

Matrizes de *score*

- As matrizes de *score*, ou matrizes de substituição, permitem obter um score para a comparação de cada um dos possíveis pares de aminoácidos
- Os valores destas matrizes expressam as diferentes tendências de conservação dos aminoácidos em posições **homólogas** de duas sequências
- Existem diferentes tipos de matrizes de *score*, baseados em diferentes análises e diferentes pressupostos sobre os mecanismos de substituição
- O processo de inserção (criação de *gaps*) é geralmente tratado separadamente (não há scores nas matrizes para alinhamento com *gaps*)

- Os dois tipos de matrizes mais usados são:

Matrizes PAM: baseadas na comparação, por alinhamento global, de famílias de sequências muito próximas

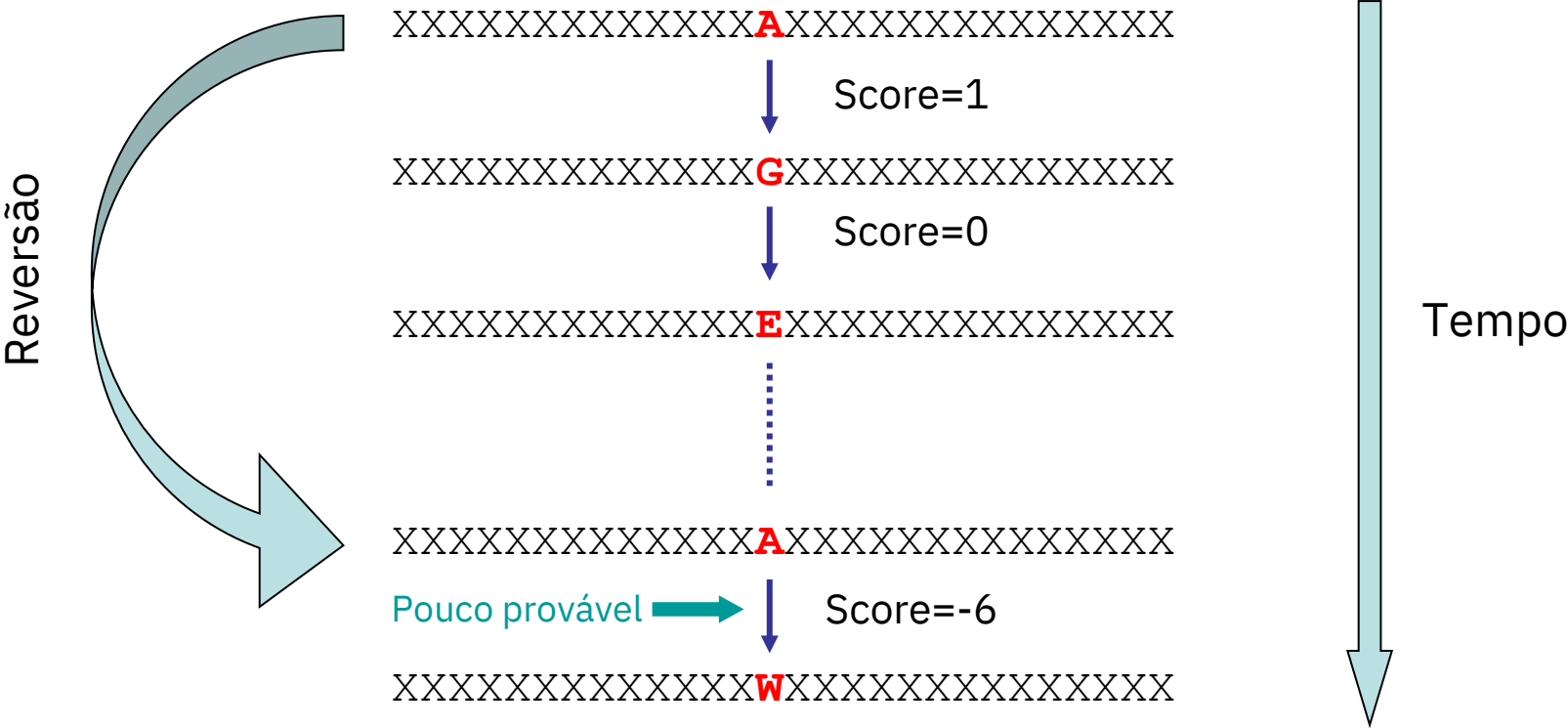
Matrizes BLOSUM: baseadas no alinhamento de regiões de elevada similaridade (blocos) em diferentes grupos de proteínas.

Percentagens de identidade para o alinhamento do cit b de primatas

Percent Identity Matrix

<input type="checkbox"/> sp Q34341 CYB_DAUMA	100.00%	86.02%	83.38%	85.49%	83.91%	82.59%	83.11%	82.85%	81.27%	79.16%	76.78%	80.74%	78.89%
<input type="checkbox"/> sp Q34876 CYB_LEMCA	86.02%	100.00%	91.03%	92.88%	90.24%	89.71%	84.17%	84.17%	78.10%	77.57%	74.93%	80.21%	78.89%
<input type="checkbox"/> sp Q9G946 CYB_PROCO	83.38%	91.03%	100.00%	97.89%	87.86%	87.60%	82.32%	81.79%	76.52%	75.46%	73.09%	77.84%	78.10%
<input type="checkbox"/> sp Q35677 CYB_PROTA	85.49%	92.88%	97.89%	100.00%	89.71%	89.71%	84.43%	83.91%	78.36%	77.31%	75.20%	79.68%	79.68%
<input type="checkbox"/> sp Q9G0S9 CYB_MICMU	83.91%	90.24%	87.86%	89.71%	100.00%	90.77%	82.59%	83.91%	77.84%	77.04%	74.14%	79.16%	78.89%
<input type="checkbox"/> sp Q3YLC2 CYB_MIRZA	82.59%	89.71%	87.60%	89.71%	90.77%	100.00%	82.59%	83.64%	77.84%	77.57%	74.14%	78.89%	76.52%
<input type="checkbox"/> sp O47488 CYB_CEPBA	83.11%	84.17%	82.32%	84.43%	82.59%	82.59%	100.00%	84.70%	78.63%	77.31%	75.20%	77.84%	78.36%
<input type="checkbox"/> sp Q35131 CYB_NYCCO	82.85%	84.17%	81.79%	83.91%	83.91%	83.64%	84.70%	100.00%	79.95%	77.31%	76.52%	78.63%	79.42%
<input type="checkbox"/> sp Q2Y067 CYB_AOTAI	81.27%	78.10%	76.52%	78.36%	77.84%	77.84%	78.63%	79.95%	100.00%	87.60%	76.78%	81.53%	79.16%
<input type="checkbox"/> sp Q35930 CYB_SAISS	79.16%	77.57%	75.46%	77.31%	77.04%	77.57%	77.31%	77.31%	87.60%	100.00%	75.20%	79.42%	78.10%
<input type="checkbox"/> sp Q50DL8 CYB_CHLAE	76.78%	74.93%	73.09%	75.20%	74.14%	74.14%	75.20%	76.52%	76.78%	75.20%	100.00%	82.11%	80.00%
<input type="checkbox"/> sp P00156 CYB_HUMAN	80.74%	80.21%	77.84%	79.68%	79.16%	78.89%	77.84%	78.63%	81.53%	79.42%	82.11%	100.00%	86.32%
<input type="checkbox"/> sp O47892 CYB_NOMLE	78.89%	78.89%	78.10%	79.68%	78.89%	76.52%	78.36%	79.42%	79.16%	78.10%	80.00%	86.32%	100.00%

Evolução por mutações sucessivas

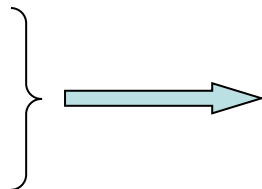


Reversão: o aminoácido numa determinada posição pode mutar sucessivas vezes e voltar a dar origem ao aminoácido inicial.

Distância PAM *versus* % de identidade

$$\text{PAM} \geq 100 - \%id$$

% de identidade
mínima para
conseguir produzir
um alinhamento



% identidade	Unidades PAM
99	1
95	5
90	11
85	17
80	23
75	30
70	38
66	47
60	56
55	67
50	80
45	94
40	112
35	133
30	159
25	195
20	246
15	328

- A percentagem de identidade não reflete exatamente o número de substituições que ocorreram numa determinada posição, pois há mutações “ocultas” mascaradas pela reversão
- O número de idades PAM de distância entre duas sequências é quase sempre superior ao número de substituições visíveis no alinhamento (100-%id).

Probabilidade para a mutação de um aminoácido a PAM1 e a PAM250

- A uma distância de PAM1, a probabilidade da fenilalanina não mutar é 99%, e a probabilidade de mutar num dos outros 19 aminoácidos é ~1%
- A uma distância de PAM250, a probabilidade de a fenilalanina ser conservada é de 32%, e de mutar em um qualquer dos outros aminoácidos é 68%
- As probabilidades individuais de mutação em cada um dos outros 19 aminoácidos são diferentes

Amino acid change	PAM1	PAM250
Phe to Ala	0.0002	0.04
Phe to Arg	0.0001	0.01
Phe to Asn	0.0001	0.02
Phe to Asp	0.0000	0.01
Phe to Cys	0.0000	0.01
Phe to Gln	0.0000	0.01
Phe to Glu	0.0000	0.01
Phe to Gly	0.0001	0.03
Phe to His	0.0002	0.02
Phe to Ile	0.0007	0.05
Phe to Leu	0.0013	0.13
Phe to Lys	0.0000	0.02
Phe to Met	0.0001	0.02
Phe to Phe	0.9946	0.32
Phe to Pro	0.0001	0.02
Phe to Ser	0.0003	0.03
Phe to Thr	0.0001	0.03
Phe to Trp	0.0001	0.01
Phe to Tyr	0.0021	0.15
Phe to Val	0.0001	0.05
SUM ^a	1.0000	1.00

Probabilidade da Fenilalanina não mutar:
 99% (1 PAM)
 32% (250 PAM)

^aApproximate since scores are rounded off.

Matriz PAM 1 – probabilidades de transição

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

(as probabilidades estão multiplicadas por 10000 para maior legibilidade)

Matriz PAM 1 – log odds scores

A	11	-26	-20	-19	-24	-20	-17	-16	-26	-22	-24	-26	-21	-27	-16	-14	-14	###	-26	-17
R	-26	14	-25	###	-26	-16	###	-30	-16	-22	-28	-13	-20	-29	-20	-18	-26	-17	-31	-26
N	-20	-25	14	-10	###	-20	-18	-19	-13	-21	-25	-15	###	-28	-24	-13	-16	-26	-20	-28
D	-19	###	-10	13	###	-19	-9	-19	-21	-26	###	-21	###	###	-29	-20	-22	###	###	-27
C	-24	-26	###	###	15	###	###	-30	-25	-24	###	###	###	###	-27	-18	-28	###	-20	-23
Q	-20	-16	-20	-19	###	14	-11	-25	-12	-27	-21	-18	-19	###	-18	-22	-23	###	###	-25
E	-17	###	-18	-9	###	-11	13	-21	-23	-22	-29	-21	-25	###	-23	-21	-25	###	-27	-24
G	-16	-30	-19	-19	-30	-25	-21	10	-30	###	-31	-26	-28	-28	-25	-16	-25	###	###	-23
H	-26	-16	-13	-21	-25	-12	-23	-30	15	-31	-23	-26	###	-23	-20	-24	-26	-26	-19	-23
I	-22	-22	-21	-26	-24	-27	-22	###	-31	14	-16	-23	-15	-17	-29	-26	-17	###	-24	-11
L	-24	-28	-25	###	###	-21	-29	-31	-23	-16	11	-27	-13	-18	-25	-28	-25	-23	-25	-18
K	-26	-13	-15	-21	###	-18	-21	-26	-26	-23	-27	11	-16	###	-25	-20	-19	###	-29	-30
M	-21	-20	###	###	###	-19	-25	-28	###	-15	-13	-16	18	-20	-28	-22	-20	###	###	-16
F	-27	-29	-28	###	###	###	###	-28	-23	-17	-18	###	-20	14	-29	-23	-29	-21	-12	-29
P	-16	-20	-24	-29	-27	-18	-23	-25	-20	-29	-25	-25	-28	-29	13	-16	-21	###	###	-23
S	-14	-18	-13	-20	-18	-22	-21	-16	-24	-26	-28	-20	-22	-23	-16	11	-13	-21	-25	-25
T	-14	-26	-16	-22	-28	-23	-25	-25	-26	-17	-25	-19	-20	-29	-21	-13	12	###	-24	-18
W	###	-17	-26	###	###	###	###	###	-26	###	-23	###	###	-21	###	-21	###	20	-22	###
Y	-26	-31	-20	###	-20	###	-27	###	-19	-24	-25	-29	###	-12	###	-25	-24	-22	15	-25
V	-17	-26	-28	-27	-23	-25	-24	-23	-23	-11	-18	-30	-16	-29	-23	-25	-18	###	-25	12
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Os valores de score são logaritmos

$$S_{ij} = \log_b (p_i M_{ij} / p_i p_j)$$

S_{ij} = score (“log odds” ratio)

M_{ij} = *score* da matriz de transição (probabilidade observada de substituição do resíduo p_i por p_j)

p_i , p_j = probabilidades de ocorrência dos aminoácidos

b = base do logaritmo (arbitrária)

Odds ratio ($p_i M_{ij} / p_i p_j$) – razão entre a probabilidade de ocorrência de uma transição, e a probabilidade de ocorrência dessa mesma transição num modelo aleatório

G A L H I V H
| | | | | | |
G G V N L V H

$p_1 * p_2 * p_3 * p_4 * p_5 * p_6 * p_7 = P_a$ (probabilidade de ocorrência do alinhamento)

dado que **$\log(a*b) = \log(a) + \log(b)$** temos:

$$\log(P_a) = \log(p_1 * p_2 * p_3 * p_4 * p_5 * p_6 * p_7) = \\ \log(p_1) + \log(p_2) + \log(p_3) + \log(p_4) + \log(p_5) + \log(p_6) + \log(p_7)$$

Assim, se usarmos $\log(p_i)$ como valor de score para cada par de resíduos, a soma destes valores produz o logaritmo do score total $\log(P_a)$!

Família de Matrizes PAM

Família de matrizes de substituição — PAM 1, PAM 2, etc. — onde PAM n é adequada à comparação de sequências que distam entre si de n PAM

$$\text{PAM } n = (\text{PAM } 1)^n$$

$$\text{PAM2} = \text{PAM1} \times \text{PAM1}$$

$$\text{PAM4} = \text{PAM2} \times \text{PAM2}$$

$$\text{PAM8} = \text{PAM4} \times \text{PAM4}$$

$$\text{PAM16} = \text{PAM8} \times \text{PAM8}$$

$$\text{PAM32} = \text{PAM16} \times \text{PAM16}$$

$$\text{PAM64} = \text{PAM32} \times \text{PAM32}$$

$$\text{PAM128} = \text{PAM64} \times \text{PAM64}$$

$$\text{PAM256} = \text{PAM128} \times \text{PAM128}$$

Não confundir com as matrizes PAM de substituição com as matrizes PAM de score.

Matrizes PAM de score são obtidas a partir das matrizes PAM de substituição calculando os logaritmos da razão das frequências observadas e esperadas (“log odds score”).

PAM2

A	10	-23	-17	-16	-21	-17	-14	-13	-23	-19	-21	-23	-18	-24	-13	-11	-11	-46	-23	-14
R	-23	14	-22	-41	-23	-13	-39	-27	-13	-19	-25	-10	-17	-26	-17	-15	-23	-14	-28	-23
N	-17	-22	14	-8	-41	-17	-15	-16	-10	-18	-22	-12	-38	-25	-21	-10	-13	-23	-17	-25
D	-16	-41	-8	13	-47	-16	-6	-16	-18	-23	-44	-18	-42	-48	-26	-17	-19	-48	-42	-24
C	-21	-23	-41	-47	15	-47	-47	-27	-22	-21	-48	-47	-46	-45	-24	-15	-25	-49	-17	-20
Q	-17	-13	-17	-16	-47	14	-9	-22	-9	-24	-18	-15	-16	-45	-15	-19	-20	-45	-43	-22
E	-14	-39	-15	-6	-47	-9	13	-18	-20	-19	-26	-18	-22	-47	-20	-18	-22	-52	-24	-21
G	-13	-27	-16	-16	-27	-22	-18	10	-27	-42	-28	-23	-25	-25	-21	-13	-22	-48	-46	-20
H	-23	-13	-10	-18	-22	-9	-20	-27	15	-28	-20	-22	-41	-20	-17	-21	-23	-23	-16	-21
I	-19	-19	-18	-23	-21	-24	-19	-42	-28	14	-13	-20	-12	-14	-26	-23	-14	-46	-21	-8
L	-21	-25	-22	-44	-48	-18	-26	-28	-20	-13	11	-24	-10	-15	-22	-25	-22	-20	-22	-15
K	-23	-10	-12	-18	-47	-15	-18	-23	-22	-20	-24	11	-13	-46	-22	-17	-16	-43	-26	-27
M	-18	-17	-38	-42	-46	-16	-22	-25	-41	-12	-10	-13	18	-17	-25	-19	-17	-45	-42	-13
F	-24	-26	-25	-48	-45	-45	-47	-25	-20	-14	-15	-46	-17	14	-26	-20	-25	-18	-9	-26
P	-13	-17	-21	-26	-24	-15	-20	-21	-17	-26	-22	-22	-25	-26	13	-13	-18	-46	-46	-20
S	-11	-15	-10	-17	-15	-19	-18	-13	-21	-23	-25	-17	-19	-20	-13	11	-10	-18	-22	-22
T	-11	-23	-13	-19	-25	-20	-22	-22	-23	-14	-22	-16	-17	-25	-18	-10	12	-44	-21	-15
W	-46	-14	-23	-48	-49	-45	-52	-48	-23	-46	-20	-43	-45	-18	-46	-18	-44	20	-19	-49
Y	-23	-28	-17	-42	-17	-43	-24	-46	-16	-21	-22	-26	-42	-9	-46	-22	-21	-19	15	-22
V	-14	-23	-25	-24	-20	-22	-21	-20	-21	-8	-15	-27	-13	-26	-20	-22	-15	-49	-22	12
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

PAM50

A	8	-8	-3	-3	-7	-4	-2	-1	-8	-5	-7	-8	-5	-10	-1	1	1	-16	-9	-2
R	-8	12	-6	-11	-10	-1	-10	-11	-1	-6	-10	2	-4	-11	-4	-3	-7	-1	-12	-9
N	-3	-6	10	3	-12	-3	-1	-3	2	-6	-9	0	-10	-10	-6	1	-1	-10	-5	-9
D	-3	-11	3	10	-17	-2	5	-3	-4	-9	-14	-4	-12	-17	-9	-3	-5	-18	-13	-9
C	-7	-10	-12	-17	14	-17	-17	-11	-9	-7	-18	-17	-16	-15	-10	-3	-9	-19	-4	-7
Q	-4	-1	-3	-2	-17	12	3	-8	3	-9	-6	-2	-4	-15	-3	-6	-6	-15	-14	-8
E	-2	-10	-1	5	-17	3	10	-4	-5	-6	-11	-4	-8	-17	-6	-4	-6	-20	-10	-7
G	-1	-11	-3	-3	-11	-8	-4	9	-10	-12	-13	-8	-10	-11	-7	-1	-6	-18	-16	-6
H	-8	-1	2	-4	-9	3	-5	-10	13	-10	-7	-6	-12	-7	-4	-6	-8	-9	-3	-7
I	-5	-6	-6	-9	-7	-9	-6	-12	-10	12	0	-7	1	-2	-10	-8	-2	-16	-7	4
L	-7	-10	-9	-14	-18	-6	-11	-13	-7	0	10	-10	2	-2	-8	-10	-8	-7	-8	-1
K	-8	2	0	-4	-17	-2	-4	-8	-6	-7	-10	9	-1	-16	-7	-4	-3	-13	-11	-10
M	-5	-4	-10	-12	-16	-4	-8	-10	-12	1	2	-1	16	-4	-9	-6	-4	-15	-13	0
F	-10	-11	-10	-17	-15	-15	-17	-11	-7	-2	-2	-16	-4	13	-12	-8	-10	-5	4	-9
P	-1	-4	-6	-9	-10	-3	-6	-7	-4	-10	-8	-7	-9	-12	11	-1	-4	-16	-16	-6
S	1	-3	1	-3	-3	-6	-4	-1	-6	-8	-10	-4	-6	-8	-1	8	2	-6	-8	-7
T	1	-7	-1	-5	-9	-6	-6	-6	-8	-2	-8	-3	-4	-10	-4	2	10	-15	-8	-2
W	-16	-1	-10	-18	-19	-15	-20	-18	-9	-16	-7	-13	-15	-5	-16	-6	-15	19	-6	-19
Y	-9	-12	-5	-13	-4	-14	-10	-16	-3	-7	-8	-11	-13	4	-16	-8	-8	-6	14	-9
V	-2	-9	-9	-9	-7	-8	-7	-6	-7	4	-1	-10	0	-9	-6	-7	-2	-19	-9	10
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

PAM150

A	4	-3	0	0	-3	-1	0	1	-3	-1	-3	-3	-2	-5	1	2	2	-9	-5	0
R	-3	9	-1	-4	-5	1	-3	-5	2	-3	-5	4	-1	-7	-1	-1	-2	2	-7	-4
N	0	-1	5	3	-6	0	1	0	2	-3	-5	1	-3	-5	-2	1	0	-6	-3	-3
D	0	-4	3	6	-8	1	5	0	0	-4	-7	-1	-5	-9	-3	0	-1	-10	-7	-4
C	-3	-5	-6	-8	13	-8	-8	-6	-5	-4	-9	-9	-8	-7	-5	0	-4	-11	-1	-3
Q	-1	1	0	1	-8	7	3	-3	4	-4	-3	0	-2	-8	0	-2	-2	-7	-6	-3
E	0	-3	1	5	-8	3	6	-1	0	-3	-5	-1	-4	-9	-2	-1	-2	-11	-6	-3
G	1	-5	0	0	-6	-3	-1	7	-4	-5	-7	-3	-5	-7	-2	1	-1	-10	-8	-3
H	-3	2	2	0	-5	4	0	-4	9	-5	-3	-1	-4	-3	-1	-2	-3	-4	-1	-4
I	-1	-3	-3	-4	-4	-4	-3	-5	-5	7	2	-3	2	1	-4	-3	0	-8	-3	5
L	-3	-5	-5	-7	-9	-3	-5	-7	-3	2	8	-5	4	1	-4	-5	-3	-3	-3	1
K	-3	4	1	-1	-9	0	-1	-3	-1	-3	-5	7	1	-8	-3	-1	-1	-6	-7	-5
M	-2	-1	-3	-5	-8	-2	-4	-5	-4	2	4	1	11	-1	-4	-3	-1	-7	-5	2
F	-5	-7	-5	-9	-7	-8	-9	-7	-3	1	1	-8	-1	11	-7	-5	-5	-1	7	-3
P	1	-1	-2	-3	-5	0	-2	-2	-1	-4	-4	-3	-4	-7	8	1	0	-9	-8	-2
S	2	-1	1	0	0	-2	-1	1	-2	-3	-5	-1	-3	-5	1	4	2	-3	-4	-2
T	2	-2	0	-1	-4	-2	-2	-1	-3	0	-3	-1	-1	-5	0	2	5	-8	-4	0
W	-9	2	-6	-10	-11	-7	-11	-10	-4	-8	-3	-6	-7	-1	-9	-3	-8	18	-2	-10
Y	-5	-7	-3	-7	-1	-6	-6	-8	-1	-3	-3	-7	-5	7	-8	-4	-4	-2	12	-4
V	0	-4	-3	-4	-3	-3	-3	-3	-4	5	1	-5	2	-3	-2	-2	0	-10	-4	7
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

PAM256

A	2	-1	0	0	-2	0	0	1	-1	0	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-1	6	0	-1	-4	1	-1	-2	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-5	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	0	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-2	0	1	-3	-1	0	5	-2	-2	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	0	-2	-2	-2	-2	-2	-2	-2	4	2	-2	2	1	-2	-1	0	-5	-1	4	
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-2	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-5	-4	-5	-5	-5	-2	1	2	-5	0	9	-4	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-2	-1	-2	-4	6	1	0	-5	-5	-1
S	1	0	1	0	0	0	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	2	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-5	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

PAM500

A	0	0	0	0	-1	0	0	1	0	0	-1	0	0	-1	0	0	0	-3	-1	0
R	0	2	0	0	-2	1	0	-1	1	-1	-1	2	0	-2	0	0	0	2	-2	-1
N	0	0	0	1	-2	1	1	0	1	-1	-1	0	-1	-2	0	0	0	-2	-1	0
D	0	0	1	1	-2	1	1	1	1	-1	-1	0	-1	-2	0	0	0	-3	-2	-1
C	-1	-2	-2	-2	9	-2	-2	-1	-2	-1	-3	-2	-2	-1	-1	0	-1	-4	1	-1
Q	0	1	1	1	-2	1	1	0	1	-1	-1	1	0	-2	0	0	0	-2	-2	-1
E	0	0	1	1	-2	1	1	0	1	-1	-1	0	-1	-2	0	0	0	-3	-2	-1
G	1	-1	0	1	-1	0	0	2	0	-1	-1	0	-1	-2	0	1	0	-3	-2	0
H	0	1	1	1	-2	1	1	0	2	-1	-1	0	-1	-1	0	0	0	-1	0	-1
I	0	-1	-1	-1	-1	-1	-1	-1	1	2	-1	1	1	-1	0	0	0	-2	0	1
L	-1	-1	-1	-1	-3	-1	-1	-1	-1	2	3	-1	2	2	-1	-1	0	-1	1	1
K	0	2	0	0	-2	1	0	0	0	-1	-1	2	0	-2	0	0	0	-1	-2	-1
M	0	0	-1	-1	-2	0	-1	-1	-1	1	2	0	2	1	-1	-1	0	-2	0	1
F	-1	-2	-2	-2	-1	-2	-2	-2	-1	1	2	-2	1	6	-2	-1	-1	1	5	0
P	0	0	0	0	-1	0	0	0	0	-1	-1	0	-1	-2	2	0	0	-3	-2	0
S	0	0	0	0	0	0	0	1	0	0	-1	0	-1	-1	0	0	0	-1	-1	0
T	0	0	0	0	-1	0	0	0	0	0	0	0	0	-1	0	0	0	-2	-1	0
W	-3	2	-2	-3	-4	-2	-3	-3	-1	-2	-1	-2	1	-3	-1	-2	-2	15	1	-3
Y	-1	-2	-1	-2	1	-2	-2	-2	0	0	1	-2	0	5	-2	-1	-1	1	6	-1
V	0	-1	0	-1	-1	-1	-1	0	-1	1	1	-1	1	0	0	0	0	-3	-1	1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
D	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	-1	-1	0	
C	0	0	0	0	5	0	0	0	0	0	-1	-1	-1	0	0	0	-1	0	0	
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
G	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	-1	-1	0	
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
L	0	0	0	0	-1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	
K	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
M	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
F	0	0	0	-1	0	0	0	-1	0	0	1	0	0	2	0	0	0	1	2	
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	
W	-1	1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	1	-1	0	-1	10	1	-1
Y	0	0	0	-1	0	0	0	-1	0	0	0	0	0	2	0	0	0	1	2	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Exemplo de entrada na base de dados BLOCKS

Block PR00808A

```
ID  AMLASEINHBTR; BLOCK
AC  PR00808A; distance from previous block=(6,42)
DE  Cereal trypsin/alpha-amylase inhibitor family signature
BL  adapted; width=15; seqs=34; 99.5%=771; strength=1203
O49864 ( 29) CAPGDALPHNPLRAC 28
O49865 ( 29) CAPGDALPHNPLRAC 28
IAAE HORVU|P01086 ( 29) CAPGDALPHNPLRAC 28
O49861 ( 29) CAPGDELPHNPLRAC 50
O49863 ( 29) CAPGDALPANPLRAC 30
Q23982 ( 32) CSPGVAFPTNLLGHC 43
IAAD HORVU|P11643 ( 32) CSPGVAFPTNLLGHC 43
O49862 ( 29) CAPGDALPANPLRAC 30
ITRF MAIZE|P01088 ( 34) CVPGWAIPHNPPLPSC 50
Q24000 ( 32) CSPGVAFPTNLHGHC 100
IA03 WHEAT|P17314 ( 29) CVPGVAFRTNLLPHC 63
O49867 ( 29) CAPGDALPANPLRAC 30
IAAB HORVU|P32936 ( 30) CTPWTATPITPLPSC 54
IA16 WHEAT|P16159 ( 30) CTPWMSTLITPLPSC 69
IA02 WHEAT|P16851 ( 30) CYPGMGLPSNPLEGC 39
IAAA HORVU|P28041 ( 30) CYAGMGLPSNPLEGC 57
Q41540 ( 30) CTPWTSTLITPLPSC 71
IAAT ELECO|P01087 ( 6) CIPGMAIPHNPPLDSC 58
RA14 ORYSA|Q01882 ( 41) CQPGMGYPMYSLPRC 35
Q40652 ( 41) CQPGMGYPMYSLPRC 35
RAG2 ORYSA|Q01885 ( 41) CQPGMGYPMYSLPRC 35
IA01 WHEAT|P16850 ( 30) CYAGMGLPINPLEGC 53
P93602 ( 32) CQPGVAFPHNALATC 43
Q40655 ( 42) CQPGMGYPMYPLPRC 30
Q40654 ( 41) CQPGIGYPTYPLPRC 36
RA05 ORYSA|Q01881 ( 41) CQPGMGYPMYSLPRC 35
RA17 ORYSA|Q01883 ( 39) CRPGISYPTYSLPQC 71
IAA HORVU|P16969 ( 33) CQLGVDFPHNPLATC 98
IAA4 SORBI|P81367 ( 7) CAPGLAIPAPPLPTC 56
IAA1 WHEAT|P01085 ( 6) CYPGQAFQVPALPAC 56
O49956 ( 37) CDPATGYKVSALTGC 71
IAA2 WHEAT|P01083 ( 7) CNPATGYKVSALTGC 77
IAA5 WHEAT|P01084 ( 6) CYPGQAFQVPALPGC 57
IAA2 HORVU|P13691 ( 37) CDPFMGHKVSPLTRC 96
//
```

Matriz	Utilização	% identidade
PAM40	Alinhamentos curtos, elevada similaridade	70-90
PAM160	Detecção de membros de uma família	50-60
PAM250	Alinhamentos de sequências distantes	~20-30
BLOSUM90	Alinhamentos curtos, elevada similaridade	70-90
BLOSUM80	Detecção de membros de uma família	50-60
BLOSUM62	Eficaz na detecção de possíveis similaridades	30-40
BLOSUM30	Alinhamentos longos, sequências distantes	<30

Matrizes de probabilidades de transição para nucleótidos

Transições: **A ↔ G** , **C ↔ T**

Transversões: **A ↔ T** , **G ↔ C**

A ↔ C , **G ↔ T**

Frequências de mutação uniformes (**1 PAM**)

	A	T	G	C
A	0.99			
T	0.0033	0.99		
G	0.0033	0.0033	0.99	
C	0.0033	0.0033	0.0033	0.99

Transições mais frequentes (3x) que transversões

	A	T	G	C
A	0.99			
T	0.0020	0.99		
G	0.0060	0.0020	0.99	
C	0.0020	0.0060	0.0020	0.99

Matrizes de *score* para nucleótidos (log odds)

$$S_{ij} = \log_b (p_i M_{ij} / p_i p_j)$$

S_{ij} = “log odds” score

M_{ij} = *score* da matriz de transição

p_i , p_j = probabilidades de ocorrência dos nucleótidos

b = base do logaritmo (arbitrária)

Frequências de mutação uniformes (**1 PAM**)

	A	T	G	C
A	2			
T	-6	2		
G	-6	-6	2	
C	-6	-6	-6	2

$$S_{A,A} = \log_2(0.25 \cdot 0.99 / 0.25 \cdot 0.25) \approx 2$$

$$S_{T,A} = \log_2(0.25 \cdot 0.0033 / 0.25 \cdot 0.25) \approx -6$$

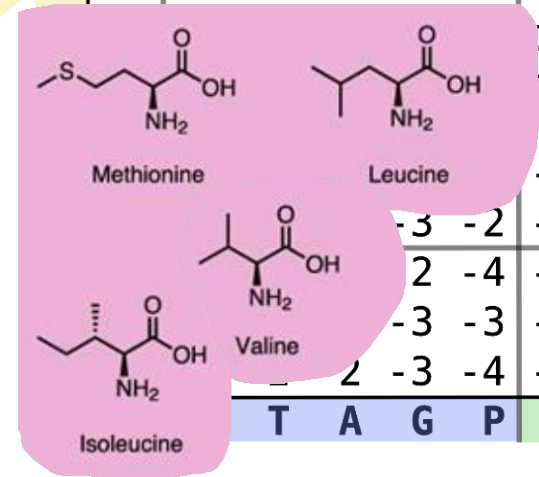
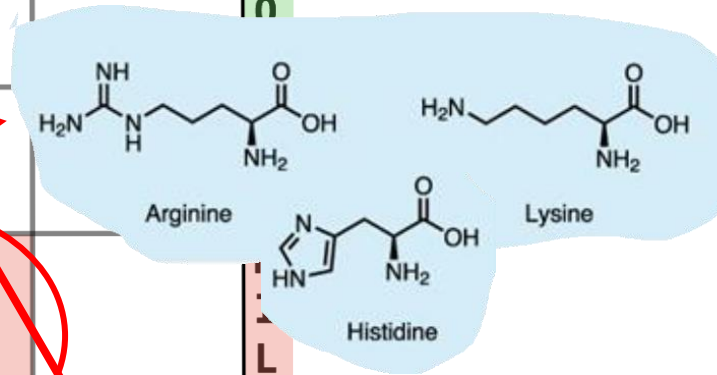
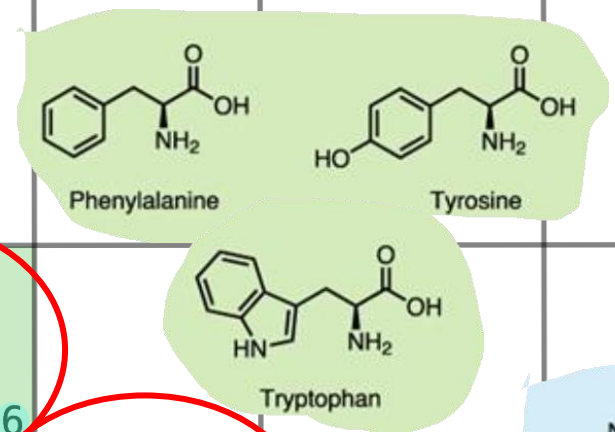
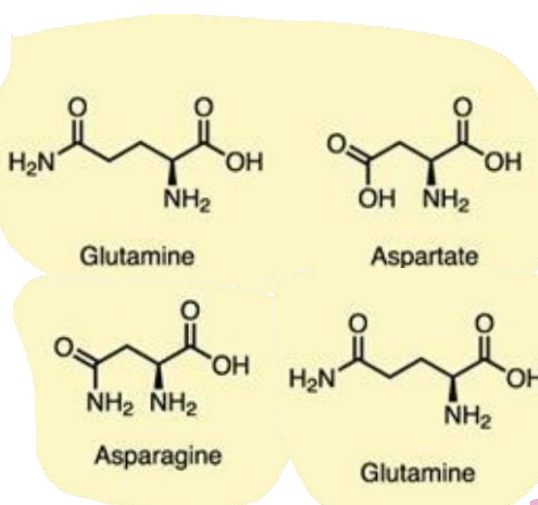
Transições mais frequentes (**3x**) que transversões

	A	T	G	C
A	2			
T	-5	2		
G	-7	-7	2	
C	-5	-7	-5	2

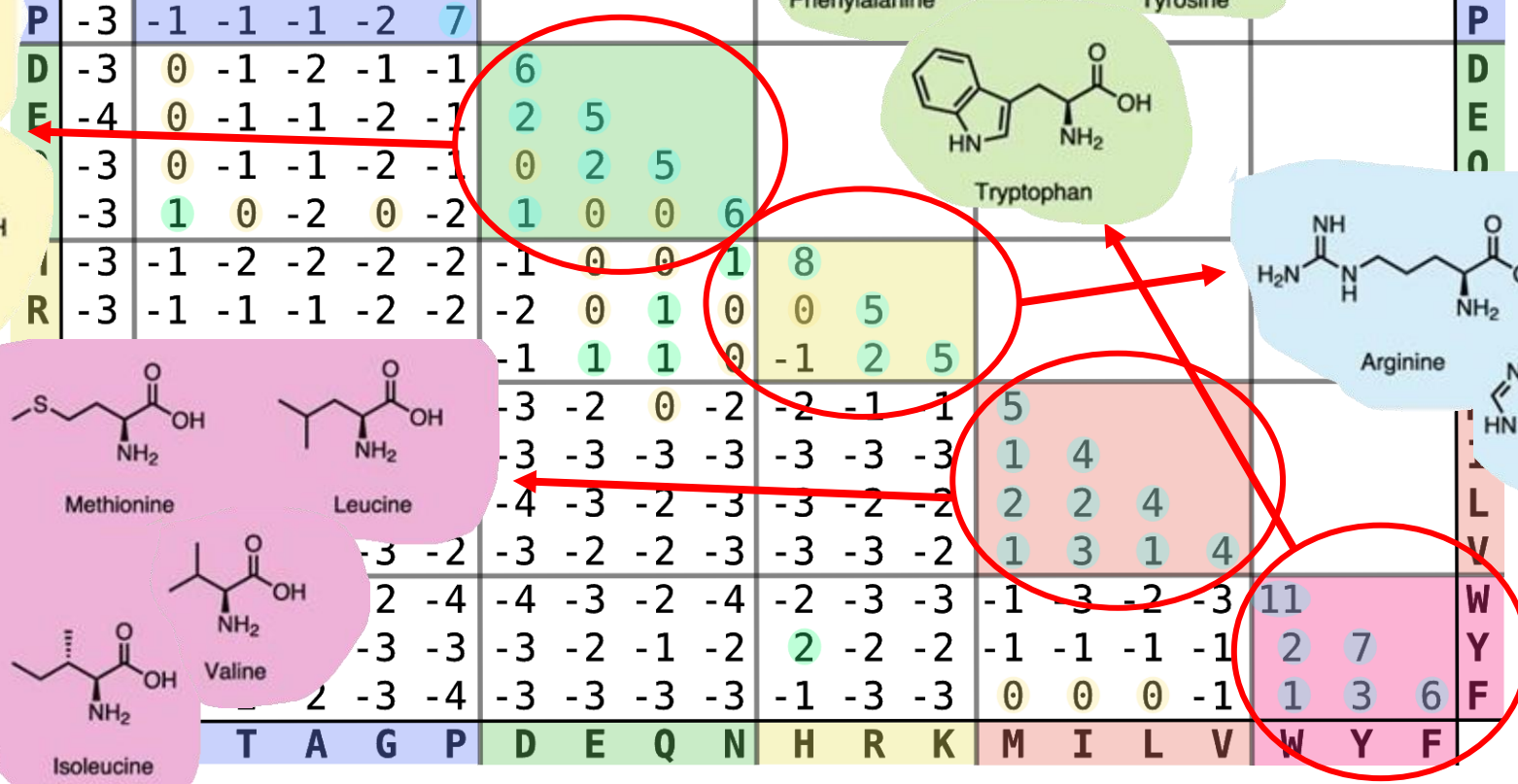
$$S_{T,A} = \log_2(0.25 \cdot 0.0020 / 0.25 \cdot 0.25) \approx -5$$

$$S_{G,A} = \log_2(0.25 \cdot 0.0060 / 0.25 \cdot 0.25) \approx -7$$

Interpretação física dos valores das matrizes de score



	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F
C	9																			
S	-1	4																		
T	-1	1	5																	
A	0	1	0	4																
G	-3	0	-2	0	6															
P	-3	-1	-1	-1	-2	7														
D	-3	0	-1	-2	-1	-1	6													
E	-4	0	-1	-1	-2	-1	2	5												
Q	-3	0	-1	-1	-2	-1	0	2	5											
N	-3	1	0	-2	0	-2	1	0	0	6										
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8									
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5								
K	-1	1	1	0	-1	2	5													
M	-3	-2	0	-2	-2	-1	1							5						
I	-3	-3	-3	-3	-3	-3	-3							1	4					
L	-4	-3	-2	-3	-3	-2	-2							2	2	4				
V	-4	-3	-2	-3	-2	-3	-3							1	3	1	4			
W	-2	-4	-4	-3	-2	-4	-2	-3	-3					-1	3	-2	-3	11		
Y	-3	-3	-3	-2	-1	-2	2	-2	-2					-1	-1	-1	-1	2	7	
F	-2	-3	-4	-3	-3	-3	-1	-3	-3					0	0	0	-1	1	3	6



Gap penalties

A inserção ou deleção de porções de uma sequência são eventos raros que conduzem a divergências de comprimento entre sequências homólogas.

Estas diferenças de comprimento implicam a necessidade inserir espaços (“gaps”) num alinhamento, mas esta inserção tem que ser pesadamente penalizada para estar de acordo com a raridade destes eventos.

Existem diferentes esquemas de penalização dos gaps (“gap penalties”), mas todos passam pela atribuição de um score negativo que está geralmente relacionado com o comprimento do gap.

Esquemas mais comuns:

- Constante: o tipo mais simples, consistem a atribuir uma penalização constante cada vez que é criado um gap num alinhamento
- Linear: a penalização é proporcional ao comprimento total dos gaps criados no alinhamento, não dependendo do seu número
- Afim (affine gap penalties): as penalizações possuem um termo constante para cada gap criado, e um termo proporcional ao comprimento do gap criado.

Penalidades afins (Affine gap penalties)

Uma representação mais realista do processo de evolução das proteínas deveria penalizar de modo diferente a *criação* e a *extensão* de um *gap*. Para entender este facto, devemos considerar que os alinhamentos entre sequências tem tendência a conter poucos gaps, mas quase sempre com vários resíduos de comprimento.

Se atribuirmos uma penalidade c para a criação de um gap e uma penalidade e para a sua extensão, temos:

$$gp = c + n \times e,$$

em que n é o comprimento do gap.

Não existe uma teoria rigorosa para a escolha de valores para estes parâmetros!

Valores usuais:

$$c = -10, e = -2 \text{ (FASTA)}$$

$$c = -5 \dots -10, e = -1, -2 \text{ (BLAST)}$$

$$c = -12, -2 \text{ (Smith-Waterman)}$$

O valor óptimo das *gap penalties* depende da matriz de score usada!

Alinhamentos sub-óptimos: o programa LALIGN

Comparison of:
 (A) f9-human.aa >F9 gi|119772|sp|P00740|FA9_HUMAN COAGULATION FA -
 (B) f12-hum.aa >F12 gi|119763|sp|P00748|FA12_HUMAN COAGULATION -
 using protein matrix

① 35.4% identity in 254 aa overlap; score: 358

```

F9      220      230      240      250      260      270
      QSFNDFTRVVGGEDAKPGQFPWQVVLNGKVDAPCGGSIVNEKWIIVTAAHCVE---TGVKI
F12     .....: : : : : : : : : : : : : : : : : : : : : : : : : : : : :
      KSLSSMTRVVGGGLVALRGAHPYIAALY-WGHSFCAGSLIAPCWVLTAAHCLQDRPAPEDL
      370      380      390      400      410      420

F9      280      290      300      310      320      330
      TVVAGEHNIEETEHETEQKRNVIRIIPHHNYNAAINKYNHDIALLELDEPL----VLNSY
F12     : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
      TVVLGQERRNHSCEPCQTLAVRSYRLHEAFSPV--SYQHDLALLRQEDADGSCALLSPY
      430      440      450      460      470      480

F9      340      350      360      370      380
      VTPICIAADKEYTNIPLKFGSGYVSGWGRVVFHKGRS-ALVLQYLRVPLVDRATCLRSTKF-
F12     : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
      VQPVCLPSGAARPSETTLCQ--VAGWGHQFEGAEYASFLQEAQVPFLSLERC SAPDVHG
      490      500      510      520      530

F9      390      400      410      420      430      440
      -TIYNNMFCAGPHEGGRDSCQDSSGGPHVTEVEGTS---FLTGIISWGEECAMKGYGIY
F12     . : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
      SSILPGMLCAGFLEGGTDACQDSSGGLVCEQAAERRLTLQGIISWGGCGDRNKPGVY
      540      550      560      570      580      590

F9      450
      TKVSRYVNWIKEKT
F12     : : : : : : : : : :
      TDVAYYLAWIREHT
      600      610
  
```

② 34.7% identity in 49 aa overlap; score: 120

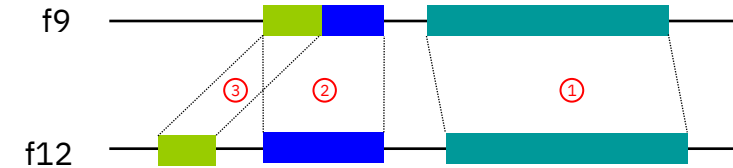
```

F9      100      110      120      130      140
      VDGQDCESNPCLNGGSKDDINSYECWCPFGFEGKNCELDVTCTNIKNGR
F12     .....: : : : : : : : : : : : : : : : : : : : : : : :
      LASQACRTNPCLHGGRCLVEGHRLCHCPVGYTGPPCDVDTKASCYDGR
      180      190      200      210      220
  
```

③ 33.3% identity in 36 aa overlap; score: 87

```

F9      100      110      120
      DQCESN-PCLNGGSKDDINSYECWCPFGFEGKNC
F12     : : : : : : : : : : : : : : : : : : : : : : : :
      DHSKHSKSPCQKGGTCVNMPSGPHCLCPQHLTGNHCC
      100      110      120      130
  
```



- Neste caso o alinhamento 1 é o alinhamento local óptimo, e os alinhamentos 2 e 3 são alinhamentos sub-óptimos identificados pelo programa LALIGN
- Muitas vezes a análise de alinhamentos sub-óptimos permite a identificação de regiões de similaridade entre duas sequências, não imediatamente reconhecíveis num alinhamento óptimo