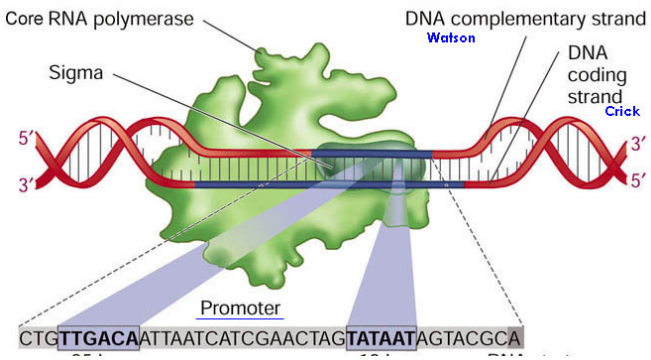
Markov Chain Models (Part 1)

BMI/CS 576
www.biostat.wisc.edu/bmi576/
Mark Craven
craven@biostat.wisc.edu
Fall 2011

Motivation for sequence modeling



these sequences are E. coli promoters

```
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca
accggaagaaaaccgtgacattttaacacgtttgttacaaggtaaaggcgacgccgc
aaattaaaatttattgacttaggtcactaaatactttaaccaatataggcatagcg
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?

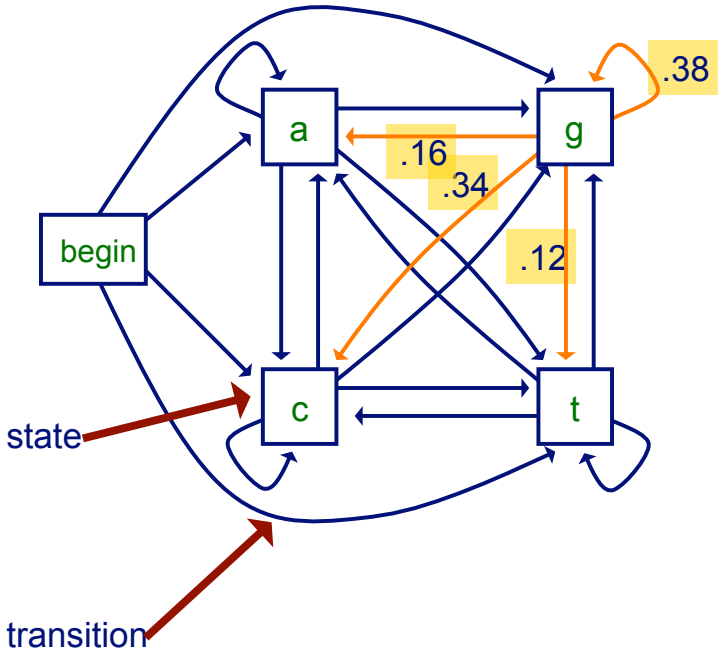
```
ccatcaaaaaaatattctcaacataaaaaactttgtgtaatacttgtaacgctacat
```

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. promoters)
  - proteins in a given family
  - etc.

- Markov models are well suited to this type of task

A Markov chain model



transition probabilities

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

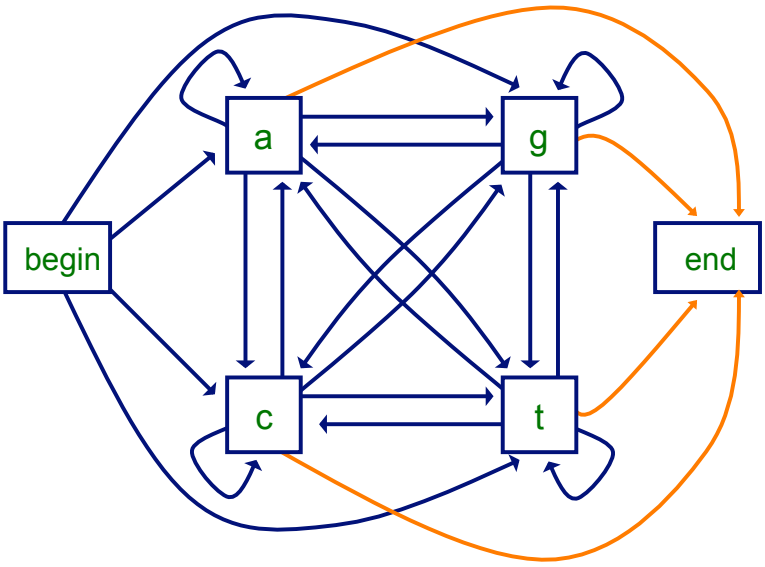
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

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

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

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

- 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*
  - 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

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

$$
\begin{aligned}
P(X) &= P(X_L, X_{L-1}, \ldots, X_1) \\
&= P(X_L \mid X_{L-1}, \ldots, X_1) \times \\
&\quad P(X_{L-1} \mid X_{L-2}, \ldots, X_1) \times \\
&\quad \vdots \\
&\quad P(X_1)
\end{aligned}
$$

Markov chain models

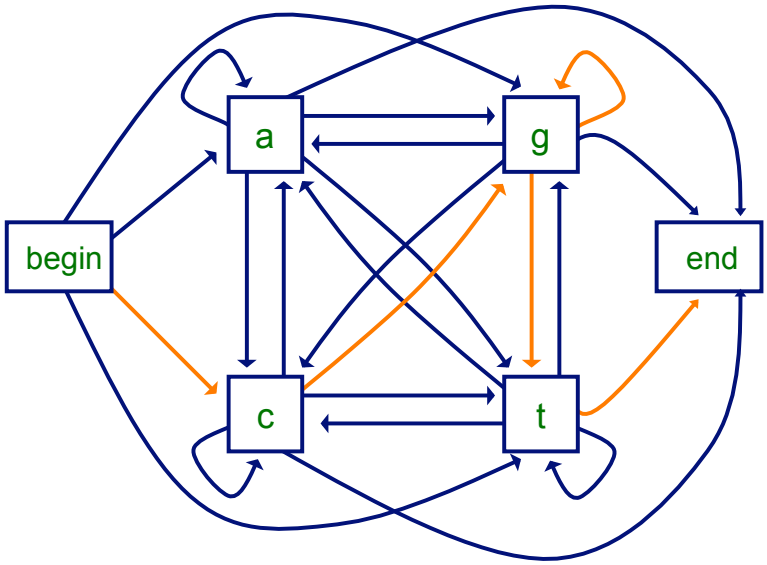
- from the chain rule we have

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

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

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

The probability of a sequence for a given Markov chain model



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

Markov chain notation

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

$$a_{x_{i-1}x_i} = P(x_i \mid x_{i-1})$$

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

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

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

Estimating the model parameters

- Given some data, 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

$$P(D \mid \theta)$$

  - i.e. make the data D look <u>as likely as possible</u> under the model

Maximum likelihood estimation

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

  accgcgctta

  gcttagtgac

  tagccgttac

$$P(a) = \frac{n_a}{\sum_i n_i}$$

- then the maximum likelihood estimates are

$$P(a) = \frac{6}{30} = 0.2 \qquad P(g) = \frac{7}{30} = 0.233$$

$$P(c) = \frac{9}{30} = 0.3 \qquad P(t) = \frac{8}{30} = 0.267$$

Maximum likelihood estimation

- suppose instead we saw the following sequences

  gccgcgcttg

  gcttggtggc

  tggccgttgc

- then the maximum likelihood estimates are

$$\boxed{P(a) = \frac{0}{30} = 0} \qquad P(g) = \frac{13}{30} = 0.433$$

$$P(c) = \frac{9}{30} = 0.3 \qquad P(t) = \frac{8}{30} = 0.267$$

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*

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

pseudocount

- where n_i represents the number of occurrences of character i
- using Laplace estimates with the sequences

  gccgcgcttg

  gcttggtggc

  tggccgttgc

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

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

A Bayesian approach

- a more general form: *m-estimates*

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

prior probability of a

number of "virtual" instances

- with m=8 and uniform priors

gccgcgcttg
gcttggtggc
tggccgttgc

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

Estimation for 1st order probabilities

- to estimate a 1st order parameter, such as *P*(*c*|*g*), we count the number of times that *g* follows the history *c* in our given sequences

- using Laplace estimates with the sequences

gccgcgcttg
gcttggtggc
tggccgttgc

$$P(a \mid g) = \frac{0+1}{12+4} \qquad P(a \mid c) = \frac{0+1}{7+4}$$

$$\vdots$$

$$P(c \mid g) = \frac{7+1}{12+4}$$

$$P(g \mid g) = \frac{3+1}{12+4}$$

$$P(t \mid g) = \frac{2+1}{12+4}$$