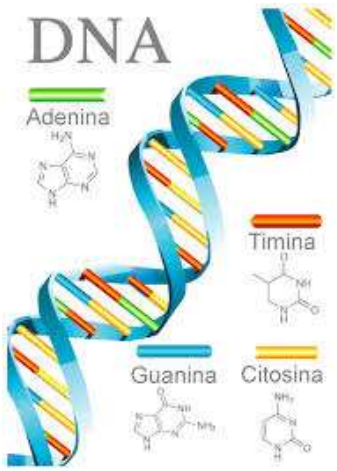
Universidade Federal do Espírito Santo - UFES
Programa de Pós-Graduação em Biotecnologia

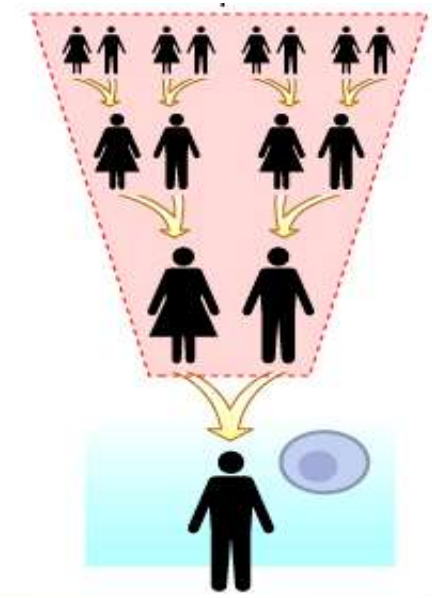
BIOINFORMÁTICA

BRUNA SOUSA DE MENDONÇA PORTELA
Prof. Dr. Teodiano Freire Bastos Filho

INTRODUÇÃO

ORGANISMOS

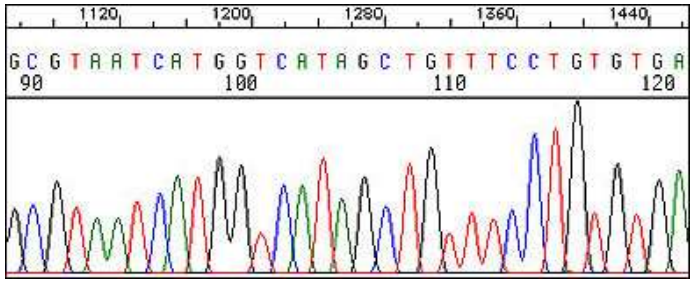




Qualquer par de organismo, não importa quão diferentes, têm um ancestral comum em algum momento do passado a partir do qual eles evoluíram

INTRODUÇÃO

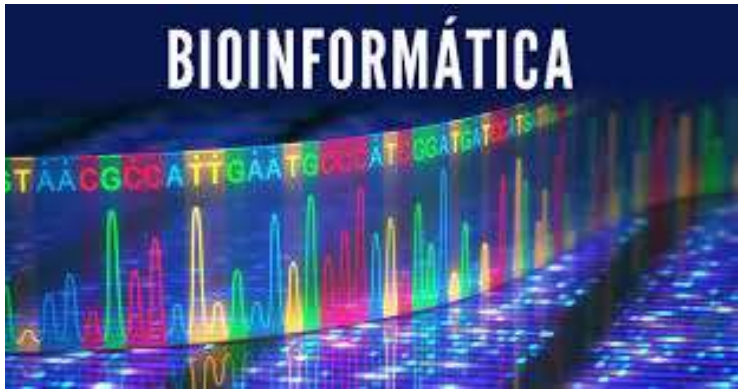
A capacidade de sequenciar o DNA de um organismo é uma das exigências mais importantes e primárias na pesquisa biológica



Informações funcionais e/ou estruturais

Identificar e agregar informação à sequência

INTRODUÇÃO



Prever informações estruturais
e funcionais dessas sequencias
nucleotídicas

Alinhamento de Sequências

ALINHAMENTO DE SEQUÊNCIAS

- **DEFINIÇÃO**
  - *Comparação de duas ou mais sequências por meio de buscas de uma série de caracteres ou padrões de caracteres que estão na mesma ordem.*

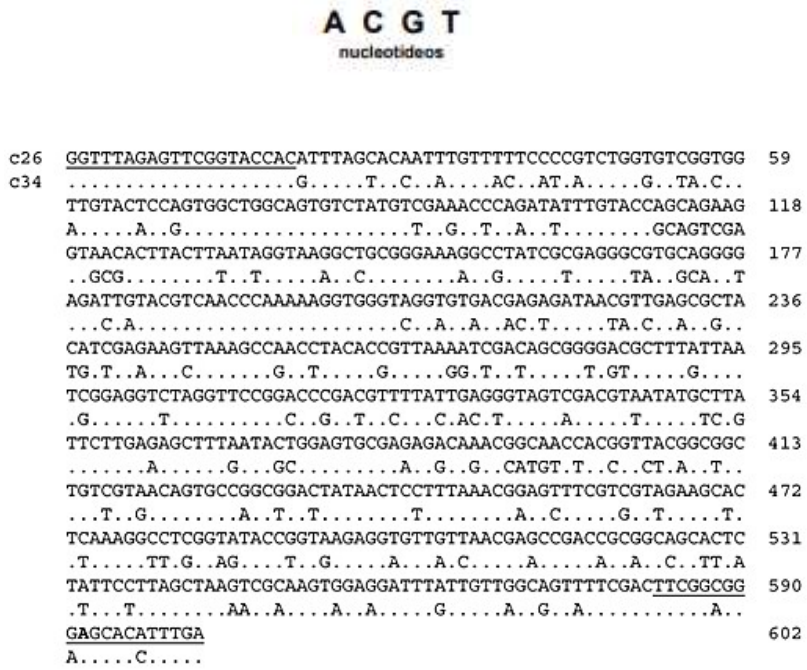
- **UTILIDADE**
  - *Comparar genes de DNA;*
  - *Estudar a estrutura de proteínas;*
  - *Estudar a evolução molecular;*
  - *Detecção de doenças, vírus, etc.*

Realizar o ALINHAMENTO facilita a computação da similaridade entre duas sequências

ALINHAMENTO DE SEQUÊNCIAS

- As sequências podem ser:

— *Sequências nucleotídicas (DNA ou RNA)*



— *Sequências de aminoácidos (proteínas)*

ALINHAMENTO DE SEQUÊNCIAS

- Durante o alinhamento, as sequências são organizadas em linhas e os caracteres biológicos integram as colunas do alinhamento.

Grupo de sequências não alinhadas

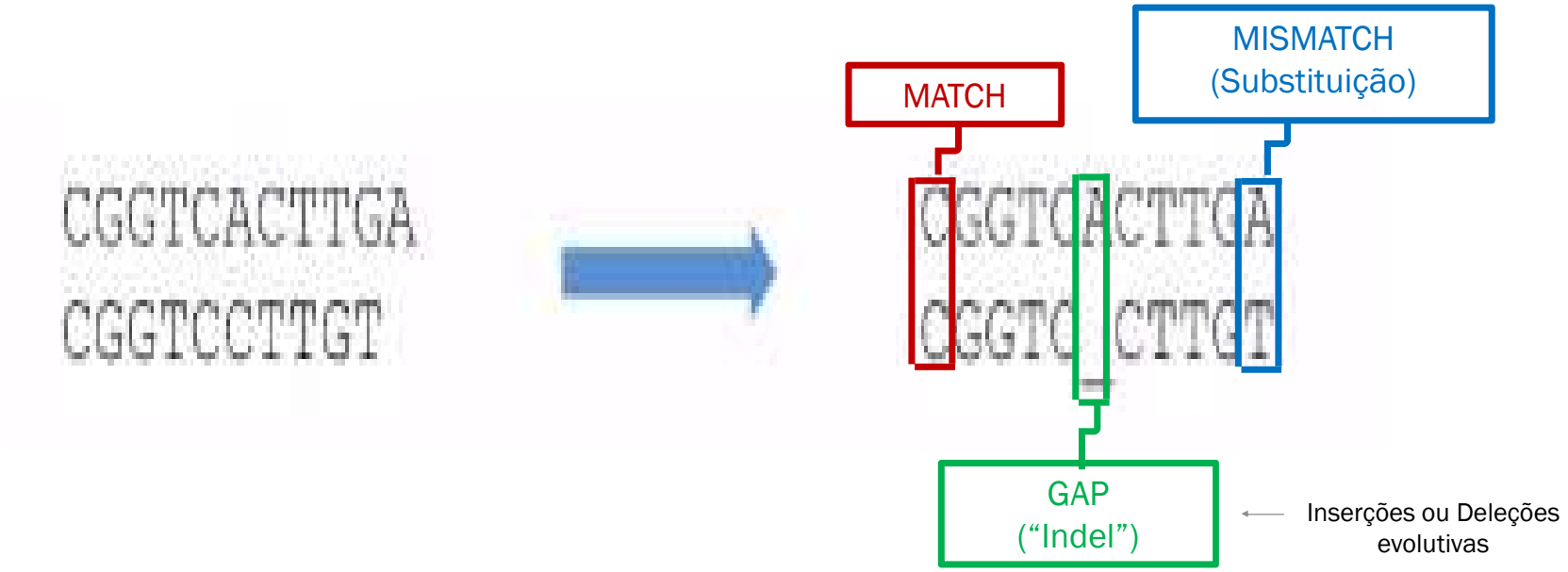
```
                    10        20        30        40        50
                    .........|.........|.........|.........|.........|.....
Sequência 01        ATATACTTGATCGATCAGCATCAGCTAGTCGAAGTTTGAGCATGCATGTGTGATG
Sequência 02        CTTGTTCCATCAGCTTCAGCTCGTCGAAGGAGCTGCATGTGTGA
Sequência 03        CTTGTTCCATCAGCTTCAGCTCGTCGAAGGAGCATGCA
Sequência 04        ATATACTTGATCGATTAGCTTCAGCTAGTGGAGCCAGTATGTGTGTG
                     *              *      **   * *
```

Grupo de sequências alinhadas

```
                    10        20        30        40        50
                    .........|.........|.........|.........|.........|.....
Sequência 01        ATATACTTGATCGATCAGCATCAGCTAGTCGAAGTTTGAGCATGCATGTGTGATG
Sequência 02        -----CTTGTTCCATCAGCTTCAGCTCGTCGAAG---GAGC-TGCATGTGTGA--
Sequência 03        -----CTTGTTCCATCAGCTTCAGCTCGTCGAAG---GAGCATGCA---------
Sequência 04        ATATACTTGATCGATTAGCTTCAGCTAGT----G---GAGCCAGTATGTGTG-TG
                          **** **  **  *** ****** **     *    ****   * *
```

ALINHAMENTO DE SEQUÊNCIAS

- O processo de rearranjo pode introduzir um ou mais espaços ou intervalos no Alinhamento.



ALINHAMENTO DE SEQUÊNCIAS

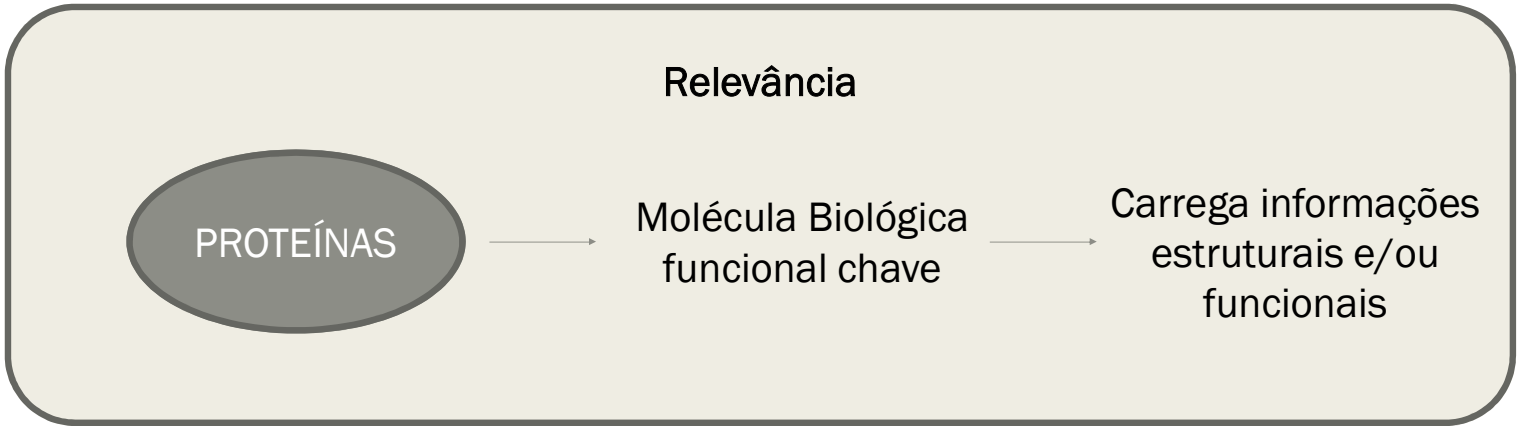
- Tipos de Alinhamentos de Sequências:
  - *Alinhamento simples ou par-a-par: considera 2 sequências de cada vez;*
  - *Alinhamento de múltiplas sequências: alinha várias sequências relacionadas.*

Pré-requisito para:

- Análises genômicas comparativas para identificação e quantificação de regiões conservadas ou motivos funcionais em uma família de sequências completas;
- Estimativa de divergência evolutiva entre sequências;
- Perfis de sequências ancestrais.

ALINHAMENTO DE SEQUÊNCIAS

- Tipos de Alinhamentos de Sequências:
  - *Alinhamento de sequências a nível de aminoácidos*



ALINHAMENTO DE SEQUÊNCIAS

- Programas baseados em Alinhamentos, independente do algoritmo subjacente, procuram correspondência de bases individuais ou aminoácidos que estão na mesma ordem em 2 ou mais sequências.

- Cada símbolo de sequência pode ser categorizado em pelo menos um dos 2 estados:

  - *Conservado/Semelhante: Correspondência*
  - *Não conservado: Incompatibilidade*

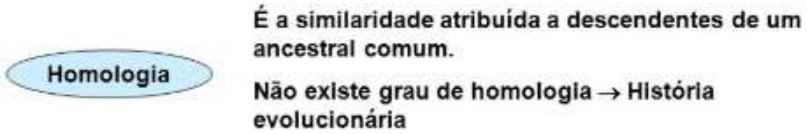
- Os programas também modelam estados: Inseridos/Excluídos (lacunas)

ALINHAMENTO DE SEQUÊNCIAS

- **Métodos de pontuação** – conhecer o nível de identidade ou similaridade.
  - *IDENTIDADE: refere-se à presença do mesmo nucleotídeo ou aminoácido na mesma posição em duas sequências alinhadas.*

  - *SIMILARIDADE: é dada pelo Alinhamento que tem maior pontuação entre todos os alinhamentos possíveis.*

  - *HOMOLOGIA: dividem a mesma ancestralidade com significado evolutivo.*



ALINHAMENTO DE SEQUÊNCIAS

Matrizes de substituição

- Sistema de pontuação biologicamente relevante;
- Para produzir alinhamentos biologicamente significativos
- Matrizes:
  - *PAM (Point Accepted Mutation)*
  - *BLOSUM (BLOcks Substituition Matrix)*

BLAST

  - *Aminoácidos*
  - *Nucleotídeos*

FASTA

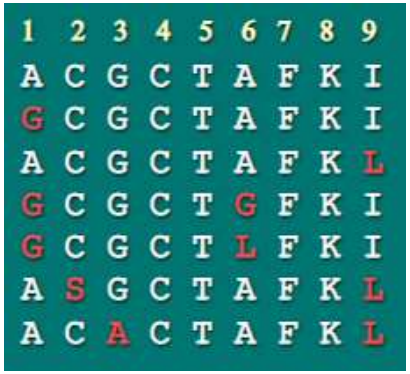
ALINHAMENTO DE SEQUÊNCIAS

Matriz PAM

- Calcula a mutabilidade relativa de cada aminoácido.

-

  - *Valor = 0: frequência de substituição entre dois aminoácidos é esperada ao acaso;*
  - *Valor < 0: frequência é menor que a esperada. Substituição de dois aminoácidos ao acaso.*
  - *Valor > 0: frequência maior que a esperada. Substituição não é ao acaso. Indica maior probabilidade de relação com ancestral.*



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

ALINHAMENTO DE SEQUÊNCIAS

Matriz BLOSUM

- Calcula a frequência de cada par de aminoácido.

-

  - *Valor = 0: probabilidade de substituição entre dois aminoácidos iguais;*
  - *Valor < 0: maior probabilidade de substituição de dois aminoácidos ser ao acaso.*
  - *Valor > 0: maior probabilidade de substituição de dois aminoácidos NÃO ser por acaso. Indica maior probabilidade de relação com ancestral.*

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

ALINHAMENTO DE SEQUÊNCIAS

PAM x BLOSUM

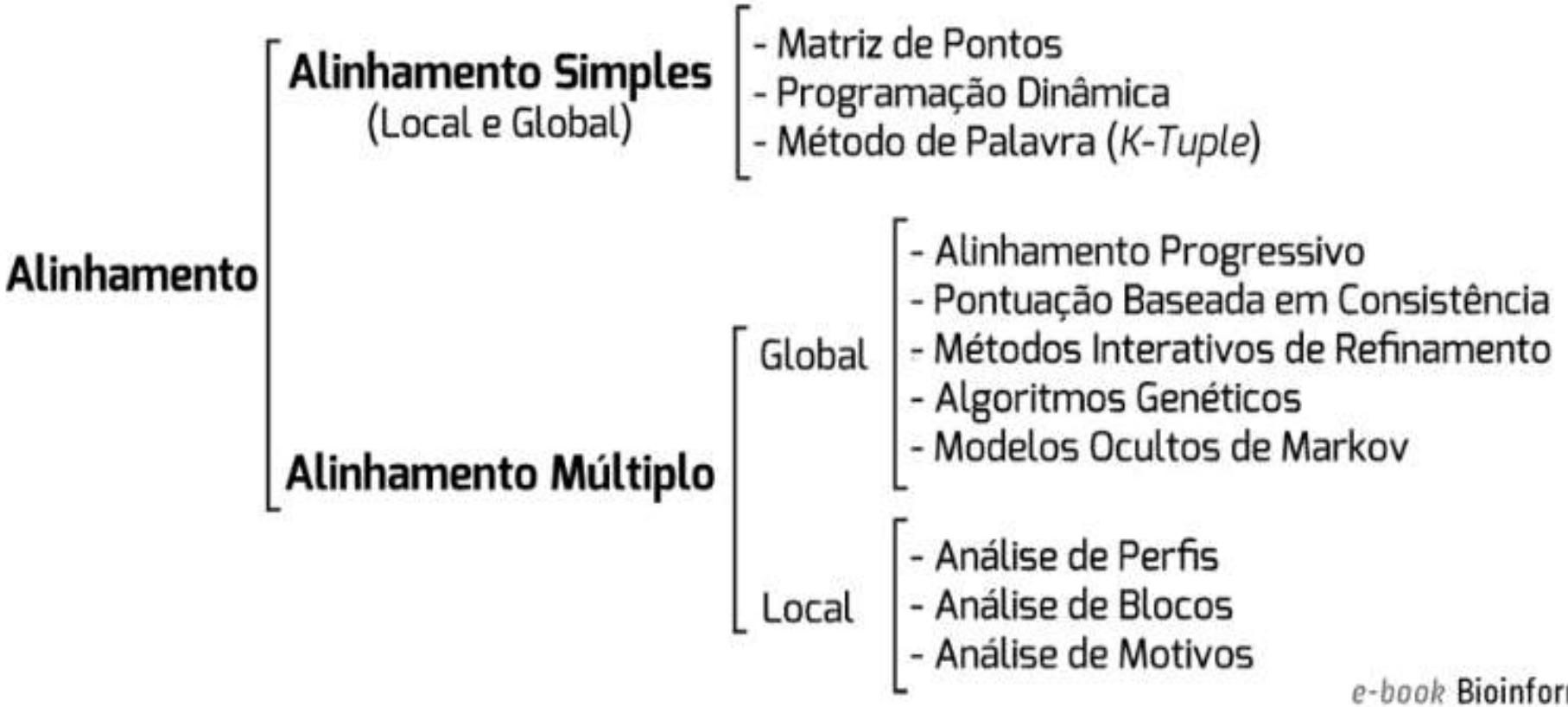
PAM: Point Accepted Mutation

- Calculada de alinhamentos globais;
- Sequências utilizadas com pelo menos 85% de similaridade;
- Usada para traçar origens da Evolução das proteínas.

BLOSUM: BLOcks Substituition Matrix

- Calculada de alinhamentos locais;
- Pode-se selecionar a similaridade entre as sequências;
- Cada matriz é gerada do resultado de uma análise;
- Usada para encontrar domínios conservados

TIPOS DE ALINHAMENTO

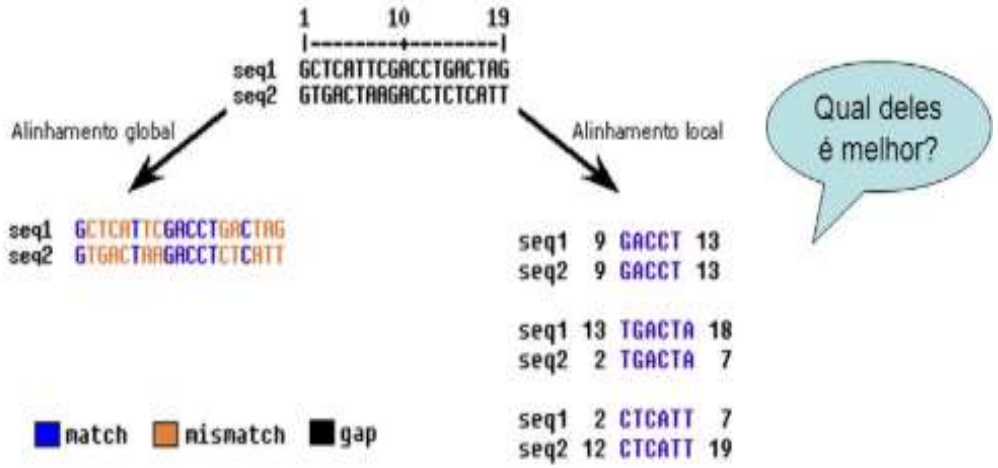


e-book Bioinformática
www.ufrgs.br/bioinfo/

ALINHAMENTO SIMPLES

- **Alinhamento Simples:** descrevem especificamente a relação de similaridade entre duas sequências quaisquer.

  - *Global: similaridade é contada em toda a extensão da sequência;*
  - *Local: buscam pequenas regiões de similaridade*



ALINHAMENTO SIMPLES

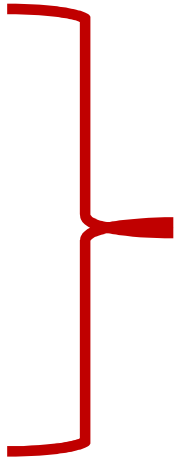
- ALINHAMENTOS ÓTIMO OU HEURÍSTICO

  - *Alinhamento Ótimo: produz o melhor resultado computacional possível;*

  - *Alinhamento Heurístico: produz um resultado o mais próximo possível do resultado ótimo, mas principalmente, produz um resultado de maneira muito veloz.*

ALINHAMENTO SIMPLES

- Como o grau de similaridade pode ser computado no Alinhamento par-a-par:
  - *Programação Dinâmica*
  - *Análise de Matriz de Pontos (Dot matrix)*
  - *Método de palavra (k-tuple)*



ALGORITMOS

PROGRAMAÇÃO DINÂMICA

Algoritmo Needleman e Wunsch

- Método mais utilizado para o Alinhamento de sequências.
  - *Capaz de encontrar o melhor alinhamento para duas sequências através da aplicação da pontuação de similaridades;*
  - *Execução relativamente rápida*
  - *Baseado no princípio de otimização de Bellmann*
    - Solução de problemas complexos através da resolução dos seus diversos subproblemas;
    - Subproblemas são resolvidos e seus resultados armazenados pelo algoritmo.

$$F(i,j) = \max \begin{cases} 1. \text{ Valor da célula na diagonal superior esquerda + pontuação da similaridade;} \\ 2. \text{ Valor da célula acima + valor da penalidade por lacuna;} \\ 3. \text{ Valor da célula à esquerda + valor da penalidade por lacuna.} \end{cases}$$

PROGRAMAÇÃO DINÂMICA

- **Problema:**
  - *Encontre o melhor alinhamento entre pares de GAATC e CATAC.*
  - *Use uma penalidade de intervalo linear de -4.*
  - *Use a seguinte matriz de substituição:*

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

PROGRAMAÇÃO DINÂMICA

- Quantas possibilidades?

GAATC CATAC	GAAT-C C-ATAC	-GAAT-C C-A-TAC
GAATC- CA-TAC	GAAT-C CA-TAC	GA-ATC CATA-C

- Quantos alinhamentos diferentes de duas sequências de comprimento n existem?

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

PROGRAMAÇÃO DINÂMICA

Inicialização

		G	A	A	T	C
	0					
C						
A						
T						
A						
C						

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Introduzindo uma lacuna

G
–

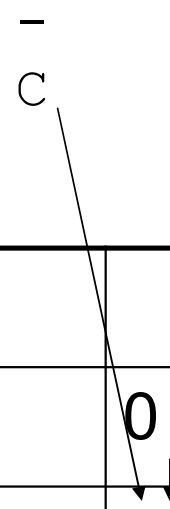
		G	A	A	T	C
	0 →	-4				
C						
A						
T						
A						
C						

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz



		G	A	A	T	C
	0 →	-4				
C	-4					
A						
T						
A						
C						

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

G
C

		G	A	A	T	C
	0	-4				
C	-4	-5				
A						
T						
A						
C						

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

\-----

CATAC

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8					
T	-12					
A	-16					
C	-20					

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8	?				
T	-12					
A	-16					
C	-20					

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

-G	G-	--G
CA	CA	CA-
-4	-9	-12

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8	-4				
T	-12					
A	-16					
C	-20					

0 -4

-4

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8	-4				
T	-12	?				
A	-16	?				
C	-20	?				

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8	-4				
T	-12	-8				
A	-16	-12				
C	-20	-16				

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5	?			
A	-8	-4	?			
T	-12	-8	?			
A	-16	-12	?			
C	-20	-16	?			

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5 →	-9			
A	-8	-4	5			
T	-12	-8	1			
A	-16	-12	2			
C	-20	-16	-2			

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Preenchimento da Matriz

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5 →	-9			
A	-8	-4	5			
T	-12	-8	1			
A	-16	-12	2			
C	-20	-16	-2			?

Encontre o alinhamento ideal e sua pontuação.

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Matriz final

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5 →	-9 →	-13	-12	-6
A	-8	-4	5 →	1 →	-3 →	-7
T	-12	-8	1	0	11 →	7
A	-16	-12	2	11 →	7	6
C	-20	-16	-2	7	11	**17**

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Indel= -4

PROGRAMAÇÃO DINÂMICA

Traceback

GA-ATC
CATA-C

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Diagonal - x_i alinha com y_j
Cima – y_i alinha com espaço
Esquerda - x_j alinha com espaço

PROGRAMAÇÃO DINÂMICA

Traceback

GAAT-C
CA-TAC

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Diagonal - x_i alinha com y_j
Cima – y_i alinha com espaço
Esquerda - x_j alinha com espaço

PROGRAMAÇÃO DINÂMICA

Traceback

GAAT-C

C-ATAC

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Diagonal - x_i alinha com y_j
Cima – y_i alinha com espaço
Esquerda - x_j alinha com espaço

PROGRAMAÇÃO DINÂMICA

Traceback

GAAT-C
-CATAC

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Diagonal - x_i alinha com y_j
Cima – y_i alinha com espaço
Esquerda - x_j alinha com espaço

PROGRAMAÇÃO DINÂMICA

Múltiplas soluções

```
GA-ATC
CATA-C

GAAT-C
CA-TAC

GAAT-C
C-ATAC

GAAT-C
-CATAC
```

- Quando um programa retorna um alinhamento de sequência, pode não ser o melhor alinhamento.

- Uso de **pontuações de distância** para o alinhamento de sequências
  - *Baseia-se em quantas mudanças são necessárias para transformar uma sequência em outra.*
  - *Quanto maior for a distância entre as sequências, maior o tempo evolutivo passado desde que as sequências divergiram de seu ancestral comum.*
  - *Pontuações de distância fornecem um método mais natural biologicamente do que as pontuações de similaridade.*

MATRIZ DE PONTOS ou MATRIZ DOT

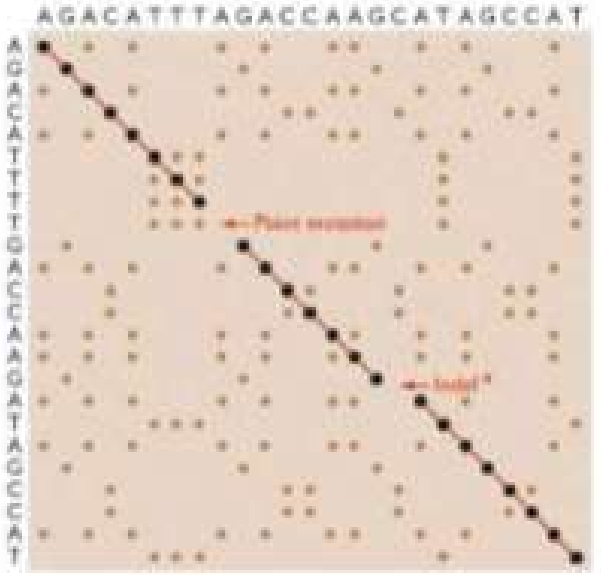
- Visualização gráfica das regiões de similaridade entre sequências – Matriz de Identidade;
- Inspeção visual de um possível alinhamento entre duas sequências;
- Identificação de regiões de pareamento intra-cadeia capazes de formar estruturas $2^{árias}$ em moléculas de RNA.
- Não fornece o alinhamento propriamente dito como resultado final
- Detecção de repetições e inversões;

- **Dotlet**
  - www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
  - Sequências curtas: até 10.000 caracteres
- **Dotter**
  - www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html
  - Sequências até 100.000 caracteres
- **EMBOSS Dottup, Dotmatcher**
  - www.emboss.org
  - Sequências maiores de 100.000 caracteres

MATRIZ DE PONTOS OU MATRIZ DOT

Exemplo

- Sequências:
  - *A) ATGCGTCGTT*
  - *B) ATCCGCGAT*



- Passos:
  - *Organize as sequências em uma matriz;*
  - *Coloque um ponto em cada lugar que houver um match entre duas bases;*
  - *Trechos diagonais (indicados por linhas) são áreas de alinhamento;*
  - *Mais de um alinhamento pode surgir*

Pontos não dispostos na diagonal representam correspondências aleatórias. Não estão relacionadas com a similaridade entre as sequências

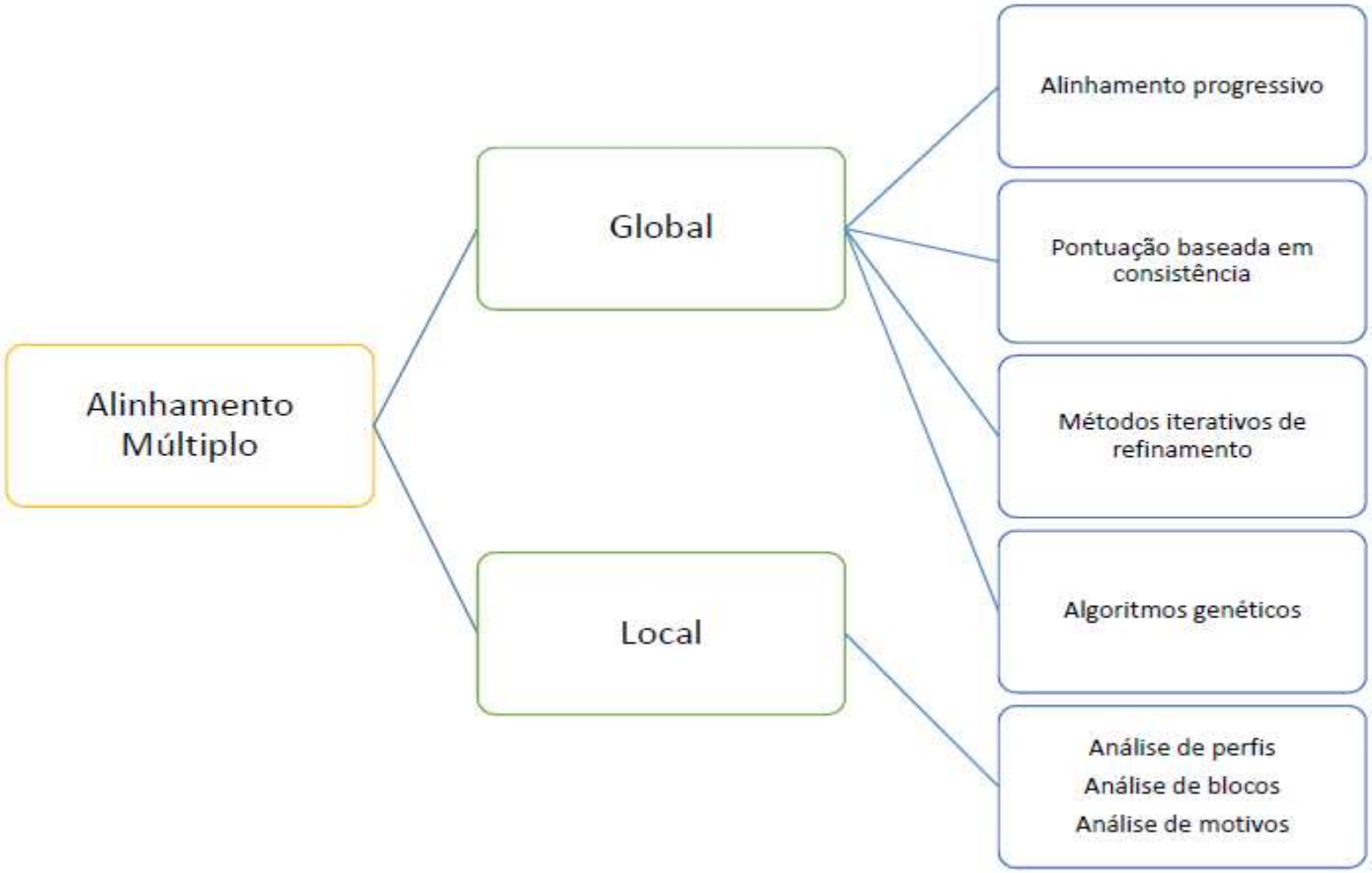
K-TUPLAS ou MÉTODO DE PALAVRAS

- Mais rápido;
- Não garante o melhor alinhamento como resultado;
- Útil em caso onde se busca similaridade de uma única sequência contra um grande conjunto de dados;
- Procuram por pequenas regiões idênticas das sequências e as unem em um alinhamento pelo método de programação dinâmica
- Através de uma matriz de penalidade, o algoritmo calculará o alinhamento que teve maior valor de pontuação.

ALINHAMENTO MÚLTIPLO

- Alinhamento simultâneo de muitas sequências de nucleotídeos ou aminoácidos.
  - *Encontrar padrões diagnósticos para caracterizar as famílias de proteínas;*
  - *Detectar ou demonstrar homologia entre novas sequências e famílias existentes de sequências;*
  - *Ajudar a prever as estruturas secundárias e terciárias de novas sequências;*
  - *Sugerir primers de oligonucleotideos para PCR;*
  - *Análise evolutiva molecular.*

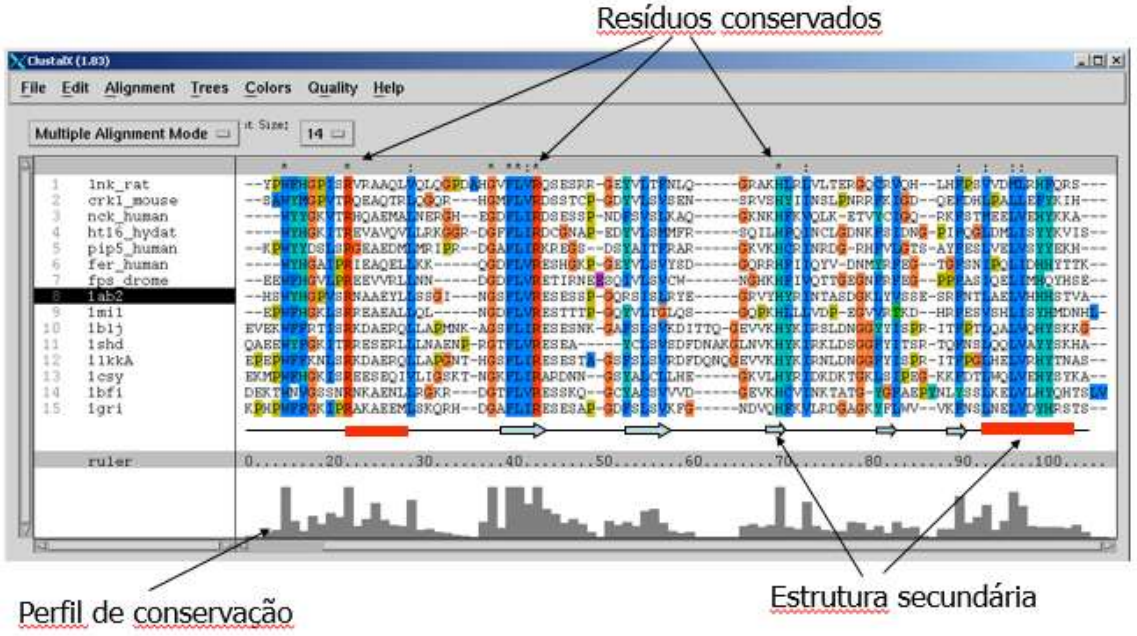
ALINHAMENTO MÚLTIPLO



ALINHAMENTO MÚLTIPLO

ALINHAMENTO PROGRESSIVO

- Leva em consideração a relação evolutiva entre as espécies;
- Utilizam as relações filogenéticas para gerar o resultado de alinhamento;
- É um método rápido e amplamente utilizado para alinhar um grande número de sequências;
- CLUSTALW e CLUSTALX.



ALINHAMENTO MÚLTIPLO
ALINHAMENTO PROGRESSIVO

- O procedimento básico é usar uma série de alinhamentos em pares para alinhar grupos maiores e maiores de sequências, seguindo a ordem de ramificação na árvore guia.

Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

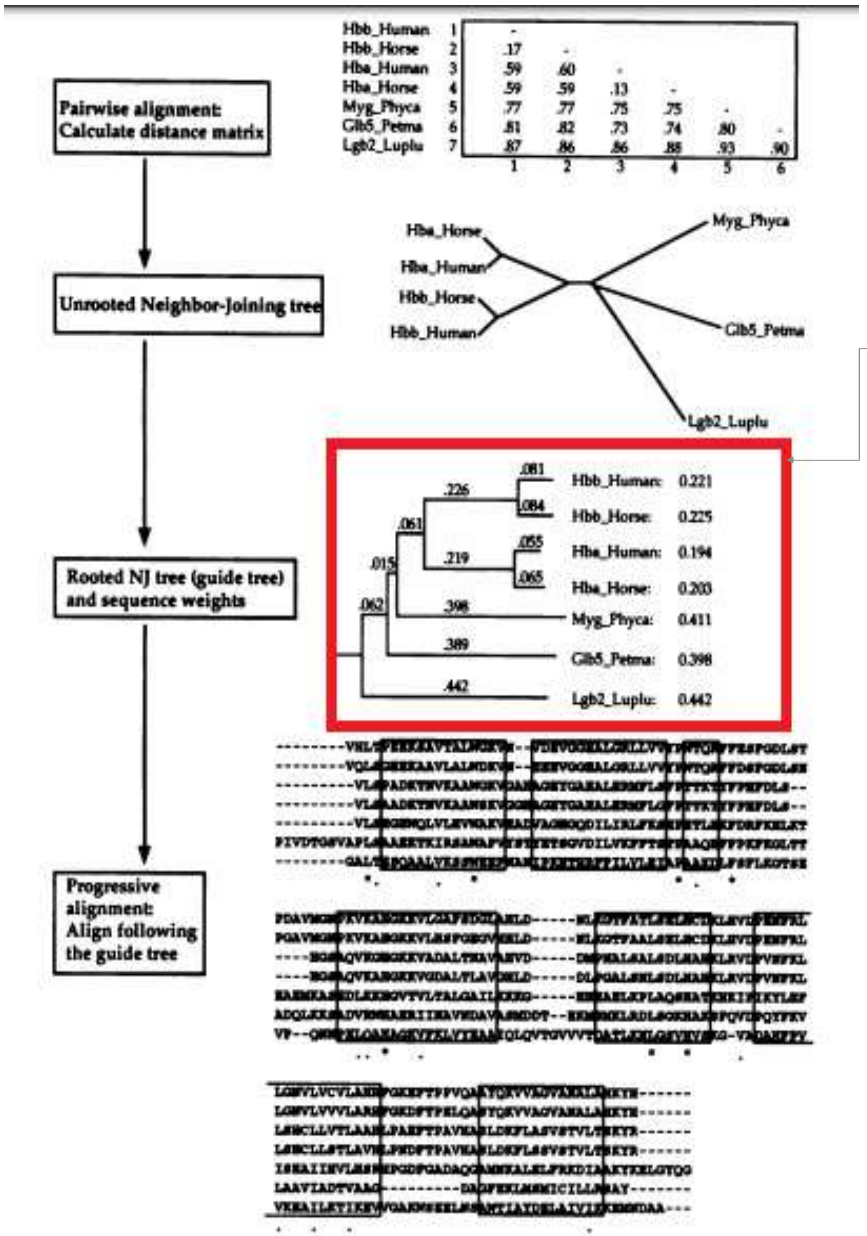
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Received July 12, 1994; Revised and Accepted September 23, 1994

ALINHAMENTO MÚLTIPLO

ALINHAMENTO PROGRESSIVO



O esquema de pontuação compara duas posições de dois alinhamentos. Duas seções de alinhamento com 4 e 2 sequências, respectivamente

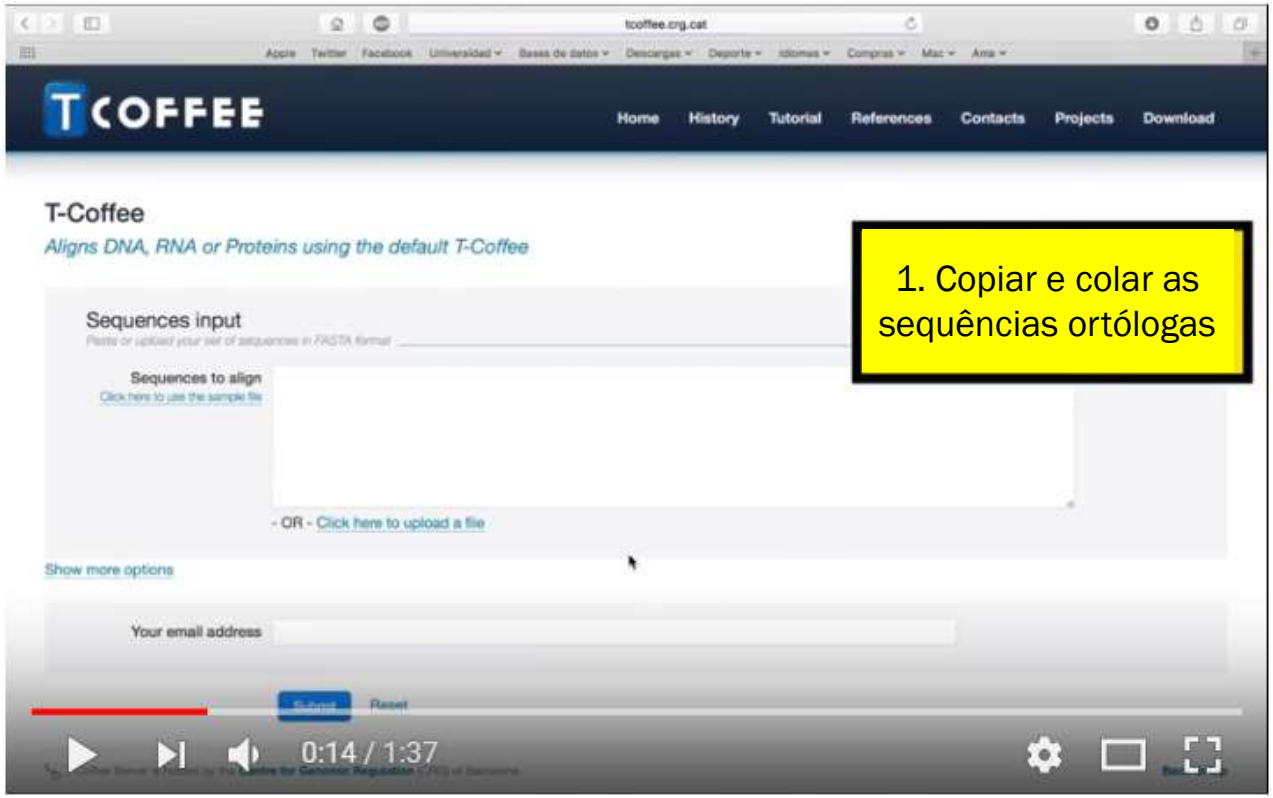
Alinha as sequências na seguinte ordem:

- β-globina humana x cavalo;
- α-globina humana x cavalo;
- As 2 α-globinas x as 2 β-globinas;
- Mioglobina vs. as hemoglobinas;
- Cianohemoglobina x as hemoglobinas + mioglobina
- Leg-hemoglobina x todo o resto.

ALINHAMENTO MÚLTIPLO

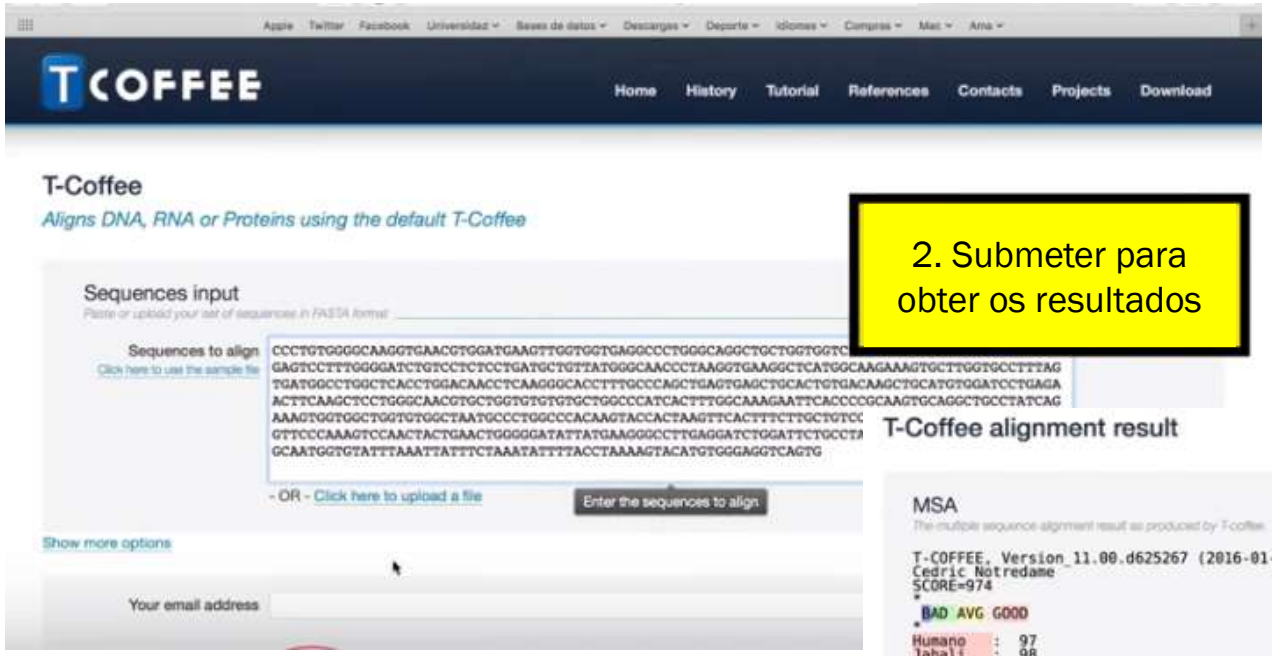
PONTUAÇÃO BASEADA EM CONSISTÊNCIA

- Baseada no algoritmo de alinhamento progressivo;
- não leva em consideração apenas o primeiro par de sequências alinhadas;
- T-coffee



ALINHAMENTO MÚLTIPLO

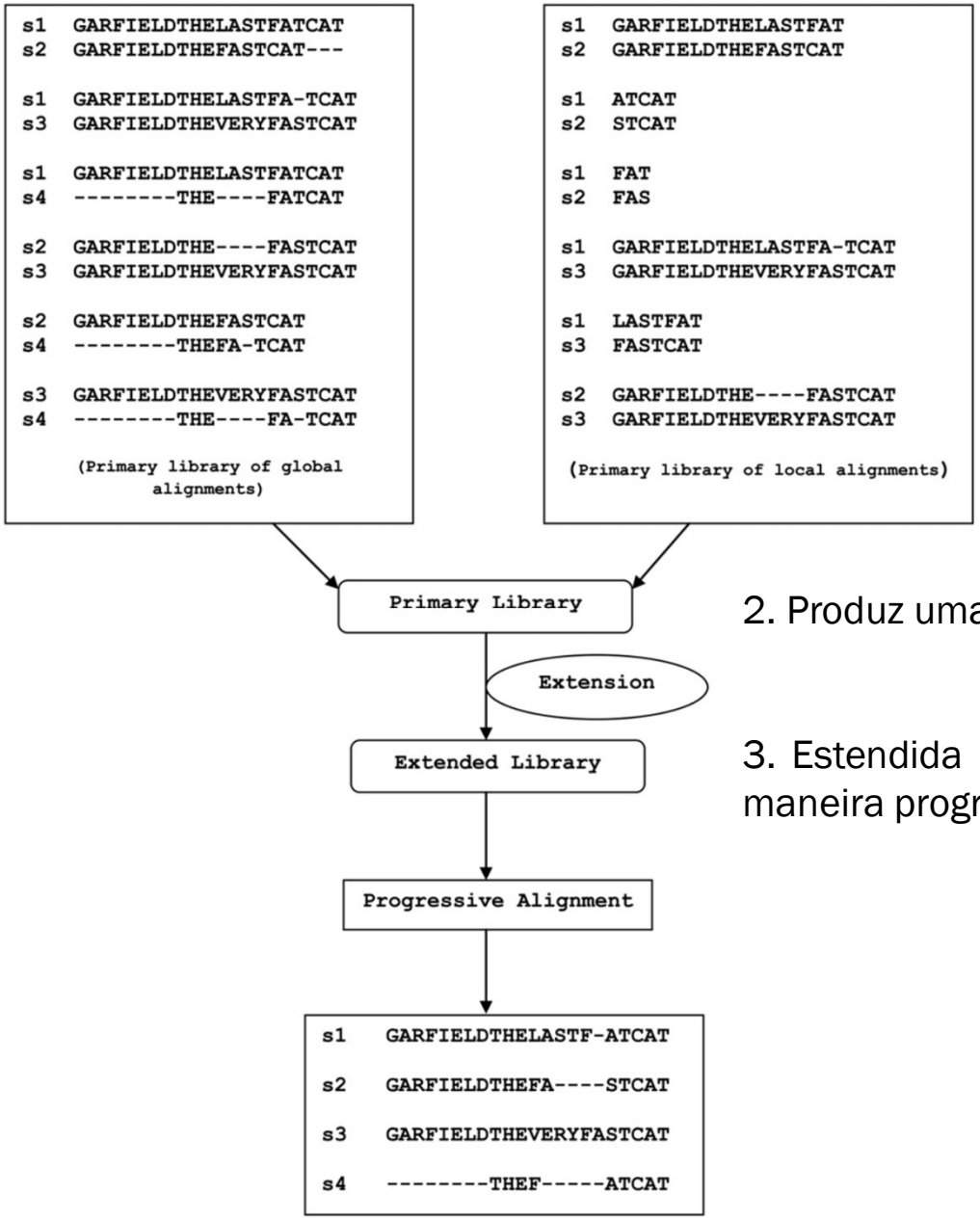
PONTUAÇÃO BASEADA EM CONSISTÊNCIA



2. Submeter para obter os resultados

ALINHAMENTO MÚLTIPLO

PONTUAÇÃO BASEADA EM CONSISTÊNCIA



1. Alinhamentos locais e globais emparelhados são computados

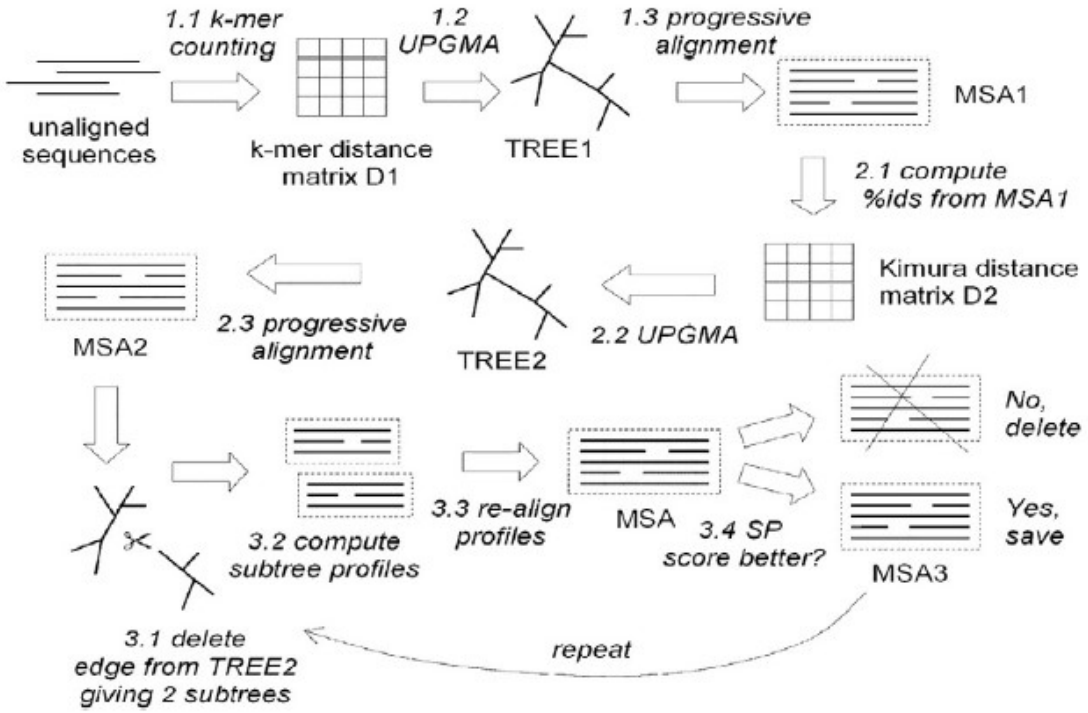
2. Produz uma biblioteca primária

3. Estendida para ser usada para a construção de ASM de maneira progressiva

ALINHAMENTO MÚLTIPLO

MÉTODOS ITERATIVOS DE REFINAMENTO

- Funcionam como os algoritmos do alinhamento progressivo;
- Grupos de sequências são realinhados constantemente ao longo das análises;
- O alinhamento inicial não define o resultado final;
- MUSCLE.



ALINHAMENTO MÚLTIPLO

ALGORITMOS GENÉTICOS

- Buscam simular o processo evolutivo no conjunto de sequências a serem alinhadas;
- Aplicam conceito de seleção e recombinação;
- Método lento – aleatoriedade do processo
- SAGA

MODELOS OCULTOS DE MARKOV

- Baseado em probabilidades estatísticas: eventos de substituição e inserção ou deleção de caracteres.
- Abordagem progressiva que usa uma combinação de modelagem probabilística e técnicas de alinhamento baseadas em consistência.

ALINHAMENTO MÚLTIPLO LOCAL

ANÁLISE DE PERFIS

- Os perfis são encontrados a partir da remoção de regiões altamente conservadas de um grupo de sequências, oriundos de um Alinhamento Múltiplo prévio.
- Uma matriz de pontuação, chamada de perfil, é então feita, e é composta de colunas que podem incluir matches ou mismatches, inserções e deleções.
- Perfil pode ser utilizado para:
  - *Alinhar sequências entre si utilizando as pontuações calculadas para avaliar a probabilidade em cada posição*
  - *Buscar sequências com o mesmo padrão em um banco de dados.*

- Desvantagem:
  - *Perfil produzido é apenas representativo da variação na família de sequências*
  - *Se várias sequências do Alinhamento Múltiplo forem similares, o perfil derivado terá um viés em favor dessas sequências.*

ALINHAMENTO MÚLTIPLO LOCAL

ANÁLISE DE BLOCOS

- Requer, inicialmente, a seleção da região de maior similaridade de um alinhamento múltiplo.
- As regiões alinhadas podem ser encontradas pela busca de seções altamente conservadas num Alinhamento Múltiplo ou pela busca de padrões do mesmo tamanho.
- Esses padrões podem incluir uma região com um ou poucos caracteres que casem (match), seguida de uma curta região separadora de caracteres que não casem (mismatch) e assim por diante (até que as sequências comecem a ficar diferentes).

ANÁLISE DE MOTIVOS

- Identificar motivos em sequências proteicas;
- Utiliza grupos de substituição de aminoácidos característicos de cada coluna de todos os alinhamentos dos bancos de dados BLOCKS e HSSP.
- A probabilidade de cada motivo é estimada a partir das frequências dos aminoácidos individuais no banco de dados SwissProt, sendo igual a o produto das somas em cada coluna

Algoritmo capaz de realizar buscas baseadas em alinhamentos

BLAST

- Basic Local Alignment Search Tool – BLAST, encontra regiões de similaridade local entre sequências;
- Compara sequências de nucleotídeos ou proteínas com sequências do banco de dados e calcula a significância estatística das similaridades;
- É um algoritmo de alinhamento **simples**, **heurístico** e **local;**
- Pode ser usado para inferir relações funcionais e evolutivas entre as sequências, bem como para ajudar a identificar membros de famílias de genes

BLAST
Definições

- **Query:** sequência submetida pelo usuário;
- **Subject:** sequência do banco de dados similar à query;
- **E-value:** controla o quanto a sequência do banco de dados deve ser similar à query para ser listada;
- **Score:** medida da perfeição do alinhamento encontrado.

SUB-PROGRAMAS BLAST

Formato da Seqüência de Entrada	Banco de dados	Formato da seqüência que é comparado	Programa BLAST adequado
Nucleotídeos	Nucleotídeos	Nucleotídeos	BLAST**n**
Proteínas	Proteínas	Proteínas	BLAST**p**
Nucleotídeos	Proteínas	Proteínas	BLAST**x**
Proteínas	Nucleotídeos	Proteínas	**T**BLAST**n**
Nucleotídeos	Nucleotídeos	Proteínas	**T**BLAST**tx**

BLAST

E-value

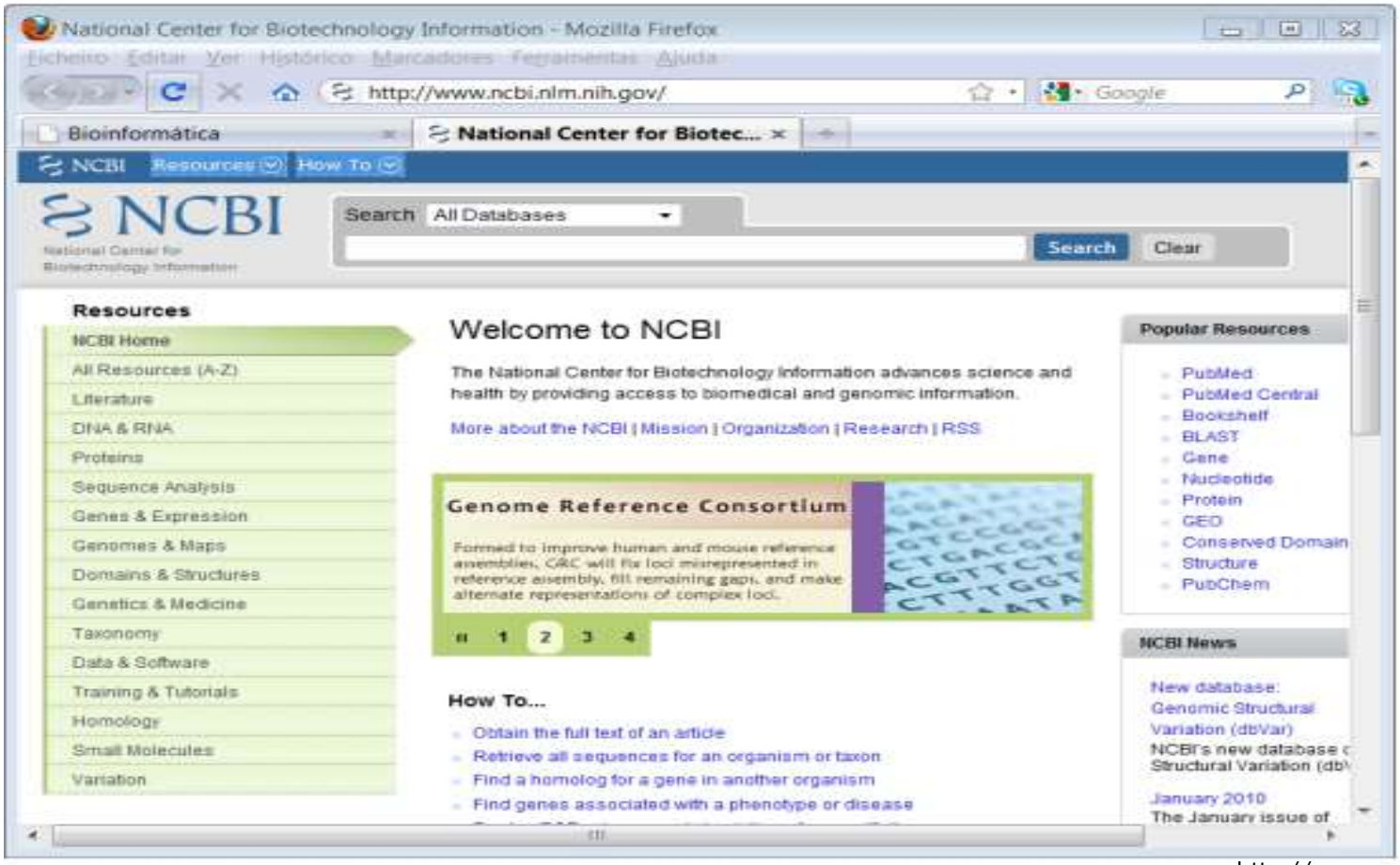
- **Significado** – determina a probabilidade do alinhamento ser devido ao acaso;

- **Observar** – quanto menor seu valor, melhor
  - *Há menor chance do alinhamento ter ocorrido ao acaso, ou seja, as sequências são homólogas.*

- **Depende** do comprimento da sequência query e do comprimento do banco de dados.

BLAST

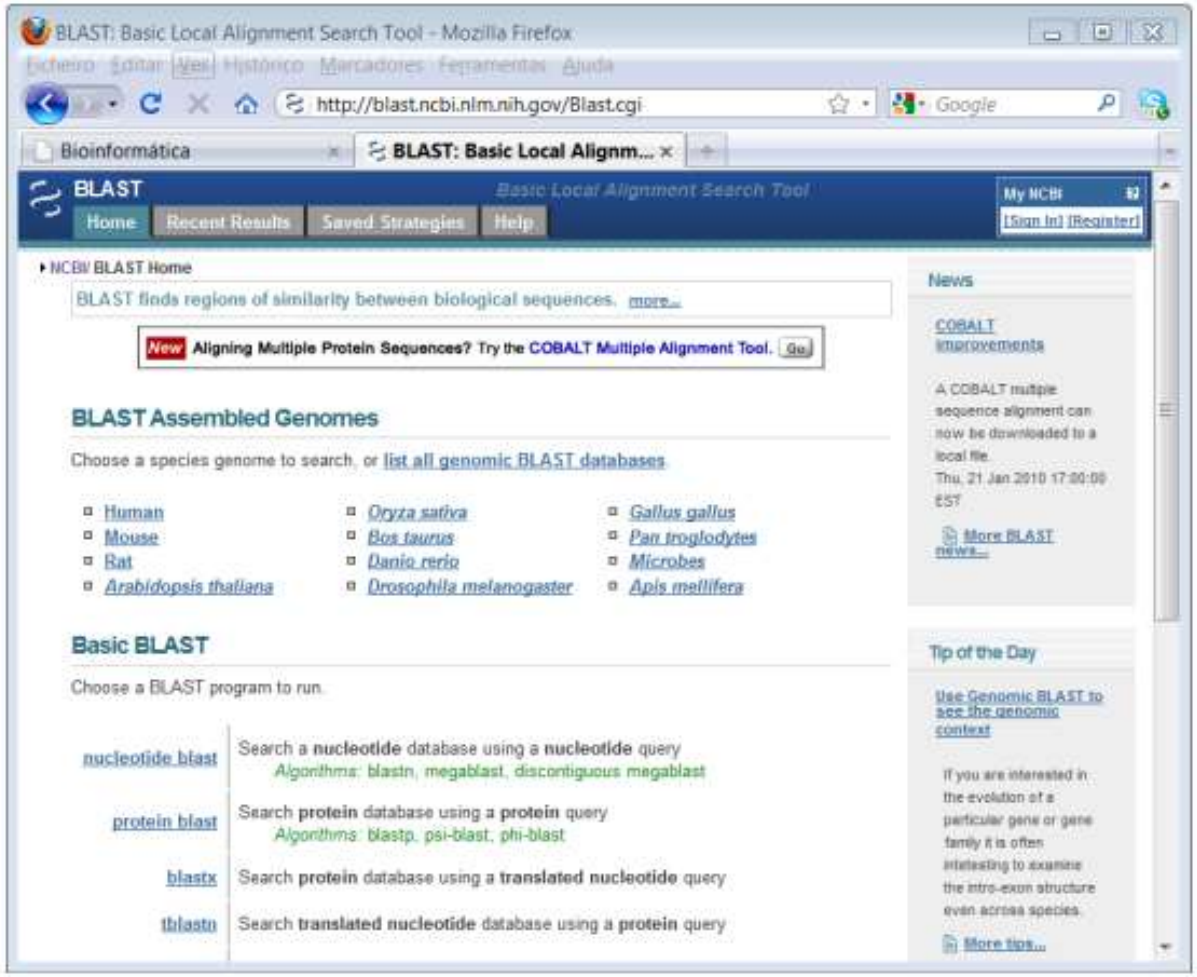
1. Abrir o site da NCBI, http://www.ncbi.nlm.nih.gov.



http://www.esa.ipb.pt/medchem/align.html

BLAST

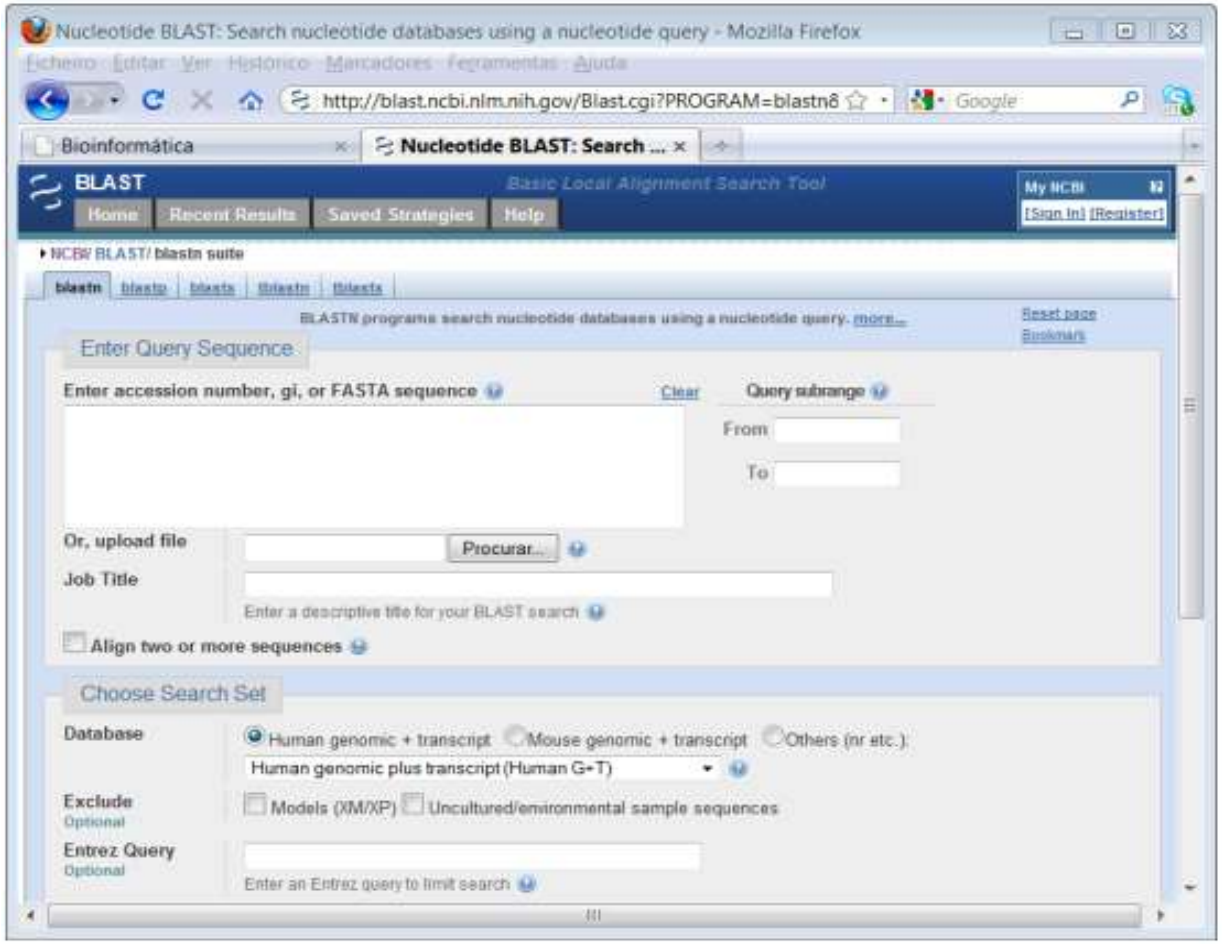
2. Clicar em BLAST na coluna do lado direito



http://www.esa.ipb.pt/medchem/align.html

BLAST

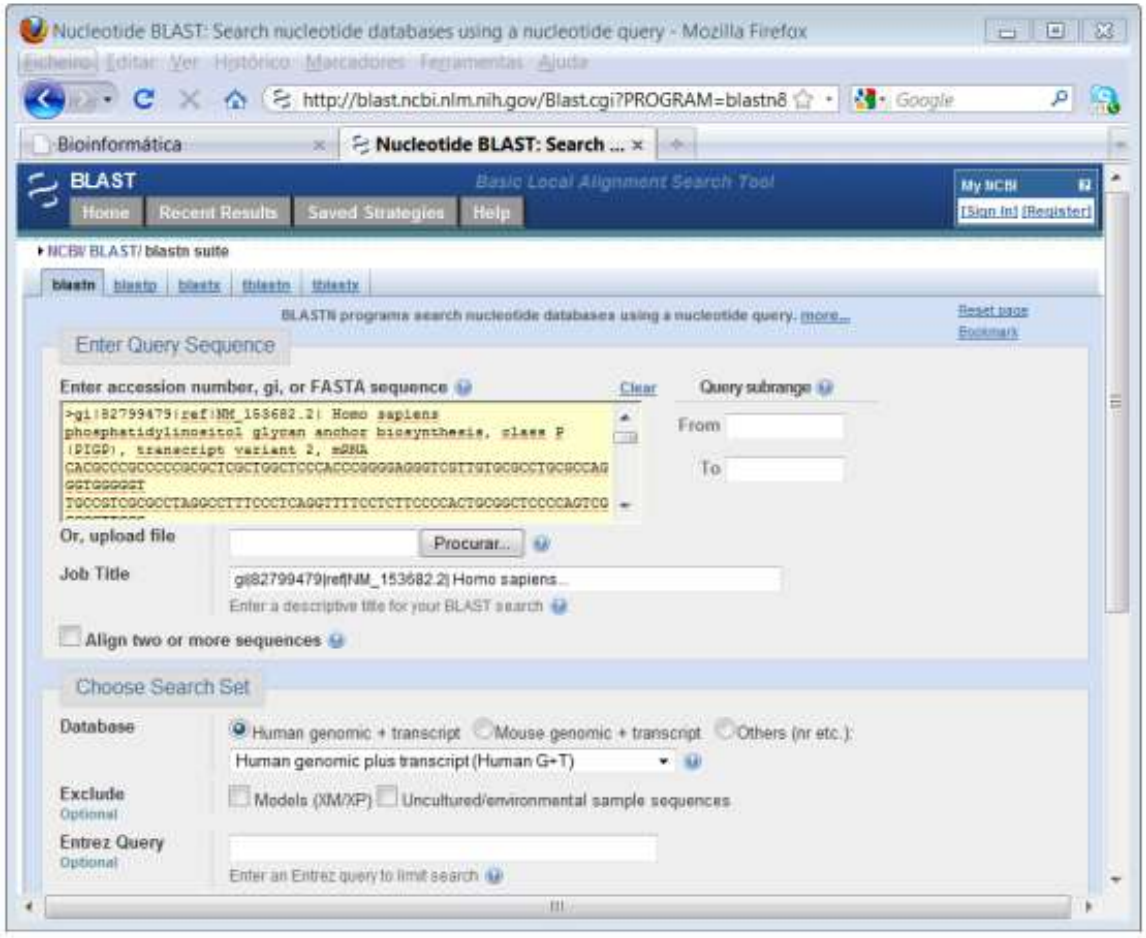
3. Clicar em Nucleotide BLAST (blastn)



http://www.esa.ipb.pt/medchem/align.html

BLAST

4. Faça "copy“ da sequência. No campo onde diz "Enter Query Sequence" faça "paste" da sequência.



A sequência apresenta o cabeçalho característico do formato FASTA

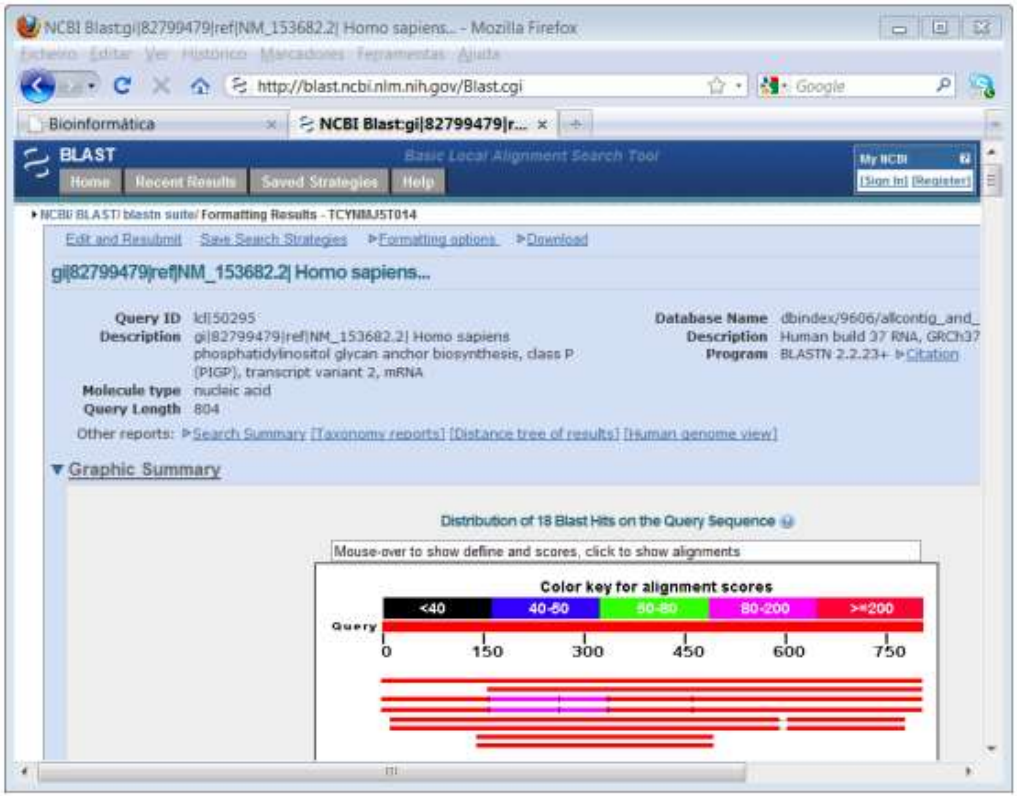
```
>Unknown sequence #1
ACTACCGCTATCAATATACTCCCACAAATATCAAGAGCCTTCCCAGTATTAAATTTGCTA
AATTCAATACGAACTTCACACTCCACAGCCTCACGCGAAATTAATAATACGTATTTAAAT
ATACCATGAACTATCGTTTAGTACATGAATTTACACACGTCAGCCCGATCAAATGTTTAT
CATTATATATGTACATTTCAGTTTGTGTATATAGACATAACATTAATGTAATAAAGACAT
TAGTACATTAATTGATTGTCCTCAAGCATATAAGCAAGTACTAGACATTCACTAGCGGTA
```

http://www.esa.ipb.pt/medchem/align.html

BLAST

5. Não alterar nenhum parâmetro e clicar o botão "BLAST".

6. Após alguns segundos aparece a representação esquemática dos resultados significativos.



A sequência de nucleotídeos do gene do exemplo possui:

- Nucleotídeos estão representados na primeira barra vermelha com a indicação do número de nucleótidos.
- As barras apresentadas abaixo representam esquematicamente as sequências de nucleotídeos que o programa pesquisou que apresentam um grau de homologia mais significativo (da mais significativa para a menos significativa).

http://www.esa.ipb.pt/medchem/align.html

BLAST

7. Logo abaixo da representação esquemática das sequências com homologia significativa, aparece a lista de todos os resultados com a identificação da sequência ("Sequence identifier") e os valores de "Score", de "E-value", e a porcentagem de identidade máxima ("Max Ident") para cada sequência.

— O **"Score"** fornece informação sobre o grau de homologia entre a sequência em questão e a sequência introduzida. Quanto maior o valor, melhor será o grau de homologia.

— O **valor "E"** ("Expected") é um indicador do grau de significância do resultado da pesquisa. Quanto menor o valor de "E" mais significativo é o resultado. O valor de "E" a partir do qual o resultado é significativo varia com o tipo de pesquisa efetuada e os respectivos parâmetros utilizados. No entanto, para uma pesquisa inicial, o valor 0,01 é um bom ponto de partida para aceitar ou rejeitar um resultado como sendo significativo.

http://www.esa.ipb.pt/medchem/align.html

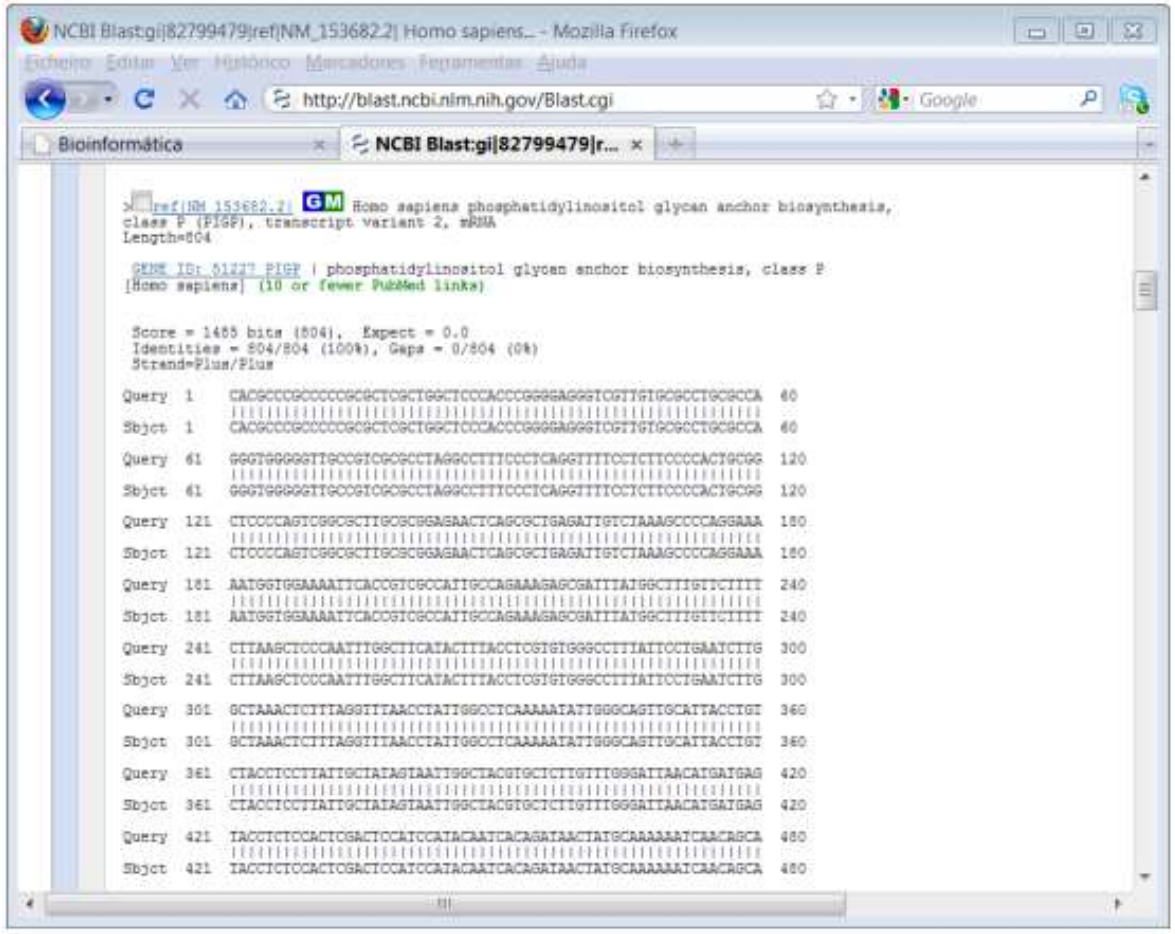
Interpretação do Valor Esperado: Evalue

- **$E<10^{-100}$ ⇒** valor muito baixo. Genes homólogos ou idênticos.
- **$E<10^{-3}$ ⇒** valor moderado. Genes podem estar relacionados.
- **$E>1$ ⇒** valor alto. Prováveis genes sem relação.
- **$0,5 < E < 1$ ⇒** Região duvidosa - "Twilight zone"

Twilight zone: nessa região, nada é garantido sobre o significado das similaridades observadas. Homologia ou não, nunca é garantida nessa área.

BLAST

8. Resultado completo do alinhamento para cada um dos resultados.



— Indicação do "Score".

— Como o alinhamento é de 100 %, o "E" tem o valor mínimo que é 0 indicando um grau de significância máximo.

— Como a sequência foi retirada da base de nucleótidos da NCBI o mais provável é que, no BLAST realizado, o primeiro resultado seja a própria sequência.

http://www.esa.ipb.pt/medchem/align.html

CONSIDERAÇÕES FINAIS

- Em sequências biomoleculares, alta similaridade de sequência frequentemente implica em:
  - *Similaridade funcional e/ou estrutural*
  - *Relação evolutiva*

- O Alinhamento fornece subsídios para a inferência, na qual é feita de maneira razoavelmente subjetiva pelo pesquisador
  - *Várias variáveis influenciam o Alinhamento e podem dar resultados diferentes*

- É preciso compreender os programas para saber o que se está analisando.

REFERÊNCIAS

- CHOWDHURY, B.; GARAI, G. A review on multiple sequence alignment from the perspective of genetic algorithm. **Genomics**, v. 109, p. 419–431, 2017.
- EDDY, S. R. What is dynamic programming? **Nature Biotechnology**, v. 22, n. 7, July 2004.
- MIR, L. (org). Genômica. São Paulo: **Ed. Atheneu**, 2004.
- VERLI, H. (org). Bioinformática da Biologia à flexibilidade Alegre, Brasil, v. 1, 2014.
- ZIELEZINSKI, A.; VINGA, S.; ALMEIDA, J.; KARLOWSKI, W. M. Alignment-free sequence comparison: benefits, applications, and tools. **Genome Biology**, v. 18, p. 186, 2017.