CISC 889 Bioinformatics (Spring 2004)

Hidden Markov Models (III)

a. Profile HMMs
b. GeneScan

Profile HMM for a family of sequences

Applications of HMM's

- Given a family of sequences, $\mathbf{O}^l=O^l_1...O^l_{K^l}$, build a hidden Markov model that best fits to this family-->Problem 3
  - Correct multiple alignment is given--> Problem 3, path known
    - MA built from structural information
    - MA obtained from other sequence based alignment procedures
  - Alignment is not assumed--> Problem 3, path not known (B-W)
- Use the obtained model to:
  - Score potential matches of new sequences-->Problem 1
  - Align new sequences--> Problem 2

Javier Garcia-Frias

CISC889, S04, Lec8, Liao

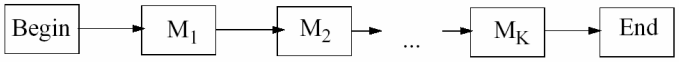
Profile HMM: Correct alignment assumed

HMM construction

Example: Assume MA given (columns marked with +)

```
A G - - -  C   O1
A GA G  -  C   O2
A -  C A C C   O3
- G L V -  C   O4
          →
  + +      +
```

- Segments of family where an alignment exists are produced by MATCH STATES



- Generation probabilities are position dependent!
- In previous example, K=3

Javier Garcia-Frias

CISC889, S04, Lec8, Liao

Profile HMM: Correct alignment assumed

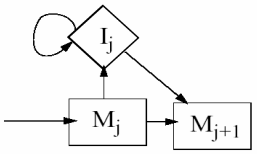
- Handling insertions: Portion of the sequences that are not aligned ---> Add INSERT STATES

Example: Assume MA given (columns marked with +)

```
A G - - -  C   O1
A GA G  -  C   O2
A -  C A C C   O3
- G L V -  C   O4
          →
  + +      +
```

- To cope with all possibilities for insertions, an insert state should be added after each match state

State I_k inserts sequence just after match state M_k (i.e., aligned column k)



O^1--> $M_1M_2M_3$
O^2--> $M_1M_2I_2I_2M_3$
O^3--> M_1? State M_2 is skipped

Javier Garcia-Frias

CISC889, S04, Lec8, Liao

Profile HMM: Correct alignment assumed

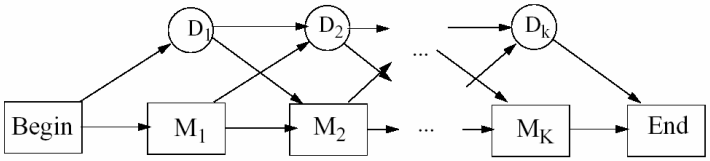
- Handling deletions: Portion of the sequences that "skips" the alignment---> Add SILENT (DELETE) STATES

Example: Assume MA given (columns marked with +)

```
A G - - -  C   O1
A GA G -  C   O2
A - C A C C   O3
- G L V -  C   O4
+ +       +
```

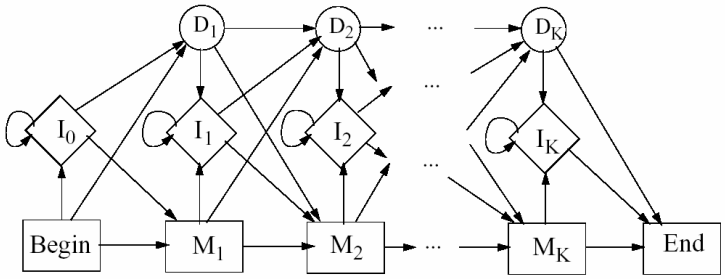
- To cope with all possibilities for deletions
  - Connect all possible match states (big complexity)
  - Add silent states (less complexity, but loss of generality)-->NO EMISSION

State D_k skips match state M_k (i.e., aligned column k)



Profile HMM: Correct alignment assumed

Resulting HMM (Profile HMM)



- Notice we have added transitions between insert and delete states

Example: Assume MA given (columns marked with +)

```
A G - - -  C   O1   M1M2M3
A GA G -  C   O2   M1M2I2I2M3
A - C A C C   O3   M1D2I2I2I2M3
- G L V -  C   O4   D1M2I2I2M3
+ +       +
```

Profile HMM: Correct alignment assumed

Key idea of profile HMM

- Transition and emission probabilities capture specific information about each position in the multiple alignment of the whole family
- Profile HMM=Statistical model representing the family

Questions

- How do we build the profile HMM that best fits to a given family? -->Problem 3 (simplified)
- How do we detect potential membership in this family (for new sequences)? --> Problem 1
- How do we align a new sequence? --> Problem 2

Javier Garcia-Frias

CISC889, S04, Lec8, Liao

Parameterization of profile HMM's: Correct alignment assumed

Profile HMM parametrization (simplified Problem 3)

- **Model length**
  - Length (and structure) completely defined when we decide which MA columns should be assigned to match states
    - Manual construction
    - Heuristic construction: e.g., column aligned if proportion of gaps is less than a threshold
    - More sophisticated methods
- **Parameter estimation**
  - Alignment is given-->Path through model is given for any sequence
  - Apply solution to Problem 3 when path is given (just count events)

Javier Garcia-Frias

CISC889, S04, Lec8, Liao

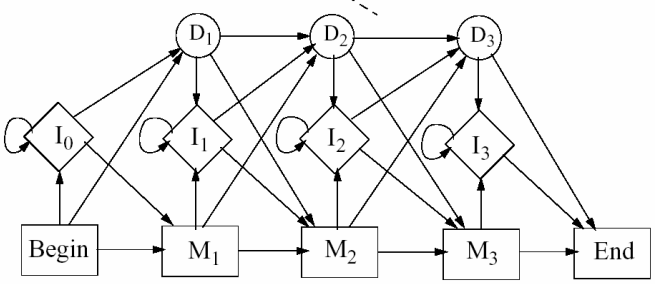
Parameterization of profile HMM's: Correct alignment assumed

Previous example

MA given
(columns marked with +)

A G - - - C O^1 $M_1M_2M_3$
A GA G - C O^2 $M_1M_2I_2I_2M_3$
A - C A C C O^3 $M_1D_2I_2I_2I_2M_3$
- G L V - C O^4 $D_1M_2I_2I_2M_3$

\+ + +



Javier Garcia-Frias

Parameterization of profile HMM's: Correct alignment assumed

Emission probabilities: Estimate from number of emissions

$N(A\|M_1)=3$ $N(other\|M_1)=0$	I_0, I_1, I_3 are not used
$N(A\|M_2)=3$ $N(other\|M_2)=0$	$N(A\|I_2)=2$ $N(C\|I_2)=2$ $N(G\|I_2)=1$
$N(C\|M_3)=4$ $N(other\|M_3)=0$	$N(L\|I_2)=1$ $N(V\|I_2)=1$ $N(other\|I_2)=0$

Transition probabilities: Estimate from number of transitions

$N(M_1\|B)=3$ $N(D_1\|B)=1$	$N(I_2\|D_2)=1$
$N(M_2\|M_1)=3$ $N(D_2\|M_1)=1$	$N(I_2\|I_2)=4$ $N(M_3\|I_2)=3$
$N(M_3\|M_2)=1$ $N(I_2\|M_2)=2$	
$N(E\|M_3)=3$	

- **If number of sequences is not high enough, estimation should be modified**

Javier Garcia-Frias

Membership in a profile HMM

Detection of potential membership, for a new sequence, in family defined by a profile HMM (Problem 1)

- Apply forward equation
- Since P(**O**|M) is length dependent, usually scoring function is modified

$$\text{Scoring}=\log\frac{P(\mathbf{O}|M)}{P(\mathbf{O}|S)}$$

S is called "standard model'': Model to use if sequences were independently distributed

- Other statistical approaches can also be used to improve the scoring system

Multiple alignment using profile HMM's

No alignment is assumed

- From an initially unaligned family of sequences, jointly perform:
  - Profile HMM estimation
  - Alignment estimation

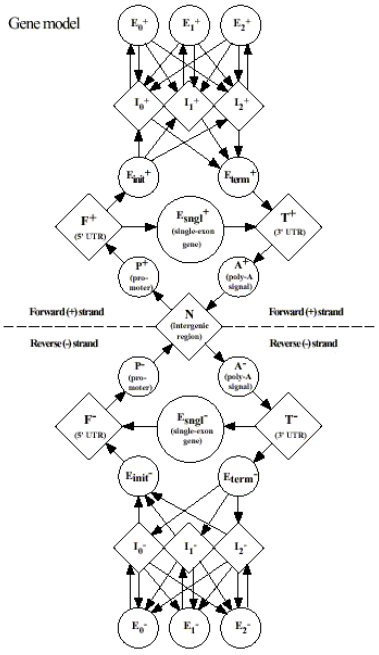
1. Initialization
- Choose length of profile HMM and initialize parameters

2. Training
- Estimate parameters of the profile HMM
- Path not known (no alignment)--> Problem 3 (Baum-Welch)

3. Alignment
- Align all sequences using Viterbi algorithm (Problem 2)

GENSCAN (generalized HMMs)

- Chris Burge, PhD Thesis '97, Stanford
- http://genes.mit.edu/GENSCAN.html
- Four components
  - A vector π of initial probabilities
  - A matrix T of state transition probabilities
  - A set of length distribution f
  - A set of sequence generating models P
- Generalized HMMs:
  - at each state, emission is not symbols (or residues), rather, it is a fragment of sequence.
  - Modified viterbi algorithm



- Initial state probabilities
  - As frequency for each functional unit to occur in actual genomic data. E.g., as ~ 80% portion are non-coding intergenic regions, the initial probability for state N is 0.80
- Transition probabilities
- State length distributions

- Training data
  - 2.5 Mb human genomic sequences
  - 380 genes, 142 single-exon genes, 1492 exons and 1254 introns
  - 1619 cDNAs

Open areas for research

- Model building
  - Integration of domain knowledge, such as structural information, into profile HMMs
  - Meta learning?
- Biological mechamism

  DNA replication
- Hybrid models
  - Generalized HMM
  - …