Парное выравнивание

Алгоритмы в биоинформатике

Антон Елисеев eliseevantoncoon@gmail.com

Что было на прошлой лекции?

- Транскрипция и трансляция.
- Свойство локальности ДНК.
- Можно считать расстояние между строками и делать выводы о свойствах организмов.
- о Сравнивать участки генома можно достаточно эффективно.

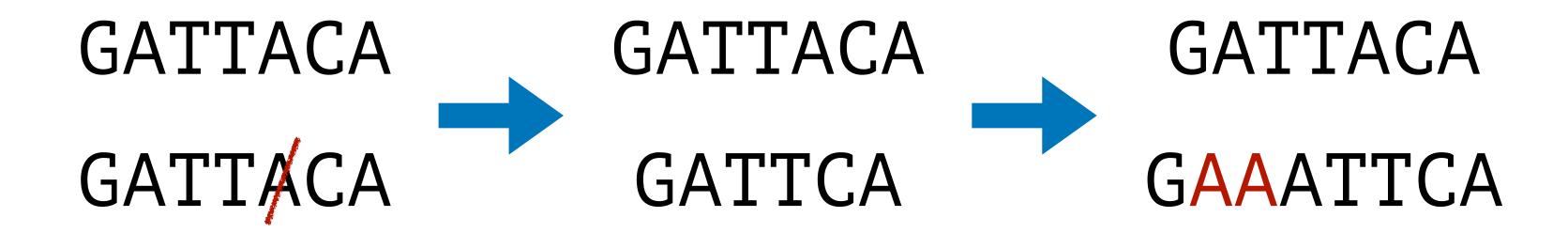
Что будет на этой лекции?

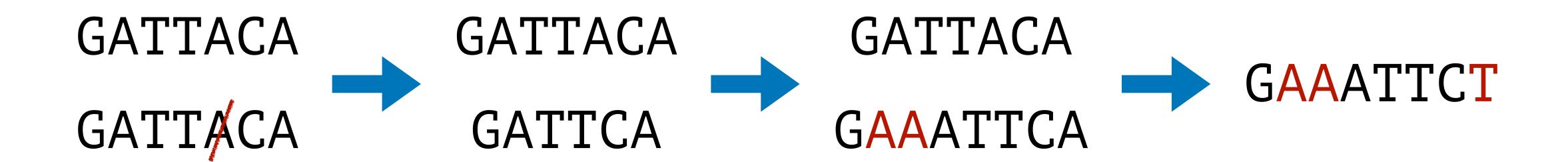
- Определение выравнивания и веса выравнивания.
- о Неравнозначные замены. Матрицы замен BLOSUM и PAM.
- Проблема гэпов. Определение аффинных штрафов за гэпы.

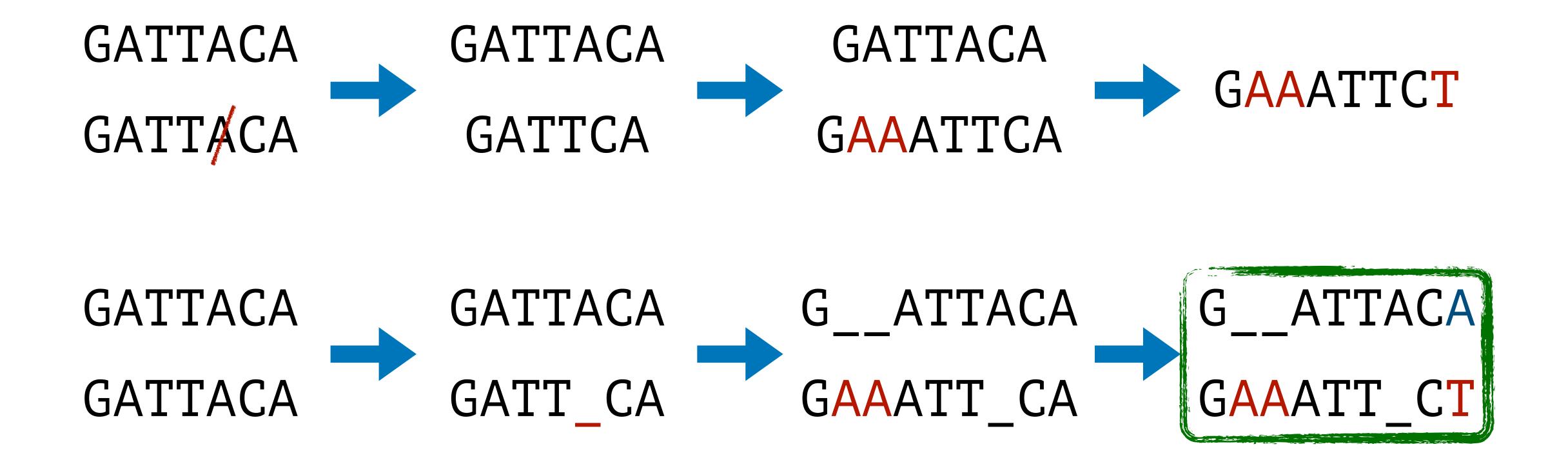
GATTACA

GATTACA

GATTACA
GATTACA
GATTACA
GATTCA







Выравнивание

Рассмотрим пару строк (a,b) где $a_i,b_i\in\mathbb{A}$

Выравнивание — такая пара строк (a^*,b^*) где $a_i^*,b_i^*\in(\mathbb{A}\ \cup\ \{_\})$, что

- 1. $|a^*| = |b^*|$
- 2. $a_i^* \neq _-$ или $b_i^* \neq _-$
- 3. При удалении всех гэпов из a^*, b^* получаем a, b

Стоимость выравнивания

Стоимостью (весом) выравнивания будем называть

$$W(a^*, b^*) = \sum_{i=1}^{|a^*|} w(a_i^*, b_i^*)$$

Где $w(a_i^*, b_i^*)$ функция $(A \cup \{_\})^2 \to \mathbb{R}$

Оптимальное выравнивание и расстояние

Оптимальным будем называть выравнивание, вес которого минимален!

А расстоянием выравнивания — вес оптимального выравнивания.

$$D(a,b) = \min_{a^*,b^*} W(a^*,b^*)$$

Где a^*, b^* — это выравнивание a, b

Редактирование и выравнивание

Для заданной $w(a_i^*,b_i^*)$ расстояние редактирование $d_w(a,b)$ равно расстоянию выравнивания $D_w(a,b)$.

Идея:

Редактирование и выравнивание

Для заданной $w(a_i^*,b_i^*)$ расстояние редактирование $d_w(a,b)$ равно расстоянию выравнивания $D_w(a,b)$.

Идея:

- $\circ d_w(a,b) \leq D_w(a,b)$: выравнивание кодирует последовательность операций редактирования.
- $^{\circ}$ $d_{w}(a,b) \geq D_{w}(a,b)$: последовательность операций редактирования порождает выравнивание такое же по весу либо меньше.

Замена
Удаление
Вставка

		G	Α	Т	Т	Α	С	Α
	0	1	2	3	4	5	6	7
Α	1	1	1	2	3	4	5	6
Α	2	2	1	2	3	3	4	5
G	3	2	2	2	3	4	4	5
Α	4	3	2	3	3	3	4	4
G	5	4	3	3	4	4	4	5
Т	6	5	4	3	3	4	5	5
Α	7	6	5	4	4	3	4	5
С	8	7	6	5	5	4	3	4

__GATTACA

AAGAGTAC

Чтобы посчитать $D_w(a,b)$, где |a|=n, |b|=m построим матрицу D, Dim(D)=(n+1,m+1) по следующим правилам:

$$O_{0,0} = 0$$

$$OD_{i,0} = D_{i-1,0} + w(a_i, _)$$

$$\circ D_{0,j} = D_{0,j-1} + w(_, b_j)$$

о
$$D_{i,j} = min \begin{cases} D_{i-1,j-1} + w(a_i,b_i) \text{ (замена)} \\ D_{i-1,j} + w(a_i,_) \text{ (удаление)} \\ D_{i,j-1} + w(_,b_j) \text{ (вставка)} \end{cases}$$

Рассмотрим выравнивание аминокислотных последовательностей

Заряженные: **D** (аспарагиновая кислота), **E** (глутаминовая кислота)

Гидрофобные: І (Изолейцин), V (Валин)

$$\circ D \rightarrow E - ?$$

$$\circ I \rightarrow V-?$$

$$\circ D \rightarrow V-?$$

Рассмотрим выравнивание аминокислотных последовательностей

Заряженные: **D** (аспарагиновая кислота), **E** (глутаминовая кислота) Гидрофобные: **I** (Изолейцин), **V** (Валин)

 $^{\circ}$ D o E — правдоподобно

- $^{\circ}~I
 ightarrow V$ правдоподобно
- $^{\circ}$ D o V— не очень то и правдоподобно

Как быть?

Как быть?

Хотелось бы отличать случайные матчи от вероятных

Как быть?

Рассмотрим выравнивание (a^*, b^*) и предположим что a^* и b^* не зависят друг от друга. Случайная модель R.

$$P(a,b \mid R) = \prod_{i=1}^{|a^*|} p_{a_i^*} \prod_{j=1}^{|b^*|} p_{b_j^*} = \prod_{i=1}^{|a^*|} p_{a_i^*} p_{b_i^*}$$

Предположим, пары встречаются не независимо. Модель сопоставления $M. \,$

$$P(a, b \mid M) = \prod_{i=1}^{|a^*|} p_{a_i^*, b_i^*}$$

Родственные к неродственным

$$\frac{P(a,b|M)}{P(a,b|R)} = \frac{\prod_{i=1}^{|a^*|} p_{a_i^*,b_i^*}}{\prod_{i=1}^{|a^*|} p_{a_i^*} p_{b_i^*}} = \frac{\prod_{i=1}^{|a^*|} p_{a_i^*,b_i^*}}{p_{a_i^*} p_{b_i^*}}$$

Хотелось бы аддитивную весовую функцию

Родственные к неродственным

$$\frac{P(a,b|M)}{P(a,b|R)} = \frac{\prod_{i=1}^{|a^*|} p_{a_i^*,b_i^*}}{\prod_{i=1}^{|a^*|} p_{a_i^*} p_{b_i^*}} = \frac{\prod_{i=1}^{|a^*|} p_{a_i^*,b_i^*}}{p_{a_i^*} p_{b_i^*}}$$

Хотелось бы аддитивную весовую функцию

$$S(a^*,b^*) = \sum_{i=1}^{|a^*|} s(a_i^*,b_i^*)$$
, где $s(a_i^*,b_i^*) = log\left(rac{p_{a_i^*,b_i^*}}{p_{a_i^*}p_{b_i^*}}
ight)$

22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

Mutation Probability Matrix for the Evolutionary Distance of One PAM

We can combine information about the individual kinds of mutations and about the relative mutability of the amino acids into one distance-dependent "mutation probability matrix" (see Figure 82). An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM.

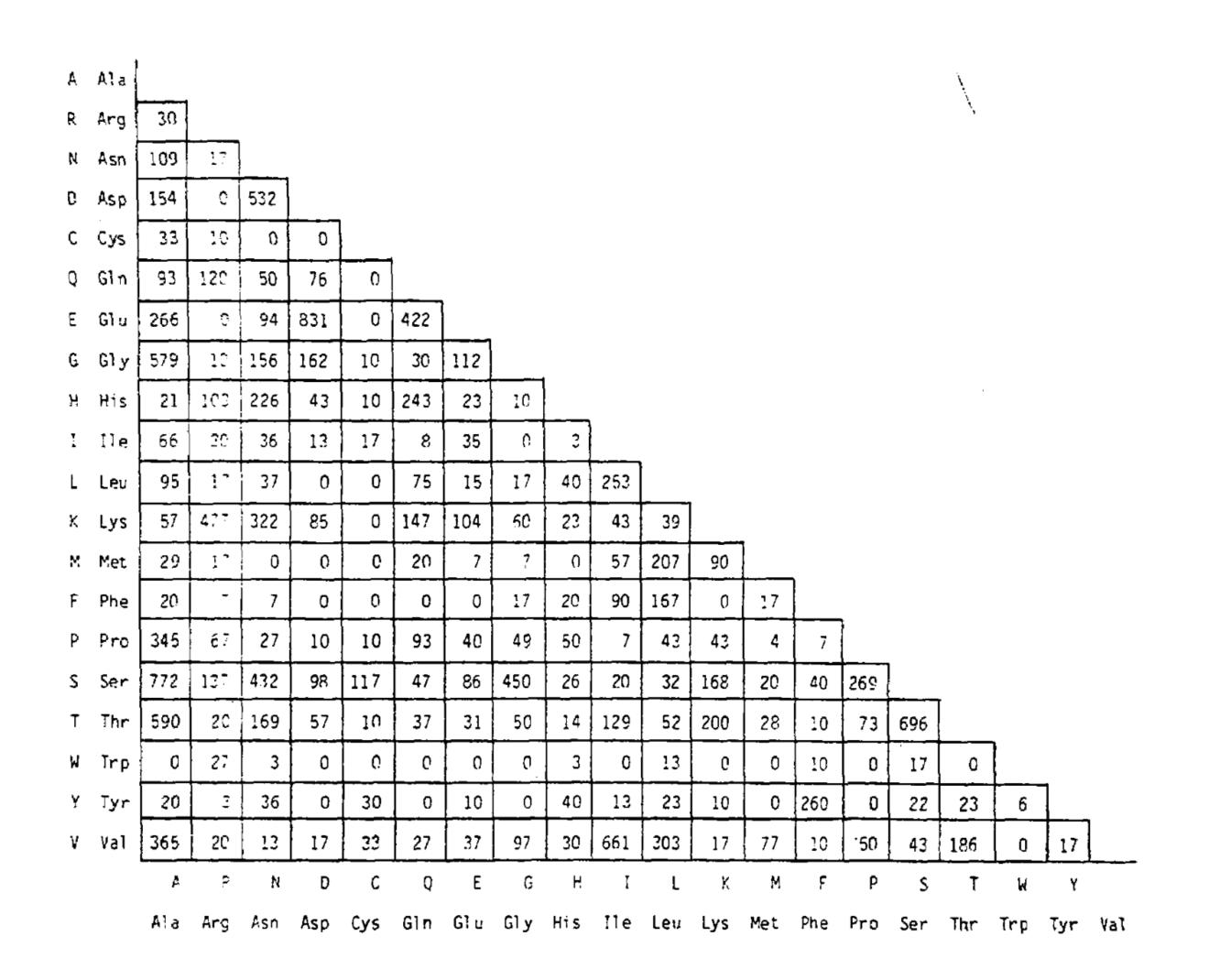


Table 22

Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041	
Ala	0.087	Asn	0.040	
Leu	0.085	Phe	0.040	
Lys	0.081	Gln	0.038	
Ser	0.070	He	0.037	
Val	0.065	His	0.034	
Thr	0.058	Cys	0.033	
Pro	0.051	Tyr	0.030	
Glu	0.050	Met	0.015	
Asp	0.047	Trp	0.010	

 m_j

The nondiagonal elements have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_{i} A_{ij}}$$

where

 A_{ij} is an element of the accepted point mutation matrix of Figure 80,

 λ is a proportionality constant, and m_j is the mutability of the jth amino acid, Table 21.

The diagonal elements have the values:

$$M_{jj} = 1 - \lambda m_j$$

_																						
ſ			A	R	N	D	С	Q	E	G	Н	I	L	K	М	F	Р	S	T	W	Y	٧
			Ala	Arg	Asn	Asp	Cys	Gln	61 u	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	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	ı	8	٥	1
	N	Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Ì	D	Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	c	o	1
1	C	Cys	1	1	0	0	9973	0	٥	0	1	. 1	0	0	0	0	1	5	1	٥	3	2
	Q	Gin	3	9	4	5	0	9876	27	1	23	1	3	6	4	o	6	2	2	o	0	1
	Ε	GT u	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
9	G	Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
AMINO ACID	Н	His	1	9	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
AMI	I	Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
1ENT	L	Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
REPLACEM	K	Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	O	1	1
REPI	М	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	o	0	4
İ	F	Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
.	Þ	Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	٥	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	O	0	0	0	0	1	0	1	0	9976	1	0
ļ	Y	Ţyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
	٧	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

$$^{\circ}$$
 $D \to E$ — правдоподобно

 $^{\circ}$ I
ightarrow V— правдоподобно

 $^{\circ}$ D o V— не очень

ſ			А	R	N	D	С	Q	Ε	G	н	I	L	K	М	F	Р	S	T	W	Υ	٧
			Ala	Arg	Asn	Asp	Cys	Gln	G] u	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	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
Ì	D	Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
١	С	Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
1	Q	Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
	Ε	Glu	10	0	,	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
- 1	I	I le	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	Đ	1	1
	М	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
١	F	Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
` 	Þ	Pro	13	5	Ì		1	8	3		1	1	2	2	1	1	9926	12	4	0	0	2
	S	Şer	١				11	4	6	16	2	2	1	7	4	3	17	9840	l	5	2	2
	Ţ	Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	. 32	9871	0:		9
	W	Trp	0	2				0	0	0	0	0	0	1		1	0	1	0	9976	1	0
}	Y	Ţyr	1	0	3	0	3	0	1	0		1	1	0		21	0	1	1	1	9945	Ì
į	V	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

$$^{\circ}$$
 $D \to E$ — правдоподобно

$$\circ$$
 $I o V$ — правдоподобно

$$^{\circ}$$
 $D o V$ — не очень

]	А	R	N	D	С	Q	E	G	Н	I	L	K	М	F	Р	S	T	W	Y	٧
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	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	ı	8	٥	1
N	Asn	4	1	9822	36	o	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	c	О	1
С	Cys	1	1	0	0	9973	0	0	. 0	1	1	0	0	0	0	1	5	1	0	3	2
Q	Gin	3	9	4	5	0	9876	27	1	23	1	3	6	4	o	6	2	2	o	0	1
Ε	G1 u	10	0	1	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	Gly	21	1	12	îì	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	٥	5
H	His	1	8	18	3	1	20	1	0	9912	0	1;	1	0	2	3	1	1	1	4	1
I	Пe	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	O	1	1
M	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	o	0	4
F	Phe	1	1	1	٥	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Þ	Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	٥	2
S	Şer	28	11.	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Ţ	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	O	0	0	0	0	1	0	1	0	9976	1	0
Y	Ţyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
٧	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901
	R N D C Q E G H I L K M F P S T W Y	R Arg N Asn D Asp C Cys Q Gln E Gly H is I leu K Lys M Met F Phe P Pro S Ser T Thr W Tyr	Ala A Ala 9867 R Arg 1 N Asn 4 D Asp 6 C Cys 1 Q Gln 3 E Glu 10 G Gly 21 H His 1 I Ile 2 L Leu 3 K Lys 2 M Met 1 P Pro 13 S Ser 28 T Thr 22 W Trp 0 Y Tyr 1	Ala Arg A Ala 9867 2 R Arg 1 9913 N Asn 4 1 D Asp 6 0 C Cys 1 1 Q Gln 3 9 E Glu 10 0 G Gly 21 1 H His 1 8 I Ile 2 2 L Leu 3 1 K Lys 2 37 M Met 1 1 F Phe 1 1 F Phe 1 1 F Phe 1 1 F Phe 1 5 S Ser 28 11 T Thr 22 2 W Trp 0 2 Y Tyr 1 0	Ala Arg Asn A Ala 9867 2 9 R Arg 1 9913 1 N Asn 4 1 9822 D Asp 6 0 42 C Cys 1 1 0 Q Gin 3 9 4 E Glu 10 0 1 G Gly 21 1 12 H His 1 8 18 I Ile 2 2 3 L Leu 3 1 3 K Lys 2 37 25 M Met 1 1 0 F Phe 1 1 1 P Pro 13 5 2 S Ser 28 11 34 T Thr 22 2 13 W Trp 0 2 0 Y Tyr 1 0 3	Ala Arg Asn Asp A Ala 9867 2 9 10 R Arg 1 9913 1 0 N Asn 4 1 9822 36 D Asp 6 0 42 9859 C Cys 1 1 0 0 G Gly 21 1 12 11 H His 1 8 18 3 I Ile 2 2 3 1 L Leu 3 1 3 0 K Lys 2 37 25 6 M Met 1 1 0 0 F Phe 1 1 0 0 F Phe 1 1 0 0 P Pro 13 5 2 1 S Ser 28 11 34 7 T Thr 22 2 13 4 W Trp 0 2 0 0 Y Tyr 1 0 3 0	Ala Arg Asn Asp Cys A Ala 9867 2 9 10 3 R Arg 1 9913 1 0 1 N Asn 4 1 9822 36 0 D Asp 6 0 42 9859 0 C Cys 1 1 0 0 9973 Q Gln 3 9 4 5 0 E Glu 10 0 56 0 G Gly 21 1 12 11 1 H His 1 8 18 3 1 I Ile 2 2 3 1 2 L Leu 3 1 3 0 0 K Lys 2 37 25 6 0 M Met 1 1 0 0 0 F Phe 1 1 1 0 0 P Pro 13 5 2 1 1 S Ser 28 11 34 7 11 T Thr 22 2 13 4 1 W Trp 0 2 0 0 0 Y Tyr 1 0 3 0 3	Ala Arg Asn Asp Cys Gln A Ala 9867 2 9 10 3 8 R Arg 1 9913 1 0 1 10 N Asn 4 1 9822 36 0 4 D Asp 6 0 42 9859 0 6 C Cys 1 1 0 0 9973 0 Q Gln 3 9 4 5 0 9876 E Glu 10 0 56 0 35 G Gly 21 1 12 11 1 3 H His 1 8 18 3 1 20 I Ile 2 2 3 1 2 1 L Leu 3 1 3 0 0 6 K Lys 2 37 25 6 0 12 M Met 1 1 0 0 0 2 F Phe 1 1 1 0 0 0 2 F Phe 1 1 1 0 0 0 0 P Pro 13 5 2 1 1 8 S Ser 28 11 34 7 11 4 T Thr 22 2 13 4 1 3 W Trp 0 2 0 0 0 0 Y Tyr 1 0 3 0 3 0	Ala Arg Asn Asp Cys Gln Glu A Ala 9867	Ala Arg Asn Asp Cys Gln Glu Gly A Ala 9867 2 9 10 3 8 17 21 R Arg 1 9913 1 0 1 10 0 0 N Asn 4 1 9822 36 0 4 6 6 D Asp 6 0 42 9859 0 6 53 6 C Cys 1 1 0 0 9973 0 0 0 Q Gln 3 9 4 5 0 9876 27 1 E Glu 10 0 7 56 0 35 9865 4 G Gly 21 1 12 11 1 3 7 9935 H His 1 8 18 3 1 20 1 0 I Ile 2 2 3 1 2 1 2 0 L Leu 3 1 3 0 0 6 1 1 K Lys 2 37 25 6 0 12 7 2 M Met 1 1 0 0 0 2 0 0 F Phe 1 1 0 0 0 2 0 0 F Phe 1 1 1 0 0 0 0 2 0 0 S Ser 28 11 34 7 11 4 6 16 T Thr 22 2 13 4 1 3 2 2 W Trp 0 2 0 0 0 0 0 0 0 Y Tyr 1 0 3 0 3 0 1 0	Ala Arg Asn Asp Cys Gln Glu Gly His A Ala 9867 2 9 10 3 8 17 21 2 R Arg 1 9913 1 0 1 10 0 0 10 N Asn 4 1 9822 36 0 4 6 6 21 D Asp 6 0 42 9859 0 6 53 6 4 C Cys 1 1 0 0 9973 0 0 0 1 Q Gln 3 9 4 5 0 9876 27 1 23 E Glu 10 0 56 0 35 9865 4 2 G Gly 21 1 12 11 1 3 7 9935 1 H His 1 2 18 3 1 20 1 0 9912 I Ile 2 2 3 1 2 1 2 0 0 L Leu 3 1 3 0 0 6 1 1 4 K Lys 2 37 25 6 0 12 7 2 2 M Met 1 1 0 0 0 0 2 0 0 0 F Phe 1 1 1 0 0 0 0 2 0 0 F Phe 1 1 1 0 0 0 0 0 0 1 2 P Pro 13 5 2 1 1 8 3 2 5 S Ser 28 11 34 7 11 4 6 16 2 T Thr 22 2 13 4 1 3 2 2 1 W Trp 0 2 0 0 0 0 0 0 0 0 Y Tyr 1 0 3 0 3 0 1 0 4	Ala Arg Asn Asp Cys Gln Glu Gly His Ile A Ala 9867	Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu A Ala 9867	Ala Arg Asn Asp Cys Gin Giu Giy His Ile Leu Lys A Ala 9867	Ala Arg Asn Asp Cys Gin Giu Giy His Ile Leu Lys Met A Ala 9867	Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe A Ala 9867	Ala Arg Asn Asp Cys Gin Giu Gly His Ile Leu Lys Met Phe Pro A Ala 9867	Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser A Ala 9867	Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr A Ala 9867	A Ala 9867	A Ala 9867

$$^{\circ}(D \to E)$$
— правдоподобно

$$\circ$$
 $I o V$ — правдоподобно

$${}^{\circ}$$
 $D o V$ — не очень

ſ			Α	R	N	D	С	Q	Ε	G	Н	I	L	K	М	F	Р	S	T	W	Υ	٧
			Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	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	ı	8	٥	1
	N	Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Ì	D	Asp	6	٥	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	o	0	1
İ	С	Cys	1	1	0	0	9973	0	0	0	1	1	0	٥	0	0	1	5	1	٥	3	2
	Q	Gin	3	9	4	5	0	9876	27	1	23	1	3	6	4	o	6	2	2	٥	0	1
	٤	GT u	10	0	-	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
ACTD	G	Gly	21	1	12	11	1	3	7	9935	1	٥	1	2	1	1	3	21	3	0	٥	5
O AC	Н	His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
AMINO	I	Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	٥	1	33
MENT	L	Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
REPLACE	K	Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	O	1	1
REPI	M	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	o	0	4
İ	F	Phe	1	1	1	٥	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	٥	2
	S	Ser	28	11.	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
	Ţ	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	O	0	0	0	0	0	O	0	0	0	0	1	0	1	0	9976	1	0
	Y	Ţyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	ì
	٧	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

PAM (point accepted mutation) и BLOSUM

- 1. База выравниваний BLOCS. Белки, разбитые на блоки [Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks."(1992)]
- 2. Кластеризация. $s_1, s_2 \in C \Leftrightarrow \frac{\#(s_{1,i} = s_{2,i})}{|s_1|} > L$
- 3. Частоты встречаемости.

Рассмотрим символы
$$ch_1, ch_2$$
 из разных кластеров. Пусть $ch_1 \in C_n^*, ch_2 \in C_m^*$, вычислим $A_{a,b} = \frac{\#(pos(a) = pos(b))}{|C_n^*| |C_m^*|}$

Тут C^* это кластер, в котором находятся символы из строк $\in C$

4. Как пользуясь $A_{a,b}$ вычислить вероятности $p_a, p_{a,b}$?

PAM и BLOSUM

- 1. База выравниваний BLOCS.
- 2. Кластеризация.
- 3. Частоты встречаемости.
- 4. Вероятности $p_a, p_{a,b}$

$$p_a = rac{\sum_b A_{a,b}}{\sum_{c,d} A_{c,d}}$$
 — символ a выровнялся для разных C_n

$$p_{a,b} = rac{A_{a,b}}{\sum_{c,d} A_{c,d}}$$
 — часть тех выравниваний где выровнялись a,b

PAM и BLOSUM

- 1. База выравниваний BLOCS.
- 2. Кластеризация.
- 3. Частоты встречаемости.
- 4. Вероятности $p_a, p_{a,b}$
- 5. Воспользуемся функцией s(a,b), чтобы получить матрицу замен!

$$s(a,b) = log\left(\frac{p_{a,b}}{p_a p_b}\right)$$

BLOSUM. Замечания.

1. Существует BLOSUM65, BLOSUM50. В чем разница?

BLOSUM. Замечания.

- 1. Существует BLOSUM62, BLOSUM50. В чем разница? Параметр L используемый для кластеризации.
- 2. Чему соответствуют меньшие/большие значения L?

BLOSUM. Замечания.

- 1. Существует BLOSUM62, BLOSUM50. В чем разница? Параметр L используемый для кластеризации.
- 2. Чему соответствуют меньшие/большие значения L? Меньшие значения L соответствуют большим эволюционным временам.
- 3. BLOSUM50 работает для выравниваний с разрывами лучше чем BLOSUM62. [Paerson 1996]

Штрафы за гэпы!

GAAATT_CT

```
GAAATT_CT
```

- В примерах выше цена выравнивания одинаковая.
- Но первое "биологически адекватнее"! Два маленьких гэпа происходят менее вероятно чем один, но длинны 2.
- о Что делать?

G_ATTACA
GAAATT_CT

```
GAAATT_CT
```

- В примерах выше цена выравнивания одинаковая.
- Но первое "биологически адекватнее"! Два маленьких гэпа происходят менее вероятно чем один, но длинны 2.
- о Что делать? Использовать аффинный штраф за гэп!

- 1. Нам нужна субаддитивная функция штрафа за гэпы:
 - $g: \mathbb{N} \to \mathbb{R}$, причем $g(n+m) \le g(n) + g(m)$
- 2. \triangleleft выравнивание (a^*, b^*) и мультимножество подстрок в нем, содержащих только гэпы Δ .
 - Вес выравнивания со штрафом за гэпы g и функцией веса замен w

$$W_{w,g}(a^*,b^*) = \sum_{i=1,a_i \neq (),b_i \neq ()}^{|a^*|} w(a_i^*,b_i^*) + \sum_{x \in \Delta} g(|x|)$$

$$W_{w,g}(a^*, b^*) = \sum_{i=1, a_i \neq (), b_i \neq ()}^{|a^*|} w(a_i^*, b_i^*) + \sum_{x \in \Delta} g(|x|)$$

На предыдущем примере:

$$a_1^*,b_1^*$$
 = (G__ATTACA,GAAATT_CT), (__,_) - мультимножество гэпов a_2^*,b_2^* =(G_A_TTACA,GAAATT_CT) => (_,_,_) - мультимножество гэпов $W_{w,g}(a_2^*,b_2^*)-W_{w,g}(a_1^*,b_1^*)=(g(1)+g(1)+g(1))-(g(2)+g(1))\geq 0$

Чтобы посчитать $D_{w,g}(a,b)$, где |a|=n, |b|=m построим матрицу D, Dim(D)=(n+1,m+1) так:

$$O_{0,0} = 0$$

$$\circ D_{i,0} = g(i)$$

$$O_{0,j} = g(j)$$

о
$$D_{i,j} = min \begin{cases} D_{i-1,j-1} + w(a_i,b_i) \text{ (замена)} \\ \min_{k=1}^i D_{i-k,j} + g(k) \text{ (удаление k символов)} \\ \min_{k=1}^j D_{i,j-k} + g(k) \text{ (вставка k символов)} \end{cases}$$

• Что хорошего в предыдущем алгоритме?

- \circ Что хорошего в предыдущем алгоритме? Можно использовать вообще для любых w,g
- Что плохого?

- \circ Что хорошего в предыдущем алгоритме? Можно использовать вообще для любых w,g
- \circ Что плохого? Сложность $O(n^3)$:(

Аффиные штрафы за гэпы.

- ^о Используем аффинную функцию g $g(k) = \alpha + \beta k$, штраф за начало гэпа α , а за его продолжение β
- \circ Можно использовать алгоритм Гота (Gotoh). Сложность $O(n^2)$

$$D_{m,n} = \text{Min} [D_{m-1,n-1} + d(\mathbf{a}_m, \mathbf{b}_n), P_{m,n}, Q_{m,n}],$$
 (1)

where

$$P_{m,n} = \min_{1 \le k \le m} [D_{m-k,n} + w_k]$$
 (2)

705

0022-2836/82/350705-04 \$03.00/0

© 1982 Academic Press Inc. (London) Ltd.

706 O. GOTOH

and

$$Q_{m,n} = \min_{1 \le k \le n} [D_{m,n-k} + w_k]. \tag{3}$$

Although $P_{m,n}$ (or $Q_{m,n}$) appears to be calculated in m-1 (or n-1) steps, it can be obtained in a single step according to the following recursion relations:

$$P_{m,n} = \operatorname{Min} \left[D_{m-1,n} + w_1, \operatorname{Min}_{2 \le k \le m} (D_{m-k,n} + w_k) \right]$$

$$= \operatorname{Min} \left[D_{m-1,n} + w_1, \operatorname{Min}_{1 \le k \le m-1} (D_{m-1-k,n} + w_{k+1}) \right]$$

$$= \operatorname{Min} \left[D_{m-1,n} + w_1, \operatorname{Min}_{1 \le k \le m-1} (D_{m-1-k,n} + w_k) + u \right]$$

$$= \operatorname{Min} \left[D_{m-1,n} + w_1, P_{m-1,n} + u \right]$$

$$= \operatorname{Min} \left[D_{m-1,n} + w_1, P_{m-1,n} + u \right]$$

$$(4)$$

and

$$Q_{m,n} = \text{Min} \left[D_{m,n-1} + w_1, Q_{m,n-1} + u \right]. \tag{5}$$

Кроме матрицы D, Dim(D) = (n+1, m+1) добавим еще матрицы A, B такого же размера.

- о $A_{i,i}$ цена лучшего выравнивания $a_{1..i}, b_{1..i}$, которое заканчивается удалением.
- о $B_{i,i}$ цена лучшего выравнивания $a_{1..i}, b_{1..i}$, которое заканчивается вставкой.

о
$$A_{i,j} = min \begin{cases} A_{i-1,j} + \beta \text{ (расширение удаления)} \\ D_{i-1,j} + g(1) \text{ (начало удаления)} \end{cases}$$

о
$$B_{i,j} = min \begin{cases} B_{i,j-1} + \beta \text{ (расширение вставки)} \\ D_{i,j-1} + g(1) \text{ (начало вставки)} \end{cases}$$

о
$$D_{i,j} = min egin{cases} D_{i-1,j-1} + w(a_i,b_i) \text{ (замена)} \\ A_{i,j} \\ B_{i,j} \end{cases}$$

^o Сложность $O(n^2)$ по времени и памяти.

Резюмируем

- Выравнивания последовательностей дают наглядное представление об эволюции.
- о Важно то, как именно вычислять стоимость замен.
- Выравнивание с аффинными гэпами вычислять не более трудно, чем с обычными.