Projet 1 : Banque de données sur les protéines

Dataset :

La banque de données Worldwide Protein Data Bank (http://www.wwpdb.org/) regroupe les données des structures macromoléculaires obtenues par diffraction aux rayons X ou par RMN.
Elle met ces données gratuitement à la disposition de tous. Leur format est parfaitement défini et conventionnel et vous trouverez plus d'infos ici : http://mmcif.wwpdb.org/ .

Objectif :

On propose pour ce projet de réaliser un script qui analyse un fichier de données de type PDB (*Protein Data Bank* : https://fr.wikipedia.org/wiki/Protein_Data_Bank).

Nous vous proposons de considérer la structure résolue pour la GFP (*Green Fluorescent Protein*, Prix Nobel 2008). Son nom d'entrée dans la base de données PDB est *1gfl*.
Sa fiche PDB est : http://www.rcsb.org/pdb/explore.do?structureId=1GFL .

Les données de toutes les protéines se trouvent ici :
ftp.ebi.ac.uk/pub/databases/pdb/data/structures/divided/mmCIF/
A vous de chercher le fichier qui correspond à la GFP !

Pour le fichier de données de la protéine Green Fluorescent, vous pourrez par exemple assurer sa lecture pour calculer :

- le barycentre de la biomolécule
- le nombre d'acides aminés ou nucléobases
- le nombre d'atomes
- la masse moléculaire
- les dimensions maximales de la protéine
- etc.

Les outils Python pour manipuler les fichiers de données sont spécifiques, il vous faut télécharger une librairie spécifique :
http://mmcif.wwpdb.org/docs/sw-examples/python/html/index.html
Afin de pouvoir utiliser ce nouveau module *pdbx*, enregistrez-le dans le dossier où vous allez écrire et lancer vos scripts. Ensuite, il faut ajouter à chaque début de script (avant de faire un import de pdbx) :

```
import sys
sys.path.append("/votre/chemin/vers/votre/dossier/du/projet/")
```

Vous trouverez des exemples d'utilisation de *pdbx* dans le lien ci-dessus.