

# **Парное выравнивание**

**Алгоритмы в биоинформатике**

**Антон Елисеев**

**[eliseevantoncoon@gmail.com](mailto: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.
- Сравнивать участки генома можно достаточно эффективно.

# Что будет на этой лекции?

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

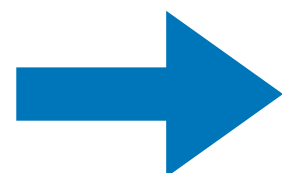
# Расстояние и выравнивание

GATTACA

GATTACA

# Расстояние и выравнивание

GATTACA      GATTACA  
GATT~~A~~CA      GATTCA



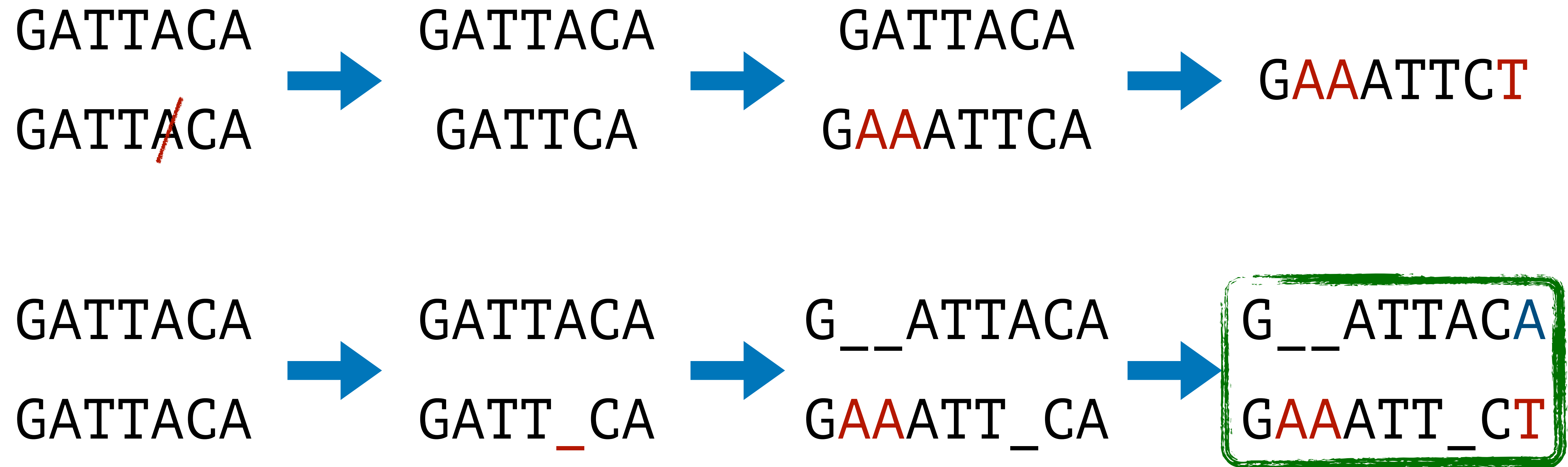
# Расстояние и выравнивание

GATTACA → GATTACA → GATTACA  
GATT~~A~~CA GATTCA GAATTCA

# Расстояние и выравнивание

GATTACA → GATTACA → GATTACA → GAAATTCT  
GATTACA GATTCA GAAATTCA

# Расстояние и выравнивание





# Выравнивание



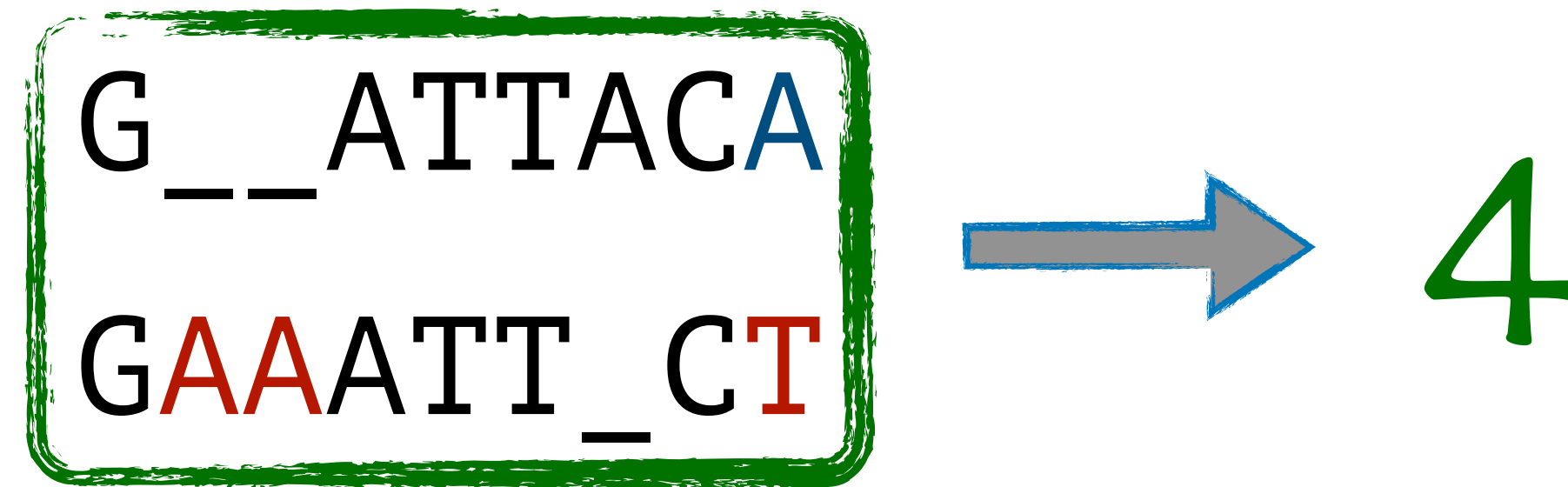
G\_\_ATTACA  
GAATT\_CT

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.  $a_i^* \neq \_ \text{ или } b_i^* \neq \_$
3. При удалении всех гэпов из  $a^*, b^*$  получаем  $a, b$

# Стоимость выравнивания

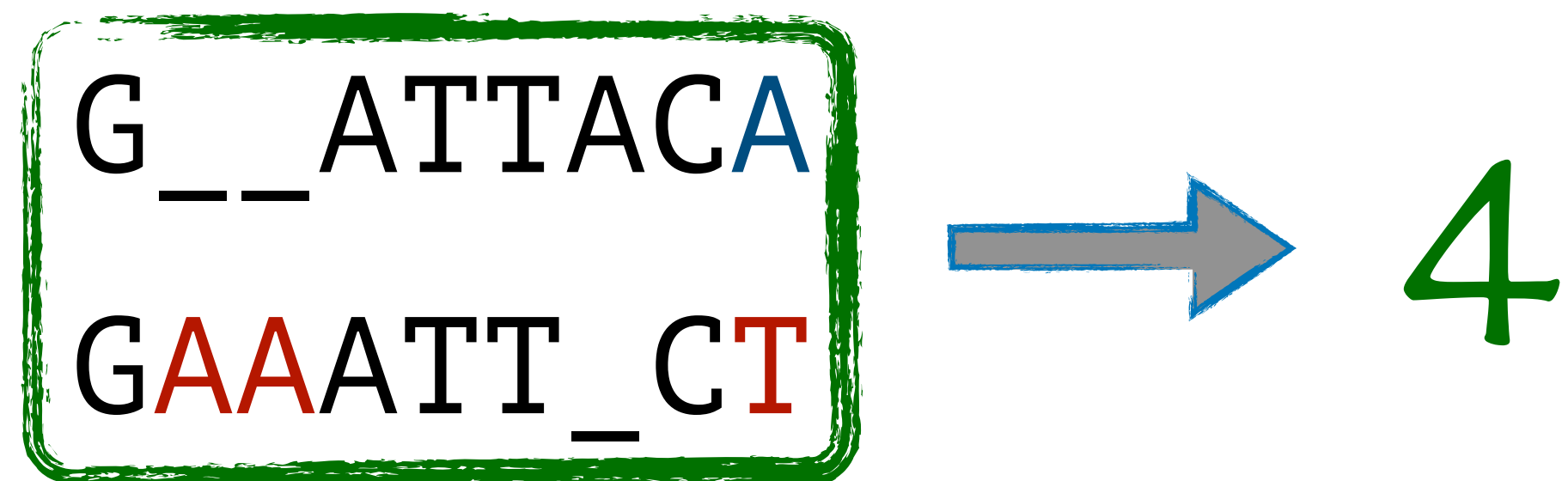


Стоимостью (весом) выравнивания будем называть

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

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

# 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 редактирования  $d_w(a, b)$  равно расстоянию выравнивания  $D_w(a, b)$ .

**Идея:**

# Редактирование и выравнивание

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

## Идея:

- $d_w(a, b) \leq D_w(a, b)$ : выравнивание кодирует последовательность операций редактирования.
- $d_w(a, b) \geq D_w(a, b)$ : последовательность операций редактирования порождает выравнивание такое же по весу либо меньше.

# Расстояние выравнивания



--GATTACA  
AAGAGTAC\_

# Расстояние выравнивания

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

- $D_{0,0} = 0$
- $D_{i,0} = D_{i-1,0} + w(a_i, \_)$
- $D_{0,j} = D_{0,j-1} + w(\_, b_j)$
- $D_{i,j} = \min \begin{cases} D_{i-1,j-1} + w(a_i, b_j) \text{ (замена)} \\ D_{i-1,j} + w(a_i, \_) \text{ (удаление)} \\ D_{i,j-1} + w(\_, b_j) \text{ (вставка)} \end{cases}$

# Замены не равноценны!

Рассмотрим выравнивание аминокислотных последовательностей

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

Гидрофобные: **I** (Изолейцин), **V** (Валин)

◦  $D \rightarrow E - ?$

◦  $I \rightarrow V - ?$

◦  $D \rightarrow V - ?$



# Замены не равноценны!

Рассмотрим выравнивание аминокислотных последовательностей

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

Гидрофобные: **I** (Изолейцин), **V** (Валин)

- $D \rightarrow E$  — правдоподобно
- $I \rightarrow V$  — правдоподобно
- $D \rightarrow V$  — не очень то и правдоподобно

# Замены не равноценны!

Как быть?

# Замены не равноценны!

Как быть?

Хотелось бы отличать случайные матчи от вероятных

# Замены не равноценны!

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  $M$ .

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

# Замены не равноценны!

Родственные к неродственным

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

Хотелось бы аддитивную весовую функцию

# Замены не равноценны!

Родственные к неродственным

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

Хотелось бы аддитивную весовую функцию

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

# Откуда узнать вероятности?

## **22 A Model of Evolutionary Change in Proteins**

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

# Откуда узнать вероятности?

## **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.



# Откуда узнать вероятности?

A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
H	His	21	100	226	43	10	243	23	10												
I	Ile	66	20	36	13	17	8	35	0	3											
L	Leu	95	17	37	0	0	75	15	17	40	253										
K	Lys	57	477	322	85	0	147	104	60	23	43	39									
M	Met	29	17	0	0	0	20	7	7	0	57	207	90								
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

**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	Ile	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

 $m_j$ 
$$A_{i,j}$$

# Откуда узнать вероятности?

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,  
 $\lambda$  is a proportionality constant, and  
 $m_j$  is the mutability of the  $j$ th amino acid, Table 21.

The diagonal elements have the values:

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

ORIGINAL AMINO ACID																					
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
	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	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	1
V	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

# Откуда узнать вероятности?

- $D \rightarrow E$  — правдоподобно
- $I \rightarrow V$  — правдоподобно
- $D \rightarrow V$  — не очень

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		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	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	1
V	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

# Откуда узнать вероятности?

- $D \rightarrow E$  — правдоподобно
- $I \rightarrow V$  — правдоподобно
- $D \rightarrow V$  — не очень

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		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	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	1
V	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901



# Откуда узнать вероятности?

- $D \rightarrow E$  — правдоподобно
- $I \rightarrow V$  — правдоподобно
- $D \rightarrow V$  — не очень

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		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	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	1
V	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 $p_a, p_{a,b}$ ?

# РАМ и BLOSUM

1. База выравниваний BLOCS.

2. Кластеризация.

3. Частоты встречаемости.

4. Вероятности  $p_a, p_{a,b}$

$$p_a = \frac{\sum_b A_{a,b}}{\sum_{cd} A_{c,d}} \text{ — символ } a \text{ выравнился для разных } C_n$$

$$p_{a,b} = \frac{A_{a,b}}{\sum_{cd} A_{c,d}} \text{ — часть тех выравниваний где выравнились } a, b$$

# РАМ 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!



G\_\_ATTACA  
GAAATT\_CT



G\_A\_TTACA  
GAAATT\_CT

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

# Штрафы за гэпы!



G\_\_ATTACA  
GAAATT\_CT



G\_A\_TTACA  
GAAATT\_CT

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

# Штрафы за гэпы!

1. Нам нужна субаддитивная функция штрафа за гэпы:

$$g : \mathbb{N} \rightarrow \mathbb{R}, \text{ причем } g(n + m) \leq g(n) + g(m)$$

2.  $\nless$  выравнивание  $(a^*, b^*)$  и мультимножество подстрок в нем, содержащих только гэпы  $\Delta$ .

Вес выравнивания со штрафом за гэпы  $g$  и функцией веса замен  $w$

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

# Штрафы за гэпы!

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

На предыдущем примере:

$a_1^*, b_1^* = (G\_ATTACA, GAAATT\_CT)$ ,  $(\_ , \_)$  - мультимножество гэпов

$a_2^*, b_2^* = (G\_A\_TTACA, GAAATT\_CT) \Rightarrow (\_ , \_ , \_)$  - мультимножество гэпов

$$W_{w,g}(a_2^*, b_2^*) - W_{w,g}(a_1^*, b_1^*) = (g(1) + g(1) + g(1)) - (g(2) + g(1)) \geq 0$$

# Штрафы за гэпы

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

- $D_{0,0} = 0$

- $D_{i,0} = g(i)$

- $D_{0,j} = g(j)$

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



# Штрафы за гэпы

- Что хорошего в предыдущем алгоритме?

# Штрафы за гэпы

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

# Штрафы за гэпы

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

# Аффинные штрафы за гэпы.

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

# Алгоритм Гота

$$D_{m,n} = \text{Min} [D_{m-1,n-1} + d(a_m, b_n), P_{m,n}, Q_{m,n}], \quad (1)$$

where

$$P_{m,n} = \text{Min}_{1 \leq k \leq m} [D_{m-k,n} + w_k] \quad (2)$$

705

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

© 1982 Academic Press Inc. (London) Ltd.

# Алгоритм Гота

706

O. GOTOH

and

$$Q_{m,n} = \text{Min}_{1 \leq k \leq n} [D_{m,n-k} + w_k]. \quad (3)$$

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:

$$\begin{aligned} P_{m,n} &= \text{Min} [D_{m-1,n} + w_1, \text{Min}_{2 \leq k \leq m} (D_{m-k,n} + w_k)] \\ &= \text{Min} [D_{m-1,n} + w_1, \text{Min}_{1 \leq k \leq m-1} (D_{m-1-k,n} + w_{k+1})] \\ &= \text{Min} [D_{m-1,n} + w_1, \text{Min}_{1 \leq k \leq m-1} (D_{m-1-k,n} + w_k) + u] \\ &= \text{Min} [D_{m-1,n} + w_1, P_{m-1,n} + u] \end{aligned} \quad (4)$$

and

$$Q_{m,n} = \text{Min} [D_{m,n-1} + w_1, Q_{m,n-1} + u]. \quad (5)$$

# Алгоритм Гота

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

- $A_{i,j}$  — цена лучшего выравнивания  $a_{1..i}, b_{1..j}$ , которое заканчивается удалением.

- $B_{i,j}$  — цена лучшего выравнивания  $a_{1..i}, b_{1..j}$ , которое заканчивается вставкой.

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

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

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

# Алгоритм Гота

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



# Резюмируем

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