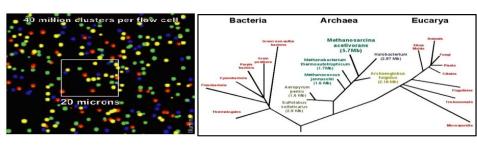


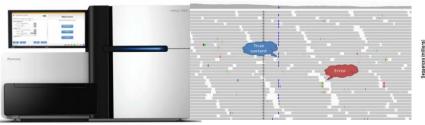
TAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAA

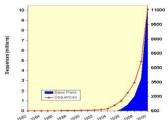


Markov Model

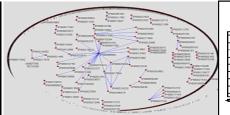
北京大学生物信息学中心 高歌 Ge Gao, Ph.D.

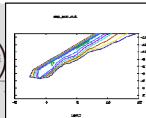
Center for Bioinformatics, Peking University





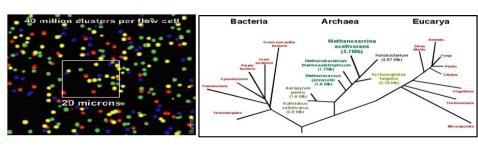








TAACCCTAACCCTAACCCTAACCCTA CCTAACCCTAACCCTAACCCTAACCC CCCTAACCCTAACCCTAACCCTAAC AACCCTAACCCTAACCCTAACCCTA ACCCTAACCCCAACCCCAACCCCAAC CTACCCTAACCCTAACCCTAACCCTA ACCCTAACCCTAACCCTAACCCTAA

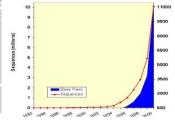


Unit 2: Hidden Markov Model

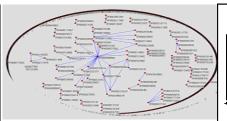
北京大学生物信息学中心 高歌 Ge Gao, Ph.D.

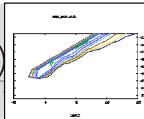
Center for Bioinformatics, Peking University





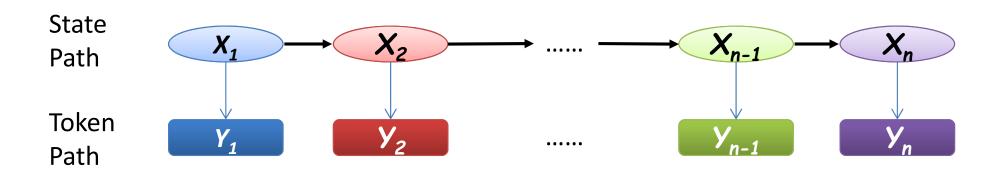






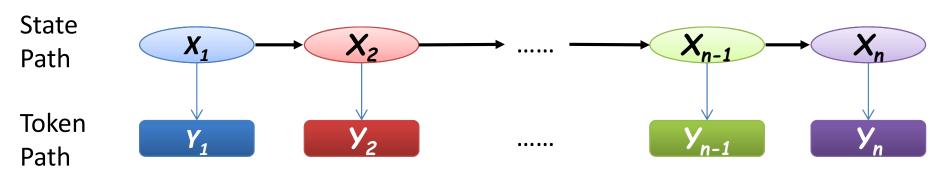
Hidden Markov Model

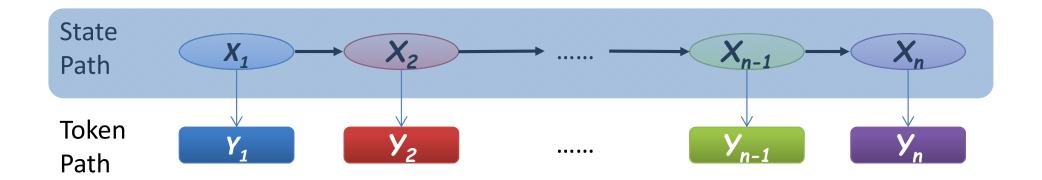
The observable symbols ("tokens", y(t)) are generated according to their corresponding states (x(t)).



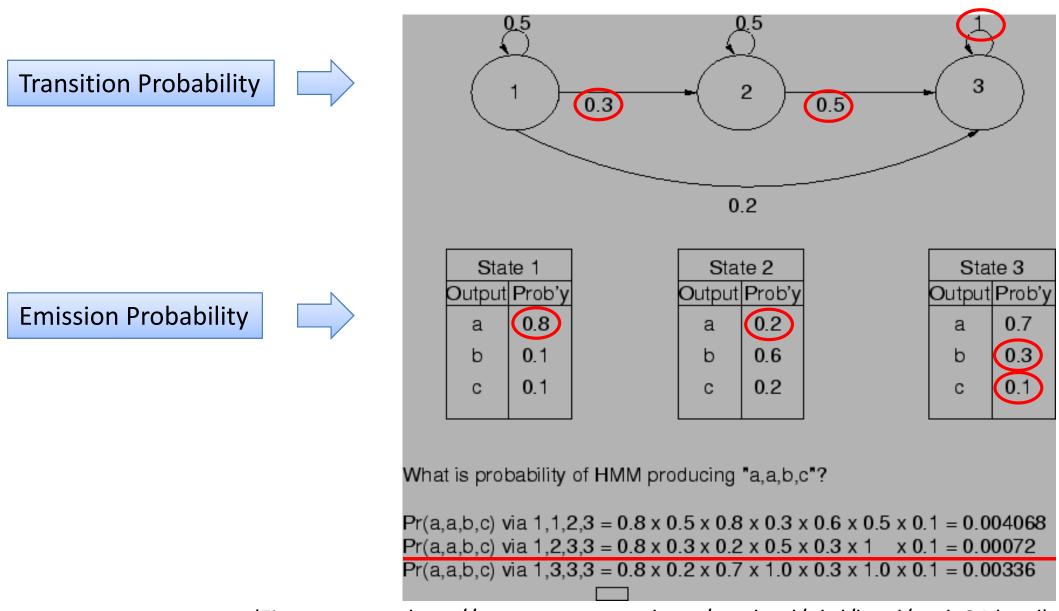
Hidden Markov Model (HMM)

- 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.



(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.

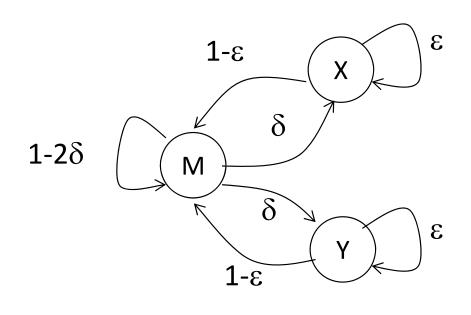


Transition Probability
$$a_{kl} = P(x_t = S_l \mid x_{t-1} = S_k)$$

Emission Probability

$$e_{k}(b) = P(y_{i} = b | x_{i} = S_{k})$$

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



Transition Probability

М	Match (<u>not necessarily identical</u>)
X	Insert at sequence X (delete at sequence Y)
Υ	Insert at sequence Y (delete at sequence X)

δ	Gap open
3	Gap Extension

M	X	Υ
1-2δ	δ	δ
1-ε	3	0
1-ε	0	0

Copyright © Peking University

Υ

M

X

	SS	SC	ST	SP	• • •	WW	S	7)
M atch							Т	P_{ab}

	С	S	•••	W	S	α
X insertion					_	\boldsymbol{q}_a

	С	S	•••	W	_	α
Y insertion					Т	\boldsymbol{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.

$$\underset{ali}{\operatorname{arg\,max}}(P(S1,S2,ali))$$

1-
$$\epsilon$$
 χ ϵ ϵ δ γ ϵ ϵ $1-\epsilon$ γ ϵ

- 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_i
- 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

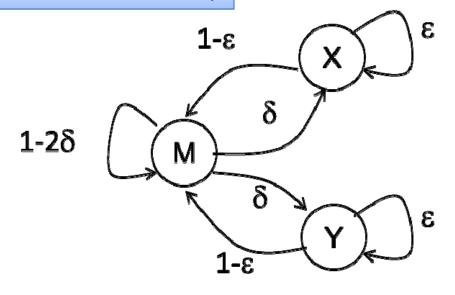
$$P_{M}(i,j) = p_{x_{i}y_{j}} \max \begin{pmatrix} (1-2\delta)P_{M}(i-1,j-1) \\ (1-\varepsilon)P_{X}(i-1,j-1) \\ (1-\varepsilon)P_{Y}(i-1,j-1) \end{pmatrix}$$

$$P_{X}(i,j) = q_{x_{i}} \max \left(\frac{\delta P_{M}(i-1,j)}{\varepsilon P_{X}(i-1,j)} \right)$$

$$P_{Y}(i,j) = q_{y_{j}} \max \left(\frac{\delta P_{M}(i,j-1)}{\varepsilon P_{Y}(i,j-1)} \right)$$

$$P(X,Y,ali) = \max(P_{M}(n,m), P_{X}(n,m), P_{Y}(n,m))$$
Copyright © Peking University

Transition Probability



M	X	Υ
1-2δ	δ	δ
1-ε	3	0
1-ε	0	0

Emission Probability

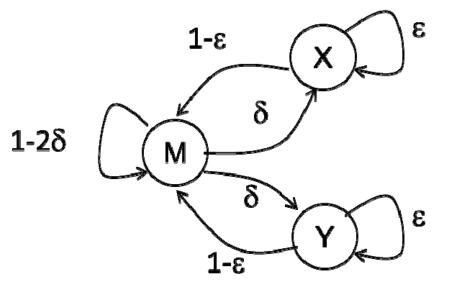
	SS	SC	ST	SP	• • •	WW
M atch						

	С	S	•••	W
X insertion				

	С	S	 W
Y insertion			

M

Probabilistic interpretation



	SS	SC	ST	SP	• • •	WW
M atch						

ORIGINAL AMINO ACID

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 5 6 21 3 1 13 0 1 2 20 9 1 4 1 0 0 Asp 6 0 42 9859 0 6 53 6 4 1 0 3 0 0 1 5 3 0 0 1 5 3 0 0 1 5 0 0 0 1 1 0 0 0 0 1 1 1 0 0 0 0	_	_																						
A Ala 9867				A	R	N	D	С	Q	Ε	G	н	1	L	K	М	F	Р	s	T	W	Y	٧	
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 35 0 4 6 6 21 3 1 13 0 1 2 20 9 1 4 1 O 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 1 0 0 9973 0 0 0 0 1 1 0 0 0 0 1 5 1 0 3 2 Q G1n 3 9 4 5 0 9876 27 1 23 1 3 6 4 0 6 2 2 0 0 1 E G1u 10 0 7 56 0 35 9865 4 2 3 1 4 1 0 3 21 3 0 0 5 H His 1 2 18 3 1 20 1 0 9912 0 1 1 0 0 3 21 3 21 3 0 0 5 H His 1 2 18 3 1 20 1 0 9912 0 1 1 0 2 3 1 1 1 4 1 4 1 I 11e 2 2 3 1 2 1 2 0 0 9876 27 9 2 12 7 0 1 7 0 1 3 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 0 2 0 0 0 5 8 4 9874 1 0 1 2 0 0 4 F Phe 1 1 1 1 0 0 0 0 2 0 0 0 5 8 4 9874 1 0 1 2 0 0 0 4 F Phe 1 1 1 1 0 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 2 1 11 2 8 6 1 5 32 9871 0 2 9 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 0 0 0 0 0 0				ala	Arg	Asn	Asp	Cys	Gin	Glu	Gly	His	He	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	İ
N Asn		A	Al a	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18	
D Asp 6 0 42 9859 0 6 53 6 4 1 0 3 0 0 1 5 3 0 0 0 1 5 3 0 0 0 1 5 3 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		R	Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1	
C Cys 1 1 1 0 0 973 0 0 0 1 1 1 0 0 0 0 0 1 5 1 0 3 2 0 0 0 1 1 1 0 0 0 0 0 1 5 1 0 3 2 0 0 1 1 2 0 0 1 1 2 0 0 1 2 0 0 1 2 0 0 0 1 2 0 0 0 0		N	Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1	
Q G1n 3 9 4 5 0 9876 27 1 23 1 3 6 4 0 6 2 2 0 0 0 1 E G1u 10 0 7 56 0 35 9865 4 2 3 1 4 1 0 3 4 2 0 1 2 G G1y 21 1 12 11 1 3 7 9935 1 0 1 2 1 1 2 1 1 3 21 3 0 0 5 5 H His 1 2 18 3 1 20 1 0 9912 0 1 1 0 2 3 1 1 1 1 4 1 1 1 4 1 1 1 1 2 2 3 1 1 3 2 1 3 0 0 0 5 5 1 1 2 2 1 1 1 3 2 1 3 0 0 0 5 5 1 1 2 2 1 1 1 3 2 1 3 0 0 0 5 5 1 1 2 2 1 1 1 3 2 1 3 0 0 0 6 1 1 1 4 22 9947 2 4 5 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 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Ì	D	Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	О	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 2 1 1 3 21 3 0 0 5 5 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8	١	¢	Cys	1	1	0	0	9973	٥	0	. 0	1	1	0	0	0	0	1	5	1	٥	3	2	
G Gly 21 1 12 11 1 3 7 9935 1 0 1 2 1 1 3 21 3 0 0 5 5 H His 1 2 18 3 1 20 1 0 9912 0 1 1 0 0 2 3 1 1 1 1 4 1 1 1 1 4 1 1 1 1 1 2 2 2 3 1 2 1 2 0 0 0 9872 9 2 12 7 0 1 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 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 0 0 0 1 1 2 8 6 0 4 9946 0 2 1 3 28 0 0 0 1 2 5 Ser 28 11 34 7 11 8 3 2 5 1 2 2 1 1 9926 12 4 0 0 2 2 5 Ser 28 11 34 7 11 4 6 16 2 2 1 7 4 3 17 9840 38 5 2 2 2 1 7 7 7 7 7 7 7 8 9870 1 0 2 9 9870 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1	1	Q	GIn	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1	ı
H His 1 8 18 3 1 20 1 0 9912 0 1 1 0 2 3 1 1 1 1 4 1 I 11e 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 0 0 0 1 0 1 0	ļ	3	G1 u	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2	ı
I 11e 2 2 3 1 2 1 2 0 0 9872 9 2 12 7 0 1 7 0 1 3 3 4 2 15 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 1 1 1 1 1 1 1 1 1 1 1 1 1 1		G	Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5	
E L Leu 3 1 3 0 0 6 1 1 4 22 9947 2 45 13 3 1 3 4 2 15 5 K Lys 2 37 25 6 0 12 7 2 2 4 1 9926 20 0 3 8 11 0 1 1 1 0 1 1 1 M Met 1 1 0 0 0 0 2 0 0 0 5 8 4 9874 1 0 1 2 0 0 4 9746 0 2 1 3 28 0 0 9 Pro 13 5 2 1 1 8 3 2 5 1 2 2 1 1 9926 12 4 0 0 2 2 5 S Ser 28 11 34 7 11 4 6 16 2 2 1 7 4 3 17 9840 38 5 2 2 1 7 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 0 0 0 0 1 0 1		Н	His	1	3	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1	١
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 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 0 0 1 0 1 0 976 1 0		1	1 le	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33	
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 2 S Ser 28 11 34 7 11 4 6 16 2 2 1 7 4 3 17 9840 38 5 2 2 7 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 0 0 1 0 1 0 976 1 0		L	Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15	į
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 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 0 0 1 0 1 0 976 1 0		K	Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1	ı
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 0 0 1 0 1 0 9976 1 0		м	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	٥	4	1
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 0 0 1 0 1 0 9976 1 0	Ì	F	Phe	1	1	1	0	0	0	0	1	2	8	6	٥	4	9946	0	2	1	3	28	0	Ì
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 0 1 0 1 0 976 1 0	4	P	Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2	١
W Trp 0 2 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 9976 1 0		S	Şer	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2	ļ
		Ţ	Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	. 32	9871	0	2	9	
Y Tyr 1 0 3 0 3 0 1 0 4 1 1 0 0 21 0 1 1 2 9945 V V V V V V V V V V V V V V V V V V V		W	Trp	0	2	0		0	0	0	0	0	0	0	0	0	1	0	1	0	l		0	
V Val 13 2 1 1 3 2 2 3 3 57 11 1 17 1 3 2 10 0 2 3701		Y	Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	D	1	1	2	9945	\mathbf{D}	
		٧	Val	13	2	1	1	3	1 2	2	3	3	57	11	1	17	1	3	2	10	0	1	3901	J

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

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

$$P_{M}(i,j) = p_{x_{i}y_{j}} \max \begin{pmatrix} (1-2\delta)P_{M}(i-1,j-1) \\ (1-\varepsilon)P_{X}(i-1,j-1) \\ (1-\varepsilon)P_{Y}(i-1,j-1) \end{pmatrix} + (1-\varepsilon)P_{X}(i-1,j-1) + (1-\varepsilon)P_{Y}(i-1,j-1)$$



$$P_{X}(i,j) = q_{x_{i}} \max \left(\frac{\delta P_{M}(i-1,j)}{\varepsilon P_{X}(i-1,j)} \right)$$

$$P_{Y}(i,j) = q_{y_{j}} \max \left(\frac{\delta P_{M}(i,j-1)}{\varepsilon P_{Y}(i,j-1)} \right)$$

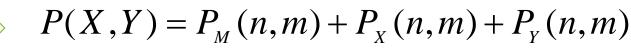


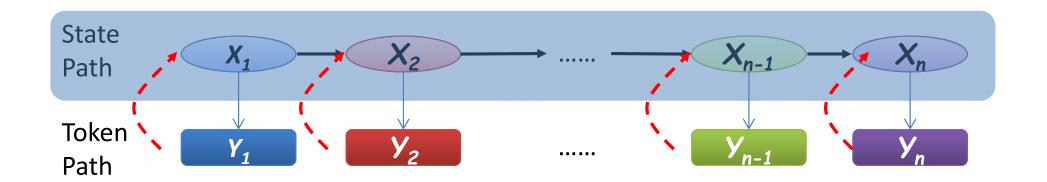
$$P_{M}(i, j) = p_{x_{i}y_{j}} \times [(1-2\delta)P_{M}(i-1, j-1) + (1-\epsilon)P_{X}(i-1, j-1) + (1-\epsilon)P_{Y}(i-1, j-1)]$$

$$P_{X}(i,j) = q_{x_{i}} \max \begin{pmatrix} \delta P_{M}(i-1,j) \\ \varepsilon P_{X}(i-1,j) \end{pmatrix} \qquad P_{X}(i,j) = q_{x_{i}} \times [\delta P_{M}(i-1,j) \\ + \varepsilon P_{X}(i-1,j)]$$

$$P_{Y}(i,j) = q_{y_{j}} \max \begin{pmatrix} \delta P_{M}(i,j-1) \\ \varepsilon P_{Y}(i,j-1) \end{pmatrix} \qquad P_{Y}(i,j) = q_{y_{j}} \times [\delta P_{M}(i,j-1) \\ + \varepsilon P_{Y}(i,j-1)]$$

$$P(X,Y,ali) = \max(P_{M}(n,m), P_{X}(n,m), P_{Y}(n,m))$$





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

