今期から始まった遠隔授業の受講者の方々にも対応するため、早めにファイルを置いていますが、授業直前に内容を修正する可能性があります。2010年4月19日

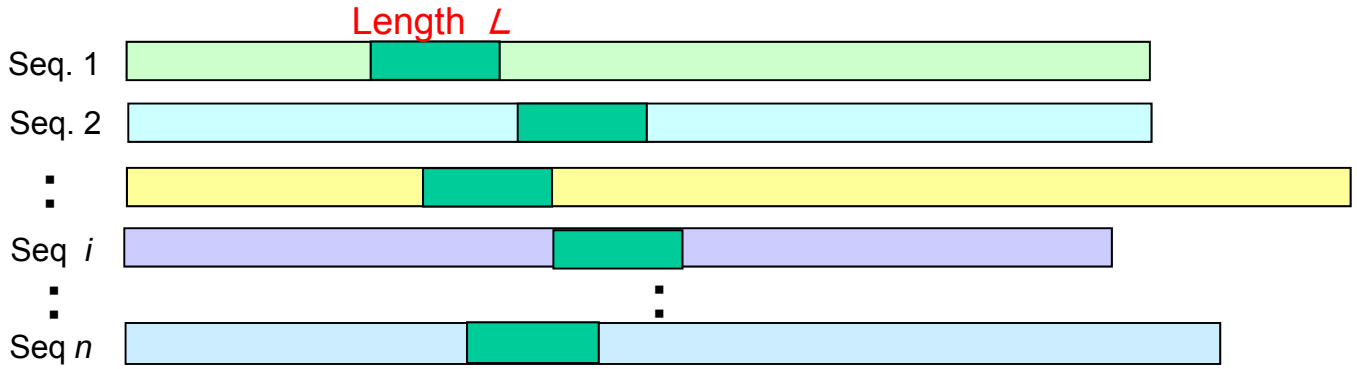
#8
HMM, Motif DB

Topics:

- Hidden Markov Model
  - Profile HMM
  - Viterbi algorithm
  - Forward-Backward algorithm
  - Baum-Welch algorithm
- Motif databases
  - PROSITE, BLOCKS, PRODOM, PFAM
  - integrated motif search system Interpro

Extracting a fixed-length motif

Review



Goal: Maximize the Relative Entropy value defined below, by sliding short windows with length *L* on each of *n* biological sequences.
OOPS: exactly One Occurrence of a motif Per Sequence
(cf. ZOOP: Zero or One Occurrence Per Sequence)

Relative Entropy $= \sum_{j=1}^{L} \sum_{a \in \Sigma} f_j(a) \cdot \log \frac{f_j(a)}{p(a)}$

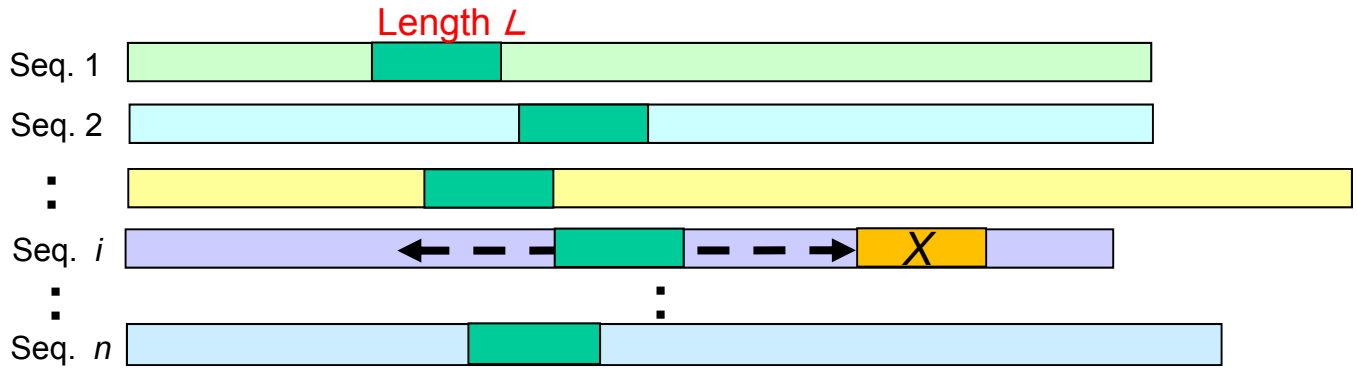
a: character, *p(a)*: background probability, *fj(a)*: frequency of character a at motif position *j*

For larger number of *n*, rigorous global optimization requires exponential time.
Approximated methods are required, just like as multiple alignment.

Extracting a fixed-length motif

Review

Approximation algorithm (using Gibbs sampling)



Step 1: Randomly choose an initial subsequence of length *L* on each seq (*1 to n).*

Step 2: (just like as the iterative improvement method of multiple alignment ...)
randomly choose one sequence from *n* sequences.（seq *i* hearinafter）

Step 3: On selected seq *i* , update the position of selected motif subseq. to X
so that X shows better similarity with other n-1 selected subsequences.。
Next position X is stochastically selected with a probability proportional to

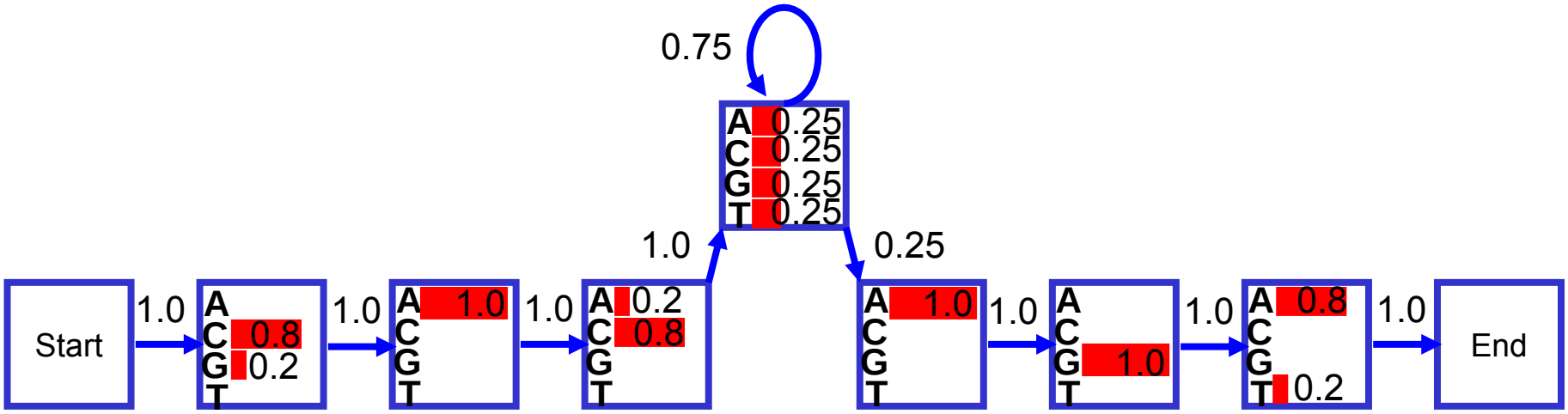
$$score(x)=\prod_{j=1}^{L}\frac{f_j(x[j])}{p(x[j])}$$

Repeat step 2 and 3 enough times, and stop when no improvement observed.
The final result depends on initial states in step1, and random numbers in step2 & 3.

Review

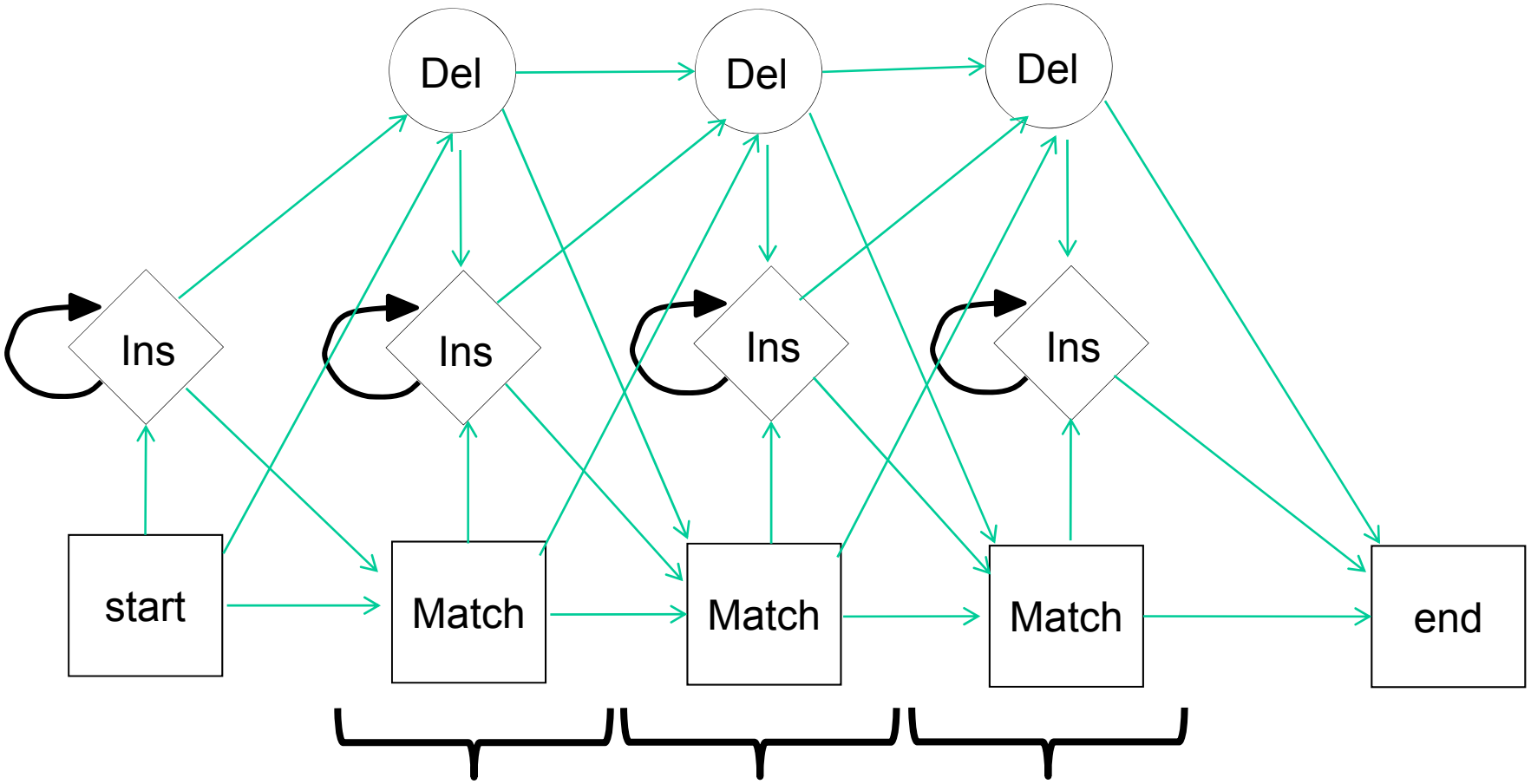
Motif Representation: HMM

1	2	3	4	5	6	7	8	9	10	11
C	A	C	a	a	a	c	g	A	G	A
C	A	C	a	t	g	g	-	A	G	T
C	A	A	t	c	t	a	-	A	G	A
G	A	C	c	g	c	t	-	A	G	A
C	A	C	a	c	t	-	-	A	G	A



Review

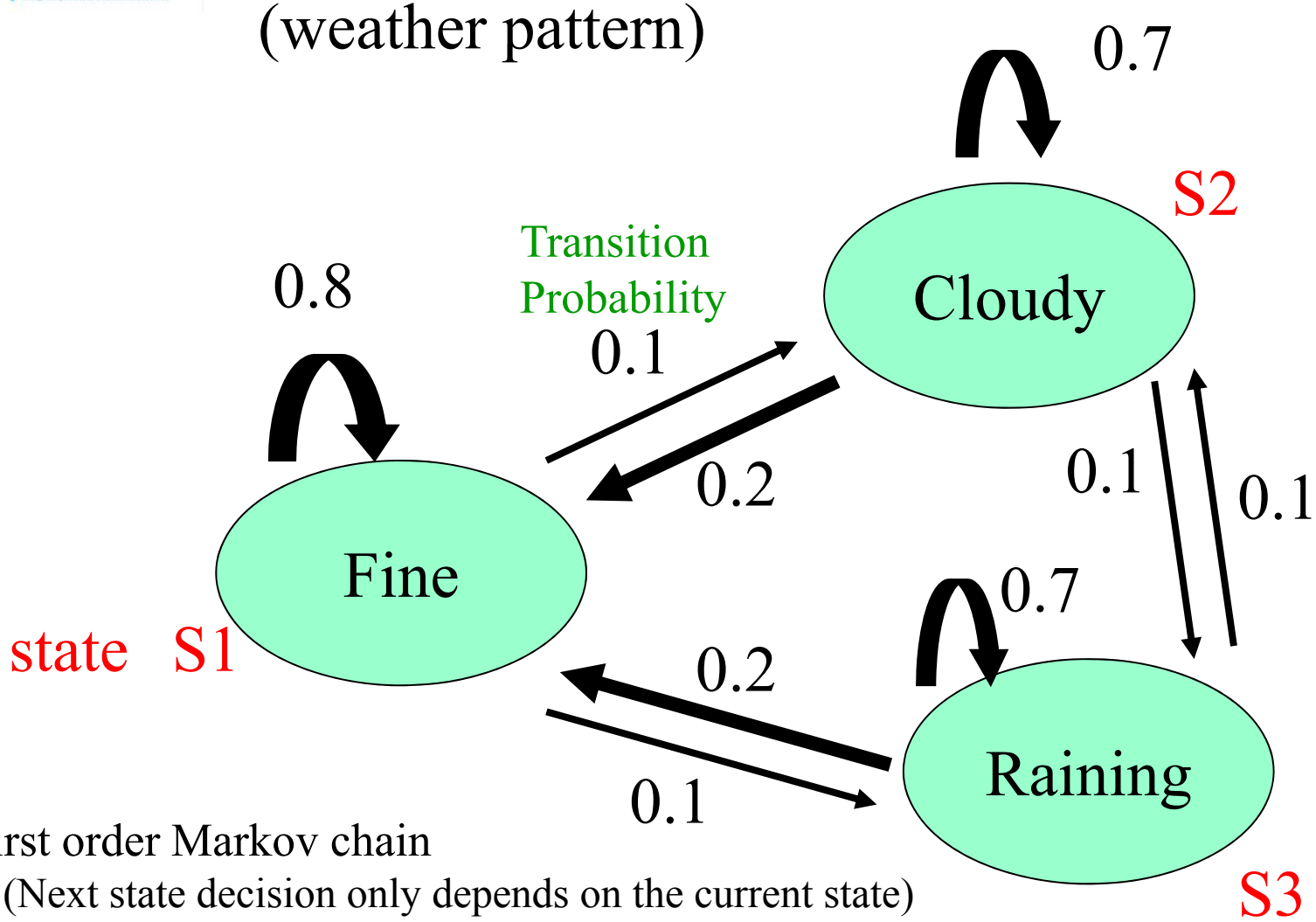
Profile HMM



corresponding to
each column position
in a sequence motif

Markov model example

(weather pattern)



first order Markov chain

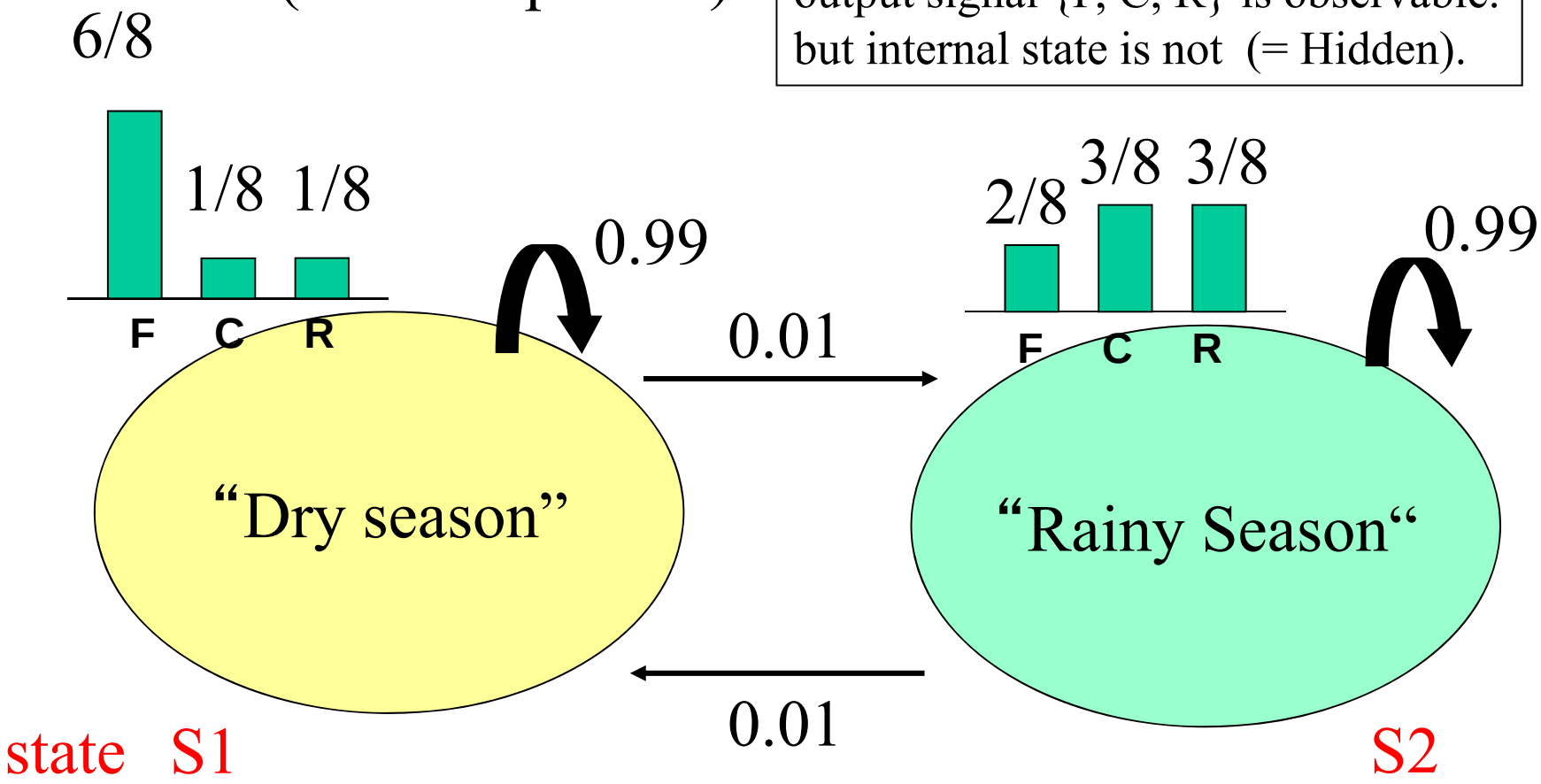
(Next state decision only depends on the current state)

$$\begin{pmatrix} P1(t+1) \\ P2(t+1) \\ P3(t+1) \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.7 \end{pmatrix} \begin{pmatrix} P1(t) \\ P2(t) \\ P3(t) \end{pmatrix} \qquad \begin{pmatrix} P1(\infty) \\ P2(\infty) \\ P3(\infty) \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \end{pmatrix}$$

Hidden Markov model example

(weather pattern)

output signal {F, C, R} is observable.
but internal state is not (= Hidden).



Each state stochastically generate output signal {F, C, R} according to its own probability distribution.

Viterbi algorithm

To find the most likely sequence path (also called as Viterbi path)

$\pi = \pi_1 \ldots \pi_n \ (\pi_t \in S)$

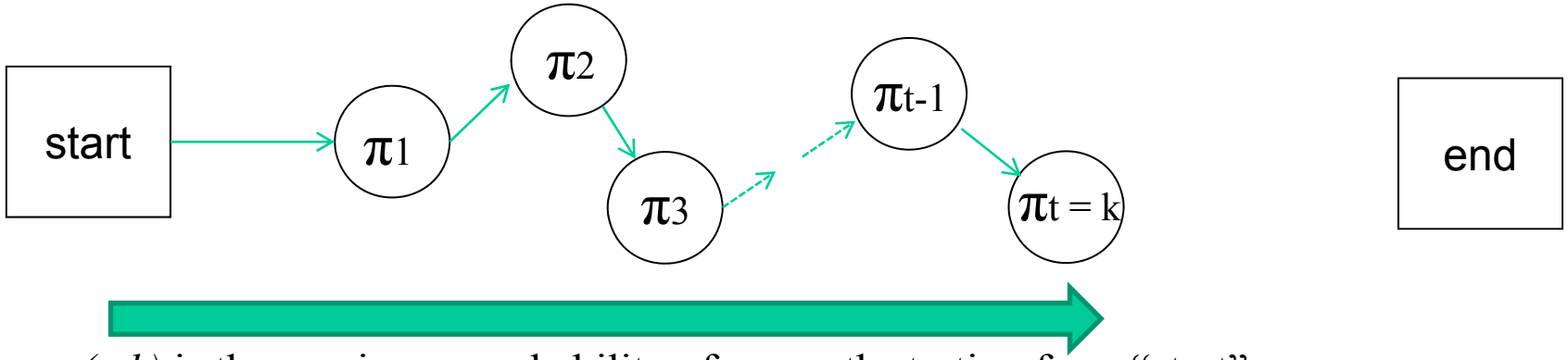
of hidden states in a HMM, when output sequence is given as

$x = x_1 \ldots x_n$

As a preparation, here we define *m(t,k)*

$$m(t,k) = \max_{\pi_i = k} p(start \rightarrow \pi_1) \cdot \prod_{j=1}^{t-1} e(x_j, \pi_j) p(\pi_j \rightarrow \pi_{j+1})$$

e(x, s) : emission probability for output ***x*** at state ***s***



m(t, k) is the maximum probability of any path starting from "start",
and passing through arbitrary *π1*, *π2*… states, and at the *t*-th step reach to the state π_t=k.

Viterbi algorithm (2)

Using

$$m(t,k) = \max_{\pi_t = k} p(start \to \pi_1) \cdot \prod_{j=1}^{t-1} e(x_j, \pi_j) p(\pi_j \to \pi_{j+1})$$

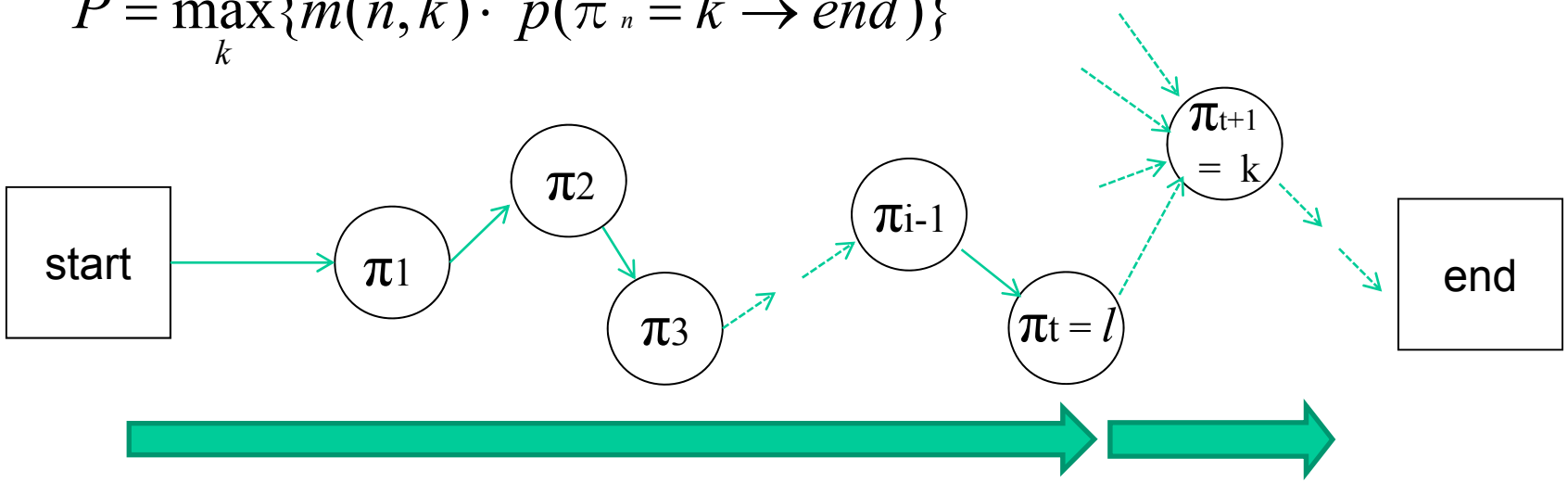
Dynamic Programming

we can get a recurrence formula as

$$m(t+1,k) = e(x_{t+1},k) \cdot \max_l \{ m(t,l) \cdot p(\pi_t = l \to \pi_{t+1} = k) \}$$

The best path (start → end) with the maximum probability P can be obtained

$$P = \max_k \{ m(n,k) \cdot p(\pi_n = k \to end) \}$$



m(t+1) can be obtained from *m(t)* by the recurrence formula

Forward-backward algorithm

Dynamic Programming

Forward algorithm

$$fwd(t+1,k) = e(x_{t+1},k) \cdot \sum_l \{fwd(t,l) \cdot p(\pi_t = l \rightarrow \pi_{t+1} = k)\}$$

not *max* but *sum*

Backward algorithm

$$bck(t,k) = \sum_l \{ p(\pi_t = i \rightarrow \pi_{t+1} = l) \cdot e(x_{t+1},l) \cdot bck(t+1,l)\}$$

$$transition(t, i \rightarrow j) = fwd(t,i) \cdot p(i \rightarrow j) \cdot e(x_{i+1}, j) \cdot bck(t+1, j)\}$$

Baum-Welch algorithm

Used to find unknown parameters for a HMM.

Enough number of observed output sequences $\{x \mid x=x_1 \ldots x_n\}$ and initial parameters are required.

B-W algorithm is a deterministic local search and will reach to a local optimum. (cf. Gibbs sampling)

A sort of "EM algorithm". Repeat E-step and M-step alternatively

1） E-step （Expectation）

With assuming current model is correct, find optimal parsing (in motif search, optimal window positions on each biological sequences).

2） M-step （Maximization）

Based on the results in previous E-step, estimate the system parameters (state transition probabilities, emission probabilities, etc.) by Bayesian low. HMM model is slightly updated.

Motif DB (4): PFAM

Sanger centre, UK
http://www.sanger.ac.uk/Software/Pfam/

First protein motif database as HMM profiles.

PFAM-A: human curated. highly reliable. (over 8957 protein families)

PFAM-B: automatically made from PRODOM. wide coverage.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy SR., Griffiths-Jones S., Howe K.L., Marshall M. and Sonnhammer E.L.:"The Pfam Protein Families Database", *Nucl. Acids Res.* 30, 1, pp.276-280, (2002).

HMMER software (by Sean Eddy, HHMI)
is used as a search engine for HMM motifs against query sequence.
http://selab.janelia.org/

InterPro: Integrated motif database

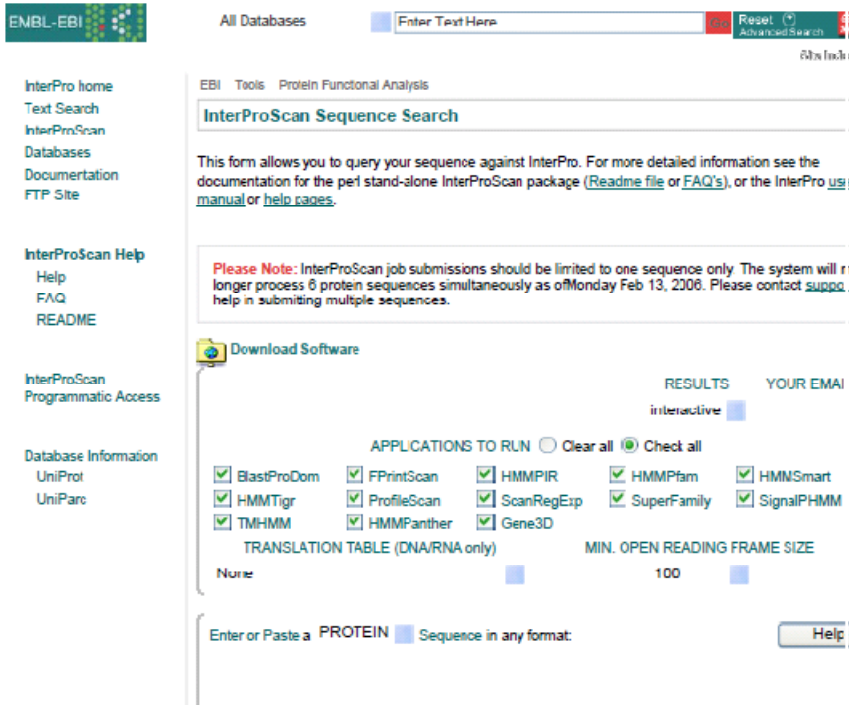
EBI, UK

http://www.ebi.ac.uk/interpro/

InterProScan Sequence search software:

PROSITE(PS)
PRINTS (PR)
SMART(SM)
PRODOM (PD)
PFAM (PF)
and so on…

simultaneous search



Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R.R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S.E., Pagni M., Peyruc D., Ponting C.P., Selengut J.D., Servant F., Sigrist C.J.A., Vaughan R, Zdobnov E.M. "The InterPro Database, 2003 brings increased coverage and new features", *Nucl. Acids. Res.*, 31, pp.315-318 (2003).