Biostatistică și Bioinformatică.

Lab 2: Prelucrări statistice asupra secvențelor de nucleotide/ aminoacizi. Teste statistice.

---

1. Pachete R pentru bioinformatică

- Bioconductor (www.bioconductor.org)
- SeqinR (http://pbil.univ-lyon1.fr/software/seqinr/home.php?lang=eng)

Pachetele Bioconductor si SeqinR conțin funcții care permit efectuarea unor prelucrări standard asupra secvențelor biologice (citire fișiere în formate specifice, analize statistice, aliniere, analiza datelor referitoare la expresia genică etc).

Pentru utilizarea unui pachet R acesta trebuie instalat și încărcat:

- Instalarea unui pachet R: Packages -> Install Packages … (se selectează site-ul de unde se descarcă și ulterior pachetul care trebuie instalat)
- Incărcarea unui pachet R: Packages -> Load Package … (se selectează pachetul) sau direct cu comanda library("nume pachet") (Ex: library("seqinr"))

Obs: pachetul SeqinR poate fi incărcat folosind procedura anterioară însă pentru încărcarea setului de pachete Bioconductor este necesară parcurgerea unei proceduri diferite:

```
>source("http://bioconductor.org/biocLite.R")
>biocLite()    # asigura incarcarea unui subset de pachete de baza
```

Pentru instalarea ulterioară a unui pachet din Bioconductor (de exemplu, Biostrings care conține funcții de prelucrare a secvențelor) se folosește comanda biocLite("nume pachet"). De exemplu prin

```
>biocLite("Biostrings")
```

se instalează pachetul Biostrings. Utilizarea pachetului necesită și încărcarea prin library("Biostrings") sau prin Packages->Load package

Obs. Există numeroase pachete R pentru bioinformatică dedicate unor prelucrări specifice dar care conțin și funcții cu caracter general. Un astfel de exemplu este pachetul ape (http://ape-package.ird.fr/) proiectat pentru analiza filogenetică însă care conține și funcții simple pentru accesarea secvențelor din GenBank. De exemplu, după încărcarea pachetului ape (library(ape)) prin read.Genbank(c(identificator),as.character=TRUE) se returnează secvența sub forma unui vector de caractere.

2. Analiza statistică simplă a unei secvențe de nucleotide

In secvențele ADN porțiunile care codifică gene diferă de porțiunile intergenice prin distribuția diferitelor tipuri de nucleotide. Astfel, secvențe cu procent mare de nucleotide A sau T corespund de regulă porțiunilor intergenice (introni) pe când secvențe cu procent mare de nucleotide C sau G indică posibila prezență a genelor. Pentru a determina frecvențele de apariție a diferitelor tipuri

de nucleotide respectiv a diferitelor tipuri de dimeri (perechi de nucleotide diferite) se pot parcurge următoarele etape:

- Preluarea secvenței dintr-un fișier și stocarea într-un vector de caractere
- Calculul frecvențelor de apariție a fiecărui tip de nucleotidă
- Construirea vectorului de dimeri corespunzători secvenței
- Calculul frecvențelor de apariție a fiecărui tip de dimer

Folosind pachetul SeqinR toate aceste prelucrări pot fi realizate foarte simplu:

```
date=read.fasta("s1.fa")          # se presupune ca fisierul se afla in folderul MyDocuments
(Obs: dacă e în alt folder trebuie specificată calea completă sau se setează directorul de lucru
      prin File->ChangeDir)
secv=date[[1]]                    # preluarea secventei de nucleotide
frecvADN=table(secv)              # construirea tabelului de frecvente
                                  # (funcția table este cea standard din R)
frecvDimeri=count(secv,2)         # determinarea frecventelor dimerilor
(Obs: functia count este din SeqinR și permite determinarea frecvențelor de apariție ale
     "cuvintelor" cu lungimea specificată în al doilea parametru al funcției
```

3. Conversia unei secvențe de nucleotide într-o secvență de aminoacizi

Corespondența dintre un triplet de nucleotide și un aminoacid este descrisă de codul genetic. Pachetul Biostrings din Bioconductor conține o structură numită GENETIC_CODE care asigură maparea corespunzătoare dintre triplete de nucleotide (specificate ca șiruri de 3 caractere) și simboluri corespunzătoare aminoacizilor.

Un exemplu de prelucrare care realizează această conversie (folosind pachetul Biostrings dar fără a folosi funcțiile din pachetul SeqinR) este:

```
# conversie secventa nucleotide in secventa de aminoacizi
fp=file("s1.fa",open="r")       # deschidere în citire a unui fisier text
linii=readLines(fp)             # citirea linie cu linie a unui fisier text
secv=linii[2]                   # concatenarea liniilor intr-un singur sir
for(i in 3:length(linii)-1) secv=paste(secv,linii[i],sep="")
# fragmentarea sirului intr-un vector de subsiruri ce contin triplete de nucleotide
rez=substring(secv, seq(1, width(secv)-2, 3),seq(3,width(secv),3))
# conversia tripletelor de nucleotide folosind codul genetic si concatenarea
secvaa=paste(GENETIC_CODE[rez])
```

4. Distribuții de probabilitate in R

In R sunt implementate principalele tipuri de repartiții utilizate în biostatistică:

- *Binomială.* Identificator: binom; Parametri: n (nr repetari), p (probabilitate de success)
- *Normală.* Identificator: norm; Parametri: m (media), s (abaterea standard)
- *Chi-pătrat.* Identificator: chisq; Parametri: df (număr de grade de libertate)

- *Fisher*. Identificator: f; Parametri: df1 (număr de grade de libertate numărător) , df2 (număr de grade de libertate numitor)
- *Student*. Identificator: t; Parametri: df (număr de grade de libertate)

Pentru fiecare dintre aceste distribuții sunt implementate funcții pentru:

- Calculul densității de probabilitate:
  - d<identificator>(x,<lista param>) returnează valoarea probabilității ca variabila aleatoare să aibă valoarea x (P(X=x)) sau valoarea corespunzătoare a funcției de densitate de probabilitate (în cazul variabilelor aleatoare continue)
  - Exemplu: dbinom(11,30,0.25) sau dnorm(1.6,0,1)

- Calculul funcției de repartiție:
  - p<identificator>(x,<lista param>) returnează valoarea probabilității ca variabila aleatoare să aibă cel mult valoarea x (P(X<=x)) sau valoarea corespunzătoare a funcției de reparție
  - Exemplu: pbinom(11,30,0.25) sau pnorm(1.68,0,1)

- Calculul cuantilelor (sau valorilor critice):
  - q<identificator>(alfa,<lista param>) returnează valoarea x care P(X<=x)=alfa
  - Exemplu: qbinom(0.95,30,0.25) sau qnorm(0.95,0,1)
  - Obs: aceste funcții pot fi utilizate pentru determinarea valorilor critice

- Generarea de valori aleatoare în conformitate cu o anumită distribuție:
  - r<identificator>(nr,<lista param>) returnează nr valori aleatoare generate în concordanță cu distribuția corespunzătoare
  - Exemple:
    - runif(10,min=0,max=1) generează 10 valori uniform distribuite în intervalul [0,1)
    - rnorm(100,0,1) generează 100 de valori având distribuția normală standard

Exercitiul 1.

a) Să se descarce din GenBank secvența ADN având identificatorul AF012130 (o secventa care intervine în controlul sintezei de insulină) și să se determine frecvența de apariție a nucleotidelor și a dimerilor.

Indicație: După conectarea la http://www.ncbi.nlm.nih.gov se specifică identificatorul și din pagina cu rezultate se selectează Nucleotide: DNA and RNA sequences după care se salvează secvența în format Fasta folosind Send.

b) Să se genereze o secvență aleatoare având aceeași dimensiune cu secvența anterioară, să se determine aceleași frecvențe și să se compare cu rezultatele anterioare.

Indicatie: pentru generarea unei secvențe aleatoare conținând L nucleotide (șir aleator peste alfabetul {a,c,g,t}) se poate proceda în felul următor:

- Se generează L valori aleatoare uniform distribuite în mulțimea {1,2,3,4}

Exemplu (pentru L=100):

```
>date=as.integer(runif(100,min=1,max=5))
                                # runif simulează o variabilă aleatoare cu repartitia uniforma
```

- Valorile generate la pasul anterior vor fi folosite ca indici în alfabetul corespunzător nucleotidelor (reprezentat ca un vector cu 4 elemente de tip caracter)

```
>alfabet=c("a","c","g","t")
>randADN=alfabet[date]
```

5.Analiza varianței (ANOVA – Analysis Of Variance)

Considerăm următoarea problemă: *se cunosc valori ale capacității respiratorii ale unor persoane grupate în următoarele categorii:*

Categoria (nr categorii=k)	***Volum eșantion (n)***	***Medie***	***Abatere standard***
Nefumător	*200*	*3.17*	*0.74*
Fumător pasiv	*200*	*2.72*	*0.71*
Fumător (nr mic de țigări)	*200*	*2.63*	*0.73*
Fumător (nr mediu de țigări)	*200*	*2.29*	*0.70*
Fumător (nr mare de țigări)	*200*	*2.12*	*0.72*

Se poate afirma pe baza acestor date că există diferențe semnificative între capacitățile pulmonare ale celor 5 categorii de persoane?

Etapele care trebuie parcurse sunt:

a) Se stabilește H_0: "nu există diferențe semnificative între cele 5 categorii"
b) Se calculează varianța medie, *Var(intra)*, din cadrul fiecărei categorii (media valorilor de pe ultima coloană)
c) Se calculează variația între categorii, *Var(inter)*, - varianța corespunzătoare valorilor medii de pe coloana a treia
d) Se calculează statistica F=*Var(inter)/Var(intra)*
e) Se determină valoarea critică corespunzătoare repartiției Fisher(k-1,n-k) și nivelului de semnificație dorit (in R se specifică qf(1-alpha,k-1,n-k)
f) Dacă valoarea statisticii depășește valoarea critică atunci ipoteza nulă se respinge

Exercitiu 2. Să se implementeze in R procedura de mai sus si sa se aplice datelor din tabel.

Indicație. Un exemplu de implementare este:

```
anova=function(medii,stdev,n,alpha)
{k=length(medii)
 varIntra=mean(stdev^2)
 medieGlobala=mean(medii)
 varInter=sum((medii-medieGlobala)^2)
 return(list("statistica",varInter/varIntra,"valoarecritica",qf(1-alpha,df1=k-1,df2=n-k)))
}

medii=c(3.17,2.72,2.63,2.29,2.12)
```

```
stdev=c(0.74,0.71,0.73,0.70,0.72)
rez=anova(medii,stdev,200,0.05)
```

Obs: Dacă se cunoaşte tabelul de date atunci se poate utiliza direct funcţia aov pentru analiza varianţei (vezi exemplul din fişierul migraineANOVA.R)

6. Verificarea unei ipoteze statistice referitoare la distribuţia nucleotidelor într-o secvenţă ADN.

Considerăm următoarea problemă: *fiind dată o secvenţă de nucleotide să se stabilească dacă secvenţa este generată aleator sau nu (distribuţia nucleotidelor este uniformă). Problema este echivalentă cu cea a verificării concordanţei dintre distribuţia secvenţei analizate şi distribuţia uniformă. Ipoteza nulă este: "distribuţia secvenţei coincide cu distribuţia uniformă" sau "secvenţa este aleatoare". In acest scop se poate folosi testul de concordanţă chi-pătrat.*

Etapele care trebuie parcurse sunt:

a) se determină frecvenţa de apariţie a fiecărei nucleotide (se foloseşte table sau count din SeqinR) şi se construieşte vectorul cu frecvenţe absolute [f_A f_C f_G f_T]

b) se calculează statistica:

$$S=(f_A-n*e_A)^2/(n*e_A)+(f_C-n*e_C)^2/(n*e_C)+(f_G-n*e_G)^2/(n*e_G)+(f_T-n*e_T)^2/(n*e_T)$$

unde n reprezintă lungimea secvenţei iar $e_A=e_C=e_G=e_T=0.25$ reprezintă probabilităţile corespunzătoare unei distribuţii uniforme.

c) se calculează valoarea critică a repartiţiei chi-pătrat pentru 3 grade de libertate şi nivelul de semnificaţie 0.05 (se poate folosi qchisq(0.95,3))

d) dacă valoarea S este mai mare decăt valoarea critică determinată la pasul c) atunci ipoteza nulă (secvenţa este generată uniform aleator) se respinge, adică nu se poate considera că secvenţa este generată aleator.

Exerciţiul 3:

1. Să se verifice ipoteza de uniformitate pentru o secvenţă încărcată din GenBank (de exemplu secvenţa cu identificatorul "AF012130.1")
2. Să se verifice ipoteza de uniformitate în cazul unei secvenţe de aceeaşi lungime dar generată aleator (folosind varianta de la exerciţiul 1) .

Indicaţie. Un exemplu simplu de implementare a testului este:

```
testUniform=function(secv, nivelSemnificatie)
{f=table(secv)
 n=length(secv)
 stat=((f[["a"]]-n/4)^2+(f[["c"]]-n/4)^2+(f[["g"]]-n/4)^2+(f[["t"]]-n/4)^2)/(n/4)
 if (stat>qchisq(1-nivelSemnificatie,3))
       {return("Se respinge ipoteza ca repartitia nucleotidelor este uniforma")}
 else
       {return("Nu se respinge ipoteza ca repartitia nucleotidelor este uniforma")}
}
```

Obs. In R testul chi-pătrat pentru verificarea uniformității poate fi aplicat simplu folosind: chisq.test(secventa,p=c(0.25,0.25,0.25,0.25)). Funcția returnează p-valoarea corespunzătoare testului (dacă p-valoarea este mai mica decât nivelul de semnificație al testului atunci ipoteza nulă este respinsă). Dacă se dorește verificarea ipotezei privind concordanța cu o altă distribuție discrete (nu neapărat uniformă) atunci se poate folosi chisq.test(secventa,p=vector_probabilitati).

7. Verificarea ipotezei de independență a nucleotidelor succesive.

Se pune problema să se verifice dacă se poate accepta ipoteza că nucleotidele succesive dintr-o secvență ADN sunt independente. In acest scop se folosește testul chi-pătrat pentru verificarea independenței. Etapele care trebuie parcurse la aplicarea testului sunt:

a) se determină tabelul frecvențelor (tabelul de contingență) corespunzătoare perechilor de nucleotide aflate pe poziții consecutive (pozițiile (i,i+1)):

	Pozitia i+1					
Pozitia i		A	C	G	T	
	A	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{1*}
	C	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{2*}
	G	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{3*}
	T	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{4*}
		Y_{*1}	Y_{*2}	Y_{*3}	Y_{*4}	Y

De exemplu Y_{11} reprezintă numărul situațiilor în care sunt două nucleotide de tip A consecutive, Y_{24} reprezintă numărul cazurilor în care o nucleotidă de tip C este urmată de o nucleotidă de tip T. Y_{i*} reprezintă suma elementelor de pe linia i iar Y_{*j} reprezintă suma elementelor de pe coloana j. Y este suma tuturor frecvențelor din tabel.

Pentru determinarea frecvențelor Y_{ij} se poate folosi count(secv,2). Pe baza valorilor returnate de această funcție se construiește matricea de mai sus.

b) se calculează statistica

$$T = \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(Y_{ij} - E_{ij})^2}{E_{ij}}$$

$$E_{ij} = \frac{Y_{i*}Y_{*j}}{Y}$$

c) se determină valoarea critică a repartiției chi-pătrat pentru 9=(4-1)*(4-1) grade de libertate și nivelul de semnificație 0.05 (se folosește qchisq(0.95,9))

d) dacă statistica T este mai mare decât valoarea critică atunci ipoteza nulă (faptul ca nucleotidele succesive sunt independente) este respinsă.

Exercitiul 4: Să se implementeze procedura de verificare descrisă mai sus și să se verifice ipoteza de independență atât în cazul unei secvențe încărcate din GenBank (de exemplu secvența AF012130) cât și în cazul unei secvențe generate aleator.

Indicație. Vezi exemplul din fisierul testIndependenta.R

Tema. Implementați în R o funcție care permite aplicarea testului semnelor (detalii în slide-uri Curs3) pentru a analiza efectul unui tratament pe baza unor măsurători efectuate înainte și după aplicarea tratamentului asupra unui set de pacienți.

Date de intrare: doi vectori de aceeaşi dimensiune conținând valorile măsurătorilor
nivelul de semnificație al testului
Date de iesire: valoarea statisticii corespunzătoare testului semnelor
decizia (se respinge sau nu ipoteza că tratamentul nu are efect)

Exemplu date de test:
V1=[2 0 0 0 2 3 3 3 0 2 1 1 0 0 3 0 0]
V2=[2 0 3 1 5 2 2 10 0 4 1 4 0 0 4 1 4]

Cele două seturi de date reprezintă valoarea unui parametru înainte de aplicarea unui tratament (V1) respectiv după aplicarea tratamentului (V2). Scopul urmărit este să se decidă dacă diferențele dintre cele două seturi sunt semnificative sau nu (pentru nivelul de semnificație egal cu 0.05).