Application of Hidden Markov Models in Bioinformatics

Dirk Husmeier
Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK
Email: dirk@bioss.ac.uk
http://www.bioss.ac.uk/~dirk
Pairwise Sequence Alignment: Motivation

- Two DNA sequences:
  - Sequence 1 $\rightarrow$ ACGTTGCA
  - Sequence 2 $\rightarrow$ ACGTGCAT

- Direct alignment:
  - $\begin{array}{cccccccc}
  A & C & G & T & T & G & C & A \\
  A & C & G & T & G & C & A & T
  \end{array}$

- Alignment with gaps:
  - $\begin{array}{cccccccc}
  A & C & G & T & T & G & C & A \\
  A & C & G & T & - & G & C & A
  \end{array}$

- Interpretation:
  - Sequence 1 $\rightarrow$ Insertion
  - Sequence 2 $\rightarrow$ Deletion
DNA Sequence Alignment

Raw sequences

A  T  C  G  T  C  A  G  C  T  C
A  C  C  C  A  G  G  T  C

After alignment

A  T  C  G  T  C  A  G  C  T  C
A  C  C  C  C  –  –  A  G  G  T  C
HMM for Pairwise Sequence Alignment

The diagram represents the Hidden Markov Model (HMM) for pairwise sequence alignment, with transitions between states labeled with probabilities:

- Match: States \( x \) and \( y \) with transition probabilities:
  - \( P(M \rightarrow M) = 1 - 2\delta \)
  - \( P(M \rightarrow \epsilon) = \delta \)
  - \( P(\epsilon \rightarrow M) = 1 - \epsilon \)
- Insert: States \( x \) and \( \epsilon \) with transition probabilities:
  - \( P(I \rightarrow I) = \epsilon \)
- Delete: States \( \epsilon \) and \( y \) with transition probabilities:
  - \( P(D \rightarrow D) = \epsilon \)
Example: Pairwise Sequence Alignment with HMMs

T
T

C
G

T
A

C
A

T
G

T
A

C
T

-
Pairwise Sequence Alignment: Concept

- Hidden states: \( S_t \in \{ \mathcal{M}, \mathcal{I}, \mathcal{D} \} \) or \( S_t \in \{ \mathcal{M}, \mathcal{I}_x, \mathcal{I}_y \} \)
- Observations: \( y_t \rightarrow (x_m, y_n) \)

<table>
<thead>
<tr>
<th>(xm)</th>
<th>C G T C A G - T</th>
</tr>
</thead>
<tbody>
<tr>
<td>(yn)</td>
<td>C C - - A G C T</td>
</tr>
<tr>
<td>(t)</td>
<td>1 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>(m)</td>
<td>1 2 3 4 5 6 6 7</td>
</tr>
<tr>
<td>(n)</td>
<td>1 2 2 2 3 4 5 6</td>
</tr>
</tbody>
</table>
Viterbi Algorithm for Pairwise Sequence Alignment

\[\gamma_{m,n}(\mathcal{M}) = P(x_m, y_n) \max \left[ (1 - 2\delta)\gamma_{m-1,n-1}(\mathcal{M}) \right. \]
\[\left. (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_x) \right. \]
\[\left. (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_y) \right]\]

\[\gamma_{m,n}(\mathcal{I}_x) = P(x_m) \max \left[ \delta\gamma_{m-1,n}(\mathcal{M}) \right. \]
\[\left. \epsilon\gamma_{m-1,n}(\mathcal{I}_x) \right]\]

\[\gamma_{m,n}(\mathcal{I}_y) = P(y_n) \max \left[ \delta\gamma_{m,n-1}(\mathcal{M}) \right. \]
\[\left. \epsilon\gamma_{m,n-1}(\mathcal{I}_y) \right]\]

\[\uparrow_{m,n}(\mathcal{M}) = \arg\max \left[ (1 - 2\delta)\gamma_{m-1,n-1}(\mathcal{M}) \right. \]
\[\left. (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_x) \right. \]
\[\left. (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_y) \right]\]

\[\uparrow_{m,n}(\mathcal{I}_x) = \arg\max \left[ \delta\gamma_{m-1,n}(\mathcal{M}) \right. \]
\[\left. \epsilon\gamma_{m-1,n}(\mathcal{I}_x) \right]\]

\[\uparrow_{m,n}(\mathcal{I}_y) = \arg\max \left[ \delta\gamma_{m,n-1}(\mathcal{M}) \right. \]
\[\left. \epsilon\gamma_{m,n-1}(\mathcal{I}_y) \right]\]
Viterbi Algorithm for Pairwise Alignment: Logarithmic Version

\[
V_{m,n}(\mathcal{M}) = \log P(x_m, y_n) + \max \left[ V_{m-1,n-1}(\mathcal{M}) + \log (1 - 2\delta) \\
V_{m-1,n-1}(\mathcal{I}_x) + \log (1 - \epsilon) \\
V_{m-1,n-1}(\mathcal{I}_y) + \log (1 - \epsilon) \right]
\]

\[
V_{m,n}(\mathcal{I}_x) = \log P(x_m) + \max \left[ V_{m-1,n}(\mathcal{M}) + \log \delta \\
V_{m-1,n}(\mathcal{I}_x) + \log \epsilon \\
V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \right]
\]

\[
V_{m,n}(\mathcal{I}_y) = \log P(y_n) + \max \left[ V_{m,n-1}(\mathcal{M}) + \log \delta \\
V_{m,n-1}(\mathcal{I}_x) + \log \epsilon \\
V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \right]
\]

\[
\uparrow_{m,n}(\mathcal{M}) = \arg \max \left[ V_{m-1,n-1}(\mathcal{M}) + \log (1 - 2\delta) \\
V_{m-1,n-1}(\mathcal{I}_x) + \log (1 - \epsilon) \\
V_{m-1,n-1}(\mathcal{I}_y) + \log (1 - \epsilon) \right]
\]

\[
\uparrow_{m,n}(\mathcal{I}_x) = \arg \max \left[ V_{m-1,n}(\mathcal{M}) + \log \delta \\
V_{m-1,n}(\mathcal{I}_x) + \log \epsilon \right]
\]

\[
\uparrow_{m,n}(\mathcal{I}_y) = \arg \max \left[ V_{m,n-1}(\mathcal{M}) + \log \delta \\
V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \right]
\]
The Full Probability of an Alignment: Summing over All Paths

- Recursion

\[
\alpha_{m,n}(M) = P(x_m, y_n) \sum \begin{bmatrix}
(1 - 2\delta)\alpha_{m-1,n-1}(M) \\
(1 - \epsilon)\alpha_{m-1,n-1}(I_x) \\
(1 - \epsilon)\alpha_{m-1,n-1}(I_y)
\end{bmatrix}
\]

\[
\alpha_{m,n}(I_x) = P(x_m) \sum \begin{bmatrix}
\delta\alpha_{m-1,n}(M) \\
\epsilon\alpha_{m-1,n}(I_x)
\end{bmatrix}
\]

\[
\alpha_{m,n}(I_y) = P(y_n) \sum \begin{bmatrix}
\delta\alpha_{m,n-1}(M) \\
\epsilon\alpha_{m,n-1}(I_y)
\end{bmatrix}
\]

- Termination

\[
P(D) = \alpha_{M,N}(M) + \alpha_{M,N}(I_x) + \alpha_{M,N}(I_y)
\]

- Initialisation

\[
\alpha_{0,0}(M) = 1
\]

\[
\alpha_{0,0}(I_x) = \alpha_{0,0}(I_y) = \alpha_{0,-1}(.) = \alpha_{-1,0}(.) = 0
\]

This corresponds to the prior probability \( P_0(M) = 1 - 2\delta, \ P_0(I_x) = P_0(I_y) = \delta \)
Significance of the Alignment

• Posterior probability of the best alignment: \( P(\hat{\pi}|D) = \frac{P(\hat{\pi},D)}{P(D)} \)

• Posterior probabilities are often very small.

• Comparison with random model:

\[
P(random, D) = \left[ (1 - \nu) \prod_{m=1}^{M} P(x_m) \right] \nu \left[ (1 - \nu) \prod_{n=1}^{N} P(y_n) \right] \nu
\]

\[
= \nu^2 (1 - \nu)^{(M+N)} \prod_{m=1}^{M} P(x_m) \prod_{n=1}^{N} P(x_n)
\]

• Log odds ratio: \( \log \left( \frac{P(\hat{\pi},D)}{P(random, D)} \right) \)
Viterbi Algorithm for Pairwise Alignment: Log-odds Version

\[ V_{m,n}(\mathcal{M}) = \log P(x_m, y_n) + \max \left[ \begin{array}{c}
V_{m-1,n-1}(\mathcal{M}) + \log(1 - 2\delta) \\
V_{m-1,n-1}(\mathcal{I}_x) + \log(1 - \epsilon) \\
V_{m-1,n-1}(\mathcal{I}_y) + \log(1 - \epsilon)
\end{array} \right] \]

\[ V_{m,n}(\mathcal{I}_x) = \log P(x_m) + \max \left[ \begin{array}{c}
V_{m-1,n}(\mathcal{M}) + \log \delta \\
V_{m-1,n}(\mathcal{I}_x) + \log \epsilon
\end{array} \right] \]

\[ V_{m,n}(\mathcal{I}_y) = \log P(y_n) + \max \left[ \begin{array}{c}
V_{m-1,n}(\mathcal{M}) + \log \delta \\
V_{m-1,n}(\mathcal{I}_y) + \log \epsilon
\end{array} \right] \]

Log-odds ratio: \( P(x_m), P(y_n) \) cancel out, \( S(x_m, y_n) = \frac{P(x_m, y_n)}{P(x_m)P(y_n)} \), appropriate definition of \( d, e \)

\[ \tilde{V}_{m,n}(\mathcal{M}) = S(x_m, y_n) + \max \left[ \begin{array}{c}
\tilde{V}_{m-1,n-1}(\mathcal{M}) \\
\tilde{V}_{m-1,n-1}(\mathcal{I}_x) \\
\tilde{V}_{m-1,n-1}(\mathcal{I}_y)
\end{array} \right] \]

\[ \tilde{V}_{m,n}(\mathcal{I}_x) = \max \left[ \begin{array}{c}
\tilde{V}_{m-1,n}(\mathcal{M}) - d \\
\tilde{V}_{m-1,n}(\mathcal{I}_x) - e
\end{array} \right] \]

\[ \tilde{V}_{m,n}(\mathcal{I}_y) = \max \left[ \begin{array}{c}
\tilde{V}_{m,n-1}(\mathcal{M}) - d \\
\tilde{V}_{m,n-1}(\mathcal{I}_y) - e
\end{array} \right] \]
Profile HMMs: Motivation

- Functional biological sequences come in families.
- Evolution, speciation $\rightarrow$ primary sequences have diverged from each other.
- However, they maintain the same or a related structure/function $\rightarrow$ homologous sequences.
Profile HMMs: Typical Problems

- **Classification**
  Does a given biological sequence belong to a certain family, e.g., is a given protein sequence a globin?

- **Database search**
  Given a set of sequences → find more sequences of the same family.

- **Multiple sequence alignment**
Profile HMMs: Improvement over Pairwise Alignment

- Family of sequences: **Database search** for more members with **pairwise alignment**, known family members = query sequences.
- Repeat with all the known family members one by one.
- **Disadvantage:** May not find sequences distantly related to the ones you have already.
- **Improvement:** Use statistical features of the **whole set of sequences** in the search.
- Multiple sequence alignment: Concentrate on features that are **conserved in the whole family** → improvement over pairwise alignment.
Architecture of a Profile HMM

- Main states $M_k, k = 1, \ldots, K$
- Insert states $I_k, k = 1, \ldots, K$
- Delete states $D_k, k = 1, \ldots, K$
Profile HMMs

- Observations: $y_t \in \{A, C, G, T\}$, $t = 1, \ldots, N$
- Hidden states: $S_t \in \{M_k, I_k, D_k\}$; $k = 1, \ldots, K$

<table>
<thead>
<tr>
<th>Observation label</th>
<th>Main state</th>
<th>Insertion</th>
<th>Deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$t \rightarrow t + 1$</td>
<td>$t \rightarrow t + 1$</td>
<td>unchanged</td>
</tr>
<tr>
<td>State label</td>
<td>$k \rightarrow k + 1$</td>
<td>unchanged</td>
<td>$k \rightarrow k + 1$</td>
</tr>
</tbody>
</table>
Forward Algorithm for Profile HMMs

\[ \alpha_t(M_k) = P(y_t|M_k) \sum \begin{bmatrix} P(M_k|M_{k-1})\alpha_{t-1}(M_{k-1}) \\ P(M_k|I_{k-1})\alpha_{t-1}(I_{k-1}) \\ P(M_k|D_{k-1})\alpha_{t-1}(D_{k-1}) \end{bmatrix} \]

\[ \alpha_t(I_k) = P(y_t|I_k) \sum \begin{bmatrix} P(I_k|M_k)\alpha_{t-1}(M_k) \\ P(I_k|I_k)\alpha_{t-1}(I_k) \\ P(I_k|D_k)\alpha_{t-1}(D_k) \end{bmatrix} \]

\[ \alpha_t(D_k) = \sum \begin{bmatrix} P(D_k|M_{k-1})\alpha_t(M_{k-1}) \\ P(D_k|I_{k-1})\alpha_t(I_{k-1}) \\ P(D_k|D_{k-1})\alpha_t(D_{k-1}) \end{bmatrix} \]
Viterbi Algorithm for Profile HMMs

\[ \gamma_t(M_k) = P(y_t|M_k) \max \begin{bmatrix} P(M_k|M_{k-1})\gamma_{t-1}(M_{k-1}) \\ P(M_k|I_{k-1})\gamma_{t-1}(I_{k-1}) \\ P(M_k|D_{k-1})\gamma_{t-1}(D_{k-1}) \end{bmatrix} \]

\[ \gamma_t(I_k) = P(y_t|I_k) \max \begin{bmatrix} P(I_k|M_k)\gamma_{t-1}(M_k) \\ P(I_k|I_k)\gamma_{t-1}(I_k) \\ P(I_k|D_k)\gamma_{t-1}(D_k) \end{bmatrix} \]

\[ \gamma_t(D_k) = \max \begin{bmatrix} P(D_k|M_{k-1})\gamma_t(M_{k-1}) \\ P(D_k|I_{k-1})\gamma_t(I_{k-1}) \\ P(D_k|D_{k-1})\gamma_t(D_{k-1}) \end{bmatrix} \]

\[ \uparrow_t(M_k) = \text{argmax} \begin{bmatrix} P(M_k|M_{k-1})\gamma_{t-1}(M_{k-1}) \\ P(M_k|I_{k-1})\gamma_{t-1}(I_{k-1}) \\ P(M_k|D_{k-1})\gamma_{t-1}(D_{k-1}) \end{bmatrix} \]

\[ \uparrow_t(I_k) = \text{argmax} \begin{bmatrix} P(I_k|M_k)\gamma_{t-1}(M_k) \\ P(I_k|I_k)\gamma_{t-1}(I_k) \\ P(I_k|D_k)\gamma_{t-1}(D_k) \end{bmatrix} \]

\[ \uparrow_t(D_k) = \text{argmax} \begin{bmatrix} P(D_k|M_{k-1})\gamma_t(M_{k-1}) \\ P(D_k|I_{k-1})\gamma_t(I_{k-1}) \\ P(D_k|D_{k-1})\gamma_t(D_{k-1}) \end{bmatrix} \]
## Parameter Estimation for Known State Sequences

- $y_t^i$: Observation at position $t$ in the $i$th sequence
- $S_t^i$: Hidden state at position $t$ in the $i$th sequence
- $o$: Label for observations. Dice: $o = 1, \ldots, 6$. DNA: $o = 1, \ldots, 4$. Proteins: $o = 1, \ldots, 20$.
- $h$: Label for hidden states. Rogue casino: $h = 1, 2$. Profile HMM of length $K$: $h = 1, \ldots, 3K + 2$.
- $y_t^i = o$: At the $t$th site in the $i$th sequence the $o$th emission symbol is observed.
- $S_t^i = h$: At the $t$th site in the $i$th sequence the hidden state takes on the $h$th state symbol.
- $N(o|h)$: Number of times the $o$th observation symbol is emitted from the $h$th state.
- $N(h'|h)$: Number of state transitions from the $h$th state to the $h'$th state.
- $D$: Data= set of all observations = $\{y_t^i\}_{t,i}$, e.g.: a given set of DNA sequences.
- $\Psi$: Set of all hidden state sequences, e.g.: annotated alignment of DNA sequences.

\[
P(D, \Psi) = \prod_i \prod_t P(y_t^i | S_t^i) P(S_t^i | S_{t-1}^i) \\
= \prod_o \prod_h P(y_t = o | S_t = h)^{N(o|h)} \prod_h \prod_{h'} P(S_t = h' | S_{t-1} = h)^{N(h'|h)}
\]


Parameter Estimation for Known State Sequences

\[ P(D, \Psi) = \prod_o \prod_h P(y_t = o|S_t = h)^{N(o|h)} \prod_h \prod_{h'} P(S_t = h'|S_{t-1} = h)^{N(h'|h)} \]

- **\( P_E(o|h) \)**  Emission probability = probability that the \( h \)th hidden state emits the \( o \)th observable symbol.
- **\( P_T(h'|h) \)**  Transition probability = probability of a transition from the \( h \)th hidden state into the \( h' \)th hidden state.
- **\( w \)**  Vector of model parameters = vector of all transition and emission probabilities.

\[ P(D, \Psi|w) = \prod_o \prod_h P_E(o|h)^{N(o|h)} \prod_h \prod_{h'} P_T(h'|h)^{N(h'|h)} \]

\[ \ln P(D, \Psi|w) = \sum_o \sum_h N(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} N(h'|h) \ln P_T(h'|h) \]

Maximum likelihood estimate:

\[ P_E(o|h) = \frac{N(o|h)}{\sum_{o'} N(o'|h)} \quad P_T(h'|h) = \frac{N(h'|h)}{\sum_{h''} N(h''|h)} \]
Parameter Estimation for Unknown State Sequences: EM Algorithm

\[ L(w) = \ln P(D|w) = \ln \sum_\Psi P(D, \Psi|w) \]

\[ F(w) = \sum_\Psi Q(\Psi) \ln \frac{P(D, \Psi|w)}{Q(\Psi)} = \sum_\Psi Q(\Psi) \ln \frac{P(\Psi|D, w)}{Q(\Psi)} + \ln P(D|w) \]

\[ L(w) = F(w) + KL[Q, P] \geq F(w) \]

\[ L \quad F \]

\[ \text{E-step} \quad \text{M-step} \]

\[ \text{E-step} \rightarrow Q(\Psi) = P(\Psi|D, w) \]

\[ \text{M-step} \rightarrow \text{Maximise } F(w) \]
Parameter Estimation for Unknown State Sequences

- $P_E(o|h)$: Emission probability
- $P_T(h'|h)$: Transition probability
- $w$: Vector of model parameters = vector of all transition and emission probabilities.
- $N_{\Psi}(o|h)$: Number of times the $o$th observation symbol is emitted from the $h$th state for known state sequences $\Psi$.
- $N_{\Psi}(h'|h)$: Number of state transitions from the $h$th state symbol to the $h'$th state symbol for known state sequences $\Psi$.

$$
\ln P(D, \Psi|w) = \sum_o \sum_h N_{\Psi}(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} N_{\Psi}(h'|h) \ln P_T(h'|h)
$$

$$
F(w) = \sum_{\Psi} Q(\Psi) \ln P(D, \Psi|w) + C
$$

$$
= \sum_o \sum_h \overline{N}(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} \overline{N}(h'|h) \ln P_T(h'|h) + C
$$

$$
\overline{N}(o|h) = \sum_{\Psi} N_{\Psi}(o|h)Q(\Psi) \quad \overline{N}(h'|h) = \sum_{\Psi} N_{\Psi}(h'|h)Q(\Psi)
$$
Parameter Estimation for Unknown State Sequences

\[ F(\mathbf{w}) = \sum_o \sum_h \overline{N}(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} \overline{N}(h'|h) \ln P_T(h'|h) + C \]

**M-step:** \( F(\mathbf{w}) \) is maximised for

\[ P_E(o|h) = \frac{\overline{N}(o|h)}{\sum_o \overline{N}(o'|h)} \quad P_T(h'|h) = \frac{\overline{N}(h'|h)}{\sum_{h''} \overline{N}(h''|h)} \]

**E-step:** \( Q(\Psi) \rightarrow P(\Psi|D, \mathbf{w}) \)

\[ \overline{N}(o|h) = \sum_{\Psi} N_{\Psi}(o|h) P(\Psi|D, \mathbf{w}) = \sum_{\Psi} \left[ \sum_t \sum_i \delta(y^i_t, o) \delta(S^i_t, h) \right] P(\Psi|D, \mathbf{w}) \]
\[ = \sum_t \sum_i \delta(y^i_t, o) P(S^i_t = h|D, \mathbf{w}) \]

\[ \overline{N}(h'|h) = \sum_{\Psi} N_{\Psi}(h'|h) P(\Psi|D, \mathbf{w}) = \sum_{\Psi} \left[ \sum_t \sum_i \delta(S^i_t, h') \delta(S^i_{t-1}, h) \right] P(\Psi|D, \mathbf{w}) \]
\[ = \sum_t \sum_i P(S^i_t = h', S^i_{t-1} = h|D, \mathbf{w}) \]