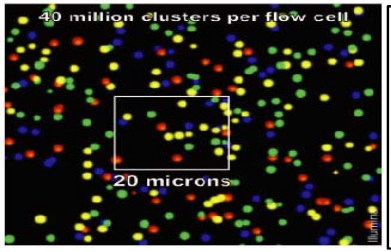
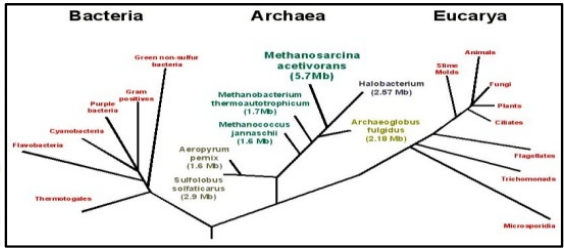
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



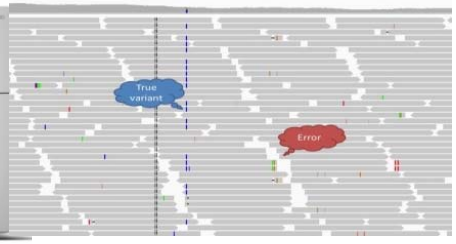


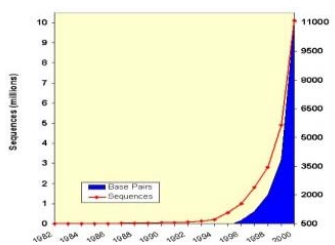
Markov Model

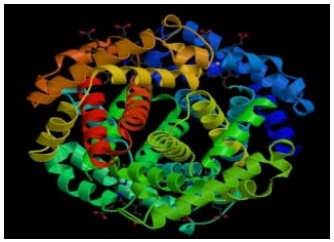
北京大学生物信息学中心 高歌

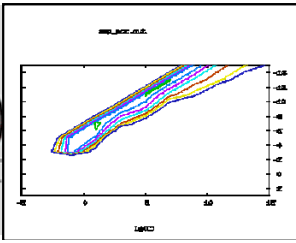
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

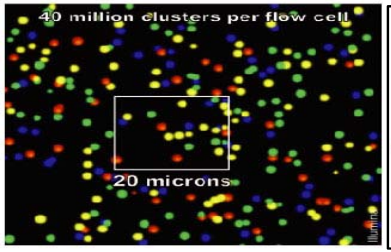


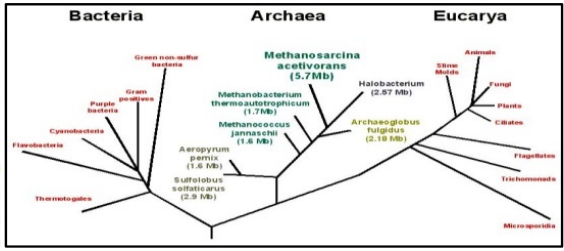






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



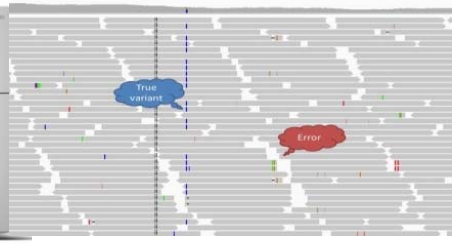


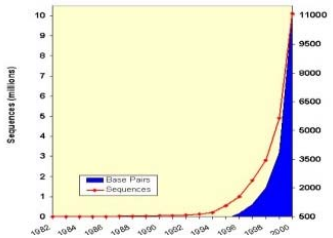
Unit 2:
Hidden Markov Model

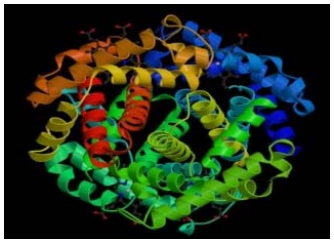
北京大学生物信息学中心 高歌

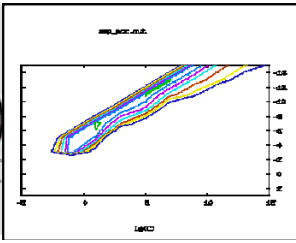
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



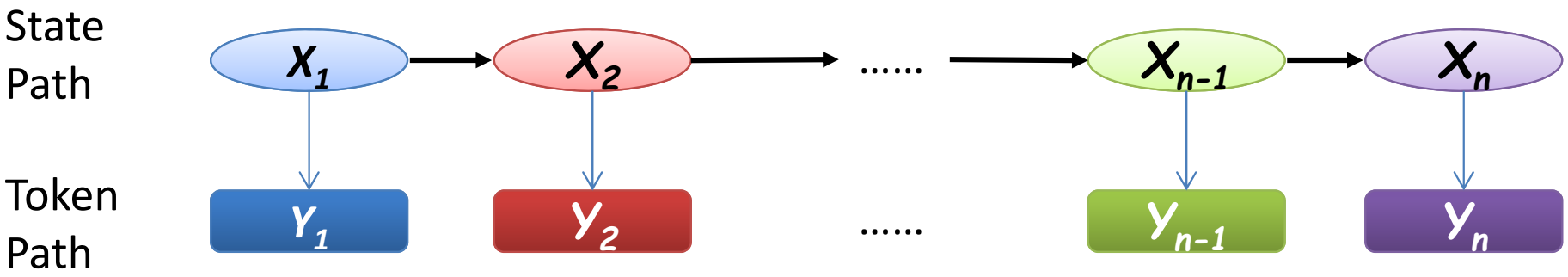






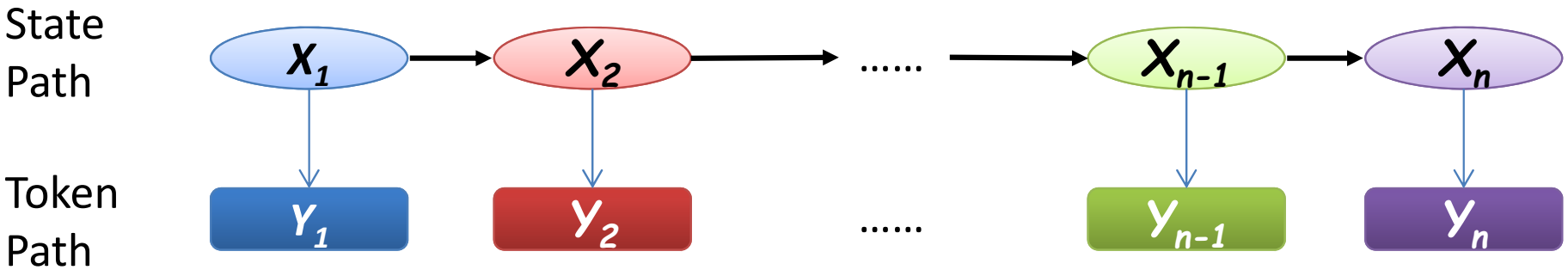
Hidden Markov Model

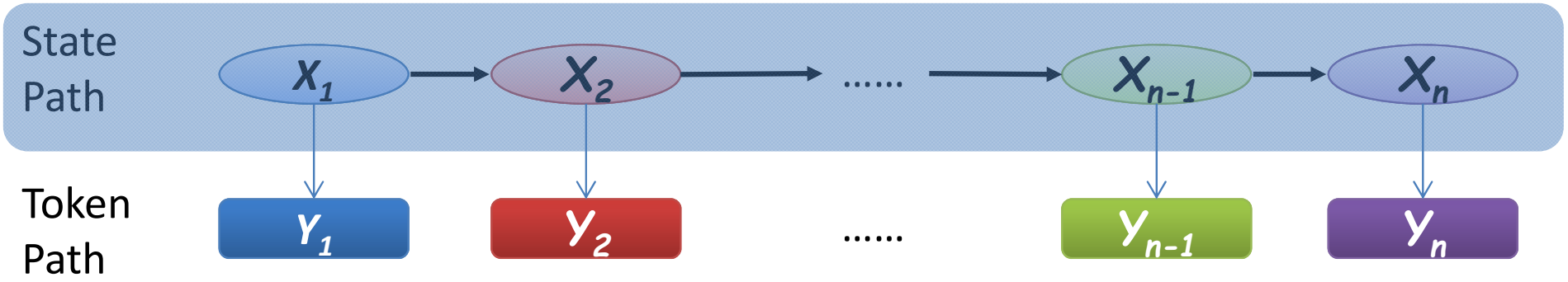
The observable symbols ("tokens", y(t)) are generated according to their corresponding states (x(t)).



Hidden Markov Model (HMM)

- In addition to State Transition Probability, each state of HMM has a probability distribution over the possible output tokens (Emission Probability).
- Thus, a HMM is consist of two strings of information.
  - The state path
  - The token path (emitted sequence).

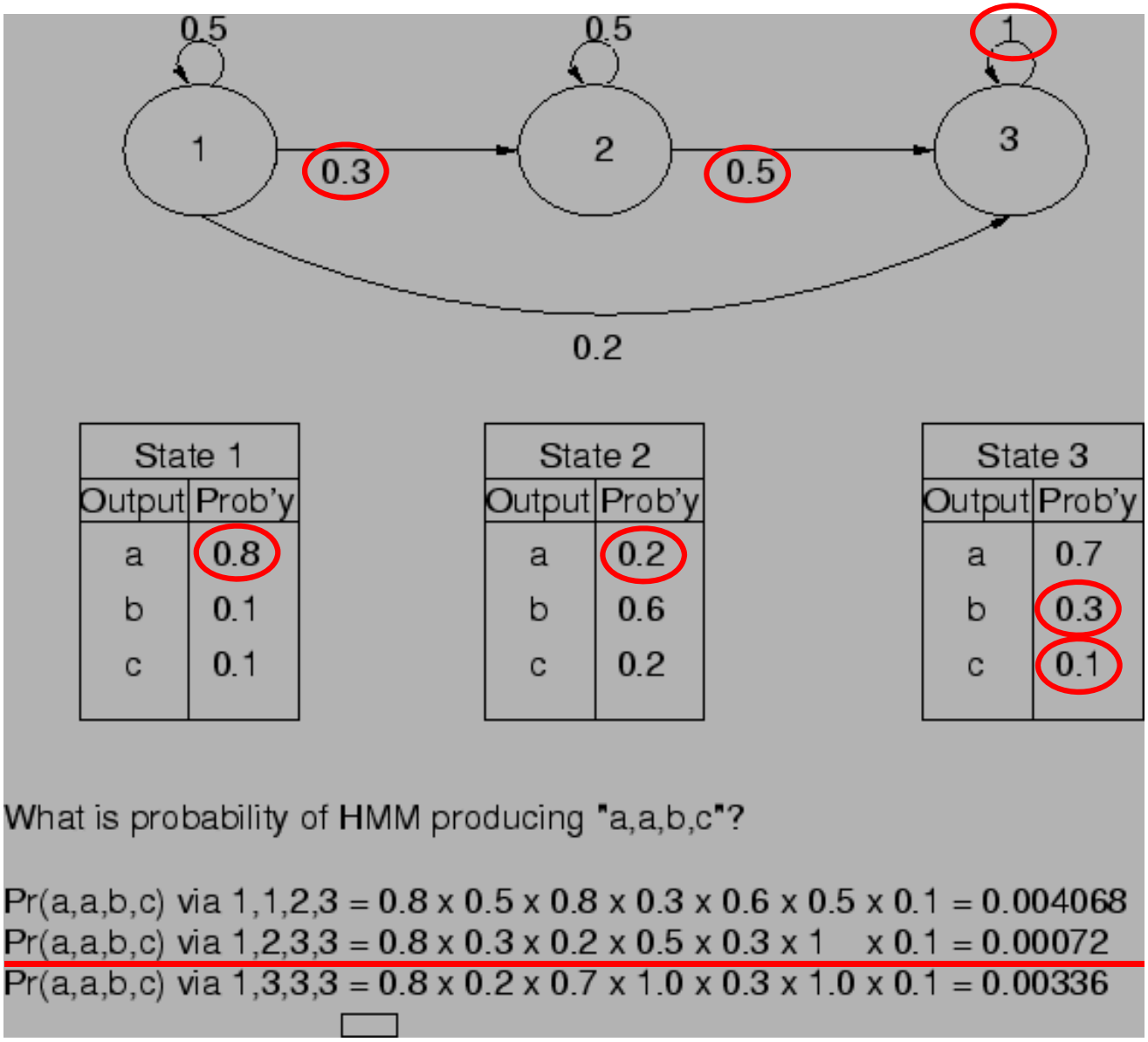




- But the state path is not directly visible
- Instead, we have to infer the underling state path, based on the observable token path.

Transition Probability

Emission Probability



(Figure source: http://www.cse.unsw.edu.au/~waleed/phd/html/node34.html)

Given a HMM, a sequence of tokens could be generated as following:

- When we “visit” a state, we emit a token from the state's emission probability distribution.
- Then, we choose which state to visit next, according to the state's transition probability distribution.

Transition Probability

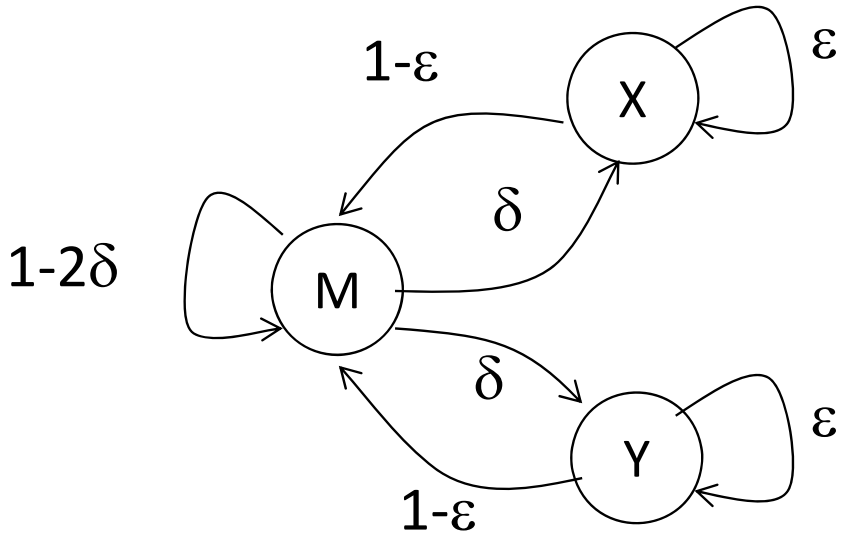
$$a_{kl} = P(x_t = S_l \mid x_{t-1} = S_k)$$

Emission Probability

$$e_k(b) = P(y_i = b \mid x_i = S_k)$$

$$P(X, Y) = \prod_{i=1}^{L} (e_{x_i}(y_i) * a_{x_i x_{i+1}})$$

M	*Match (**not necessarily identical**)*
X	*Insert* at sequence X (delete at sequence Y)
Y	*Insert* at sequence Y (delete at sequence X)

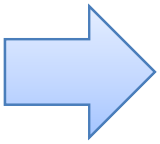


δ	Gap open
ε	Gap Extension

Transition Probability

	M	X	Y
M	1-2δ	δ	δ
X	1-ε	ε	0
Y	1-ε	0	0

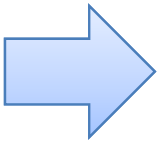
	SS	SC	ST	SP	…	WW
Match						



S
T

p_{ab}

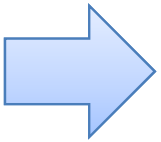
	C	S	…	W
X insertion				



S
-

q_a

	C	S	…	W
Y insertion				



-
T

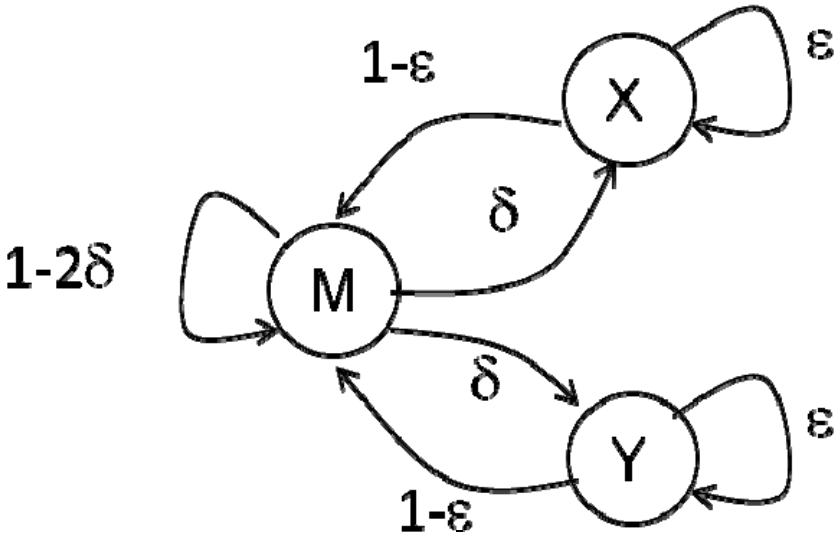
q_a

Sequence alignment with HMM

- Each "token" of the HMM is an aligned pair of two residues (M state), or of a residue and a gap (X or Y state).
  - Transition and emission probabilities define the probability of each aligned pair of sequences.

- Based on the HMM, each alignment of two sequences can be assigned with a probability
  - Given two input sequences, we look for an alignment with the maximum probability.

$$\arg\max_{ali}(P(S1, S2, ali))$$



- P_M (i,j) is the probability of the best alignment between $x_{1...i}$ and $y_{1...j}$, given ***x*$_i$ aligned to *y*$_j$**
- P_X (i,j) is the probability of the best alignment between $x_{1...i}$ and $y_{1...j}$, given ***x*$_i$ aligned to *a gap***
- P_Y (i,j) is the probability of the best alignment between $x_{1...i}$ and $y_{1...j}$, given ***y*$_j$ aligned to *a gap***

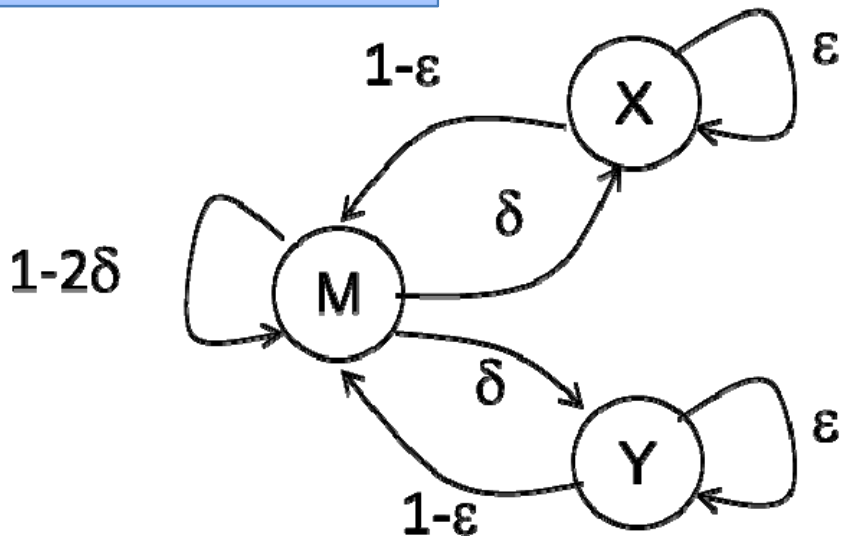
$$P_M(i,j) = p_{x_i y_j} \max \begin{pmatrix} (1-2\delta)P_M(i-1,j-1) \\ (1-\varepsilon)P_X(i-1,j-1) \\ (1-\varepsilon)P_Y(i-1,j-1) \end{pmatrix}$$

$$P_X(i,j) = q_{x_i} \max \begin{pmatrix} \delta P_M(i-1,j) \\ \varepsilon P_X(i-1,j) \end{pmatrix}$$

$$P_Y(i,j) = q_{y_j} \max \begin{pmatrix} \delta P_M(i,j-1) \\ \varepsilon P_Y(i,j-1) \end{pmatrix}$$

$$P(X,Y,ali) = \max(P_M(n,m), P_X(n,m), P_Y(n,m))$$

Transition Probability



	M	X	Y
M	1-2δ	δ	δ
X	1-ε	ε	0
Y	1-ε	0	0

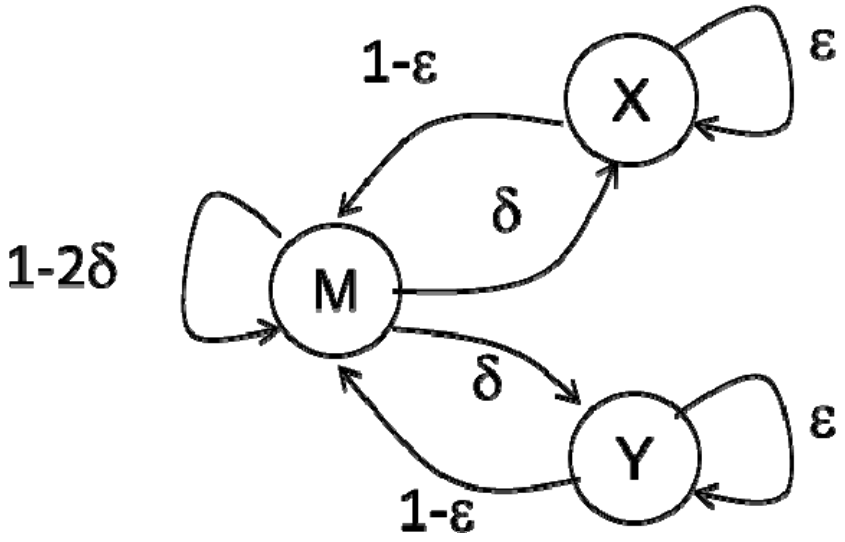
Emission Probability

	SS	SC	ST	SP	…	WW
Match						

	C	S	…	W
X insertion				

	C	S	…	W
Y insertion				

Probabilistic interpretation



	SS	SC	ST	SP	…	WW
Match						

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	[illegible]	9901

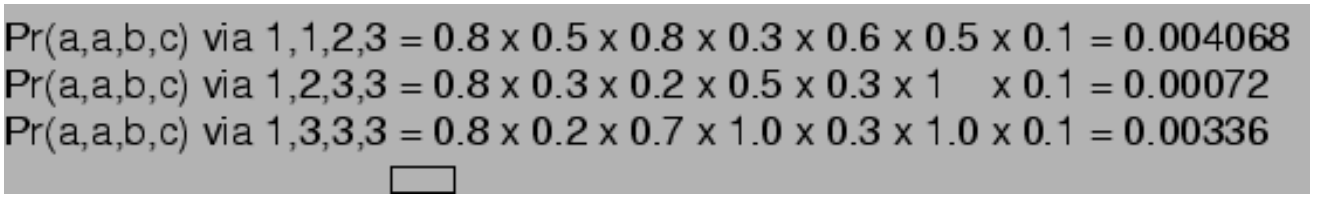
(Dayhoff 1978)

Probabilistic inference

For example, to calculate the probability that a given pair of sequences are related by *any* (unspecified) alignment

- Or, what's the best likelihood we can expect for given two sequences?

Given the nature of HMM, many different state paths can give rise to the same token sequence



(Figure source: http://www.cse.unsw.edu.au/~waleed/phd/html/node34.html)

So we can simply sum up them together to get the ***full probability of a given token sequence***

$$P(X,Y) = \sum_{ali} P(X,Y,ali)$$

$$P_M(i,j) = p_{x_i y_j} \max \begin{pmatrix} (1-2\delta)P_M(i-1,j-1) \\ (1-\varepsilon)P_X(i-1,j-1) \\ (1-\varepsilon)P_Y(i-1,j-1) \end{pmatrix}$$

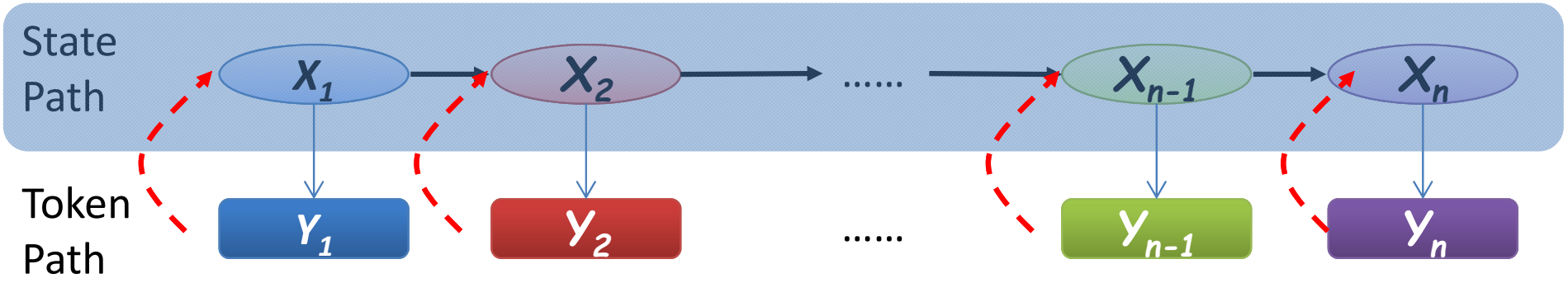
$$P_X(i,j) = q_{x_i} \max \begin{pmatrix} \delta P_M(i-1,j) \\ \varepsilon P_X(i-1,j) \end{pmatrix}$$

$$P_Y(i,j) = q_{y_j} \max \begin{pmatrix} \delta P_M(i,j-1) \\ \varepsilon P_Y(i,j-1) \end{pmatrix}$$

$$\Rightarrow \begin{cases} P_M(i,j) = p_{x_i y_j} \times [(1-2\delta)P_M(i-1,j-1) \\ \quad + (1-\varepsilon)P_X(i-1,j-1) \\ \quad + (1-\varepsilon)P_Y(i-1,j-1)] \\ P_X(i,j) = q_{x_i} \times [\delta P_M(i-1,j) \\ \quad + \varepsilon P_X(i-1,j)] \\ P_Y(i,j) = q_{y_j} \times [\delta P_M(i,j-1) \\ \quad + \varepsilon P_Y(i,j-1)] \end{cases}$$

$$P(X,Y,ali) = \max(P_M(n,m), P_X(n,m), P_Y(n,m))$$

$$\Rightarrow P(X,Y) = P_M(n,m) + P_X(n,m) + P_Y(n,m)$$



Hidden Markov Model: as a predictor

Summary Questions

- Could you name a few Markov Chain and Hidden Markov Model?

- Did we construct the global alignment or local alignment in this Unit? Explain.

生物信息学：导论与方法
Bioinformatics: Introduction and Methods