Trabalho de Sistemas Distribuídos

Setembro 2018

Atenção! Leia atentamente a entrada e a saída do problema.

Problema B [1]

Subsequencias de DNA

O formato FASTA é um arquivo de texto que armazena sequências de DNA/RNA, no qual as bases hidrogenadas são representadas utilizando caracteres únicos. Na bioinformática, esse formato é utilizado, entre outras coisas, em aplicações de alinhamento de sequencias e comparações genéticas. Um mesmo arquivo FASTA representa diferentes sequências, delimitadas por uma linha de descrição iniciada pelo carácter maior que ('>'). A descrição é seguida por várias linhas contendo as bases hidrogenadas. As sequencias de base são compostas apenas pelos caracteres ('A', 'T', 'C', 'G') e são divididas em linhas de no máximo 80 caracteres. O arquivo termina com o marcado *EOF*. Escreva um programa paralela para encontrar uma subsequencia de DNA dentro de uma base FASTA. Se a *string* de busca for semelhante a diferentes sequências, cada resultado deve ser reportado. Se a *string* de busca for semelhante com múltiplas posições em uma mesma sequência, apenas a primeira posição deve ser reportada.

[1] Problema obtido na maratona de 2009 do SBAC

Entrada

A entrada é lida de dois arquivos diferentes, ambos no padrão FASTA. O tamanho das bases é de até 1,000,000 caracteres.
A base FASTA deve ser lida de um arquivo chamado dna.in As subsequencias, isto é, as *strings* de busca devem ser lidas do arquivo query.in

Saída

A saída contém as *strings* que foram encontradas. Para cada *string*, o programa deve escrever sua descrição, seguida pelo retorno da busca. Se a *string* de busca for encontrada dentro da base FASTA, o retorno deve conter a descrição da sequência seguida da posição de inicio da *substring*. Se a *string* de busca não for encontrada em nenhuma sequencia, a mensagem 'NOT FOUND' deve ser impressa.
A saída deverá ser escrita no arquivo dna.out.

Exemplo

Entrada
FASTA database

>Escherichia coli partial genome (1)
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAA
AAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGT
AAATTAAAATTTTATTGACTTAGGTCA
>Escherichia coli partial genome (2)
CTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTA
CAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCAC
CATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAAC
ACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAA
AGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACA

Entrada
Arquivo de Busca

>Query string 1
TATAGG
>Query string 2
TTTT
>Query string 3
ATCG
>Query string 4
AACTGG

Saída

>Query string 1
>Escherichia coli partial genome (2)
17
>Query string 2
>Escherichia coli partial genome (1)
3
>Escherichia coli partial genome (2)
178
>Query string 3
NOT FOUND
>Query string 4
>Escherichia coli partial genome (1)
75