_____ / _____ / _____

# Bioinformática - 1

Em 26 de junho de 2000, o primeiro ministro Tony Blair e o presidente Bill Clinton anunciaram o término do primeiro esboço do sequenciamento do genoma humano. Estamos falando de algo como 3 bilhões de bases. Desde que os tribunais disseram que sequências de genes eram patenteáveis, começou uma corrida entre as empresas e a academia para ver quem chegava antes. A estratégia da academia é disponibilizar em bancos de dados públicos o quanto antes para impedir a proteção via patente. Existem 3 bancos mundiais, que são diariamente sincronizados: o GenBank em Maryland, USA; o Nucleotide Sequence Database em Hinxton, Reino Unido e o DNA Databank of Japan.

A genética se baseia nos conceitos de genótipo e fenótipo e o motor da evolução é a assimetria entre ambos. O genótipo é formado por uma longa cadeia de DNA. Existem 4 nucleotídeos no DNA: a=adenina, g=guanina, c=citosina e t=timina (esta última vira u=uracil na transcrição via RNA). Note que os nucleotídeos são escitos em minúscula. Os nucleotídeos são organizados de 3 em 3 formando 20 aminoácidos. A descoberta deste fato rendeu o prêmio Nobel de medicina em 1968. Os 20 aminoácidos são: G=glicina, A=alanina, P=prolina, V=valina, I=isoleucina, L=leucina, F=fenilalanina, M=metionina, S=serina, C=cisteína, T=treonina, N=asparagina, Q=glutamina, H=histinina, Y=tirosina, W=triptofano, D=ácido aspártico, E=ácido glutâmico, K=lisina e R=arginina. Os aminoácidos são escritos em maiúsculo.

Neste ponto, vá a um micro com internet acesse o Swiss Institute for Bioinformatics: `http://www.uniprot.org/` e depois demande `cat pancreatic ribonuclease`. Você acabou de pedir os aminoácidos da ribonuclease pancreática do gato. Eis a resposta:

```
MAVAVARLVF LQLAFGPALV VDIEMQIAIK DFHMLHVDYP  40
RVHYPKGFQG YCNGLMAYVR GRKQKSWYCP QIHYMVHAPW  80
REVQKFCKYS ESFCENYNEY CTFTEDSFPI TICSLAPNQP 120
PTSCYYNSTL TNQRLYLLCS GKRDAEPIDI IGVY       154
```

Para saber mais: LESK, Arthur. **Introdução à Bioinformática**. Editora Artmed.

## Algoritmos

Existem inúmeros algoritmos que manipulam cadeias (ou strings) que vem a ser apenas sequências de caracteres retirados de um determinado universo. Vamos estudar três deles:

- Determinar a maior subsequência comum em duas cadeias. (nesta folha 536b)
- Custo para transformar uma cadeia em outra e
- Determinar a ocorrência de um padrão em um texto (nosso conhecido CTRL-F). Estes 2 últimos na próxima folha 536c.

**Maior subsequência comum** Começa-se definindo sequência que vem a ser uma lista de caracteres na qual a ordem é importante. Para facilitar a prosa, vamos chamar sequência de cadeia. Assim, a cadeia ATCCG é diferente da cadeia TAGCC. Uma subsequência de uma cadeia é a própria cadeia com a possível eliminação de caracteres. Por exemplo, se $X$ é a cadeia ATC, então ela porta 8 subsequências a saber: ATC, AT, AC, TC, A, T, C, e a subsequência nula, que vem a ser a cadeia vazia. Se $X$ e $Y$ são cadeias e $Z$ é uma subsequência comum, então $Z$ pertence a ambas cadeias. Define-se aqui subcadeia, que é uma subsequência de caracteres contíguos. Por exemplo, na cadeia CATCGA a subsequência ATCG é subcadeia, mas a subsequência CTCA não é subcadeia. O objetivo do algoritmo a seguir é, dadas duas cadeias $X$ e $Y$, determinar a subsequência $Z$ mais longa que existe. Supondo $t(x)$ o tamanho da cadeia $X$ e $t(y)$ o tamanho de $Y$, a primeira coisa a fazer é construir uma matriz M, de $t(x)+1$ linhas por $t(y)+1$ colunas, colocando $X$ na vertical da matriz e $Y$ na horizontal, começando ambas na segunda linha e coluna. A primeira linha e a primeira coluna têm ambas o valor 0. O resto da matriz é preenchida usando a seguinte regra:

```
1: M matriz de t(x) + 1 linhas por t(y) + 1 colunas contendo zeros
2: para i = 2, i ≤ t(x), i + +
3:     para j = 2, j ≤ t(y), j + +
4:         se x[i] = y[j]
5:             M[i, j] ← M[i − 1, j − 1] + 1
6:         senão
7:             M[i, j] ← o maior entre M[i, j − 1] e M[i − 1, j]
8:         fim{se}
9:     fim{para}
10: fim{para}
```

O comprimento da maior subsequência está dado na célula $M[1+t(x), 1+t(y)]$. Agora, para obter a MSC (maior subsequência comum) execute-se o seguinte algoritmo:

```
1: MSC (M, X, Y)
2: RESP ← cadeia-vazia
3: i ← 1 + t(x)
4: j ← 1 + t(y)
5: enquanto M[i; j] ≠ 0
6:     se X[i − 1] = Y[j − 1]
7:         RESP ← X[i − 1] + RESP
8:         i − −
9:         j − −
10:    senão
11:        se M[i; j − 1] < M[i − 1; j]
12:            i − −
13:        senão
14:            j − −
15:        fim{se}
16:    fim{se}
17: fim{enquanto}
18: retorne RESP
```

Em RESP está a Maior Subsequência Comum. .

## Exemplos

1. Seja $x =' LMMNPQ'$ e $y =' LALAMNMMMPAQASA'$. A matriz $M$ é

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2
0 1 1 1 1 2 2 3 3 3 3 3 3 3 3 3
0 1 1 1 1 2 3 3 3 3 3 3 3 3 3 3
0 1 1 1 1 2 3 3 3 3 3 4 4 4 4 4
0 1 1 1 1 2 3 3 3 3 4 4 5 5 5 5
```

e a MSC é $LMNPQ$.

2. Seja $x =' GGGADFEABGDDCAD'$ e
$y =' CBGDDFCCDBAEBGCGABBABDFGAAABCE'$ eis como ficaria a matriz $M$

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
0 0 0 1 2 2 2 2 2 2 2 2 2 2 2 3 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5
0 0 0 1 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 5 6 6 6 6 6 6 6 6 6
0 0 0 1 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 6 6 6 6 6 6 6 6 7
0 0 0 1 2 2 3 3 3 3 3 4 4 4 4 4 4 5 5 5 5 5 6 6 7 7 7 7 7 7 7
0 0 1 1 2 2 3 3 3 3 4 4 4 5 5 5 5 5 6 6 6 6 6 6 6 7 7 7 8 8 8
0 0 1 2 2 2 3 3 3 3 4 4 4 5 6 6 6 6 6 6 6 6 6 6 7 7 7 7 8 8 8
0 0 1 2 3 3 3 3 4 4 4 4 5 6 6 6 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8
0 0 1 2 3 4 4 4 4 4 4 4 5 6 6 6 6 6 6 6 6 7 7 7 7 7 7 8 8 8 8
0 1 1 2 3 4 4 5 5 5 5 5 5 5 6 7 7 7 7 7 7 7 7 7 7 7 7 8 9 9 9
0 1 1 2 3 4 4 5 5 5 5 6 6 6 6 7 7 8 8 8 8 8 8 8 8 8 8 8 9 9 9
0 1 1 2 3 4 4 5 5 6 6 6 6 6 6 7 7 8 8 8 8 8 9 9 9 9 9 9 9 9 9
```

Como se viu acima a maior subsequência tem tamanho 9 e ela é $MSC = GDFABGCAD$.

## ☞ Para você fazer

1. Suponha $x = GFADCEFBEEBBBED$
   e $y = EBEDEGDBDBFEECGGDFEAGGEAECDGDD$
   Ache a maior subsequência comum e responda

| tamanho | qual é ela |
|---|---|
| | |

2. Suponha $x = DCDCCEFCCBFBBFC$
   e $y = FFBECCFDEACCBABDCFFDAFABAFFECE$
   Ache a maior subsequência comum e responda

| tamanho | qual é ela |
|---|---|
| | |