Alinhamento de Sequências

SCC02713 - Introdução à Bioinformática

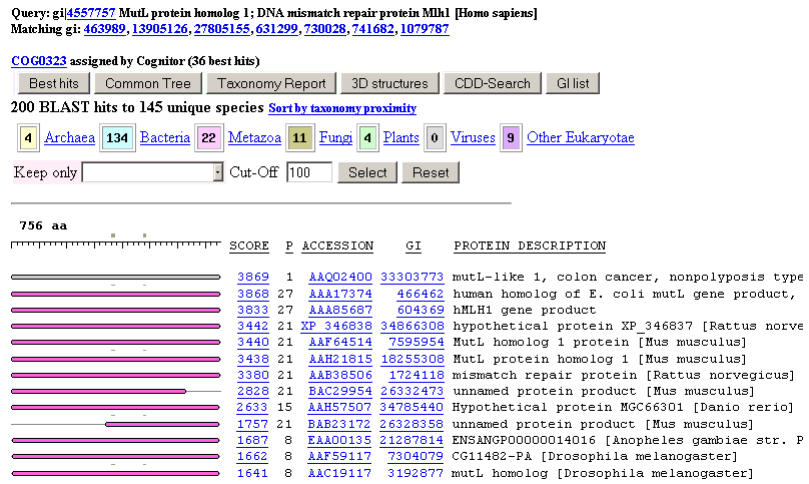
https://www.youtube.com/watch?v=8A-msg23u0w&list=PLpPXw4zFa0uLMHwSZ7DMeLGjIUgo1IBbn&index=30
https://www.youtube.com/watch?v=Ge00nXqIIow
https://www.youtube.com/watch?v=uGhZygAMQik

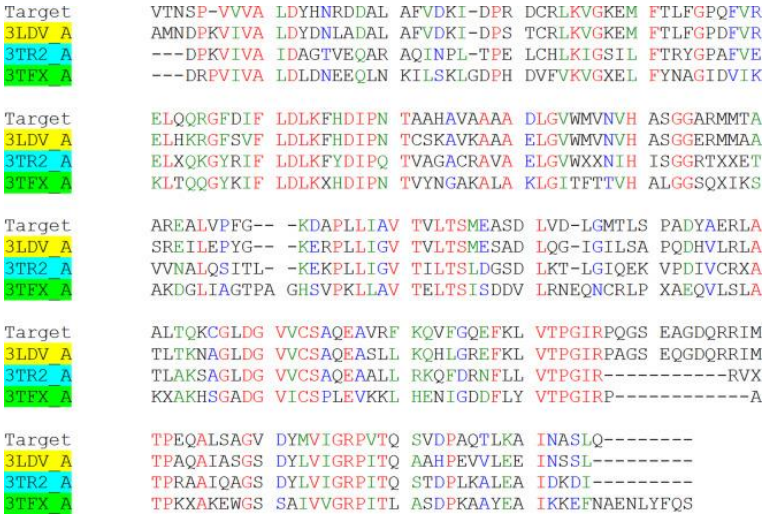
Alinhamento de Sequências

gaps and mismatches exist

Sequence alignment algorithms:

- E se quisermos alinhar várias sequências?
  - Uma contra muitas
  - Muitas contra muitas

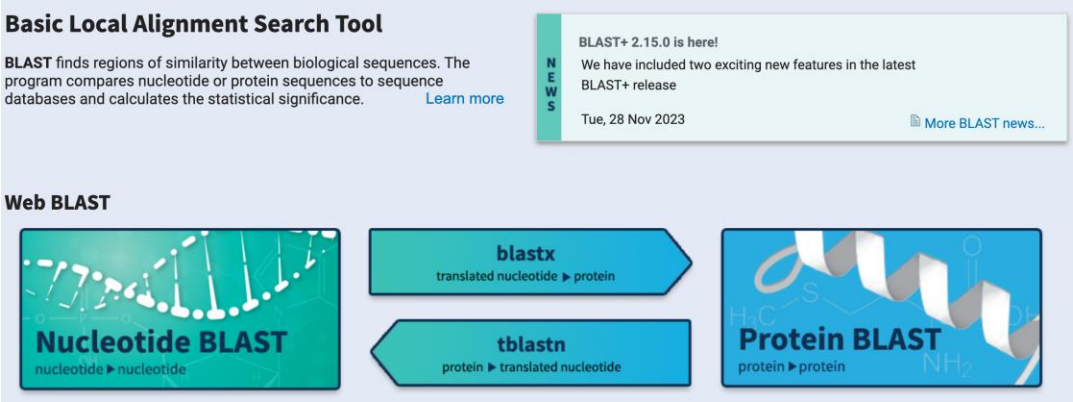




https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blink.html
https://www.sciencedirect.com/science/article/pii/B9780128096338204846

Alinhamento de Sequências

- BLAST - Basic Local Alignment Search Tool



jmb Journal of Molecular Biology
Volume 215, Issue 3, 5 October 1990, Pages 403-410

Basic local alignment search tool

Stephen F. Altschul [1], Warren Gish [1], Webb Miller [2], Eugene W. Myers [3], David J. Lipman [1]

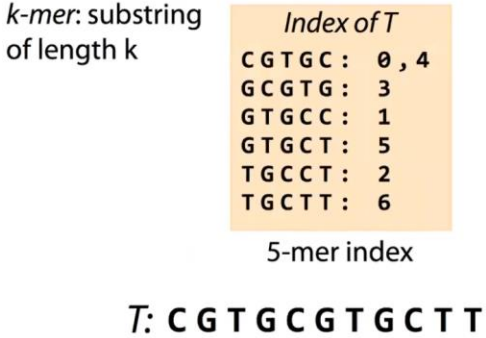
[1] National Center for Biotechnology Information National Library of Medicine, National Institutes of Health Bethesda, MD 20894, U.S.A.
[2] Department of Computer Science The Pennsylvania State University, University Park, PA 16802, U.S.A.
[3] Department of Computer Science University of Arizona, Tucson, AZ 85721, U.S.A.

Received 26 February 1990, Accepted 15 May 1990, Available online 6 February 2007.

https://blast.ncbi.nlm.nih.gov/Blast.cgi
https://www.sciencedirect.com/science/article/pii/S0022283605803602?via%3Dihub

Alinhamento de Sequências

- BLAST - Heurística
  - Quebra as sequências em palavras (k-mers) e faz hash de suas localizações para acelerar pesquisas posteriores

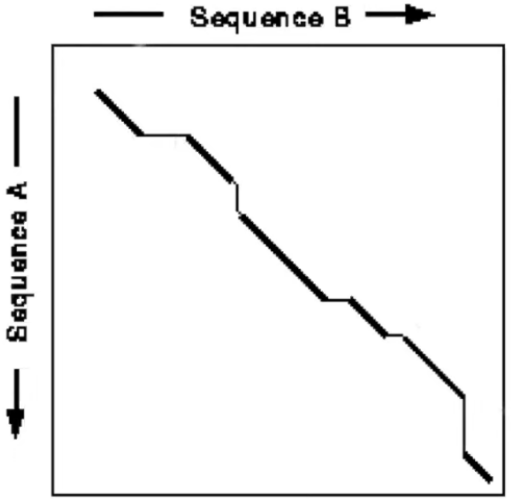


  - Para cada k-mer na query, encontre os k-mers no banco de dados que têm bom match
  - Somente k-mers com uma pontuação de alinhamento maior que um limiar são mantidas

https://www.youtube.com/watch?v=Ge00nXqIlow

Alinhamento de Sequências

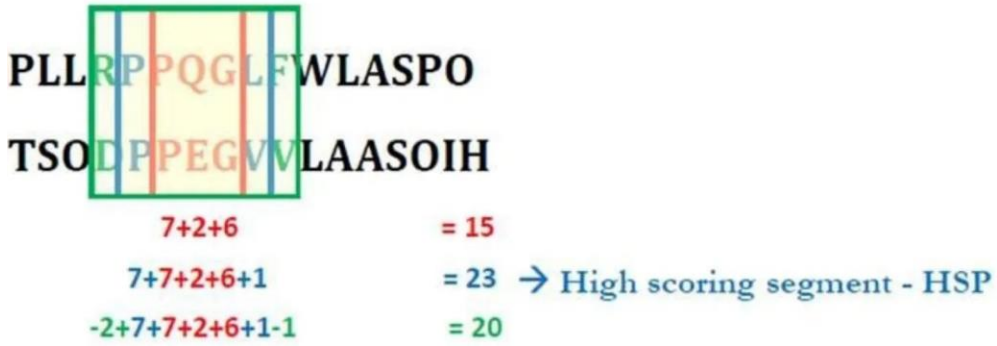
- BLAST - Heurística
  - Para cada trecho de alta pontuação no banco de dados, tenta estender o alinhamento para ambos os lados
  - Forma HSP (High-scoring Segment Pairs)
  - Mantém apenas os HPS estatisticamente significantes
    - Baseado no score de alinhar duas sequências aleatórias
  - Usa o algoritmo Smith-Waterman para unir os HSP e obter o alinhamento final ótimo



https://www.youtube.com/watch?v=Ge00nXqIlow

Alinhamento de Sequências

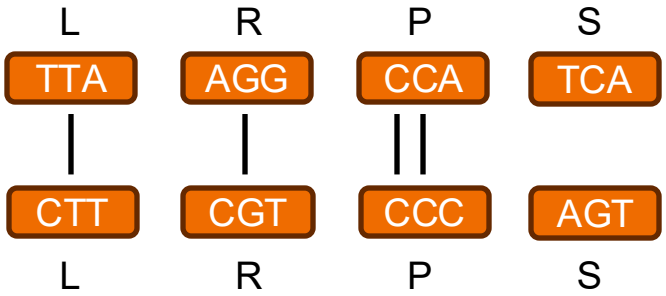
- BLAST – Seed and Extend
  - BLAST 1: Estende os k-mers para a esquerda e para a direita usando alinhamentos sem gaps
  - A extensão continua enquanto a pontuação não cair abaixo de um limiar determinado
  - BLAST 2: Estende os HSPs usando alinhamento com gaps



https://www.youtube.com/watch?v=Ge00nXqIlow
https://medium.com/@gokcenazakyol/biological-sequence-analysis-ii-multiple-sequence-alignment-alignment-algorithms-tools-6dac6d87119a

Alinhamento de Sequências

- DNA x Proteínas
  - Diferentes códons codificam o mesmo aminoácido



Standard DNA codon table [edit]

Amino-acid biochemical properties	Nonpolar (np)	Polar (p)	Basic (b)	Acidic (a)

Termination: stop codon *

Initiation: possible start codon ⇒

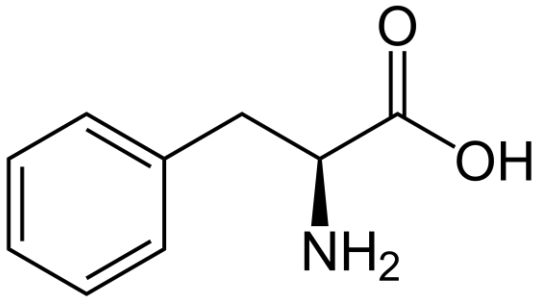
Standard genetic code[17][note 3]

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine (np)	TCT	(Ser/S) Serine (p)	TAT	(Tyr/Y) Tyrosine (p)	TGT	(Cys/C) Cysteine (p)	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine (np)	TCA		TAA	Stop (*Ochre*) *[note 2]	TGA	Stop (*Opal*) *[note 2]	A
	TTG ⇒		TCG		TAG	Stop (*Amber*) *[note 2]	TGG	(Trp/W) Tryptophan (np)	G
C	CTT	(Leu/L) Leucine (np)	CCT	(Pro/P) Proline (np)	CAT	(His/H) Histidine (b)	CGT	(Arg/R) Arginine (b)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine (p)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine (np)	ACT	(Thr/T) Threonine (p)	AAT	(Asn/N) Asparagine (p)	AGT	(Ser/S) Serine (p)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine (b)	AGA	(Arg/R) Arginine (b)	A
	ATG ⇒	(Met/M) Methionine (np)	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine (np)	GCT	(Ala/A) Alanine (np)	GAT	(Asp/D) Aspartic acid (a)	GGT	(Gly/G) Glycine (np)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid (a)	GGA		A
	GTG ⇒		GCG		GAG		GGG		G

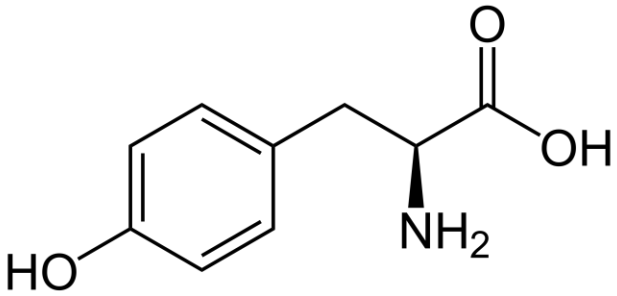
https://en.wikipedia.org/wiki/DNA_and_RNA_codon_tables

Alinhamento de Sequências

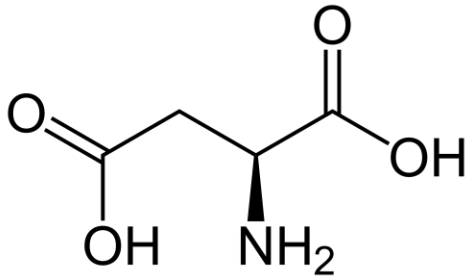
- DNA x Proteínas
  - Diferenças nas estruturas dos aminoácidos têm influência no alinhamento



Fenilalanina (TTT ou TTC)



Tirosina (TAT ou TAC)



Ácido aspártico (GAT ou GAC)

https://en.wikipedia.org/wiki/Tyrosine
https://en.wikipedia.org/wiki/Aspartic_acid
https://en.wikipedia.org/wiki/Phenylalanine
https://www.youtube.com/watch?v=K7i2XbFZv6Y&list=PLpPXw4zFa0uLMHwSZ7DMeLGjIUgo1IBbn&index=32

Alinhamento de Sequências

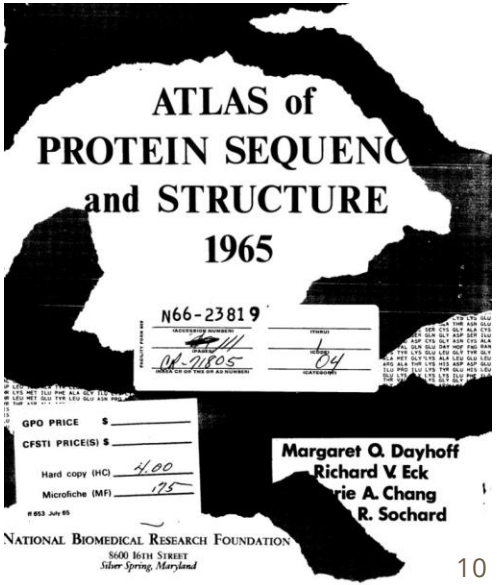
- Matrizes de Pontuação
  - Baseadas em taxas de substituição observadas, derivadas de frequências de substituição observadas em alinhamentos múltiplos de sequências
  - Mais usadas:
    - PAM – Point Accepted Mutation: baseado em alinhamento global e considera processo evolutivo
    - BLOSUM – Block Substitution: baseado em alinhamento local e considera similaridades entre domínios conservados

https://www.youtube.com/watch?v=uGhZygAMQik

Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Compilação de alinhamentos de sequências de 71 famílias de proteínas relacionadas (>= 85% de similaridade)
  - Dayhoff verificou 1572 mudanças em aminoácidos nessas sequências e calculou suas frequências
  - Isso definiu então a frequência com que uma mudança é aceita (Point Accepted Mutation - PAM)
  - Portanto: PAM é baseada na divergência evolutiva entre sequências de proteínas
    - PAM100 – 100 mutações a cada 100 aminoácidos
    - PAM120 – 120 mutações a cada 100 aminoácidos
    - PAM250 – 250 mutações a cada 100 aminoácidos (2,5 mutações por aminoácido)

Margaret Dayhoff

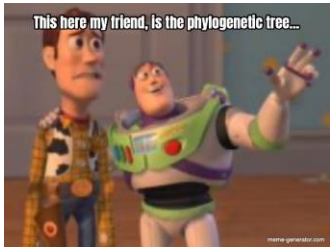


https://ntrs.nasa.gov/api/citations/19660014530/downloads/19660014530.pdf

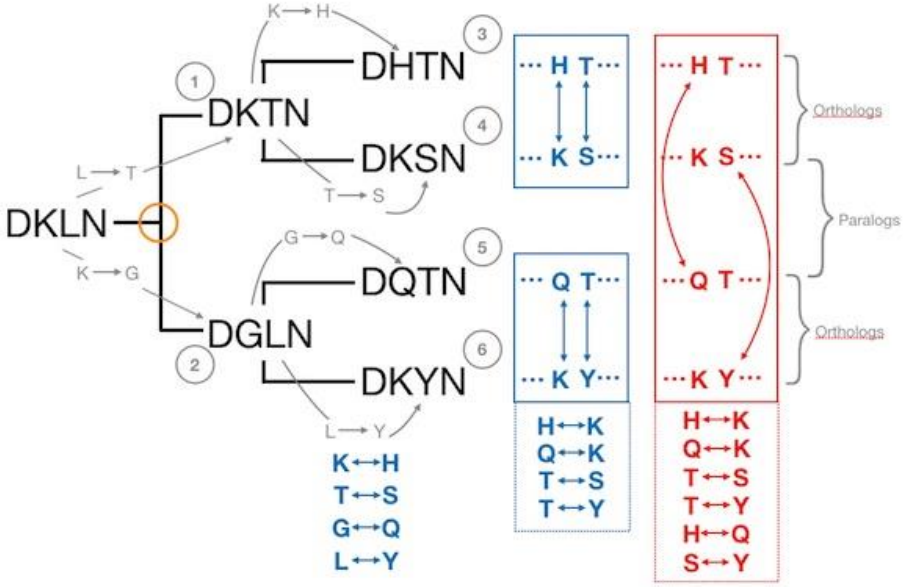
Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Cada linha e coluna representa um dos 20 aminoácidos padrão. Assim, as matrizes são usadas como matrizes de substituição para pontuar alinhamentos de sequências de proteínas
  - Dessa forma, a PAM é uma mudança de um aminoácido para outro aceita pela evolução (seleção natural)
  - Para ser aceita: o novo aminoácido deve funcionar de maneira similar ao antigo
  - Os dados de mutação foram obtidos de pares de sequências relacionadas e árvores filogenéticas

Alinhamento de Sequências



- Matrizes de Pontuação – PAM
  - Árvore Filogenética



Accepted Point Mutations

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

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html

Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Contagem total de 1.572 mutações pontuais aceitas (PAMs) de 71 famílias de proteínas
  - Das 1572 trocas, o maior número, 83, foi observado entre Asp e Glu, dois aminoácidos quimicamente muito semelhantes com códons diferindo por um nucleotídeo
  - Dayhoff e colaboradores multiplicaram os números por 10 para simplificação de cálculos
  - Trocas fracionárias ocorrem quando sequências ancestrais são ambíguas

Número de mutações pontuais aceitas (X10) acumuladas a partir de sequências intimamente relacionadas

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	50	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Dayhoff e colegas também calcularam a mutabilidade relativa dos aminoácidos
  - Contagem do número de vezes que cada aminoácido é substituído dividido pelo número de vezes que ocorre em um intervalo observado
  - Ex: para o alinhamento entre AGLL e AGAV o cálculo da mutabilidade relativa é o seguinte:

```
                                A G L L
                                A G A V

Amino acids:                    A       G       L       V
Changes:                        1       0       2       1
Frequency of occurrence:        3       2       2       1
Relative mutability:            0.33    0       1       1
```

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html

Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Dayhoff e colegas também calcularam a mutabilidade relativa dos aminoácidos
  - Contagem do número de vezes que cada aminoácido é substituído dividido pelo número de vezes que ocorre em um intervalo observado

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

- Matrizes de Pontuação – PAM
  - Podemos combinar informações sobre os tipos individuais de mutações e sobre a mutabilidade relativa dos aminoácidos em uma “matriz de probabilidade de mutação” dependente da distância
  - Um elemento desta matriz, M_{ij}, fornece a probabilidade de que o aminoácido na coluna *j* seja substituído pelo aminoácido na linha *i* após um determinado intervalo evolutivo, neste caso 1 PAM

Valores multiplicados por 10000

Original amino acid

Replacement amino acid	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	Sum
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18	1.007
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1	0.998
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1	0.996
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1	0.999
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2	0.999
Q	3	9	4	5	0	9876	27	1	23	1	2	6	4	0	6	2	2	0	0	1	0.997
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2	1.000
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5	1.003
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1	0.998
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33	0.996
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15	1.008
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1	1.007
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4	0.990
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0	1.000
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2	0.999
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2	1.004
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9	1.000
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0	0.998
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1	0.999
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901	1.004
Sum	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

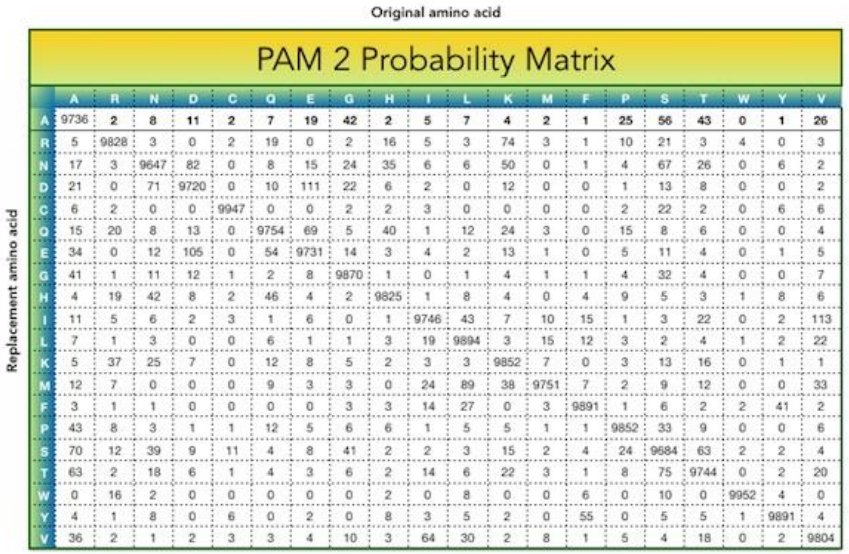
- Matrizes de Pontuação – PAM – Extrapolando a PAM1
  - Matriz PAM1: baseada em sequências 99% idênticas. Distância evolutiva onde 1% dos aminoácidos foram alterados. Portanto ótimas para comparar sequências 99% idênticas
  - Existe uma matriz de probabilidade de mutação diferente para cada intervalo evolutivo. Podemos usar multiplicação de matrizes
  - Multiplicamos PAM1 por si mesma muitas vezes para chegar a alguma medida de distância evolutiva específica em PAMs

A	B	C		J	K	L		[AJ + BM + CP]	[AK + BN + CQ]	[AL + BO + CR]
D	E	F	X	M	N	O	=	[DJ + EM + FP]	[DK + EN + FQ]	[DL + EO + FR]
G	H	I		P	Q	R		[GJ + HM + IP]	[GK + HN + IQ]	[GL + HO + IR]

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

- Matrizes de Pontuação – PAM2
  - Em que ordem devemos multiplicar PAM1 por si mesma para obter, por exemplo, PAM 3?
  - PAM2 x PAM5, por exemplo, não é o mesmo que PAM5 x PAM2
  - A resposta é que sempre mantemos PAM1 no lado esquerdo, portanto PAM1 x PAM249 = PAM250



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9736	2	8	11	2	7	19	42	2	5	7	4	2	1	25	56	43	0	1	26
R	5	9828	3	0	2	19	0	2	16	5	3	74	3	1	10	21	3	4	0	3
N	17	3	9647	82	0	8	15	24	35	6	6	50	0	1	4	67	26	0	6	2
D	21	0	71	9720	0	10	111	22	6	2	0	12	0	0	1	13	8	0	0	2
C	6	2	0	0	9947	0	0	2	2	3	0	0	0	0	2	22	2	0	6	6
Q	15	20	8	13	0	9754	69	5	40	1	12	24	3	0	15	8	6	0	0	4
E	34	0	12	105	0	54	9731	14	3	4	2	13	1	0	5	11	4	0	1	5
G	41	1	11	12	1	2	8	9870	1	0	1	4	1	1	4	32	4	0	0	7
H	4	19	42	8	2	46	4	2	9825	1	8	4	0	4	9	5	3	1	8	6
I	11	5	6	2	3	1	6	0	1	9746	43	7	10	15	1	3	22	0	2	113
L	7	1	3	0	0	6	1	1	3	19	9894	3	15	12	3	2	4	1	2	22
K	5	37	25	7	0	12	8	5	2	3	3	9852	7	0	3	13	16	0	1	1
M	12	7	0	0	0	9	3	3	0	24	89	38	9751	7	2	9	12	0	0	33
F	3	1	1	0	0	0	0	3	3	14	27	0	3	9891	1	6	2	2	41	2
P	43	8	3	1	1	12	5	6	6	1	5	5	1	1	9852	33	9	0	0	6
S	70	12	39	9	11	4	8	41	2	2	3	15	2	4	24	9684	63	2	2	4
T	63	2	18	6	1	4	3	6	2	14	6	22	3	1	8	75	9744	0	2	20
W	0	16	2	0	0	0	0	0	2	0	8	0	0	6	0	10	0	9952	4	0
Y	4	1	8	0	6	0	2	0	8	3	5	2	0	55	0	5	5	1	9891	4
V	36	2	1	2	3	3	4	10	3	64	30	2	8	1	5	4	18	0	2	9804

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

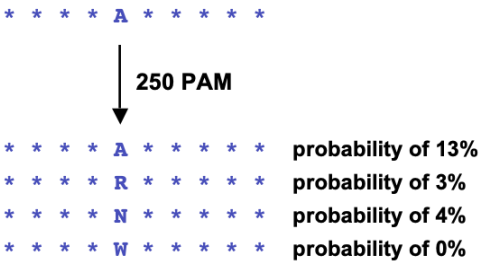
- Matrizes de Pontuação – PAM2
  - A célula A-A na matriz PAM2 é a probabilidade de A não sofrer mutação ou a probabilidade de observarmos A-A
  - Assim, para obter o valor de A-A após PAM1 x PAM1 ou um período evolutivo PAM1 após PAM1, precisamos primeiro calcular o seguinte:
    - 1. A probabilidade de A permanecer A, que é (A-A) x (A-A)
    - 2. A probabilidade de qualquer um dos outros aminoácidos sofrer mutação para A: soma de todas as probabilidades na primeira linha (R-A, N-A, D-A,..., V-A), exceto A-A
    - 3. A probabilidade de A sofrer mutação para qualquer um dos outros aminoácidos: todas as probabilidades na primeira coluna (A-R, A-N, A-D,..., A-V), exceto A-A
  - Portanto, o valor de A-A na matriz PAM2 é a soma das probabilidades 1+2+3 = 9735.802
    - O mesmo valor de A-A na matriz PAM2 depois de multiplicar PAM1 x PAM1

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html
https://courses.cs.duke.edu/spring24/compsci260/resources/AlignmentPapers/1978.dayhoff.schwartz.orcutt.pdf

Alinhamento de Sequências

- Matrizes de Pontuação – PAM250
  - Valores multiplicados por 100

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17



https://rob-p.github.io/CSE549F17/lectures/Lec10.pdf

Alinhamento de Sequências

- Matrizes de Pontuação – PAM250
  - Para o alinhamento vamos da matriz de probabilidade para a matriz de substituição

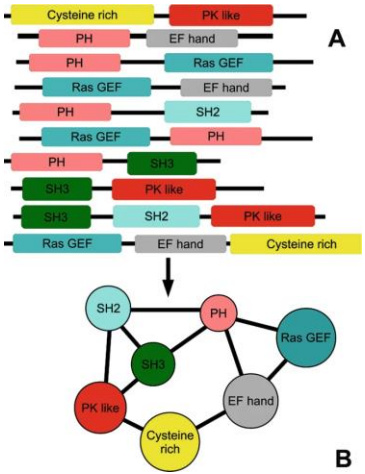
$$s_n(i,j) = \log \frac{M^n_{ji}}{f_j} = \log \frac{P_{ji,n}}{f_i f_j}$$

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

https://www.thegpm.org/BIOML/aainfo/pam250.htm

Alinhamento de Sequências

- Matrizes de Pontuação – BLOSUM
  - Utiliza domínios conservados



Jorja Henikoff

Steven Henikoff

Proc Natl Acad Sci U S A. 1992 Nov 15; 89(22): 10915–10919.
doi: 10.1073/pnas.89.22.10915

PMCID: PMC50453
PMID: 1438297

Amino acid substitution matrices from protein blocks.

S Henikoff and J G Henikoff

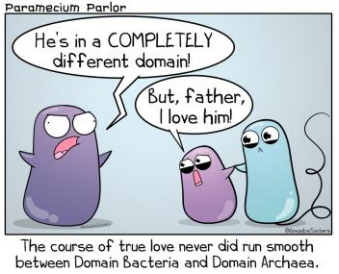
Author information · Copyright and License information · PMC Disclaimer

Abstract

Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50453/
https://research.fredhutch.org/henikoff/en/members.html
https://fiuglaser.fiu.edu/?p=290
https://link.springer.com/protocol/10.1007/978-1-4939-9074-0_15

Alinhamento de Sequências



- Matrizes de Pontuação – BLOSUM
  - Henikoff e Henikoff analisaram regiões conservadas de sequências de proteínas relacionadas que obtiveram do banco de dados BLOCKS
    - Protein Domains / Blocks
  - Examinaram 2000 blocos sem gaps e 500 grupos de proteínas relacionadas contando o número de matchs e mismatches de cada tipo dos 20 aminoácidos diferentes
  - Henikoff e Henikoff criaram uma tabela de frequência e, usando essas frequências, calcularam a probabilidade de cada tipo de match e mismatch
  - Converteram as probabilidades em logaritmo de razões de probabilidade (log odds ratios)

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices.html

Alinhamento de Sequências

- Matrizes de Pontuação – BLOSUM
  - Henikoff e Henikoff dividiram os grupos em subgrupos por sua porcentagem de similaridade
  - Essa divisão resultou em uma família de matrizes BLOSUM
  - BLOSUM65 significa que as pontuações são de um conjunto de sequências onde as sequências são pelo menos 65% semelhantes
  - BLOSUM80: pontuações são de conjuntos com pelo menos 80% de similaridade, e assim por diante

Alinhamento de Sequências

- Matrizes de Pontuação – BLOSUM62
  - Matriz padrão do BLAST
  - Por isso é a mais usada

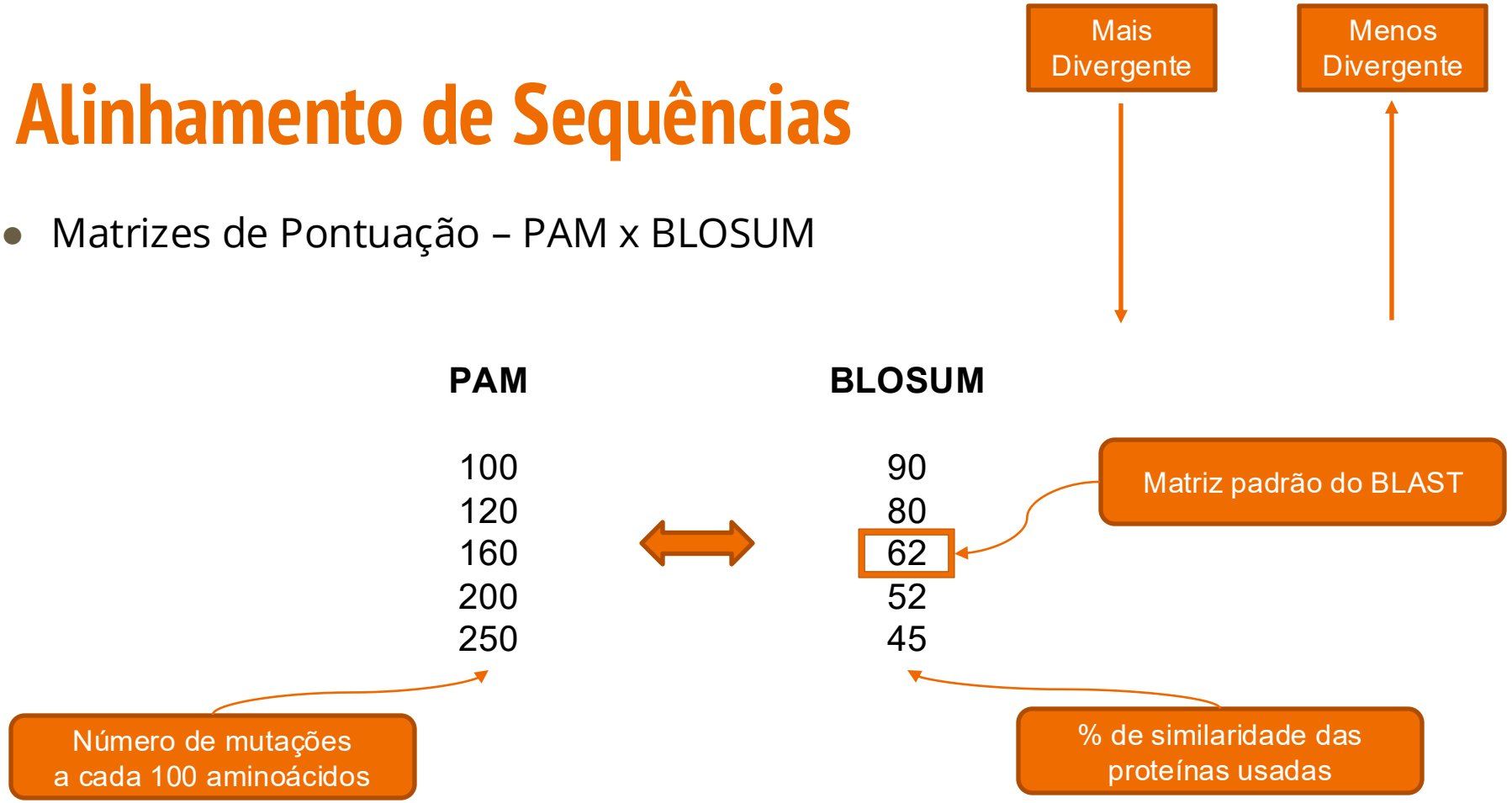
$$S_{ij} = \left(\frac{1}{\lambda}\right) log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

https://pt.wikipedia.org/wiki/BLOSUM

Alinhamento de Sequências

- Matrizes de Pontuação – PAM x BLOSUM



É isso aí!