Algoritmo BLAST

Aluno: Marllus de Melo Lustosa

Sumário:

1. Alinhamento de sequências

- Encontrar um grau de similaridade entre sequências de nucleotídeos ou proteínas.
  - Definir sequências homólogas;
  - Definir fragmentos similares entre sequências;
  - Determinar características entre sequências;

1.1. Alinhamento global

- Sequências são alinhadas de ponta a ponta;
- Pode incluir grandes pedaços com baixa similaridade;
- Útil para comparar sequências cujas semelhanças sejam esperadas em toda a sua extensão;

```
LGPSSKQTGKGS-SRIWDN
|     |   |||       |
LN-ITKSAGKGAIMRLGDA
```

Fig1. Exemplo de alinhamento global entre duas sequências. [1]

1.2. Alinhamento local

- São alinhados um ou mais segmentos com alta similaridade entre as sequências;
- Útil quando não se tem nenhum conhecimento sobre a semelhança entre as sequências a comparar;

```
                    tccCAGTTATGTCAGgggacacgagcatgcagagac
                       ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Fig2. Exemplo de alinhamento local entre duas sequências. [2]

1.3. Global x Local

Global: As sequências são alinhadas de ponta a ponta;
Local: Pedaços das sequências é que são comparados;

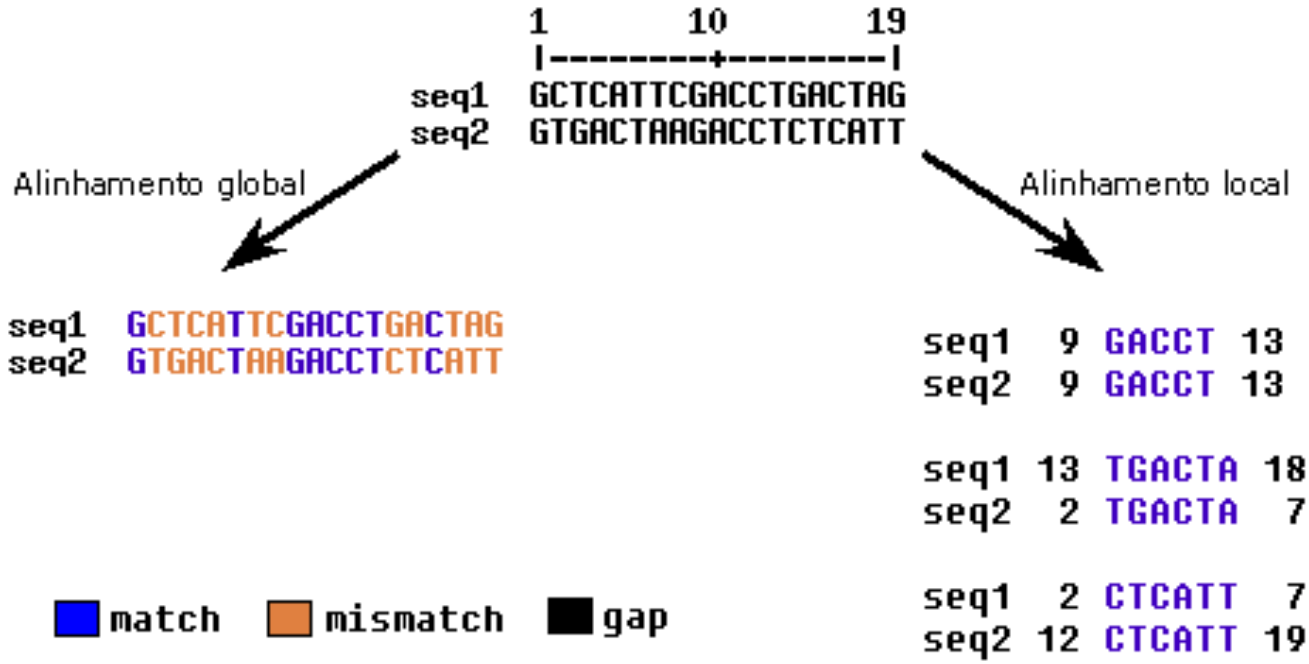


Fig3. Exemplo de alinhamento global e local entre duas sequências. [3]

1.4. Alinhamento ótimo x heurístico

heurística -- do dicionário *Houaiss*

Acepções
¦ substantivo feminino
1 arte de inventar, de fazer descobertas; ciência que tem por objeto a descoberta dos fatos
1.1 Rubrica: história.
ramo da História voltado à pesquisa de fontes e documentos
1.2 Rubrica: informática.
método de investigação baseado na aproximação progressiva de um dado problema
1.3 Rubrica: pedagogia.
método educacional que consiste em fazer descobrir pelo aluno o que se lhe quer ensinar

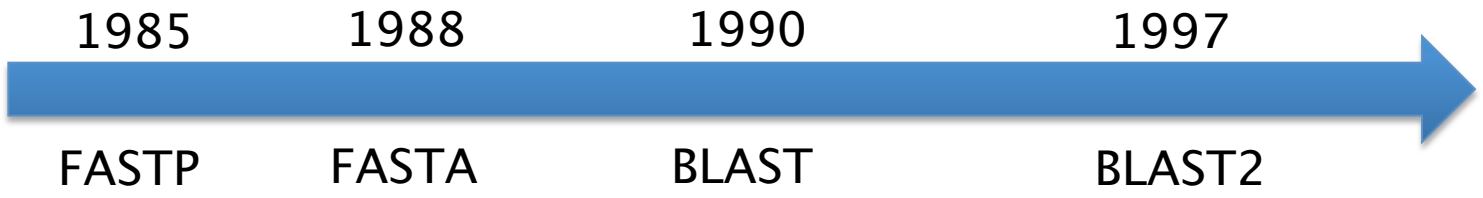
- Alinhamento **ótimo**: produz o melhor resultado computacionalmente possível;
- Alinhamento **heurístico**: produz um resultado o mais próximo possível do resultado ótimo, mas, principalmente, produz um resultado de maneira muito veloz;

1.5. Ferramentas de alinhamento

Programa	Tipo de Alinhamento	Precisão do Alinhamento	Número de seqüências a serem alinhadas
BLAST2Sequences	Local	Heurístico	2
SWAT (Smith-Waterman)	Local	Ótimo	2
ClustalW	Global	Heurístico	N
Multalin	Global	Heurístico	N
Needleman-Wunsch	Global	Ótimo	2

Tab1. Exemplo de alinhamento global e local entre duas sequências. [3]

2. FASTA vs BLAST



- BLAST e FASTA são algoritmos de alinhamento local;
- BLAST é mais rápido que o FASTA;
- BLAST é mais preciso que o FASTA;
- BLAST é mais versátil e mais amplamente utilizado que o o FASTA;
- Partem da ideia básica: Um bom alinhamento contém subsequências de identidade absoluta (pequenas palavras de similaridade exata) [5].

3. BLAST

- Basic Local Alignment Search Tool;
- Ferramenta de alinhamento mais utilizada no mundo;
- O artigo onde a ferramenta foi publicada é o mais citado da história das ciências biológicas;
- É um algoritmo de alinhamento simples, heurístico e local;
- Alinha um sequência de entrada contra uma base de dados desejada;

3. BLAST

Programa	Seqüência de entrada	Tipo de seqüência alvo
blastp	proteína	proteína
blastn	nucleotídeo	nucleotídeo
blastx	nucleotídeo traduzido	proteína
Tblastn	proteína	nucleotídeo traduzido
Tblastx	nucleotídeo traduzido	nucleotídeo traduzido

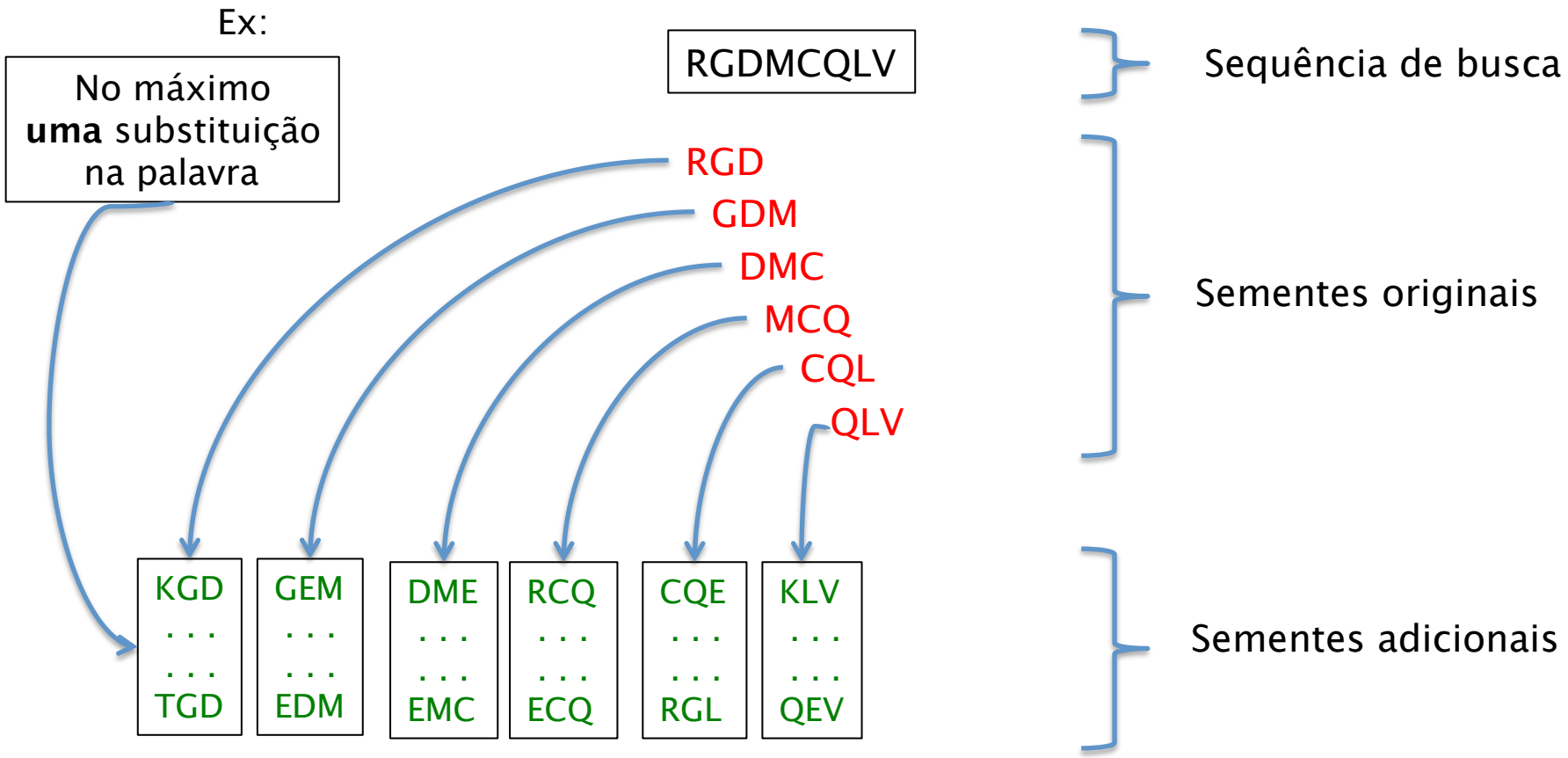
Tab2. Família de programas BLAST. [4 – Adaptada]

3.1. Funcionamento do algoritmo

- Consiste em 3 etapas heurísticas:
  - **Semeadura**;
    - Separa a sequência de busca em palavras;
    - Identifica onde começa o alinhamento;
  - **Extensão**;
    - Extende o alinhamento das sementes;
  - **Avaliação**;
    - Determina quais alinhamentos são significantes;

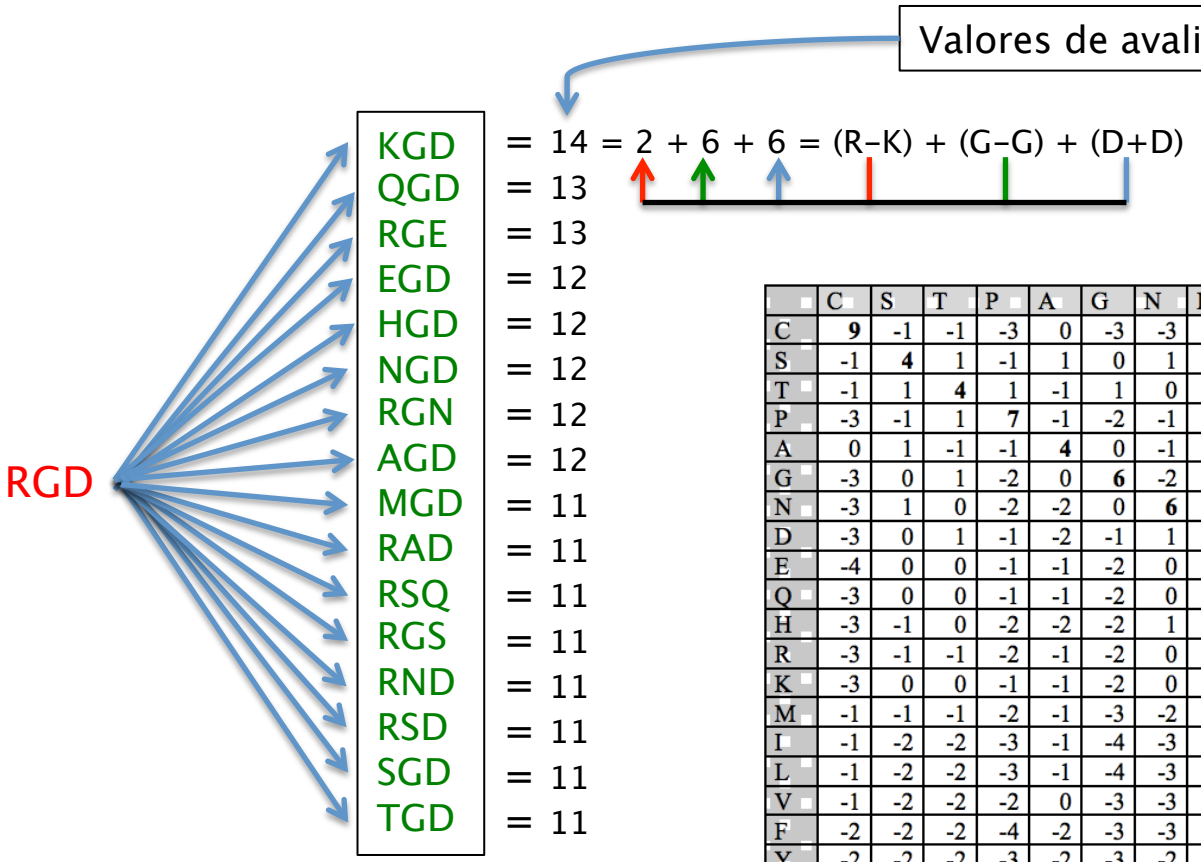
3.1.1. Semeadura

- Dada uma sequência de entrada, identifique todas as sequências de um tamanho específico (sementes).



3.1.1. Semeadura

- Comparar as sementes adicionais com a semente original correspondente, utilizando uma matriz de substituição (recomenda-se a matriz BLOSUM62 [6]).

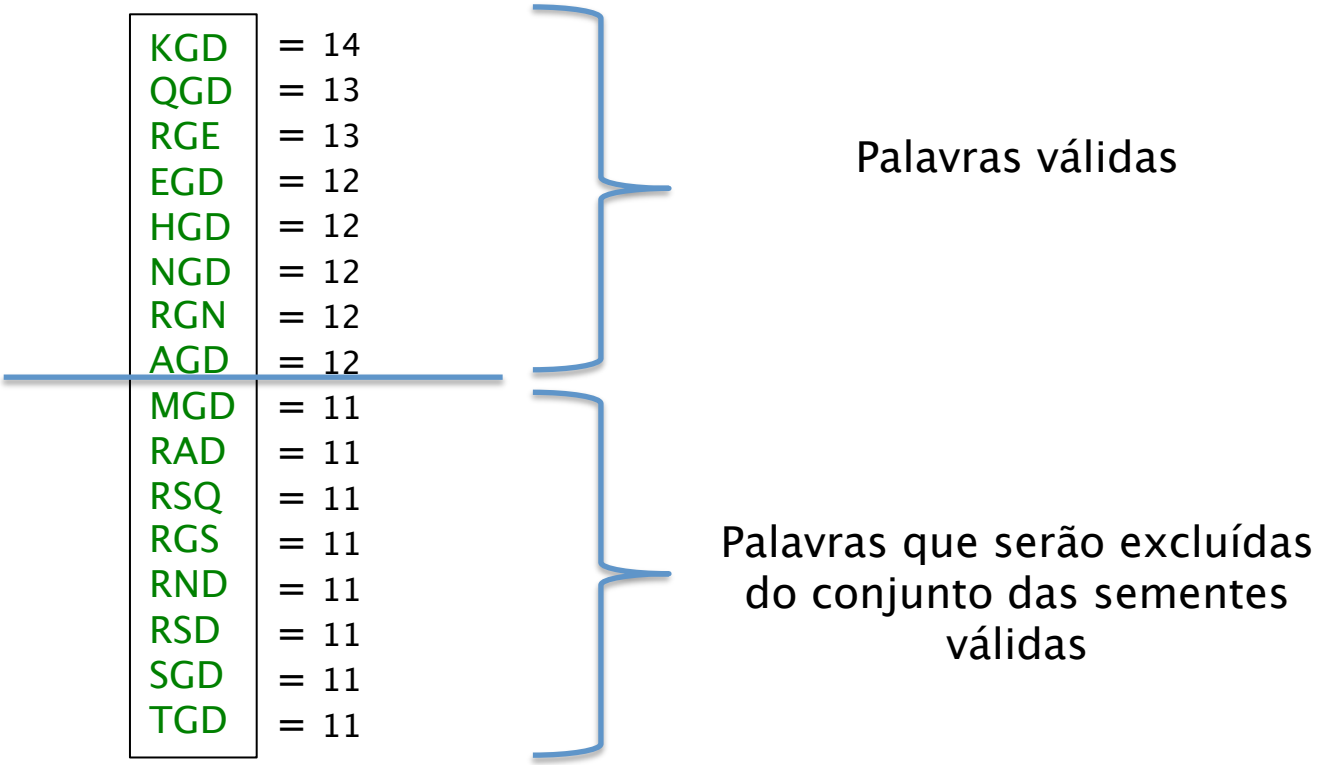


BLOSUM62

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

3.1.1. Semeadura

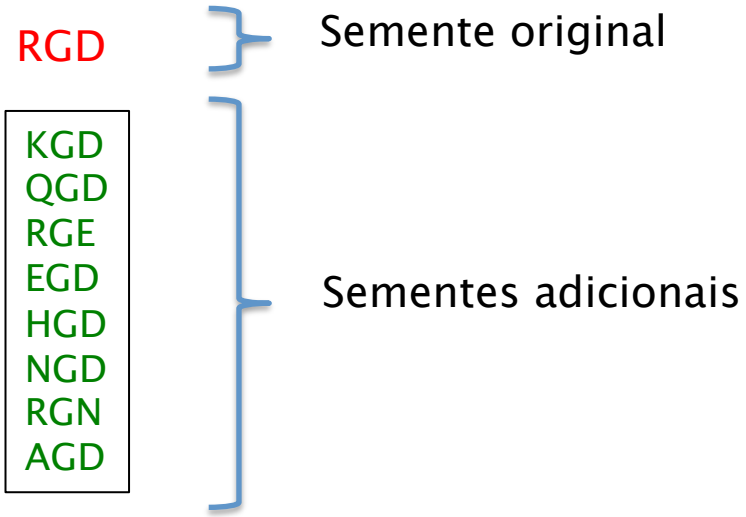
- Definir um valor mínimo de avaliação na seleção das sementes adicionais;
  - Padrão BLAST, em geral, é utilizado o valor 12 para palavras de tamanho 3;



3.1.1. Semeadura

- Conjunto de sementes válidas para a busca no banco de dados:

Sementes originais + sementes adicionais



3.1.1. Semeadura

- Realizar a busca pelas sementes no banco de dados (prioridade para sementes originais);

Ex:

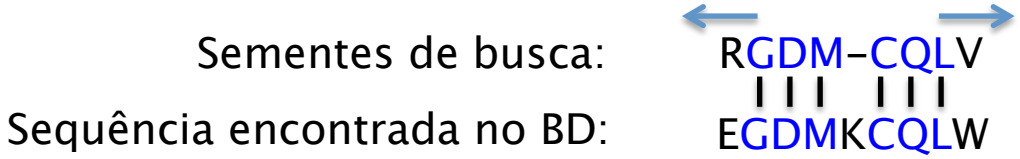
Sementes de busca: GDM CQL

Sequência encontrada no BD: EGDMKCQLW

3.1.2. Extensão

- Extender o alinhamento das sementes;

Ex:



- Extende cada semente para a direita e para esquerda, considerando os seguintes critérios [7]:
  - A pontuação (*score*) da semente for maior que um valor **T**;
  - Possuir outra semente a uma certa distância máxima entre elas;
  - O *score* da extensão com *gaps* tem pontuação normalizada de pelo menos S_g' *bits*;

- Muitas vezes é necessário adicionar *gaps* (buracos) para corrigir o alinhamento;
  - *Gaps* são vistos pelo BLAST como penalidades;
  - Quanto menos *gaps,* melhor o alinhamento;

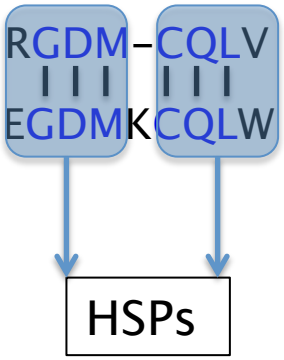
3.1.2. Extensão

- Extender o alinhamento das sementes;

Ex:

Sementes de busca: RGDM–CQLV

Sequência encontrada no BD: EGDMKCQLW



- HSPs (*High-scoring Segment Pair*): São alinhamentos locais que atingem os *scores* mais altos em uma busca;
- MSPs *(Maximal-scoring Segment Pair):* São os maiores HSPs encontrados na busca;

3.1.3. Avaliação

- ***Score* bruto**:

  S = soma(*matches*) – soma(*mismatches*) – soma(penalidades de *gap*)

- ***Score* normalizado (Bit score)**:

  $S' = (\lambda S - \ln K) / \ln 2$

- ***E-value*** (probabilidade de alinhamentos terem ocorrido ao acaso [2])***:***

  $E(S) = Kmne^{-\lambda S}$

  ou

  $E(S') = mn2^{-S'}$

legenda

m = Tamanho do banco de dados
n = Tamanho da sequência de entrada
λ = Escala da matriz de *scores*
K = Escala do tamanho do espaço de busca

Penalidades de *gap*:

(*gap open + gap extension*)

Gap open: é definido um valor
Gap extension: é definido um valor

3.1.3. Avaliação

Sementes de busca: RGDM-CQLV
Sequência encontrada no BD: EGDMKCQLW

Definir *Gap Costs:*

Gap open: 5

Gap extension: 2

De acordo com [8], valores de:

Match/mismatches = 1/-3 => Sequências 99% conservadas

***Match/mismatches* = 1/-2** => Sequências 95% conservadas

Match/mismatches = 1/-1 => Sequências 75% conservadas

Matches: 6*1 = 6

Mismatches: 2*2 = 4

Gap open: 1*5 = 5

Gap extension: 1*2 = 2

Resultado parcial

Similaridade: 6/9 (67%)

S = 6 – 4 – 5 –2 = **–5**

- Outro exemplo, adaptado de [2]:

Resultado do BLAST

```
Score = 68.0 bits (35),  Expect = 4e-09
 Identities = 55/60 (92%), Gaps = 2/60 (3%)
 Strand=Plus/Plus

Query  1    TTACTAGCAGAGCTTGGTGTTATTGCTGCCAC-ATACCATTT-TTTTAGGAACTGCTTTA  58
            ||||||||||||||||||||||||||||||| ||||||    ||||||||||||||||
Sbjct  723  TTACTAGCAGAGCTTGGTGTTATTGCTGCCACTATACCAGGAGTTTTAGGAACTGCTTTA  782
```

Matches: 55*1 = 55
Mismatches: 3*(2) = 6
Gap Open: 2*5 = 10
Gap extension: 2*2 = 4

S = 55 - 6- 10 - 4 = 35

S' = [λR – ln(K)]/ln(2)
[1.33*35-ln(0.6209)]/ln(2) = 68

$E = mn2^{-S}$
$34*33036625752*2^{-68}$ = 4e-09

Valores calculados na "mão"

Referências

[1] http://www.bdtd.ucb.br/tede/tde_arquivos/13/TDE-2004-11-12T134348Z-143/Publico/Dissert_Felipe%20Liberman.pdf

[2] http://minerva.ufpel.edu.br/~lmaia.faem/aula4.pdf

[3] http://www.biotecnologia.com.br/revista/bio29/bioinf.pdf

[4] http://www.bdtd.ucb.br/tede/tde_arquivos/13/TDE-2004-11-12T134348Z-143/Publico/Dissert_Felipe%20Liberman.pdf

[5] http://csb.stanford.edu/class/public/readings/Bioinformatics_I_Lecture6/Altschul_JMB_90_BLAST_Sequence_alignment.pdf

[6] Henikoff, S. & Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA 89:10915-10919.

[7] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/pdf/253389.pdf

[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC55814/

OBRIGADO

Contatos
www.marllus.com
marlluslustosa@gmail.com