Alinhamento de Seqüências Biológicas

Profª Drª Silvana Giuliatti
Departamento de Genética – FMRP/USP
silvana@rge.fmrp.usp.br

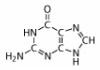
O que se compara?

- A comparação de seqüências de **DNA**, RNA e proteínas é uma das bases da bioinformática.


Citosina


Timina

A C G T
nucleotídeos


Guanina


Adenina

O que se compara?

- A comparação de seqüências de DNA, **RNA** e proteínas é uma das bases da bioinformática.


Citosina


Uracila

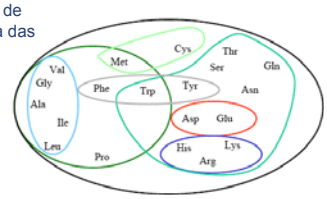
A C G U


Guanina


Adenina

O que se compara?

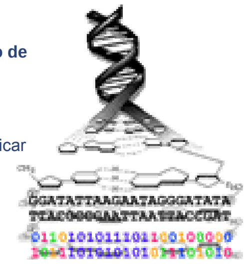
- A comparação de seqüências de DNA, RNA e **proteínas** é uma das bases da bioinformática.

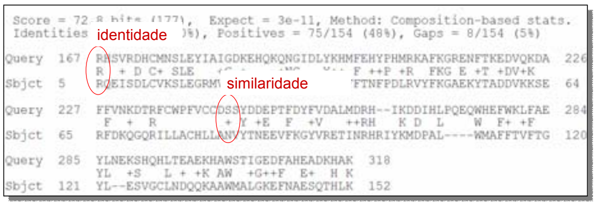


G A S T C V I L P F Y M W N Q H D E K R
Aminoácidos

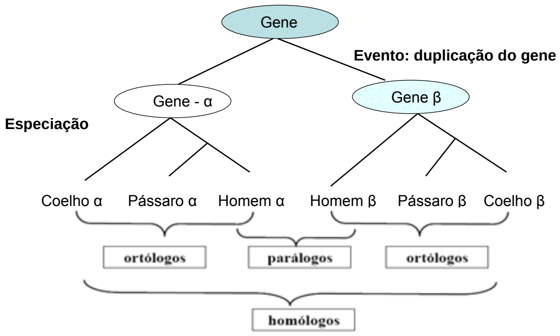
Comparação de Seqüências

- É necessário realizar um **alinhamento de sequências**.
- **Questão:** como alinhar e como classificar esse alinhamento?





Homologia



Alinhamentos de Seqüências

- **Alinhamentos podem ser entre:**
  - Pares de seqüências
  - Múltiplas seqüências
- **Global**: similaridade é considerada ao longo de toda as seqüências

```
SSGYTCA--SKMGLG
| |    |    |    | |
SSKQTGKGSSR- ILG
```

- **Local:** similaridade é considerada apenas nas melhores regiões

```
------SSGYT------
| |   | |
------SSKYT------
```

Histórico

1970
- **Matrizes de Pontos**
- **Programação Dinâmica**
  - **Alinhamento Global**

1981
- **Programação Dinâmica**
  - **Alinhamento local**

1988
- **Alinhamento com Banco de dados**
  - **FASTA**

1990
  - **BLAST**

1994
- **Alinhamento múltiplo**
  - **CLUSTAL**

Matrizes de Pontos

- Descrito pela primeira vez por Gibbs e McIntyre (1970).
- Método usado para alinhar duas seqüências
- Todas as regiões possíveis de serem alinhadas são encontradas
- Não permite a inclusão de gaps

Matrizes de Pontos

- Colocar uma seqüência em uma linha e a outra seqüência em uma coluna
- Colocar um ponto em todas as posições onde houver similaridade
- Diagonais revelam a similaridade entre as duas seqüências

Matrizes de Pontos

	A	C	G	T	A	C	G	T	A	C
G			o				o			
G			o				o			
T				o				o		
T				o				o		
A	o				o				o	
C		o				o				o
G			o				o			
G			o				o			
T				o				o		
C		o				o				o

Programas para Matrizes de Pontos

- **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

Programação Dinâmica

- **"Problema do caixeiro viajante"**
- **Procura por todas as soluções possíveis**
  - **Encontra a solução ótima**

Alinhamento com Programação Dinâmica

- **Alinhamento de pares de seqüências**
  - **Global e Local**
  - **Pode considerar lacunas (gaps) ao longo do alinhamento**
- **Encontrar o melhor alinhamento possível – alinhamento ótimo**
  - **Pode existir mais de um alinhamento ótimo**
- **Limitação:** pode se tornar lento dependendo do tamanho das seqüências

Score

- **Considerar as seqüências ACGGACT e ATCGGATCT**

```
A – C G G A – C T
|   | | | |   | |
A T C G G A T C T
```

```
A – C –  G G – A CT
|   |    |       | |
A T C G G A T – C T
```

Qual destes é o melhor alinhamento?

Score (Pontuação)

- **Score ou Pontuação: Medida pela qual os alinhamentos são quantificados**
- **Considere o seguinte esquema simples de pontuação**

+1 para igualdade	(match)
-1 para desigualdade	(mismatch)
-2 para lacunas	(gap)

Score (Pontuação)

Qual é o melhor alinhamento?

Alinhamento 1

```
A – C G G A – C T
|   | | | |   | |
A T C G G A T C T
```

+1 -2 +1 +1 +1 +1 -2 +1 +1 = +2

Alinhamento 2

```
A – C –  G G – A C T
|   |    |       | |
A T C G G A T – C T
```

+1 -2 +1 -2 +1 -1 -2 -2 +1+1 = - 4

Melhor Alinhamento: Alinhamento 1

Programação Dinâmica

- **Alinhamento Global** - Algoritmo de Needleman-Wunsch (1970).
- **Alinhamento Local** – Smith-Waterman (1981)
  - Modificação do Algoritmo de Needleman-Wunsch
- Gerar uma matriz
- Encontrar todos os possíveis alinhamentos

Programação Dinâmica

Alinhamento Global

		0	1	2	3	4
		-	G	T	A	A
0	-	0	-2	-4	-6	-8
1	C	-2	-1	-3	-5	-7
2	T	-4	-3	0	-2	-4
3	T	-6	-5	-2	-1	-3
4	A	-8	-7	-4	-1	0
5	G	-10	-7	-6	-3	-2
6	A	-12	-9	-8	-5	-2

```
- G T A - A
    | |   |
C T T A G A
```

-2 -1 +1 +1 -2 +1 = **-2**

Programação Dinâmica

Alinhamento Local

		0	1	2	3	4
		-	G	T	A	A
0	-	0	0	0	0	0
1	C	0	0	0	0	0
2	T	0	0	1	0	0
3	T	0	0	1	0	0
4	A	0	0	0	2	1
5	G	0	1	0	0	1
6	A	0	0	0	1	1

Matrizes de Substituição

- Sistema de pontuação biologicamente relevantes
- Para produzir alinhamentos biologicamente significativos
- Matrizes
  - PAM
  - BLOSUM
  - Aminoácidos
  - Nucleotídeos

Matrizes de Substituição

Matrizes de Substituição PAM

- PAM – Percent Accept Mutation
- Desenvolvida por Margaret Dayhoff *et al* (1978)
- Considerou seqüências de aminoácidos com pelo menos 85% de similaridade
- As substituições de aminoácidos foram estimados
  - 1572 mudanças em 71 grupos de seqüências de proteínas.
- Matriz mais utilizada PAM 250

Matrizes de Substituição

Matriz PAM 250

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

Matrizes de Substituição

Matriz PAM

- **Valor Zero →** freqüência de substituição entre dois aminoácidos é esperada ao acaso
- **Valor menor que zero →** freqüência é menor que a esperada. Substituição de dois aminoácidos ao acaso.
- **Valor maior que Zero →** freqüência maior que a esperada. Substituição não é ao acaso. Indica maior probabilidade de relação com ancestral

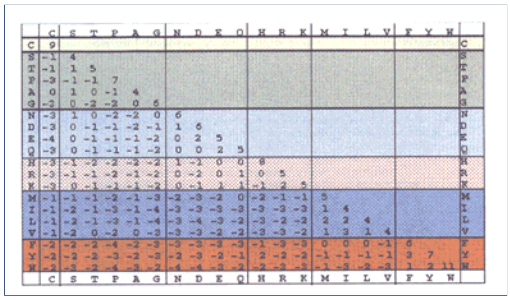
Matrizes de Substituição

Matriz de Substituição BLOSUM

- **BLOSUM – Blocks Substitution Matrix**
- **Desenvolvidas por Henikoff e Henikoff, (1992)**
- **Aminoácidos são organizados em blocos**
- **Utilizou-se 500 famílias de proteínas**
- **Matriz mais utilizada BLOSUM62**

Matrizes de Substituição

Matriz BLOSUM62



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

Matrizes de Substituição

Matriz BLOSUM

- **Valor Zero →** probabilidade de substituição entre dois aminoácidos iguais
- **Valor menor que zero →** maior probabilidade de substituição de dois aminoácidos ser ao acaso
- **Valor maior que Zero →** maior probabilidade de substituição entre dois aminoácidos não ser por acaso. Indica maior probabilidade de relação com ancestral

Matrizes de Substituição

PAM	X	BLOSUM
• Calculadas de alinhamentos globais		• Calculadas de alinhamentos locais
• Seqüências utilizadas com pelo menos 85% de similaridade		• Pode-se selecionar a similaridade entre as seqüências
• As matrizes são extrapolações da PAM 1		• Cada matriz é gerada do resultado de uma análise
• Usada para traçar origens da Evolução das proteínas		• Usadas para encontrar domínios conservados

Métodos de Palavras

- Alinham sequências mais rapidamente.
- Procuram por partes curtas idênticas (palavras ou k-tuplas).
- Pesquisas em bancos de dados: FASTA e BLAST
- Seguem um método heurístico.

FASTA

- Desenvolvido por Pearson e Lipman (1988).
- Uma sequência de proteína ou DNA com todas as sequências num **banco de dados**.
- Apresenta os alinhamentos locais da sequência analisada com as sequências do banco.

FASTA

- Algoritmo escrito em linguagem C.
- Mais lento que BLAST.
- Procura por um número k de consecutivas letras (aminoácidos ou nucleotídeos): palavras ou k-tuplas.

FASTA

A sequência de entrada deve estar no formato FASTA.

O algoritmo pode ser dividido em 4 etapas:

a) seleção das 10 melhores regiões.

b) re-classificação das 10 melhores regiões.

c) seleção das seqüências mais semelhantes.

d) alinhamento das seqüências selecionadas.

Métodos de Palavras
BLAST
Basic Local Aligment Sequence Tool

- Alinhamento de uma sequência de proteína ou DNA com todas as sequências num banco de dados.
- Apresenta os alinhamentos locais da sequência analisada com as sequências do banco.
- Mais rápido que FASTA.
- Algoritmo escrito em linguagem C.
- Procura por um número k de consecutivas letras (aminoácidos ou nucletídeos): palavras ou k-tuplas.

BLAST

- Procura por palavras que são mais significantes
- A significância é incorporada ao algoritmo através de matrizes de pontuação.
- Buscar por identidades de comprimento k:
  - **11 para nucleotídeos**
  - **3 para aminoácidos**

BLAST

- O algoritmo pode ser dividido em 4 etapas:

a) montagem da lista de palavras.

b) procura pelas palavras em cada sequência do banco.

c) extensão.

d) alinhamento das seqüências.

BLAST

d) Alinhamento das seqüências

Determina se cada HSP é estatisticamente significante.

Depois de determinar se HSP é estatisticamente significante, faz o alinhamento dos melhores segmentos.

EVALUE

O mais usado score é o **Evalue:** proporciona uma estimativa do número de falsos positivos esperados.

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.

Alinhamento Múltiplo

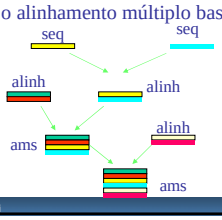
- O alinhamento múltiplo de sequências identifica resíduos ou regiões conservadas ou equivalentes em estruturas.

CLUSTALW

- Método heurístico
- Rápido e eficiente.
- Faz alinhamento progressivo dos perfis e sequências mais distantes
- O mais usado: ClustalW (Thompson et al, 1994)

CLUSTALW

- **Algoritmo de 3 etapas:**
  - Alinhamento em pares de todas as sequências para determinar similaridade entre elas.
  - Definir a ordem do alinhamento progressivo baseado na similaridade.
  - Construir o alinhamento múltiplo baseado na ordem definida.



Qual a ordem do alinhamento?

CLUSTALW

Etapa 1: Alinhamento em pares de todas sequências para determinar similaridade entre elas.

- Usa método de alinhamento (global) de pares de sequências
- Usa matriz de substituição e penalidade por gaps.

CLUSTALW

- Usa os alinhamentos em pares para calcular uma "distância genética" entre todos os pares de sequências.
- Constrói uma matriz de valores de distâncias.

CLUSTALW

Etapa 2: Definir a ordem do alinhamento progressivo baseado na similaridade.

- Definir as sequências mais próximas: árvore de similaridade.
  - Usa matriz de distâncias para calcular a árvore.
  - Método de junção por vizinhos (neighbor-joining)

CLUSTALW

Etapa 3: Construir o alinhamento múltiplo baseado na ordem definida.

- Combinar os alinhamentos começando com os grupos mais próximos para os mais distantes

Referências

- http://www.sxc.hu
- http://creative.gettyimages.com/source/home/home.aspx
- Mount, D. W., Bioinformatics - Sequence and Genome Analysis, ed. CSHL, 2ª edição.
- Gibas, C e Jambeck, P., Desenvolvendo a Bioinformática. Ed. Campus.