Biostatistică și Bioinformatică.

Lab 3: Identificarea șabloanelor

Șabloanele (motivele) sunt secvențe relativ scurte de nucleotide (sau aminoacizi) care au un anumit rol funcțional (de exemplu marchează începutul unei regiuni codante sau poziții unde se pot "lega" anumite enzime). Identificarea șabloanelor ("motif finding") este o problemă ce poate fi abordată pe diferite nivele de dificultate:

- Varianta 1: se cunoaște șablonul și se urmăreste identificarea tuturor pozițiilor în care acesta apare. Este o problema tipică de "pattern matching" care poate fi rezolvată utilizând algoritmi tradiționali (ex: Knuth-Morris-Pratt, Aho-Corasick).
- Varianta 2: șablonul nu se cunoaște exact însă se știe că respectă o anumită regulă (de exemplu, poate fi descris printr-o expresie regulată). In cazul acesta șabloanele identificate nu au toate aceeași lungime.
- Varianta 3: șablonul are lungime fixă însă poate să apară alterat de mutații; în acest caz problema principală este de a stabili o măsură a similarității între diferite șabloane și de a construi un șablon reprezentativ (șablon consensual). In aceste cazuri nu se știe ce șablon se caută însă se identifică subsecvențele frecvente suficient de similare între ele (și implicit cu șablonul profil).
- Varianta 4: nu se cunoaște nici dimensiunea nici conținutul șablonului. Se caută subsecvențe cu scor mare de similaritate. Este varianta cea mai dificilă.

In continuare sunt prezentate câteva exemple corelate cu primele trei variante.

1. Determinarea prezențelor unui șablon fixat

Problema: se caută toate aparițiile unui șablon cunoscut într-un șir

Pentru problema căutării unui subșir (șablon) într-un șir se pot folosi diferite funcții R. Cele mai comune sunt:

- match(sablon,sir) - face parte din comenzile de bază
- matchPattern(sablon,sir) – face parte din pachetul Biostrings (necesita încărcare prealabilă). Dacă se dorește să se determine doar de câte ori apare sablonul în șir atunci se poate folosi countPattern

Exercitiul 1. Sa se preia din GenBank secvența ADN cu *identificatorul* NC_012920 (sau secventa din fisierul sequence.fasta). Să se determine numarul de apariții ale subsecventeleor 'TATA' (subsecvență cu rol de promoter în procesul de sinteză a proteinelor) respectiv 'TATATA'.

Indicatie. Daca s1 este secvența analizată atunci prin

```
poz = matchPattern("TATA",s1)
```

se determina pozițiile de start ale șablonului, iar prin

```
countPattern("TATA",s1)
```

se determină câte prezențe ale șablonului căutat au fost întâlnite

Exercitiul 2. Să se genereze 30 de secvențe aleatoare, să se determine pentru fiecare dintre ele numărul de apariții ale șablonului "TATA" respectiv "TATATA" și să se calculeze valoarea medie și abaterea standard a numărului de apariții pentru fiecare dintre cele două șabloane.

Indicație. Se scrie o funcție ce conține un ciclu în care se generează 30 de secvențe aleatoare (folosind funcția pentru generarea secvențelor aleatoare descrisă în Lab 2). Pentru calculul mediei se poate folosi funcția **mean** iar pentru calculul abaterii standard se poate folosi funcția **std.**

2. Determinarea șabloanelor care respectă o expresie regulată

Problema: se caută toate aparițiile unui șablon care respectă o expresie regulată
Specificarea expresiilor regulate in R: în specificarea expresiilor regulate se pot utiliza următoarele simboluri și construcții lexicale:

. (punct) – un simbol arbitrar
\w - caracter alfanumeric (litera, cifra sau _)
c? - un simbol care apare o dată sau deloc
c* - succesiune de lungime mai mare sau egala cu 0, constituită prin repetarea simbolului c
c+ - succesiune de lungime mai mare sau egală cu 1, constituită prin repetarea simbolului c
c{m,n} – succesiune cu lungimea cuprinsă între m și n
[c1 c2 c3 …] – unul dintre simbolurile specificate
[^ c1 c2 c3 …] – nici unul dintre simbolurile specificate

In locul unui simbol c se poate specifica **.** (punct) pentru a indica un simbol arbitrar sau (secventa) pentru a specifica o secvență de simboluri.

Pentru a analiza prezența unui șablon specificat printr-o expresie regulată se poate folosi funcțiile:

- grep (expresie regulata, secventa)
- regexpr (expresie regulata, secventa)

Exercitiul 3. Pentru secvența din GenBank și una dintre secvențele aleatoare folosite la exercițiile anterioare să se determine numarul de sabloane de forma:
a) GAGGAGGAG... GAG
b) cel mai lung șablon de forma TA....AT (între TA și AT pot fi orice caractere)

3. Construirea matricii profil, a șablonului consensual și a scorului de potrivire

Se consideră un set de t secvențe ADN, fiecare de lungime L. *Matricea profil* a unei astfel de mulțimi conține 4 linii și L coloane:

Linia 1: pe fiecare poziție se află numărul de aparitii ale nucleotidei 'A' în poziția corespunzătoare în toate secvențele

Linia 2: pe fiecare poziție se află numărul de aparitii ale nucleotidei 'C' în poziția corespunzătoare în toate secvențele

Linia 3: pe fiecare poziție se află numărul de aparitii ale nucleotidei ‘G’ în poziția corespunzătoare în toate secvențele

Linia 4: pe fiecare poziție se află numărul de aparitii ale nucleotidei ‘T’ în poziția corespunzătoare în toate secvențele

Sablonul consensual are L elemente și se construiește punând pe fiecare poziție nucleotida ce apare cel mai frecvent pe poziția corespunzătoare în setul de secvențe.

Scorul de potrivire se obține adunând valorile maxime de pe coloanele matricii profil.

Exercitiul 4. Să se construiască matricea profil, șablonul consensual și scorul de potrivire pentru setul de secvențe:
c(“AGGTACTT”,”CCATACGT”,”ACGTTAGT”,”ACGTCCAT”,”CCGTACGG”})

Indicatie. Vezi fisierul construireProfil.R

Tema 1: Să se modifice funcția de construire a șablonului consensual astfel încât în cazul în care sunt mai multe nucleotide care apar cu aceeași frecvență acestea să apară specificate în șablon ca alternative. De exemplu daca matricea profil este:

A: 2 1 3 0 2 4
C: 1 3 3 4 3 1
G: 2 3 0 2 2 1
T: 2 0 1 1 0 1

atunci șablonul consensual ar trebui să fie: {A|G|T}{C|G}{A|C}CA

4. Determinarea șablonului consensual de scor maxim și a localizării lui în cazul a două secvențe.

Exercitiul 5. Să se determine șablonul consensual de lungime 8 având scor maxim corespunzător secventelor: s(1:50) și s(100:200), unde s este secvența ADN de la Exercitiul 1.

Indicatie: se parcurg secvențele si pentru fiecare pereche de indici de start i și j se calculează matricea profil respectiv scorul de potrivire. Șablonul corespunzător scorului maxim precum și pozițiile de unde începe în fiecare dintre secvențe poate fi determinat cu ajutorul funcției descrise in fisierul gasireSablon.R

Un exemplu de apel este:

```
rez = gasireSablon(substring(s,1,50),substring(s,100,200),8)
```

5. Determinarea șablonului consensual de scor maxim și a localizarii lui în cazul unui numar arbitrar de secvențe folosind algoritmul bazat pe tehnica greedy.

Exercitiul 6. Să se determine sablonul consensual pentru cele 5 exemple de secvențe de la Cursul 5.

Indicatie: Se determină matricea profil și pozițiile șablonului de scor maxim pentru primele două secvențe după care pentru fiecare dintre secvențele următoare se determină poziția șablonului de scor maxim (actualizând adecvat matricea profil). Detalii in fisierul ConsensusGreedy.R

Exemplu de apel pentru secvențele de la curs:

```
set=c("cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgtttccaaccat","agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc","aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt","agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgtacgtataca","ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgttaacgtacgtc")

rez = consensusGreedy(set, 8)
```

Tema 2. Să se implementeze varianta de algoritm (similara algoritmului Consensus) descrisă în Curs 5 (slide 29) .