Εργασία 4η

- Κουντουρόγιαννης Δημήτριος
- A.M. : 02399
- GitHub : dkountour02399
- https://github.com/dkountour02399/erg4.git

- Βιοπληροφορική ΙΙ, ΣΤ' εξάμηνο

Πρόγραμμα σε Perl:

```
#!/usr/bin/perl
use strict;
use warnings;

# Ρυθμίσεις αρχικές για τη προσομοίωση ακολουθιών DNA
my $num_seq = shift || 1000;        # Πλήθος τυχαίων ακολουθιών DNA (default: 1000)
my $seq_len = shift || 1_000_000;   # Μήκος κάθε ακολουθίας (default: 1.000.000 βάσεις)
# Έβαλα shift για να επιταχύνω τις δοκιμές όπως (perl erg4.pl 10 10000)

# Ορισμός κατανομών βάσεων
my @katanomes = (
  { name => 'iso',       freqs => { A=>0.25, T=>0.25, G=>0.25, C=>0.25 } },
#"ισοπίθανες" βάσεις
  { name => 'at30_gc70', freqs => { A=>0.15, T=>0.15, G=>0.35, C=>0.35 } },
#A+T=30%  G+C=70%
  { name => 'gc30_at70', freqs => { A=>0.35, T=>0.35, G=>0.15, C=>0.15 } },
#A+T=70%  G+C=30%
);
# Πίνακας κωδικονίων λήξης
my %stop = map { $_=>1 } qw(TAA TAG TGA);

# Συμπληρωματική Αλυσίδα DNA (complement DNA sequences)
my %comp = (A=>'T', T=>'A', G=>'C', C=>'G');

# Συνάρτηση για reverse complement (για εύρεση ORFs και στην άλλη αλυσίδα)
sub revcomp {
    my $seq = shift;
    $seq = reverse $seq;
```

```
    $seq =~ tr/ACGT/TGCA/;
    return $seq;
}

# Κύρια διαδικασία ανά κατανομή βάσεων
for my $k (@katanomes) {
    print "\n=== Κατανομή: $k->{name} ===\n";
    my @lengths;  # Λίστα για τα μήκη των ORFs

    # Προσομοίωση DNA ακολουθιών
    for (1..$num_seq) {
        my $dna = '';
        for (1..$seq_len) {
            my $r = rand();
            $dna .= $r < $k->{freqs}->{A} ? 'A'
                  : $r < $k->{freqs}->{A} + $k->{freqs}->{T} ? 'T'
                  : $r < $k->{freqs}->{A} + $k->{freqs}->{T} + $k->{freqs}->{G} ? 'G'
                  : 'C';
        }

        # Εύρεση ORFs και στις δύο κατευθύνσεις
        foreach my $strand ($dna, revcomp($dna)) {
            for my $frame (0,1,2) {  # 3 δυνατά reading frames ανά κατεύθυνση
                for (my $i = $frame; $i < length($strand)-2; $i += 3) {
                    if (substr($strand,$i,3) eq 'ATG') {
                        for (my $j = $i+3; $j < length($strand)-2; $j += 3) {
                            my $codon = substr($strand,$j,3);
                            if ($stop{$codon}) {
                                push @lengths, $j+3-$i;  # Καταγραφή μήκους ORF
                                last;
                            }
                        }
                    }
                }
            }
        }
    }

    # Στατιστικά ανάλυσης
```

```
    my $n = scalar @lengths;
    my $sum = 0; $sum += $_ for @lengths;
    my $mean = $n ? $sum/$n : 0;
    my $sq = 0; $sq += ($_-$mean)**2 for @lengths;
    my $var = $n>1 ? $sq/($n-1) : 0;

    printf "Σύνολο ORFs: %d\nΜέσος όρος: %.2f\nΔιασπορά: %.2f\n", $n, $mean, $var;

    # Υπολογισμός για Ιστόγραμμα
    my ($min,$max) = ($lengths[0], $lengths[0]);
    for (@lengths) {
        $min = $_ if $_ < $min;
        $max = $_ if $_ > $max;
    }

    my $bins = 50;

    # Αν όλα τα μήκη είναι ίδια, δείξε μια μπάρα και προχώρα
    if ($max == $min) {
        print "\n Ιστόγραμμα μηκών ORFs: \n";
        printf "%6s - %6s | %6d %s \n", $min, $max, $n, '*' x 50;
        next;
    }

    my $bin_size = ($max-$min)/$bins;
    my @hist = (0) x $bins;
    $hist[int(($_-$min)/$bin_size)]++ for @lengths;

    # Εμφάνιση Ιστογράμματος σε ASCII
    print "\n Ιστόγραμμα μηκών ORFs: \n";
    for my $i (0..$bins-1) {
        my $low  = sprintf("%.0f", $min + $i*$bin_size);
        my $high = sprintf("%.0f", $min + ($i+1)*$bin_size);
        my $bar  = '*' x int(($hist[$i]/$n)*50 + 0.5);
        printf "%6s - %6s | %6d %s\n", $low, $high, $hist[$i], $bar;
    }
}
```

Τρέχοντας τον κώδικα πήρα:

```
=== Κατανομή: iso ===
Σύνολο ORFs: 31263466
Μέσος όρος: 67.00
Διασπορά: 3905.36

 Ιστόγραμμα μηκών ORFs:
     6 -     29 | 9969302 ****************
    29 -     52 | 6792070 ***********
    52 -     74 | 4141500 *******
    74 -     97 | 3303600 *****
    97 -    120 | 2013358 ***
   120 -    143 | 1607024 ***
   143 -    166 | 1095589 **
   166 -    188 | 668493 *
   188 -    211 | 532708 *
   211 -    234 | 326273 *
   234 -    257 | 259614
   257 -    280 | 176495
   280 -    302 | 107690
   302 -    325 |  86244
   325 -    348 |  52505
   348 -    371 |  41820
   371 -    394 |  28557
   394 -    416 |  17003
   416 -    439 |  13706
   439 -    462 |   8471
   462 -    485 |   6693
   485 -    508 |   4680
   508 -    530 |   2819
   530 -    553 |   2320
   553 -    576 |   1397
   576 -    599 |   1146
   599 -    622 |    767
   622 -    644 |    491
   644 -    667 |    359
   667 -    690 |    251
   690 -    713 |    170
   713 -    736 |    100
   736 -    758 |     78
   758 -    781 |     55
   781 -    804 |     41
   804 -    827 |     24
   827 -    850 |     19
   850 -    872 |     10
   872 -    895 |      8
   895 -    918 |      3
   918 -    941 |      1
   941 -    964 |      4
   964 -    986 |      2
   986 -   1009 |      2
  1009 -   1032 |      0
  1032 -   1055 |      1
  1055 -   1078 |      1
  1078 -   1100 |      1
  1100 -   1123 |      0
  1123 -   1146 |      0
```

```
=== Κατανομή: at30_gc70 ===
Σύνολο ORFs: 15744426
Μέσος όρος: 159.80
Διασπορά: 24093.94

 Ιστόγραμμα μηκών ORFs:
     6 -     56 | 4405226 **************
    56 -    106 | 3177201 **********
   106 -    156 | 2167111 *******
   156 -    205 | 1678111 *****
   205 -    255 | 1209451 ****
   255 -    305 | 825342 ***
   305 -    355 | 638419 **
   355 -    405 | 438315 *
   405 -    455 | 337473 *
   455 -    505 | 243211 *
   505 -    554 | 165800 *
   554 -    604 | 129007
   604 -    654 |  92695
   654 -    704 |  62896
   704 -    754 |  48649
   754 -    804 |  33370
   804 -    854 |  25853
   854 -    903 |  18556
   903 -    953 |  12754
   953 -   1003 |   9886
  1003 -   1053 |   7069
  1053 -   1103 |   4821
  1103 -   1153 |   3648
  1153 -   1203 |   2478
  1203 -   1252 |   2015
  1252 -   1302 |   1446
  1302 -   1352 |   1011
  1352 -   1402 |    754
  1402 -   1452 |    495
  1452 -   1502 |    399
  1502 -   1552 |    288
  1552 -   1602 |    154
  1602 -   1651 |    147
  1651 -   1701 |     94
  1701 -   1751 |     71
  1751 -   1801 |     70
  1801 -   1851 |     40
  1851 -   1901 |     28
  1901 -   1951 |     29
  1951 -   2000 |      7
  2000 -   2050 |      9
  2050 -   2100 |      6
  2100 -   2150 |      5
  2150 -   2200 |      5
  2200 -   2250 |      2
  2250 -   2300 |      4
  2300 -   2349 |      1
  2349 -   2399 |      2
  2399 -   2449 |      1
  2449 -   2499 |      0
```

```
=== Κατανομή: gc30_at70 ===
Σύνολο ORFs: 36749147
Μέσος όρος: 40.67
Διασπορά: 1306.07

 Ιστόγραμμα μηκών ORFs:
     6 -     20 | 12483736 ****************
    20 -     33 | 8238884 ***********
    33 -     47 | 4528213 ******
    47 -     61 | 3905431 *****
    61 -     74 | 2145558 ***
    74 -     88 | 1847236 ***
    88 -    102 | 1017104 *
   102 -    115 | 877828 *
   115 -    129 | 578651 *
   129 -    143 | 317960
   143 -    156 | 275026
   156 -    170 | 150869
   170 -    184 | 130262
   184 -    198 |  71281
   198 -    211 |  61863
   211 -    225 |  33526
   225 -    239 |  29248
   239 -    252 |  18968
   252 -    266 |  10735
   266 -    280 |   8955
   280 -    293 |   4999
   293 -    307 |   4441
   307 -    321 |   2339
   321 -    334 |   2091
   334 -    348 |   1108
   348 -    362 |    970
   362 -    375 |    605
   375 -    389 |    358
   389 -    403 |    315
   403 -    416 |    169
   416 -    430 |    143
   430 -    444 |     82
   444 -    457 |     59
   457 -    471 |     41
   471 -    485 |     20
   485 -    498 |     29
   498 -    512 |     12
   512 -    526 |     10
   526 -    540 |      6
   540 -    553 |      4
   553 -    567 |      4
   567 -    581 |      3
   581 -    594 |      2
   594 -    608 |      0
   608 -    622 |      2
   622 -    635 |      0
   635 -    649 |      0
   649 -    663 |      0
   663 -    676 |      0
   676 -    690 |      0
```

Συμπέρασμα:

Για κατανομή "ισοπίθανη": Σύνολο ORFs: 31263466 | Μέσος όρος: 67.00 | Διασπορά: 3905.36
Για κατανομή "A+T=30%" : Σύνολο ORFs: 15744426 | Μέσος όρος:159.80| Διασπορά: 24093.94
Για κατανομή "A+T=70%" : Σύνολο ORFs: 36749147 | Μέσος όρος: 40.67 | Διασπορά: 1306.07

Στην ''Ισοπίθανη'' κατανομή (25% κάθε βάσης) έχουμε ουδέτερη συμπεριφορά, ενδιάμεσο πλήθος και μέγεθος ORFs.

Στην ''GC-πλούσια'' (A+T=30%, G+C=70%) έχουμε λιγότερα ORFs, αλλά πολύ μεγαλύτερα κατά μέσο όρο, λόγω σπανιότητας των stop codons.

Στην ''AT-πλούσια'' (A+T=70%, G+C=30%) έχουμε τα περισσότερα ORFs, αλλά πολύ μικρού μήκους, γιατί τα stop codons (πλούσια σε A/T) «κόβουν» νωρίς τα πλαίσια ανάγνωσης.

Συμπέρασμα κλεισίματος:

Οι αλληλουχίες DNA με υψηλό AT περιεχόμενο δημιουργούν πολλά, αλλά σύντομα, ORFs, ενώ αυτές με υψηλό GC οδηγούν σε λιγότερα, αλλά πολύ μακρύτερα, ORFs. Η ισοπίθανη κατανομή, ως ουδέτερη συνθήκη, δίνει αποτελέσματα ενδιάμεσα των δύο άκρων. Αυτά τα ευρήματα υπογραμμίζουν πώς η σύσταση βάσεων επηρεάζει τόσο την έκταση όσο και τον αριθμό ανοιχτών πλαισίων ανάγνωσης, με σημαντικές βιολογικές συνέπειες στη δομή και λειτουργία του γονιδιώματος.»