Biostatistică și Bioinformatică.

Lab 4: Alinierea secvențelor

- **Analiza vizuală a similarității dintre secvențe (matrici de puncte)**
- **Aliniere globală (algoritmul Needleman-Wunsch)**
- **Aliniere locala (algoritmul Smith-Waterman)**

Scopul alinierii a două secvențe este de a estima similaritatea dintre ele. In general secvența analizată este comparată cu secvențe al căror rol este cunoscut. Alinierea este un instrument util pentru căutarea în bazele de date genomice.

Alinierea a două secvențe (de nucleotide sau de aminoacizi), care nu au neapărat aceeași lungime, presupune eventuala inserare de spații (gap-uri) în cadrul secvențelor pentru a maximiza potrivirea dintre ele. Potrivirea dintre două secvențe se măsoară pe baza unor scoruri care cuantifică potrivirea dintre perechi de nucleotide sau aminoacizi și penalizează inserția unor gap-uri. Scorul unei perechi de elemente aliniate (care ocupă aceeași poziție în cadrul secvențelor aliniate) este maxim dacă elementele sunt identice. Dacă elementele sunt diferite scorul depinde de probabilitatea ca unul dintre elemente să fie transformat în celălalt prin mutație. Mutația poate consta nu doar în înlocuirea unei nucleotide sau aminoacid cu o altă nucleotidă sau aminoacid ci și în inserția sau eliminarea unei nucleotide/aminoacid (corespunde introduducerii de gap-uri în secvențele de aliniat).

In cazul secvențelor de nucleotide scorul de potrivire este stocat într-o matrice 5x5 care conține, de regulă, trei tipuri de valori pentru a recompensa potrivirea exactă, pentru a penaliza nepotrivirea și pentru a penaliza introducerea unui gap.

In cazul secvențelor de aminoacizi există două modalități principale de a stabili scorul de potrivire a unei perechi: folosind matrici de tip PAM (Point Accepted Mutation) sau BLOSUM (BLOck SUbstitution Matrix). Ambele matrici sunt construite prin estimarea statistică a probabilității de mutație pornind de la secvențe biologice aparținând unor specii înrudite.

Din punctul de vedere al scopului urmărit, alinierea poate să fie:

- *Globală:* se obțin două secvențe aliniate având aceeași lungime (se folosește în cazul secvențelor scurte sau a celor pentru care similaritatea dintre ele este mare)
- *Locală:* se aliniază doar porțiuni din secvențele inițiale

Din punctul de vedere al tehnicii utilizate, alinierea poate să fie obținută prin:

- *Metode exacte*: algoritmul Needleman-Wunsch (pt. aliniere globală) și algoritmul Smith-Waterman (pt. aliniere locală); ambii algoritmi se bazează pe tehnica programării dinamice și ordinul de complexitate este dat de produsul lungimilor secvențelor care se aliniază; aceste metode pot fi folosite pentru secvențe scurte.
- *Metode aproximative*: Fasta și BLAST. Sunt tehnici euristice care se folosesc în principal la căutarea unei secvențe într-o bază de date. Ideea de bază este identificarea unor subsecvențe scurte (cuvinte) și extinderea lor atâta timp cât scorul de potrivire este suficient de mare.

Problema alinierii se poate pune și pentru cazul mai multor secvențe, caz în care este denumită aliniere multiplă și necesită tehnici specifice.

1. Analiza vizuală a similarității dintre două secvențe

Analiza vizuală a similarităților dintre două secvențe se poate realiza prin intermediul unei matrici cu puncte (*dot matrix*). Fiind date două secvențe s1 și s2, prima având n1 elemente iar a doua n2 elemente se construiește o matrice cu n1 linii (elementele din secvența s1 vor eticheta liniile) și cu n2 coloane (elementele din secvența s2 vor eticheta coloanele) având elemente de forma:

m(i,j) = 1 dacă s1(i)=s2(j) și m(i,j)=0 dacă s1(i) !=s2(j).

Obs. In calculul valorii m(i,j) se poate ține cont nu doar de valorile de pe pozițiile i respectiv j ci și de potrivirile din vecinătatea acestor poziții. În acest caz se folosește o "fereastră de parcurgere" a secvențelor. In cazul unei ferestre de dimensiune w se contorizează numărul de potriviri din subsecvențele s1(i-w:i+w) și s2(j-w:j+w), iar dacă numărul de potriviri depășește un prag atunci valoarea lui m(i,j) se pune pe 1.

Pentru vizualizarea matricii de puncte se poate utiliza funcția dotPlot din pachetul seqinr.

Mod de apel: dotPlot(sir1,sir2) unde sir1 si sir2 sunt vectori de caractere
Parametri suplimentari:

- wsize (dimensiunea ferestrei de parcurgere a secvențelor)
- wstep (dimensiunea pasului de parcurgere)
- nmatch (numărul de corespondențe din ferestre peste care se consideră că e potrivire)

Valorile implicite ale parametrilor sunt egale cu 1.
Exemplu de apel: dotPlot(sir1,sir2,wsize=8,wstep=1,nmatch=4)

Exercitiul 1. Să se vizualizeze matricea de puncte pentru:

a) secvențele de nucleotide din GenBank având identificatorii 'NM_000520' (http://www.ncbi.nlm.nih.gov/nuccore/NM_000520) respectiv 'AK080777' (http://www.ncbi.nlm.nih.gov/nuccore/AK080777). Ambele secvențe corespund unei gene ce controlează sinteza unei proteine – hexasaminidase A – ce intervine în boala Tay-Sachs, însă prima este din genomul uman iar cea de a doua de la șoarece;
b) două secvențe aleatoare de aminoacizi (având aceleași lungimi ca secvențele anterioare).

Indicatie. Se va încerca construirea matricii de puncte pornind atât de la secvențele de nucleotide cât și de la secvențele corespunzătoare de aminoacizi (obținute cu funcția din fișierul nt2aa.R). De asemenea se va construi matricea atât pe baza potrivirilor individuale cât și pe baza potrivirilor determinate pe baza unei ferestre de parcurgere (de exemplu dimensiunea ferestrei este 8 iar numărul minim de potriviri este 4)

a) După descărcarea fișierelor de la http://www.ncbi.nlm.nih.gov/ șirurile corespunzătoare pot fi prelucrate cu

```
sir1=read.fasta("AK080777.fasta")  # se specifica numele fisierului in care s-a salvat din
                                   # GenBank
sir2=read.fasta("NM_000520.fasta")
dotPlot(sir1[[1]],sir2[[1]])
```

Pentru a verifica ce se obține în cazul secvențelor de aminoacizi se construiesc șirurile corespunzătoare de aminoacizi:

```
sir1a =nt2aa (sir1[[1]]); sir2a =nt2aa (sir2[[1]]);
```

și se apelează funcția dotPlot în două variante:

```
dotPlot(sir1a,sir2a)
```

respectiv

```
dotPlot(sir1a,sir2a,wsize=8,wstep=1,nmatch=4)
```

b) Secvențele aleatoare pot fi generate folosind funcțiile de simulare a variabilelor aleatoare (vezi Lab 2) sau prin selecția aleatoare a etichetelor matricilor de scor (de exemplu BLOSUM50 definită în subpachetul Biostrings din Bioconductor):

```
library("biostrings")        # incarcare pachet
data(BLOSUM50)               # incarcare matrice cu scoruri
srand1=sample(rownames(BLOSUM50),length(sir1a),replace=TRUE)
srand2=sample(rownames(BLOSUM50),length(sir2a),replace=TRUE)
```

Tema 1. Să se implementeze (într-un limbaj de programare la alegere) algoritmul de construire a unei matrici cu valori de similaritate în varianta caracterizată prin potriviri la nivelul unor subsecvențe. Algoritmul va avea ca date de intrare: secvențele și lungimea subsecvenței utilizată la analiza similarității între două poziții. Valoarea similarității se calculează ca fiind raportul dintre numărul de potriviri între cele două subsecvențe și lungimea subsecventei. Matricea cu valori de similaritate va fi vizualizată grafic (valoarea unui element va corespunde nivelului de gri al celulei asociate în vizualizarea grafică).

2. Alinierea globală a două secvențe. Algoritmul Needleman-Wunsch

Alinierea globală a două secvențe s1 și s2 (folosind algoritmul Needleman-Wunsch – care este bazat pe tehnica programării dinamice) presupune construirea unei matrici conținând scoruri de aliniere partială. Scorul alinierii va fi reprezentat de elementul de pe ultima linie și ultima coloană a matricii. Elementele acestei matrici se calculează folosind relația de recurență:

$S(i,j)=\max\{S(i-1,j-1)+m(s1(i),s2(j)),S(i-1,j)-g,S(i,j-1)-g\}$ pentru $i=1..length(s1)$, $j=1..length(s2)$
$S(i,0)= - i*g$
$S(0,j)= - j*g$

In relatia de recurență m(s1(i),s2(j)) reprezintă scorul de potrivire dintre simbolurile aflate pe pozițiile i respectiv j din cele două secvențe, iar g reprezintă penalizarea pentru introducerea unui gap.

In cazul secvențelor de nucleotide se poate utiliza un scor comun de potrivire (sp), un scor de penalizare a nepotrivirii (-sn) și un scor de penalizare a inserției gap-urilor.

In cazul secvențelor de aminoacizi, scorurile de potrivire sunt cuprinse în matrici specifice calculate folosind estimări statistice pornind de la secvențe considerate suficient de similare (vezi curs 7). Cele mai frecvent folosite matrici sunt BLOSUM50, BLOSUM62 și PAM250.

Pachetul Biostrings conține funcția pairwiseAlignment care permite alinierea a două secvențe (de nucleotide sau de aminoacizi) specificate ca șiruri de caractere. Cel mai simplu mod de apel al funcției de aliniere este: pairwiseAlignment(secv1,secv2).

Observație. Lista parametrilor de apel ai funcției pairwiseAlignment poate fi extinsă cu:

- substitutionMatrix
  - In cazul secvențelor de nucleotide matricea de substituție poate fi definită folosind funcția nucleotideSubstitutionMatrix(match=sp,mismatch=-sn,baseOnly=TRUE)
  - In cazul secvențelor de aminoacizi se pot folosi matricile de substituție clasice: "BLOSUM40", "BLOSUM50", "BLOSUM62", "BLOSUM80", "BLOSUM100", "PAM250" etc.
- gapOpening (reprezintă valoarea penalizării aplicate în cazul inițierii unei secvențe de gap-uri)
- gapExtension (reprezintă valoarea penalizarii aplicate în cazul extinderii unei secvențe de gap-uri – valoarea absolută a penalizării extinderii este mai mică decât valoarea absolută a penalizării inițierii secvenței de gap-uri)
- scoreOnly (valori posibile: TRUE / FALSE; dacă opțiunea este setată pe TRUE atunci se returnează doar scorul alinierii).

Exercitiul 2. Aliniați secvențele de nucleotide/aminoacizi de la Ex.1. folosind funcția pairwiseAlignment. Analizați efectul schimbării matricii de substituție și a valorilor de penalizare pentru gap-uri.

Exercițiul 3. Implementați în R algoritmul Needleman-Wunsch pentru alinierea secvențelor de nucleotide pentru cazul în care se utilizează penalizarea liniară a gap-urilor (detalii în curs 6).

Indicație. O variantă de implementare este descrisă în fișierul NWscor.R (pentru calculul matricii de scor) și în NWaliniere.R (pentru construirea unei alinieri).

Exemplu de apel:
matriceScor=NWscor("GAATTC","GATTA",2,-1,-2)
aliniere=NWaliniere("GAATTC","GATTA",matriceScor,-2)

3. Alinierea locală a două secvențe. Algoritmul Smith-Waterman.

Principala diferență dintre algoritmii de aliniere locală și cei de aliniere globală este faptul că în cazul alinierii locale matricea de scor conține doar valori pozitive, relația de recurență fiind:

$S(i,j)=\max\{0,S(i-1,j-1)+m(s1(i),s2(j)),S(i-1,j)-g,S(i,j-1)-g\}$ pentru i=1..length(s1), j=1..length(s2)
$S(i,0)= 0$
$S(0,j)= 0$

Scorul alinierii este reprezentat de valoarea maxima din matricea de scor.

Pentru a obține o aliniere locală se poate folosi tot funcția pairwiseAlignment însă trebuie specificată opțiunea type=”local”.

Exercitiul 4. Să se realizeze alinierea locala a secvențelor de la Ex. 1a și de la Ex 1b și să se compare lungimile și scorurile de aliniere obținute cu cele de la Ex. 2 (corespunzător alinierii globale). Se vor testa rezultatele obținute cu următoarele matrici de scor: PAM30, PAM40, PAM70,PAM120,PAM250, BLOSUM50 si BLOSUM62.

Exercitiul 5. Implementați în R algoritmul Smith-Waterman pentru alinierea secvențelor de nucleotide pentru cazul în care se utilizează penalizarea liniară a gap-urilor (detalii în curs 6).

Indicație. O variantă de implementare este descrisă în SWscor.R (pentru calculul matricii de scor) și în SWaliniere.R (pentru construirea unei alinieri).

Exemplu de apel:
matriceScor=SWscor("GAATTC","GATTA",2,-1,-2)
aliniere=SWaliniere("GAATTC","GATTA",matriceScor,-2)

Tema 2. Alinierea de tip overlap se caracterizează prin faptul că se ignoră penalizarea gap-urilor de la extremități urmărindu-se alinierea porțiunii finale dintr-o secvență cu porțiunea initial din altă secvență. In acest caz alinierea e de forma:

xxxxxxxxxxxxxxxxxxxx
yyyyyyyyyyyyyyyyyyyyyyy

sau

xxxxxxxxxxxxxxxxxxxx
yyyyyyyyyyyyyyyyyyyyyyy

(doar porțiunile suprapuse sunt luate în considerare în calculul scorului de aliniere.

La construirea matricii de scor prima linie și prima coloană este inițializată cu 0, iar celelalte elemente se completează în aceeași manieră în care se completează pentru alinierea globală. La construirea alinierii se pornește de la valoarea maxima aflată fie pe ultima coloană, fie pe ultima linie a matricii.

Tema 3 (suplimentară). Să se modifice algoritmii de aliniere globală și locală pentru cazul penalizării afine a gap-urilor. Se pot folosi următoarele valori ale scorurilor:

Potrivire nucleotide: r=6
Nepotrivire: p=-2
Initiere gap: d=-6
Extindere gap: e=-4