# Genome Annotation

## Introduction to Markov Chains

# Outline

- Motivation for Markov models
- Markov chain definition
- Example application to CpG islands

# Motivation for Markov Models in Computational Biology

- there are many cases in which we would like to represent the statistical regularities of some class of sequences
  - genes
  - various regulatory sites in DNA (e.g., where RNA polymerase and transcription factors bind)
  - proteins in a given family
- Markov models are well suited to this type of task
  - They allow for the modeling of *dependencies* between nearby positions

# Markov models have wide applications

- Genome annotation
  - Given a genome sequence find functional units of the genome
    - Genes, CpG islands, promoters..
- Sequence classification
  - represent a family of proteins or DNA/RNA sequences
- Sequence alignment
- Time series analysis
  - e.g., analysis of gene expression over time

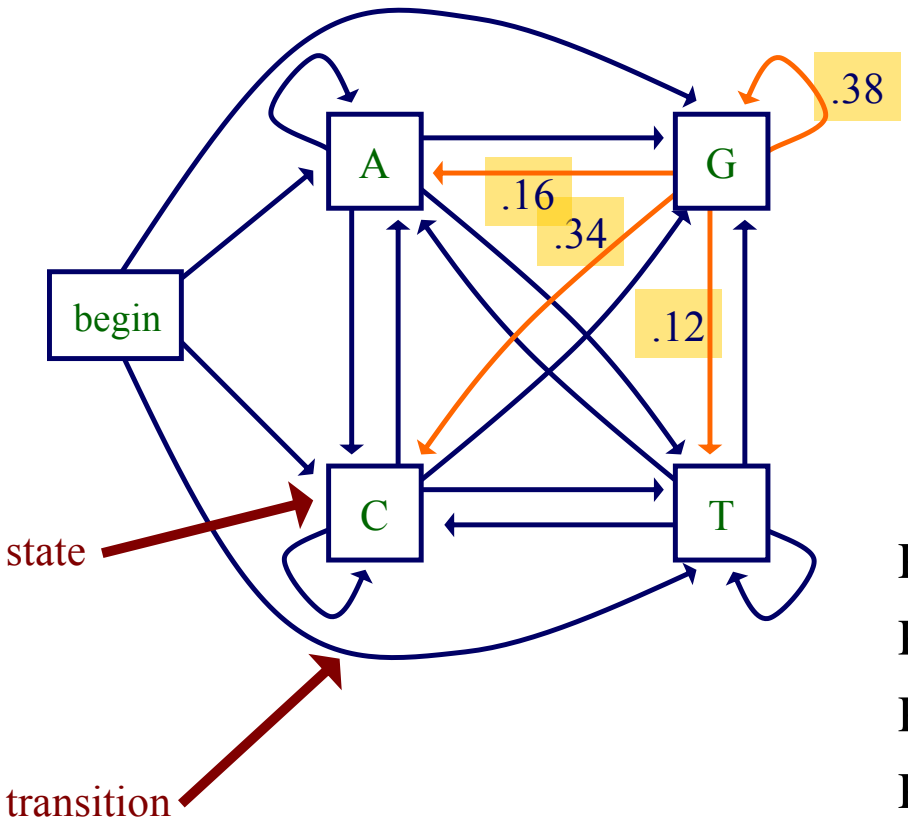
# Applications outside of molecular biology

- Any sort of time series or one-dimensional positional data
- Important in natural language processing
  - speech recognition
  - parsing and understanding of written text

# Markov Chain Models

- a Markov chain model is defined by
  - a set of **states**
    - we'll think of a traversal through the states as *generating* a sequence
    - each state adds to the sequence (except for *begin* and *end*)
  - a set of **transitions** with associated probabilities
    - the transitions emanating from a given state define a distribution over the possible next states

# A Markov Chain Model for DNA




transition probabilities

$$\Pr(X_i = a \mid X_{i-1} = g) = 0.16$$

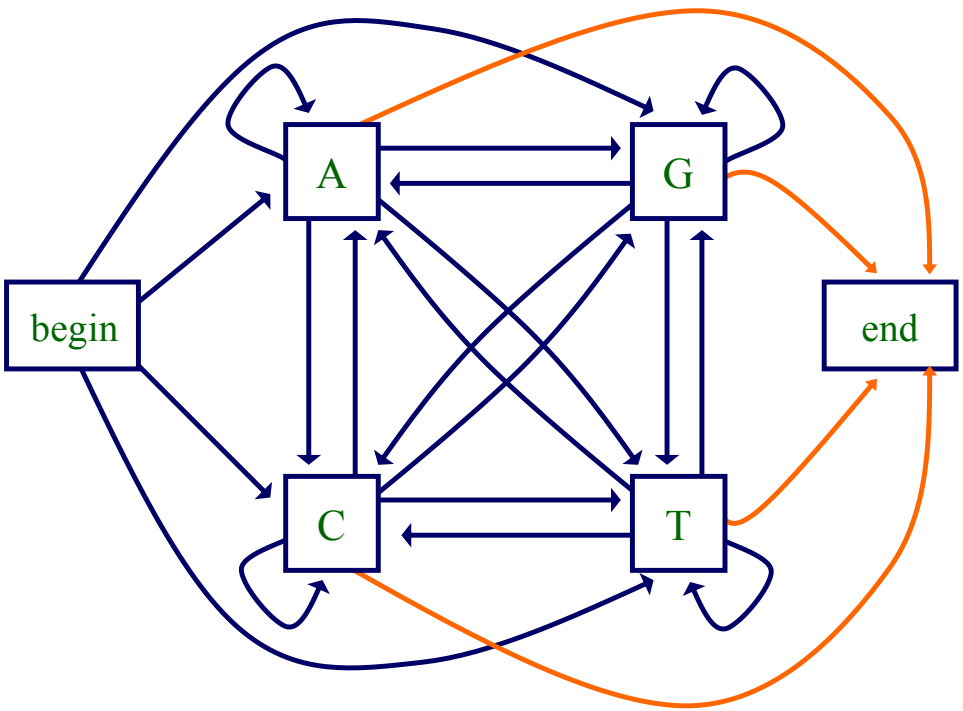
$$\Pr(X_i = c \mid X_{i-1} = g) = 0.34$$

$$\Pr(X_i = g \mid X_{i-1} = g) = 0.38$$

$$\Pr(X_i = t \mid X_{i-1} = g) = 0.12$$

# 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 Models

- Let X be a sequence of *L* random variables $X_1 \ldots X_L$ representing a biological sequence generated through some process
- We can ask how probable the sequence is given our model
- for any probabilistic model of sequences, we can write this probability as (recall the "Chain Rule of Probability")

$$\Pr(X) = \Pr(X_L, X_{L-1}, \ldots, X_1)$$

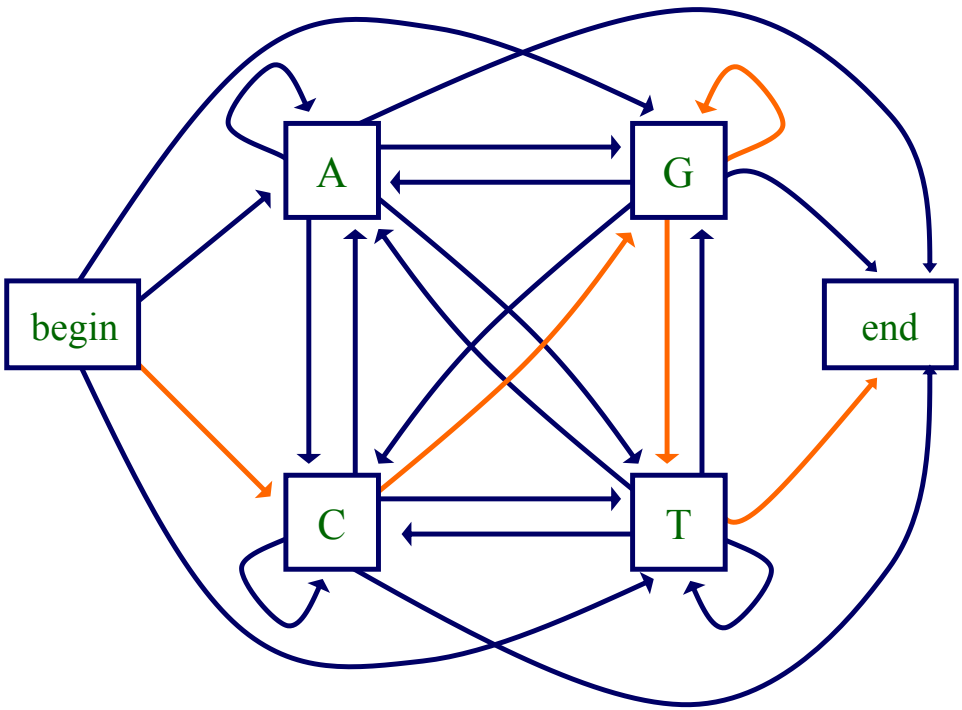
$$= \Pr(X_L / X_{L-1}, \ldots, X_1) \Pr(X_{L-1} \mid X_{L-2}, \ldots, X_1) \ldots \Pr(X_1)$$

- key property of a (1st order) Markov chain: $X_i$ is conditionally independent of $X_1, X_2, \ldots, X_{i-2}$ given $X_{i-1}$

$$\Pr(X) = \Pr(X_L / X_{L-1}) \Pr(X_{L-1} \mid X_{L-2}) \ldots \Pr(X_2 \mid X_1) \Pr(X_1)$$

$$= \Pr(X_1) \prod_{i=2}^{L} \Pr(X_i \mid X_{i-1})$$

# The Probability of a Sequence for a Given Markov Chain Model




$$\Pr(cggt) = \Pr(c \mid begin)\Pr(g \mid c)\Pr(g \mid g)\Pr(t \mid g)\Pr(\text{end} \mid t)$$

# Markov Chain Notation

- the transition parameters will be denoted by $a_{s,t}$ where

$$a_{s,t} = \Pr(X_i = t | X_{i-1} = s)$$

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

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

where $a_{\mathrm{B}x_1}$ represents the transition from the *begin* state

- This gives a probability distribution over sequences of length $L$

# 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

# 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:

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

# Markov Chains for Discrimination

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

$\Pr(c \mid a)$

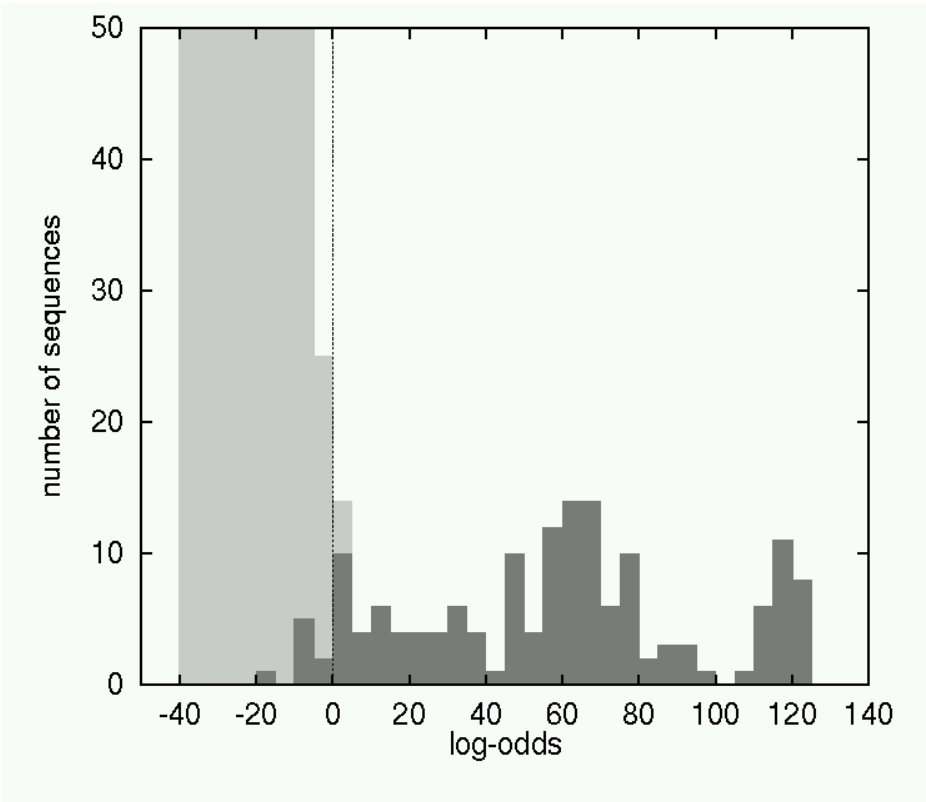
| + | *a* | *c* | *g* | *t* |
|---|---|---|---|---|
| *a* | .18 | .27 | .43 | .12 |
| *c* | .17 | .37 | .27 | .19 |
| *g* | .16 | .34 | .38 | .12 |
| *t* | .08 | .36 | .38 | .18 |

CpG

| - | *a* | *c* | *g* | *t* |
|---|---|---|---|---|
| *a* | .30 | .21 | .28 | .21 |
| *c* | .32 | .30 | .08 | .30 |
| *g* | .25 | .24 | .30 | .21 |
| *t* | .18 | .24 | .29 | .29 |

null

# 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

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

- Bayes' rule tells us

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

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

- if we're not taking into account prior probabilities of two classes ($\Pr(CpG)$ and $\Pr(null)$) then we just need to compare $\Pr(x \mid CpG)$ and $\Pr(x \mid null)$

# Summary

- Markov chains are natural models for sequence-like data
- Markov chains are defined by two components
  - A set of states
  - Transition probabilities between states
- A Markov chain defines a probability distribution over sequences
- Markov chains can be used to discriminate between classes of sequences