CSE 549
Lecturer: Sael Lee

AMINO ACID SEQUENCE ALIGNMENT II

Slides provided by courtesy of Dr. D. Kihara @ Purdue
REF: Biological sequence analysis: Probabilistic models of proteins and nucleic acids Richard Durbin et al.

Slides by SNU BioIntelligence Lab. (http://bi.snu.ac.kr)
Slides by D. Kihara @ Purdue
PROFILE HMM

- An HMM which model a multiple sequence alignment of a protein family
- Concentrate on features that are conserved in the whole family (consensus modeling):
  + Improves alignment of distantly related sequence of the same family.
  + Able to characterize the family.
Let’s start by considering only the ungapped regions

Probability of new seq. $x$ according to the emission probabilities $e_i(a)$ and assuming independence between positions.

$$P(x \mid M) = \prod_{i=1}^{L} e_i(x_i)$$

Log-odd ratio for testing for membership in the family:

$$S = \sum_{i=1}^{L} \log \frac{e_i(x_i)}{q_{x_i}}$$
**UNGAPPED**

- **Match states**
  - Emission probabilities of observing AA $a$ in position $M_i$
    \[ e_{M_i}(a) \]
  - Transition probabilities all 1

![Diagram showing process](attachment:image.png)
**ADD INSERTIONS**

- Introduce insert states $I_i$
  - Emission prob. $e_{I_i}(a)$
    - Normally set to equal background distribution $q_a$
  - Transition prob. For
    - $M_i$ to $I_i$
    - $I_i$ to itself (multiple insertion)
    - $I_i$ to $M_{i+1}$
  - Log-odds score of a gap of length $k$
    - Assuming that $e_{I_i}(a) = q_a$, there is no log-odds from emission
      \[
      \log a_{M_jI_j} + \log a_{I_jM_{j+1}} + (k - 1) \log a_{I_jI_j}
      \]
ADD DELETION

- Introduce delete states (silent state)
  - No emission prob.
  - Cost of a deletion sum cost of
    - $M \rightarrow D$ transition
    - $D \rightarrow D$ transitions
    - $D \rightarrow M$ transition
  - Each $D \rightarrow D$ might be different prob. Unlike $I \rightarrow I$ that have same prob.
COMPONENTS OF PROFILE HMMS (5)

- Combining all parts

Figure 5.2 The transition structure of a profile HMM.

Improbable but can create problem when left out.
DERIVING PROFILES HMM FROM MSA

Assume correct multiple seq. alignment is given

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HBA_HUMAN</td>
<td>...VGA--HAGEY...</td>
<td></td>
</tr>
<tr>
<td>HBB_HUMAN</td>
<td>...V----NVDEV...</td>
<td></td>
</tr>
<tr>
<td>MYG_PHYCA</td>
<td>...VEA--DVAGH...</td>
<td></td>
</tr>
<tr>
<td>GLB3_CHITP</td>
<td>...VKG------D...</td>
<td></td>
</tr>
<tr>
<td>GLB5_PETMA</td>
<td>...VYS--TYETS...</td>
<td></td>
</tr>
<tr>
<td>LGB2_LUPLU</td>
<td>...FNA--NIPKH...</td>
<td></td>
</tr>
<tr>
<td>GLB1_GLYDI</td>
<td>...IAGADNGAGV...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>*****</td>
</tr>
</tbody>
</table>

Figure 5.3 Ten columns from the multiple alignment of seven globin protein sequences shown in Figure 5.1 The starred columns are ones that will be treated as ‘matches’ in the profile HMM.
HMMS FROM MULTIPLE ALIGNMENTS

- Basic profile HMM parameterization
  - Aim: generate distribution peak around members of the family

- Parameters
  - Probabilities values: various ways to do it but let assume independent samples aligned independently to the HMM

\[ a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad \quad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')} \]

- Length of the model: heuristics or systematic way
  - Deciding which MSA columns to assign to match states and which to insert states.
  - One Heuristics: columns that are more than half gap should be modelled buy inserts.
Main usage of profile HMMs

- Detecting potential membership in a family
- By (global) matching a sequence to the profile HMMs
- Scoring a match:
  - Viterbi equations – gives h most probable alignment of a seq together with its probability
  - Forward equation – calculates the full probabilities of seq summed overall possible paths.
- Either case, what we want is the log-odd ratio $x$ being the family compared to the random model

$$P(x \mid R) = \prod_{i} q_{x_i}$$
Many state paths generate the same symbol sequence

Choose the highest probability path, $\pi^*$ for a sequence:

$$
\pi^* = \operatorname{argmax}_\pi P(x, \pi)
$$

**Initialization:**

\[ 
V_0^M(0) = 0; \quad V_{j>0}^M(0) = -\infty; \quad V_0^M(i > 0) = -\infty; \\
V_j^I(0) = -\infty; \\
V_0^D(i) = -\infty. 
\]

**Termination:**

\[ 
V = \max[V_L^M(N), V_L^I(N), V_L^D(N)] 
\]
let $V_j^M(i)$ be the log-odds score of the best path matching subseq. $x_1....x_i$ to the submodel up to state $j$ ending with $x_i$ being emitted by state $M_j$

$$V_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \max \left\{ V_{j-1}(i-1) + \log a_{M_{j-1}M_j}, V_{j-1}(i-1) + \log a_{I_{j-1}M_j}, V_{j-1}(i-1) + \log a_{D_{j-1}M_j} \right\}$$
Let $V_j^I(i)$ be the score of the best path ending in $x_i$ being emitted by $I_j$

$$V_j^I(i) = \log \frac{e_{I_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_j^M(i-1) + \log a_{M_j I_j}, \\ V_j^I(i-1) + \log a_{I_j I_j}, \\ V_j^D(i-1) + \log a_{D_j I_j}; \end{cases}$$

Can be removed: No emission score set to equal background distribution in $I$.

Not likely to happen.
Let $V_j^D(i)$ be the best path ending in state Dj.

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i) + \log a_{M_{j-1}D_j}, \\
V_{j-1}^I(i) + \log a_{I_{j-1}D_j}, & \text{Not likely to happen} \\
V_{j-1}^D(i) + \log a_{D_{j-1}D_j}; \end{cases}$$
The best path visiting state $k$ at time $i$ extends best path ending at state $l$ at time $i+1$: recursive calculation (DP can be used!)

\[ v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl}) \]
**ALGORITHM: VITERBI**

Initialisation ($i = 0$): $v_0(0) = 1$, $v_k(0) = 0$, for $k > 0$

Recursion ($i = 1..L$):

$v_l(i) = e_l(x_i) \max_k (v_k(i-1)a_{kl});$

tracing path: $ptr_i(l) = \arg \max_k (v_k(i-1)a_{kl})$

Termination:

$P(x, \pi^*) = \max_k (v_k(L)a_{k0});$

$\pi^*_L = \arg \max_k (v_k(L)a_{k0})$
FORWARD ALGORITHM

- Given an HMM, what is the probability of obtaining sequence $x$? (I don’t care which path)

$$P(x) = \sum_{\pi} P(x, \pi)$$

- Suppose the probability of the observed sequence up to and including $x_i$, requiring that $p_i = k$ is known, $f_k(i) = P(x_1 x_2 \ldots x_i, \pi_i = k)$. Then recursively,

$$f_l(i + 1) = e_l(x_{i+1}) \sum_k f_k(i) a_{kl}$$
SEARCHING WITH PROFILE HMMS: FORWARD ALGORITHM

\[ F_j^M (i) = \log \frac{e_{M_j} (x_i)}{q_{x_i}} + \log [a_{M_{j-1}M_j} \exp(F_{j-1}^M (i-1)) \]
\[ + a_{I_{j-1}M_j} \exp(F_{j-1}^I (i-1)) + a_{D_{j-1}M_j} \exp(F_{j-1}^D (i-1))] ; \]

\[ F_j^I (i) = \log \frac{e_{I_j} (x_i)}{q_{x_i}} + \log [a_{M_{j-I_j}} \exp(F_{j-I}^M (i-1)) \]
\[ + \log a_{I_{j-I_j}} \exp(F_{j-I}^I (i-1)) + a_{D_{j-I_j}} \exp(F_{j-I}^D (i-1))] ; \]

\[ F_j^D (i) = \log [a_{M_{j-D_j}} \exp(F_{j-D}^M (i)) + \log a_{I_{j-D_j}} \exp(F_{j-D}^I (i)) \]
\[ + a_{D_{j-D_j}} \exp(F_{j-D}^D (i))] ; \]

Initialization:

\[ V_0^M (0) = 0; \quad V_{j>0}^M (0) = -\infty; \quad V_0^M (i > 0) = -\infty; \quad F = \log [\exp(F_{L}^M (N)) + \exp(F_{L}^I (N)) + \exp(F_{L}^D (N))] \]

\[ V_0^I (0) = -\infty; \]

Termination:

\[ V_0^D (i) = -\infty. \]
ALGORITHM: FORWARD ALGORITHM

Initialisation \((i = 0)\):

\[
f_0(0) = 1, \; f_k(0) = 0 \text{ for } k > 0
\]

Recursion \((i = 1..L)\):

\[
f_i(i) = e_l(x_i) \sum_k f_k(i - 1) a_{kl}
\]

Termination: \(P(x) = \sum_k f_k(L) a_{k0}\)
Given an observed sequence, the probability that $x_i$ came from state $k$: (posterior probability)

$$P(\pi_i = k \mid x)$$

Ex. Given an observation of the sequence of rolls, the probability that the die was loaded at the $i$-th roll


\[ P(x, \pi_i = k) = P(x)P(\pi_i = k \mid x) \]

\[ \therefore P(\pi_i = k \mid x) = \frac{P(x, \pi_i = k)}{P(x)} \]

\[
P(x, \pi_i = k) = P(x_1x_2\ldots x_i, \pi_i = k)P(x_{i+1}x_{i+2}\ldots x_L \mid x_1x_2\ldots x_i, \pi_i = k)
\]

\[ = P(x_1x_2\ldots x_i, \pi_i = k)P(x_{i+1}x_{i+2}\ldots x_L \mid \pi_i = k) \]

\[ = f_k(i)P(x_{i+1}x_{i+2}\ldots x_L \mid \pi_i = k) \]

- Let’s write \[ P(x_{i+1}x_{i+2}\ldots x_L \mid \pi_i = k) = b_k(i) \] and calculate it recursively
**Backward Algorithm**

$b_k(i)$: the probability that the state $k$ is used for the step $i$, and the sequences $x_{i+1} \ldots X_L$ is observed

$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$
**Algorithm: Backward Algorithm**

**Initialisation** \((i = L)\):

\[
b_k(L) = a_{k0}, \text{ for all } k
\]

**Recursion** \((i = L - 1 .. 1)\):

\[
b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)
\]

**Termination**: \(P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)\)
PARAMETER ESTIMATION FOR HMMS

- Use a set of example sequences (training sequences)
- When the paths are known for all the examples:
  - E.g. CpG Island, protein secondary structure

\[ a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \quad \ldots \quad (A) \]

where \( A_{kl}, E_k(b) \): # of counts in the training data

- Pseudo-counts: \( r_{kl}, r_k(b) \)
  - \( A_{kl} = \# \text{ of transitions } k \text{ to } l \text{ in the training data} + r_{kl} \)
  - \( E_k(b) = \# \text{ of emissions of } b \text{ from } k \text{ in the training data} + r_k(b) \)
Example of parameter estimation

AACACCTCATCTTAGCC
aagaggagaaagagg

Pseudocount: 1 for $E$ and $a$

Transition Probabilities:
- $\#\alpha\alpha$: 3; $\#\alpha\gamma$: 5; $\#\gamma\alpha$: 4; $\#\gamma\gamma$: 2
- $a_{\alpha\alpha} = (5+1)/(5+1+3+1) = 6/10$
- $a_{\alpha\gamma} = (3+1)/(3+1+5+1) = 4/10$
- $a_{\gamma\alpha} = (4+1)/(4+1+2+1) = 5/8$
- $a_{\gamma\gamma} = (2+1)/(4+1+2+1) = 3/8$

Emission Probabilities:
- $E_\alpha(G)$: 1; $E_\alpha(C)$: 2; $E_\alpha(A)$: 3; $E_\alpha(T)$: 2
- $e_\alpha(G) = (1+1)/(8+4) = 2/12$
- $e_\alpha(C) = (2+1)/(8+4) = 3/12$
- $e_\alpha(A) = (3+1)/(8+4) = 4/12$
- $e_\alpha(T) = (2+1)/(8+4) = 3/12$