Markov chain models

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on Mark Craven's class at University of Wisconsin

http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Overview

- Motivation for statistical models in computational biology
  - to represent the statistical regularities of some class of sequences,
  - the sequences could be genes, various regulatory sites in DNA (e.g. promoters), proteins in a given family,
- Markov models
  - Markov property
    - given the present, the future does not depend on the past,
  - trade-off between simplicity and veracity,
  - Markov chains
    - the model states are observable,
    - one-to-one link between the states and the sequence symbols,
  - hidden Markov models.
    - the relationship between states and symbols remains hidden.

Motivation for sequence modeling

these sequences are E. coli promoters

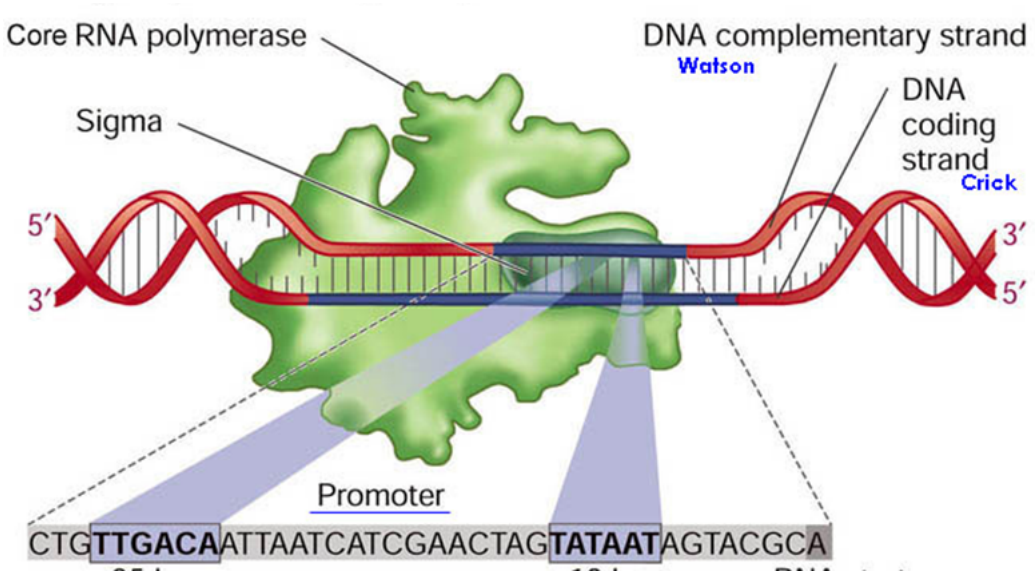
```
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca
accggaagaaaaccgtgacattttaacacgtttgttacaaggtaaaggcgacgccgc
aaattaaaatttttattgacttaggtcactaaatactttaaccaatataggcatagcg
ttgtcataatcgacttgtaaaccaaattgaaaagatttaggtttacaagtctacacc
catcctcgcaccagtcgacgacggtttacgctttacgtatagtggcgacaatttttt
tccagtataatttgttggcataattaagtacgacgagtaaaattacatacctgcccg
acagttatccactattcctgtggataaccatgtgtattagagttagaaaacacgagg
```

these sequences are not promoters

```
atagtctcagagtcttgacctactacgccagcattttggcggtgtaagctaaccatt
aactcaaggctgatacggcgagacttgcgagccttgtccttgcggtacacagcagcg
ttactgtgaacattattcgtctccgcgactacgatgagatgcctgagtgcttccgtt
tattctcaacaagattaaccgacagattcaatctcgtggatggacgttcaacattga
aacgagtcaatcagaccgctttgactctggtattactgtgaacattattcgtctccg
aagtgcttagcttcaaggtcacggatacgaccgaagcgagcctcgtcctcaatggcc
gaagaccacgcctcgccaccgagtagacccttagagagcatgtcagcctcgacaact
```

How can we tell the difference? Is this sequence a promoter?

```
ccatcaaaaaaatattctcaacataaaaaactttgtgtaatacttgtaacgctacat
```

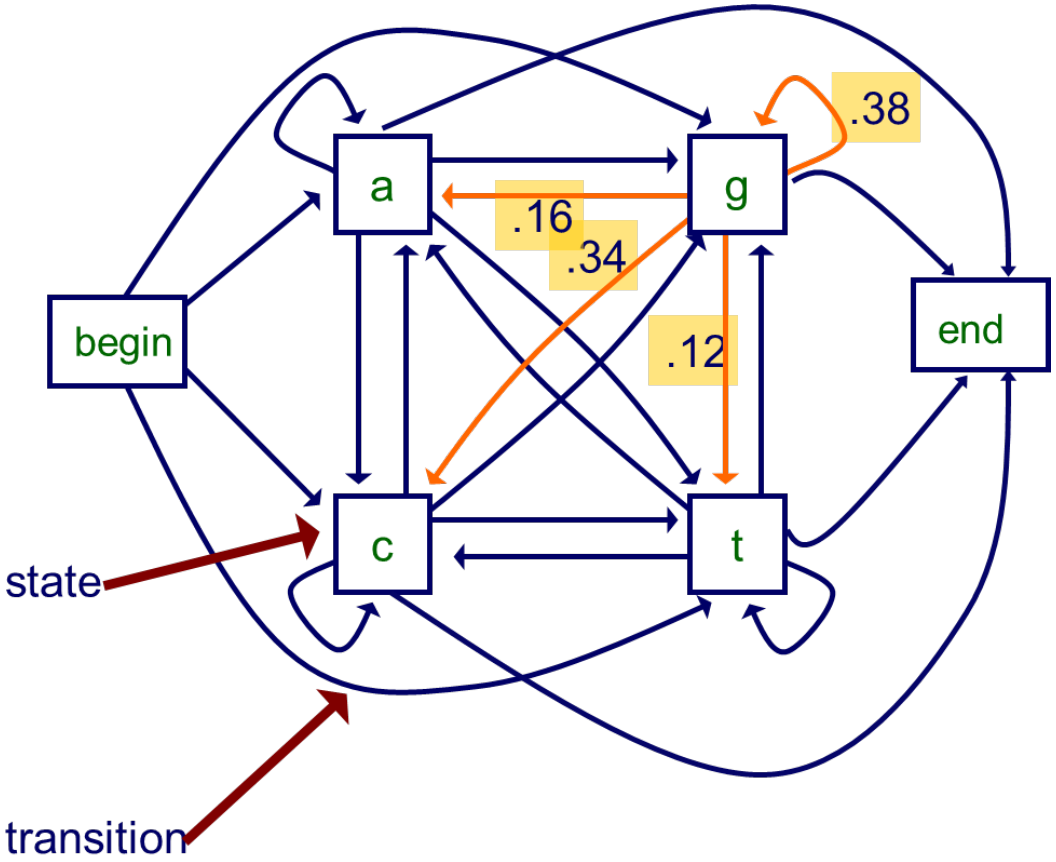


http://helicase.pbworks.com/

Markov chain models

- a Markov chain model is defined by
  - a set of states
    - some states **emit** symbols,
    - other states (e.g., the begin and end states) are **silent**,
    - in our case, the silent states allow the model to represent
      - preferences for beginning and ending sequences with certain symbols,
      - a distribution over sequences of different lengths,
  - 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



the set of states:

$S = \{begin, end, a, c, g, t\}$

the transition probabilities:

$P(x_i = a|x_{i-1} = g) = 0.16$

$P(x_i = c|x_{i-1} = g) = 0.34$

$P(x_i = g|x_{i-1} = g) = 0.38$

$P(x_i = t|x_{i-1} = g) = 0.12$

...

Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Markov property

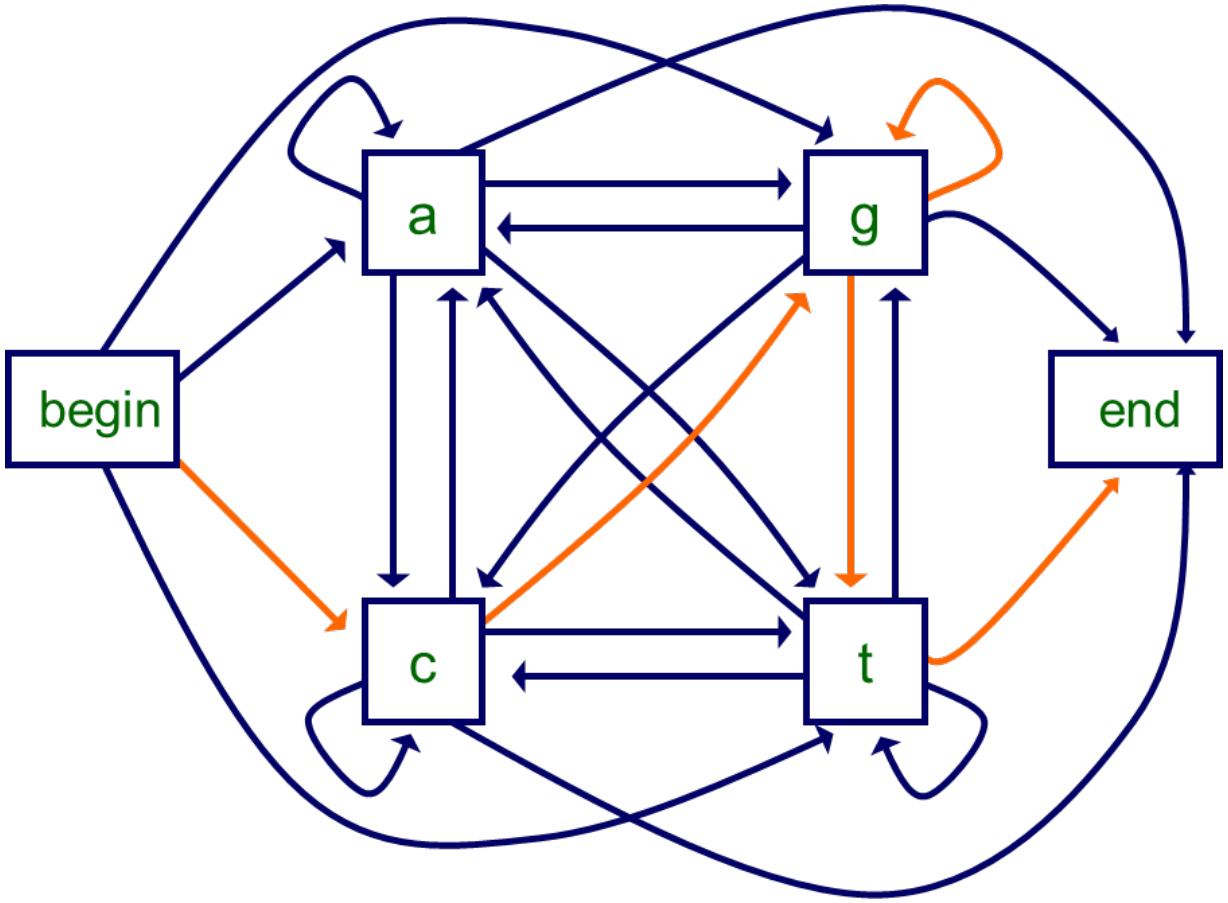
- Let X be a sequence of random variables $X_1 \dots X_L$ representing a biological sequence,
- from the chain rule of probability

$$\begin{aligned} P(X) &= P(X_L, X_{L-1}, \dots, X_1) = \\ &= P(X_L|X_{L-1}, \dots, X_1)P(X_{L-1}|X_{L-2}, \dots, X_1) \dots P(X_1) \end{aligned}$$

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

$$\begin{aligned} P(X) &= P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \dots P(X_2|X_1)P(X_1) = \\ &= P(X_1)\prod_{i=2}^{L} P(X_i|X_{i-1}) \end{aligned}$$

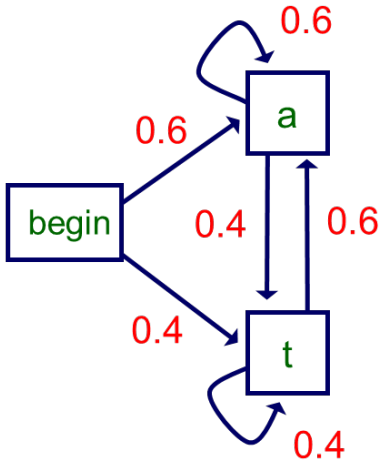
The probability of a sequence for a given Markov chain

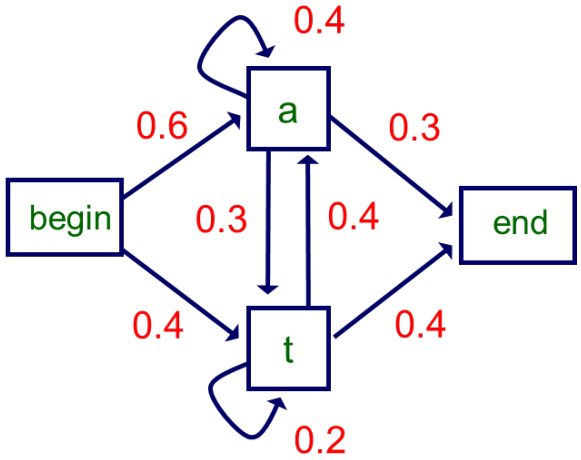


$$P(cggt) = P(c)P(g|c)P(g|g)P(t|g)P(end|t)$$

The role of the end state

- The end state defines a distribution over varying sequence lengths.





P(A) = 0.6	P(AA) = 0.36	P(A) = 0.18	P(AA) = 0.072
P(T) = 0.4	P(AT) = 0.24	P(T) = 0.18	P(AT) = 0.072
	P(TA) = 0.24		P(TA) = 0.048
	P(TT) = 0.16		P(TT) = 0.032
P(L=1) = 1	P(L=2) = 1	P(L=1) = 0.36	P(L=2) = 0.224

Estimating the model parameters

- Given some data, how can we determine the probability parameters of our model?
- one approach: **maximum likelihood estimation** (MLE)
  - given a set of data D,
  - set the parameters θ to maximize $P(D|\theta)$,
  - i.e. make the data D look as likely as possible under the model,
- suppose that we are given the following set of DNA sequences

  $D = \{$**accgcgctta**,**gcttagtgac**,**tagccgttac**$\}$
  - what parameters do we have to find?
  - how can we compute them?
  - is MLE the best approach?

Maximum likelihood estimation

- We have to estimate transition probabilities
  - initial probabilities: $P(a)$, $P(c)$, $P(g)$, $P(t)$,
  - 16 1st order probabilities: $P(a|a)$, $P(a|c)$, $\ldots$, $P(t|t)$,
- MLE implemented via relative frequencies

$$P(x) = \frac{n_x}{\sum_{i \in \{a,c,g,t\}} n_i} \quad \text{where } n_x \text{ is frequency of } x$$

$$P(a) = \frac{6}{30} = 0.2, \quad P(c) = \frac{9}{30} = 0.3, \quad P(g) = \frac{7}{30} = 0.233, \quad P(t) = \frac{8}{30} = 0.267$$

$$P(x|y) = \frac{n_{yx}}{\sum_{i \in \{a,c,g,t\}} n_{yi}} \quad \text{where } n_{yx} \text{ is frequency of the subsequence } yx$$

$$P(a|g) = \frac{1}{7}, \quad P(c|g) = \frac{4}{7}, \quad P(t|g) = \frac{2}{7}, \quad P(g|g) = \frac{0}{7}$$

- do we really want to have zero probabilities?

A Bayesian approach

- Start with some prior belief for each parameter
  - instead of estimating parameters strictly from the data,
  - maximize posterior probability instead of the likelihood

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- **Laplace estimates** represent the way of smoothing for discrete variables

$$P(x) = \frac{n_x + 1}{\sum_{i \in \{a,c,g,t\}}(n_i + 1)}$$ where 1 is a pseudocount

- **m-estimates** represent its more general form

$$P(x) = \frac{n_x + p_x m}{\sum_{i \in \{a,c,g,t\}}(n_i) + m}$$

where m is the number of virtual instances and p_x is a prior probability of x.

A Bayesian approach

- Remember the data: $D = \{$**accgcgctta**,**gcttagtgac**,**tagccgttac**$\}$,
- regularize $P(a|g)$ by Laplace estimate

$$P(x|y) = \frac{n_{yx} + 1}{\sum_{i \in \{a,c,g,t\}} (n_{yi} + 1)}$$

$$P(g|g) = \frac{0 + 1}{7 + 4} = 0.091$$

- regularize $P(a|g)$ by m-estimate with $m = 8$ and uniform priors

$$P(x|y) = \frac{n_{yx} + p_x m}{\sum_{i \in \{a,c,g,t\}} (n_{yi}) + m}$$

$$P(g|g) = \frac{0 + 0.25 \times 8}{7 + 8} = 0.133$$

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 nth order Markov model

$$P(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) = P(X_i|X_{i-1}, \ldots, X_{i-n})$$

- 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, ...?)

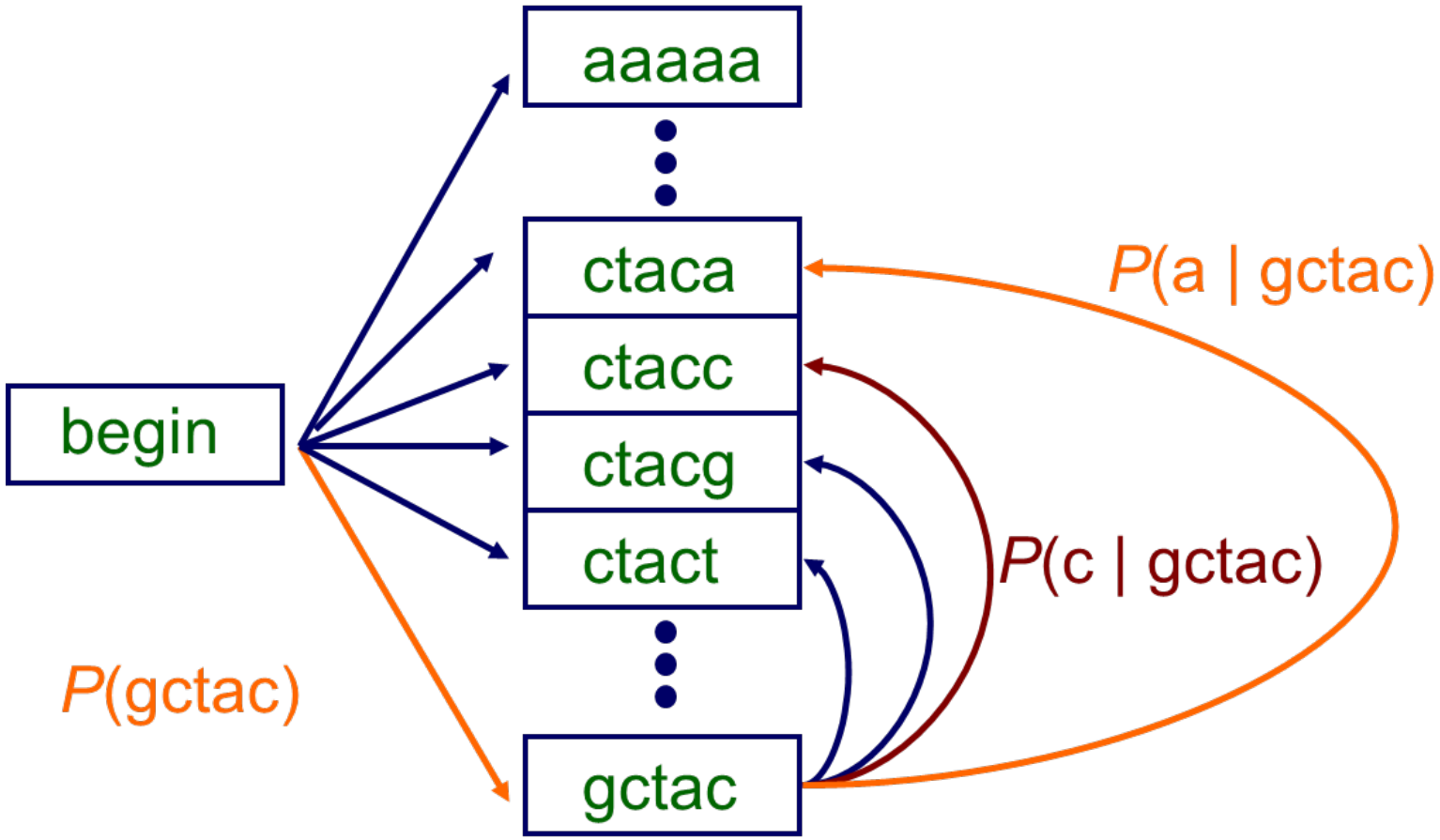
Selecting the order of a Markov chain model

- The order of a Markov chain is a trade-off between simplicity and veracity,
- the number of parameters grows **exponentially** with the order
  - for modeling DNA we need $\mathcal{O}(4^{n+1})$ parameters for an nth order model,
- the higher the order, the less reliable the parameter estimates
  - estimating the parameters of a 2nd 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 8th order chain, we'd see each word $\approx$ 5 times on average.

Higher order Markov chains

- an nth order Markov chain over some alphabet Σ is equivalent to a first order Markov chain over the alphabet Σ^n of n-tuples,
- 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
- caveat: we process a sequence one character at a time
  - a sequence **A C G G T** processed as **A C** $\rightarrow$ **C G** $\rightarrow$ **G G** $\rightarrow$ **G T**,

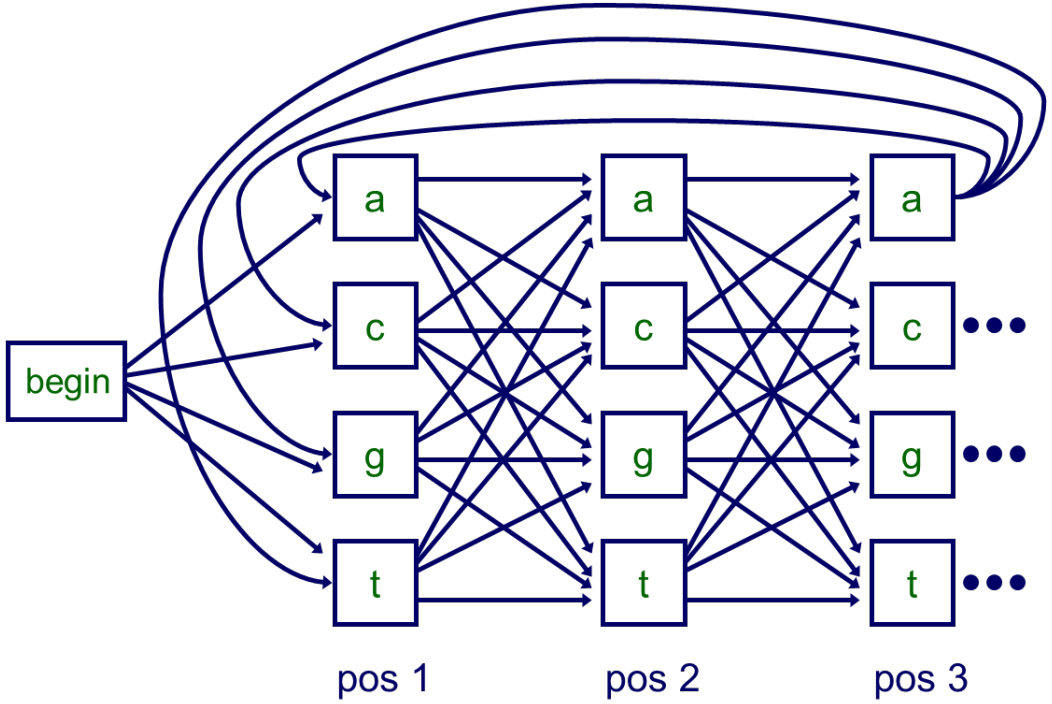
A fifth-order Markov chain



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

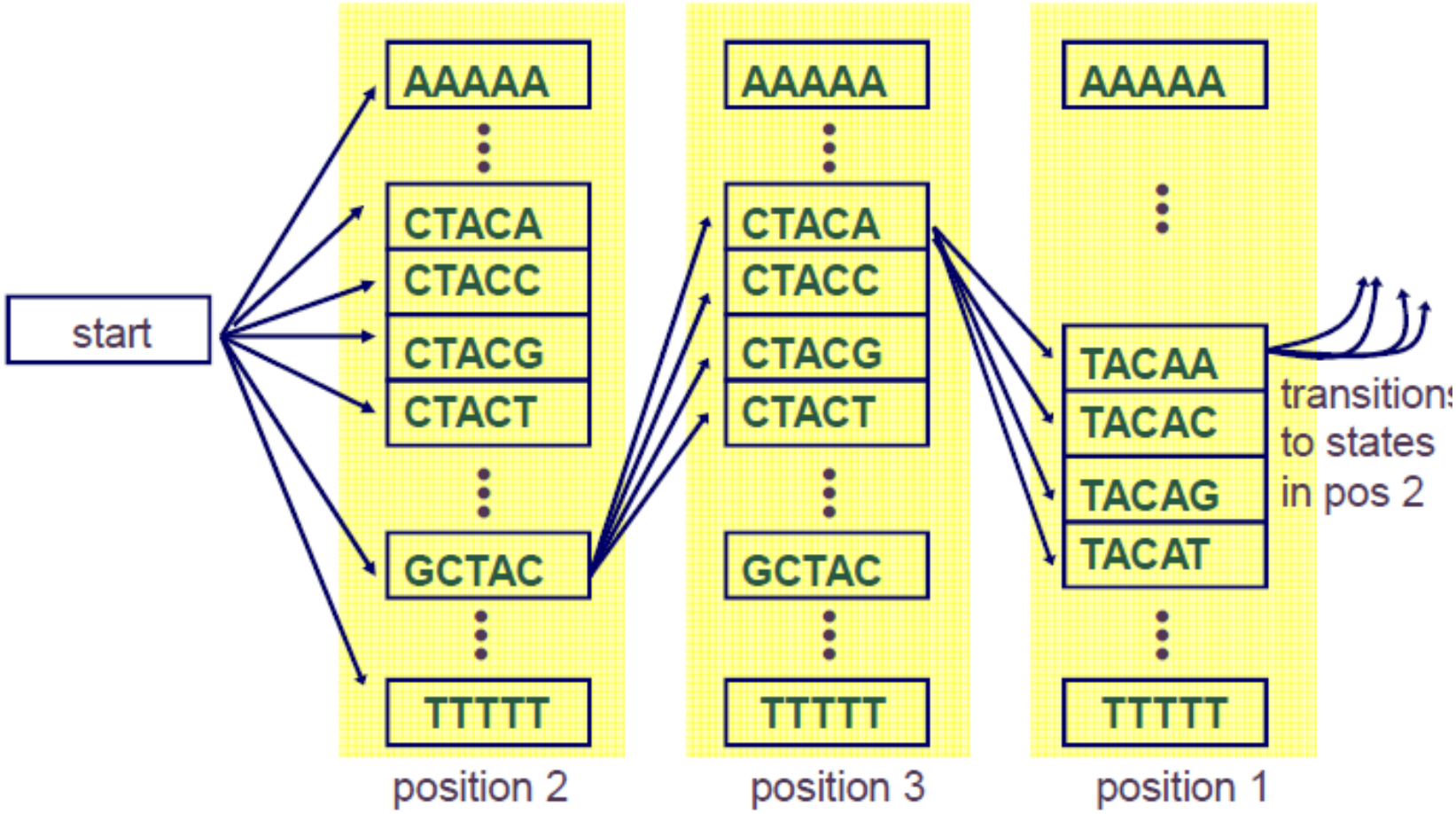
Inhomogenous Markov chains

- in an **inhomogeneous** Markov model, we can have different distributions at different positions in the sequence,
- consider modeling codons in protein coding regions.



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

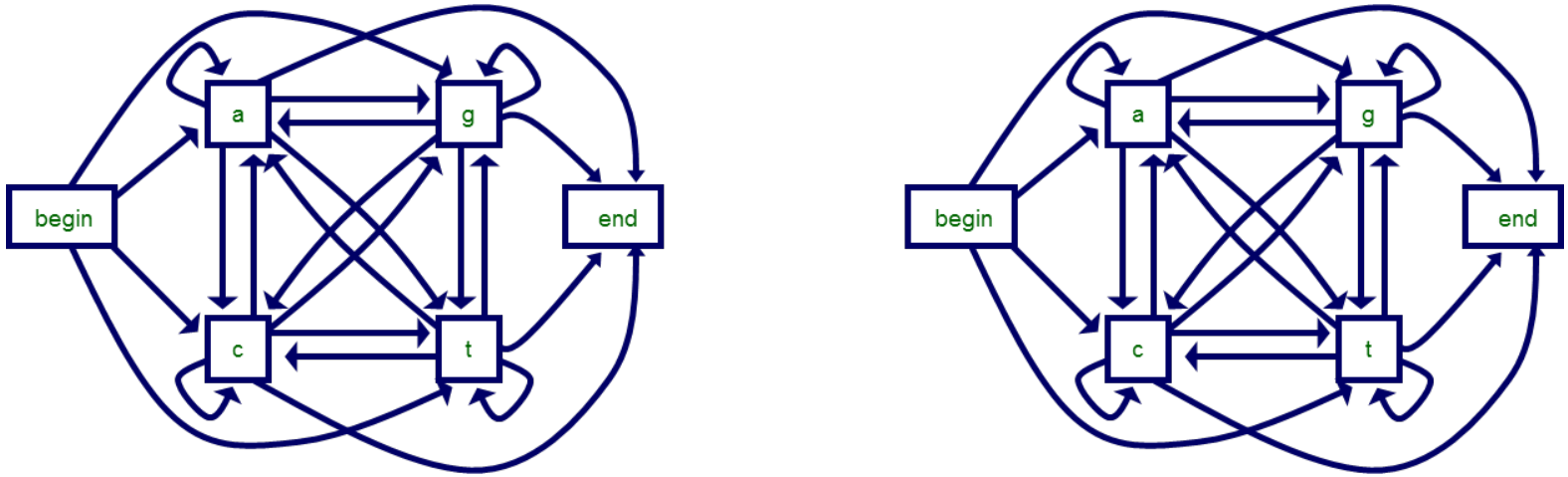
A fifth-order inhomogenous Markov chain



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Example Markov chain application

- CpG islands
  - **CG** dinucleotides are rarer in eukaryotic genomes than expected given the marginal probabilities of **C** and **G**,
  - CpG islands = the regions upstream of genes rich in CG dinucleotides,
  - useful evidence for finding genes,
- could classify CpG islands with Markov chains
  - one to represent CpG islands, one to represent the rest of the genome.



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

CpG islands as a classification task

- train a CpG chain and a null chain
  - parameters estimated from sample sequences,
  - in here, human sequences with 48 CpG islands, 60000 nucleotides,

$P(c \mid a)$

+	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

Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

- given a test sequence X, use two models to
  - determine its probability given both the models,
  - classify the sequence = compare the posterior probabilities.

Markov chains for discrimination

- compare the posterior probabilities, use Bayes' rule

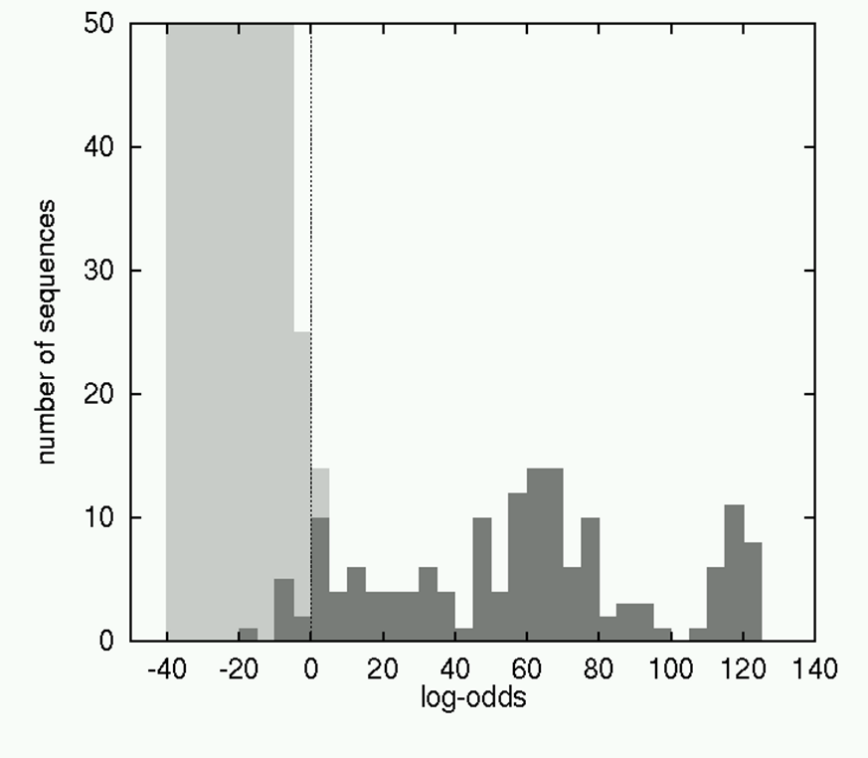
$$P(CpG|X) = \frac{P(X|CpG)P(CpG)}{P(X)} = \\ = \frac{P(X|CpG)P(CpG)}{P(X|CpG)P(CpG) + P(X|null)P(null)}$$

- if we do not know prior probabilities of two classes ($P(CpG)$ and $P(null)$) then we just need to compare $P(X|CpG)$ and $P(X|null)$
  - i.e, the probabilities derived from the chains,
- often shown and compared in terms of log odds

$$log\frac{P(CpG|X)}{P(null|X)} = logP(CpG|X) - logP(null|X) \geq 0$$

Markov chains for discrimination

- light bars represent negative sequences,
- dark bars represent positive sequences (e.g., CpG islands),
- however, the figure here is not from a CpG island discrimination task.



Krogh et al.: An Introduction to Hidden Markov Models for Biological Sequences.

Summary

- DNA and protein Markov chains
  - simple stochastic models representing local sequential regularities,
  - could be used for sequence generation as well as their discrimination,
- key terms
  - the order of the chain
    - the size of memory of the process,
    - a trade-off between informedness and the size of the model,
  - homogeneity of the chain
    - do we distinguish different positions in the sequence?
  - regularization
    - do we learn the chain parameters purely from observed data?
    - YES = no regularization = MLE,
    - NO = regularization = pseudocounts, prior beliefs.