Profile HMMs for Sequence Families

COMP 571
Luay Nakhleh, Rice University

1

Sequence Families

- Functional biological sequences typically come in families
- Sequences in a family have diverged during evolution, but normally maintain the same or a related function
- Thus, identifying that a sequence belongs to a family tells about its function

2

HMM Profile

- Consensus modeling of the family using a probabilistic model
- Built from a given multiple alignment (assumed to be correct)

3

Sequences from a Globin Family

Alignment of 7 globins

The 8 alpha helices are shown as A-H above the alignment

4

Ungapped Score Matrices

- A natural probabilistic model for a conserved region would be to specify independent probabilities $e_i(a)$ of observing amino acid a in position i
- The probability of a new sequence x according to this model is

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

5

Log-odds Ratio

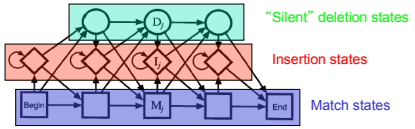
- We are interested in the ratio of the probability to the probability of x under the random model

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

Position specific score matrix (PSSM)

6

Adding Indels to Obtain a Profile HMM



Profile HMMs generalize pairwise alignment

7

Deriving Profile HMMs from Multiple Alignments

- Essentially, we want to build a model representing the consensus sequence for a family, rather than the sequence of any particular member
- Non-probabilistic profiles and profile HMMs

8

Non-probabilistic Profiles

- Gribskov, McLachlan, and Eisenberg 1987
- No underlying probabilistic model, but rather assigned position specific scores for each match state and gap penalty
- The score for each consensus position is set to the average of the standard substitution scores from all the residues in the corresponding multiple sequence alignment column

9

Non-probabilistic Profiles

10

```
HBA_HUMAN   ...VGA--HAGEY...
HBB_HUMAN   ...V----NVDEV...
MYG_PHYCA   ...VEA--DVAGH...
GLB3_CHITP  ...VKG------D...
GLB5_PETMA  ...VYS--TYETS...
LGB2_LUPLU  ...FNA--NIPKH...
GLB1_GLYDI  ...IAGADNGAGV...
              **  *****
```

The score for residue 'a' in column 1 → $\frac{5}{9}s(\mathrm{V},a)+\frac{1}{9}s(\mathrm{F},a)+\frac{1}{9}s(\mathrm{I},a)$

s(a,b) : standard substitution matrix

Non-probabilistic Profiles

11

- They also set gap penalties for each column using a heuristic equation that decreases the cost of a gap according to the length of the longest gap observed in the multiple alignment spanning the column

Problem with the Approach

12

- If we had an alignment with 100 sequences, all with a cysteine (C), at some position, the probability distribution for that column for an "average" profile would be exactly the same as would be derived from a single sequence
- Doesn't correspond to our expectation that the likelihood of a cysteine should go up as we see more confirming examples

Similar Problem with Gaps

13

```
HBA_HUMAN   ...VGA--HAGEY...
HBB_HUMAN   ...V----NVDEV...
MYG_PHYCA   ...VEA--DVAGH...
GLB3_CHITP  ...VKG------D...
GLB5_PETMA  ...VYS--TYETS...
LGB2_LUPLU  ...FNA--NIPKH...
GLB1_GLYDI  ...IAGADNGAGV...
              ***  *****
```

Scores for a deletion in columns 2 and 4 would be set to the same value

More reasonable to set the probability of a new gap opening to be higher in column 4

Basic Profile HMM Parameterization

14

- A profile HMM defines a probability distribution over the whole space of sequences
- The aim of parameterization is to make this distribution peak around members of the family
- Parameters: probabilities and the length of the model

Model Length

15

- A simple rule that works well in practice is that columns that are more than half gap characters should be modeled by inserts

Probability Values

16

$$a_{k\ell} = \frac{A_{k\ell}}{\sum_{\ell'} A_{k\ell'}} \qquad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')}$$

k, ℓ : indices over states

$a_{k\ell}, e_k(a)$: transition and emission probabilities

$A_{k\ell}, E_k(a)$: transition and emission frequencies

Problem with the Approach

17

- Transitions and emissions that don't appear in the training data set would acquire zero probability (would never be allowed)
- Solution: add pseudo-counts to the observed frequencies
- Simples pseudo-count is Laplace's rule: add one to each frequency

Example

18

```
HBA_HUMAN   ...VGA--HAGEY...
HBB_HUMAN   ...V----NVDEV...
MYG_PHYCA   ...VEA--DVAGH...
GLB3_CHITP  ...VKG------D...
GLB5_PETMA  ...VYS--TYETS...
LGB2_LUPLU  ...FNA--NIPKH...
GLB1_GLYDI  ...IAGADNGAGV...
               ***  *****
```

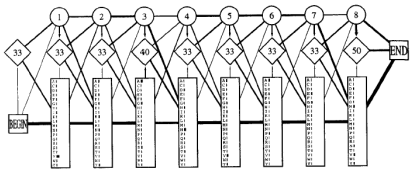
$e_{M_1}(\mathrm{V}) = 6/27, e_{M_1}(\mathrm{I}) = e_{M_1}(\mathrm{F}) = 2/27$

$e_{M_1}(a) = 1/27$ for all residue types a other than V, I, F

$a_{M_1M_2} = 7/10, a_{M_1D_2} = 2/10$ and $a_{M_1I_1} = 1/10$

Example: Full Profile HMM

19



Searching with Profile HMMs

20

- One of the main purposes of developing profile HMMs is to use them to detect potential membership in a family
- We can either use Viterbi algorithm to get the most probable alignment or the forward algorithm to calculate the full probability of the sequence summed over all possible paths

Viterbi Algorithm

21

$$V_j^{\mathrm{M}}(i) = \log \frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_{j-1}^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_{j-1}\mathrm{M}_j}; \end{cases}$$

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

$$V_j^{\mathrm{D}}(i) = \max \begin{cases} V_{j-1}^{\mathrm{M}}(i) + \log a_{\mathrm{M}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{I}}(i) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{D}}(i) + \log a_{\mathrm{D}_{j-1}\mathrm{D}_j}. \end{cases}$$

Forward Algorithm

$$
\begin{aligned}
F_j^{\mathrm{M}}(i) &= \log \frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \log\big[a_{\mathrm{M}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{M}}(i-1)\big) \\
&\quad + a_{\mathrm{I}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{I}}(i-1)\big) + a_{\mathrm{D}_{j-1}\mathrm{M}_j} \exp\big(F_{j-1}^{\mathrm{D}}(i-1)\big)\big]; \\
F_j^{\mathrm{I}}(i) &= \log \frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \log\big[a_{\mathrm{M}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{M}}(i-1)\big) \\
&\quad + a_{\mathrm{I}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{I}}(i-1)\big) + a_{\mathrm{D}_j\mathrm{I}_j} \exp\big(F_j^{\mathrm{D}}(i-1)\big)\big]; \\
F_j^{\mathrm{D}}(i) &= \log\big[a_{\mathrm{M}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{M}}(i)\big) + a_{\mathrm{I}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{I}}(i)\big) \\
&\quad + a_{\mathrm{D}_{j-1}\mathrm{D}_j} \exp\big(F_{j-1}^{\mathrm{D}}(i)\big)\big].
\end{aligned}
$$

22

Questions?

23