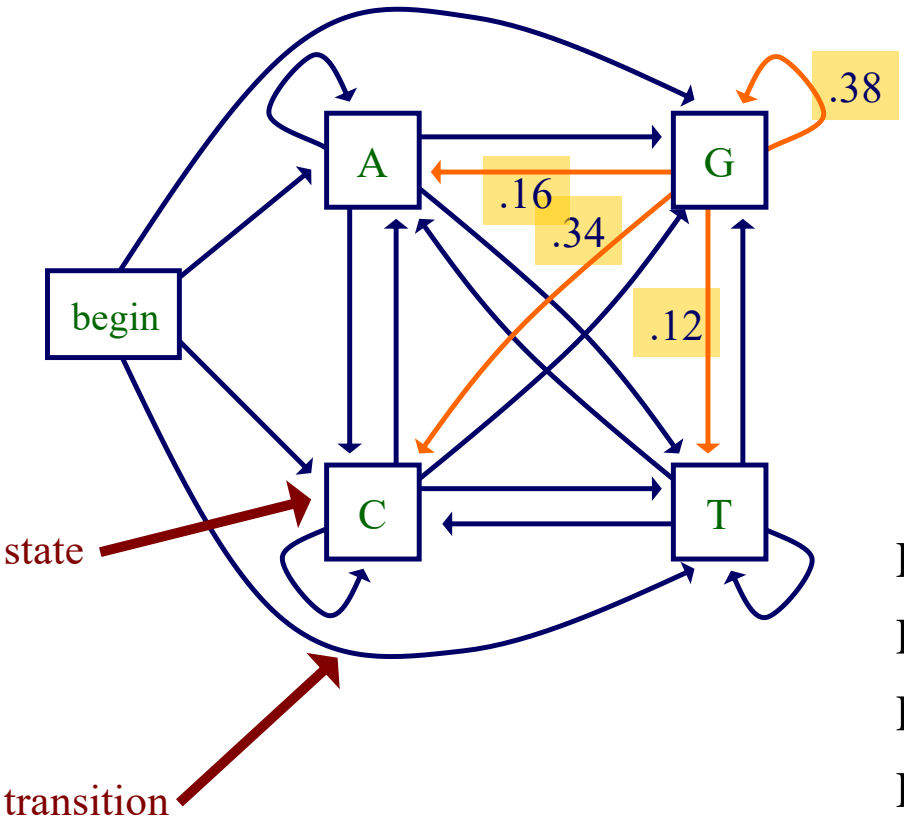
Genome Annotation

Markov Chain parameter estimation

Outline

- Estimating the parameters of a Markov chain
  - Maximum likelihood estimation
  - Bayesian approach

A Markov Chain Model for DNA



transition probabilities

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

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

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

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

Estimating the Model Parameters

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

$$\Pr(D \mid \theta)$$

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

Maximum Likelihood Estimation

- Let's use a very simple sequence model (even simpler than a Markov chain)
  - every position is independent of the others
  - every position generated from the same categorical distribution (multinomial distribution with $n=1$, a single trial)
- we want to estimate the parameters Pr(a), Pr(c), Pr(g), Pr(t)
- and we're given the sequences

  accgcgctta

  gcttagtgac

  tagccgttac

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

- then the maximum likelihood estimates are the observed frequencies of the bases

$$\Pr(a) = \frac{6}{30} = 0.2 \quad \Pr(g) = \frac{7}{30} = 0.233$$

$$\Pr(c) = \frac{9}{30} = 0.3 \quad \Pr(t) = \frac{8}{30} = 0.267$$

Maximum Likelihood Estimation

- suppose instead we saw the following sequences

  gccgcgcttg

  gcttggtggc

  tggccgttgc

- then the maximum likelihood estimates are

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

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

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

- where n_i represents the number of occurrences of character i

- 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

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

Estimation for Markov chains

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

- using Laplace estimates with the sequences

gccgcgcttg
gcttggtggc
tggccgttgc

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

$$\vdots$$

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

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

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

Summary

- Estimation of parameters by maximum likelihood
- Estimates for categorical distributions are simply the frequencies of events in the training set
- Bayesian approaches to handle small training sets
  - Laplace estimates
  - M-estimates
- For Markov chains of DNA, estimates are obtained by counting the frequency at which each base follows another base in the training set