Advanced Algorithms and Models for Computational Biology

-- a machine learning approach

Computational Genomics III:

Gibbs motif sampler & advanced motif detection algorithms

Eric Xing

Lecture 8, February 13, 2005

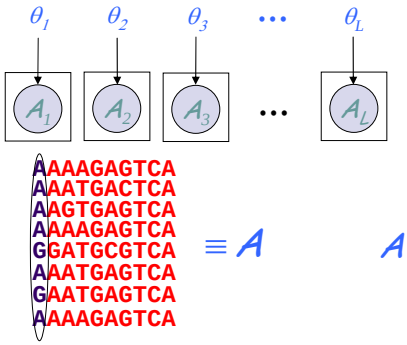
Reading: Chap. 9, DTW book, additional papers

The product multinomial model

[Lawrence *et al*. Science 1993]

- Positional specific multinomial distribution:

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

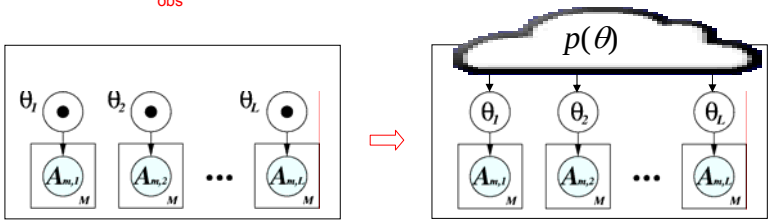


$\equiv A$ $\qquad A \sim P_{\mathrm{PM}}(A \mid \theta)$

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

Bayesian approach

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



- Bayes' rule: $p(\theta|\mathbf{A}) \propto p(\mathbf{A}|\theta)p(\theta)$
  posterior likelihood prior
- Bayesian estimation: $\theta_{Bayes} = \int \theta p(\theta|\mathbf{A})d\theta$

Bayesian missing data problem

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

$$p(\theta \mid \mathbf{X}) = \Pi_{i=1,\ldots,n} p(x_i \mid \theta)\, p(\theta) / p(\mathbf{X})$$

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

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

  - This integration is usually hard obtain in close-form → Imputation

Data Imputation

- Multiple values, $z^{(1)}, \ldots, z^{(m)}$ are drawn from $p(Z \mid \mathbf{Y})$ to form m complete data sets.
- With these **imputed** data sets and the ergodicity theorem,

$$p(\theta \mid \mathbf{Y}) \approx 1/m \cdot \{p(\theta \mid \mathbf{Y}, z^{(1)}) + \ldots + p(\theta \mid \mathbf{Y}, z^{(m)})\}$$

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

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 $p(\theta, Z \mid \mathbf{Y})$
- DA algorithm
  - Notice that: $p(Z \mid \mathbf{Y}) = \int p(\theta, Z \mid \mathbf{Y}) d\theta = \int p(Z \mid \theta, \mathbf{Y}) p(\theta \mid \mathbf{Y}) d\theta$
    $\approx 1/m \cdot \{p(Z \mid \theta^{(1)}, \mathbf{Y}) + \ldots + p(Z \mid \theta^{(m)}, \mathbf{Y})\}$
  - I-step $z^{(m)} \sim p(Z \mid \theta^{(m)}, \mathbf{Y})$
  - Recall that: $p(\theta \mid \mathbf{Y}) \approx 1/m \cdot \{p(\theta \mid \mathbf{Y}, z^{(1)}) + \ldots + p(\theta \mid \mathbf{Y}, z^{(m)})\}$
  - P-step $\theta^{(m)} \sim p(\theta \mid \mathbf{Y}, z^{(m)})$
- By iterating between drawing θ from $p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z})$ and drawing Z from $p(\mathbf{Z} \mid \boldsymbol{\theta}, \mathbf{Y})$, DA constructs a Markov chain whose equilibrium distribution is $p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y})$

Collapsed Gibbs Sampler (J. Liu)

- Consider Sampling from $p(\theta \mid D)$, $\theta=(\theta_1,\theta_2,\theta_3)$
  - Original Gibbs Sampler

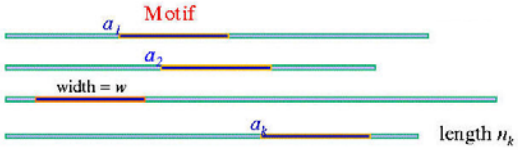
$$\text{(i)}\quad \theta_1 \sim p(\theta_1 \mid \theta_2,\theta_3,D)$$
$$\text{(ii)}\quad \theta_2 \sim p(\theta_2 \mid \theta_1,\theta_3,D)$$
$$\text{(iii)}\quad \theta_3 \sim p(\theta_3 \mid \theta_1,\theta_2,D)$$

  - Collapsed Gibbs Sampler (J. Liu):

$$\text{(i)}\quad (\theta_1,\theta_2) \sim p(\theta_1,\theta_2 \mid D)$$
$$\text{(ii)}\quad \theta_3 \sim p(\theta_3 \mid \theta_1,\theta_2,D)$$

Back to *de novo* Motif Detection
— Bayesian missing data problem

- The oops model (one occurrence per sequence)



- Parameter of Interest

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

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

The Gibbs Motif Sampler

- **Standard Gibbs Sampler (Iterative sampling):** $p(\theta|A,Y)$; $p(A|\theta,Y)$
  - Draw from [θ | A, Data], then draw from [A | θ, Data]
- **Collapsed Gibbs Sampler (Predictive Updating)**: $A \sim p(A \mid Y)$
  - pretend that K-1 motif instances have been found. We *stochastically* predict for the K-th instance!!
    - Step0. choose an arbitrary starting point

      $\mathbf{A^{(0)}=(a_1^{(0)},a_2^{(0)},\ldots,a_K^{(0)})}$;
    - Step1. Generate $A^{(t+1)}=(a_1^{(t+1)}, a_2^{(t+1)},\ldots,a_N^{(t+1)})$ as follows:

      Generate $\mathbf{a_1^{(t+1)} \sim p(a_1 \mid a_2^{(t)},\ldots,a_K^{(t)}, Y)}$;

      Generate $\mathbf{a_2^{(t+1)} \sim p(a_2 \mid a_1^{(t+1)}, a_3^{(t)},\ldots,a_K^{(t)}, Y)}$;

      …

      Generate $\mathbf{a_K^{(t+1)} \sim p(a_N \mid a_1^{(t+1)}, a_2^{(t+1)},\ldots,a_{K-1}^{(t+1)}, Y)}$;

      Generate $\mathbf{\theta^{(t+1)} \sim p(\theta \mid a_1^{(t+1)}, a_2^{(t+1)},\ldots,a_K^{(t+1)}, Y)}$;
    - Step2. Set **t=t+1**, and go to step 1

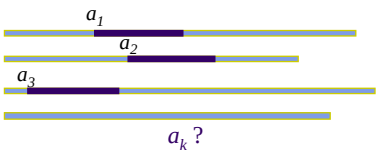
The Predictive Update Version

- Initialized by choosing random starting positions

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

- 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:

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

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

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

Predictive Distribution

- The predictive distribution for a_k:

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

- Predictive update:

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

assuming: $\theta_0 \sim$ Dirichlet(α), $\theta \sim$ Product Dirichlet(B), B = ($\beta_{l,j}$), a ~ 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.

Derivation

$$p(a_k | a_1, \cdots, a_{k-1}, a_{k+1}, \cdots, a_K, Y) = \frac{p(Y \mid a_k, \mathcal{A}_{-k}) p(\mathcal{A})}{\sum_i p(Y \mid a_k = i, \mathcal{A}_{-k}) p(\mathcal{A})} = \frac{p(Y \mid a_k, \mathcal{A}_{-k})}{\sum_i p(Y \mid a_k = i, \mathcal{A}_{-k})} \qquad (1)$$

$$p(Y \mid \mathcal{A}) = \int p(Y \mid \theta, \theta_0, \mathcal{A}) p(\theta) p(\theta_0) d\theta d\theta_0 \qquad (2)$$

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

$$\begin{aligned} p(Y|\mathcal{A}) &= \int p(Y \mid \theta, \theta_0, \mathcal{A}) p(\theta) p(\theta_0) d\theta d\theta_0 \\ &= \left(\frac{\Gamma(\sum_j \beta_{0,j})}{\prod_j \Gamma(\beta_j)} \times \frac{\prod_j \Gamma(\beta_j + c_{0,j}(\mathcal{A}))}{\Gamma(\sum_j \beta_{0,j} + c_{0,j}(\mathcal{A}))} \right) \times \left(\prod_l \frac{\Gamma(\sum_j \beta_{l,j})}{\prod_j \Gamma(\beta_{l,j})} \times \frac{\prod_j \Gamma(\beta_{l,j} + c_{l,j}(\mathcal{A}))}{\Gamma(\sum_j \beta_{l,j} + c_{l,j}(\mathcal{A}))} \right) \\ &= \left(\frac{\Gamma(\sum_j \beta_{0,j})}{\prod_j \Gamma(\beta_j)} \times \frac{\prod_j \Gamma(\beta_j + c_{0,j}(\mathcal{A}_{-k})) + c_{0,j}(a_k))}{\Gamma(\sum_j \beta_{0,j} + c_{0,j}(\mathcal{A}_{-k})) + c_{0,j}(a_k))} \right) \times \left(\prod_l \frac{\Gamma(\sum_j \beta_{l,j})}{\prod_j \Gamma(\beta_{l,j})} \times \frac{\prod_j \Gamma(\beta_{l,j} + c_{l,j}(\mathcal{A}_{-k})) + \delta(c_{l,j}(a_k))}{\Gamma(1 + \sum_j \beta_{l,j} + c_{l,j}(\mathcal{A}_{-k}))} \right) \end{aligned} \qquad (4)$$

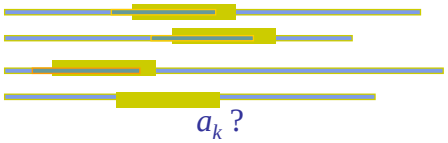
…

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

Phase-shift and fragmentation

- Sometimes get stuck in a local shift optimum

: True motif locations

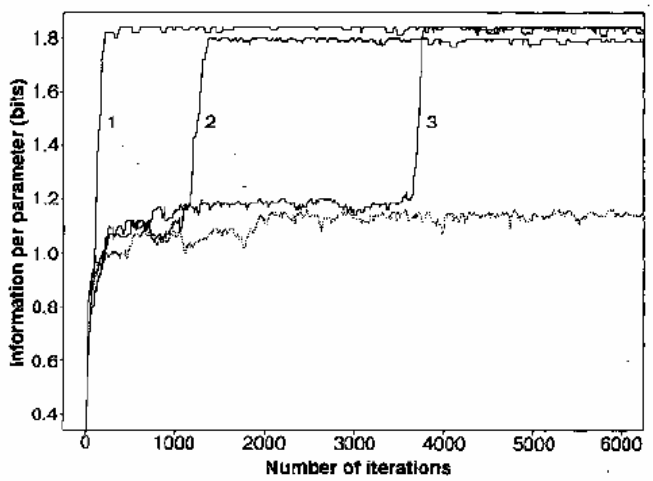


- How to "escape" from this local optimum?
  - Simultaneous move: A $\rightarrow$A+δ, A+δ={a_1+δ, … , a_K+δ}
  - Use a Metropolis step: accept the move with prob=p,

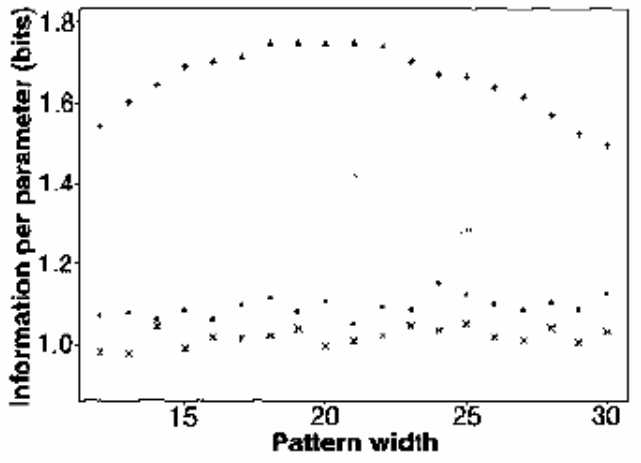
$$r = \min\{1, \frac{p(A+\delta \mid Y)}{p(A \mid Y)}\}$$

Compare entropies between new columns and left-out ones.

Convergence



Model Selection



Natural Extensions to Basic Model I

- **Multiple Pattern Occurances 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 *e* to start a binding site:

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

    width = *w*

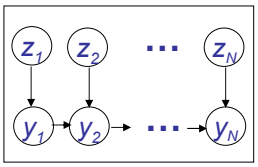
    

  - **Recall**

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

    - **augmented proposal distribution for the Gibbs motif sampler**

Back to the binary indicator model



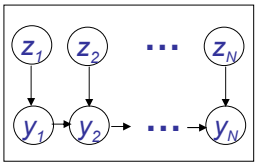
$Z \in \{0, 1\}^N$

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

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

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

Any problem with the model?



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

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

Natural Extensions to Basic Model II

- **Composite Patterns:**

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

- Multiply the motif *p*-values

Combined p-value = 0.00001 * 0.0035 * 0.007 * 0.00000005
= $1.225 * 10^{-17}$

The product of *p*-values

- Theorem: The probability Fn(*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.

Natural Extensions to Basic Model III

- **Correlated in Nucleotide Occurrence in Motif:**

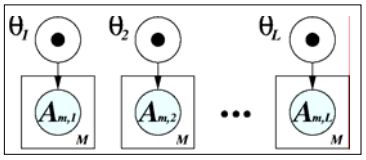
  Modeling within-motif dependence for transcription factor binding site predictions. ***Bioinformatics*, 6,** 909-916.

- **Insertion-Deletion**

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

Recall the PM Model

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

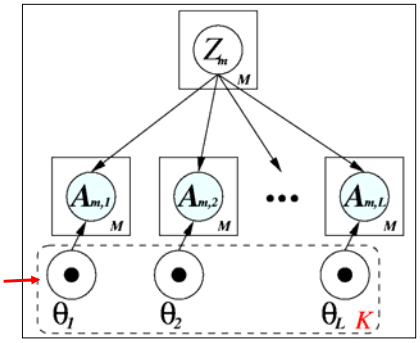


The nucleotide distributions at different sites are independent !

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

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

Mixture of PM models

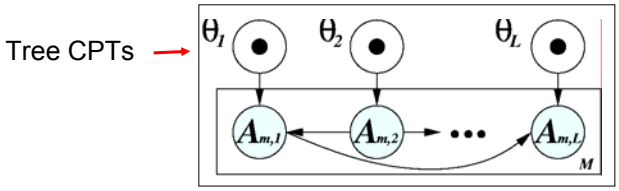


The nt-distributions at different sites are conditionally independent but marginally dependent !

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

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

Tree models



The nt-distributions at different sites are pairwisely dependent !

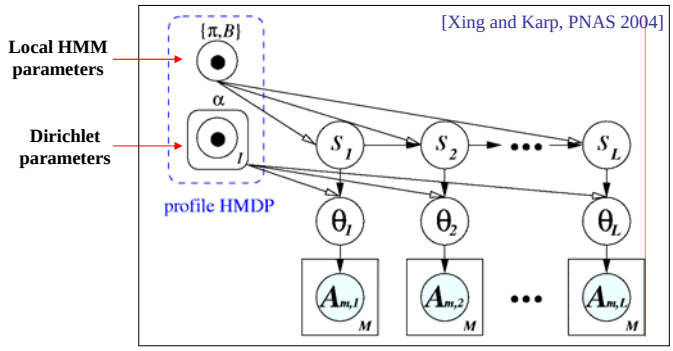
The **likelihood** of a candidate substring: **AAAAGAGTCA**

$$P(x=\{\text{AAAAGAGTCA}\}) = \prod_l p(x_l \mid pi(x_l)) = p(x_1 \mid x_2)\, p(x_l \mid x_2) \cdots p(x_m \mid x_1)$$

Mixture of trees

…

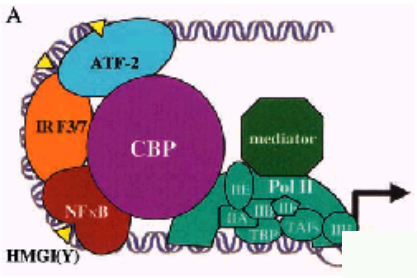
MotifPrototyper: A Bayesian Markovian model

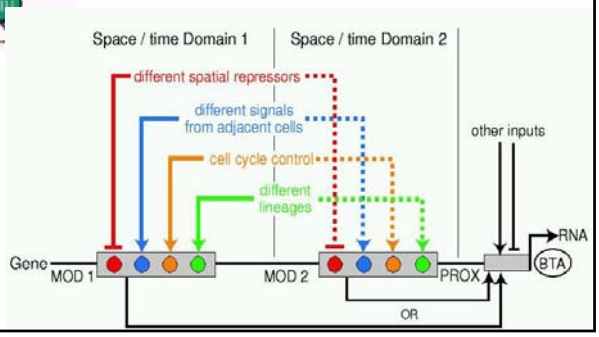


Learning: empirical Bayes estimation:

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

Enhanceosome: Cis-Regulatory Modules



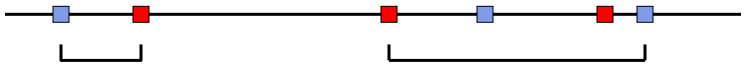


Cluster Finding Methods

- Poisson 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:

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

"Pearson type III distribution":

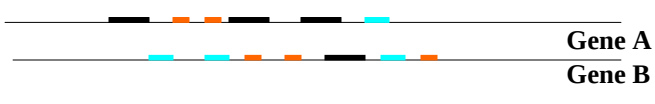
$$pdf\left(x\right) = \frac{a}{(k-2)!}\left(ax\right)^{k-2} e^{-ax}$$

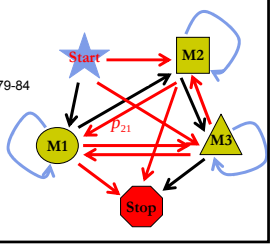
- Regulatory Modules:

Cister & Comet

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

LOGOS





Cister & Comet: Introduction

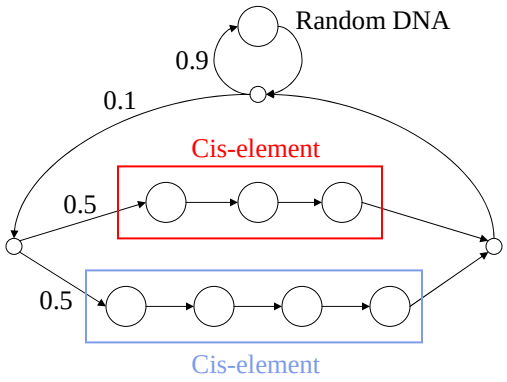
DNA sequence segment

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

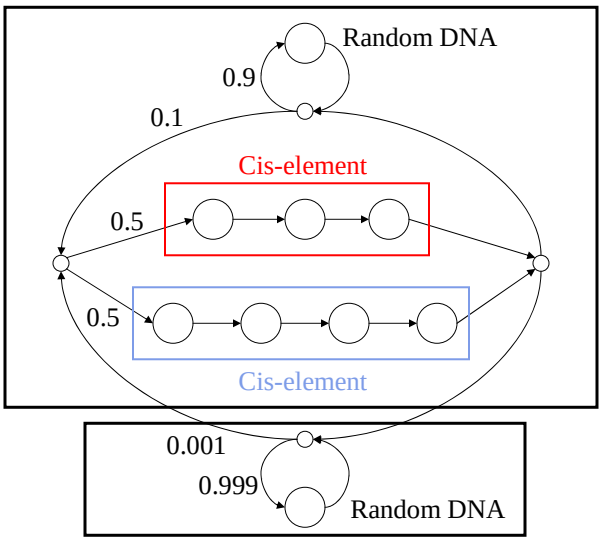
Cluster model:

Poisson-distributed cis-elements, embedded in random DNA

Hidden Markov Model



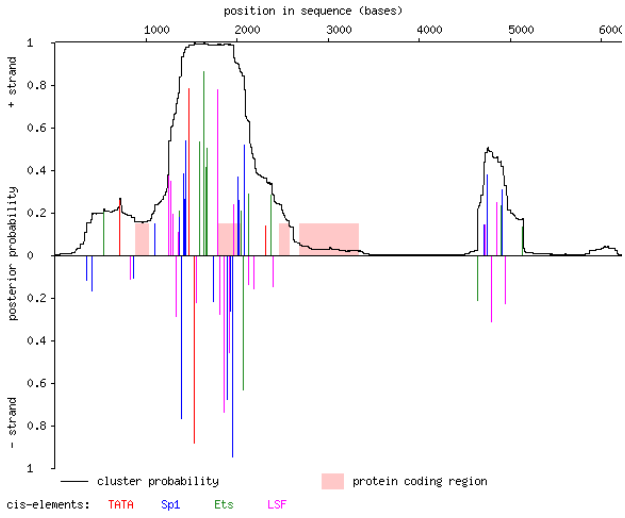
Cister



Cluster model

Random model

Cister applied to Human c-fos



Comet

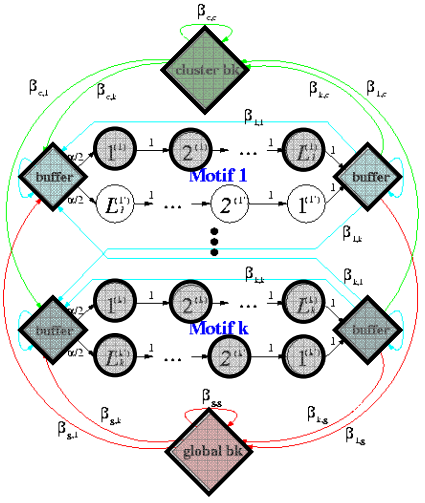
DNA sequence segment

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

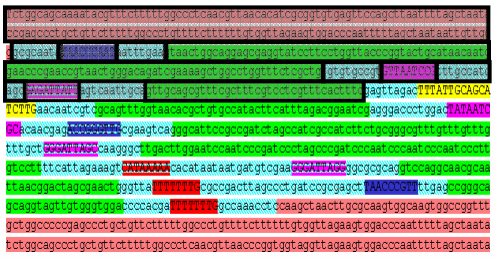
Parse:

One particular arrangement of cis-elements in the segment

The *CisModuler* global HMM



- Space of hidden states
  - all possible functional annotations
- Parameters
  - transition parameters $\{\beta\}$
- Empirical priors for $\{\beta\}$
- Posterior decoding



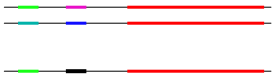
[Xing, Wu, Jordan and Karp. JBCB 2004]

Systems Biology Approach:
Combining Signals and other Data

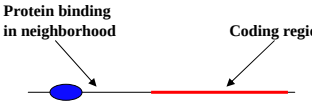
- Expresssion and Motif Regression:
  Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci. 100.3339-44

1. Rank genes by E=log2(expression fold change)
2. Find "many" (hundreds) candidate motifs
3. For each motif pattern *m*, compute the vector ***Sm*** of matching scores for genes with the pattern
4. Regress E on *Sm*

$$y_g = \alpha + \beta_m S_{mg} + \varepsilon_g$$

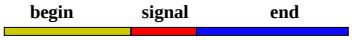


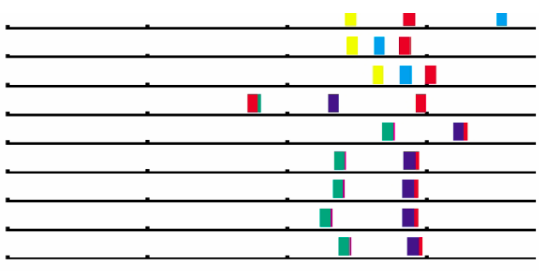
- ChIP-on-chip - 1-2 kb information on protein/DNA interaction:
  An Algorithm for Finding Protein-DNA Interaction Sites with Applications to Chromatin Immunoprecipitation Microarray Experiments *Nature Biotechnology*, 20, 835-39

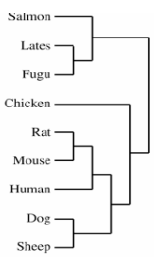


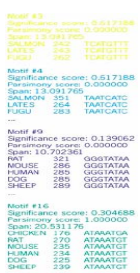
Phylogenetic Footprinting (homologous detection)

- 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.**









Blanchette and Tompa (2003) "FootPrinter: a program designed for phylogenetic footprinting" NAR 31:13:3840-

Databases

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

TRANSFAC

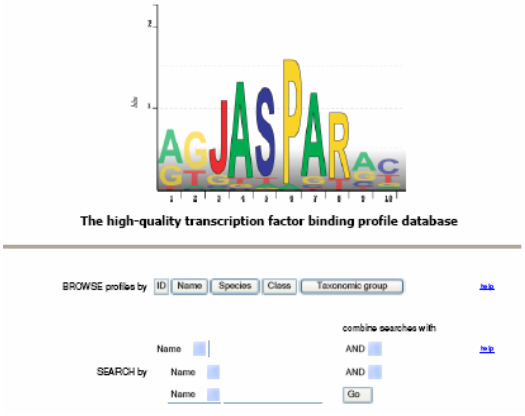
TRANSFAC FACTOR TABLE, Release 7.0 - public - 2005-09-30, (C) Biobase G

```
AC   T00302
XX
ID   T00302
XX
DT   15.10.1992 (created); ewi.
DT   26.08.2002 (updated); hom.
CO   Copyright (C), Biobase GmbH.
XX
FA   GAL4
XX
SY   GAL4; YPL248C.
XX
OS   yeast, Saccharomyces cerevisiae
OC   Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
OC   Saccharomycetaceae; Saccharomyces.
XX
SQ   MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESR
SQ   LERLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPL
SQ   TLRQHRISATSSSEESSNNGSQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDM
SQ   SDGLPFLKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTT
SQ   SRLLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTD
SQ   IDVFYYQNAKSHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLG
SQ   LNRDLPSSFSDSSILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTT
SQ   TGPTIYHGIIETARLLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVWRQAPKFLQ
SQ   MDISTTALTNLLKEHPWLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQS
SQ   YEVKRCSIMLSDAAQRTVMSVSSYMDNHNVTPYFAWNCSYYFFNAVLVPIKTLLSNSKSN
SQ   AENNETAQLLQQINTVLMLLKKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYN

MX   M00049 FSGAL4_01.
MX   M00198 FSGAL4_C.
XX
BS   R00501 ASSGAL4_01; Quality: 1.
BS   R04203 ASSGAL4_02; Quality: 6.
```

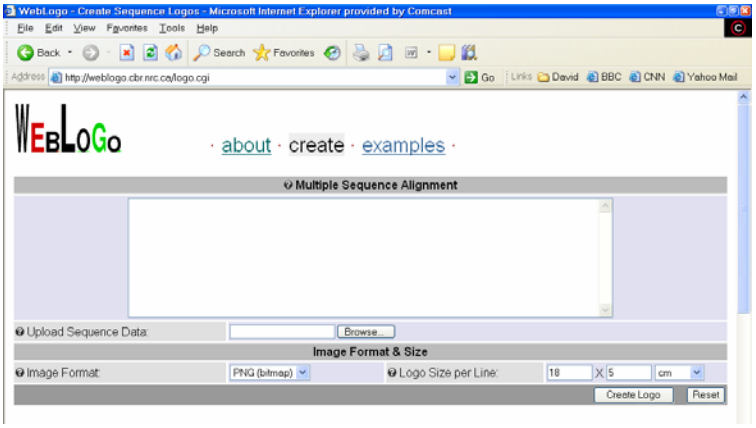
Binding Sites

More Databases



- 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/

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

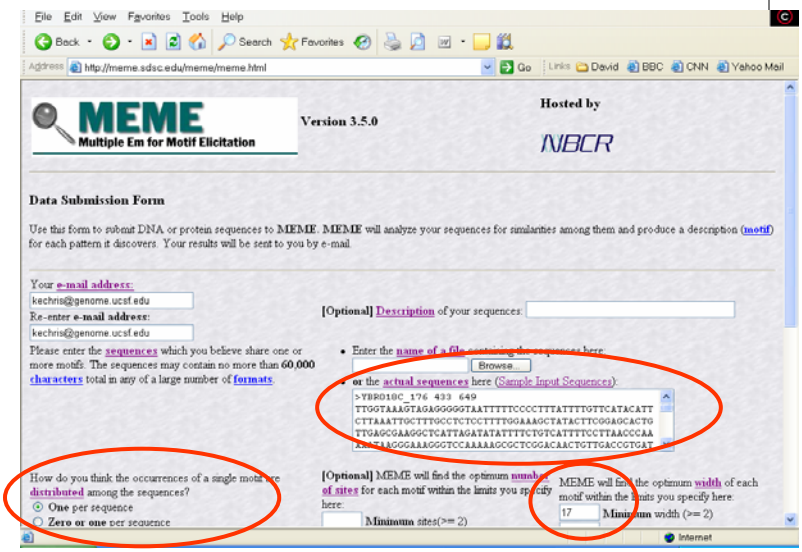


Gibbs Motif Sampler

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

MEME

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



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.

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