CCM128 EXERCÍCIO-PROGRAMA 1

GENES

Y. KOHAYAKAWA

Data de entrega: 17/3/2014 (23:55)

1. Introdução e definições básicas

Este EP trata de processamento de cadeias de caracteres (*strings*). Supomos fixo um *alfabeto* Σ, cujos membros chamamos de *letras* (um *alfabeto* é simplesmente um conjunto finito). Nossas cadeias de caracteres, ou *palavras*, sobre Σ são sequências finitas de letras. O conjunto das palavras sobre Σ é denotado Σ^*. O comprimento $|s|$ de uma palavra $s \in \Sigma^*$ é o número de letras que a compõe. Dadas duas palavras s e $t \in \Sigma^*$, a *concatenação* st destas palavras é uma nova palavra sobre Σ; claramente $|st| = |s| + |t|$. A palavra vazia (aquela de comprimento 0) é denotada por λ.

2. Genes

Uma palavra f é um *fator* de uma palavra s se $s = s_1 f s_2$ com s_1 e $s_2 \in \Sigma^*$. O conjunto dos fatores de uma palavra s é denotado $\mathrm{Fat}(s)$.

No que segue, consideramos sempre $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Ademais, as palavras $s = \mathtt{ATG}$ e $t = \mathtt{TAG}$ terão papéis especiais. Seja dada uma palavra $D \in \Sigma^*$. Um *gene* em D é um fator $g \neq \lambda$ de D de comprimento divisível por 3 e tal que $sgt \in \mathrm{Fat}(D)$. Assim, um gene em D é um fator de comprimento multiplo de 3, precedido de $s = \mathtt{ATG}$ e seguido de $t = \mathtt{TAG}$. Note que D pode conter um dado gene g em vários locais diferentes.

3. Seu programa

Seu EP1 deve ter como entrada uma palavra $D \in \Sigma^*$. No modo mais simples, seu programa deve ter como saída o número *total* de genes que D contém, contando com multiplicidade (se um dado gene ocorre m vezes, este gene contribui com m nesse total). Por exemplo, na entrada

```
CATGCTATAGATGTAGTAGAGCATGTAGCTAAGCTAGATGCTATAGATGAGCATGCTAG
```

seu programa deve ter como saída o inteiro 13:

```
yoshi@renyi ~/TMP $ java-introcs GeneFindAll < ex.txt
13
yoshi@renyi ~/TMP $
```

(`ex.txt` contém o exemplo acima). Seu programa deve também ter um modo, indicado pela opção `-v`, em que ele imprime todos os genes encontrados:

```
yoshi@renyi ~/TMP $ java-introcs GeneFindAll -v < ex.txt
CTA
CTATAGATG
CTATAGATGTAG
```

Versão de 16 de abril de 2014, 19:35.

```
CTATAGATGTAGTAGAGCATG
CTATAGATGTAGTAGAGCATGTAGCTAAGC
CTATAGATGTAGTAGAGCATGTAGCTAAGCTAGATGCTA
TAG
TAGTAGAGCATG
TAGTAGAGCATGTAGCTAAGC
TAGTAGAGCATGTAGCTAAGCTAGATGCTA
TAGCTAAGC
TAGCTAAGCTAGATGCTA
CTA
yoshi@renyi ~/TMP $
```

Instâncias. Você deve supor que seu EP1 será tipicamente executado com palavras D bastante longas (no modo padrão (sem `-v`)). Por exemplo, D poderia ser a sequência completa de nucleotídeos que compõe um cromossomo de um organismo simples. Por exemplo, veja o cromossomo III do *C. elegans* em `http://www.ncbi.nlm.nih.gov/nuccore/NC_003281.10?report=fasta&log$=seqview` ($>$ 20MB). (Ao visitar a URL acima, você perceberá que é necessário remover mudanças de linha.)

Observações. Seguem algumas observações importantes.

1. *Este EP é estritamente individual.* Programas semelhantes receberão nota 0.
2. Seja cuidadoso com sua programação (correção, documentação, apresentação, clareza de código, etc). A correção não é baseada apenas na correção de seu programa.
3. Comparem entre vocês o desempenho de seus programas.
4. Entregue seu EP no Paca.
5. Não deixe de incluir em seu material um *relatório* para discutir seu EP: faça quaisquer comentários que você achar interessante e inclua exemplos de execuções de seu programa. *Quando mais o monitor vir as qualidades de seu programa, melhor será o seu humor e melhor será sua nota.*

Observação final. Enviem dúvidas para a lista de discussão da disciplina.

Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, 05508–090 São Paulo, SP

Endereço eletrônico: `yoshi@ime.usp.br`