

Markov Chain Models

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

February 2002

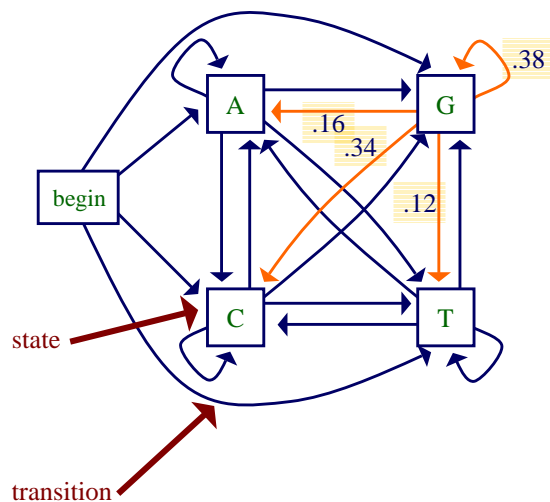
Announcements

- no office hours tomorrow
- interest in basic probability tutorial? (Wed, Thurs evening)
- HW #1 out; due March 11
 - 3 free late days for semester
 - homeworks docked 10 percentage points/day after late days used
- next reading: Salzberg et al., Microbial Gene Identification Using Interpolated Markov Models
- “Biomodule” class: *Introduction to GCG Computing and Sequence Analysis in Unix and Xwindows Environments*
 - taught by Ann Palmenberg and Jean-Yves Sgro
 - April 16 and 17
 - see <http://www.virology.wisc.edu/acp/> for more details

Topics for the Next Few Weeks

- Markov chain models (1st order, higher order and inhomogenous models; parameter estimation; classification)
- interpolated Markov models (and back-off models)
- Expectation Maximization (EM) methods (applications to motif finding)
- Gibbs sampling methods (applications to motif finding)
- hidden Markov models (forward, backward and Baum-Welch algorithms; model topologies; applications to gene finding and protein family modeling)

Markov Chain Models



Markov Chain Models

- a Markov chain model is defined by
 - a set of states
 - some states *emit* symbols
 - other states (e.g. the *begin* state) are *silent*
 - a set of transitions with associated probabilities
 - the transitions emanating from a given state define a distribution over the possible next states

Markov Chain Models

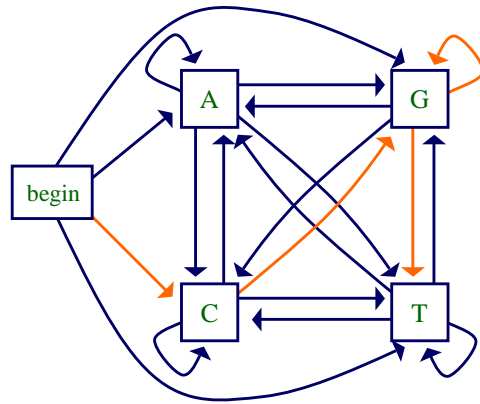
- given some sequence x of length L , we can ask how probable the sequence is given our model
- for any probabilistic model of sequences, we can write this probability as

$$\begin{aligned}\Pr(x) &= \Pr(x_L, x_{L-1}, \dots, x_1) \\ &= \Pr(x_L | x_{L-1}, \dots, x_1) \Pr(x_{L-1} | x_{L-2}, \dots, x_1) \dots \Pr(x_1)\end{aligned}$$

- key property of a (1st order) Markov chain: the probability of each x_i depends only on the value of x_{i-1}

$$\begin{aligned}\Pr(x) &= \Pr(x_L | x_{L-1}) \Pr(x_{L-1} | x_{L-2}) \dots \Pr(x_2 | x_1) \Pr(x_1) \\ &= \Pr(x_1) \prod_{i=2}^L \Pr(x_i | x_{i-1})\end{aligned}$$

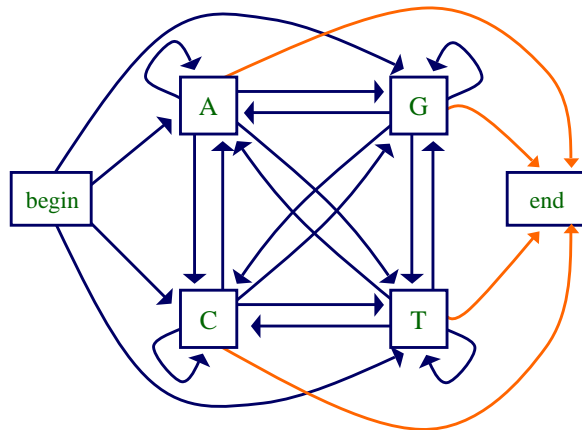
Markov Chain Models



$$\Pr(cggt) = \Pr(c) \Pr(g | c) \Pr(g | g) \Pr(t/g)$$

Markov Chain Models

- can also have an *end* state; allows the model to represent
 - a distribution over sequences of different lengths
 - preferences for ending sequences with certain symbols



Markov Chain Notation

- the transition parameters can be denoted by $a_{x_{i-1}x_i}$ where

$$a_{x_{i-1}x_i} = \Pr(x_i | x_{i-1})$$

- similarly we can denote the probability of a sequence x as

$$a_{\text{B}x_i} \prod_{i=2}^L a_{x_{i-1}x_i} = \Pr(x_1) \prod_{i=2}^L \Pr(x_i | x_{i-1})$$

where $a_{\text{B}x_i}$ represents the transition from the *begin* state

Example Application

- CpG islands
 - CG dinucleotides are rarer in eukaryotic genomes than expected given the independent probabilities of C, G
 - but the regions upstream of genes are richer in CG dinucleotides than elsewhere – *CpG islands*
 - useful evidence for finding genes
- could predict CpG islands with Markov chains
 - one to represent CpG islands
 - one to represent the rest of the genome

Estimating the Model Parameters

- given some data (e.g. a set of sequences from CpG islands), how can we determine the probability parameters of our model?
- one approach: *maximum likelihood estimation*
 - given a set of data D
 - set the parameters θ to maximize

$$\Pr(D | \theta)$$

- i.e. make the data D look likely under the model

Maximum Likelihood Estimation

- suppose we want to estimate the parameters $\Pr(a)$, $\Pr(c)$, $\Pr(g)$, $\Pr(t)$
- and we're given the sequences

accgcgctta

gcttagtgac

tagccgttac

- then the maximum likelihood estimates are

$$\Pr(a) = \frac{6}{30} = 0.2$$

$$\Pr(g) = \frac{7}{30} = 0.233$$

$$\Pr(c) = \frac{9}{30} = 0.3$$

$$\Pr(t) = \frac{8}{30} = 0.267$$

Maximum Likelihood Estimation

- suppose instead we saw the following sequences

gccgcgcttg

gcttggtggc

tggccgttgc

- then the maximum likelihood estimates are

$$\Pr(a) = \frac{0}{30} = 0$$

$$\Pr(g) = \frac{13}{30} = 0.433$$

$$\Pr(c) = \frac{9}{30} = 0.3$$

$$\Pr(t) = \frac{8}{30} = 0.267$$

do we really want to set this to 0?

A Bayesian Approach

- instead of estimating parameters strictly from the data, we could start with some prior belief for each
- for example, we could use *Laplace estimates*

$$\Pr(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}$$

pseudocount

- using Laplace estimates with the sequences

gccgcgcttg

gcttggtggc

tggccgttgc

$$\Pr(a) = \frac{0+1}{34}$$

$$\Pr(c) = \frac{9+1}{34}$$

A Bayesian Approach

- a more general form: *m-estimates*

$$\Pr(a) = \frac{n_a + p_a m}{\left(\sum_i n_i \right) + m}$$

prior probability of *a*

number of “virtual” instances

- with $m=8$ and uniform priors

gccgcgcttg

gcttggtggc

tggccgttgc

$$\Pr(c) = \frac{9 + 0.25 \times 8}{30 + 8} = \frac{11}{38}$$

Markov Chains for Discrimination

- suppose we want to distinguish CpG islands from other sequence regions
- given sequences from CpG islands, and sequences from other regions, we can construct
 - a model to represent CpG islands
 - a *null model* to represent the other regions
- can then score a test sequence by:

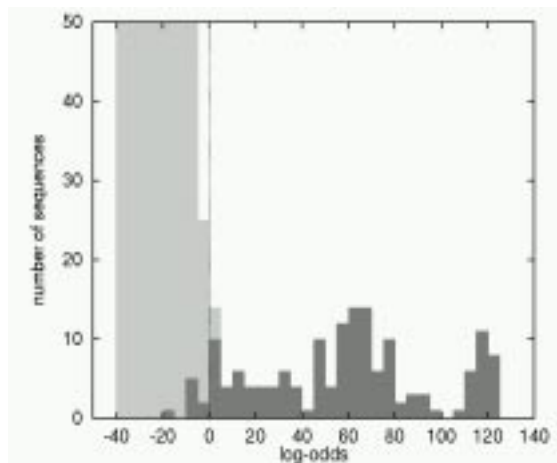
$$\text{score}(x) = \log \frac{\Pr(x \mid \text{CpG model})}{\Pr(x \mid \text{null model})}$$

Markov Chains for Discrimination

- parameters estimated for CpG and null models

+	A	C	G	T	-	A	C	G	T
A	.18	.27	.43	.12	A	.30	.21	.28	.21
C	.17	.37	.27	.19	C	.32	.30	.08	.30
G	.16	.34	.38	.12	G	.25	.24	.30	.21
T	.08	.36	.38	.18	T	.18	.24	.29	.29

Markov Chains for Discrimination



- light bars represent negative sequences
- dark bars represent positive sequences
- the actual figure here is not from a CpG island discrimination task, however

Figure from A. Krogh, "An Introduction to Hidden Markov Models for Biological Sequences" in Computational Methods in Molecular Biology, Salzberg et al. editors, 1998.

Markov Chains for Discrimination

- why use

$$score(x) = \log \frac{\Pr(x | CpG)}{\Pr(x | null)}$$

- Bayes' rule tells us

$$\Pr(CpG | x) = \frac{\Pr(x | CpG) \Pr(CpG)}{\Pr(x)}$$

$$\Pr(null | x) = \frac{\Pr(x | null) \Pr(null)}{\Pr(x)}$$

- if we're not taking into account priors, then just need to compare $\Pr(x | CpG)$ and $\Pr(x | null)$

Higher Order Markov Chains

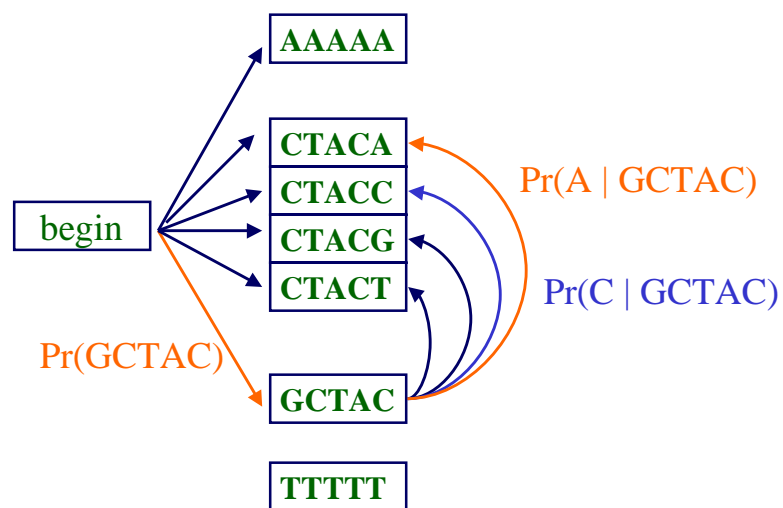
- the Markov property specifies that the probability of a state depends only on the probability of the previous state
- but we can build more “memory” into our states by using a higher order Markov model
- in an n th order Markov model

$$\Pr(x_i | x_{i-1}, x_{i-2}, \dots, x_1) = \Pr(x_i | x_{i-1}, \dots, x_{i-n})$$

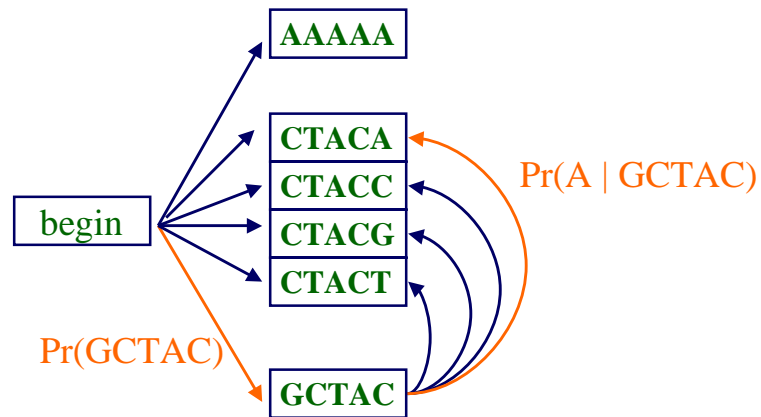
Higher Order Markov Chains

- an n th order Markov chain over some alphabet A is equivalent to a first order Markov chain over the alphabet of n -tuples A^n
- example: a 2nd order Markov model for DNA can be treated as a 1st order Markov model over alphabet
AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT,
TA, TC, TG, TT

A Fifth Order Markov Chain



A Fifth Order Markov Chain



$$\Pr(\text{gctaca}) = \Pr(\text{gctac}) \Pr(\text{a} \mid \text{gctac})$$

Inhomogenous Markov Chains

- in the Markov chain models we have considered so far, the probabilities do not depend on where we are in a given sequence
- in an *inhomogeneous* Markov model, we can have different distributions at different positions in the sequence
- consider modeling codons in protein coding regions

Inhomogeneous Markov Chains

