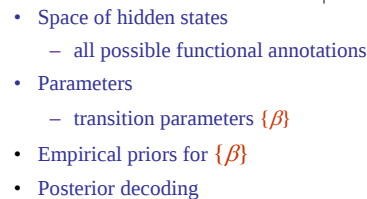
Parse:



One particular arrangement of cis-elements in the segment

© Eric Xing @ CMU, 2005-2009

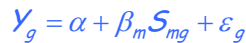
42



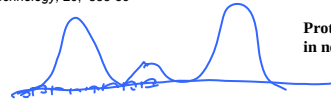
© Eric Xing @ CMU, 2005-2009

43

- Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci. 100.3339-44



- An Algorithm for Finding Protein-DNA Interaction Sites with Applications to Chromatin Immunoprecipitation Microarray Experiments *Nature Biotechnology*, 20, 835-39



Coding regions

© Eric Xing @ CMU, 2005-2009

44

Phylogenetic Footprinting (homologous detection)

- Term originated in 1988 in Tagle et al. **Blanchette et al.:** For unaligned sequences related by phylogenetic tree, find all segments of length k with a history costing less than d . Motif loss an option.



Discussion

- Model versus algorithm:
 - Model: e.g., oops, zoops, HMM,
 - Algorithms: EM, Gibbs, heuristics ...
- Different algorithms can be used for solving the same model!
 - Need to be clear whether improvement/loss is due to model or algorithm (many paper/author got confused with these two aspects, and discuss results in convolved way!)
 - Fix one, and analyze the other, one at a time!!

Databases

TRANSFAC: <http://www.gene-regulation.com/pub/databases.html#transfac>



TRANSFAC FACTOR TABLE, Release 7.0 - public - 2005-09-30, (C) Biobase G

```

AC T00302
XX
ID T00302
XX
DT 15.10.1992 (created); ewl.
DT 26.09.2002 (updated); hom.
CC Copyright (C), Biobase GmbH.
XX
FA GAL4
XX
SY GAL4; YPL248C.
XX
OS yeast, Saccharomyces cerevisiae
CC Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
CC Saccharomycetaceae; Saccharomyces.
XX
SQ MKLLSIEQACDICRLKKLCKSEKPKCAKCLNNWECRYSPKTRSPTRAHLTEVER
SQ LERLEQLFLIPFREDLEMIKMDSLQDIALLTGLFVQNVNNDVATORLASVETDML
SQ TLRQHRIGATSSSEESNKGQGLTVSIOAAHNDSTIFLQFMFRAHGFWEEDCM
SQ SDGLSPFKTDPNNNGFFDGGSLICLRLSICGKPNVTHSNVRLPTMTDRYTLASRST
SQ SRLQSYNNFRPYCPIVHSPILMMLYNNQIEIASKDQWQLFNCILAIGAWCIEGST
SQ IDVFYQNAKSHLTQKVFESGSIILVIALHLLSRYTQWRQNTSYNHFHSIRMAISLG
SQ LMRDLPSSFSDSSILEQRRRIWBSVYSWEIQLSLYGRSILQSNITISFFSVQVQRTT
SQ TQPTIVSGHIIETARLLQVPTIYELEKTVYAKSPICAKSLMICNIEIEVWQAAPFLQ
SQ MDISTATLNLKEHPWLSFTFELKWKQLSLIIVYVLEDFNTFTQKKSQLEQDQNDHQS
SQ YEVKRCSIMLSDAQAQRTVMNSVSYMDNHNVTPTFAWNCYFFFAVLVPIKILLNSKSN
SQ AENNETACLOOINTVIMLLKRLATFKIQTCEYIOVLEEVCAFFLLSUCAIPLFHSYN
XX
MX M00049 FSGAL4_01.
MX M00198 FSGAL4_C.
XX
ES R00501 ASFGAL4_01; Quality: 1.
ES R04203 ASFGAL4_02; Quality: 6.

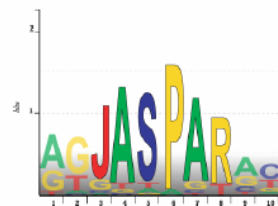
```

Binding Sites

© Eric Xing @ CMU, 2005-2009

47

More Databases



The high-quality transcription factor binding profile database

BROWSE profiles by ID Name Species Class Taxonomic group [link](#)

SEARCH by Name Name Name

combine searches with [link](#)

Species-specific:

- SCPD (yeast) <http://rulai.cshl.edu/SCPD/>
- DPInteract (e. coli) <http://arep.med.harvard.edu/dpinteract/>
- Drosophila DNase I Footprint Database (v2.0) <http://www.flyreg.org/>

© Eric Xing @ CMU, 2005-2009

48

Logos: <http://weblogo.berkeley.edu/>



© Eric Xing @ CMU, 2005-2009

49

Gibbs Motif Sampler

<http://bayesweb.wadsworth.org/gibbs/gibbs.html>



© Eric Xing @ CMU, 2005-2009

50

MEME

<http://meme.sdsc.edu/meme/website/meme.html>



MEME
Multiple Em for Motif Elicitation

Version 3.5.0

Hosted by
NCBIR

Data Submission Form

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (motif) for each pattern it discovers. Your results will be sent to you by e-mail.

Your e-mail address:
kechris@genome.ucsf.edu

Re-enter e-mail address:
kechris@genome.ucsf.edu

Please enter the sequences which you believe share one or more motifs. The sequences may contain no more than 60,000 characters total in any of a large number of formats.

[Optional] Description of your sequences:

Enter the name of a file containing sequences here:
Browse...

or the actual sequences here (Sample Input Sequences):

```
>TBR010C 176 433 649
TTGGTAAAGTAGAGGGGTAATTTTCCCTTTATTTTGTTCATACATT
CTTAAATTGCTTTGCTCTGCTTTTGGAAAGCTATAGTTGGAGCACTG
TTAGGGGAAGGCTGATTAGATATATTTCTGTGATTCTTCTTAAGCCAA
TATTAAGGGAAAGGCTCCAAAAGGCTCGGCAACTGTTGACCTGAT
```

How do you think the occurrences of a single motif are distributed among the sequences?

☒ One per sequence

☐ Zero or one per sequence

[Optional] MEME will find the optimum number of sites for each motif within the limits you specify here:

Minimum sites (>= 2): 17

MEME will find the optimum width of each motif within the limits you specify here:

Minimum width (>= 2): 17

© Eric Xing @ CMU, 2005-2009

51

References



- Reviews
 - Stormo GD (2000), *Bioinformatics*, 16:16-23
 - Bulyk (2003), *Genome Biology* 5:201
- Logos
 - Schneider & Stephens (1990), *Nucleic Acids Res.* 18:6097-6100
- Probabilistic
 - GMS: Lawrence *et al.* (1993), *Science*, 262:208-214
 - MEME: Bailey & Elkan (1995), *Machine Learning*, 21:51-80
 - Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. J. S. Liu, A. F. Neuwald, and C. E. Lawrence. *Journal of the American Statistical Association*, 90, 1156-1170, 1995.

© Eric Xing @ CMU, 2005-2009

52

References



- Structural Constraints
 - Modeling dependencies in protein-dna binding sites, Y. Barash et al. *ISMB* 2003
 - MotifPrototyper: a profile Bayesian model for motif family. E.P. Xing and R. M. Karp, *Proc. Natl. Acad. Sci. vol. 101, no. 29, 10523-10528, 2004.*
 - Kechris et al. (2004), *Genome Biology*, 5(7):R50.
- Regulatory models
 - Frith MC, Hansen U, Weng Z (2001) *Bioinformatics* 17(10): 878-889
 - Frith MC, Spouge JL, Hansen U, Weng Z (2002) *Nucleic Acids Research* (in press)
 - LOGOS: A modular Bayesian model for *de novo* motif detection. E.P. Xing, et al., *Journal of Bioinformatics and Computational biology* 2004, 2(1), 127-154.
 - CisModule: *De novo* discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA*, 101: 12114-12119.
 - *De novo* cis-regulatory module elicitation for eukaryotic genomes M. Gupta and J. Liu, *PNAS*, vol. 101, 7079-7084