TP de Manipulation de Fichiers PDB

Documentation concernant le format des fichiers PDB :
http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

Guide : Vous testerez vos fonctions et programmes sur les PDBs suivants : 3PDZ et 1FCF.

A. Lecture d'un fichier PDB

(1) Ecrire une fonction qui lit un fichier PDB et renvoie (*i*) le nom de la méthode expérimentale (chaîne de caractères), (*ii*) la résolution de la structure, en Angströms (réel, pas toujours disponible), (*iii*) le nombre de modèles (entier, uniquement en RMN).

Aide : Les champs `EXPDTA` et `NUMMDL` d'un fichier PDB décrivent respectivement la méthode expérimentale et le nombre de modèles fournis (RMN) ; la résolution est mentionnée dans le champ `REMARK`.

(2) Ecrire une fonction qui lit un fichier PDB et renvoie les informations associées à la séquence de référence de chaque chaîne (nucléotidique ou protéique) : base de données et numéro d'accès, correspondance avec la numérotation PDB des résidus.

Aide : Voir description du champ `DBREF`.

(3) Ecrire une fonction qui lit un fichier PDB et renvoie les coordonnées 3D des atomes de chaque chaîne du fichier. Dans le cas de plusieurs modèles, par défaut, la fonction renverra les coordonnées du premier modèle uniquement. Optionnellement, la fonction renverra les coordonnées de chaque modèle.

Aide : Voir description des champs `MODEL` et `ATOM`.

(4) Ecrire une fonction qui lit un fichier PDB et renvoie, pour chaque atome de chaque chaîne, le nom du résidu auquel il appartient, le numéro du résidu et le nom de l'atome.

(5) Ecrire un programme principal qui fait appel à toutes ces fonctions.

B. Comparaison de deux structures protéiques

(1) Ecrire une fonction qui calcule le RMSD entre deux structures de protéines *P1* et *P2* alignées, données en entrée. Les sélections de résidus *sel_P1* et *sel_P2* mises en correspondance pour le calcul du RMSD seront données en paramètres d'entrée, sous forme de listes. Le RMSD sera calculé uniquement sur les C-alpha.

(2) Modifier le programme principal en conséquence.

(3) La structure 1FCF a été alignée sur 3PDZ avec Pymol, via la commande **align** (alignement basé sur la séquence). Télécharger le fichier correspondant *1fcf_aliSeq.pdb*. Calculer le RMSD entre *3pdz.pdb* et *1fcf_aliSeq.pdb*, en prenant comme sélections de résidus les résidus alignés dans la figure ci-dessous (3PDZ en rouge, 1FCF en bleu). Quelle est la valeur du RMSD ?

```
    21   26   31   36   41         46   51   56   61   66   71
----SVTGGVNTSVRHGGIYVKAVI-----PQGAAESDGRIHKGDRVLAVNGVSLEGATHKQ
 156  161   166  171  176  181  186  191   196  201  206  211
GTAGSVTG-VGLEITYDGGSGKDVVVLTPAPGGPAEKAGA-RAGDVIVTVDGTAVKGLSLYD
```

(4) Ouvrir les 2 fichiers PDB dans Pymol. Aligner 1FCF sur 3PDZ avec la commande **super** (alignement structural). Que constate-t-on ? Quelle est la valeur de RMSD correspondante ? Sélectionner les résidus 21 de 3PDZ et 159 de 1FCF pour les visualiser sur les structures. Sont-ils superposés ? Que peut-on en déduire ?

C. Cartes de contacts

(1) Ecrire une fonction qui calcule toutes les distances entre deux sélections de résidus dans une protéine, en considérant uniquement les C-alpha, et affiche le résultat sous forme de matrice colorée.

Aide : Pour l'affichage de la matrice, le module pylab et la fonction pcolor peuvent être utilisés.

(2) Proposer une mesure de dissimilarité entre deux cartes de contact et écrire une fonction qui calcule la dissimilarité entre deux cartes de contact données en entrée.

D. Variance circulaire

La variance circulaire est une mesure géométrique qui rend compte de l'enfouissement des résidus dans les structures de protéines. Etant donné un atome *i*, la variance circulaire de *i* est égale à 1 moins la résultante des vecteurs qui partent de *i* et pointent vers les autres atomes de la protéine, dans un rayon r_c :

$$CV_i = 1 - \frac{1}{n_i}\left| \sum_{j \neq i, r_i \leq r_c} \frac{\vec{r}_{ij}}{\left|\vec{r}_{ij}\right|} \right|$$

Si la résultante est nulle, alors l'atome est enfoui dans la protéine ; si la résultante est grande, alors l'atome est protubérant. Pour déterminer le niveau d'enfouissement d'un résidu, on calcule la moyenne des variances circulaires des atomes qui le composent.

(1) Ecrire une fonction qui calcule la valeur CV pour chaque résidu d'une protéine, étant donné un rayon donné en entrée. Typiquement, le rayon doit être supérieur ou égal à 20 Å.

(2) Ecrire une fonction qui détermine les x% de résidus les plus enfouis, et les x% les plus protubérants d'une protéine, étant donné un rayon donnée en entrée.

(3) Ecrire une fonction qui lit un fichier PDB et des valeurs associées à chaque résidu ou chaque atome, et écrit un fichier PDB avec les valeurs données en entrée dans la colonne des B-facteurs (12ème colonne). Utiliser cette fonction pour visualiser les valeurs de CV sous pymol

(4) Comparer les valeurs CV des chaînes A et B de la structure 2BBM, quand on considère chaque chaîne séparément, ou bien le complexe entier.

E. Champ de force

Un champ de force est une expression analytique (ou fonctionnelle) qui représente les interactions inter-atomiques d'un système moléculaire. Il permet d'estimer l'énergie mécanique moléculaire d'une protéine ou d'un complexe, dans le vide ou solvaté.

Pour une protéine *P*, dans le vide, l'énergie est exprimée comme :

$$E_{tot}(P) = E_{bonded}(P) + E_{non-bonded}(P,P) = E_{int}(P)$$

Où E_{bonded} correspond à l'énergie d'interaction entre les atomes liés covalemment et $E_{non-bonded}$ correspond à l'énergie d'interaction entre les atomes non liés (distant de plus de 4 atomes).

Pour un complexe formé par deux protéines *R* et *L*, dans le vide, l'énergie totale du système peut s'exprimer comme :

$$E_{tot}\,(R \cdot L) = E_{int}(R) + E_{int}(L) + E_{non-bonded}(R,L)$$

L'énergie d'interaction associée à la formation du complexe vaut :

$$\Delta E_{inter}(R \cdot L) = E_{tot}\,(R \cdot L) - E_{tot}(R) - E_{tot}(L) = E_{non-bonded}(R,L)$$

avec :

$$E_{non-bonded}(R,L) = \sum_{i,j} \frac{A_{ij}}{R_{ij}^8} - \frac{B_{ij}}{R_{ij}^6} + f\,\frac{q_i q_j}{20 R_{ij}}$$

où *i* et *j* sont les indices des atomes de *R* et *L*, et *f=332.0522*. Voir le fichier `ForceField.py` (cf Cornell *et al.* 1995) pour les paramètres.

(1) Ecrire une fonction qui calcule l'énergie d'interaction entre deux chaînes contenues dans un fichier PDB.

(2) Calculer l'énergie d'interaction entre la calmoduline et son peptide cible à partir du fichier 2BBM.