Genomics

Freie Universität Berlin, Institut für Informatik
Peter Robinson
Wintersemester 2014/2015

2. Übungsblatt vom 28. Oktober 2014
Diskussion am 6 November 2014

Aufgabe 1.

Geben Sie den Karyotyp für Down-Syndrom auf Grund einer freien Trisomie 21 an

Aufgabe 2.

69,XXX ist eine Form der . . .

Aufgabe 3.

Das Nichttrennen von Schwesterchromatiden durch eine Störung der Metaphase während der Meiose II bezeichnet man als . . .

Aufgabe 4.

Die erste Subsubbande der ersten Subbande der ersten Bande des langen Arms des zwölften Chromosoms

Aufgabe 5.

Ein Prozess mit zwei aufeinander folgenden Zellteilungen, aber nur einer Runde der DNA-Replikation

Aufgabe 6.

Eine Deletion auf dem langen Arm des zweiten Chromosoms, welche von der zweiten Bande (3. Subbande) zur dritten Bande (2. Subbande) reicht.

Aufgabe 7.

Wofür steht H3K16ac?

Kernlokalisationssequenz (praktische Übung #1)

Die Kernlokalisationssequenz (Nuclear localisation sequence, NLS) ist ein kurzes Sequenzmotiv vieler für den Kern bestimmter Proteine. Im Folgenden werden drei Beispiele gezeigt.

Protein	Sequenzmotiv
Glucocorticoidrezeptor	RKxxxxxxxxxRKxKK
Androgenrezeptor	RKxxxxxxxxxRKxKK
p53	KRxxxxxxxxxxxKKK

Die Suche mittels regulärer Ausdrücke wird in der Bioinformatik vielfach angewendet. Für diese Aufgabe sollen Sie auf der prosite-Webseite nach den hier angegebenen regulären Ausdrücken bei diesen drei Proteinen suchen (`http://prosite.expasy.org/scanprosite/`). Lesen Sie bitte dort über die Syntax der regulären Ausdrücke bei Prosite[1] Zum Beispiel:

Ausdruck	Bedeutung
x(3)	x-x-x (drei beliebige Aminosäuren)
x(2,4)	x-x oder x-x-x oder x-x-x-x (2 bis 4 beliebige Aminosäuren)
A(3)	A-A-A (drei Alaninreste)
`[CH]`	ein Cystein oder ein Histidin

Schreiben Sie einen regulären Ausdruck, um die NLS bei p53 zu identifizieren. Hierzu brauchen Sie auch noch die Accessionnummer von p53 bei der UniProtKB (`P53_HUMAN`), diese wird im linken Fenster eingetragen. Tragen Sie Ihren regulären Ausdruck im rechten Fenster ein.

- Geben Sie nun die genaue Aminosäuresequenz der NLS in p53 sowie deren Position (Aminosäurenzahl) innerhhalb des Proteins an.
- Schreiben Sie nun einen einzigen regulären Ausdruck, welcher alle drei oben genannten NLS bei den entsprechenden Proteinen findet (hierzu müsssen Sie nach den Accessionnummern der Proteine in der UniProtKB suchen).

Dot-Plots und Low-Copy-Repeats (praktische Übung #2)

In dieser Aufgabe wollen wir die Skriptsprache R verwenden, um die Sequenzähnlichkeit zwischen einem Gen und einem entsprechenden Pseudogen mittels Dot-Plot zu visualisieren. Wir werden R für mehrere Aufgaben im Verlauf des Kurses verwenden. R ist relativ einfach zu lernen, hat aber eine Reihe von Besonderheiten für Java- oder C-Programmierer. Wir werden die für diese Aufgabe erfoderlichen Grundlagen im Folgenden erklären.

Installation

R kann unter allen Betriebssystemen kostenlos installiert werden. Für weitere Einzelheiten s. `http://http://www.r-project.org/`. R kann durch zahlreiche Pakete ergänzt werden, die jedoch nachinstalliert werden müssen. Für die Übung werden wir das Paket `Biostrings` installieren. Unter Windows und Mac können Paketmanager verwendet werden (s. die Dokumentation für Ihr System). Unter allen Betriebssystemen können Sie das Paket mit dem folgenden Kommando installieren:

```
> install.packages("Biostrings")
```

Es erscheint ein Dialogfenster, Sie können den Server angeben, von dem Sie die Library beziehen wollen (in der Regel innerhhalb Deutschlands). Unter Umständen müssen Sie administrative Rechte haben, um Pakete zu installieren.

[1] der Verweis “pattern(s)” bringt Sie zur Seite: `http://prosite.expasy.org/scanprosite/scanprosite-doc.html#pattern_syntax`.

FASTA Sequenzen einlesen

Laden Sie bitte zwei FASTA Sequenzen von der `NCBI Nucleotide Datenbank` herunter[2]:

- Homo sapiens RAD17 homolog (S. pombe) (RAD17), transcript variant 1, mRNA (Suche nach `NM_133338.2`)
- Homo sapiens RAD17 homolog (S. pombe) pseudogene 2 (RAD17P2) on chromosome 13 (Suche nach `NG_002928.5`)

Das Pseudogen von RAD17 weist eine etwa 91% Identität über einen Teilabschnitt des Gens auf. Bemerken Sie, dass man das Format "FASTA" einstellen muss. Bitte fügen Sie beider Seqeunzen zu einer einzigen FASTA-Datei hinzu, welche Sie "RAD17.fa" benennen sollen.

Wir starten R von einem Verzeichnis, in dem die Datei "RAD17.fa" zu finden ist. Eventuell müssen Sie den Pfad ändern, damit R die Datei auf Ihrem System finden kann.

```
library(Biostrings)

filename <- "RAD17.fa"
dna <- readFASTA(file = filename, strip.descs=TRUE)
seq <- sapply(dna, function(x) x$seq)
s1 <- seq[1]
s2 <- seq[2]
```

Sie sollen jetzt prüfen, dass die Variablen `s1` und `s2` die entsprechenden DNA-Sequenzen aufweisen. Es folgt nun der Code derenFUnktion `makeDotPlot`, womit ein Dot-Plot erzeugt werden kann.

```
makeDotPlot <- function(seq1,seq2,windowsize=7)
  {
    dotsize <- 1
    length1 <- nchar(seq1) - windowsize
    length2 <- nchar(seq2) - windowsize

    x <- 1
    y <- 1
     plot(x,y,ylim=c(1,length2),xlim=c(1,length1),col="white",
          xlab = "Sequence 1", ylab = "Sequence 2")
    ## erzeugt einen leeren Plot
    ## Die For-Schleifen vergleichen alle Subsequenzen
    ##    der Laenge windowsize
     for (i in 1:length1)
     {
        sseq1 <- substr(seq1,i,i+windowsize)
        for (j in 1:length2)
        {
           sseq2 <- substr(seq2,j,j+windowsize)
           if (sseq1 == sseq2)
           {
              points(x=i,y=j,cex=dotsize,col="blue",pch=20)
           }
        }
```

[2] http://www.ncbi.nlm.nih.gov/nuccore/

```
    }
  title(main = paste("Dot plot, window size = ",windowsize),
        sub = NULL, line = NA, outer = FALSE)
}
```

- Interpretieren Sie das Ergebnis!