PROJET 4 : Evaluation de la structure tridimensionnelle d'une protéine par des potentiels de force moyenne

Introduction :

Nos connaissances actuelles sur la physique du repliement n'est pas suffisante pour nous permettre de prédire la stabilité d'une protéine à partir de sa seule séquence et sa structure. Cependant, les méthodes statistiques peuvent produire des modèles utiles à partir de données expérimentales. Les potentiels statistiques décrivant la propension d'une paire de résidus d'être à une certaine distance spatiale, sont appelés potentiels de force moyenne.
L'application de ces potentiels recouvre de nombreux aspects essentiels tels que l'optimisation de l'alignement de séquences lorsqu'une structure est disponible, l'évaluation de modèles tridimensionnels de protéines, le repliement *ab initio,* la détection du repliement correspondant à une séquence cible (*threading*), ou encore la prédiction de l'impact énergétique, structural ou fonctionnel de mutations ponctuelles.

But du projet :

Le but est de générer des potentiels statistiques à partir d'un ensemble de structures représentatives de la PDB. Ensuite à partir d'une structure cible, il faut créer un programme qui calcule son énergie totale à partir des potentiels déterminés précédemment.

Références :

- Protein Sci. 2006 Nov;15(11):2507-24. Statistical potential for assessment and prediction of protein structures. Shen MY, Sali A.
- J Mol Biol. 1990 Jun 20;213(4):859-83.Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. Sippl MJ.

Exercices :

1) Créez un programme qui, à partir d'un fichier PDB, lu en argument, affiche en sortie pour chaque carbone alpha (CA) le numéro de l'acide aminé, le type d'acide aminé et les coordonnées cartésiennes du Carbone alpha.
   Exemple :

```
          x        y        z
1 ALA -11.956   8.450   7.858
2 VAL -12.364  11.175   5.203
3 ARG -11.707  11.147   1.438
```

   Pour filtrer le fichier PDB, utilisez le champ ATOM et le champ spécifiant le type d'atome (CA).

2) Créez un programme qui, à partir d'un fichier PDB, calcule les distances inter carbone alpha séparant chaque acide aminé.

   La distance est calculée selon la formule suivante :

$$\text{Distance entre 1 et 2} = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)^2}$$

L'affichage devra avoir le format suivant :

```
Position de l'acide aminé 1 sur la séquence
| Acide aminé 1
|  |    Position de l'acide aminé 2 sur la séquence
|  |    | Acide aminé 2
|  |    |    Distance en Angstrom séparant les carbones
|  |    |    |  alpha des deux acides aminés
|  |    |    |
1 ALA 2 VAL 2.5
1 ALA 2 ARG 3.5
. ... . ... ...
```

3) Créez une version modifiée du programme précédent qui, à partir d'un fichier PDB, calcule les distances inter carbone alpha séparant chaque acide aminé et qui ajoute deux filtres : Le premier filtre est basé sur la distance, au niveau de la séquence séparant deux acides aminés, qui doit être supérieure à 4.
Exemple :
Dans la séquence suivante MKPLEKHYSVVLPTR, si je dois comptabiliser les distances à partir du carbone alpha de l'Histidine (H) je ne prendrais en compte que les acides aminés M en position -6, K en position -5, L en position +5, P en position +6 …

```
Position relative 6543210 12345678…
SEQUENCE          MKPLEKHYSVVLPTR…
```

Un deuxième filtre est appliqué : il est basé sur la distance au niveau spatial séparant les deux carbones alpha, seuls les résidus situés entre 0 et 15 Å sont conservés.
Le programme affiche ensuite les résultats selon le même format précédemment décrit.
Exemple :

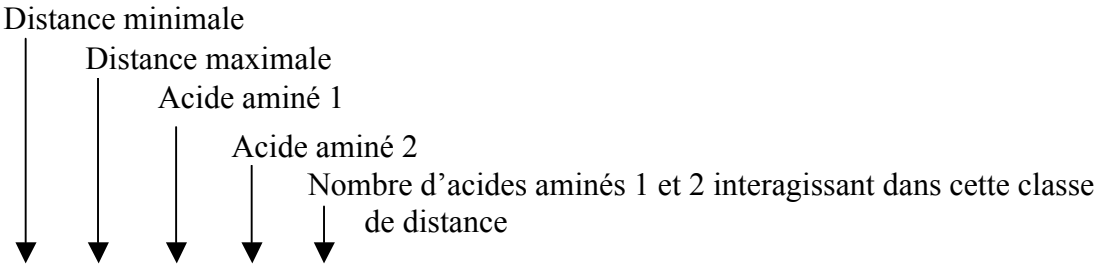
```
1 ALA 5 ILE 6.5
1 ALA 6 LEU 9.5
. ... . ... ...
```

4) Créez un programme qui va chercher un fichier PDB à partir du code à 4 lettres du fichier et à partir du RCSB par FTP (ftp://ftp.rcsb.org/pub/pdb/data/structures/all/pdb/) et le télécharge dans votre répertoire.

5) Interfacez ce programme avec le programme précédent de sorte qu'avec un code pdb (code à 4 lettres comme par exemple 1atn), le programme télécharge automatiquement le PDB, le décompresse (grâce à *zcat* ou *gunzip –d* ou fonction interne de perl), traite celui-ci avec votre programme de calcul de distance et vous donne les distances filtrées séparant les différents acides aminés dans le format détaillé initialement.

6) Créez un programme qui, à partir d'une liste de fichier PDB (à télécharger ici : http://www.dsimb.inserm.fr/~gelly/doc/list_rs126), utilise le programme précédent pour générer automatiquement une liste contenant les distances de séparation de tous les acides aminés pour tous les fichiers de la liste (pour des raisons de place et de temps de calcul, vous pouvez prendre une vingtaine de fichiers seulement).

7) Créez un programme qui, à partir du fichier précédent, comptabilise et classe les différents couples d'acides aminés selon la distance les séparant. Il existe 30 classes qui sont définies de 0 Å à 15 Å par pas de 0,5 Å (0.0 à 0.5 Å; 0.5 à 1.0 Å; …. ; 14,5 à 15.0 Å).




```
0.0 0.5 ALA ALA 0
0.5 1.0 ALA ALA 0
... ... ... ... ...
3.5 4.0 ALA CYS 112
... ... ... ... ...
3.5 4.5 GLY VAL 112
```

8) Créez un programme qui, à partir du fichier obtenu précédemment, dérive une énergie pour chaque type d'interaction.
L'énergie d'interaction entre un acide aminé *i* et un acide aminé *j* pour une classe de distance *d* est obtenue par la formule suivante :

$$E_{(ij,d)} = -kT \times \ln \left(f_{(ij,d)} / f_{(r,d)} \right)$$

Avec
k = constante de Boltzmann = 1.38066×10^{-23} J/K (joule/Kelvin)
T= température en degrés Kelvin = 298
Pour des raisons pratiques vous pouvez considérer que kT = 1. L'énergie ne s'exprimera donc pas en joule.

$f_{(ij,d)}$= fréquence de l'interaction ij à la distance d observée
$f_{(r,d)}$= fréquence d'interaction de référence à la distance d

Calcul de $f_{(ij,d)}$

$$f_{(ij|d)} = \frac{\text{Nombre d'interaction } ij \text{ à la distance } d}{\text{Nombre d'interaction } ij \text{ pour toutes les distances}}$$

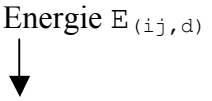
Exemple pour l'interaction ALA ILE à la distance 3,5 - 4,0
$f_{(ALA\ ,ILE,1)} = 150 / 1500 = 0.1$

Calcul de $f_{(r,d)}$

$$f_{(r,d)} = \frac{\text{Nombre d'interaction pour tous les couples ij à la distance } d}{\text{Nombre d'interaction pour tous les couples ij pour toutes les distances}}$$

Remarque : Si il n'y a aucune interaction pour une distance donnée, l'énergie sera de 10.0 par convention. Il s'agira aussi de l'énergie maximale possible (toute énergie supérieure sera remplacé par 10.0)

En sortie le fichier aura le format suivant :




```
0.0 0.5 ALA ALA  10.0
0.5 1.0 ALA ALA  -2.3
... ... ... ...   ...
3.5 4.0 ALA CYS  -5.1
... ... ... ...   ...
3.5 4.0 GLY VAL   5.5
```

9) Créez un programme qui, à partir du fichier obtenu précédemment, et d'un fichier PDB fournit en argument, calcule l'énergie de la protéine.
   Pour cela, dans une première étape, utilisez le programme 3 pour obtenir les différentes distances pour chaque interaction entre acides aminés de la protéine. Ensuite, dans une deuxième étape, remplacez ces distances par les énergies calculées grâce au fichier de sortie du programme 8. Faire une sommation de toutes les énergies afin d'obtenir l'énergie totale de la molécule que vous afficherez.

Pour aller plus loin :

- Proposez un programme qui permette le « protein design », c'est-à-dire que pour chaque position de la protéine, vous proposerez l'acide aminé qui aurait la meilleure énergie à cette position (attention plus une énergie est négative, meilleure elle sera).
- Proposez un outil de *threading* qui calcule le score d'adéquation d'une structure avec une séquence donnée (par double programmation dynamique si vous êtes à l'aise).
- Proposez une interface graphique en perlTK ou perlGTK.
- D'autres extensions sont possibles selon votre inspiration.