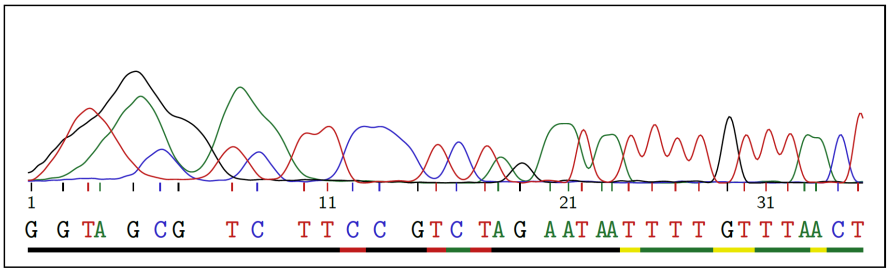
PMR3201 - Computação para Mecatrônica

Prof. Thiago de castro Martins
Prof. Newton Maruyama
Prof. Marcos de S.G. Tsuzuki
Monitor: Pietro Teruya Domingues

Exercício Programa 1 - Versão 2017

Algoritmo de busca de subcadeias de caracteres: uma aplicação em biologia molecular



- **DATA FINAL DE ENTREGA: Sexta-feira 24/04/2017**
- **O exercício deve ser feito individualmente.**
- **Submeta através do sistema MOODLE:**
  - **O relatório com os resultados em formato PDF.**
  - **Código fonte e relatório em um arquivo compactado.**

1 Introdução

Nesse Exercício Programa será desenvolvido um algoritmo de busca de subcadeias de caracteres, i.e., busca de um padrão definido por uma subcadeia de caracteres dentro de uma cadeia de caracteres. Algoritmos de busca de subcadeias são utilizados em diversas áreas como: editores de texto, máquinas de busca para *web*, filtros de *spam*, biologia computacional (busca de padrões de sequencia de DNA ou de proteínas), detecção de *features* em imagens, etc.

Sejam as seguintes definições:

- Texto ou cadeia de caracteres T de comprimento $|T| = n$,
- Subcadeia de caracteres denominada padrão P, de comprimento $|P| = m$.

Obviamente, devemos ter $n \geq m$ para que o algoritmo possa ser executado. O problema de busca exata de subcadeia de caracteres (*Perfect string matching*) consiste em determinar todas as posições

$s \in \{0, \ldots, n-m\}$ em T aonde se inicia uma subcadeia de caracteres que coincide com o padrâo P, ou seja, matematicamente devemos ter:

$$P[i] = T[s+i], \forall i \in \{0, \ldots, m\}. \tag{1}$$

O algoritmo mais simples para a busca exata de subcadeia de caracteres é um algoritmo do tipo força bruta denominado usualmente como *the Naïve algorithm*. O algoritmo possui complexidade $O(m \times n)$. O algoritmo consiste em percorrer as posições do texto T e a cada posição verificar se esta é o início de uma subcadeia de caracteres coincidente com o padrão P.

Por exemplo dado o Padrão P e o texto T ilustrado abaixo. O algoritmo deve encontrar a posição $s = 3$ como resposta única.

Padrão P

0	1	2	3
A	T	A	T

Texto T

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	T	G	A	T	A	T	T	T	G	T	C	G	G	C	A	T	T	A	C

Figura 1: Exemplo de busca de subcadeia de caracteres exata.

Além de algoritmos de busca exata existe um interesse crescente em algoritmos de busca inexata ou aproximada (*Partial string matching*). Dentro desse contexto é necessário a definição de uma métrica que estabeleça uma medida de distância d entre duas subcadeias de caracteres. Várias definições distintas de métrica podem ser estabelecidas.

Vamos estabelecer aqui uma noção de distância baseada no número de caracteres do padrão P que são coincidentes com uma subcadeia de mesmo tamanho contida no texto T.

Se m caracteres forem coincidentes então a distância $d = 0$, se $m - 1$ caracteres forem coincidentes então a distância $d = 1$ e assim por diante.

Utilizando como exemplo o Padrão P e o Texto T ilustrados na Figura 1. Os seguintes resultados são obtidos:

s	subcadeia	d		s	subcadeia	d
0	"ATGA"	2		10	"TCGG"	4
1	"TGAT"	2		11	"GCGC"	4
2	"GATA"	4		12	"GGCA"	4
3	"ATAT"	0		13	"GCAT"	2
4	"TATT"	3		14	"CATT"	3
5	"ATTT"	1		15	"ATTA"	2
6	"TTTG"	3		16	"TTAC"	3
7	"TTGT"	2				
8	"TGTC"	4				
9	"GTCG"	3				

Na área de biologia molecular a busca aproximada de subcadeias de caracteres é parte importante do processo de análise de sequência do material genético (DNA, RNA) ou análise de sequência de peptídeos para o entendimento de suas características, função ou evolução. O programa de domínio público BLAST (*Basic Local Alignment Search Tool*) é a ferramenta de análise mais utilizada no momento. O genoma

humano possui uma estimativa de cerca de 3 bilhões de pares de nucleotídeos. Análises desse genoma só são possíveis através de algoritmos eficientes.

O sequenciamento de DNA é realizado através de máquinas que utilizam a técnica de cromatografia para identificar a sequência de nucleotídeos base: `G` (Guanine), `C` (Cytosine), `A` (Adenosine) e `T` (Thymine).

Através da comparação de sequências de DNA de diferentes organismos e medindo o número de mudanças (mutações) é possível determinar se as espécies estão relacionadas de maneira próxima ou distante indicando um possível caminho evolucionário.

Por exemplo, veja a similaridade entre as sequencias de DNA de *cytochrome c* (Presente nas membranas de mitocondrias) de humanos, chimpanzes e ratos.

```
#Human
ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCCCAGTGCCACACC
GTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTTTGGGCGGAAG
ACAGGTCAGGCCCCTGGATACTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTGG
GGAGAGGATACACTGATGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAA
ATGATCTTTGTCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTATCTCAAA
AAAGCTACTAATGAGTAA
#Chimpanzee
ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCCCAGTGCCATACC
GTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCTTCGGGCGGAAG
ACAGGTCAGGCCCCTGGATATTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTGG
GGAGAGGATACACTGATGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAA
ATGATATTTGTCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTATCTCAAA
AAAGCTACTAATGAGTAA
#Mouse
ATGGGTGATGTTGAAAAAGGCAAGAAGATTTTTGTTCAGAAGTGTGCCCAGTGCCACACT
GTGGAAAAGGGAGGCAAGCATAAGACTGGACCAAATCTCCACGGTCTGTTCGGGCGGAAG
ACAGGCCAGGCTGCTGGATTCTCTTACACAGATGCCAACAAGAACAAAGGCATCACCTGG
GGAGAGGATACCCTGATGGAGTATTTGGAGAATCCCAAAAAAGTACATCCCTGGAACAAAA
ATGATCTTCGCTGGAATTAAGAAGAAGGGAGAAAGGGCAGACCTAATAGCTTATCTTAAA
AAGGCTACTAATGAGTAA
```

Seríamos então descendentes de ratos ?

2 Especificações

1. Deseja-se pesquisar a existência de determinadas sequencias de DNA em genomas de Vírus da Dengue, mais especificamente utiliza-se a variedade BA05i do Vírus da Dengue Tipo 2 originário da Indonésia [1] e a variedade TB55i do Vírus da Dengue Tipo 3 originário da Malasia [2]. Ambos os genomas podem ser encontrados no banco de dados do NCBI *The National Center for Biotechnology Information*.

2. Os dois genomas se encontram nos arquivos `DengueVirus2StrainBA05i_Jakarta.txt` e `DengueVirus3StrainTB55i_KualaLumpur.txt`. Os dois genomas foram editados para terem o mesmo número de nucleotídeos. O formato de cada arquivo é organizado em sete colunas sempre a primeira se referindo ao índice da posição do primeiro nucleotídeo da linha em questão. As seis colunas adicionais são agrupamentos de dez nucleotídeos. Um trecho do arquivo `DengueVirus2StrainBA05i_Jakarta.txt` é apresentado a seguir.

[1] `http://www.ncbi.nlm.nih.gov/nuccore/AY858035.2`
[2] `http://www.ncbi.nlm.nih.gov/nuccore/AY858048.2`

```
1 agttgttagt ctacgtggac cgacaaagac agattctttg aggaagctaa gcttaacgta
61 gttctaacag ttttttaatt agagagcaga tctctgatga ataaccaacg gaaaaaggcg
121 agaaatacgc ctttcaatat gctgaaacgc gagagaaacc gcgtgtcaac tgtgcagcag
181 ctgacaaaga gattctcact tggaatgcta cagggacgag gaccattgaa actgttcatg
241 gccctggtgg cattccttcg tttcctaaca atcccgccaa cagcagggat attaaaaaga
301 tggggaacaa tcaaaaaatc aaaggctatc aatgtcttga gagggttcag gaaagagatt
361 ggaaggatgc tgaacatctt gaacaggaga cgcagaacag caggtataat tattatgatg
...
```

O algoritmo exige que o Padrão P e o Texto T sejam armazenados numa variável do tipo `String`. Os arquivos de genomas de vírus da Dengue farão o papel de Texto T Uma possível subrotina que faz a leitura do arquivo e armazena a sequência de DNA numa `String` é apresentado em seguida.

Listing 1: Subrotina para leitura dos arquivos.

```
def LeArquivoDNA(filename):
    files=open(filename, 'r')
    lists=files.readlines() #this is a matrix of nlines
    nlines=len(lists)     # number of lines

    a = lists[0].rstrip('\n').split(' ') # separa a linhas em colunas
                                # observa o espaco ' ' como
                                # caracter de separacao
                                # descarta \n
    cadeiacompleta = a[1] + a[2] + a[3] + a[4] + a[5] + a[6]
    # concatena as colunas 1..6 ignora a[0]
    # agora que cadeiacompleta nao e'vazio faca ate o final
    for i in range(1,nlines):
        a=lists[i].rstrip('\n').split(' ')
        cadeiacompleta = cadeiacompleta + a[1]+a[2]+a[3]+a[4]+a[5]+a[6]
    #retorna a string completa
    return(cadeiacompleta)
```

3. Os seguintes Padrões P devem ser utilizados para busca de cadeias de caracteres aproximada:
   (a) `"actgcttctg"`
   (b) `"aggaggctgg"`
   (c) `"tacatgccat"`
   (d) `"cctcagcatc"`
   (e) `"gcaacgttca"`
   (f) `"gacattgact"`

4. Um parâmetro que define a máxima distância de aproximação d_{max} entre subcadeias de caracteres deve ser definido. Este parâmetro pode ser escolhido pelo usuário do programa. Por exemplo, se $d_{max} = 2$ somente serão considerados os resultados para as distâncias $d = \{0, 1, 2\}$.

5. Os resultados devem resumidos em um arquivo do tipo *.txt cujo nome pode ser escolhido pelo usuário. Você deve indicar para cada Padrão P e para cada Texto T a posição s em que o Padrão P é encontrado de maneira exata ou aproximada e a medida de distância d entre as subcadeias de caracteres.

   Um possível formato do arquivo de resultados é apresentado abaixo.

```
Results for partial matching on DNA Sequences
Maximum distance considered = 2
```

```
Test results for DNA Sequence of Virus Type 2
Length of DNA Sequence of Virus Type 2 = 10680

Test - Pattern P = <actgcttctg> - DNA Sequence Virus Type 2
Distance bettween Pattern P and DNA Sequence Virus Type 2
s=3324 <actgcttct> Distance = 0
s=3524 <actgttcct> Distance = 2
s=4013 <actgctttt> Distance = 2
s=5996 <aatgcttct> Distance = 2

Test - Pattern P = <aggaggctgg> - DNA Sequence Virus Type 2
Distance bettween Pattern P and DNA Sequence Virus Type 2
s=363 <aggatgctg> Distance = 2
s=779 <aggggcctg> Distance = 2
s=986 <aggaagctg> Distance = 1
s=3215 <aggaccctg> Distance = 2
s=3401 <tggatgctg> Distance = 2
s=3506 <aggagtctt> Distance = 2
s=4760 <aggaggctg> Distance = 0
s=4778 <aggagaatg> Distance = 2
s=5159 <acgaggctt> Distance = 2
s=7820 <ggggggctg> Distance = 2
s=9659 <aagaggatg> Distance = 2
s=10249 <aagaggcag> Distance = 2
s=10256 <aggagtctt> Distance = 2
```

3 Referências

1. Algorithms, Robert Sedgewick and Kevin Wayne, Addison-Wesley Professional, 4th Edition, 2011.