

## Pairwise Sequence Alignment: Motivation

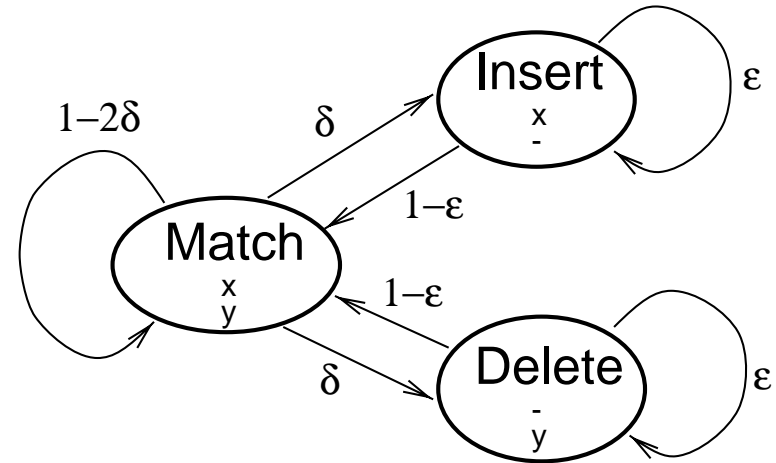
- Two DNA sequences: Sequence 1  $\rightarrow$  ACGTTGCA  
Sequence 2  $\rightarrow$  ACGTGCAT
- Direct alignment:
 

A	C	G	T	T	G	C	A
A	C	G	T	G	C	A	T
- Alignment with gaps:
 

A	C	G	T	T	G	C	A
A	C	G	T	-	G	C	A
- Interpretation: Sequence 1  $\rightarrow$  Insertion  
Sequence 2  $\rightarrow$  Deletion

slide-2

## HMM for Pairwise Sequence Alignment



slide-4

## DNA Sequence Alignment

Raw sequences

A	T	C	G	T	C	A	G	C	T	C
A	C	C	C	A	G	G	T	C		

After alignment

A	T	C	G	T	C	A	G	C	T	C
A	C	C	C	-	-	A	G	G	T	C

slide-1

slide-3

## Application of Hidden Markov Models in Bioinformatics

Dirk Husmeier  
 Biomathematics and Statistics Scotland  
 at the Scottish Crop Research Institute  
 Invergowrie, Dundee DD2 5DA, UK  
 Email: dirk@bioss.ac.uk  
<http://www.bioss.ac.uk/~dirk>

## Pairwise Sequence Alignment: Concept

- Hidden states:  $S_t \in \{\mathcal{M}, \mathcal{I}, \mathcal{D}\}$  or  $S_t \in \{\mathcal{M}, \mathcal{I}_x, \mathcal{I}_y\}$
- Observations:  $y_t \rightarrow (x_m, y_n)$

$x_m$	C	G	T	C	A	G	-	T
$y_n$	C	C	-	-	A	G	C	T
$t$	1	2	3	4	5	6	7	8
$m$	1	2	3	4	5	6	6	7
$n$	1	2	2	2	3	4	5	6

slide-6

## Viterbi Algorithm for Pairwise Alignment: Logarithmic Version

$$V_{m,n}(\mathcal{M}) = \log P(x_m, y_n) + \max \begin{bmatrix} V_{m-1,n-1}(\mathcal{M}) + \log(1 - 2\delta) \\ V_{m-1,n-1}(\mathcal{I}_x) + \log(1 - \epsilon) \\ V_{m-1,n-1}(\mathcal{I}_y) + \log(1 - \epsilon) \end{bmatrix}$$

$$V_{m,n}(\mathcal{I}_x) = \log P(x_m) + \max \begin{bmatrix} V_{m-1,n}(\mathcal{M}) + \log \delta \\ V_{m-1,n}(\mathcal{I}_x) + \log \epsilon \end{bmatrix}$$

$$V_{m,n}(\mathcal{I}_y) = \log P(y_n) + \max \begin{bmatrix} V_{m,n-1}(\mathcal{M}) + \log \delta \\ V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \end{bmatrix}$$

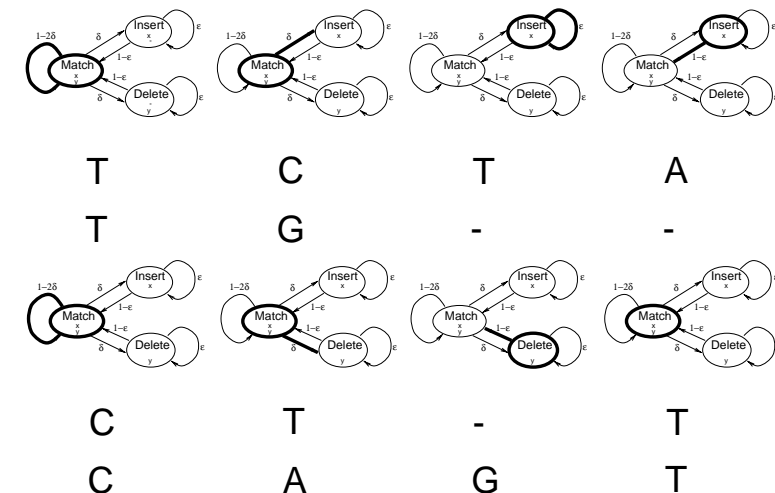
$$\uparrow_{m,n}(\mathcal{M}) = \operatorname{argmax} \begin{bmatrix} V_{m-1,n-1}(\mathcal{M}) + \log(1 - 2\delta) \\ V_{m-1,n-1}(\mathcal{I}_x) + \log(1 - \epsilon) \\ V_{m-1,n-1}(\mathcal{I}_y) + \log(1 - \epsilon) \end{bmatrix}$$

$$\uparrow_{m,n}(\mathcal{I}_x) = \operatorname{argmax} \begin{bmatrix} V_{m-1,n}(\mathcal{M}) + \log \delta \\ V_{m-1,n}(\mathcal{I}_x) + \log \epsilon \end{bmatrix}$$

$$\uparrow_{m,n}(\mathcal{I}_y) = \operatorname{argmax} \begin{bmatrix} V_{m,n-1}(\mathcal{M}) + \log \delta \\ V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \end{bmatrix}$$

slide-8

## Example: Pairwise Sequence Alignment with HMMs



slide-5

## Viterbi Algorithm for Pairwise Sequence Alignment

$$\gamma_{m,n}(\mathcal{M}) = P(x_m, y_n) \max \begin{bmatrix} (1 - 2\delta)\gamma_{m-1,n-1}(\mathcal{M}) \\ (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_x) \\ (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_y) \end{bmatrix}$$

$$\gamma_{m,n}(\mathcal{I}_x) = P(x_m) \max \begin{bmatrix} \delta\gamma_{m-1,n}(\mathcal{M}) \\ \epsilon\gamma_{m-1,n}(\mathcal{I}_x) \end{bmatrix}$$

$$\gamma_{m,n}(\mathcal{I}_y) = P(y_n) \max \begin{bmatrix} \delta\gamma_{m,n-1}(\mathcal{M}) \\ \epsilon\gamma_{m,n-1}(\mathcal{I}_y) \end{bmatrix}$$

$$\uparrow_{m,n}(\mathcal{M}) = \operatorname{argmax} \begin{bmatrix} (1 - 2\delta)\gamma_{m-1,n-1}(\mathcal{M}) \\ (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_x) \\ (1 - \epsilon)\gamma_{m-1,n-1}(\mathcal{I}_y) \end{bmatrix}$$

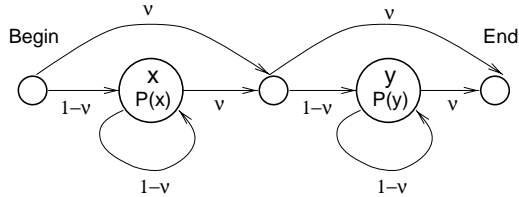
$$\uparrow_{m,n}(\mathcal{I}_x) = \operatorname{argmax} \begin{bmatrix} \delta\gamma_{m-1,n}(\mathcal{M}) \\ \epsilon\gamma_{m-1,n}(\mathcal{I}_x) \end{bmatrix}$$

$$\uparrow_{m,n}(\mathcal{I}_y) = \operatorname{argmax} \begin{bmatrix} \delta\gamma_{m,n-1}(\mathcal{M}) \\ \epsilon\gamma_{m,n-1}(\mathcal{I}_y) \end{bmatrix}$$

slide-7

## Significance of the Alignment

- Posterior probability of the **best alignment** :  $P(\hat{\pi}|D) = \frac{P(\hat{\pi}, D)}{P(D)}$
- Posterior probabilities are often very small.
- Comparison with **random model** :



$$\begin{aligned}
 P(\text{random}, D) &= \left[ (1 - \nu) \prod_{m=1}^M P(x_m) \right] \nu \left[ (1 - \nu) \prod_{n=1}^N P(y_n) \right] \nu \\
 &= \nu^2 (1 - \nu)^{(M+N)} \prod_{m=1}^M P(x_m) \prod_{n=1}^N P(y_n)
 \end{aligned}$$

- Log odds ratio:  $\log \left( \frac{P(\hat{\pi}, D)}{P(\text{random}, D)} \right)$

slide-10

## Profile HMMs: Motivation

- Functional biological sequences come in **families** .
- Evolution, speciation  $\rightarrow$  primary sequences have **diverged** from each other.
- However, they maintain the same or a related structure/function  $\rightarrow$  **homologous sequences** .

slide-12

## The Full Probability of an Alignment: Summing over All Paths

- Recursion

$$\alpha_{m,n}(\mathcal{M}) = P(x_m, y_n) \sum \begin{bmatrix} (1 - 2\delta)\alpha_{m-1,n-1}(\mathcal{M}) \\ (1 - \epsilon)\alpha_{m-1,n-1}(\mathcal{I}_x) \\ (1 - \epsilon)\alpha_{m-1,n-1}(\mathcal{I}_y) \end{bmatrix}$$

$$\alpha_{m,n}(\mathcal{I}_x) = P(x_m) \sum \begin{bmatrix} \delta\alpha_{m-1,n}(\mathcal{M}) \\ \epsilon\alpha_{m-1,n}(\mathcal{I}_x) \end{bmatrix}$$

$$\alpha_{m,n}(\mathcal{I}_y) = P(y_n) \sum \begin{bmatrix} \delta\alpha_{m,n-1}(\mathcal{M}) \\ \epsilon\alpha_{m,n-1}(\mathcal{I}_y) \end{bmatrix}$$

- Termination

$$P(D) = \alpha_{M,N}(\mathcal{M}) + \alpha_{M,N}(\mathcal{I}_x) + \alpha_{M,N}(\mathcal{I}_y)$$

- Initialisation

$$\alpha_{0,0}(\mathcal{M}) = 1$$

$$\alpha_{0,0}(\mathcal{I}_x) = \alpha_{0,0}(\mathcal{I}_y) = \alpha_{0,-1}(\cdot) = \alpha_{-1,0}(\cdot) = 0$$

This corresponds to the prior probability  $P_0(\mathcal{M}) = 1 - 2\delta$ ,  $P_0(\mathcal{I}_x) = P_0(\mathcal{I}_y) = \delta$

slide-9

## Viterbi Algorithm for Pairwise Alignment: Log-odds Version

$$V_{m,n}(\mathcal{M}) = \log P(x_m, y_n) + \max \begin{bmatrix} V_{m-1,n-1}(\mathcal{M}) + \log(1 - 2\delta) \\ V_{m-1,n-1}(\mathcal{I}_x) + \log(1 - \epsilon) \\ V_{m-1,n-1}(\mathcal{I}_y) + \log(1 - \epsilon) \end{bmatrix}$$

$$V_{m,n}(\mathcal{I}_x) = \log P(x_m) + \max \begin{bmatrix} V_{m-1,n}(\mathcal{M}) + \log \delta \\ V_{m-1,n}(\mathcal{I}_x) + \log \epsilon \end{bmatrix}$$

$$V_{m,n}(\mathcal{I}_y) = \log P(y_n) + \max \begin{bmatrix} V_{m,n-1}(\mathcal{M}) + \log \delta \\ V_{m,n-1}(\mathcal{I}_y) + \log \epsilon \end{bmatrix}$$

Log-odds ratio:  $P(x_m), P(y_n)$  cancel out,  $S(x_m, y_n) = \frac{P(x_m, y_n)}{P(x_m)P(y_n)}$ , appropriate definition of  $d, e$

$$\tilde{V}_{m,n}(\mathcal{M}) = S(x_m, y_n) + \max \begin{bmatrix} \tilde{V}_{m-1,n-1}(\mathcal{M}) \\ \tilde{V}_{m-1,n-1}(\mathcal{I}_x) \\ \tilde{V}_{m-1,n-1}(\mathcal{I}_y) \end{bmatrix}$$

$$\tilde{V}_{m,n}(\mathcal{I}_x) = \max \begin{bmatrix} \tilde{V}_{m-1,n}(\mathcal{M}) - d \\ \tilde{V}_{m-1,n}(\mathcal{I}_x) - e \end{bmatrix}$$

$$\tilde{V}_{m,n}(\mathcal{I}_y) = \max \begin{bmatrix} \tilde{V}_{m,n-1}(\mathcal{M}) - d \\ \tilde{V}_{m,n-1}(\mathcal{I}_y) - e \end{bmatrix}$$

slide-11

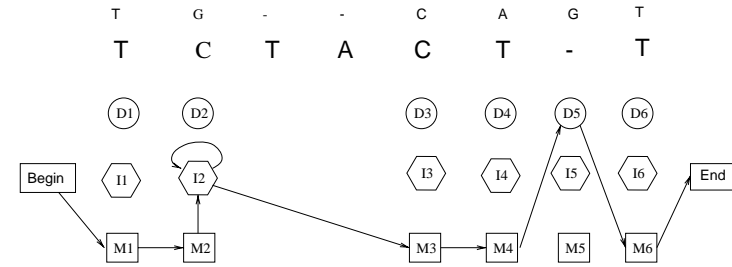
## Profile HMMs: Improvement over Pairwise Alignment

- Family of sequences: **Database search** for more members with **pairwise alignment**, known family members = query sequences.
- Repeat with all the known family members one by one.
- **Disadvantage:** May not find sequences distantly related to the ones you have already.
- **Improvement:** Use statistical features of the **whole set of sequences** in the search.
- Multiple sequence alignment: Concentrate on features that are **conserved in the whole family** → improvement over pairwise alignment.

slide-14

## Profile HMMs

- **Observations:**  $y_t \in \{A, C, G, T\}$ ,  $t = 1, \dots, N$
- **Hidden states:**  $S_t \in \{\mathcal{M}_k, \mathcal{I}_k, \mathcal{D}_k\}$ ;  $k = 1, \dots, K$



		Main state	Insertion	Deletion
Observation label	$t$	$t \rightarrow t+1$	$t \rightarrow t+1$	unchanged
State label	$k$	$k \rightarrow k+1$	unchanged	$k \rightarrow k+1$

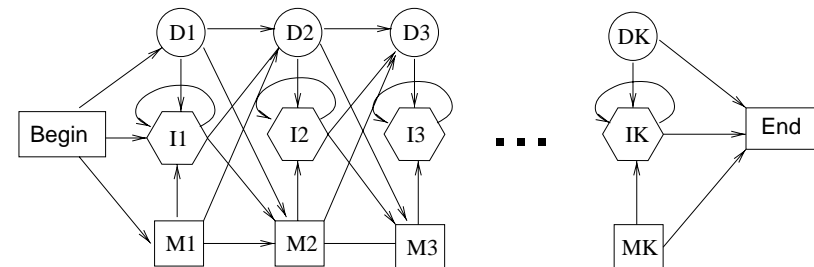
slide-16

## Profile HMMs: Typical Problems

- **Classification**  
Does a given biological sequence belong to a certain family, e.g. is a given protein sequence a globin?
- **Database search**  
Given a set of sequences → find more sequences of the same family.
- **Multiple sequence alignment**

slide-13

## Architecture of a Profile HMM



- Main states  $\mathcal{M}_k$ ,  $k = 1, \dots, K$
- Insert states  $\mathcal{I}_k$ ,  $k = 1, \dots, K$
- Delete states  $\mathcal{D}_k$ ,  $k = 1, \dots, K$

slide-15

## Viterbi Algorithm for Profile HMMs

$$\gamma_t(\mathcal{M}_k) = P(y_t|\mathcal{M}_k) \max \begin{bmatrix} P(\mathcal{M}_k|\mathcal{M}_{k-1})\gamma_{t-1}(\mathcal{M}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{I}_{k-1})\gamma_{t-1}(\mathcal{I}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{D}_{k-1})\gamma_{t-1}(\mathcal{D}_{k-1}) \end{bmatrix}$$

$$\gamma_t(\mathcal{I}_k) = P(y_t|\mathcal{I}_k) \max \begin{bmatrix} P(\mathcal{I}_k|\mathcal{M}_k)\gamma_{t-1}(\mathcal{M}_k) \\ P(\mathcal{I}_k|\mathcal{I}_k)\gamma_{t-1}(\mathcal{I}_k) \\ P(\mathcal{I}_k|\mathcal{D}_k)\gamma_{t-1}(\mathcal{D}_k) \end{bmatrix}$$

$$\gamma_t(\mathcal{D}_k) = \max \begin{bmatrix} P(\mathcal{D}_k|\mathcal{M}_{k-1})\gamma_t(\mathcal{M}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{I}_{k-1})\gamma_t(\mathcal{I}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{D}_{k-1})\gamma_t(\mathcal{D}_{k-1}) \end{bmatrix}$$

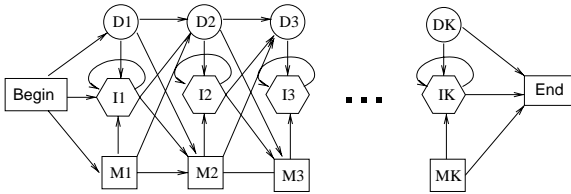
$$\uparrow_t(\mathcal{M}_k) = \operatorname{argmax} \begin{bmatrix} P(\mathcal{M}_k|\mathcal{M}_{k-1})\gamma_{t-1}(\mathcal{M}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{I}_{k-1})\gamma_{t-1}(\mathcal{I}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{D}_{k-1})\gamma_{t-1}(\mathcal{D}_{k-1}) \end{bmatrix}$$

$$\uparrow_t(\mathcal{I}_k) = \operatorname{argmax} \begin{bmatrix} P(\mathcal{I}_k|\mathcal{M}_k)\gamma_{t-1}(\mathcal{M}_k) \\ P(\mathcal{I}_k|\mathcal{I}_k)\gamma_{t-1}(\mathcal{I}_k) \\ P(\mathcal{I}_k|\mathcal{D}_k)\gamma_{t-1}(\mathcal{D}_k) \end{bmatrix}$$

$$\uparrow_t(\mathcal{D}_k) = \operatorname{argmax} \begin{bmatrix} P(\mathcal{D}_k|\mathcal{M}_{k-1})\gamma_t(\mathcal{M}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{I}_{k-1})\gamma_t(\mathcal{I}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{D}_{k-1})\gamma_t(\mathcal{D}_{k-1}) \end{bmatrix}$$

slide-18

## Forward Algorithm for Profile HMMs



$$\alpha_t(\mathcal{M}_k) = P(y_t|\mathcal{M}_k) \sum \begin{bmatrix} P(\mathcal{M}_k|\mathcal{M}_{k-1})\alpha_{t-1}(\mathcal{M}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{I}_{k-1})\alpha_{t-1}(\mathcal{I}_{k-1}) \\ P(\mathcal{M}_k|\mathcal{D}_{k-1})\alpha_{t-1}(\mathcal{D}_{k-1}) \end{bmatrix}$$

$$\alpha_t(\mathcal{I}_k) = P(y_t|\mathcal{I}_k) \sum \begin{bmatrix} P(\mathcal{I}_k|\mathcal{M}_k)\alpha_{t-1}(\mathcal{M}_k) \\ P(\mathcal{I}_k|\mathcal{I}_k)\alpha_{t-1}(\mathcal{I}_k) \\ P(\mathcal{I}_k|\mathcal{D}_k)\alpha_{t-1}(\mathcal{D}_k) \end{bmatrix}$$

$$\alpha_t(\mathcal{D}_k) = \sum \begin{bmatrix} P(\mathcal{D}_k|\mathcal{M}_{k-1})\alpha_t(\mathcal{M}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{I}_{k-1})\alpha_t(\mathcal{I}_{k-1}) \\ P(\mathcal{D}_k|\mathcal{D}_{k-1})\alpha_t(\mathcal{D}_{k-1}) \end{bmatrix}$$

slide-17

## Parameter Estimation for Known State Sequences

$$P(D, \Psi) = \prod_o \prod_h P(y_t = o | S_t = h)^{N(o|h)} \prod_h \prod_{h'} P(S_t = h' | S_{t-1} = h)^{N(h'|h)}$$

$P_E(o|h)$  Emission probability =

probability that the  $h$ th hidden state emits the  $o$ th observable symbol.

$P_T(h'|h)$  Transition probability =

probability of a transition from the  $h$ th hidden state into the  $h'$ th hidden state.

$\mathbf{w}$  Vector of model parameters = vector of all transition and emission probabilities.

$$P(D, \Psi | \mathbf{w}) = \prod_o \prod_h P_E(o|h)^{N(o|h)} \prod_h \prod_{h'} P_T(h'|h)^{N(h'|h)}$$

$$\ln P(D, \Psi | \mathbf{w}) = \sum_o \sum_h N(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} N(h'|h) \ln P_T(h'|h)$$

Maximum likelihood estimate:

$$P_E(o|h) = \frac{N(o|h)}{\sum_{o'} N(o'|h)} \quad P_T(h'|h) = \frac{N(h'|h)}{\sum_{h''} N(h''|h)}$$

slide-20

## Parameter Estimation for Known State Sequences

$y_t^i$  Observation at position  $t$  in the  $i$ th sequence

$S_t^i$  Hidden state at position  $t$  in the  $i$ th sequence

$o$  Label for observations.

Dice:  $o = 1, \dots, 6$ . DNA:  $o = 1, \dots, 4$ . Proteins:  $o = 1, \dots, 20$ .

$h$  Label for hidden states.

Rogue casino:  $h = 1, 2$ . Profile HMM of length  $K$ :  $h = 1, \dots, 3K + 2$ .

$y_t^i = o$  At the  $t$ th site in the  $i$ th sequence the  $o$ th emission symbol is observed.

$S_t^i = h$  At the  $t$ th site in the  $i$ th sequence the hidden state takes on the  $h$ th state symbol.

$N(o|h)$  Number of times the  $o$ th observation symbol is emitted from the  $h$ th state.

$N(h'|h)$  Number of state transitions from the  $h$ th state to the  $h'$ th state

$D$  Data = set of all observations =  $\{y_t^i\}_{t,i}$ , e.g.: a given set of DNA sequences.

$\Psi$  Set of all hidden state sequences, e.g.: annotated alignment of DNA sequences.

$$P(D, \Psi) = \prod_i \prod_t P(y_t^i | S_t^i) P(S_t^i | S_{t-1}^i)$$

$$= \prod_o \prod_h P(y_t = o | S_t = h)^{N(o|h)} \prod_h \prod_{h'} P(S_t = h' | S_{t-1} = h)^{N(h'|h)}$$

slide-19

## Parameter Estimation for Unknown State Sequences

$P_E(o h)$	Emission probability
$P_T(h' h)$	Transition probability
$\mathbf{w}$	Vector of model parameters = vector of all transition and emission probabilities.
$N_\Psi(o h)$	Number of times the $o$ th observation symbol is emitted from the $h$ th state for known state psequences $\Psi$ .
$N_\Psi(h' h)$	Number of state transitions from the $h$ th state symbol to the $h'$ th state symbol for known state sequences $\Psi$ .

$$\ln P(D, \Psi|\mathbf{w}) = \sum_o \sum_h N_\Psi(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} N_\Psi(h'|h) \ln P_T(h'|h)$$

$$F(\mathbf{w}) = \sum_\Psi Q(\Psi) \ln P(D, \Psi|\mathbf{w}) + C$$

$$= \sum_o \sum_h \bar{N}(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} \bar{N}(h'|h) \ln P_T(h'|h) + C$$

$$\bar{N}(o|h) = \sum_\Psi N_\Psi(o|h) Q(\Psi) \quad \bar{N}(h'|h) = \sum_\Psi N_\Psi(h'|h) Q(\Psi)$$

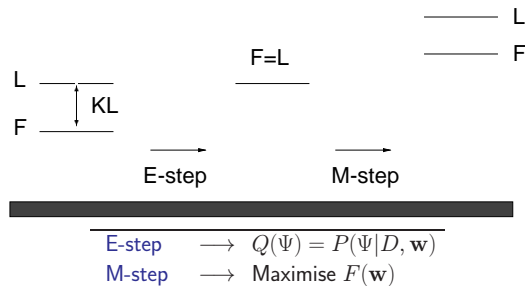
slide-22

## Parameter Estimation for Unknown State Sequences: EM Algorithm

$$L(\mathbf{w}) = \ln P(D|\mathbf{w}) = \ln \sum_\Psi P(D, \Psi|\mathbf{w})$$

$$F(\mathbf{w}) = \sum_\Psi Q(\Psi) \ln \frac{P(D, \Psi|\mathbf{w})}{Q(\Psi)} = \sum_\Psi Q(\Psi) \ln \frac{P(\Psi|D, \mathbf{w})}{Q(\Psi)} + \ln P(D|\mathbf{w})$$

$$L(\mathbf{w}) = F(\mathbf{w}) + KL[Q, P] \geq F(\mathbf{w})$$



slide-21

## Parameter Estimation for Unknown State Sequences

$$F(\mathbf{w}) = \sum_o \sum_h \bar{N}(o|h) \ln P_E(o|h) + \sum_h \sum_{h'} \bar{N}(h'|h) \ln P_T(h'|h) + C$$

M-step:  $F(\mathbf{w})$  is maximised for

$$P_E(o|h) = \frac{\bar{N}(o|h)}{\sum_{o'} \bar{N}(o'|h)} \quad P_T(h'|h) = \frac{\bar{N}(h'|h)}{\sum_{h''} \bar{N}(h''|h)}$$

E-step:  $Q(\Psi) \rightarrow P(\Psi|D, \mathbf{w})$

$$\bar{N}(o|h) = \sum_\Psi N_\Psi(o|h) P(\Psi|D, \mathbf{w}) = \sum_\Psi \left[ \sum_t \sum_i \delta(y_t^i, o) \delta(S_t^i, h) \right] P(\Psi|D, \mathbf{w})$$

$$= \sum_t \sum_i \delta(y_t^i, o) P(S_t^i = h|D, \mathbf{w})$$

$$\bar{N}(h'|h) = \sum_\Psi N_\Psi(h'|h) P(\Psi|D, \mathbf{w}) = \sum_\Psi \left[ \sum_t \sum_i \delta(S_t^i, h') \delta(S_{t-1}^i, h) \right] P(\Psi|D, \mathbf{w})$$

$$= \sum_t \sum_i P(S_t^i = h', S_{t-1}^i = h|D, \mathbf{w})$$

slide-23