Biostatistică și Bioinformatică.

Lab 7:

Predicția structurii proteinelor – Modelul HP

---

1. Predicția structurii proteinelor. Modelul HP

Aranjarea spațială a aminoacizilor constituenți are un rol important în stabilirea funcției pe care o îndeplinește o proteină, iar erorile din această aranjare (erori de împachetare) pot fi cauza unor boli. Predicția structurii unei proteine presupune determinarea modului în care sunt amplasați aminoacizii constituenți în cadrul unei structuri laticiale astfel încât să fie minimizată energia structurii respective.

Unul dintre cele mai simple modele de descriere a structurii proteinelor este modelul HP (Hidrofobic(non-polar)-Hidrofilic(polar)) în care cei 20 de aminoacizi sunt împărțiți în două categorii:

- Hidrofobi (H)
- Hidrofili (P)

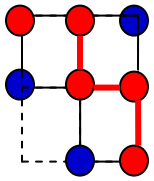
și sunt plasați în nodurile unei grile. Cazul cel mai mult studiat este cel al grilelor bidimensionale pătratice.

În acest caz scopul urmărit este ca pornind de la o secvența de aminoacizi să se determine pozițiile acestora în nodurile grilei astfel încât numărul de *contacte* să fie maximizat. Pentru o secvență de aminoacizi având elementele plasate în noduri ale grilei *un contact este reprezentat de două elemente de tip H care sunt vecine în cadrul grilei dar nu sunt vecine în cadrul secvenței*.

Exemplu:

1	2	3	4	5	6	7	8
H	P	H	P	H	H	P	H

Reprezentare binară: 1 0 1 0 1 1 0 1



Număr contacte (muchii rosii): 3

Proprietăți:

- Contactele pot exista doar între aminoacizi hidrofobi plasați pe poziții de paritate diferită în secvența inițială (distanța dintre doi aminoacizi care formează contact trebuie să fie număr impar).
- Numărul maxim de contacte corespunzătoare unei secvențe S în cazul unei grile bidimensionale pătratice este:

$$C_{2D}(S) = 2 \min(O(S),E(S))$$

unde O(S) este numărul aminoacizilor de tip H aflați pe poziții impare în secvența initială iar E(S) e numărul celor aflați pe poziții pare.

Algoritmul Hart-Istrail

Este un algoritm aproximativ de complexitate liniară care permite construirea de împachetări care conțin cel puțin $C_{2D}(S)/4$ contacte (deci rata de aproximare a algoritmului este 0.25).

Ideea algoritmului:

- Se identifică un indice p din {1,2,…,N} (N fiind lungimea secvenței) astfel încât cel puțin jumătate dintre aminoacizii de tip H aflați pe poziții pare se află în stânga poziției p iar cel puțin jumătate dintre aminoacizii de tip H aflați pe poziții impare se află în dreapta lui p. O astfel de poziție, numită *poziție de balansare*, există întotdeauna întrucât este suficient să se aleagă p ca fiind poziția cu proprietatea că în stânga sa se află jumătate dintre aminoacizii H de tip par iar în dreapta se află cealaltă jumătate de aminoacizi H de tip par.
- După identificarea poziției p se plasează elementele din secvență (pornind de la poziția p) astfel încât aminoacizii H de pe poziții pare dintr-una dintre părți să fie învecinați cu aminoacizi H de pe poziții impare din cealaltă parte.

Exemplu:

```
        1 2 3 4 5 6 7 8
S:      H P H P H H P H
Binar:  1 0 1 0 1 1 0 1
```

Pozițiile impare pe care se află H: {1,3,5} => O(S)= 3

Pozițiile pare pe care se află H: {6,8} => E(S)= 2

Număr maxim de contacte: $C_{2D}(S) = 2 \min(O(S),E(S))=4$

Poziția de balansare: p=7

Impachetare (prin crearea unei "bucle" care conține punctul de balansare =3 contacte):

```
    H==P
    |  ||
P ==H _ H
||   |   ||
H==H==P
```

Direcții de deplasare (U(p)-sus, D(own)-jos, L(eft)-stânga, R(ight)-dreapta) pornind de la primul element din secvență (cel marcat cu roșu): RDDLLUR

Exercițiul 1:

Scrieți funcții R care primesc o secvență S de aminoacizi în codificare HP (sau binară) și returnează:

a) O(S), E(S), $C_{2D}(S)$
b) o poziție de balansare p

Indicatie: exemple de implementare in ProteinFolding_modelHP.R (functiile countH si findP)

Exercițiul 2:

a) Scrieți o funcție R care primește o secvență de direcții de deplasare (de exemplu LDRRUUL) și returnează o listă cu coordonatele pozițiilor aminoacizilor (considerând că primul element din secvență se află pe poziția (0,0)).
b) Folosind secventa initiala si pozitiile din grila vizualizati grafic impachetarea si determinati care dintre elementele din secvență formează contacte.

Indicație. Deplasările se calculează folosind regulile:

U: (i,j) -> (i,j-1) L: (i,j)-> (i-1,j)
D: (i,j)-> (i,j+1) R: (i,j)->(i+1,j)

Exemple de implementare sunt in ProteinFolding_modelHP.R (functiile pozitieGrila si grafic)

Exercițiul 3:

Urmăriți animații/simulari ale procesului de împachetare:

a. http://www.youtube.com/watch?v=yZ2aY5lxEGE
b. http://www.youtube.com/watch?v=gFcp2Xpd29I
c. http://www.youtube.com/watch?v=SxAqyw9hr4k