EECS 730
Introduction to Bioinformatics
Sequence Alignment

Luke Huan
Electrical Engineering and Computer Science
http://people.eecs.ku.edu/~jhuan/
HMM

- $\prod_i$ is a set of states
- Transition Probabilities
  \[ a_{kl} = \Pr(\pi_i = l \mid \pi_{i-1} = k) \]
  - Probability of transition from state $k$ to state $l$
- Emission Probabilities
  \[ e_k(b) = \Pr(x_i = b \mid \pi_i = k) \]
  - Probability of emitting character $b$ in state $k$
- HMM topology
  - A fully connected graph (i.e. clique) contains too many parameters
HMM

- $\Pi_i = \{ S_j, \text{begin}, \text{end} \} \quad j = [1,2]$

$$a_{kl} = \Pr(\pi_i = l | \pi_{i-1} = k) \quad e_k(b) = \Pr(x_i = b | \pi_i = k)$$
Components of profile HMMs

- From bioinformatics

The transition structure of a profile HMM.
Trivial questions:

- What is the probability that we will observe the state-path (path) $b, S_1, S_2, e$?
- Given a path $b, S_1, S_2, e$, what is the probability that we will observe the sequence “01”?
Slightly involved questions:

- What is the probability that we will observe the sequence “01” by going through the path $b, S_1, S_2, e$?
- What is the probability that we will observe the sequence “01” with $M$?
- What is the most likely path, when we observe the sequence “01” from $M$?
A hard question:

- Given a set of sequences (assuming they are generated by a HMM), how do we estimate the parameters (and the structure) of the related HMM?
Why do we care? Assign membership

- Given a HMM $M$, building from a protein family $P$, and a new sequence $s$, the probability $P(s|M)$ tells us how likely the sequence $s$ belongs to $P$ and hence have the same function as proteins in $P$.
- Questions: what if we have two families $P_1$ and $P_2$ and we are not sure which family I should assign the sequence to?
Why do we care? Find the alignment

- Given a HMM $M$, building from a protein family $P$, and a new sequence $s$, the most likely path of events $T = \max P(s|P)$ ($P$ is a valid path in $M$) tells us how should we align $s$ to $M$. 
Why do we care? Build a HMM

- Given a set of protein sequences $S$, build the HMM that mostly likely generates $S$. 
Three Important Questions

- How likely is a given sequence?
  - The Forward algorithm
- What is the most probable “path” for generating a given sequence?
  - The Viterbi algorithm
- How can we learn the HMM parameters given a set of sequences?
  - The Forward-Backward (Baum-Welch) algorithm
Searching with profile HMMs

- Main usage of profile HMMs
  - Detecting potential membership in a family
  - Matching a sequence to the profile HMMs
  - Viterbi algorithm
    - Based on Dynamic Programming
  - Maintaining log-odd ratio compared with random model

$$P(x \mid R) = \prod_{i} q_{x_i}$$
Viterbi Algorithm

The best way to get to E is either:

- To go to N5 via the best way to it from S and then to E, or
- To go to N6 via the best way to it from S and then to E, or
- To go to N7 via the best way to it from S and then to E.

The best way to get to N5 is either:

- To go to N2 via the best way to it from S and then to N5
- etc., etc.,
- In practice:
  - Calculate best route to N1, then N2, N3, N4, N5, N6, N7 & E
Viterbi equation

\[ V^M_j(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \max \left\{ \begin{array}{l}
V^M_{j-1}(i-1) + \log a_{M_{j-1}M_j}, \\
V^I_{j-1}(i-1) + \log a_{I_{j-1}M_j}, \\
V^D_{j-1}(i-1) + \log a_{D_{j-1}M_j}; \\
\end{array} \right. \]

\[ V^I_j(i) = \log \frac{e_{I_j}(x_i)}{q_{x_i}} + \max \left\{ \begin{array}{l}
V^M_j(i-1) + \log a_{M_jI_j}, \\
V^I_j(i-1) + \log a_{I_jI_j}, \\
V^D_j(i-1) + \log a_{D_jI_j}; \\
\end{array} \right. = 0 \]

\[ V^D_j(i) = \max \left\{ \begin{array}{l}
V^M_{j-1}(i) + \log a_{M_{j-1}D_j}, \\
V^I_{j-1}(i) + \log a_{I_{j-1}D_j}, \\
V^D_{j-1}(i) + \log a_{D_{j-1}D_j}; \\
\end{array} \right. \]
Example Calculation

Best path to N1 scores
max\{0.1*0.1\} = 0.01 from S

Best path to N2 scores
max\{0.01 * 1, 0.2*1\} = 0.2 from S

Best path to N3 scores
max\{0.7*0.5, 0.01 *0.9 *0.5\} = 0.035 from S

Best path to N4 scores
max\{0.035 *0.1, 0.2 *0.1\} = 0.02 from N2

and so on...

As with Needleman-Wunsch, we must record the nodes from which the best path came
HMMs from multiple alignments

Key idea behind profile HMMs

- Use the same structure, with different transition and emission probabilities, to capture specific information about each position in the multiple alignment of the whole family.
- Model representing the consensus for the family.
- Not the sequence of any particular member.

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

Ten columns from the multiple alignment of seven globin protein sequences. The starred columns are ones that will be treated as ‘matches’ in the profile HMM.
Multiple alignment by profile HMM training-
Multiple alignment with a known profile HMM

Before we estimate a model and a multiple alignment simultaneously we consider the simpler problem of obtaining a multiple alignment from a known model.

When we have a multiple alignment and a model of a small representative set of sequences in a family, and we wish to use that model to align a large member of other family members altogether.
Multiple alignment by profile HMM training-

Multiple alignment with a known profile HMM

- We know how to align a sequence to a profile HMM-
  Viterbi algorithm

- Construction a multiple alignment just requires calculating a Viterbi alignment for each individual sequence.
  - Residues aligned to the same profile HMM match state are aligned in columns.
Multiple alignment by profile HMM training-
Multiple alignment with a known profile HMM

- Importance difference with other MSA programs
  - Viterbi path through HMM identifies inserts
  - Profile HMM does not align inserts
  - Other multiple alignment algorithms align the whole sequences.

- HMM doesn’t attempt to align residues assigned to insert states.
  - The insert state residues usually represent part of the sequences which are atypical, unconserved, and not meaningfully alignable.
  - This is a biologically realistic view of multiple alignment.
Example Alignment, given a learned HMM for 3 sequences

**ACSA**

- Best path: 
  - MSA so far: ACGA

**AST**

- Best path: 
  - MSA so far: ACSA A-ST

**ACCST**

- Best path: 
  - MSA so far: AC-SA A--ST ACCST
Another Example

**ATSA**

- Best path:
  - MSA so far: ATSA

**ACCA**

- Best path:
  - MSA so far:
    - AT-SA
    - A-CCA

**ACAST**

- Best path:
  - MSA so far:
    - AT--GA
    - A-C--CA
    - AC--AST
HMMs from multiple alignments

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

- Parameters
  - the probabilities values: emission probabilities, transition probabilities
  - length of the model: heuristics or systematic way
Training from an existing alignment

- Start with a predetermined number of states in your HMM.
- For each position in the model, assign a column in the multiple alignment that is relatively conserved.
- Emission probabilities are set according to amino acid counts in columns.
- Transition probabilities are set according to how many sequences make use of a given delete or insert state.

\[
a_{kl} = \frac{A_{kl}}{\sum_l A_{kl'}} \quad \quad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')}
\]
More on estimation of prob. (1)

- Maximum likelihood (ML) estimation
  - given observed freq. $c_{ja}$ of residue $a$ in position $j$.

$$e_{M_j}(a) = \frac{c_{ja}}{\sum_{a'} c_{ja'}}$$

- Problem of ML estimation
  - If observed cases are absent?
  - Specially when observed examples are somewhat few.
More on estimation of prob. (2)

- Simple pseudocounts
  - $q_a$: background distribution
  - $A$: weight factor

$$e_{M_j}(a) = \frac{c_{ja} + Aq_a}{A + \sum_{a'} c_{ja'}}$$

- Laplace’s rule: $Aq_a = 1$
A simple example

- Chose six positions in model.
- Highlighted area was selected to be modeled by an insert due to variability.
Profile HMM training from unaligned sequences

- Harder problem – estimating both a model and a multiple alignment from initially unaligned sequences.
  - Initialization: Choose the length of the profile HMM and initialize parameters.
  - Training: Estimate the model using the Baum-Welch algorithm or the Viterbi alternative.
  - Multiple Alignment: Align all sequences to the final model using the Viterbi algorithm and build a multiple alignment as described in the previous section.
Profile HMM training from unaligned sequences

- **Initial Model**
  - The only decision that must be made in choosing an initial structure for Baum-Welch estimation is the length of the model M.
  - A commonly used rule is to set M be the average length of the training sequence.
  - We need some randomness in initial parameters to avoid local maxima.
Find appropriate parameters

- Baum-Welch algorithm
  - Instance of EM (Expectation-Maximization) algorithms.

Flow of the B-W algorithm:

1. Set initial parameters at random.
2. Update parameters
3. Increase of likelihood < ε
   - yes: Output parameters
   - no: Update parameters

The update always increases the likelihood $P(X|\eta)$, where $X$ is a set of sequences.
Find appropriate parameters

- The Viterbi alternative
  - Start with a model whose length matches the average length of the sequences and with random emission and transition probabilities.
  - Align all the sequences to the model.
  - Use the alignment to alter the emission and transition probabilities.
  - Repeat. Continue until the model stops changing.
Multiple alignment by profile HMM training

- Avoiding Local maxima
  - Baum-Welch algorithm is guaranteed to find a LOCAL maxima.
    - Models are usually quite long and there are many opportunities to get stuck in a wrong solution.
  - Multidimensional dynamic programming finds global optima, but is not practical.
- Solution
  - Start again many times from different initial models.
  - Use some form of stochastic search algorithm, e.g. simulated annealing.
Profile HMM training from unaligned sequences

- **Advantages:**
  - You take full advantage of the expressiveness of your HMM.
  - You might not have a multiple alignment on hand.

- **Disadvantages:**
  - HMM training methods are local optimizers, you may not get the best alignment or the best model unless you’re very careful.
  - Can be alleviated by starting from a logical model instead of a random one.
Profile HMM Summary

● Advantages:
  ● Very expressive profiling method
  ● Transparent method: You can view and interpret the model produced
  ● A consistent theory behind gap and insertion scores
  ● Very effective at detecting remote homolog

● Disadvantages:
  ● Slow – full search on a database of 400,000 sequences can take 15 hours
  ● Have to avoid over-fitting and locally optimal models