TODAY’S TOPICS

- Profile hidden Markov models
- Protein family/domain search (Pfam, SMART, InterPro, etc.)
- Distance estimation
  - for DNA sequences

What is hidden Markov model?

Markov chain
Fair coin vs. loaded coin [states: H and T]

Initial Probability

Transition Probability

$P_0 = \begin{pmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{pmatrix}$

$P_f = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$

If you see three heads in a row, which coin do you have more likely?

Observation: HHH
State predicted: Fair coin
What is hidden Markov model?

• Markov chain
  If fair and loaded coins are mixed!
  \[
  P_H = \begin{bmatrix}
  0.5 & 0.5 \\
  0.5 & 0.5 
  \end{bmatrix}
  \]
  \[
  P_L = \begin{bmatrix}
  0.4 & 0.6 \\
  0.4 & 0.6 
  \end{bmatrix}
  \]
  \[
  P_{LO} = \begin{bmatrix}
  0.4 & 0.6 \\
  0.4 & 0.6 
  \end{bmatrix}
  \]

• Observation: HHH
  \[P(H_H_H_H_H)\text{ or } P(H_H_H_H_L)\text{ or } P(L_H_H_H_H)\text{ or } \ldots \]

What is hidden Markov model?

• State sequence is unknown (hidden): Fair or Loaded

  Observation: H T H T T T T

  Can we guess the hidden state sequence?

  What we know:
  
  Transition probabilities:
  Fair: \[
  \begin{bmatrix}
  0.5 & 0.5 \\
  0.5 & 0.5 
  \end{bmatrix}
  \]
  Loaded: \[
  \begin{bmatrix}
  0.4 & 0.6 \\
  0.4 & 0.6 
  \end{bmatrix}
  \]

  Initial probabilities: (0.5, 0.5)

  Emission probabilities:
  Fair: H 0.5, T 0.5
  Loaded: H 0.4, T 0.6
**What is hidden Markov model?**

- State sequence is unknown (hidden): Fair or Loaded

---

**Profile hidden Markov models**

- Probabilistic models of multiple alignments (Krogh et al., 1994; Krogh 1998; Eddy 2004)
  - Closely related to standard profiles (PSSMs) introduced by Gribskov
  - Used in e.g., PFAM, SMART, Superfamily, Panther (databases of multiple alignments and profile HMMs for domains and protein families)
  - States: Insertion, Deletion, and Match
  - Transition probabilities: between states
  - Emission probabilities: for 20 amino acids or 4 nucleotides

---

**Basic architecture of a profile HMM**

- Any sequence can be represented by a path through the model
- Multiple paths are possible to model a sequence

---

**Basic architecture of a profile HMM**

- Observed sequence: ACCY
- P(sequence | path): the product of emission and transition probabilities
  \[ P(\text{observed sequence} | \text{path}) = 0.4 \times 0.3 \times 0.46 \times 0.01 \times 0.97 \times 0.5 \times 0.97 \times 0.01 \times 0.7 \]
Basic architecture of a profile HMM

Observed sequence: ACCY

P(sequence | path): the product of emission and transition probabilities

$$P(\text{sequence} | \text{path}) = P(\text{sequence} | \text{path}_1) + P(\text{sequence} | \text{path}_2) + \ldots$$

→ Probability of the sequence given the model

Profile HMM search

• Sequence vs. Profile HMM database

Query sequence

HMM database

Search result

Find functional domains from a sequence. Identify its functions or protein family membership.

Profile HMM search by PFAM/HMMER3

Like PSI-blast

Like blastp

Like FASTA

The Pfam protein families database

http://pfam.xfam.org

A collection of multiple sequence alignments and profile HMMs, each representing a protein family or domain.
Clan: a collection of families

Score: \( \log \left( \frac{P(\text{seq} | \text{HMM})}{P(\text{seq} | \text{null})} \right) \)

Switch between "Standard" and "Advanced" views

Clust: a collection of families

Information Content (bits)

For each column, \( p_i \): Probability of AA

\( \sum_{i=1}^{n} p_i \log \left( \frac{p_i}{q_i} \right) \)

\( \text{Stack height:} \)

\( \text{Letter height} \propto \text{letter frequency} \)

\( \text{Red} \rightarrow \text{more insertions} \)

\( \text{Blue} \rightarrow \text{more deletions} \)

Occupancy

Insert Probability

Expected insert length

\( \text{Red} \rightarrow \text{more insertions} \)
Profile HMM entry

- Pfam: a database of multiple alignments of protein domains or conserved protein regions [http://pfam.xfam.org](http://pfam.xfam.org)
- InterPro: Protein sequence analysis & classification [https://www.ebi.ac.uk/interpro/](https://www.ebi.ac.uk/interpro/)
- HMMER3: profile HMM [http://www.ebi.ac.uk/Tools/hmmer/](http://www.ebi.ac.uk/Tools/hmmer/)
- HHblits: Homology detection by iterative HMM-HMM comparison [MPI Bioinformatics Toolkit](https://toolkit.tuebingen.mpg.de)

Profile HMM databases for domain/protein families

- Pfam: a database of multiple alignments of protein domains or conserved protein regions
- SMART: Simple Modular Architecture Research Tool
- Superfamily: HMM library and genome assignments server
- Panther: Protein ANalysis THrough Evolutionary Relationships
- InterPro: Protein sequence analysis & classification
- CDD: Conserved domain database
- HMMER3: profile HMM
- HHblits: Homology detection by iterative HMM-HMM comparison

Distance estimation

- Number of substitutions per site (\( p \)) or degree of divergence
  \[ p = \frac{n_d}{L} \]
  \[ V(p) = \frac{p(1-p)}{L} \]
  \[ \sigma(p) = \sqrt{\frac{p(1-p)}{L}} \]

\( n_d \): Number of differences between the two sequences
\( L \): Number of nucleotides (or amino acids) compared
\( V(p) \): Variance or standard error of the mean (\( p \)) for binomial distribution

- Can be used for both nucleotide and amino acid substitutions

Distance estimation

Ancestral sequence?

ACTGAGGATCGC [Ancestral]

ACTGAGGATCGC

AATGAAAGATCGC

AATGAAAGATCGC

AATGAAAGATCGC

AATGAAAGATCGC

Distance estimation

AATGAGGATCGC [Ancestral]

ACTGAGGATCGC

AATGAAAGATCGC

ACTGAGGATCGC

AATGAAAGATCGC

AATGAAAGATCGC

Distance estimation

3 substitutions: No hidden substitution
Distance estimation for nucleotide substitutions

**Jukes-Cantor (one-parameter) method**

Jukes and Cantor (1969)

- **A**: Expected number of nucleotide substitutions per site or **Distance**
- **p**: Proportion of nucleotide differences (observed)
- **n**: number of differences, **l**: number of nucleotides compared

\[ k = -\frac{3}{4} \ln(1 - \frac{4}{3}p) \]

\[ V(l) = \frac{2p(1-p)}{(1-2p)^2} \]

\[ \phi(l) = \frac{3}{(3-4p)} \sqrt{\frac{l(1-p)}{l}} \]

**Without multiple-hit correction (p-distance):**

\[ p = \frac{n}{l} \]

- **n**: number of differences, **l**: number of nucleotides compared

- **With multiple-hit correction by Jukes-Cantor method:**

\[ k = -\frac{3}{4} \ln(1 - \frac{4}{3}p) \]

\[ p_c (Observed) \text{ proportion of nucleotide differences} = 0.214 \]

\[ k = -\frac{3}{4} \ln(1 - \frac{4}{3} \times 0.214) = 0.252 \]
Distance estimation for nucleotide substitutions

**Kimura two-parameter method** (Kimura 1980)

\[
Q = \frac{1}{2} \ln \left( \frac{1}{1 - 2Q} \right)
\]

**Distance estimation for nucleotide substitutions**

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>α</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>T</td>
<td>β</td>
<td>β</td>
<td>α</td>
</tr>
</tbody>
</table>

Distance estimation for nucleotide substitutions

\[
k = \frac{1}{2} \ln \left( \frac{1}{1 - 2P - Q} \right) + \frac{1}{4} \ln \left( 1 - 2Q \right)
\]

- **P**: Proportion of transitional (Ts) differences
- **Q**: Proportion of transversional (Tv) differences

- **Number of differences**: 3
- **Ts = 2, Tv = 1
- **Alignment length**: 14
- **Without multiple-hit correction (p-distance)**: 
  \[p = 0.214 \pm 0.110\]
- **Jukes-Cantor distance**:
  \[
k = -3/4 \ln(1 - 4x0.214/3) = 0.252 \pm 0.154
\]
- **Kimura 2-parameter distance**:
  \[
k = 2/14, Q = 1/14
\]
  \[k = 0.259 \pm 0.143\]