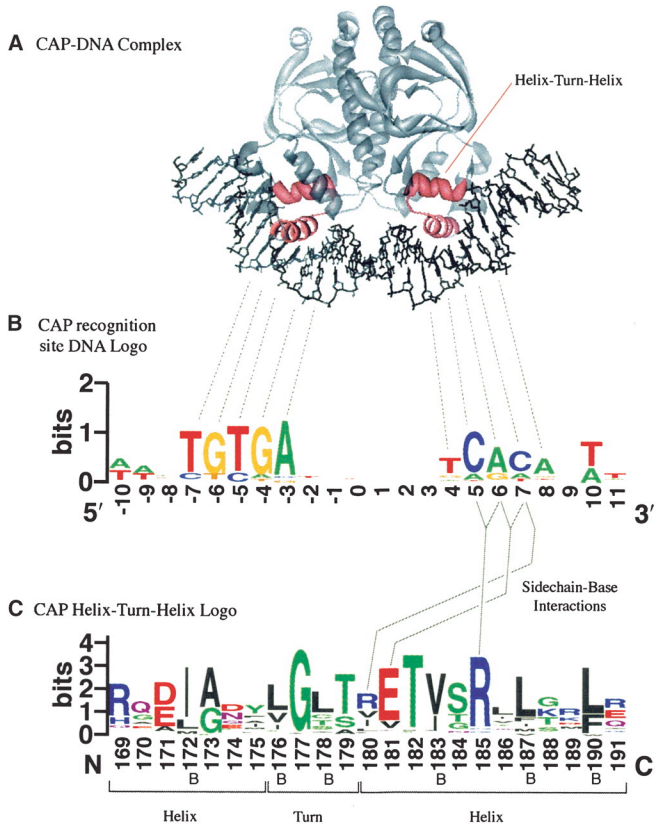
Learning Sequence Motif Models Using Expectation Maximization (EM) and Gibbs Sampling

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2009
Mark Craven
craven@biostat.wisc.edu

Sequence Motifs

- what is a sequence *motif* ?
  - a sequence pattern of biological significance
- examples
  - protein binding sites in DNA
  - protein sequences corresponding to common functions or conserved pieces of structure

Sequence Motifs Example

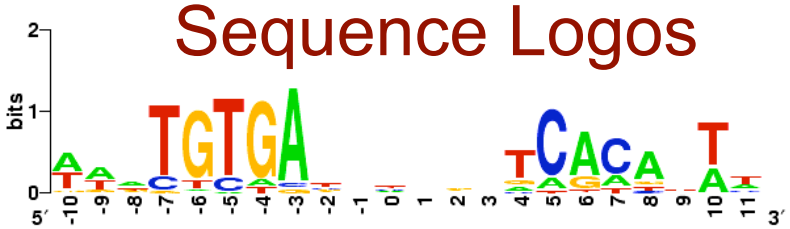


CAP-binding motif model based on 59 binding sites in E.coli

helix-turn-helix motif model based on 100 aligned protein sequences

Figure from Crooks et al., *Genome Research* 14:1188-90, 2004.

Sequence Logos



- based on entropy (H) of a random variable (X) representing distribution of character states at each position

$$H(X) = -\sum_{x} P(x) \log_2 P(x)$$

- height of logo at a given position determined by decrease in entropy (from maximum possible)

$$H_{\max} - H(X) = -\log_2\left(\frac{1}{N}\right) - \left(-\sum_{x} P(x) \log_2 P(x)\right)$$

# of characters in alphabet

- height of each character x is proportional to $P(x)$

The Motif Model Learning Task

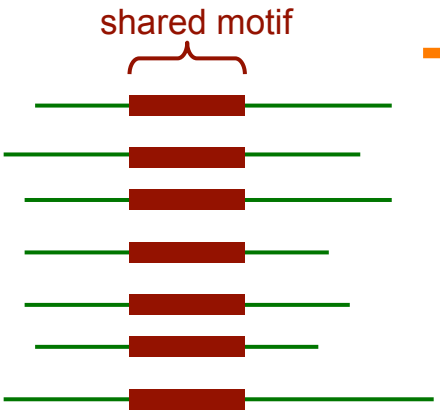
given: a set of sequences that are thought to contain an unknown motif of interest

do:

- infer a model of the motif
- predict the locations of the motif in the given sequences

Motifs and *Profile Matrices* (a.ka. *Position Weight Matrices*)

- given a set of aligned sequences, it is straightforward to construct a profile matrix characterizing a motif of interest



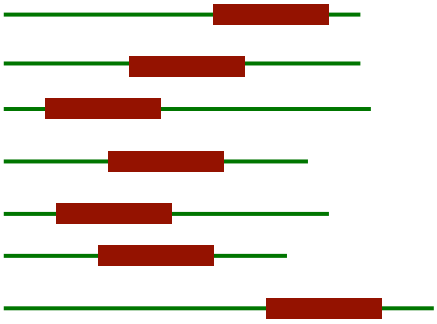
sequence positions

	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

- each element represents the probability of given character at a specified position

Motifs and Profile Matrices

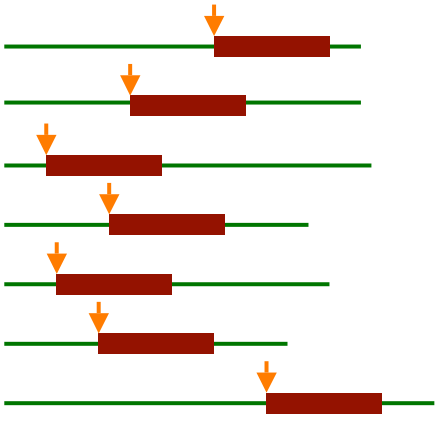
- how can we construct the profile if the sequences aren't aligned?
  - in the typical case we don't know what the motif looks like



- use an Expectation Maximization (EM) algorithm

The EM Approach

- EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
- in our problem, the hidden state is where the motif starts in each training sequence



The MEME Algorithm

- Bailey & Elkan, 1993, 1994, 1995
- uses EM algorithm to find multiple motifs in a set of sequences
- first EM approach to motif discovery: Lawrence & Reilly 1990

Representing Motifs in MEME

- a motif is
  - assumed to have a fixed width, *W*
  - represented by a matrix of probabilities: $p_{c,k}$ represents the probability of character c in column k
- also represent the "background" (i.e. outside the motif) probability of each character: $p_{c,0}$ represents the probability of character c in the background

Representing Motifs in MEME

- example: a motif model of length 3

$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

background

Basic EM Approach

- the element $Z_{i,j}$ of the matrix Z represents the probability that the motif starts in position j in sequence i
- example: given DNA sequences of length 6, where *W*=3

G C T G T A
G C T G T A
G C T G T A
G C T G T A

$Z =$

	1	2	3	4
seq1	0.1	0.1	0.2	0.6
seq2	0.4	0.2	0.1	0.3
seq3	0.3	0.1	0.5	0.1
seq4	0.1	0.5	0.1	0.3

Basic EM Approach

given: length parameter *W*, training set of sequences

set initial values for p

do

re-estimate Z from p (E –step)

re-estimate p from Z (M-step)

until change in p < ε

return: p, Z

The Probability of a Sequence Given a Hypothesized Starting Position

$$P(X_i \mid Z_{i,j} = 1, p) = \prod_{k=1}^{j-1} p_{c_k,0} \prod_{k=j}^{j+W-1} p_{c_k,k-j+1} \prod_{k=j+W}^{L} p_{c_k,0}$$

before motif | motif | after motif

X_i is the i th sequence

$Z_{i,j}$ is 1 if motif starts at position j in sequence i

c_k is the character at position k in sequence i

Example

X_i = G C **T G T** A G

$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$$P(X_i \mid Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

The E-step: Estimating Z

- to estimate the starting positions in Z at step t

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t)}) P(Z_{i,j} = 1)}{\sum_{k=1}^{L-W+1} P(X_i \mid Z_{i,k} = 1, p^{(t)}) P(Z_{i,k} = 1)}$$

- this comes from Bayes' rule applied to

$$P(Z_{i,j} = 1 \mid X_i, p^{(t)})$$

The E-step: Estimating Z

- assume that it is equally likely that the motif will start in any position

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t)}) \cancel{P(Z_{i,j} = 1)}}{\sum_{k=1}^{L-W+1} P(X_i \mid Z_{i,k} = 1, p^{(t)}) \cancel{P(Z_{i,k} = 1)}}$$

Example: Estimating Z

X_i = G C T G T A G

p =

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$$Z_{i,1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i,2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$$\vdots$$

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{i,j} = 1$

The M-step: Estimating p

- recall $p_{c,k}$ represents the probability of character c in position k ; values for $k=0$ represent the background

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1}=c\}} Z_{i,j} & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \end{cases}$$

total # of c's in data set

Example: Estimating p

A C A G C A

$Z_{1,1} = 0.1, \quad Z_{1,2} = 0.7, \quad Z_{1,3} = 0.1, \quad Z_{1,4} = 0.1$

A G G C A G

$Z_{2,1} = 0.4, \quad Z_{2,2} = 0.1, \quad Z_{2,3} = 0.1, \quad Z_{2,4} = 0.4$

T C A G T C

$Z_{3,1} = 0.2, \quad Z_{3,2} = 0.6, \quad Z_{3,3} = 0.1, \quad Z_{3,4} = 0.1$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \ \ldots \ + Z_{3,3} + Z_{3,4} + 4}$$

Representing Motifs in MEME

- example: a motif model of length 3

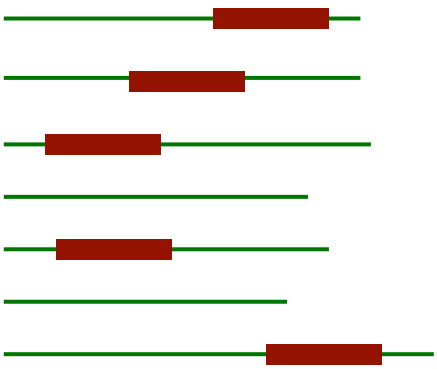
$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

background

The ZOOPS Model

- the approach as we've outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model
- the ZOOPS model assumes *zero or one* occurrences per sequence



E-step in the ZOOPS Model

- we need to consider another alternative: the ith sequence doesn't contain the motif
- we add another parameter (and its relative)

λ

- prior probability that any position in a sequence is the start of a motif

$\gamma = (L - W + 1)\lambda$

- prior probability of a sequence containing a motif

E-step in the ZOOPS Model

$$Z_{i,j}^{(t)} = \frac{P(X_i \mid Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}{P(X_i \mid Q_i = 0, p^{(t)})(1 - \gamma^{(t)}) + \sum_{k=1}^{L-W+1} P(X_i \mid Z_{i,k} = 1, p^{(t)})\lambda^{(t)}}$$

- Q_i is a random variable for which Q_i = 1 if sequence X_i contains a motif, $Q_i = 0$ otherwise

$$P(Q_i = 1) = \sum_{j=1}^{L-W+1} Z_{i,j}$$

$$P(X_i \mid Q_i = 0, p) = \prod_{j=1}^{L} p_{c_j,0}$$

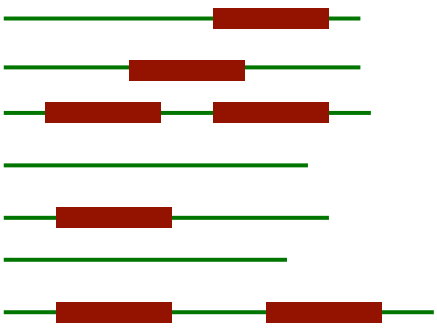
M-step in the ZOOPS Model

- update p same as before
- update γ as follows:

$$\gamma^{(t+1)} \equiv \lambda^{(t+1)}(L - W + 1) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{L-W+1} Z_{i,j}^{(t)}$$

The TCM Model

- the TCM (two-component mixture model) assumes *zero or more* motif occurrences per sequence



Likelihood in the TCM Model

- the TCM model treats each length *W* subsequence independently
- to determine the likelihood of such a subsequence:

$$P(X_{i,j} \mid Z_{i,j} = 1, p) = \prod_{k=j}^{j+W-1} p_{c_k,k-j+1}$$

assuming a motif starts there

$$P(X_{i,j} \mid Z_{i,j} = 0, p) = \prod_{k=j}^{j+W-1} p_{c_k,0}$$

assuming a motif doesn't start there

E-step in the TCM Model

$$Z_{i,j}^{(t)} = \frac{P(X_{i,j} \mid Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}{P(X_{i,j} \mid Z_{i,j} = 0, p^{(t)})(1-\lambda^{(t)}) + P(X_{i,j} \mid Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}$$

subsequence isn't a motif

subsequence is a motif

- M-step same as before

Extending the Basic EM Approach in MEME

- How to choose the width of the motif?
- How to find multiple motifs in a group of sequences?
- How to choose good starting points for the parameters?
- How to use background knowledge to bias the parameters?

Choosing the Width of the Motif

- try various widths
  - estimate the parameters each time
  - apply a likelihood ratio test based on
    - probability of data under motif model
    - probability of data under *null* model
  - penalized by # of parameters in the model

Finding Multiple Motifs

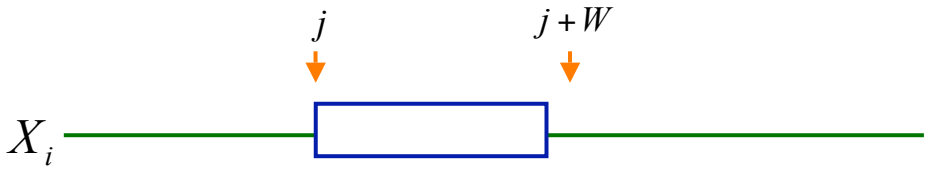
- we might want to find multiple motifs in a given set of sequences
- how can we do this without
  - rediscovering previously learned motifs

  - discovering a motif that substantially overlaps with previously learned motifs

Finding Multiple Motifs

- basic idea: discount the likelihood that a new motif starts in a given position if this motif would overlap with a previously learned one
- when re-estimating $Z_{i,j}$, multiply by $P(V_{i,j}=1)$

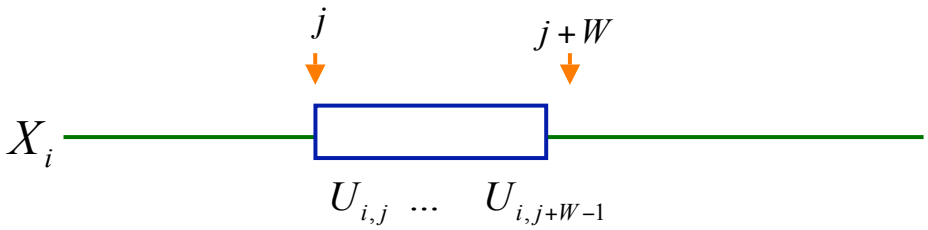
$$V_{i,j} = \begin{cases} 1, & \text{no previous motifs in } [X_{i,j},\ldots,X_{i,j+w-1}] \\ 0, & \text{otherwise} \end{cases}$$



Finding Multiple Motifs

- to determine $P(V_{i,j} = 1)$ need to take into account individual positions in the window

$$U_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \notin \text{previous motif occurrence} \\ 0, & \text{otherwise} \end{cases}$$



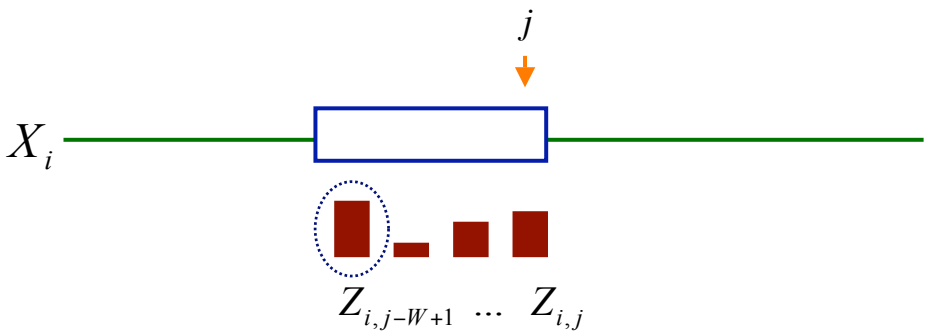
Finding Multiple Motifs

- Updating U after each motif-finding pass

$$U_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \notin \text{previous motif occurrence} \\ 0, & \text{otherwise} \end{cases}$$

"pass" m

$$U_{i,j}^{(m)} = U_{i,j}^{(m-1)} \left(1 - \max\left(Z_{j-W+1}, \ldots, Z_j\right)\right)$$

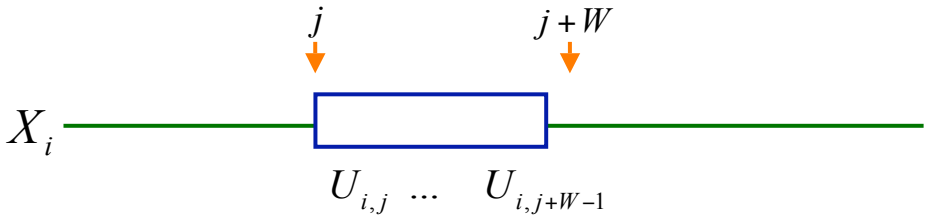


Finding Multiple Motifs

- updating the probability that a motif in position *j* would not overlap *any* previous motif

$$P(V_{i,j} = 1) = \min\left(P(U_{i,j} = 1), \ \ldots \ , P(U_{i,j+W-1} = 1)\right)$$

$$= \min\left(U_{i,j}^{(m)}, \ \ldots \ , U_{i,j+W-1}^{(m)}\right)$$



Starting Points in MEME

- EM is susceptible to local maxima
- for every distinct subsequence of length *W* in the training set
  - derive an initial *p* matrix from this subsequence
  - run EM for 1 iteration
- choose motif model (i.e. *p* matrix) with highest likelihood
- run EM to convergence

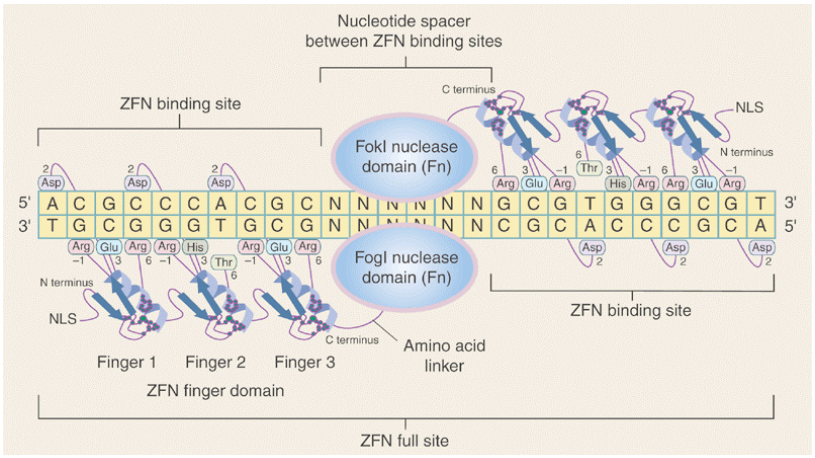
Using Subsequences as Starting Points for EM

- set values corresponding to letters in the subsequence to some value π
- set other values to $(1-\pi)/(M-1)$ where M is the length of the alphabet
- example: for the subsequence TAT with $\pi = 0.5$

$$p = \begin{array}{c|ccc} & 1 & 2 & 3 \\ A & 0.17 & 0.5 & 0.17 \\ C & 0.17 & 0.17 & 0.17 \\ G & 0.17 & 0.17 & 0.17 \\ T & 0.5 & 0.17 & 0.5 \end{array}$$

Using Background Knowledge to Bias the Parameters

- accounting for palindromes that are common in DNA binding sites



- using Dirichlet mixture priors to account for biochemical similarity of amino acids

Representing Palindromes

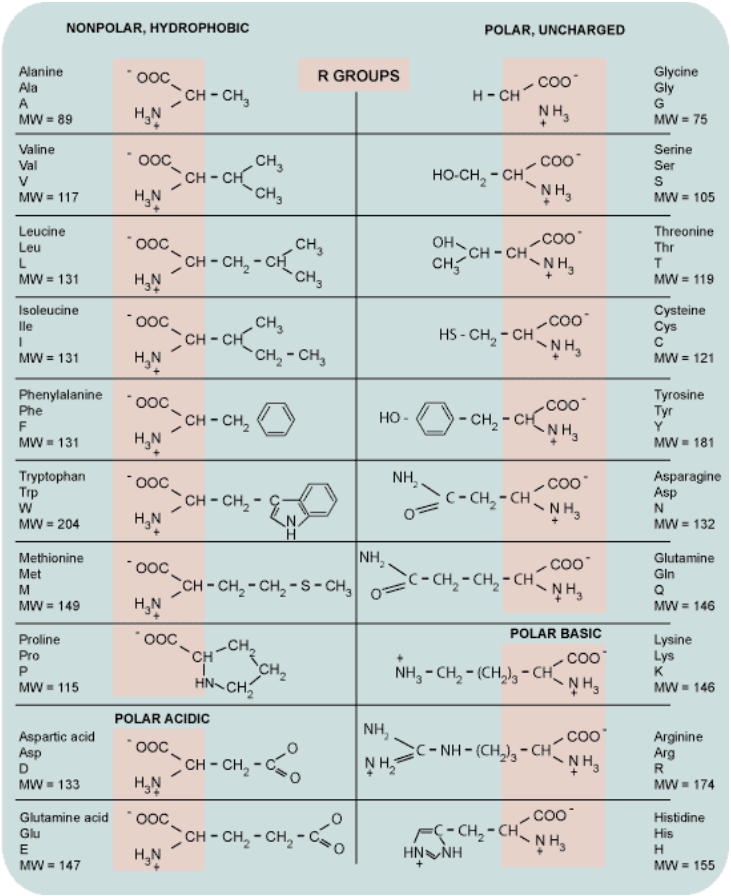
- parameters in probabilistic models can be “tied” or “shared”

$$\begin{bmatrix} p_{a,0} & p_{a,1} & \cdots & p_{a,W} \\ p_{c,0} & p_{c,1} & \cdots & p_{c,W} \\ p_{g,0} & p_{g,1} & \cdots & p_{g,W} \\ p_{t,0} & p_{t,1} & \cdots & p_{t,W} \end{bmatrix}$$

- during motif search, try tying parameters according to palindromic constraint; accept if it increases likelihood test (half as many parameters)

Amino Acids

- Can we encode prior knowledge about amino acid properties into the motif finding process?



Using Dirichlet Mixture Priors

- recall that the M-step updates the parameters by:

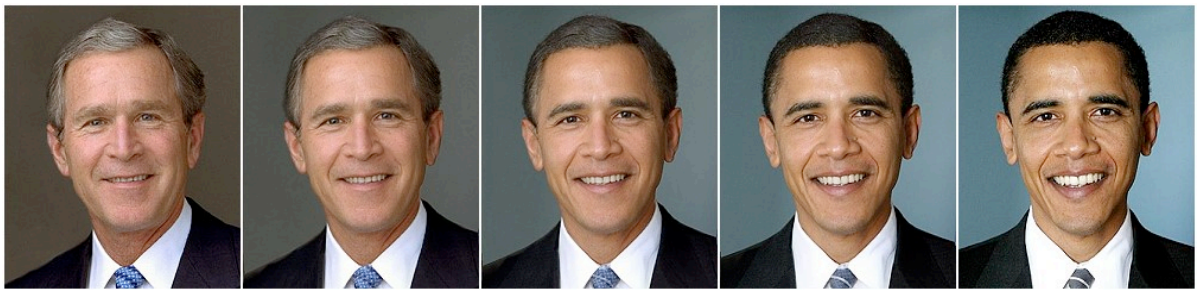
$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

- we can set the pseudocounts using a mixture of Dirichlets:

$$d_{c,k} = \sum_j P(\delta^{(j)} \mid \mathbf{n}_k) \delta_c^{(j)}$$

- where $\delta^{(j)}$ is the j^{th} Dirichlet component

Mixture Example



$\delta^{(\text{Bush})}$ $\quad 0.5 \times \delta^{(\text{Bush})} + 0.5 \times \delta^{(\text{Obama})}$ $\quad \delta^{(\text{Obama})}$

Mixture of Dirichlets

- we'd like to have Dirichlet distributions characterizing amino acids that tend to be used in certain "roles"
- Brown et al. [ISMB '95] induced a set of Dirichlets from trusted protein alignments
  - "large, charged and polar"
  - "polar and mostly negatively charged"
  - "hydrophobic, uncharged, nonpolar"
  - etc.

The Beta Distribution

- suppose we're taking a Bayesian approach to estimating the parameter θ of a weighted coin
- the Beta distribution provides an appropriate prior

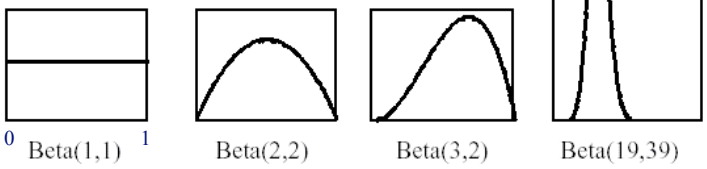
$$P(\theta) = \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

where

α_h # of "imaginary" heads we have seen already

α_t # of "imaginary" tails we have seen already

Γ continuous generalization of factorial function



The Beta Distribution

- suppose now we're given a data set D in which we observe M_h heads and M_t tails

$$P(\theta \mid D) = \frac{\Gamma(\alpha + M_h + M_t)}{\Gamma(\alpha_h + M_h)\Gamma(\alpha_t + M_t)} \theta^{\alpha_h + M_h - 1}(1-\theta)^{\alpha_t + M_t - 1}$$

$$= \text{Beta}(\alpha_h + M_h, \alpha_t + M_t)$$

- the posterior distribution is also Beta: we say that the set of Betas distributions is a *conjugate* family for binomial sampling

The Dirichlet Distribution

- for discrete variables with more than two possible values, we can use *Dirichlet* priors

- Dirichlet priors are a *conjugate* family for multinomial data

$$P(\theta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$$

- if $P(\theta)$ is Dirichlet($\alpha_1, \ldots, \alpha_K$), then $P(\theta|D)$ is Dirichlet($\alpha_1+M_1, \ldots, \alpha_K+M_K$), where M_i is the # occurrences of the i^{th} value

Using Dirichlet Mixture Priors

$$d_{c,k} = \sum_j P(\delta^{(j)} \mid \mathbf{n}_k)\ \delta_c^{(j)}$$

likelihood of "observed" counts under j^{th} Dirichlet

parameter for character c n j^{th} Dirichlet

Gibbs Sampling: An Alternative to EM

- a general procedure for sampling from the joint distribution of a set of random variables $P(U_1 \ldots U_n)$ by iteratively sampling from $P(U_j \mid U_1 \ldots U_{j-1}, U_{j+1} \ldots U_n)$ for each j
- application to motif finding: Lawrence et al. 1993
- can view it as a stochastic analog of EM for this task
- in theory, less susceptible to local minima than EM

Gibbs Sampling Approach

- in the EM approach we maintained a distribution Z_i over the possible motif starting points for each sequence
- in the Gibbs sampling approach, we'll maintain a specific starting point for each sequence a_i but we'll keep randomly resampling these

Gibbs Sampling Approach

given: length parameter W, training set of sequences

choose random positions for a

do

pick a sequence X_i

estimate p given current motif positions a (update step)

(using all sequences but X_i)

sample a new motif position a_i for X_i (sampling step)

until convergence

return: p, a

Sampling New Motif Positions

- for each possible starting position, $a_i = j$, compute a weight

$$A_j = \frac{\prod_{k=j}^{j+W-1} p_{c_k,\, k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k,\, 0}}$$

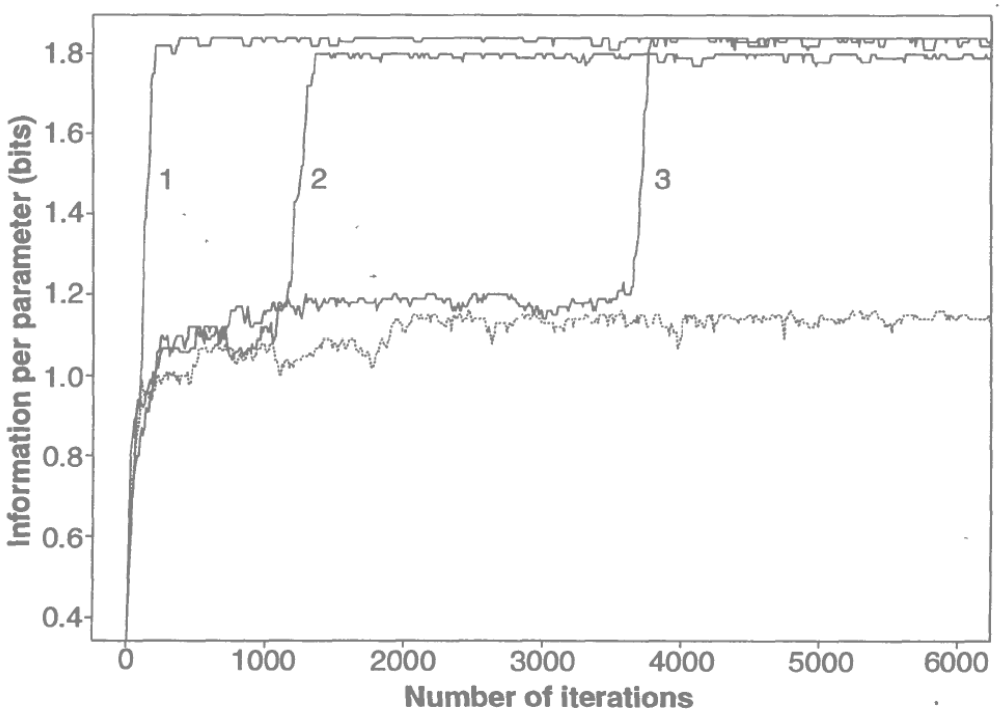
- randomly select a new starting position a_i according to these weights

The Phase Shift Problem

- Gibbs sampler can get stuck in a local maxima that corresponds to the correct solution shifted by a few bases

- Solution : add a special step to shift the a values by the same amount for all sequences. Try different shift amounts and pick one in proportion to its probability score.

Convergence of Gibbs



Markov Chain Monte Carlo

- method for sampling from some probability distribution
- construct Markov chain with stationary distribution equal to distribution of interest; by sampling can find most probable states
- detailed balance:

$$P(x)\tau(y \mid x) = P(y)\tau(x \mid y)$$

probability of state x

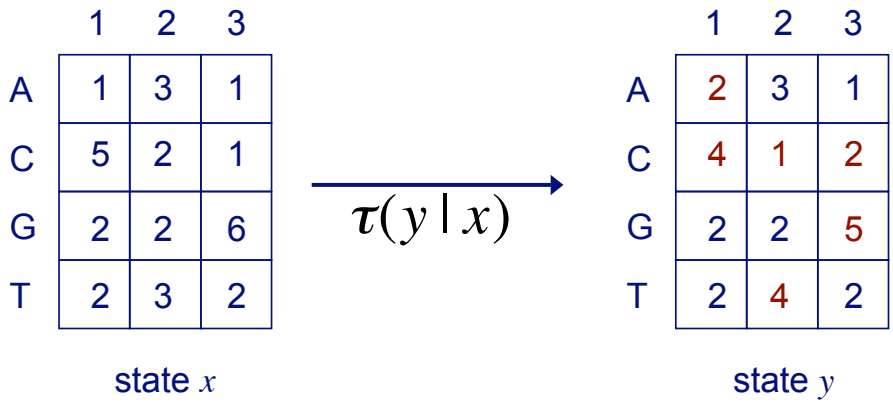
probability of transition $x \rightarrow y$

- when detailed balance holds:

$$\frac{1}{N} \lim_{N \to \infty} count(x) = P(x)$$

Markov Chain Monte Carlo

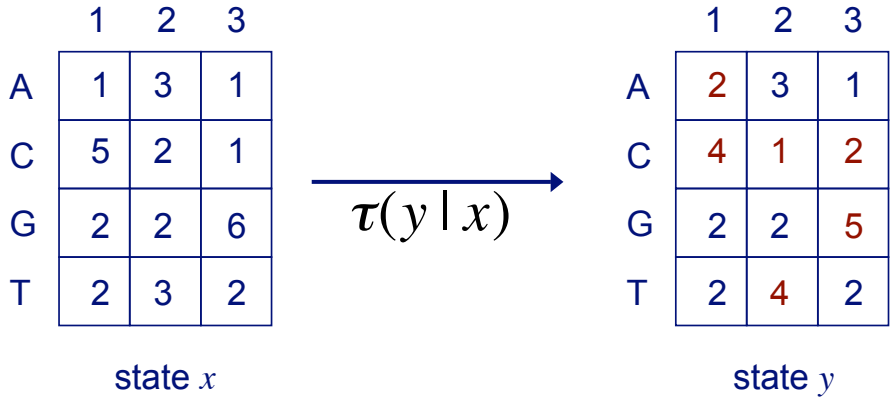
- in our case, a state corresponds to counts of the characters observed in motif occurrences for a given a



Markov Chain Monte Carlo

- the probability of a state is given by

$$P(x) \propto \prod_{c} \prod_{j=1}^{W} \left(\frac{p_{c,j}(x)}{p_{c,0}} \right)^{n_{c,j}(x)}$$



Motif Finding: EM and Gibbs

- these methods compute *local*, *multiple* alignments
- both methods try to optimize the likelihood of the sequences
- EM converges to a local maximum
- Gibbs will converge to a global maximum, *in the limit;* in a reasonable amount of time, probably not
- MEME can take advantage of background knowledge by
  - tying parameters
  - Dirichlet priors

- there are many other methods for motif finding
- in practice, motif finders often fail
  - motif “signal” may be weak
  - large search space, many local minima