SUNY Korea
The State University of New York
한국뉴욕주립대학교

Stony Brook University

CSE 549

Lecturer: Sael Lee

AMINO ACID SEQUENCE ALIGNMENT II

Slides provided by courtesy of Dr. D. Kihara @ Purdue

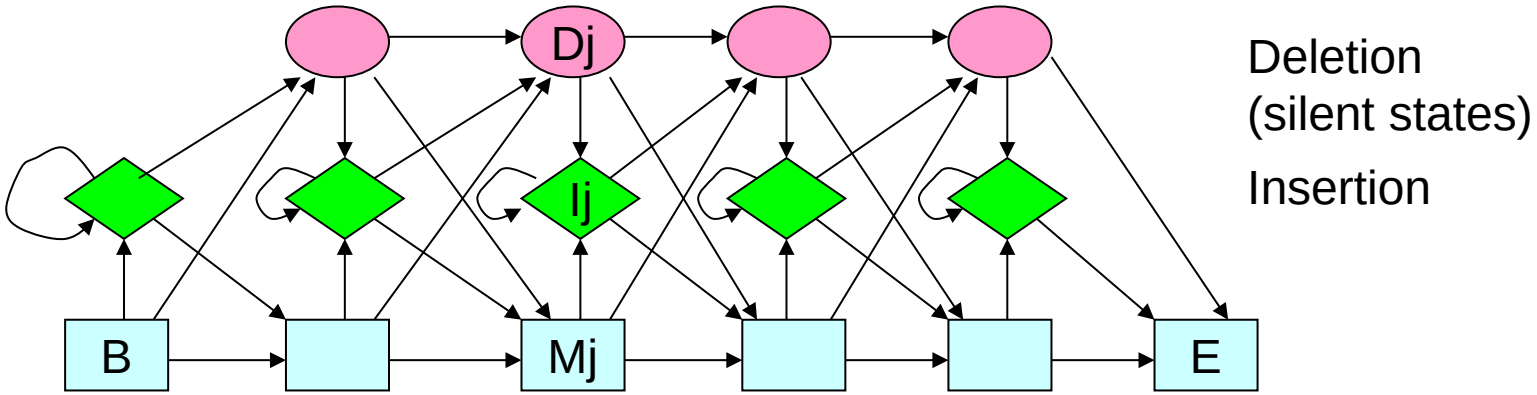
PROFILE HIDDEN MARKOV MODEL

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)
Sildes 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 wh ole family (consensus modeling):
  - Improves alignment of distantly related sequence of the sa me family.
  - Able to characterize the family.



UNGAPPED SCORE MATRICES

- Let's start by considering only the ungapped regions
- Probability of new seq. ***x*** according to the emission probabilities *ei(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:

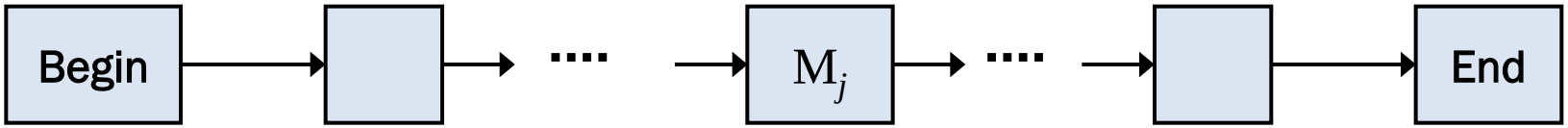
$$S = \sum_{i=1}^{L} \log \frac{e_i(x_i)}{q_{x_i}}$$

UNGAPPED

- Match states
  - Emission probabilities of observing AA *a* in position M*i*

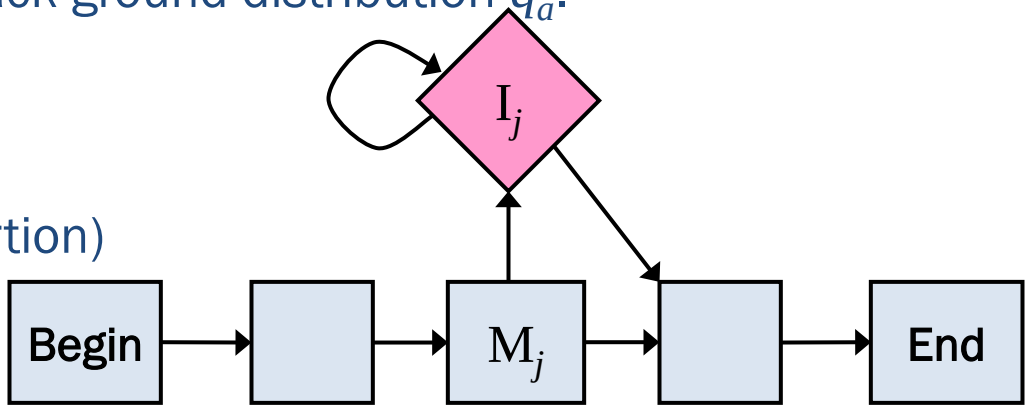
$$e_{M_i}(a)$$

  - Transition probabilities all 1



ADD INSERTIONS

- Introduce insert states I_i
  - Emission prob. $e_{I_i}(a)$
    - Normally set to equal back ground distribution q_a.
  - Transition prob. For
    - M_i to I_i,
    - I_i to itself (multiple insertion)
    - I_i to M_{i+1}

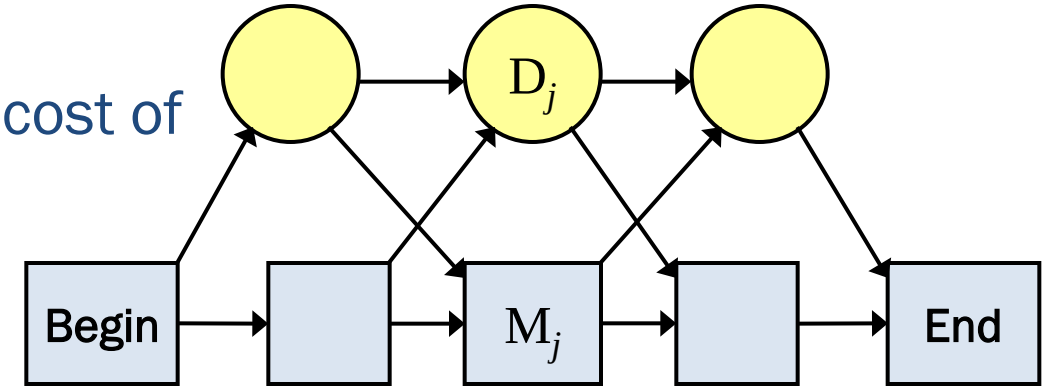


  - Log-odds score of a gap of length k
    - Assuming that $e_{I_i}(a) = q_{a\_}$ there is no logg-odds from emission

$$\log a_{M_j I_j} + \log a_{I_j M_j+1} + (k-1)\log a_{I_j I_j}$$

ADD DELETION

- Introduce delete states (silent state)
  - No emission prob.
  - Cost of a deletion sum cost of
    - M→D transition
    - D→D transitions
    - D→M transition



  - Each D→D might be different prob. Unlike I->I that have same prob.

COMPONENTS OF PROFILE HMMS (5)

- Combining all parts

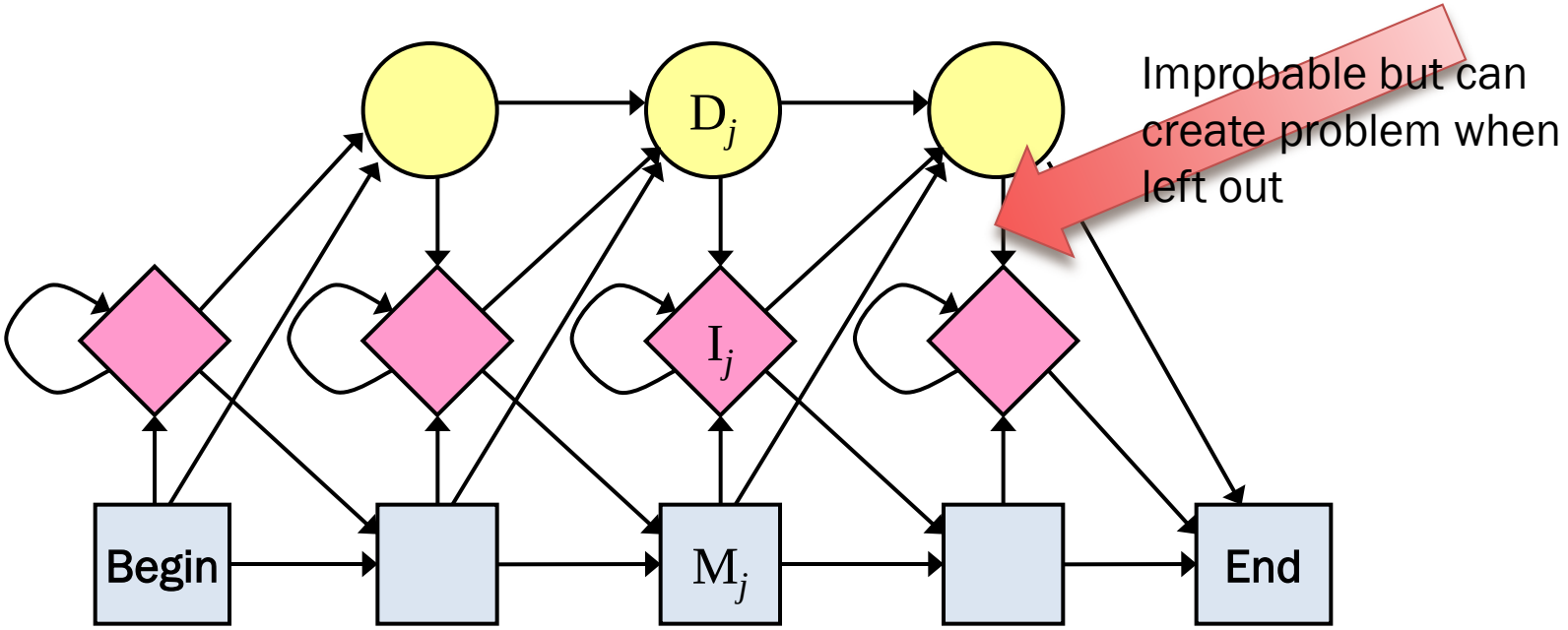


Figure 5.2 *The transition structure of a profile HMM.*

DERIVING PROFILES HMM FROM MSA

- Assume correct multiple seq. alignment is given

```
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...
                  ***   *****
```

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

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

  - 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.

SEARCHING WITH PROFILE HMMS (1)

- 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

$$P(x \mid R) = \prod_i q_{x_i}$$

DECODING THE MOST PROBABLE STATE PATH: THE VITERBI ALGORITHM

- Many state paths generate the same symbol sequence
- Choose the highest probability path, π^* for a sequence:

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

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:

$$V = \max\left[V_L^M(N), V_L^I(N), V_L^D(N)\right]$$

SEARCHING WITH PROFILE HMMS: VITERBI EQUATION

- let $V_j^{\mathrm{M}}(i)$ be the log-odds score of the best path matching subseq. x1….xi to the submodel up to state j ending with xi being emitted by state Mj

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

SEARCHING WITH PROFILE HMMS: VITERBI EQUATION

- Let $V_j^{\mathrm{I}}(i)$ be the score of the best path ending in xi be ing emittedb y Ij

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

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

Not likely to happen

SEARCHING WITH PROFILE HMMS: VITERBI EQUATION

- Let $V_j^{\mathrm{D}}(i)$ be the best path ending in state Dj

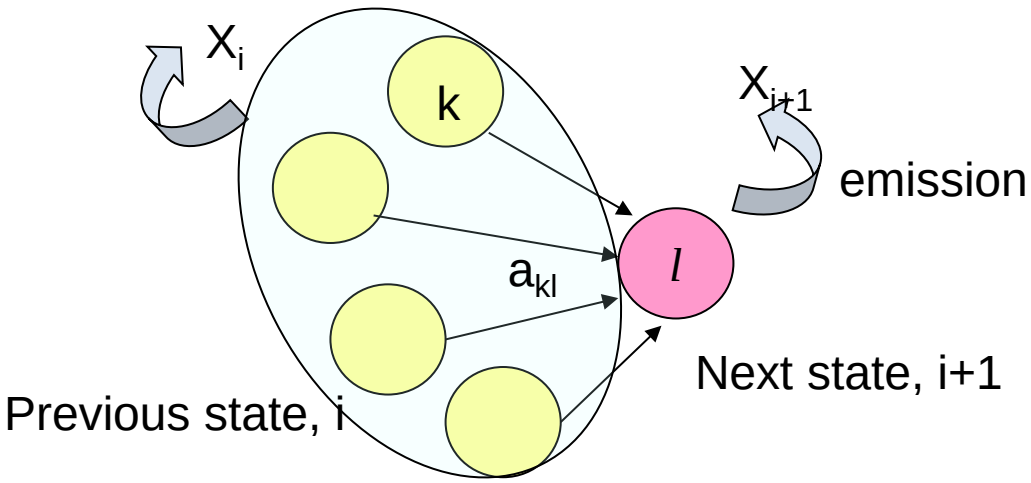
$$V_j^{\mathrm{D}}(i) = \max \begin{cases} V_{j-1}^{\mathrm{M}}(i) + \log a_{\mathrm{M}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{I}}(i) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{D}}(i) + \log a_{\mathrm{D}_{j-1}\mathrm{D}_j}; \end{cases}$$

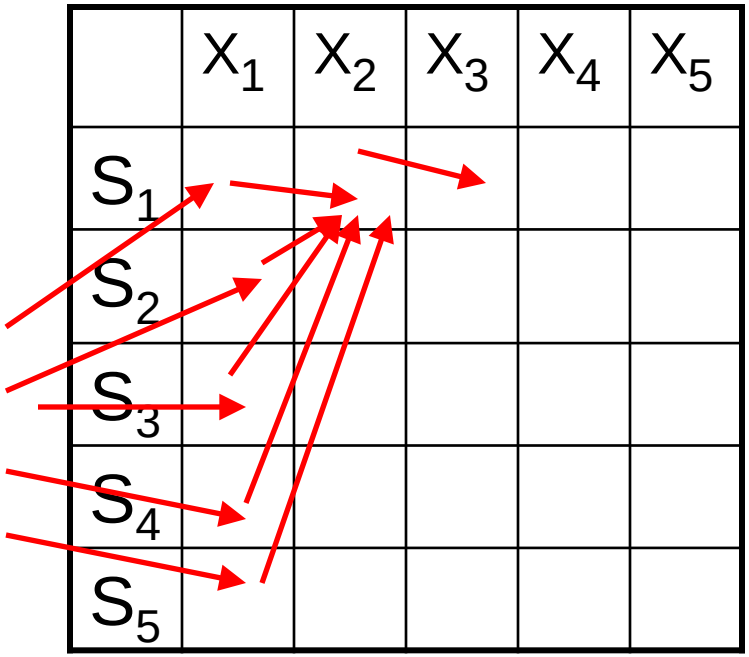
Not likely to happen

THE VITERBI ALGORITHM

- The best path visiting state k at time i extends best path ending at state l at time i+1 : recursive calcula tion (DP can be used!)

$$v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl})$$





ALGORITHM: VITERBI

$$\underline{\text{Initialisation } (i=0)}: v_0(0)=1, v_k(0)=0, \text{ for } k>0$$

$$\underline{\text{Recursion } (i=1..L)}:$$

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

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

$$\underline{\text{Termination}}:$$

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

$$\pi_L^*=\arg\max_k(v_k(L)a_{k0})$$

FORWARD ALGORITHM

- Given an HMM, what is the probability of obtaining s equence x ? (I don't care which path)

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

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

  Then recursively,

$$f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i) a_{kl}$$

SEARCHING WITH PROFILE HMMS: FORWARD ALGORITHM

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

$$F_j^{\mathrm{I}}(i) = \log \frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \log[a_{\mathrm{M}_j\mathrm{I}_j} \exp(F_j^{\mathrm{M}}(i-1)) + \log a_{\mathrm{I}_j\mathrm{I}_j} \exp(F_j^{\mathrm{I}}(i-1)) + a_{\mathrm{D}_j\mathrm{I}_j} \exp(F_j^{\mathrm{D}}(i-1))];$$

$$F_j^{\mathrm{D}}(i) = \log[a_{\mathrm{M}_{j-1}\mathrm{D}_j} \exp(F_{j-1}^{\mathrm{M}}(i)) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j} \exp(F_{j-1}^{\mathrm{I}}(i)) + a_{\mathrm{D}_{j-1}\mathrm{D}_j} \exp(F_{j-1}^{\mathrm{D}}(i))];$$

Initialization:

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

Termination:

$$F = \log\left[\exp\left(F_L^M(N)\right) + \exp\left(F_L^I(N)\right) + \exp\left(F_L^D(N)\right)\right]$$

ALGORITHM: FORWARD ALGORITHM

$$\underline{\text{Initialisation } (i = 0)}:$$

$$f_0(0) = 1,\ f_k(0) = 0 \text{ for } k > 0$$

$$\underline{\text{Recursion}(i = 1..L)}:$$

$$f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl}$$

$$\underline{\text{Termination}}: P(x) = \sum_k f_k(L) a_{k0}$$

BACKWARD ALGORITHM

- Given an observed sequence, the probability that x_i c ame 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

PREPARATION

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

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

Forward algorithm

$$\begin{aligned} P(x, \pi_i = k) &= P(x_1 x_2 .. x_i, \pi_i = k) P(x_{i+1} x_{i+2} .. x_L \mid x_1 x_2 .. x_i, \pi_i = k) \\ &= P(x_1 x_2 .. x_i, \pi_i = k) P(x_{i+1} x_{i+2} .. x_L \mid \pi_i = k) \\ &= f_k(i) P(x_{i+1} x_{i+2} .. x_L \mid \pi_i = k) \end{aligned}$$

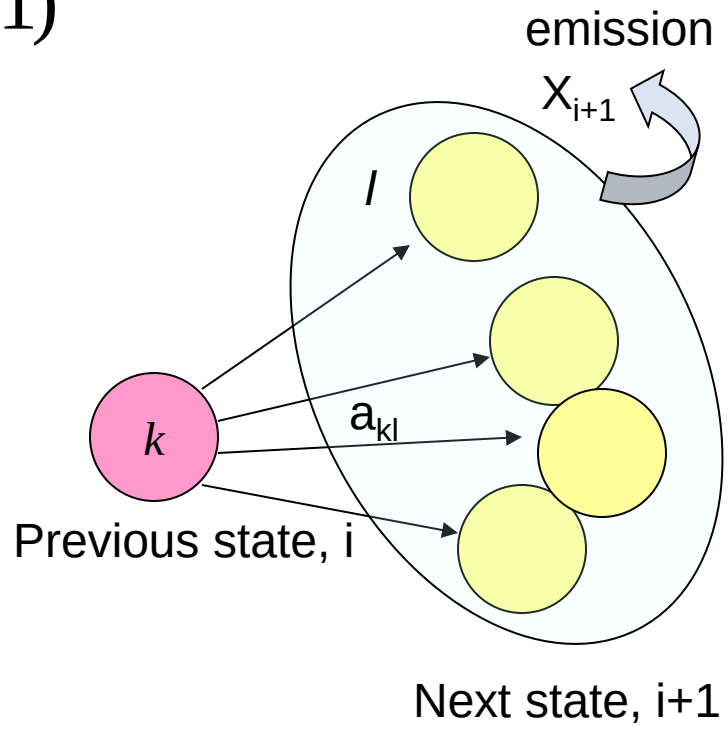
- Let's write $P(x_{i+1} x_{i+2} .. x_L \mid \pi_i = k) = b_k(i)$ and calculate it recursively

BACKWARD ALGORITHM

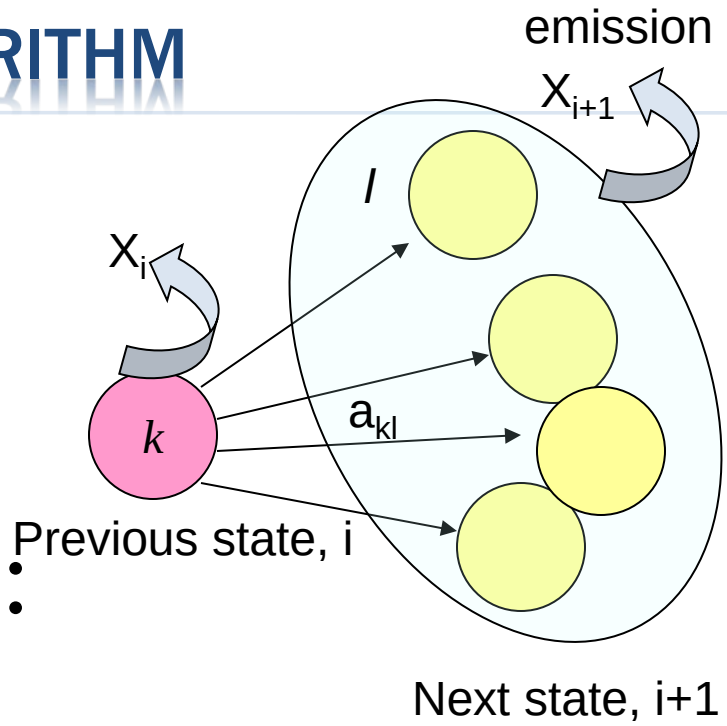
$b_k(i)$: the probability that the state k is used for the step i, and the sequences x_{i+1} .. X_L is observed

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

	X_1	X_2	X_3	X_4	X_5
S_1					
S_2					
S_3					
S_4					
S_5					



ALGORITHM: BACKWARD ALGORITHM



$$\underline{\text{Initialisation } (i = L)}:$$

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

$$\underline{\text{Recursion}(i = L-1..1)}:$$

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

$$\underline{\text{Termination}}: P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$$

PARAMETER ESTIMATION FOR HMMS

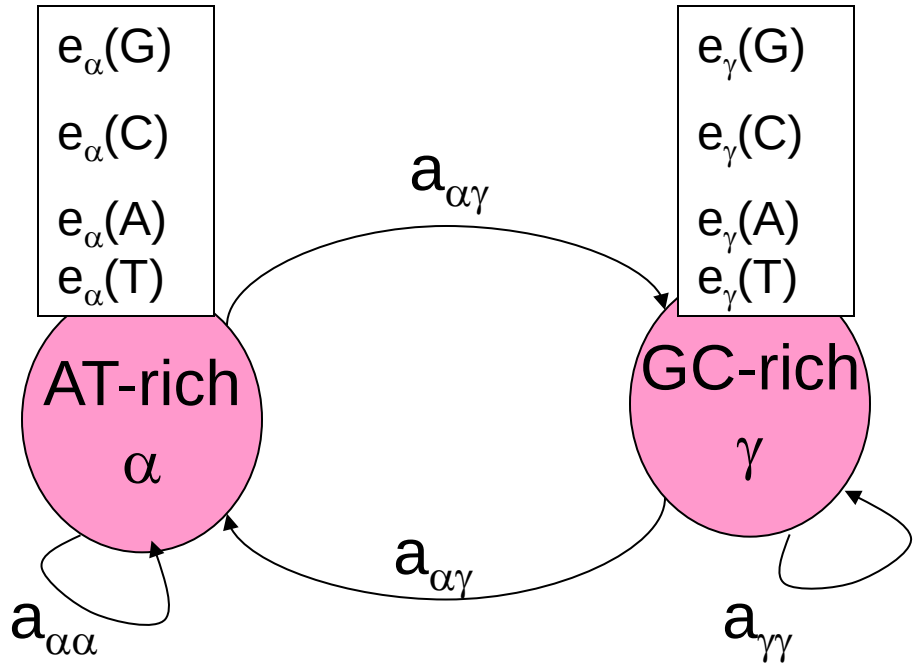
- Use a set of example sequences (training sequences)
- When the paths are known for all the examples:
  - E.g. CpG Island, protein secondary structure

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \qquad \text{..... (A)}$$

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

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

EXAMPLE OF PARAMETER ESTIMATION



Emission Probabilities:
Eα(G): 1; Eα(C): 2; Eα(A): 3; Eα(T): 2 ;
eα(G) = (1+1)/(8+4) = 2/12
eα(C) = (2+1)/(8+4) = 3/12
eα(A) = (3+1)/(8+4) = 4/12
eα(T) = (2+1)/(8+4) = 3/12

AACACTCATCTAGCC
aagaggagaaagagg

Pseudocount: 1 for E and a

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