

# Markov Chain Models (Part 2)

BMI/CS 576

[www.biostat.wisc.edu/bmi576/](http://www.biostat.wisc.edu/bmi576/)

Mark Craven

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

Fall 2011

## 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

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

## Selecting the order of a Markov chain model

- higher order models remember more “history”
- additional history can have predictive value
- example:
  - predict the next word in this sentence fragment  
“... the\_\_” (duck, end, grain, tide, wall, ...?)
  - now predict it given more history  
“... against the \_\_” (duck, end, grain, tide, wall, ...?)
  - “swim against the \_\_” (duck, end, grain, tide, wall, ...?)

## Selecting the order of a Markov chain model

- but the number of parameters we need to estimate grows exponentially with the order
  - for modeling DNA we need  $O(4^{n+1})$  parameters for an  $n$ th order model
- the higher the order, the less reliable we can expect our parameter estimates to be
  - estimating the parameters of a 2<sup>nd</sup> order Markov chain from the complete genome of *E. Coli*, we’d see each word > 72,000 times on average
  - estimating the parameters of an 8<sup>th</sup> order chain, we’d see each word ~ 5 times on average

## Higher order Markov chains

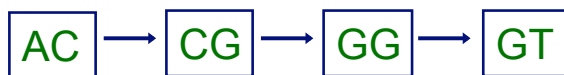
- an  $n$ th order Markov chain over some alphabet  $A$  is equivalent to a first order Markov chain over the alphabet  $A^n$  of  $n$ -tuples

- example: a 2<sup>nd</sup> order Markov model for DNA can be treated as a 1<sup>st</sup> order Markov model over alphabet

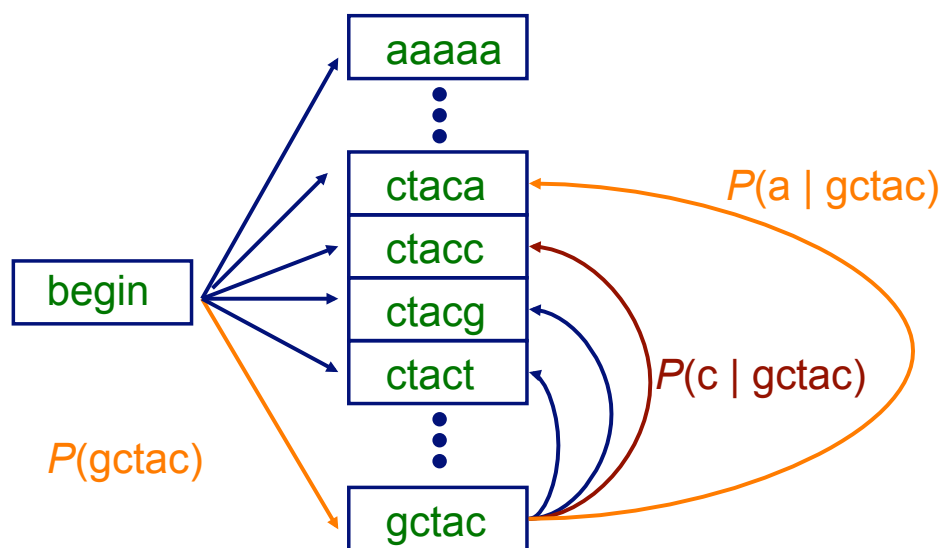
AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT

- caveat: we process a sequence one character at a time

A C G G T



## A fifth-order Markov chain

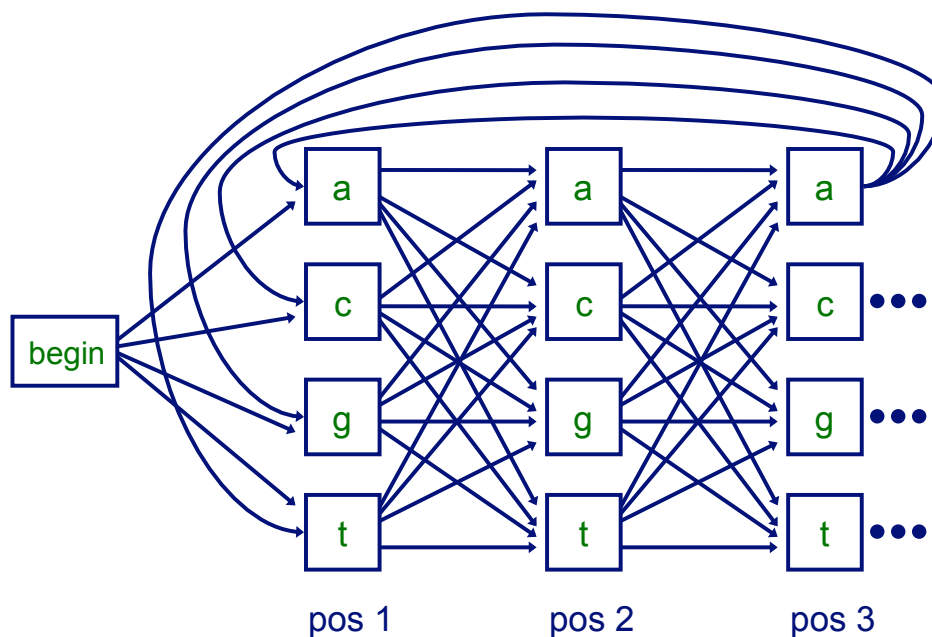


$$P(gctaca) = P(gctac)P(a | gctac)$$

# Inhomogenous Markov chains

- in the Markov chain models we have considered so far, the probabilities do not depend on our position 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

## An inhomogeneous Markov chain

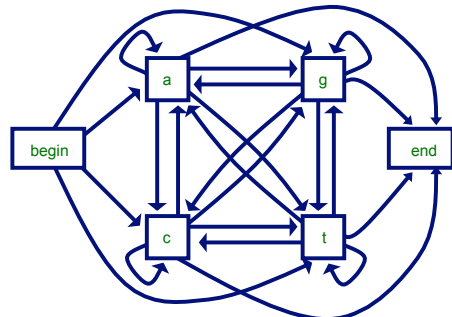
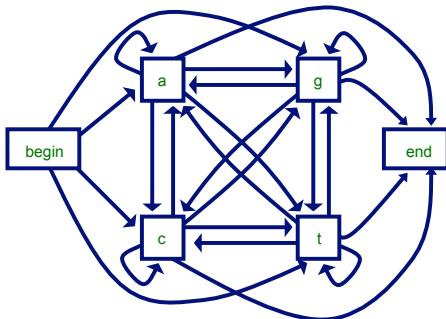


## Example application

- CpG islands
  - CG dinucleotides are rarer in eukaryotic genomes than expected given the marginal probabilities of C and 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

## CpG islands as a classification task

1. train two Markov models: one to represent CpG island sequence regions, another to represent other sequence regions (*null*)



2. given a test sequence, use two models to
  - determine probability that sequence is a CpG island
  - classify the sequence (*CpG* or *null*)

## Markov chains for discrimination

- parameters estimated for CpG and null models
  - human sequences containing 48 CpG islands
  - 60,000 nucleotides

		$P(c   a)$			
+		<i>a</i>	<i>c</i>	<i>g</i>	<i>t</i>
<i>a</i>	.18	.27	.43	.12	
<i>c</i>	.17	.37	.27	.19	
<i>g</i>	.16	.34	.38	.12	
<i>t</i>	.08	.36	.38	.18	

CpG

		<i>a</i>	<i>c</i>	<i>g</i>	<i>t</i>
-	<i>a</i>	.30	.21	.28	.21
<i>c</i>	.32	.30	.08	.30	
<i>g</i>	.25	.24	.30	.21	
<i>t</i>	.18	.24	.29	.29	

null

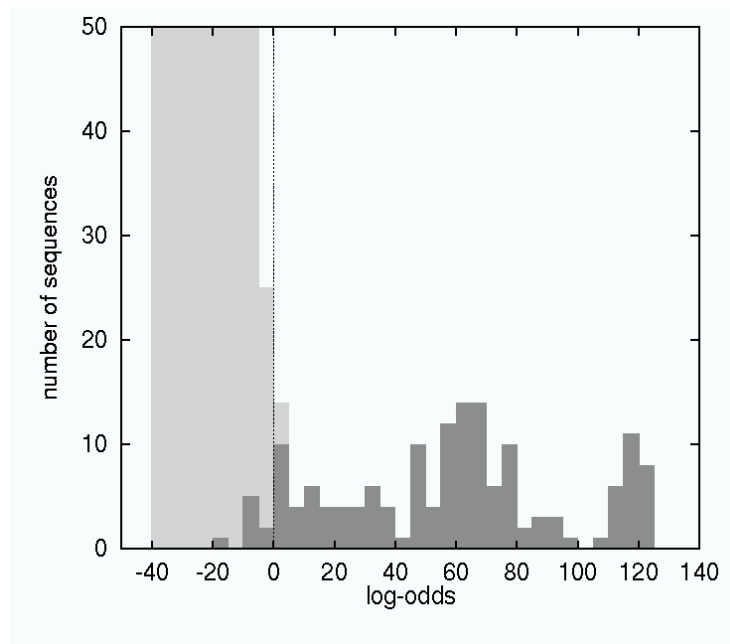
## Markov chains for discrimination

- using Bayes' rule tells us

$$\begin{aligned}
 P(\text{CpG} | x) &= \frac{P(x | \text{CpG})P(\text{CpG})}{P(x)} \\
 &= \frac{P(x | \text{CpG})P(\text{CpG})}{P(x | \text{CpG})P(\text{CpG}) + P(x | \text{null})P(\text{null})}
 \end{aligned}$$

- if we don't take into account prior probabilities of two classes ( $P(\text{CpG})$  and  $P(\text{null})$ ) then we just need to compare  $P(x | \text{CpG})$  and  $P(x | \text{null})$

# Markov chains for discrimination



- light bars represent negative sequences
- dark bars represent positive sequences (i.e. CpG islands)
- 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.