

Learning Sequence Motif Models Using Gibbs Sampling

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2015

Colin Dewey

cdewey@biostat.wisc.edu

Goals for Lecture

the key concepts to understand are the following

- Markov Chain Monte Carlo (MCMC) and Gibbs sampling
- Gibbs sampling applied to the motif-finding task
- parameter tying
- incorporating prior knowledge using Dirichlets and Dirichlet mixtures

Gibbs Sampling: An Alternative to EM

- EM can get trapped in local maxima
- one approach to alleviate this limitation: try different (perhaps random) initial parameters
- Gibbs sampling exploits randomized search to a much greater degree
- can view it as a stochastic analog of EM for this task
- in theory, Gibbs sampling is less susceptible to local maxima than EM
- [Lawrence et al., *Science* 1993]

Gibbs Sampling Approach

- in the EM approach we maintained a distribution $Z_i^{(t)}$ over the possible motif starting points for each sequence at iteration t
- 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 Algorithm for Motif Finding

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

(using all sequences but X_i) (predictive update step)

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

until convergence

return: p, a

Markov Chain Monte Carlo (MCMC)

- Consider a Markov chain in which, on each time step, a grasshopper randomly chooses to stay in its current state, jump one state left or jump one state right.

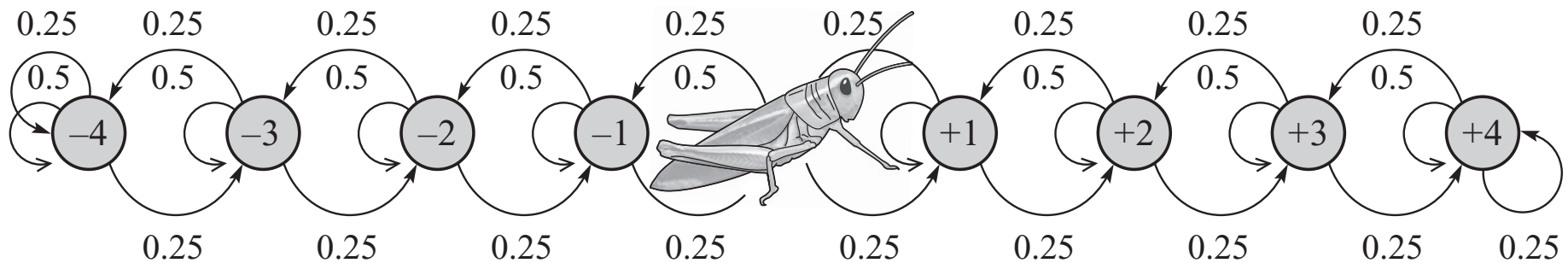


Figure from Koller & Friedman, Probabilistic Graphical Models, MIT Press

- let $P^{(t)}(u)$ represent the probability of being in state u at time t in the random walk

$P^{(0)}(0) = 1$	$P^{(0)}(+1) = 0$	$P^{(0)}(+2) = 0$
$P^{(1)}(0) = 0.5$	$P^{(1)}(+1) = 0.25$	$P^{(1)}(+2) = 0$
$P^{(2)}(0) = 0.375$	$P^{(2)}(+1) = 0.25$	$P^{(2)}(+2) = 0.0625$
\vdots	\vdots	\vdots
$P^{(100)}(0) \approx 0.11$	$P^{(100)}(+1) \approx 0.11$	$P^{(100)}(+2) \approx 0.11$

The Stationary Distribution

- let $P(u)$ represent the probability of being in state u at any given time in a random walk on the chain

$$P^{(t)}(u) \approx P^{(t+1)}(u)$$

$$P^{(t+1)}(u) = \sum_v P^{(t)}(v) \tau(u | v)$$

probability of
state v

probability of
transition $v \rightarrow u$

- the stationary distribution is the set of such probabilities for all states

Markov Chain Monte Carlo (MCMC)

- we can view the motif finding approach in terms of a Markov chain
- each state represents a configuration of the starting positions (a_i values for a set of random variables $A_1 \dots A_n$)
- transitions correspond to changing selected starting positions (and hence moving to a new state)

ACATCCG
CGACTAC
ATTGAGC
CGTTGAC
GAGTGAT
TCGTTGG
ACAGGAT
TAGCTAT
GCTACCG
GGCCTCA

state u

$\xrightarrow{\tau(v | u)}$

ACATCCG
CGACTAC
ATTGAGC
CGTTGAC
GAGTGAT
TCGTTGG
ACAGGAT
TAGCTAT
GCTACCG
GGCCTCA

state v

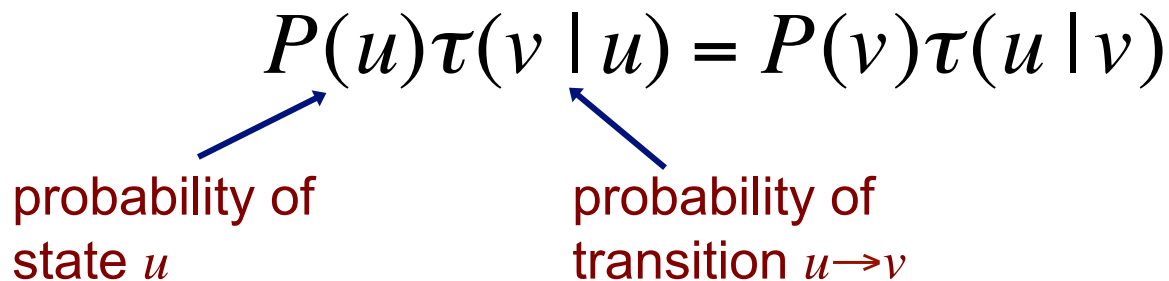
Markov Chain Monte Carlo

- for the motif-finding task, the number of states is enormous
- key idea: construct Markov chain with stationary distribution equal to distribution of interest; use sampling to find most probable states
- detailed balance:

$$P(u)\tau(v \mid u) = P(v)\tau(u \mid v)$$

probability of state u

probability of transition $u \rightarrow v$

The diagram shows the detailed balance equation $P(u)\tau(v \mid u) = P(v)\tau(u \mid v)$. Below the equation, there are two labels in red text. The label "probability of state u " has a blue arrow pointing to the $P(u)$ term in the equation. The label "probability of transition $u \rightarrow v$ " has a blue arrow pointing to the $\tau(v \mid u)$ term in the equation.

- when detailed balance holds:

$$\frac{1}{N} \lim_{N \rightarrow \infty} count(u) = P(u)$$

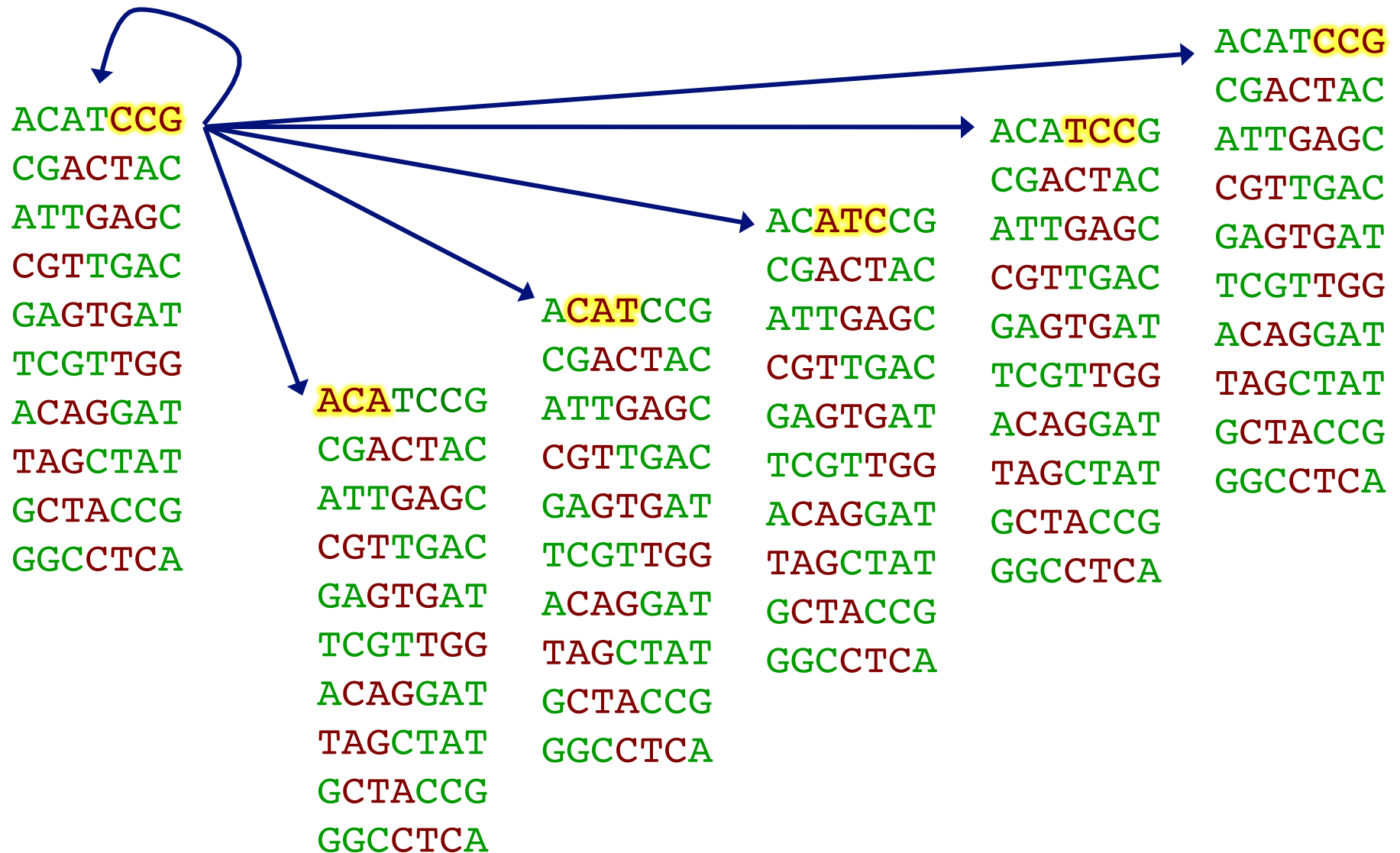
MCMC with Gibbs Sampling

Gibbs sampling is a special case of MCMC in which

- Markov chain transitions involve changing one variable at a time
- transition probability is conditional probability of the changed variable given all others
- i.e. we sample the joint distribution of a set of random variables $P(A_1 \dots A_n)$ by iteratively sampling from $P(A_i \mid A_1 \dots A_{i-1}, A_{i+1} \dots A_n)$

Gibbs Sampling Approach

- possible state transitions when first sequence is selected



Gibbs Sampling Approach

- How do we get the transition probabilities when we don't know what the motif looks like?

Gibbs Sampling Approach

- the probability of a state is given by

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

background probability for character c

probability of c in motif position j

count of c in motif position j

u

ACATCCG
CGACTAC
ATTGAGC
CGTTGAC
GAGTGAT
TCGTTGG
ACAGGAT
TAGCTAT
GCTACCG
GGCCTCA

$n(u)$

	1	2	3
A	1	3	1
C	5	2	1
G	2	2	6
T	2	3	2

Sampling New Motif Positions

- for each possible starting position, $A_i = j$, compute the likelihood ratio (leaving sequence i out of estimates of p)

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

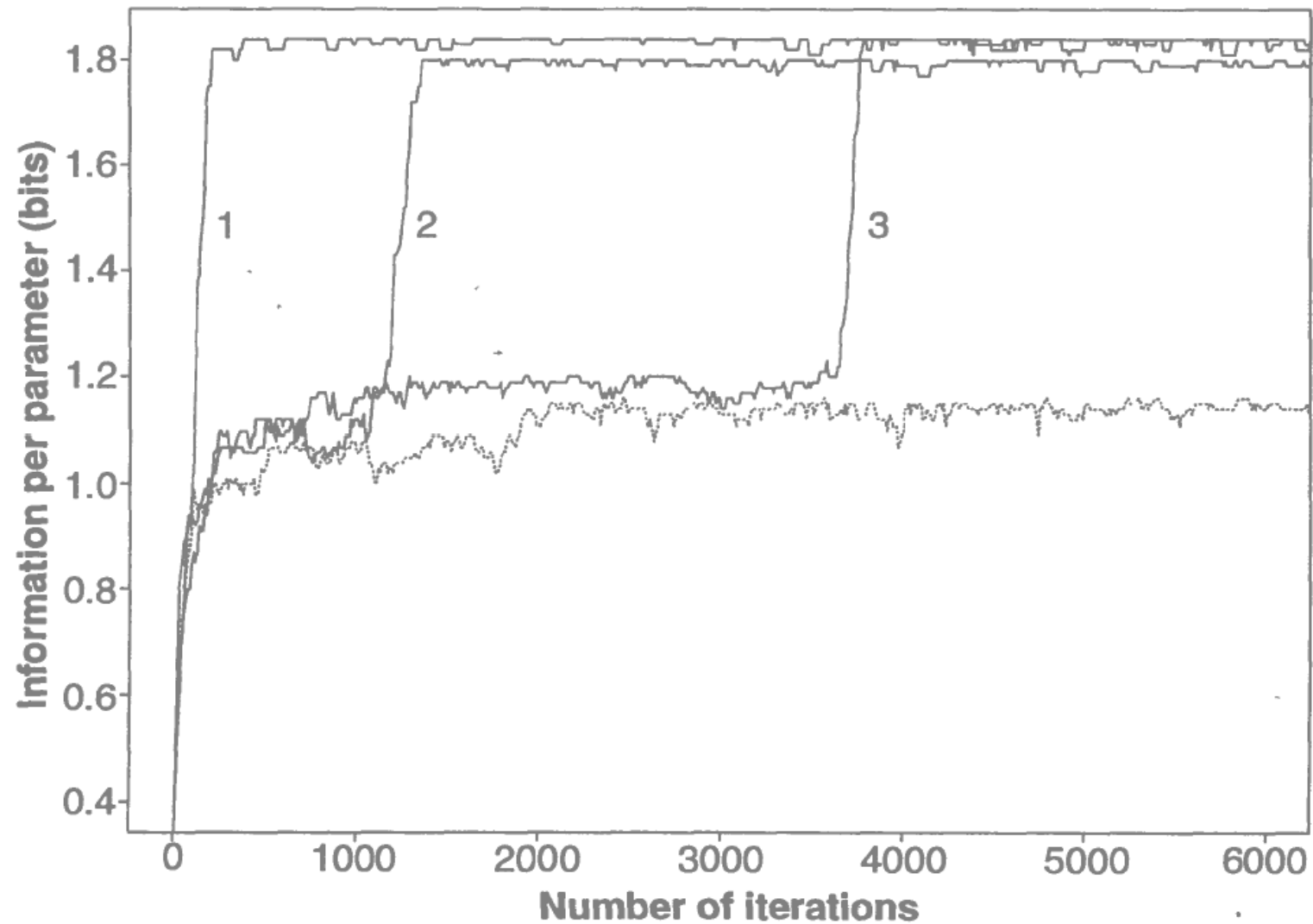
- randomly select a new starting position $A_i = j$ with probability

$$\frac{LR(j)}{\sum_{k \in \{\text{starting positions}\}} LR(k)}$$

The Phase Shift Problem

- Gibbs sampler can get stuck in a local maximum 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



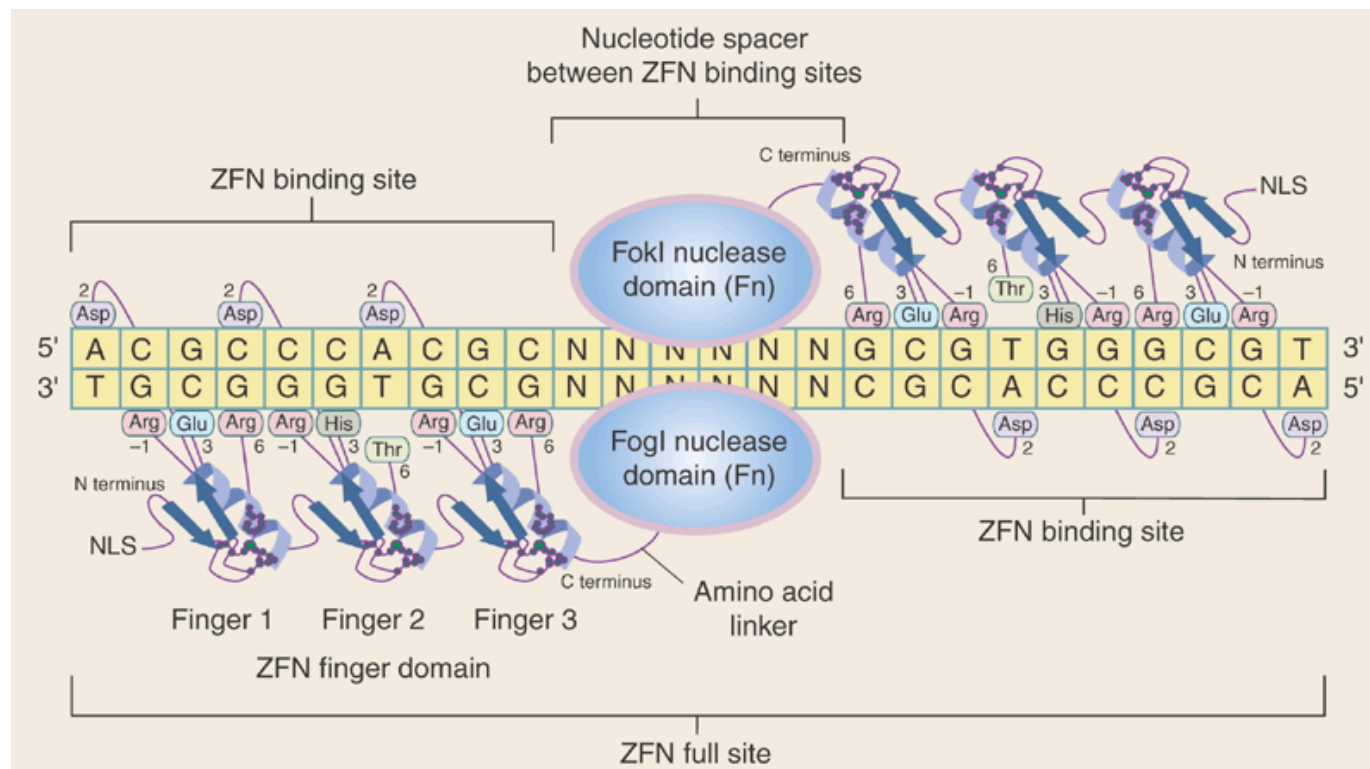
Using Background Knowledge to Bias the Parameters

let's consider two ways in which background knowledge can be exploited in the motif finding process

1. accounting for palindromes that are common in DNA binding sites
2. using Dirichlet mixture priors to account for biochemical similarity of amino acids

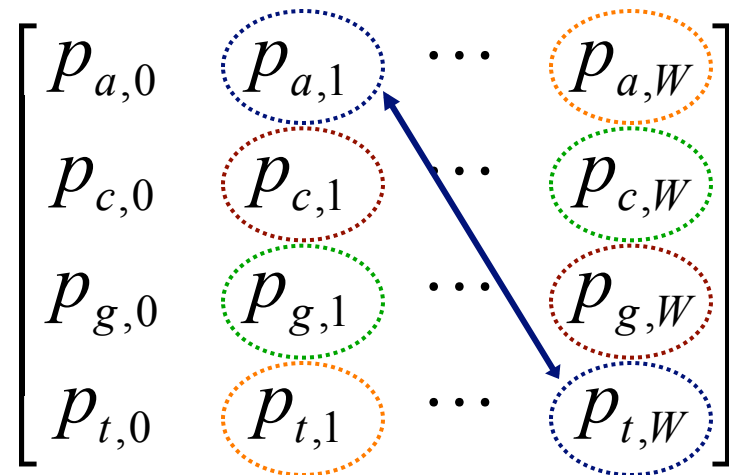
Using Background Knowledge to Bias the Parameters

- Many DNA motifs have a palindromic pattern because they are bound by a protein *homodimer*: a complex consisting of two identical proteins



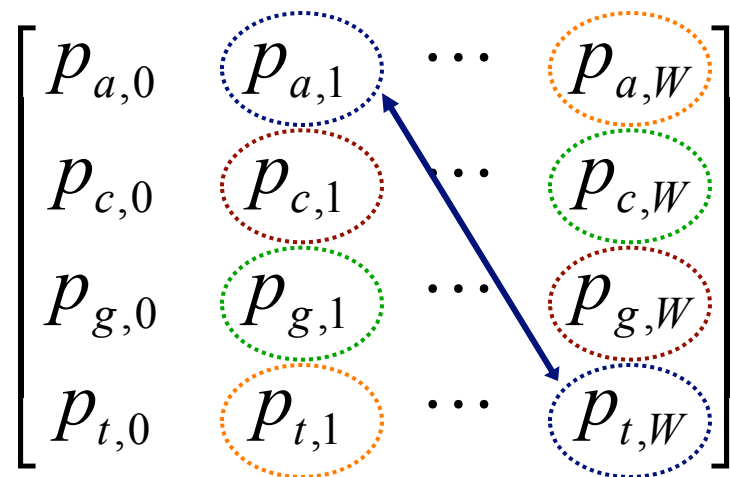
Representing Palindromes

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



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

Updating Tied Parameters



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

The diagram shows a 4x4 matrix of parameters. The first column contains parameters $p_{a,0}, p_{c,0}, p_{g,0}, p_{t,0}$. The second column contains $p_{a,1}, p_{c,1}, p_{g,1}, p_{t,1}$. The third column contains ellipses. The fourth column contains $p_{a,W}, p_{c,W}, p_{g,W}, p_{t,W}$. The parameters $p_{a,1}$ and $p_{t,W}$ are highlighted with blue dotted circles. The parameters $p_{c,1}$ and $p_{c,W}$ are highlighted with red dotted circles. The parameters $p_{g,1}$ and $p_{g,W}$ are highlighted with green dotted circles. The parameters $p_{t,1}$ and $p_{a,W}$ are highlighted with orange dotted circles. A blue arrow points from $p_{a,1}$ to $p_{t,W}$, indicating they are tied parameters.

$$p_{a,1} \equiv p_{t,W} = \frac{n_{a,1} + n_{t,W} + d_{a,1} + d_{t,W}}{\sum_b (n_{b,1} + d_{b,1}) + \sum_b (n_{b,W} + d_{b,W})}$$

Using Dirichlet Mixture Priors

- recall that the EM/Gibbs update the parameters by:

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

- Can we use background knowledge to guide our choice of pseudocounts ($d_{c,k}$)?
- suppose we're modeling protein sequences...

Amino Acids

- Can we encode prior knowledge about amino acid properties into the motif finding process?
- there are classes of amino acids that share similar properties

NONPOLAR, HYDROPHOBIC		POLAR, UNCHARGED	
		R GROUPS	
Alanine Ala A MW = 89	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_3 \end{array}$		Glycine Gly G MW = 75
Valine Val V MW = 117	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3)_2 \end{array}$		Serine Ser S MW = 105
Leucine Leu L MW = 131	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}(\text{CH}_3)_2 \end{array}$		Threonine Thr T MW = 119
Isoleucine Ile I MW = 131	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3) - \text{CH}_2 - \text{CH}_3 \end{array}$		Cysteine Cys C MW = 121
Phenylalanine Phe F MW = 131	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_6\text{H}_5 \end{array}$		Tyrosine Tyr Y MW = 181
Tryptophan Trp W MW = 204	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_8\text{H}_6\text{N}_2 \end{array}$		Asparagine Asp N MW = 132
Methionine Met M MW = 149	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{S} - \text{CH}_3 \end{array}$		Glutamine Gln Q MW = 146
Proline Pro P MW = 115	$\begin{array}{c} \text{OOC}^- \\ \\ \text{CH} - \text{CH}_2 - \text{CH}_2 \\ \quad \quad \\ \text{HN} \quad \quad \text{CH}_2 \end{array}$		Lysine Lys K MW = 146
POLAR ACIDIC		POLAR BASIC	
Aspartic acid Asp D MW = 133	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{array}$		Arginine Arg R MW = 174
Glutamic acid Glu E MW = 147	$\begin{array}{c} \text{OOC}^- \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{array}$		Histidine His H MW = 155

Using Dirichlet Mixture Priors

- since we're estimating multinomial distributions (frequencies of amino acids at each motif position), a natural way to encode prior knowledge is using Dirichlet distributions
- let's consider
 - the Beta distribution
 - the Dirichlet distribution
 - mixtures of Dirichlets

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



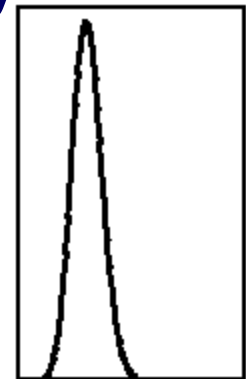
0 Beta(1,1) 1



Beta(2,2)



Beta(3,2)



Beta(19,39)

The Beta Distribution

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

$$P(\theta | D) = \frac{\Gamma(\alpha + D_h + D_t)}{\Gamma(\alpha_h + D_h)\Gamma(\alpha_t + D_t)} \theta^{\alpha_h + D_h - 1} (1 - \theta)^{\alpha_t + D_t - 1}$$
$$= \text{Beta}(\alpha_h + D_h, \alpha_t + D_t)$$

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

The Dirichlet Distribution

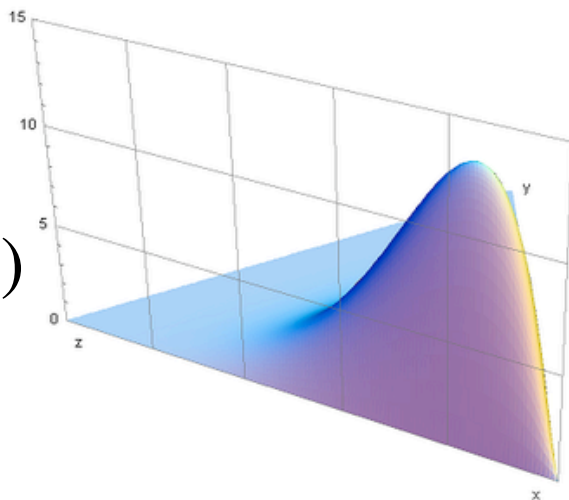
- for discrete variables with more than two possible values, we can use *Dirichlet* priors
- Dirichlet priors are a *conjugate* family for multinomial data

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

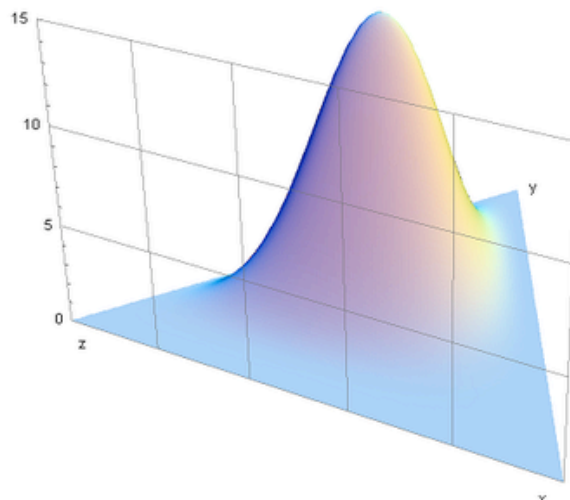
- if $P(\theta)$ is $\text{Dirichlet}(\alpha_1, \dots, \alpha_K)$, then $P(\theta|D)$ is $\text{Dirichlet}(\alpha_1 + D_1, \dots, \alpha_K + D_K)$, where D_i is the # occurrences of the i^{th} value

Dirichlet Distributions

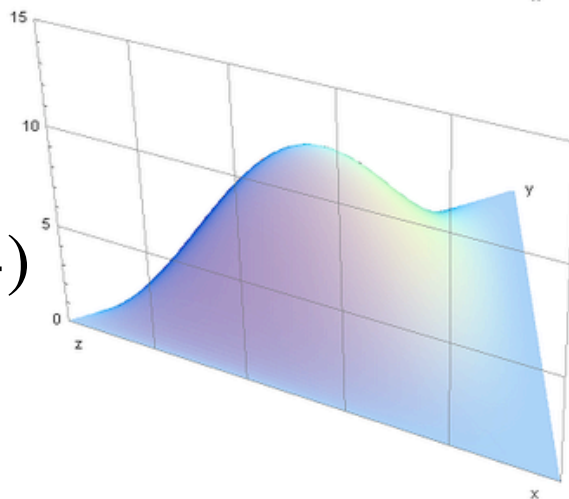
$$\alpha = (6, 2, 2)$$



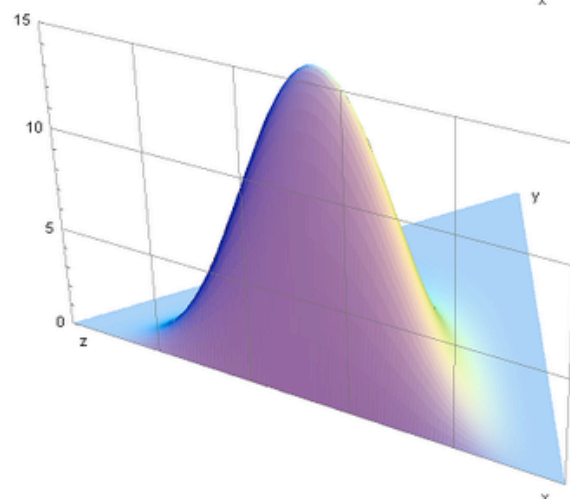
$$\alpha = (3, 7, 5)$$



$$\alpha = (2, 3, 4)$$



$$\alpha = (6, 2, 6)$$



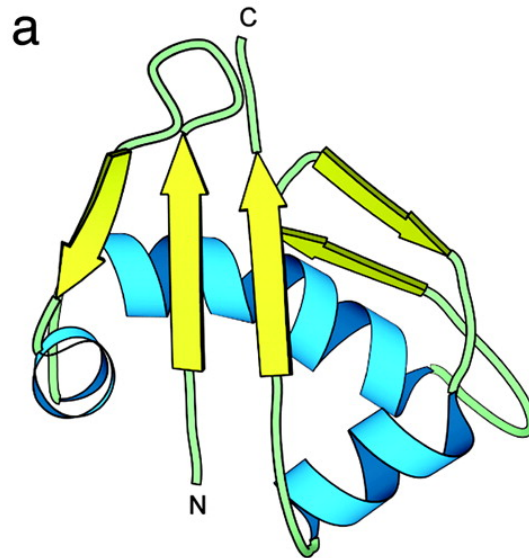
probability density (shown on a simplex) of Dirichlet distributions for $K=3$ and various parameter vectors α

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 '93*] induced a set of Dirichlets from “trusted” protein alignments
 - “large, charged and polar”
 - “polar and mostly negatively charged”
 - “hydrophobic, uncharged, nonpolar”
 - etc.

Trusted Protein Alignments

- a trusted protein alignment is one in which known protein structures are used to determine which parts of the given set of sequences should be aligned



C

(a)	2580558	Hs	886	HL	SLIVRFPNQGRQVDELDIWSHTND	TIGSV	RRCIVNRIKA-N	927
	6678523	Mm	885	HL	SFIVRFPNQGRQVDDLEVWSHTND	TIGSV	RRCILNRIKA-N	926
	22507351	Mm	885	HL	SFTVRFPNQGRQVDELDIWSHTND	TIGSV	RRCILNRMNV-N	926
	31235452	Ag	835	Q	VELIVKFQTPGRQLDDIELLSHSNET	TMH	SFKRNLLRRIKVLK	877
	24651755	Dm	979	NT	ILYIRFQNPGRSIDDMEIVTHSNET	TMA	AFKRNLLKRIKGTS	1021

Using Dirichlet *Mixture* Priors

- recall that the EM/Gibbs update the parameters by:

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

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

$$d_{c,k} = \sum_j P(\alpha^{(j)} | \mathbf{n}_k) \alpha_c^{(j)}$$

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

Using Dirichlet Mixture Priors

$$d_{c,k} = \sum_j P(\alpha^{(j)} | \mathbf{n}_k) \alpha_c^{(j)}$$

probability of j^{th} Dirichlet
given observed counts

parameter for character c
in j^{th} Dirichlet

- we don't have to know which Dirichlet to pick
- instead, we'll hedge our bets, using the observed counts to decide how much to weight each Dirichlet

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
- 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