



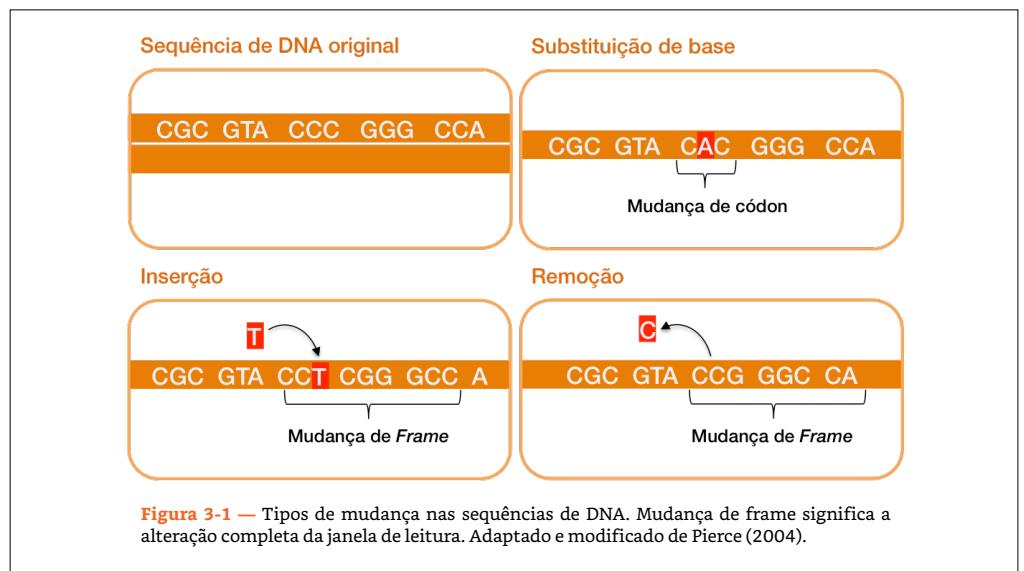
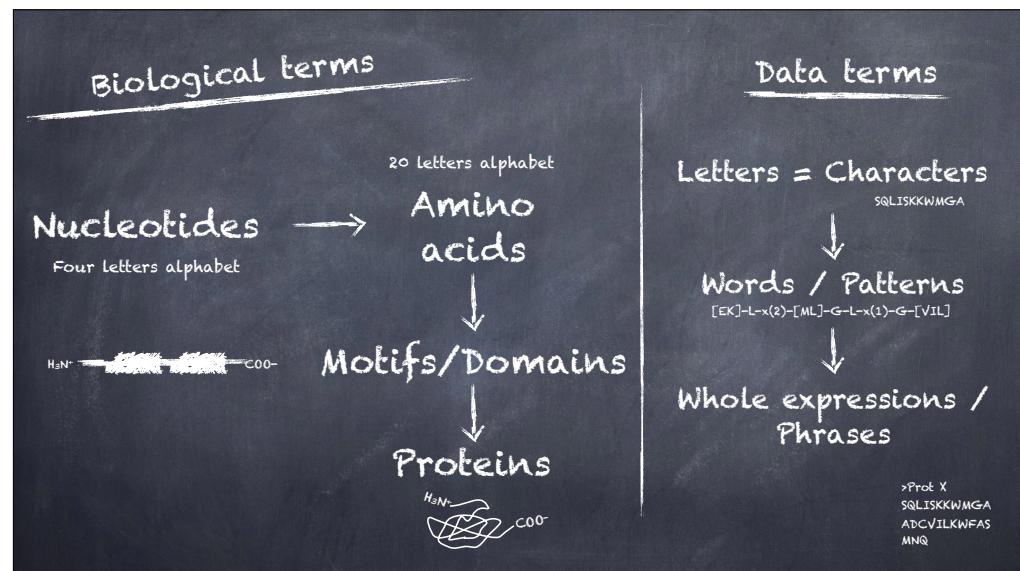
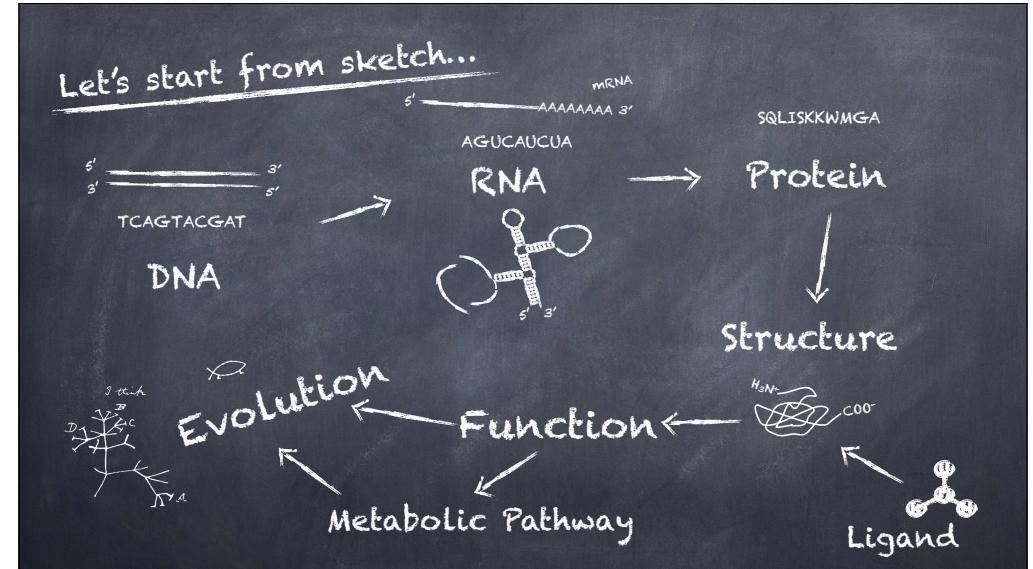
Bioinformatics
Multidisciplinary
Environment

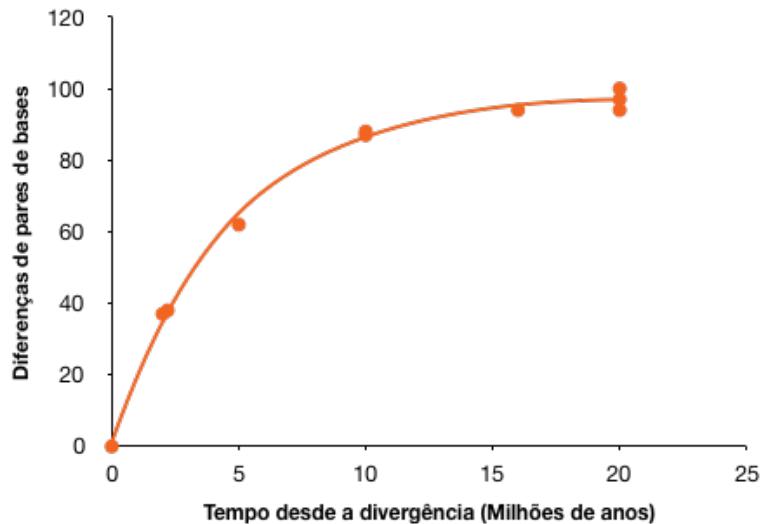
Centro
Multiusuário
de Bioinformática



Alinhamento de Sequências Proteicas

Prof. João Paulo Matos
Dept. de Bioquímica - Centro de Biociências
Bioinformatics Multidisciplinary Environment - BioME
Universidade Federal do Rio Grande do Norte





Buscas de similaridades

- A comparação de sequências é um método poderoso e confiável para responder questões biológicas e entender relações evolucionárias entre genes e proteínas.
- Depende de:
 - Busca por homologia.
 - Do alinhamento de sequências
- Onde conseguir sequências?
 - Em experimentos de laboratório.
 - Em Bancos de Dados Biológicos:
 - Armazenam e gerenciam a informação biológica, de uma forma em que ela pode ser facilmente preservada e acessada, permitindo a inclusão de novos dados.

Homologia

- Similaridade que é um resultado de uma herança a partir de um ancestral comum;
- Duas seqüências similares são homólogas se elas descenderam de um único ancestral.

Alinhamento de Seqüências

- O alinhamento é uma hipótese de homologia posicional entre os nucleotídeos (DNA) e entre os aminoácidos (Proteínas).

Objetivos do Alinhamento

- Gerar um sumário dos dados de seqüências conciso, rico em informações;
- Algumas vezes usados para ilustrar a dissimilaridade entre um grupo de seqüências;
- Alinhamentos podem ser tratados como modelos que podem ser usados para testar hipóteses;
- Se este modelo reflete com precisão a evidências biológicas conhecidas.

Alinhamento de Seqüências

GCGGGCCA	TCAGGTAGTT	GGTGG
GCGGGCCA	TCAGGTAGTT	GGTGG
GCCTTCCA	TCAGCTGGTT	GGTGG
GCGTCCA	TCAGCTAGTT	GGTGG
GCAGCGCA	TTAGCTAGTT	GGTGA
*****	*****	*****

Fácil

TTGACATG	CCGGGG---A	AACCG
TTGACATG	CCGGTG---GT	AAGCC
TTGACATG	-CTAGG---A	ACGCG
TTGACATG	-CTAGGGAAC	ACGCG
TTGACATC	-CTCTG---A	ACGCG
*****	???????????	*****

Difícil
(indels)

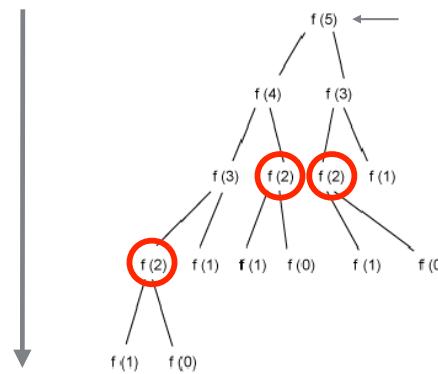
Métodos

- Programação dinâmica:
 - Considerar 2 sequências proteicas de 100 aminoácidos:
 - Se para alinhar 2 sequências se leva 100^2 segundos, 3 sequências serão 100^3 s, 4 sequências 100^4 s, etc.
 - Levaria mais tempo do que a própria vida para alinhar 20 sequências.
- Exemplo:
 - Sequência Fibonacci (A000045):
 - 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, etc.

Algoritmo Fibonacci 1: Recursivo

```
public int fibonacci1(int n) {  
    if (n == 0) {  
        return 0;  
    } else if (n == 1) {  
        return 1;  
    } else {  
        return fibonacci1(n - 1) + fibonacci1(n - 2);  
    }  
}
```

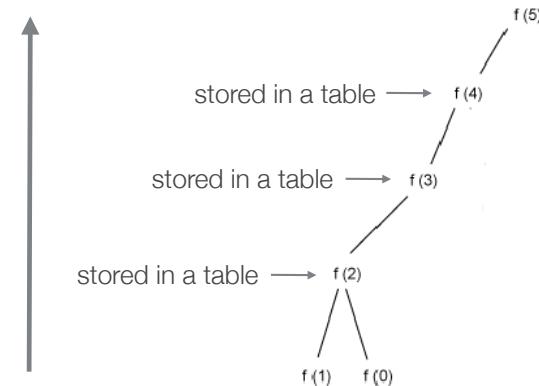
Algoritmo Fibonacci 1: Recursivo



Algoritmo Fibonacci2: Iterativo

```
public int fibonacci2(int n) {  
    int[] table = new int[n + 1];  
    for (int i = 0; i < table.length; i++) {  
        if (i == 0) {  
            table[i] = 0;  
        } else if (i == 1) {  
            table[i] = 1;  
        } else {  
            table[i] = table[i - 2] + table[i - 1];  
        }  
    }  
  
    return table[n];  
}
```

Algoritmo Fibonacci2: Iterativo



Programação Dinâmica

- Tempo Fibonacci 1: Exponencial a n ;
- Tempo Fibonacci 2: $0 + (n)$ tempo.
 - Desvantagem?
- Algoritmos conhecidos:
 - Smith-Waterman (swat);
 - Needleman-Wunsch.

Alinhamento de Seqüências

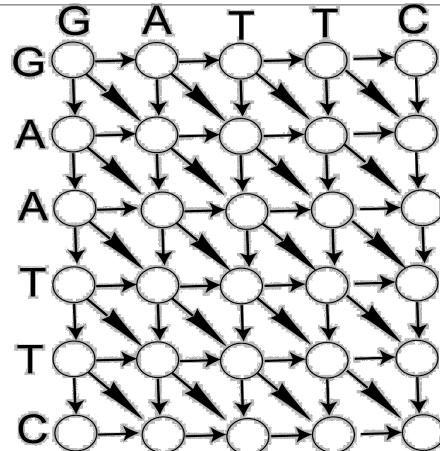
GC GG CCCA TC AGGTAGTT GGTGG
GC GG CCCA TC AGGTAGTT GGTGG
GC GTTCCA TC AGCTGGTT GGTGG
GC GTCCC A TC AGCTAGTT GGTGG
GC GG CGCA TT AGCTAGTT GGTGA
***** * ***** * *****

Fácil

TT GACATG CCGGGG---A AACCG
TT GACATG CCGGTG---GT AAGCC
TT GACATG -CTAGG---A ACGCG
TT GACATG -CTAGGGAAC ACGCG
TT GACATC -CTCTG---A ACGCG
***** ?????????? *

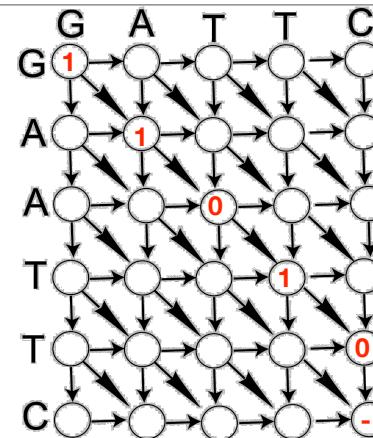
Difícil
(indels)

Alinhamento entre 2 Seqüências



Seq. A → G ATTC
Seq. B → GA ATT C

Alinhamento Possível:

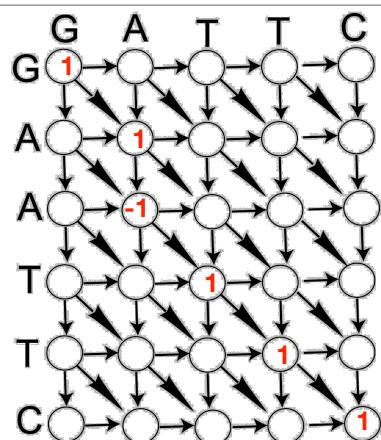


Esquema de Pontuação:
Pareamento: +1
Despareamento: 0
Gap: -1

Alinhamento Obtido:
G A T T C –
G A A T T C

Escore = 2

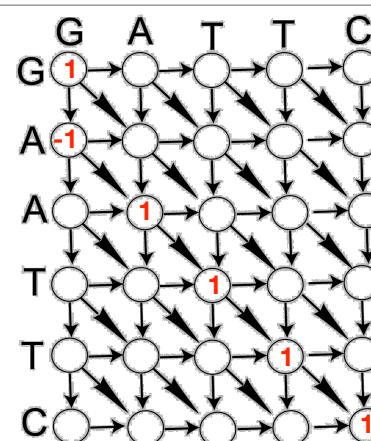
Alinhamento Ótimo 1:



Esquema de Pontuação:
Pareamento: +1
Despareamento: 0
Gap: -1
Alinhamento Obtido:
G A - T T C
G A A T T C

Escore = 4

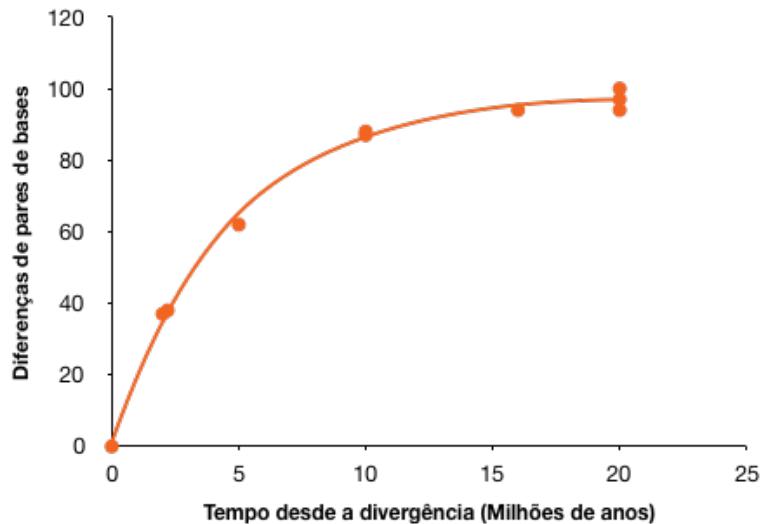
Alinhamento Ótimo 2:



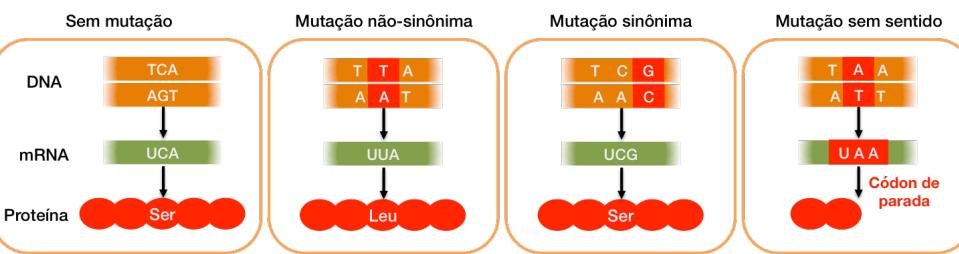
Esquema de Pontuação:
Pareamento: +1
Despareamento: 0
Gap: -1

Alinhamento Obtido:
G - A T T C
G A A T T C

Escore = 4



Em sequências codificantes



First letter of codon (5' end)

Second letter of codon

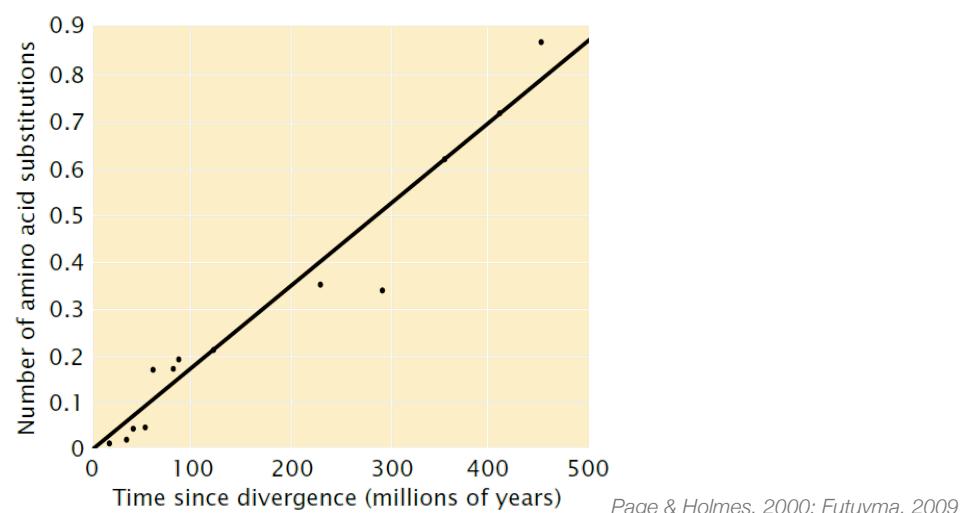
↓

	U	C	A	G
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GAA Glu GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGG Gly GGG Gly

Legend: Red boxes indicate mutations. The first two rows show non-synonymous mutations, while the last two rows show synonymous mutations.

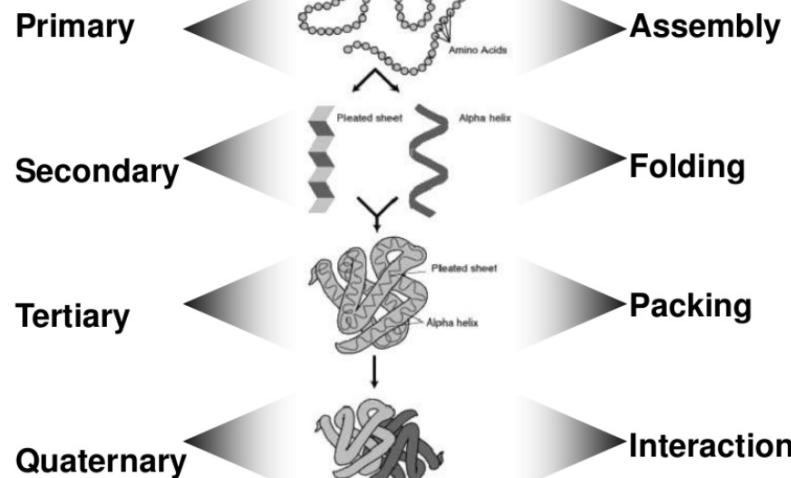
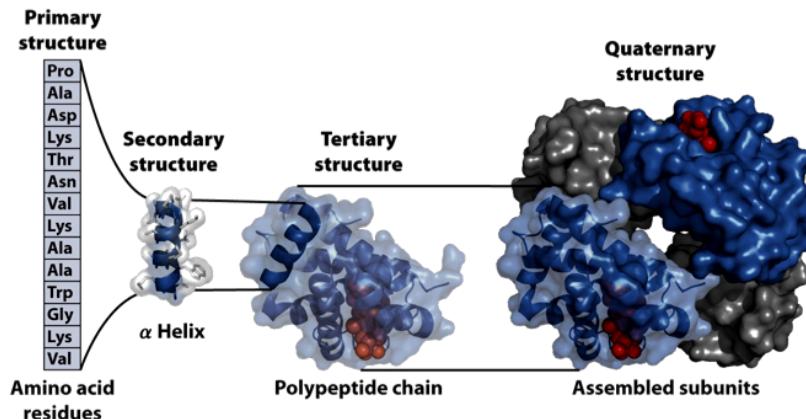
Chemical structures of amino acid side chains:

- Nonpolar, aliphatic R groups:** Glycine (H₃C-H), Alanine (CH₃-H), Valine (CH₃-CH₂-H).
- Aromatic R groups:** Phenylalanine (CH₃-C₆H₄-H), Tyrosine (CH₃-C₆H₄-OH), Tryptophan (CH₃-C₆H₄-NH).
- Polar, uncharged R groups:** Serine (H₃N-C₂H₅-OH), Threonine (H₃N-C₃H₇-OH), Cysteine (H₃N-C₂H₅-SH).
- Positively charged R groups:** Lysine (H₃N-C₂H₅-NH₂), Arginine (H₃N-C₃H₇-NH₂), Histidine (H₃N-C₂H₅-NH-C(=O)-CH₃).
- Negatively charged R groups:** Aspartate (H₃N-C₂H₅-COO⁻), Glutamate (H₃N-C₃H₇-COO⁻).



Page & Holmes, 2000; Futuyma, 2009

Níveis de Arquitetura das Proteínas



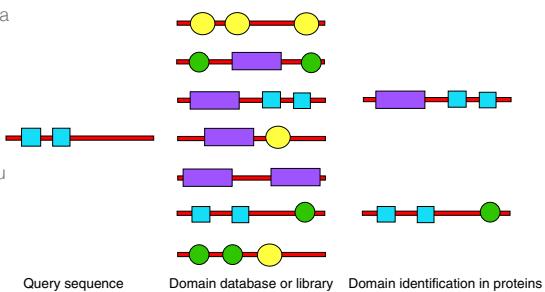
© Lippincott, Raven and company press, 5th edition, pp.

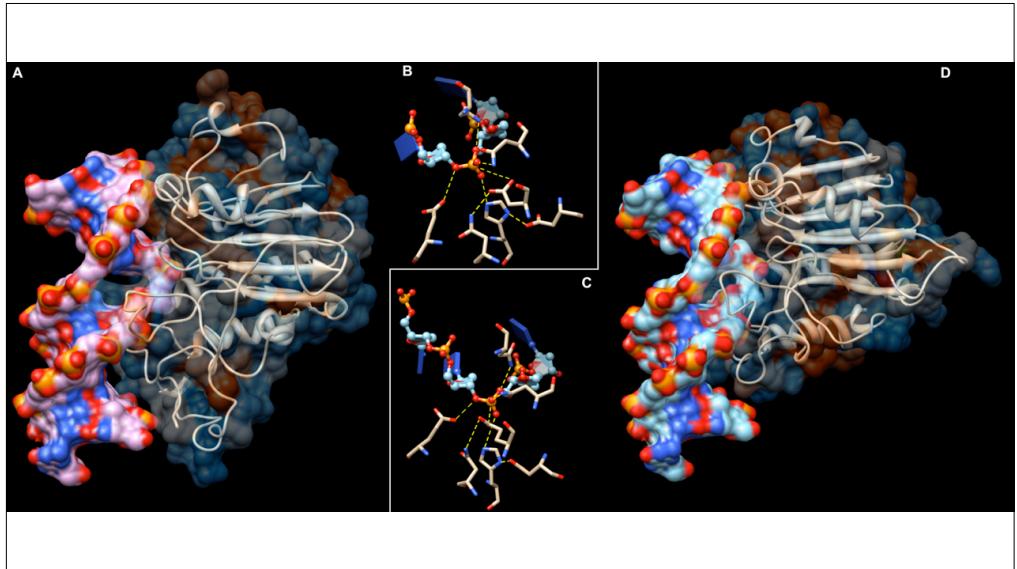
Componentes da estrutura terciária

- Fold (dobramento):**
 - Utilizado de forma diferente em contextos distintos.
 - De forma ampla é um arranjo 3D reproduzível e reconhecível.
- Motivos (Motifs):**
 - Também conhecidos como estruturas supersecundárias.
 - Um sub-componente reconhecível do fold.
 - Muitos motivos normalmente englobam um domínio.
- Domínio (Domain):**
 - Um componente da proteína compacto e que sofre *self folding*, que normalmente representa uma unidade estrutural e funcional discreta.
 - São aceitos como as unidades funcionais ou evolucionárias das estruturas de proteínas.

Natureza modular das proteínas

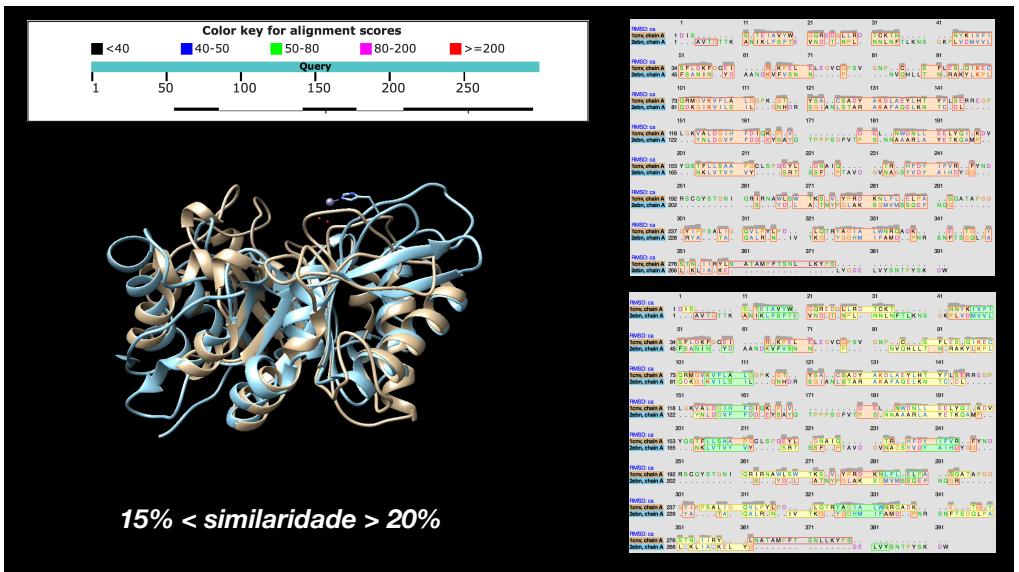
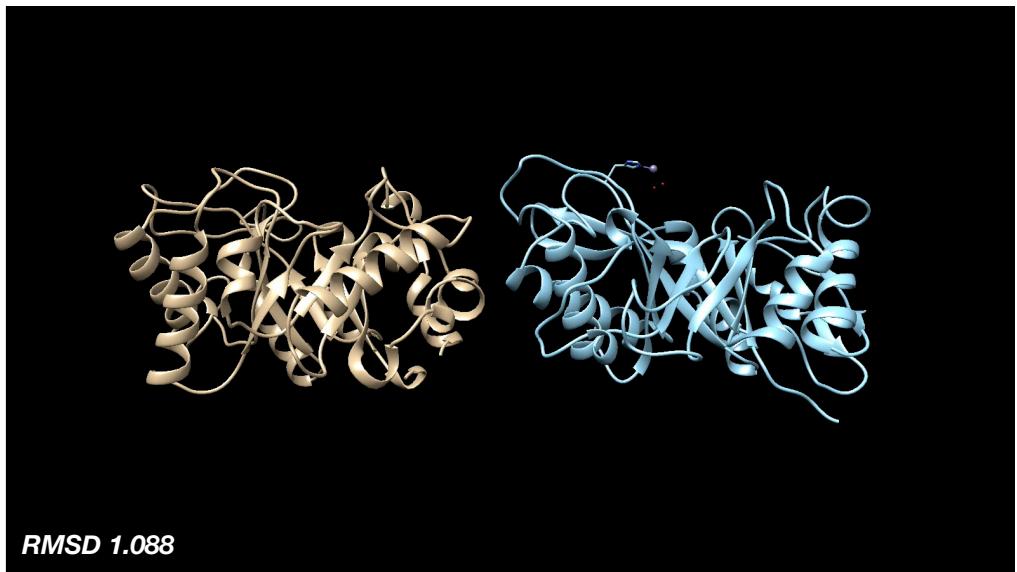
- As estruturas de proteínas são normalmente inherentemente flexíveis e podem ser moldadas para desempenhar um amplo espectro de funções.
 - O desenvolvimento de um novo sistema ou função pode ser alcançado por meio da reengenharia ou combinação de sistemas existentes.
- A modularidade torna as estruturas "evoluíveis", ou capazes de lidar com pressões seletivas distintas, diminuindo o número de restrições.
- Os sistemas modulares também promovem a emergência de novas funções por rearranjos.
- A modularidade nas estruturas protéicas pode ser vista como as ligações dentro das unidades funcionais.

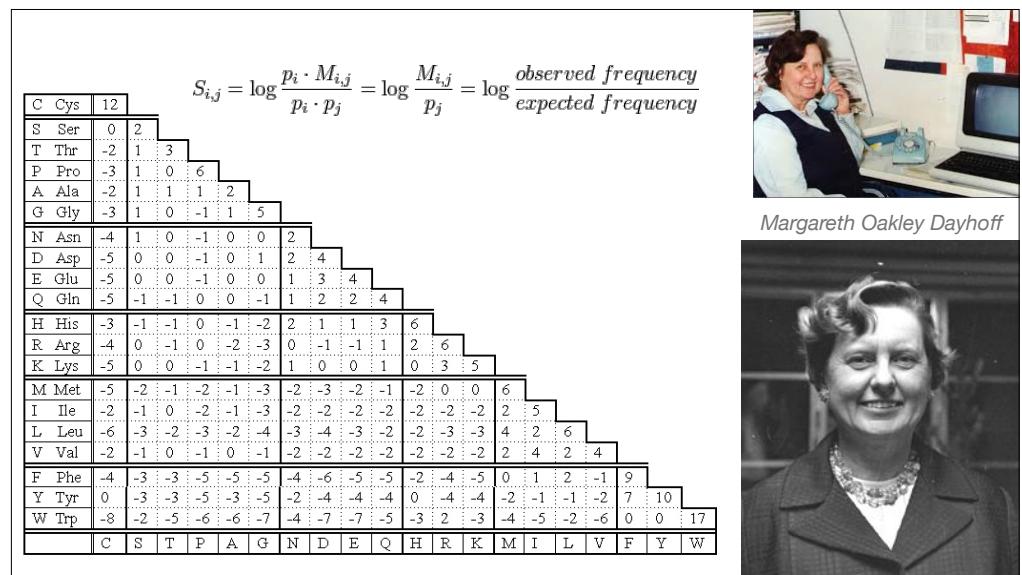




Alinhamento de sequências codificantes para proteína

- Necessita de cuidados adicionais.
 - Nem sempre é sensível alinhar sequências de DNA que codificam proteínas. Isto dependerá bastante do grau de conservação de sequência.
- ATGCTGTAGGG → ATGCT-GTTAGGG
 ATGCTCGTAGGG → ATGCTCGTA-GGG
- O resultado pode ser nada plausível e pode não refletir nenhum processo biológico conhecido ou possível para aquelas sequências.
 - Neste caso é muito mais sensível que as sequências sejam traduzidas para as suas sequências de aminoácidos respectivas, alinhar a sequências de AAs e depois colocar os gaps nas sequências de DNA de acordo onde eles são encontrados no alinhamento de aminoácidos.
 - Ou realizar um alinhamento baseado em códons - *Codon-based alignment*.

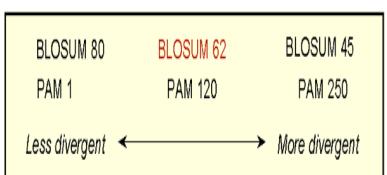




BLOSUM 62																					
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V		
A	4																				
R	1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	9																	
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-2	-3	-4	-1	-2	-3	-4	-3	2	4											
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V		

Matriz BLOSUM62

Escolha das Matrizes:



Protein Query Length	Matrix	Open Gap	Extend Gap
>300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
>300	PAM250	-10	-2
85-300	PAM120	-16	-4
35-85	MDM40	-12	-2
<=35	MDM20	-22	-4
<=10	MDM10	-23	-4

PAM100 ==> Blosum90
PAM120 ==> Blosum80
PAM160 ==> Blosum60
PAM200 ==> Blosum52
PAM250 ==> Blosum45

Objetivos do Alinhamento

◆ Predição de Estruturas:

- Deduzir a estrutura secundária e terciária de um produto de um gene a partir do conhecimento da sequência do gene.

◆ Comparação de Sequências:

- Estimativa de consensus e distância genética; predição e anotação de sítios funcionais; predição de genes, identificação e validação; desenho de primers e de drogas; classificação de famílias de proteínas e RNAs.

◆ Busca em Banco de Dados:

- Maximizar a distinção entre sequências homólogas e não-homólogas.

◆ Filogenia:

- Produzir hipóteses plausíveis de homologia evolucionária entre os resíduos.

(a) Structure

1smlA MAGHTPGSTAWTWTDRNGKPVRIAYADS---LSA
 1qh5A TPCHTSGHICYFVSK- PGGSEPPAVFTGDTLF???
 Structure --BBBTT-SSSSS-- TT----SSSS--TSS---

(b) Function

1smlA MAGHTPGSTAWTWTDRNGKPVRIAYADSLSA
 1qh5A TPCHTSGHICYFVSK- PGGSEPPAVFTGDTLF
 Function -A-ZAA-----A--A

(c) Database searching

1smlA ??MAGHTPGSTAWTWTDT---RNGKPVRIAYADSLSA
 1qh5A TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF?????
 Consensus TPHPGHGPVHVYVYLGGR---KVLFTGDLLFSGGCGR

(d) Phylogeny

1smlA MAGHTPGSTAWTWTDRNGKPVRIAYADSLSA
 1qh5A TPCHTSGHICYFVSKPG- GSEPPAVFTGDTLF

(e) Global similarity

1smlA MAGHTPGSTAWTWTDRNGKPVRIAYADSLSA
 1qh5A TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF?

(f) Local similarity

1smlA ---magHTPGSTAWTWTDRNGKPVRIAYADSLSA
 1qh5A tpc---HTSGHICYFVSKPGGSEPPAVFTGDTLF?

Como utilizar estas informações nos BDs?

- Usando apenas a sequência:
 - Realizando buscas de similaridade (como as realizadas com o BLAST).
- Usando a sequência e conservação:
 - Construindo um perfil:
 - Motivos lineares (como por exemplo as assinaturas PROSITE).
 - Composição geral de sequência (construção de uma nova matriz de posição específica - PSSM - utilizada no PSI-BLAST ou HMMs).



Figure 6.16 Pairwise similarity search of the databank using single 'query' sequence

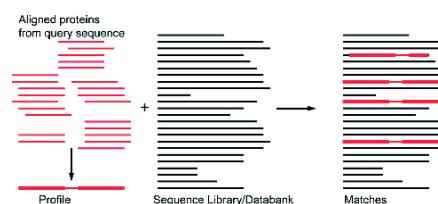
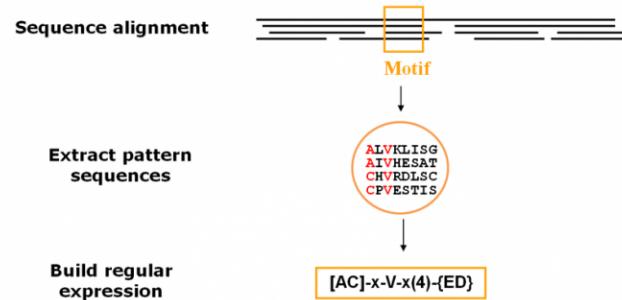


Figure 6.17 The use of a profile established from aligned proteins to improve quality of matches

Padrões

- Padrões:

- Regiões da sequência com alguns aminoácidos que são essenciais para a função proteica. Podem ser representados por expressão regular.



Assinaturas PROSITE



C-x(5)-PVCC-x(1,4)-G-x(1,6)-T-x(2)-N-x(1)-C-x(7,14)-G-x(1)-C-x(1,5)-[HN]-x(4)-P

C-x(5)-PVC-x(4,10)-[TS]-x(14,20)-G-x(1)-C-x(14,68)-P-[SV]-C

<A-x-[ST](2)-x(0,1)-{V}>

Assinaturas

- O seguinte padrão:

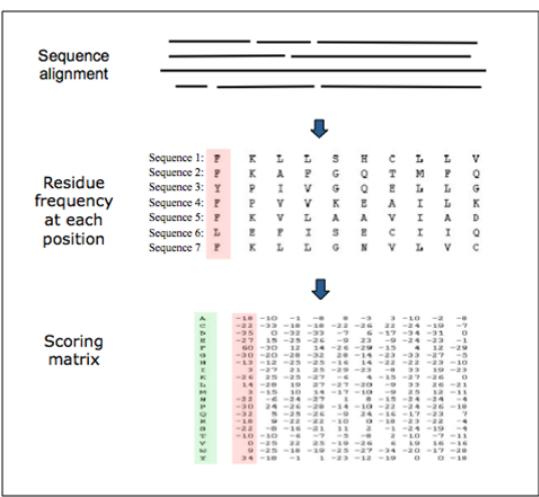
<A-x-[ST](2)-x(0,1)-{V}>

- Significa:

- Uma Ala no N-terminal.
- Seguida por qualquer aminoácido.
- Seguida por uma Ser ou Thr duas vezes.
- Seguida ou não por qualquer resíduo.
- Seguida por qualquer aminoácido exceto Val.

Profiles (Perfis)

- Perfis:
 - Utilizados para modelar domínios e famílias de proteínas. Eles são construídos a partir da conversão de alinhamento múltiplos de sequências em sistemas de pontuação específicos de posição (*position-specific scoring systems - PSSMs*).
 - Aminoácidos em cada posição no alinhamento são pontuados de acordo com a frequência que eles ocorrem.



PSSMs

Alignment

PFM
Position Frequency Matrix

$$M = \begin{bmatrix} A & 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ C & 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ G & 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ T & 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

PPM
Position Probability Matrix

$$M = \begin{bmatrix} A & 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ C & 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ G & 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ T & 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^N I(X_{i,j} = k)$$

PSSMs

- Probabilidade da sequência S = GAGGTAAAC:
 - $p(S|M) = 0.1 \times 0.6 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 0.7 \times 0.2 \times 0.2 = 0.0007056.$

PSSMs (PWM - Position Weight Matrix)

$$M_{k,j} = \log_2 (M_{k,j}/b_k)$$

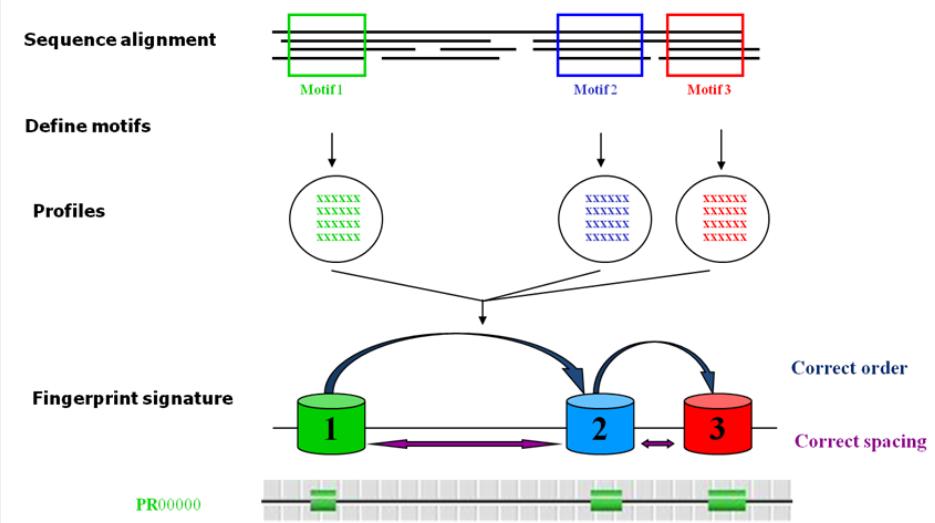
- $b_k = 1/k$ para todos os símbolos do alfabeto (0.25 para nucleotídeos and 0.05 para aminoácidos).

$$M = \begin{matrix} A & [0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32] \\ C & [-0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32] \\ G & [-1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32] \\ T & [0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26] \end{matrix}$$

- As entradas $-\infty$ na matriz são pseudocontagens.

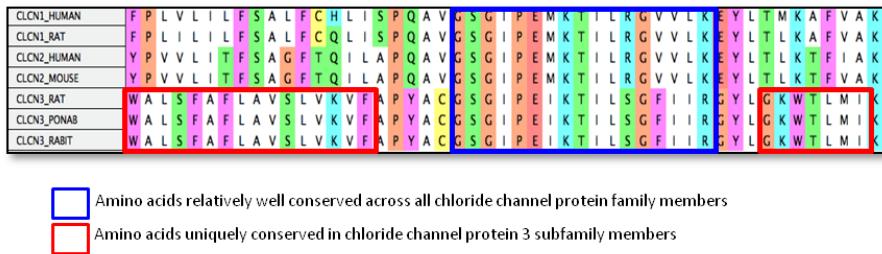
Fingerprints

- Quando algumas famílias proteicas são caracterizadas por mais de uma região conservada, que ocorrem em uma certa ordem.
- A identificação destas regiões é o princípio das *Fingerprints*:
 - Descrição dos múltiplos motivos conservados, retirados a partir do alinhamento da sequência.
 - Cada motivo é convertido em um motivo individual.



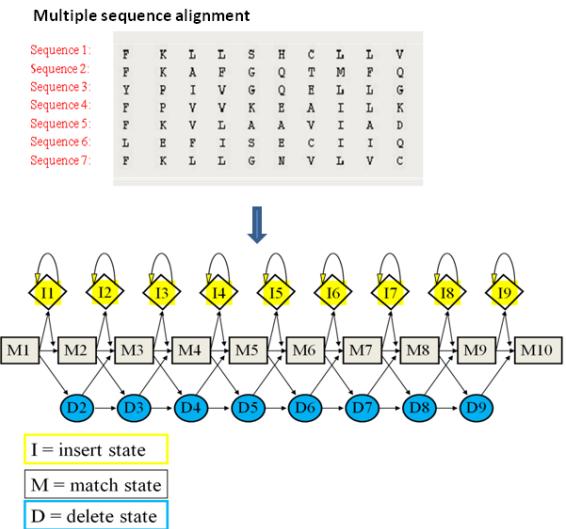
Fingerprints

- As fingerprints são também úteis para identificar pequenas diferenças em proteínas proximamente relacionadas.

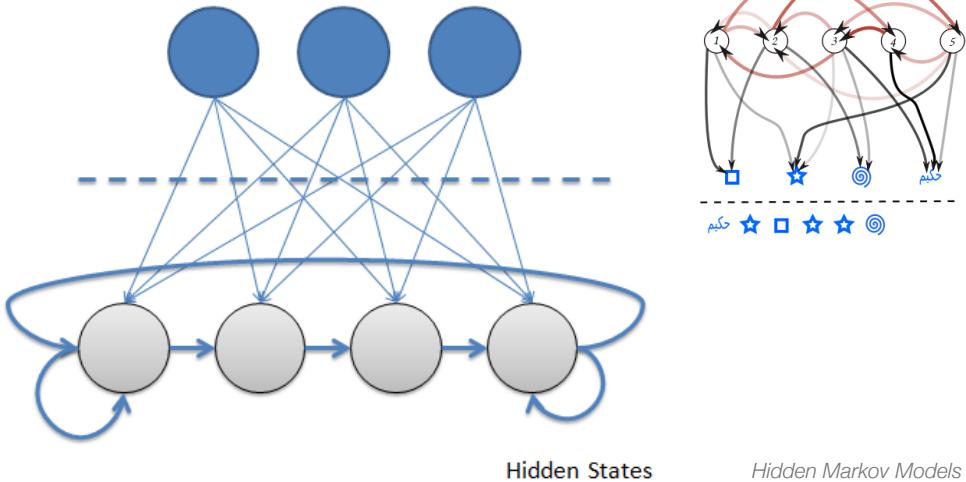


HMMs

- HMMs:
 - É um modelo estatístico para qualquer sistema que pode ser representado como uma sucessão de transições entre estados discretos.
 - HMMs representam inserções e remoções de aminoácidos, em um sentido que eles podem modelar alinhamentos inteiros, incluindo regiões divergentes.
 - São bem adequadas para busca de similaridades em bancos de dados.



Observable States

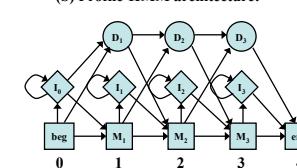


HMMs a partir do alinhamento

- Ideia principal:
 - Modelo que representa o consenso para o alinhamento de sequência da mesma família.
 - Não a seqüência de qualquer membro em particular.

(a) Multiple alignment:

(b) Profile-HMM architecture:



HBA_HUMAN	...VGA--HAGEY...
HBB_HUMAN	...V---NVDEV...
MYG_PHYCA	...VEA--DVAGH...
GLB3_CHITP	...VKG-----D...
GLB5_PETMA	...VYS--TYETS...
LGB2_LUPLU	...FNA--NIPKH...
GLB1_GLYDI	...IAGADNGAGV...

(c) Observed emission/transition counts

	0	1	2	3
Emission from M	A	0	4	0
	C	-	0	4
	G	-	0	5
	T	-	0	0
Emission from I	A	0	0	6
	C	0	0	0
	G	0	0	1
	T	0	0	0
Transition probabilities	M-M	4	4	2
	M-D	1	0	0
	M-I	0	0	3
	I-M	0	0	2
	I-I	0	0	4
	D-M	-	1	0
	D-D	-	0	0

HMMs a partir do alinhamento

	1	2	3	4	5	6	7	8
Alignment	A	C	D	E	F A C A	D	F	
	A	F	D	A	— — C	C	F	
Alignment	A	—	—	E	F D — F	D	C	
	A	C	A	E	F — — A	—	C	
Alignment	A	D	D	E	F A A A	D	F	

1 2 3 4 5 6 7 8

A	C	D	E	F	A C A	D	F
A	F	D	A	— — C	C	C	F
Alignment	A	—	—	E	F D — F	D	C
	A	C	A	E	F — — A	—	C
	A	D	D	E	F A A A	D	F

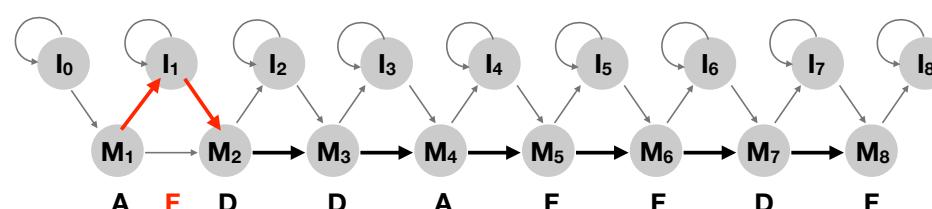
Remova as colunas se a fração de inserções excede o limiar de frações máximas de inserções, θ .

1 2 3 4 5 6 7 8

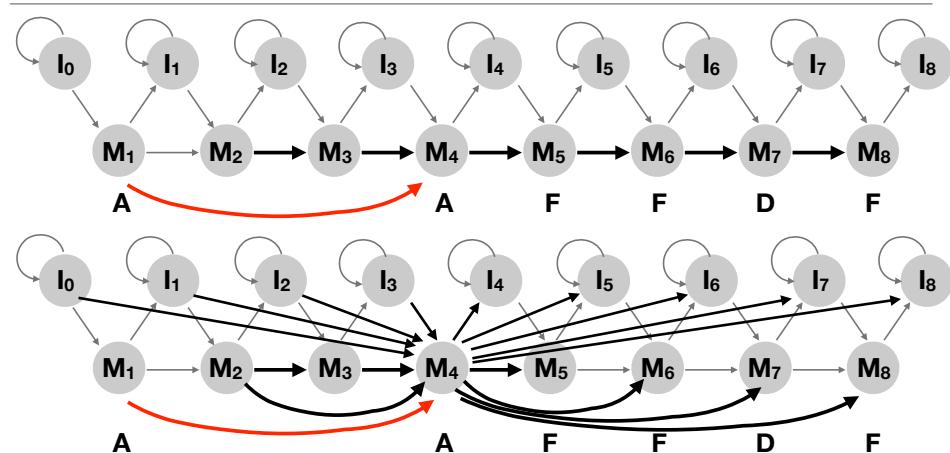
A	C	D	E	F	A	D	F	
A	F	D	A	—	C	C	F	
Alignment [†]	A	—	—	E	F	F	D	C
	A	C	A	E	F	A	—	C
	A	D	D	E	F	A	D	F

1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Alignment	A	C	D	E	F A C A	D	F	A	C	D	E	F	A	D	F
Alignment [†]	A	—	—	E	F D — F	D	C	A	—	E	F	F	D	C	
Profile (Alignment*) [†]	A	—	—	E	F D — F	D	C	A	—	E	F	F	D	C	
Profile (Alignment*)D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Profile (Alignment*)A	1/4	3/4	0	0	0	0	0	0	0	0	0	0	0	0	0
Profile (Alignment*)C	2/4	0	0	0	0	0	0	1/5	1/5	1/4	2/5	0	0	0	0
Profile (Alignment*)E	0	0	0	4/5	0	0	0	0	0	0	0	0	0	0	0
Profile (Alignment*)F	0	1/4	0	0	0	1	0	1/5	0	0	3/5	0	0	0	0
HMM Diagram	M ₁	→	M ₂	→	M ₃	→	M ₄	→	M ₅	→	M ₆	→	M ₇	→	M ₈
	A	D	D	A	F	F	D	A	F	F	D	D	F	F	
	1	0.25	0.75	0.20	1	0.20	0.75	0.60							

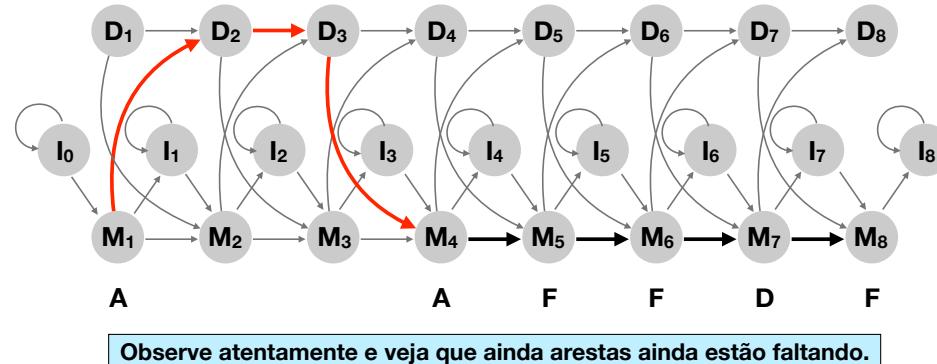
Modelando as inserções



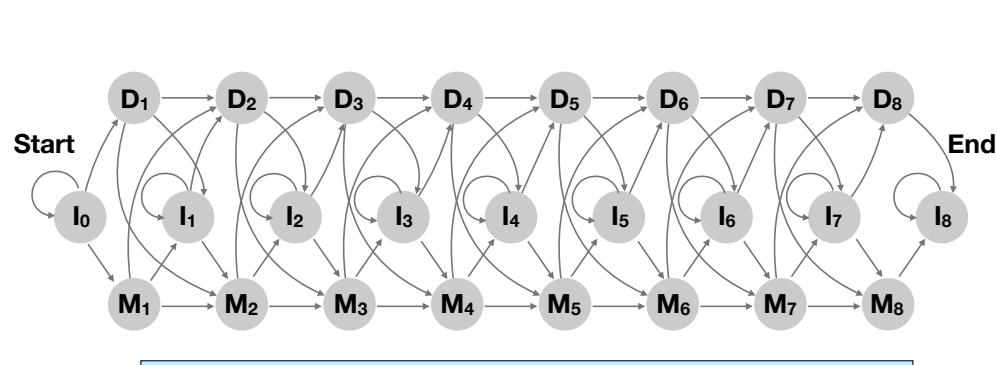
Modelando as remoções



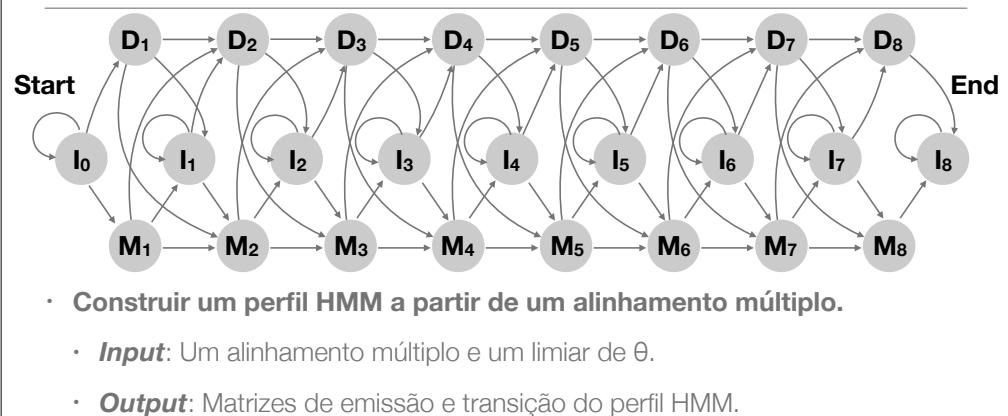
Adicionando estados de remoções

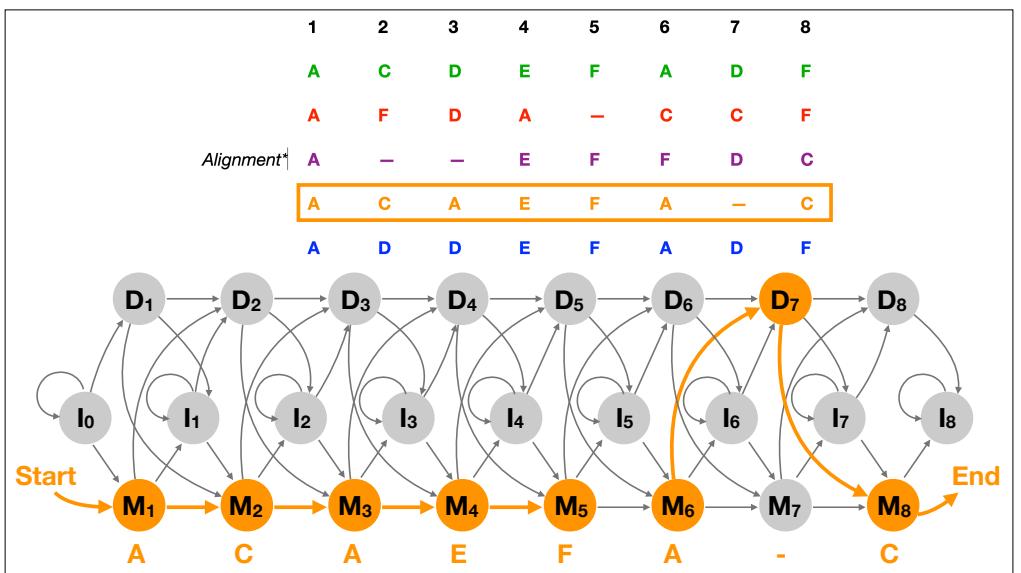
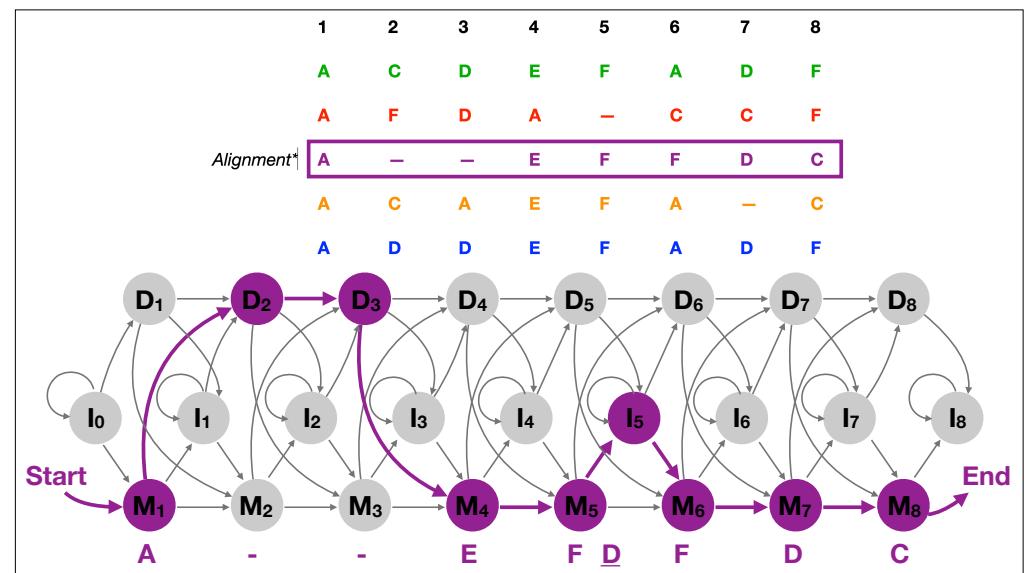
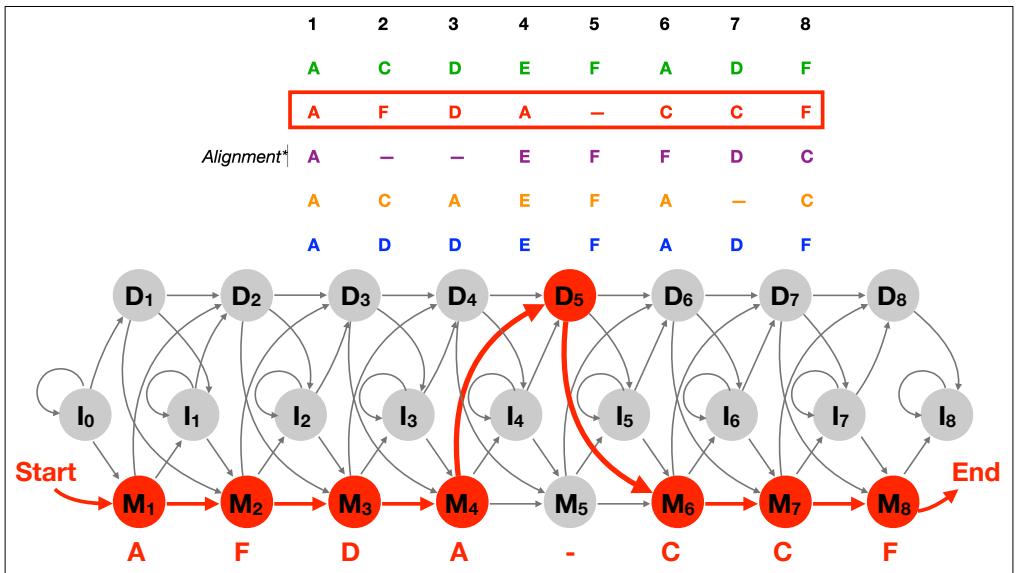
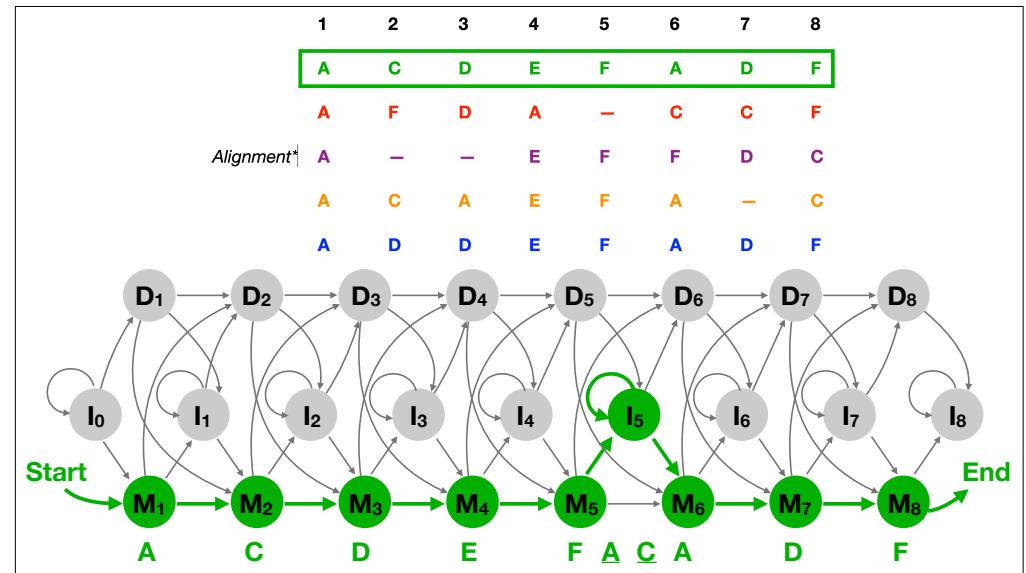


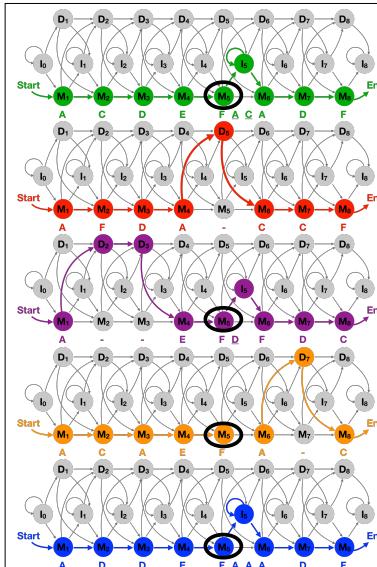
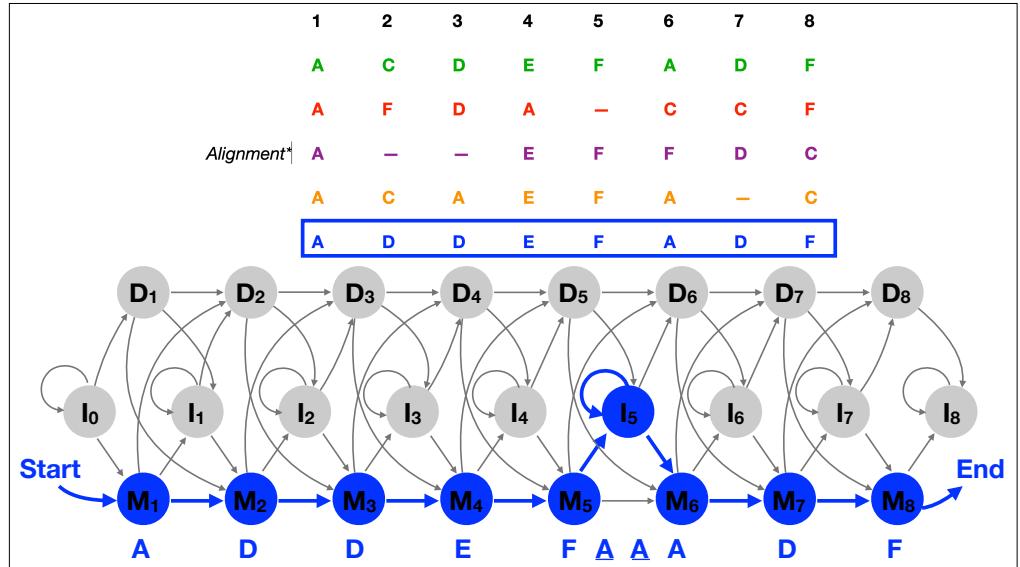
Adicionando arestas entre inserções e remoções



Adicionando arestas entre inserções e remoções







Probabilidades de transição:

4 Transições a partir de M₅:

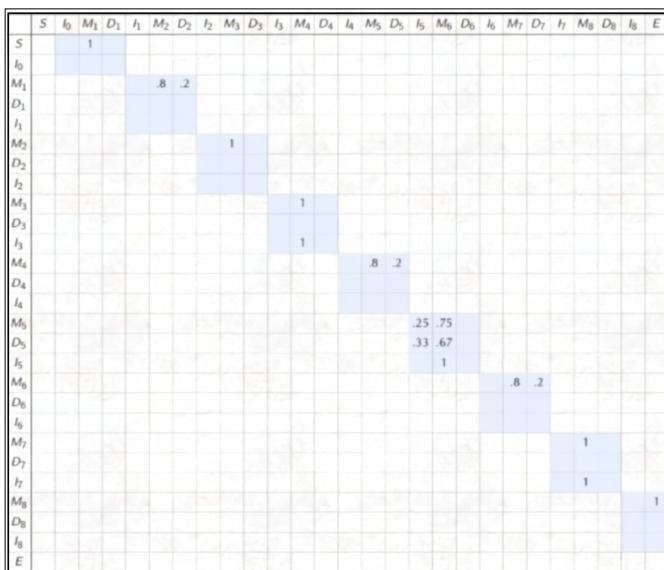
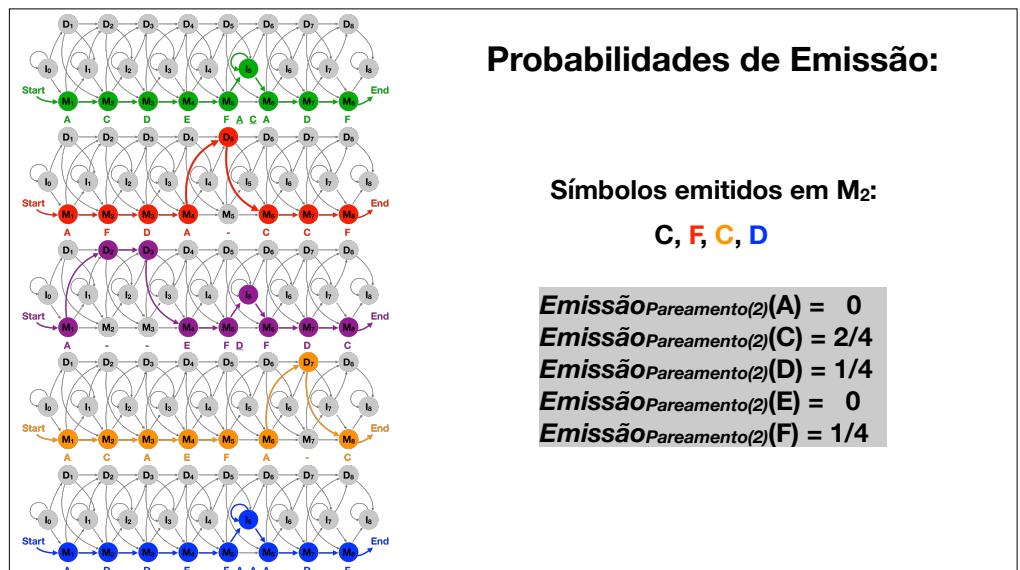
1 para M₆

0 para D₆

transição Pareamento(5), Inserção(5) = 3/4

transição Pareamento(5), Pareamento(6) = 1/4

transição Pareamento(5), Remoção(6) = 0



Probabilidades de Emissão:

Símbolos emitidos em M_2 :

C. F. C. D

$$Emiss\~ao_{Pareamento(2)}(A) = 0$$

$$Emiss\~ao_{Pareamento(2)}(C) = 2/4$$

Emissão Pareamento(2)(D) = 1/4

Emissão Pareamento(2)(E) = 0

$$Emissão_{Pareamento(2)}(E) = 0$$

Transições não permitidas

Células cinzas:

Arestas no diagrama HMM.

Células cinzas:

Transições não permitidas.

**Não esquecer as
pseudocontagens!**

Busca em Bancos de Dados Biológicos

- Ferramenta BLAST (Altschul et al.):
 - Basic Local Alignment Search Tool:
 - Programa mais utilizado para busca de similaridade em bancos de dados;
 - A partir de uma seqüência query o programa procura achar todas as seqüências do banco (subjects) que têm alinhamento com significância estatística;
 - É uma heurística. Realiza uma comparação local. Utiliza sementes e procura estender esses casamentos ao longo de suas respectivas diagonais;
 - A extensão do alinhamento é controlada pela pontuação. Se esta cai abaixo de um certo limite, a extensão é interrompida.

BLAST

- **Parâmetros de Avaliação da qualidade do alinhamento:**
 - **Bit score:**
 - É a pontuação do alinhamento obtido.
 - É um escore normalizado, que permite a comparação entre diferentes alinhamentos, mesmo que estes tenham sido obtidos com diferentes parâmetros de busca.
 - **E-value:**
 - Dá o número de diferentes alinhamentos (não é uma probabilidade), com escores tão bons quanto o obtido, que se espera que ocorram neste particular banco de seqüências, com esta particular matriz de pontuação, que sejam devidos ao acaso.
 - Quanto menor o E-value, maior a significância daquele resultado.

BLAST

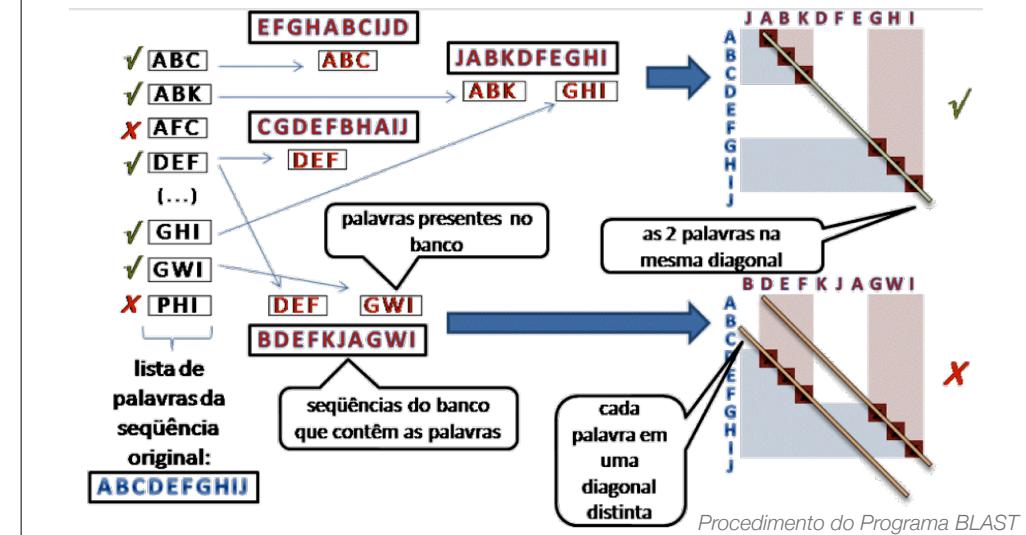
- Parâmetros de Avaliação da qualidade do alinhamento:
 - **Query coverage:**
 - É a porcentagem da sequência inicial que foi alinhada com a sequência do banco.
 - Dá uma noção do tamanho e extensão do alinhamento, não da qualidade.
 - **Identidade (Identity):**
 - Porcentagem que reflete o número de resíduos exatamente iguais entre a sequência inicial e a sequência do banco.
 - Dá uma noção da qualidade do alinhamento.
 - Em alinhamento de seqüências de AAs é dividida em um outro parâmetro:
 - *Positives:* Resíduos que não são iguais, mas a troca mantém a característica química do aminoácido.

BLAST – Versões

- BLASTn (nucleotide blast):
 - A partir de uma seqüência de nucleotídeos (DNA ou RNA), busca similaridade em um banco de seqüências de nucleotídeos.
- BLASTp (protein blast):
 - A partir de uma seqüência de aminoácidos, busca similaridade em um banco de seqüências proteicas.
- BLASTX:
 - A partir de uma seqüência de nucleotídeos, realiza a tradução nos 6 frames e busca similaridade em um banco de seqüências de proteínas.

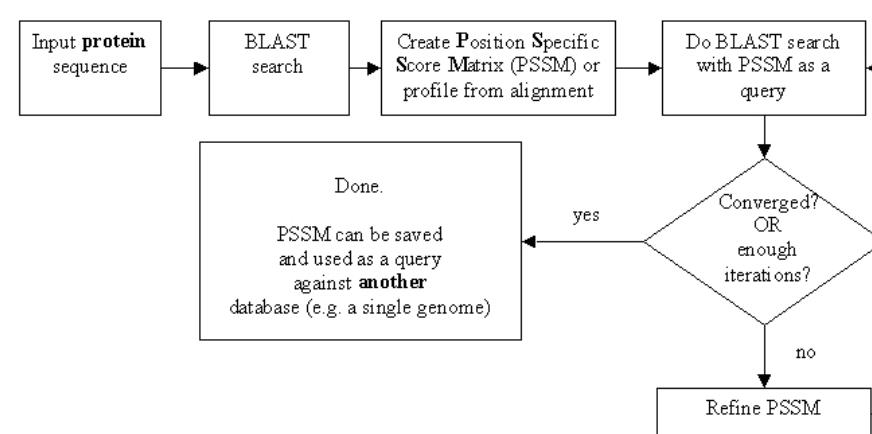
BLAST - Versões

- tBLASTn:
 - A partir de uma seqüência de aminoácidos, realiza busca em um banco de seqüências de nucleotídeos traduzida.
- tBLASTx:
 - A partir de uma seqüência de nucleotídeos traduzida, realiza busca em um banco de seqüências de nucleotídeos traduzidas.



BLAST - Outras Versões

- PSI-BLAST:
 - *Position Specific Iterative BLAST*
 - Deriva uma matriz de pontuação posição-específica (PSSM) ou um perfil a partir do alinhamento múltiplo das seqüências, acima de um limiar de um dado score, utilizando BLAST proteína-proteína.
 - A PSSM é utilizada para procurar adicionalmente novos hits no banco e é atualizada para iterações subsequentes com as novas seqüências detectadas.
 - É um meio eficiente de detectar relações distantes entre proteínas;
 - "Procura um perfil contra um banco de seqüências".

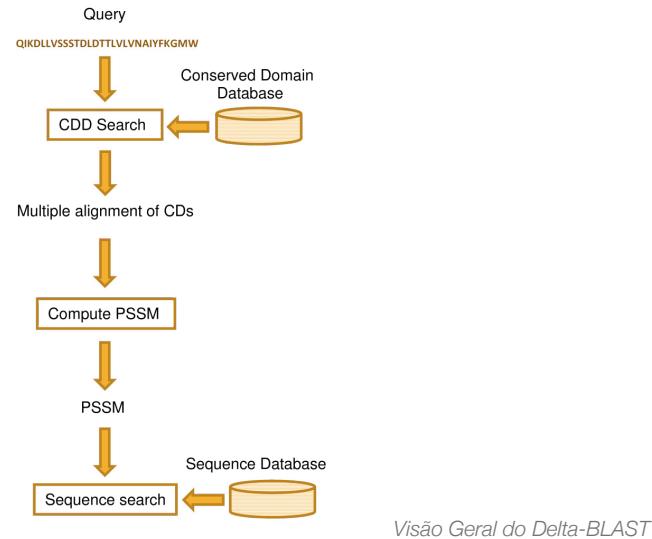


BLAST - Outras Versões

- PHI-BLAST:
 - *Pattern hit initiated BLAST*
 - Utiliza expressões regulares, reforçando a busca por um motivo específico;
 - As expressões regulares devem estar de acordo com as regras do PROSITE;
 - Muito útil na identificação de uma família de proteínas formalmente definida, assim como membros bem dissimilares de cada família;
 - Exemplo:
 - W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P

BLAST - Outras Versões

- DELTA-BLAST:
 - *Domain Enhanced Lookup Time Accelerated BLAST*
 - Também realiza uma busca a partir de PSSM. Corre uma rápida busca RPSBLAST para construir a PSSM;
 - Reverse PSI-BLAST: busca uma sequência contra um banco de perfis.
 - Utiliza a PSSM construída e procura contra um banco BLAST;
 - Mais preciso do que o BLASTP e similar a algumas iterações utilizando o PSI-BLAST.



Melhoramento do Alinhamento

(a)	taxon10... ...20... ...30... ...40... ...50
Fu	Nosema.40928	QPGLFSPEIIRRASSVAVLIR--YPTLNG--VIIKESGLVCAGHFGHIELVK
Fu	Aspergillus.	QPGLFSPEIIRRMSVHVRE--YPTMDEORGRRTKGLECPGHFGHIELAT
Ap	Plasmodium.3	ELGVLDPEIIRRMSVCEIVN--NVDIYKDC--FIREGGLYCPGHFGHIELAK
An	Cricetulus.2	QPGVLSPDELKRMNSVTEGCKYKPETT--GORIQLGLLECPGHFGHIELAK
An	Homo.7434727	QPGVLSPDELKRMNSVTEGCKYKPETT--GORIQLGLLECPGHFGHIELAK
An	Drosophila.133	QPGLSPPEIIRRMSVH--VEHSEPTMDESG_RURVGCLDCPCHFGHIELAK
Fu	Spombe.54881	QPGLSPPEIIRRMSVH--IEFPETMDESG_RURVGCLDCPCHFGHIELAK
P1	Athaliana.40	QPGLSPDELKRMNSVH--VEHSEPTEKKK1KVCGOLECPGHFGCYIELAK
My	Odискоидеум.	-----ECPGHFGHIELAK
Rh	Porphyra.316	-----ECPGHFGFIELAK
Kt	Tbrucei.1021	QPEIFKERQIKSYAVG_LVEHAKSYANAA---AUSGEAACPGHFGYIELAB
Kt	Leishmania.7	QPEVFKEAOIKAYAKCQ_IIEHAKSYEHG---QIVRGGIECPGHFGYVELAB

(b)	taxon10... ...20... ...30... ...40... ...50
Fu	Nosema.40928	QPGLFSPEIIRRASSVAVLIR--YPTLNG--VIIKESGLVCAGHFGHIELVK
Fu	Aspergillus.	QPGLFSPEIIRRMSVHVRE--YPTMDEORGRRTKGLECPGHFGHIELAT
Fu	Spombe.54881	QPGLSPPEIIRRMSVH--IEFPETMDESG_RURVGCLDCPCHFGHIELAK
Ap	Plasmodium.3	ELGVLDPEIIRRMSVCEIVN--NVDIYKDC--FIREGGLYCPGHFGHIELAK
An	Cricetulus.2	QPGVLSPDELKRMNSVTEGCKYKPETT--GORIQLGLLECPGHFGHIELAK
An	Homo.7434727	QPGVLSPDELKRMNSVTEGCKYKPETT--GORIQLGLLECPGHFGHIELAK
An	Drosophila.9	QPGLSPPEIIRRMSVH--VEHSEPTMDESG_RURVGCLDCPCHFGHIELAK
An	Celebens.133	QPGLSPPEIIRRMSVH--VEHSEPTMDESG_RURVGCLDCPCHFGHIELAK
P1	Athaliana.40	QPGLSPDELKRMNSVH--VEHSEPTEKKK1KVCGOLECPGHFGCYIELAK
My	Oдискоидеум.	-----ECPGHFGHIELAK
Rh	Porphyra.316	-----ECPGHFGFIELAK
Kt	Tbrucei.1021	QPEIFKERQIKSYAVG_LVEHAKSYANAA---AUSGEAACPGHFGYIELAB
Kt	Leishmania.7	QPEVFKEAOIKAYAKCQ_IIEHAKSYEHG---QIVRGGIECPGHFGYVELAB