DEA BBSG. Examen d'Informatique

Décembre 2000

Durée: 2 heures Daniel GAUTHERET
Documents autorisés Denis THIEFFRY

1. (2 pt)

Donnez une commande Unix permettant de trouver dans un fichier de séquence au format Fasta les occurrences de l'expression Prosite suivante :

```
C-x(2)-C-x-[DE]-x(5)-[HN]-[FY]-x(4)-C-x(2)-C-x(2)-F-F-x-R
```

Note sur la syntaxe Prosite: parenthèses = nombre de répétitions ; crochets = choix ; « x » = résidu indéterminé ; tirets = séparateurs.
On ignorera le problème des sauts de ligne.

2. (5 pt)

Indentifiez parmi les commandes Unix et Perl suivantes celles qui comportent une erreur. Corrigez l'erreur.

```
a.    if ($rep != /(oui|non)/) {print "recommencez";}
b.    print "le fichier %f n'existe pas\n";
c.    cp ~gauthere/fichier1
d.    if ($x = 3) {$a = $a + 7;}
e.    rm –r ../Temp/*
f.    ls –l > toto
g.    while (ligne = <F>) {}
h.    $thisentry  =~ /^LOCUS\s+(\w+);
i.    print FOUT $libnum, "\t", $library, "\n";
j.    $options .= @ARGV[$i];
```

3. (6 pt)

Considérez le programme Perl en annexe 1.

a. Quelle est la fonction de la partie de code marquée « 1 » ?
b. Quelle est la fonction de la partie de code marquée « 2 » ?
c. A quoi sert la variable $totnt ?
d. A quoi sert la variable $Aprop ?
e. A quoi sert le programme ?

f. Quel serait la sortie du programme si on lui fournit en argument le fichier de l'Annexe 2 ?

Note : L'expression « `$x =~ tr/y/y/` » est utilisée ici pour renvoyer le nombre d'occurrences de y dans $x.

4. (4 pt)

Le fragment de programme Perl en Annexe 3 lit et affiche les séquences d'un fichier Fasta. Transformez-le pour lire des séquences au format «GENBANK». Un exemple de format Genbank est fourni en Annexe 4.

5. (3 pt)

La variable `$seq` est une chaîne de caractères contenant une séquence nucléotidique de taille quelconque. Réalisez une Boucle Perl permettant de parcourir cette séquence codon par codon et de stocker les codons dans un tableau `@codons`. On utilisera la fonction `substr` pour parcourir la séquence et on admettra que le premier codon débute en position zéro. (la commande `substr ($s, x, y)` renvoie la sous chaîne de `$s` à partir de la position `x` sur une longueur `y`)

ANNEXE 1

```
#!/usr/bin/perl

if ($#ARGV != 0) {
    print "Programme <Fichier sequence Fasta> \n";
    exit;
}

open(F, $ARGV[0]) || die ("\"$ARGV[0]\" introuvable\n");

while ($line = <F>){
    if ($line =~ /^>/) {
        $totseq++;
    }else{
        chop $line;
        $line =~ tr/[\s]//d;
        $line =~ tr/[a-z]/[A-Z]/;
        $line =~ tr/[T]/[U]/;
        $seq = $line;

        $l = length ($seq);
        $totnt += $l;

        $Acnt += $seq =~ tr/A/A/;
        $Ucnt += $seq =~ tr/U/U/;
        $Gcnt += $seq =~ tr/G/G/;
        $Ccnt += $seq =~ tr/C/C/;
        $Ncnt += $seq =~ tr/N/N/;
    }
}

$base_nonN= $Acnt + $Ucnt + $Gcnt + $Ccnt;

$Aprop = $Acnt / $base_nonN * 100;
$Uprop = $Ucnt / $base_nonN * 100;
$Gprop = $Gcnt / $base_nonN * 100;
$Cprop = $Ccnt / $base_nonN * 100;
$Nprop = $Ncnt / $totnt * 100;

print "Totseq   totnt   moy-nt    A     U     G     C    N\n";
printf ("%5d   %7d  %5d  %5.1f %5.1f %5.1f %5.1f %5.1f\n",
        $totseq, $totnt, $totnt/$totseq,
        $Aprop, $Uprop, $Gprop, $Cprop, $Nprop);
```

1 — lignes 3 à 7

2 — ligne 9

ANNEXE 2

```
>seq 1
AAAGGGAAACCC
>seq 2
GGGAAAAAACCC
```

Annexe 3

```
$line = <F>;
while ($line) {
    if ($line =~ /^>/) {
        $name = $line;
        $seq = "";
        while (($line = <F>) && ($line !~ /^>/)) {
            chop $line;
            $seq .=  $line;
        }

        print $name;
        print $seq;

    } else {
        $line = <F>;
    }
}
```

Annexe 4

Séquence au format Genbank. Plusieurs enregistrements de ce type peuvent se trouver à la suite dans le même fichier.

```
LOCUS       AF299079     1052 bp    DNA             BCT       22-OCT-2000
DEFINITION  Bartonella henselae elongation factor EF-Tu (tuf) gene, partial
            cds.
ACCESSION   AF299079
VERSION     AF299079.1  GI:10945626
KEYWORDS    .
SOURCE      Bartonella henselae.
  ORGANISM  Bartonella henselae Bacteria; Proteobacteria; alpha subdivision; Rhizobiaceae
       group;
            Bartonellaceae; Bartonella.
REFERENCE   1  (bases 1 to 1052)
  AUTHORS   Chow,V.T.K., Yeo,W., Soong,P.L. and Nasirudeen,A.M.A.
  TITLE     Sequence and Evolutionary Characterization of the Elongation Factor
            Tu Gene in Bartonella henselae
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1052)
  AUTHORS   Chow,V.T.K., Yeo,W., Soong,P.L. and Nasirudeen,A.M.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-AUG-2000) Microbiology, National University of
            Singapore, 5 Science Drive 2, Singapore 117597, Singapore
FEATURES             Location/Qualifiers
     source          1..1052
                     /organism="Bartonella henselae"
                     /strain="ATCC49882"
                     /db_xref="taxon:38323"
                     /db_xref="ATCC:49882"
     gene <1..>1052
                     /gene="tuf"
     CDS <1..>1052
                     /gene="tuf"
                     /note="mediates elongation of amino acid chain during
                     protein synthesis"
                     /codon_start=1
                     /transl_table=11 /product="elongation factor EF-Tu"
                     /protein_id="AAG24621.1"
                     /db_xref="GI:10945627"
                     /translation="GTIGHVDHGKTSLTAAITKYFGEFKAYDQIDAAPEERARGITIS
                     TAHVEYETEKRHYAHVDCPGHADYVKNMITGAAQMDGAILVVSAADGPMPQTREHILL
                     ARQVGVPAIVVFLNKVDQVDDAELLELVELEVRELLSKYDFPGDDIPIVKGSALAALE
                     DKDKSIGEDAVRLLMSEVDNYIPTPERPVDQPFLMPIEDVFSISGRGTVVTGRVERGV
                     IKVGEEVEIIGIRPTSKTTVTGVEMFRKLLDQGQAGDNIGALLRGIDREGIERGQVLA
                     KPASVTPHTRFKAEAYILTKDEGGRHTPFFTNYRPQFYFRTTDVTGIVTLPEGTEMVM
                     PGDNVAMDVSLIVPIA"
BASE COUNT      253 a    178 c    292 g    329 t
ORIGIN
        1 ggtacgattg gtcacgttga ccatgggaag acctcgttga cggcagcgat tacgaaatat
       61 tttggtgaat ttaaggccta tgaccaaatt gatgcagcgc ctgaggagcg tgcacgtgga
      121 attactattt ctacagcgca tgttgaatat gaaacagaga agcggcatta tgcacatgtt
      181 gattgtccag gtcacgcgga ttatgtgaag aacatgatca cgggcgcggc gcaaatggat
      241 ggtgcgattt tggttgtttc agctgctgat ggtccgatgc ctcaaacacg tgagcatatt
      301 cttcttgccc gtcaggttgg tgttccagcg attgttgttt ttcttaataa ggttgatcag
      361 gttgatgatg ctgagctttt ggagcttgtt gagcttgaag ttcgggagtt attgtcgaaa
      421 tatgattttc caggagacga tattccgatc gttaaaggtt ctgctttggc agcgcttgaa
      481 gataaagata aaagcattgg tgaagatgcg gttcgtcttt tgatgagtga agttgataat
      541 tatataccga cgcctgaacg tcctgttgat cagccgtttt tgatgccaat tgaagatgtt
      601 ttttcgattt cgggtcgtgg aactgttgtg acgggtcgtg ttgagcgtgg tgttattaag
      661 gttggtgaag aagttgagat tatcggcatt cgtccaactt ctaagacaac agttacaggg
      721 gttgaaatgt tccgcaagct tttagatcag gggcaagcgg gtgataatat tggagcgctg
      781 cttcgtggta ttgatcgtga agggattgag cgtggacaag ttttggcgaa gcctgcttcg
      841 gttacacctc atacgagatt taaagcagag gcttacattt tgacgaaaga tgaaggtggt
      901 cgtcatactc catttttcac gaattatcgt cctcagtttt atttccgtac tacggatgta
      961 acgggaattg ttacgcttcc agaaggtaca gagatggtta tgcctggtga taatgttgct
     1021 atggatgtct ctctgattgt tccaattgcc at
//
```