

生物信息学：导论与方法

Bioinformatics: Introduction and Methods



<https://www.coursera.org/course/pkubioinfo>



生物信息学：导论与方法

Bioinformatics: Introduction and Methods

北京大学生物信息学中心 高歌、魏丽萍

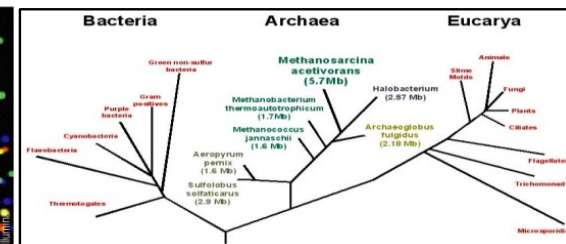
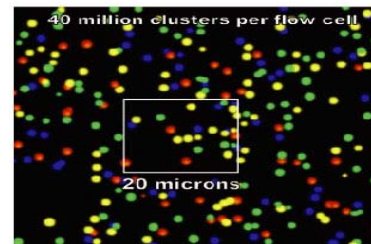
Ge Gao & Liping Wei

Center for Bioinformatics, Peking University





TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
 AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
 CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA

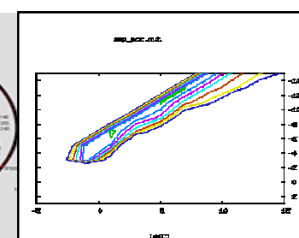
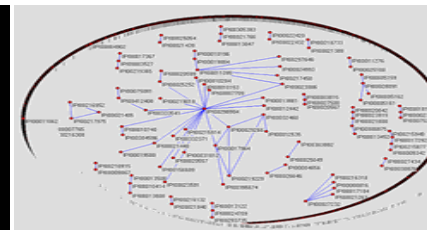
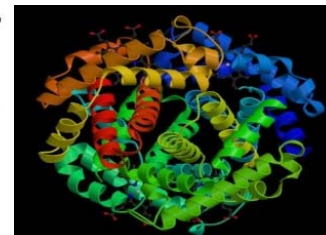
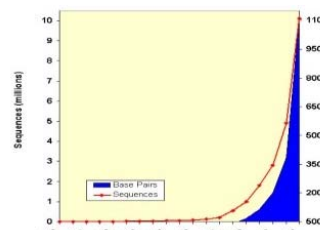
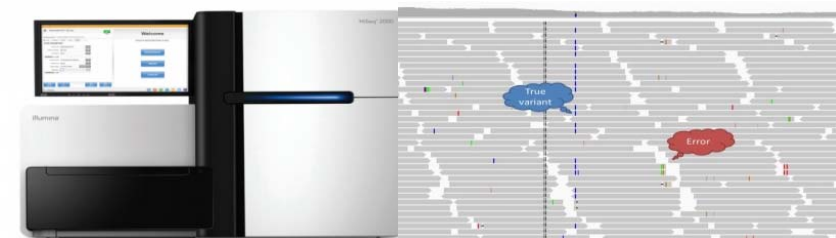


Markov Model

北京大学生物信息学中心 高歌

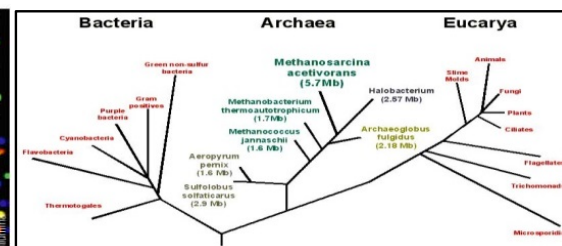
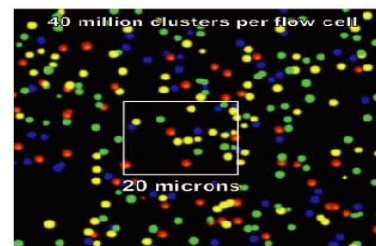
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University





TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACC
CCCTAACCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAA

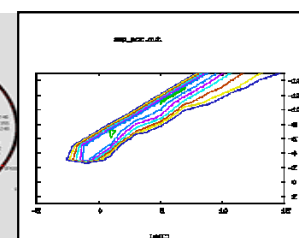
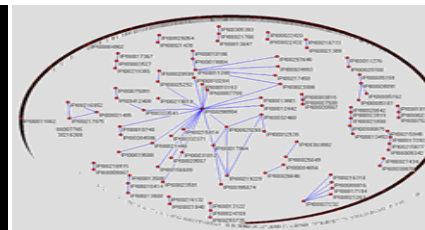
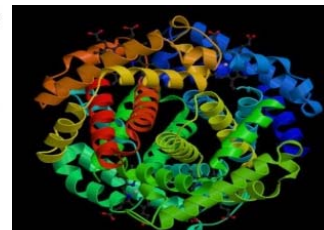
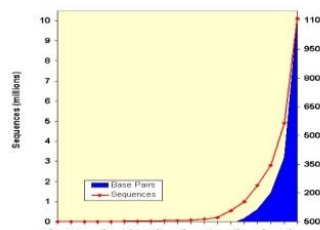
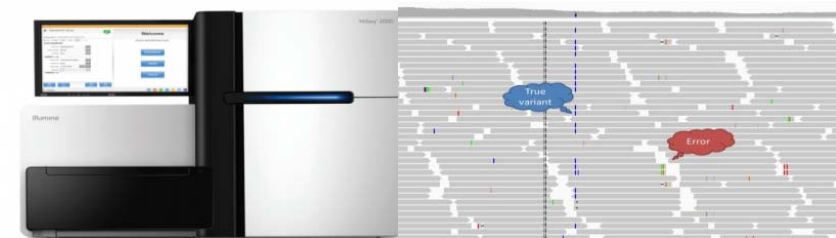


Unit 1: From States to Markov Chain

北京大学生物信息学中心 高歌

Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



Affine gap penalty: **opening** a gap receives a score of **d**; **extending** a gap receives a score of **e**.

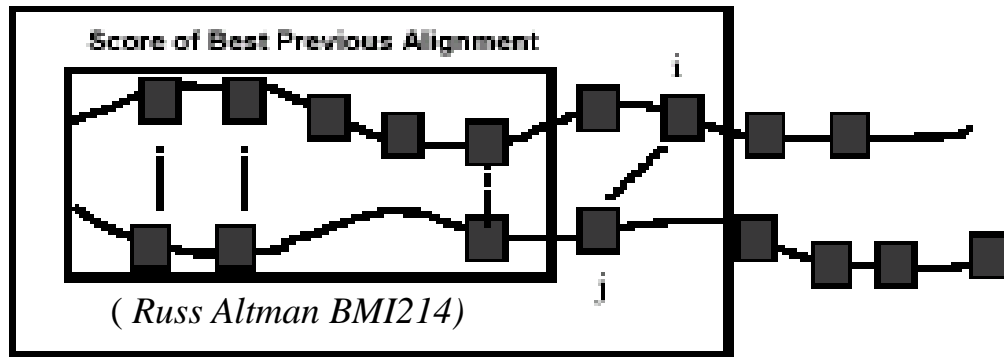
$$\text{Penalty} = d + (n-1) * e$$

```
#=====
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#=====
```

HBA_HUMAN	1	MV-LSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-D	48
		: .: .: . :.. . . :..: . :..: .	
HBB_HUMAN	1	MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD	48
HBA_HUMAN	49	LS-----HGSAQVKGHGKKVADALTNVAHVDDMPNALSALSDLHAHKLR	93
		. .: .: :..: . :..: . :..: .	
HBB_HUMAN	49	LSTPDAMVGPNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTATLSELHCDKLH	98
HBA_HUMAN	94	VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR	142
		. : .: .: :..: . :..: . :..: .	
HBB_HUMAN	99	VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147

Alignment as (a series of) state(s)

New Best Alignment = Previous Best + Local Best



		A	A	G
	0	-5		
A		2	-3	
G				-1
C				-6

L S P -

- T P E



X M M Y

M *Match (not necessarily identical)*

X *Insert at sequence X
(delete at sequence Y)*

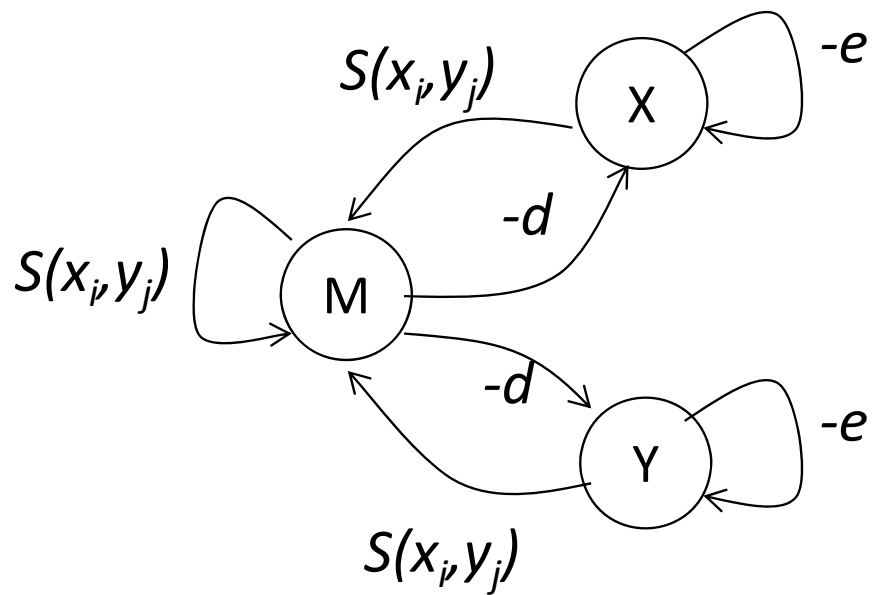
Y *Insert at sequence Y
(delete at sequence X)*

A A G -

A - G C



M X M Y

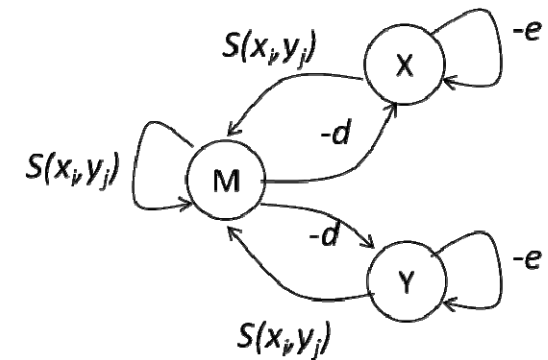


M	Match (<u>not necessarily identical</u>)
X	Insert at sequence X (delete at sequence Y)
Y	Insert at sequence Y (delete at sequence X)

d	Gap open
e	Gap Extension

- $M(i,j)$ is the score of the best alignment between $x_{1\dots i}$ and $y_{1\dots j}$, given **x_i aligned to y_j**
- $X(i,j)$ is the score of the best alignment between $x_{1\dots i}$ and $y_{1\dots j}$, given **x_i aligned to a gap**
- $Y(i,j)$ is the score of the best alignment between $x_{1\dots i}$ and $y_{1\dots j}$, given **y_j aligned to a gap**

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + s(x_i, y_j) \\ X(i-1, j-1) + s(x_i, y_j) \\ Y(i-1, j-1) + s(x_i, y_j) \end{cases}$$

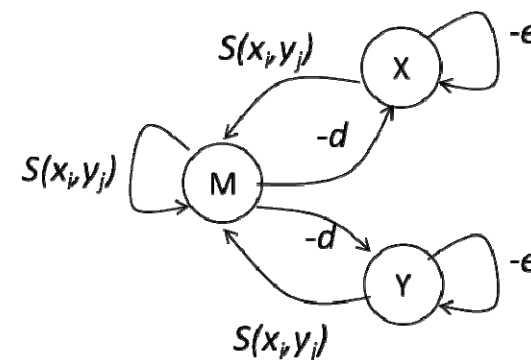


$$X(i, j) = \max \begin{cases} M(i-1, j) - d \\ X(i-1, j) - e \end{cases}$$

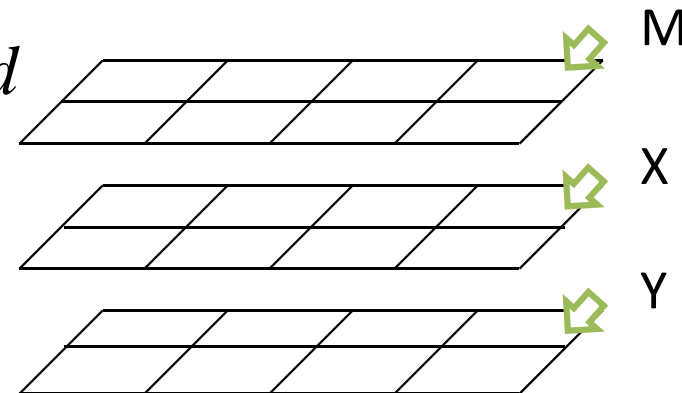
$$Y(i, j) = \max \begin{cases} M(i, j-1) - d \\ Y(i, j-1) - e \end{cases}$$

x_i aligned to y_j $M(i, j) = \max \left\{ \begin{array}{l} M(i-1, j-1) + s(x_i, y_j) \quad \text{after a match} \\ X(i-1, j-1) + s(x_i, y_j) \\ Y(i-1, j-1) + s(x_i, y_j) \end{array} \right\} \quad \text{after a gap}$

x_i aligned to a gap $X(i, j) = \max \left\{ \begin{array}{l} M(i-1, j) - d \\ X(i-1, j) - e \end{array} \right.$



y_j aligned to a gap $Y(i, j) = \max \left\{ \begin{array}{l} M(i, j-1) - d \\ Y(i, j-1) - e \end{array} \right.$



Andrei Andreyevich Markov (1856~1922)

The Mathematics
Genealogy
Project



Ph.D. [University of St. Petersburg](#) 1884



Dissertation: *On Some Applications of Algebraic Continuous Fractions*

Advisor: [Pafnuty Chebyshev](#)

Student(s):

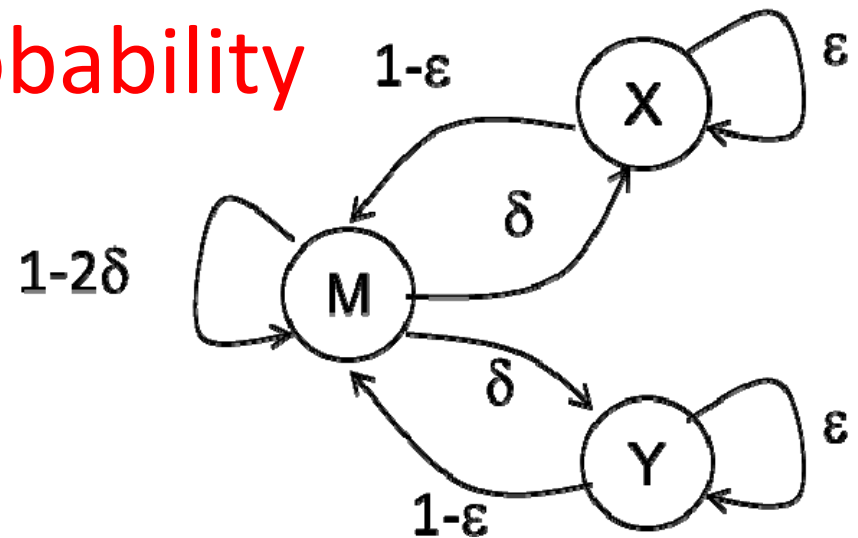
Click [here](#) to see the students listed in chronological order.

Name	School	Year	Descendants
Abram Besicovitch	University of St. Petersburg	1912	34
Nikolai Gunter	University of St. Petersburg	1915	124
Jacob Tamarkin	University of St. Petersburg	1917	778
Georgy Voronoy	University of St. Petersburg	1897	2100

According to our current on-line database, Andrei Markov has 4 [students](#) and 3039 [descendants](#).

Markov Chain

- A Markov chain describes a **discrete stochastic process** at **successive times**. The transitions from one state to any of all states, including itself, are governed by **a probability distribution**.



Markov Chain

- $P(x_t \mid x_1 \dots x_{t-1}) = P(x_t \mid x_{t-m} \dots x_{t-1})$
 - $X_t = f(X_{t-1}, X_{t-2}, \dots, X_{t-m})$
- A chain of random variables in which the next one depends (only) on the current one
 - $P(x_t \mid x_1 \dots x_{t-1}) = P(x_t \mid x_{t-1})$



Transition Probability

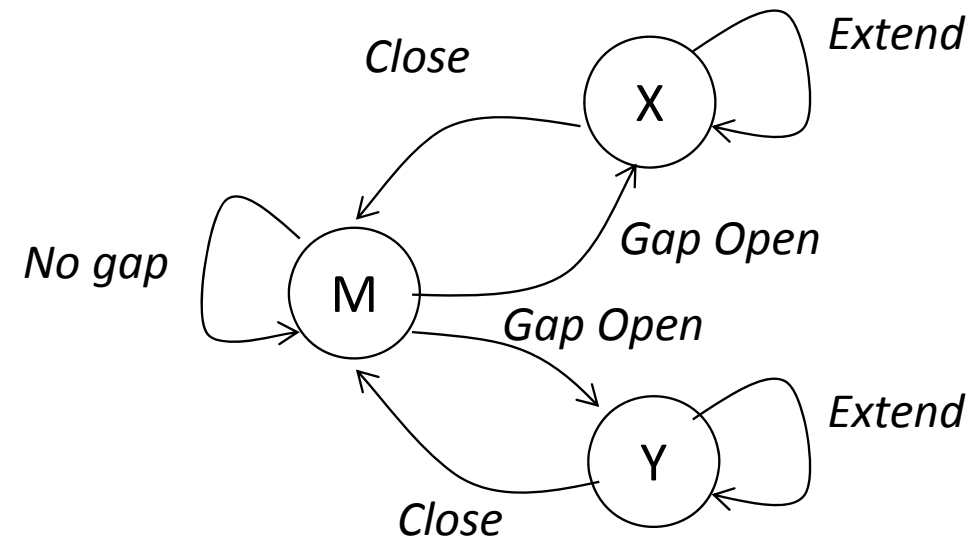
$$a_{kl} = P(X_t = S_l \mid X_{t-1} = S_k)$$

$$a_{lk} = P(X_t = S_k \mid X_{t-1} = S_l)$$

	S_1	...	S_k	...	S_l	...	S_n
S_1							
...							
S_k					a_{kl}		
...							
S_l			a_{lk}				
...							
S_n							

Affine gap penalty: **opening** a gap receives a score of **d**; **extending** a gap receives a score of **e**.

$$\text{Penalty} = d + (n-1)^* e$$

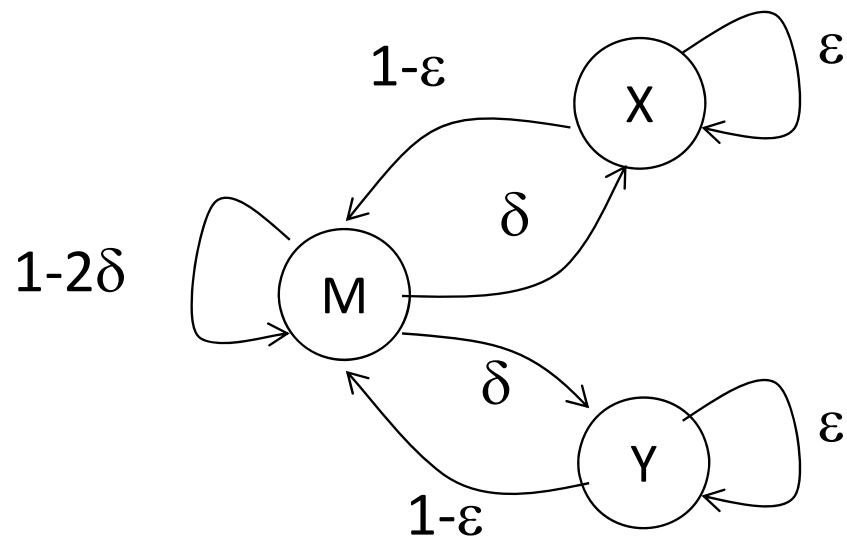


```

#
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#
#=====

```

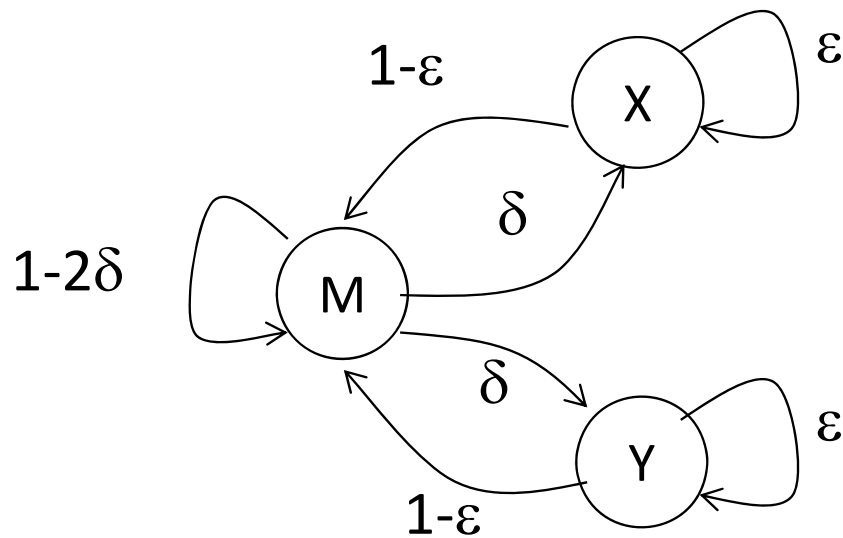
HBA_HUMAN	1 MV-LSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHF-D	48
HBB_HUMAN	1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD	48
HBA_HUMAN	49 LS-----HGSAQVKGHGGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR . .: .:. . . :.. . . .: .: . .:. .:. .	93
HBB_HUMAN	49 LSTPDAAVMGNPKVKAHGKKVLGAFSDDLGLAHLNLKGTFTLSELHCCKLH	98
HBA_HUMAN	94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFASVSVLTLSKYR	142
HBB_HUMAN	99 VDPENFRLLGNVLVCVLAHHFGKEFTPVPQAAYQKVAVGVANALAHHKYH	147



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

δ	Gap open
ε	Gap Extension

	M	X	Y
M	1-2δ	δ	δ
X	1-ε	ε	0
Y	1-ε	0	ε



L S P -
- T P E

→ X M M Y

$$P(XMMY) = a_{XM} a_{MM} a_{MY}$$

$$= (1 - \varepsilon)(1 - 2\delta)\delta$$

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

δ	Gap open
ε	Gap Extension

	M	X	Y
M	1-2δ	δ	δ
X	1-ε	ε	0
Y	1-ε	0	ε

Summary Questions

- There isn't direct transition between X and Y in our model, do you think it's reasonable? Explain your answer.
- Is it possible to have global alignment based on the rephrased Markov Chain-based model? Explain your answer.

生物信息学：导论与方法

Bioinformatics: Introduction and Methods



<https://www.coursera.org/course/pkubioinfo>