Lecture 3
Probabilistic Sequence Models

Burr Settles
IBS Summer Research Program 2008
bsettles@cs.wisc.edu
www.cs.wisc.edu/~bsettles/ibs08/
Probability 101

• *frequentist* interpretation: the probability of an event is the proportion of the time events of same kind will occur in the long run

• examples
  – the probability my flight to Chicago will be on time
  – the probability this ticket will win the lottery
  – the probability it will rain tomorrow

• always a number in the interval \([0,1]\)
  0 means “never occurs”
  1 means “always occurs”
Sample Spaces

- *sample space*: a set of possible outcomes for some event

- **examples**
  - flight to Chicago: \{on time, late\}
  - lottery: \{ticket 1 wins, ticket 2 wins, ..., ticket \(n\) wins\}
  - weather tomorrow:
    - \{rain, not rain\} or
    - \{sun, rain, snow\} or
    - \{sun, clouds, rain, snow, sleet\} or...
Random Variables

• *random variable*: a variable representing the outcome of an experiment

• example:
  – $X$ represents the outcome of my flight to Chicago
  – we write the probability of my flight being on time as $\Pr(X = \text{on-time})$
  – or when it’s clear which variable we’re referring to, we may use the shorthand $\Pr(\text{on-time})$
Notation

- uppercase letters and capitalized words denote random variables
- lowercase letters and uncapitalized words denote values
- we’ll denote a particular value for a variable as follows
  \[ \Pr(X = x) \quad \Pr(\text{Fever} = \text{true}) \]
- we’ll also use the shorthand form
  \[ \Pr(x) \quad \text{for} \quad \Pr(X = x) \]
- for Boolean random variables, we’ll use the shorthand
  \[ \Pr(\text{fever}) \quad \text{for} \quad \Pr(\text{Fever} = \text{true}) \]
  \[ \Pr(\neg \text{fever}) \quad \text{for} \quad \Pr(\text{Fever} = \text{false}) \]
Probability Distributions

- If $X$ is a random variable, the function given by $\Pr(X = x)$ for each $x$ is the \textit{probability distribution} of $X$.

- Requirements:

  \[ \Pr(x) \geq 0 \quad \text{for every } x \]

  \[ \sum_x \Pr(x) = 1 \]
Joint Distributions

- **joint probability distribution**: the function given by \( \Pr(X = x, Y = y) \)
- read “\( X \) equals \( x \) and \( Y \) equals \( y \)”
- example

<table>
<thead>
<tr>
<th>( x, y )</th>
<th>( \Pr(X = x, Y = y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

probability that it’s sunny and my flight is on time
Marginal Distributions

- The marginal distribution of $X$ is defined by
  \[ \Pr(x) = \sum_y \Pr(x, y) \]
  "the distribution of $X$ ignoring other variables"

- This definition generalizes to more than two variables, e.g.
  \[ \Pr(x) = \sum_y \sum_z \Pr(x, y, z) \]
# Marginal Distribution Example

<table>
<thead>
<tr>
<th>$x, y$</th>
<th>$\Pr(X = x, Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\Pr(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>rain</td>
<td>0.5</td>
</tr>
<tr>
<td>snow</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The joint distribution shows the probabilities of weather conditions and whether they are on-time or late. The marginal distribution for $X$ shows the probabilities of each weather condition, regardless of whether it is on-time or late.
Conditional Distributions

- the *conditional distribution* of $X$ given $Y$ is defined as:

$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

“the distribution of $X$ given that we know $Y$”
### Conditional Distribution Example

**joint distribution**

<table>
<thead>
<tr>
<th>$x, y$</th>
<th>$\Pr(X = x, Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**conditional distribution for $X$ given $Y=$on-time**

| $x$                | $\Pr(X = x|Y=\text{on-time})$ |
|--------------------|--------------------------------|
| sun                | $0.20/0.45 = 0.444$            |
| rain               | $0.20/0.45 = 0.444$            |
| snow               | $0.05/0.45 = 0.111$            |
Independence

- two random variables, $X$ and $Y$, are *independent* if
  \[ \Pr(x, y) = \Pr(x) \times \Pr(y) \quad \text{for all } x \text{ and } y \]
## Independence Example #1

<table>
<thead>
<tr>
<th>joint distribution</th>
<th>$x, y$</th>
<th>$\Pr(X = x, Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, on-time</td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td></td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>marginal distributions</th>
<th>$x$</th>
<th>$\Pr(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td>rain</td>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>snow</td>
<td></td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
<th>$\Pr(Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>on-time</td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td>late</td>
<td></td>
<td>0.55</td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent here?  **NO.**
**Independence Example #2**

<table>
<thead>
<tr>
<th>joint distribution</th>
<th>Pr($X = x, Y = y$)</th>
<th>marginal distributions</th>
<th>Pr($X = x$)</th>
<th>Pr($Y = y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x, y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sun, fly-United</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain, fly-United</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>snow, fly-United</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sun, fly-Northwest</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain, fly-Northwest</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>snow, fly-Northwest</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent here? **YES.**
Conditional Independence

- two random variables $X$ and $Y$ are \textit{conditionally independent} given $Z$ if
  \[
  \Pr(X \mid Y, Z) = \Pr(X \mid Z)
  \]
  “once you know the value of $Z$, knowing $Y$ doesn’t tell you anything about $X$”

- alternatively
  \[
  \Pr(x, y \mid z) = \Pr(x \mid z) \times \Pr(y \mid z) \text{ for all } x, y, z
  \]
### Conditional Independence Example

<table>
<thead>
<tr>
<th>Flu</th>
<th>Fever</th>
<th>Vomit</th>
<th>Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>0.04</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>0.04</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>true</td>
<td>0.01</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>0.01</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>true</td>
<td>0.009</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>0.081</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>0.081</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>0.729</td>
</tr>
</tbody>
</table>

Fever and Vomit are not independent: \( \Pr(\text{fever, vomit}) \neq \Pr(\text{fever}) \times \Pr(\text{vomit}) \)

Fever and Vomit are conditionally independent given Flu:

\[
\Pr(\text{fever, vomit} | \text{flu}) = \Pr(\text{fever} | \text{flu}) \times \Pr(\text{vomit} | \text{flu})
\]

\[
\Pr(\text{fever, vomit} | \neg \text{flu}) = \Pr(\text{fever} | \neg \text{flu}) \times \Pr(\text{vomit} | \neg \text{flu})
\]

etc.
Bayes Theorem

\[ \Pr(x \mid y) = \frac{\Pr(y \mid x) \Pr(x)}{\Pr(y)} = \frac{\Pr(y \mid x) \Pr(x)}{\sum_x \Pr(y \mid x) \Pr(x)} \]

- this theorem is extremely useful
- there are many cases when it is hard to estimate \( \Pr(x \mid y) \) directly, but it’s not too hard to estimate \( \Pr(y \mid x) \) and \( \Pr(x) \)
Bayes Theorem Example

- MDs usually aren’t good at estimating $\Pr(\text{Disorder} \mid \text{Symptom})$
- they’re usually better at estimating $\Pr(\text{Symptom} \mid \text{Disorder})$
- if we can estimate $\Pr(\text{fever} \mid \text{flu})$ and $\Pr(\text{flu})$ we can use Bayes’ Theorem to do diagnosis

$$
\Pr(\text{flu} \mid \text{fever}) = \frac{\Pr(\text{fever} \mid \text{flu}) \Pr(\text{flu})}{\Pr(\text{fever} \mid \text{flu}) \Pr(\text{flu}) + \Pr(\text{fever} \mid \neg\text{flu}) \Pr(\neg\text{flu})}
$$
Expected Values

• the *expected value* of a random variable that takes on numerical values is defined as:

\[
E[X] = \sum_x x \times \Pr(x)
\]

this is the same thing as the *mean*

• we can also talk about the expected value of a function of a random variable

\[
E[g(X)] = \sum_x g(x) \times \Pr(x)
\]
Expected Value Example

• Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

\[
E[gain(Lottery)] = gain(\text{winning}) \cdot \Pr(\text{winning}) + gain(\text{losing}) \cdot \Pr(\text{losing}) = (\$100 - \$1) \times 0.001 - \$1 \times 0.999 = -\$0.90
\]
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
Probability Of A Sequence

- given some sequence \( x \) of length \( L \), we want to compute its probability (likelihood)
- one way to compute this is the joint probability of all the characters in the sequence:

\[
\Pr(x) = \Pr(x_1, x_2, \ldots, x_L)
\]

\[
= \Pr(x_1) \Pr(x_2 | x_1) \ldots \Pr(x_L | x_1, \ldots, x_{L-1})
\]

- for example:

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

- **problem**: biological sequences tend to be very long; that’s too many conditional probabilities to estimate!
The Markov Assumption

- trick: assume the probability of a character is only dependent on the previous character, not the entire prefix

\[
\Pr(x) = \Pr(x_1, x_2, \ldots, x_L) \\
\approx \Pr(x_1) \Pr(x_2 \mid x_1) \ldots \Pr(x_{L-1} \mid x_{L-2}) \Pr(x_L \mid x_{L-1}) \\
= \Pr(x_1) \prod_{i=2}^{L} \Pr(x_i \mid x_{i-1})
\]

- now our probabilities are easier to estimate:

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

- this trick is called the Markov assumption, and a statistical process that uses it is called a Markov chain
Markov Chain Models

transition probabilities

\[ \text{Pr}(x_i = a \mid x_{i-1} = g) = 0.16 \]
\[ \text{Pr}(x_i = c \mid x_{i-1} = g) = 0.34 \]
\[ \text{Pr}(x_i = g \mid x_{i-1} = g) = 0.38 \]
\[ \text{Pr}(x_i = t \mid x_{i-1} = g) = 0.12 \]
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

- 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 Notation

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

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

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

\[
a_{Bx_1} \prod_{i=2}^{L} a_{x_{i-1}x_i} = \Pr(x_1) \prod_{i=2}^{L} \Pr(x_i \mid x_{i-1})
\]

where \( a_{Bx_1} \) represents the transition from the \textit{begin} state
The Probability of a Sequence for a Given Markov Chain Model

\[ \Pr(cggt) = \Pr(c) \Pr(g \mid c) \Pr(g \mid g) \Pr(t \mid g) \Pr(\text{end} \mid t) \]
Estimating the Model Parameters

• given some data (e.g. a set of sequences), how can we determine the probability parameters of our model?

• one approach: maximum likelihood estimation
  – given a set of data $D$
  – set the parameters $\theta$ to maximize

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

  – i.e. make the data $D$ look as likely as possible under the model $\theta$
Maximum Likelihood Estimation

- suppose we want to estimate the parameters: Pr(a), Pr(c), Pr(g), Pr(t)
- and we’re given the sequences
  accgcgctta
gcttagtgac
tagccgttac
- then the maximum likelihood estimates are

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

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

\[
Pr(c) = \frac{9}{30} = 0.3 \quad Pr(t) = \frac{8}{30} = 0.267
\]
Maximum Likelihood Estimation

• suppose instead we saw the following sequences
  gccgcgcttg
  gcttggtggc
  tggccgttgc
• then the maximum likelihood estimates are

\[
\begin{align*}
\Pr(a) &= \frac{0}{30} = 0 \\
\Pr(c) &= \frac{9}{30} = 0.3 \\
\Pr(g) &= \frac{13}{30} = 0.433 \\
\Pr(t) &= \frac{8}{30} = 0.267
\end{align*}
\]

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*

\[
\Pr(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}
\]

• where \( n_i \) represents the number of occurrences of character \( i \)

• using Laplace estimates with the sequences
  
  \[
  \begin{align*}
  \text{gccgcgccttg} & \quad \Pr(a) = \frac{0 + 1}{34} \\
  \text{gcttggtgge} & \quad \Pr(c) = \frac{9 + 1}{34}
  \end{align*}
  \]
A Bayesian Approach

- a more general form: \( m\)-estimates

\[
Pr(a) = \frac{n_a + p_a m}{\left( \sum_i n_i \right) + m}
\]

- with \( m=8 \) and uniform priors

\[
\begin{align*}
gccgcgcttg & \\
gctttggtggc & \\
tggccgcttgc & \\
\end{align*}
\]

\[
Pr(c) = \frac{9 + 0.25 \times 8}{30 + 8} = \frac{11}{38}
\]
Estimation for 1\textsuperscript{st} Order Probabilities

- to estimate a 1\textsuperscript{st} order parameter (where each character depends on 1 previous character), 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:

\[
\begin{align*}
gccgcgcttg & \quad \Pr(a \mid g) = \frac{0+1}{12+4} \quad \Pr(a \mid c) = \frac{0+1}{7+4} \\
gcttggtggc & \quad \Pr(c \mid g) = \frac{7+1}{12+4} \\
tggccgttgc & \quad \Pr(g \mid g) = \frac{3+1}{12+4} \\
\end{align*}
\]
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

- higher order models remember more “history”
- additional history can have predictive value
- example:
  - predict the next word in this sentence fragment
    “…finish _____” (up, it, first, last, …?)
  - now predict it given more history
    “nice guys finish _____”
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 8th order chain, we’d see each “word” about 5 times on average
Higher Order Markov Chains

• an $n$th order Markov chain over some alphabet $A$ is equivalent to a first order Markov chain over the alphabet of $n$-tuples $A^n$

• example: a 2$^\text{nd}$ order Markov model for DNA can be treated as a 1$^\text{st}$ order Markov model over alphabet

\[\text{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}\]

• caveat: we process a sequence one character at a time

\[\text{A C G G T}\]

\[\text{AC} \rightarrow \text{CG} \rightarrow \text{GG} \rightarrow \text{GT}\]
A Fifth Order Markov Chain

begin

Pr(GCTAC)

Pr(A | GCTAC)

Pr(C | GCTAC)

Pr(GCTAC)

AAAAA

CTACA

CTACC

CTACG

CTACT

GCTAC

TTTTT
A Fifth Order Markov Chain

\[ \text{Pr}(gctaca) = \text{Pr}(gctac) \text{Pr}(a \mid gctac) \]
Example Application

• language classification
• given:
  – passages of text from different languages
  – e.g. newspaper articles written in English, French, Spanish, German, and Italian
• do:
  – learn a Markov chain model for each language
  – use these models to determine the most likely language for some new passage of text

• http://pages.cs.wisc.edu/~bsettles/webtoys/polyglot/
Example Biological 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
Example Biological Application

- 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:

\[
\text{score}(x) = \log \frac{\Pr(x \mid \text{CpG model})}{\Pr(x \mid \text{null model})}
\]
Example Biological Application

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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>c</th>
<th>g</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.18</td>
<td></td>
<td>.27</td>
<td>.43</td>
</tr>
<tr>
<td>c</td>
<td>.17</td>
<td>.37</td>
<td>.27</td>
<td>.19</td>
</tr>
<tr>
<td>g</td>
<td>.16</td>
<td>.34</td>
<td>.38</td>
<td>.12</td>
</tr>
<tr>
<td>t</td>
<td>.08</td>
<td>.36</td>
<td>.38</td>
<td>.18</td>
</tr>
</tbody>
</table>

Pr(c | a)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>c</th>
<th>g</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>.30</td>
<td>.21</td>
<td>.28</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>.32</td>
<td>.30</td>
<td>.08</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>.25</td>
<td>.24</td>
<td>.30</td>
</tr>
<tr>
<td>t</td>
<td></td>
<td>.18</td>
<td>.24</td>
<td>.29</td>
</tr>
</tbody>
</table>

CpG  null
Example Biological Application

- 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.
Example Biological Application

• why use

\[
\text{score}(x) = \log \frac{\Pr(x \mid \text{CpG})}{\Pr(x \mid \text{null})}
\]

• Bayes’ rule tells us

\[
\Pr(\text{CpG} \mid x) = \frac{\Pr(x \mid \text{CpG}) \Pr(\text{CpG})}{\Pr(x)}
\]

\[
\Pr(\text{null} \mid x) = \frac{\Pr(x \mid \text{null}) \Pr(\text{null})}{\Pr(x)}
\]

• if we’re not taking into account prior probabilities of two classes (\( \Pr(\text{CpG}) \) and \( \Pr(\text{null}) \)) then we just need to compare \( \Pr(x \mid \text{CpG}) \) and \( \Pr(x \mid \text{null}) \)
• given say a $T$ in our input sequence, which state emitted it?
Hidden State

• we’ll distinguish between the *observed* parts of a problem and the *hidden* parts

• in the Markov models we’ve considered previously, it is clear which state accounts for each part of the observed sequence

• in this example, there are multiple states that could account for each part of the observed sequence
  – this is the *hidden* part of the problem
  – *hidden Markov models* (HMMs) are Markov chain models with hidden state
Simple HMM for Gene Finding

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences
HMM Applications

• classification
  – *given*: a set of models representing different sequence classes (e.g. protein families), and a test sequence
  – *do*: determine which model/class best explains the sequence
  – use Forward algorithm to calculate probability of sequence under each model

• segmentation
  – *given*: a model representing different sequence classes, a test sequence
  – *do*: segment the sequence into subsequences, predicting the state labels for each subsequence
  – use Viterbi algorithm to find most probable path for sequence
Example: Protein Classification

given: amino-acid sequence of a protein
do: predict the family to which it belongs

GDLSTPDAMGPKVKAHGKKVLGAFSDGLALHDLNKGTFATLSELHCDKLIHVDPENFRLLGNVCVLAHFFGKEFTPPVQAAYAKVVAVCAGNALHKYH
Alignment of Globin Family Proteins

- The sequences in a family may vary in length
- Some positions are more conserved than others
Profile HMMs

- profile HMMs are commonly used to model families of sequences

*Delete states* are silent; they account for characters missing in some sequences.

*Insert states* account for extra characters in some sequences.

*Match states* represent key conserved positions.

*Insert and match states* have emission distributions over sequence characters.
Profile HMM Accuracy

- classifying 2447 proteins into 33 families
- \( x \)-axis represents the median \# of negative sequences that score as high as a positive sequence for a given family’s model

Figure from Jaakola et al., ISMB 1999
Example: Gene Finding

given: an uncharacterized DNA sequence
do: locate the genes in the sequence, including the coordinates of individual *exons* and *introns*

image from the UCSC Genome Browser
http://genome.ucsc.edu/
Eukaryotic Gene Structure
The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc.
Example: Information Extraction From Biomedical Literature

given: a passage of text from a scientific article
do: identify mentions of genes or proteins, annotate the article with this information in a database
Next Time…

- basic molecular biology
- sequence alignment
- probabilistic sequence models
- gene expression analysis
- protein structure prediction
  - by Ameet Soni