A Markov Chain Model

- Transition probabilities
  - \( \Pr(x_i=a|x_{i-1}=g)=0.16 \)
  - \( \Pr(x_i=c|x_{i-1}=g)=0.34 \)
  - \( \Pr(x_i=g|x_{i-1}=g)=0.38 \)
  - \( \Pr(x_i=t|x_{i-1}=g)=0.12 \)

\[ \sum \Pr(x_i \mid x_{i-1} = g) = 1 \]

- Exercise
  - Can “acgt” be generated from the MC model?
  - How about “cagt”, “gcat”, “tcga”?
Definition of Markov Chain Model

• A Markov chain model is defined by
  • a set of states
    • some states emit symbols
    • other states (e.g., the begin state) 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: Properties

- Given some sequence $x$ of length $L$, we can ask the probability of the sequence with the given model.
- For any probabilistic model of sequences, we can write this probability as

$$
\text{Pr}(x) = \text{Pr}(x_L, x_{L-1}, \ldots, x_1)
$$

$$
= \text{Pr}(x_L / x_{L-1}, \ldots, x_1) \text{Pr}(x_{L-1} / x_{L-2}, \ldots, x_1) \ldots \text{Pr}(x_1)
$$

- Key property of a (1st order) Markov chain: the probability of each $x_i$ depends only on the value of $x_{i-1}$

$$
\text{Pr}(x) = \text{Pr}(x_L / x_{L-1}) \text{Pr}(x_{L-1} / x_{L-2}) \ldots \text{Pr}(x_2 / x_1) \text{Pr}(x_1)
$$

$$
= \text{Pr}(x_1) \prod_{i=2}^{L} \text{Pr}(x_i / x_{i-1})
$$
The Probability of a Sequence for a Markov Chain Model

\[
Pr(cggt) = Pr(c)Pr(g|c)Pr(g|g)Pr(t|g)
\]
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

\[
\Pr( x_i \mid x_{i-1}, x_{i-2}, \ldots, x_1 ) = \Pr( x_i \mid x_{i-1}, \ldots, x_{i-n} )
\]
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 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 8-th order chain, we’d see each word ~ 5 times on average.
Hidden Markov Model: A Simple HMM

Given observed sequence GAA, which state emits every item?
Hidden Markov Model: A Simple HMM

Given observed sequence GAA, which state emits every item?
References

- Lecture notes@M. Craven’s website: www.biostat.wisc.edu/~craven
- I. Korf, M. Yandell, and J. Bedell. BLAST. O'Reilly, 2003