Implémentation d'un outil de visualisation et d'analyse en protéomique quantitative

BETHEGNIES-GENTY Fabien
Master de Bioinformatique, Université Paris Diderot

Enseignant référent : Mme Catherine Etchebest

Projet proposé par: Mr Pierre Poulain

Année 2017 -2018

Table des matières

I. Introduction

L'objectif de ce projet est d'implémenter un outil de visualisation pour l'analyse de données en protéomique quantitative. Cet outil est intégré dans un serveur virtuel permettant aux utilisateurs de réaliser leurs analyses en ligne.
Les données à la base d'expériences sont des données protéomiques quantitatives.Elles proviennent d'une expérience de protéomique quantitative entre deux conditions A et B.Pour chaque protéine, l'expérience permet de déterminer le rapport des abondances entre A et B et p-value.
Le rapport des abondances mesurées par spectrométrie de masse dans les conditions A et B quantifie la différence d'expression protéique entre A et B et la p-value quantifie la confiance que l'on peut avoir dans l'éventuelle variation mesurée entre A et B.

On représente pour toutes les protéines identifiées le logarithme du rapport des abondances (log fold change) en fonction du logarithme de la p-value (log pvalue). Le graphique obtenu est appelé volcano plot.

II. Matériels et méthodes

II.2. Volcano plot

Il est utilisé pour identifier rapidement les changements dans de grands ensembles de données composées de données répliquées.[1]
Un volcano plot combine une mesure de significativité statistique avec l'ampleur du changement, permettant une identification visuelle rapide des points ayant une significativité statistique importante.

Le graphique est construit en traçant le $\log_{10}$ négatif de la pvalue (logpv) sur l'axe des ordonnées contre le $\log_2$ du fold-change (logfc) sur l'axe des abscisses.
Cela permet de monter que les pvalues faibles qui sont hautement significatives apparaissant vers le haut de la courbe.

Le tracé permet d'obtenir deux nuages de points permettant de mettre en évidence deux régions d'intérêt sur le graphique.Les points qui se trouvent vers le haut graphe et qui sont éloignés vers la gauche ou la droite.Ceux-ci représentent des mesures ayant de grands changements de fold change c'est à dire avec un rapport d'abondance très grand ou très petit (donc étant à gauche ou à droite du centre) ainsi qu'une haute signification statistique (donc vers le haut).

On peut donc pointer deux régions d'intéret en haut à gauche et en haut à droite comme on peut le voir dans la figure 1

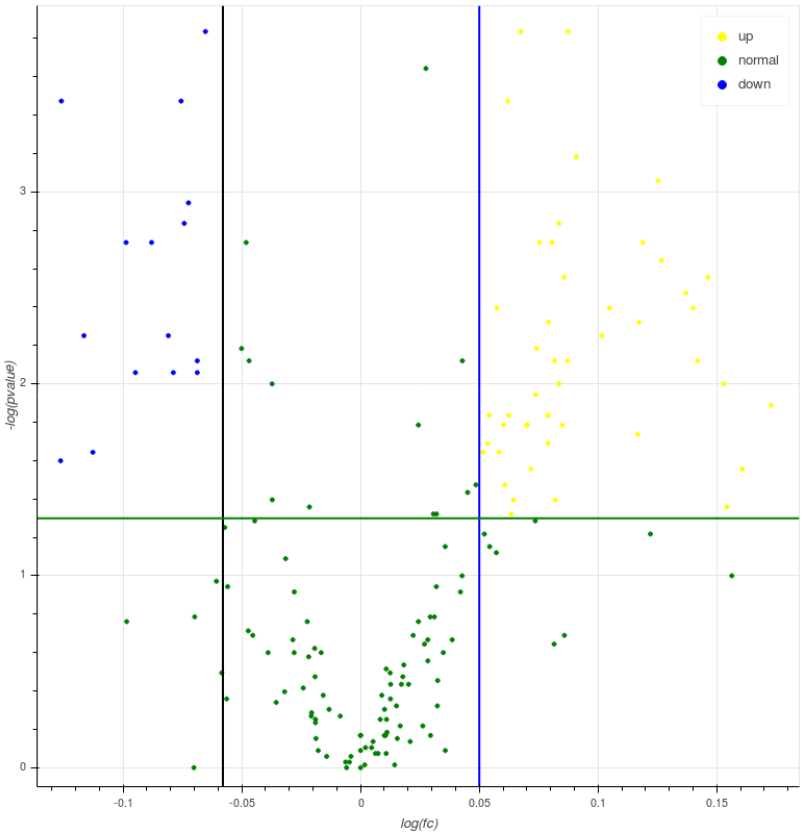

Figure 1 : Volcano plot issue de l'application développée. Les points d'intérêts sont de couleurs bleu et jaune.Ils sont situés en haut à droite et en haut à gauche

II.2. Pandas

Pandas est une librairie Python destinée à la manipulation et à l'analyse de données. En particulier, il propose des structures permettant de manipuler aisément des grand jeux de données. Dans ce projet, pandas a particulièrement été utilisé pour la récupération de données issues de fichiers CSV. De plus il permet de gérer des tableaux de données avec des formats de données différents pour chaque colonne.

II.3. Bokeh

Bokeh est une librairie de visualisation interactive Python basée sur l'utilisation de navigateur web pour la présentation des données. Son but est de fournir une construction élégante et concise de nouveaux graphismes.[2]

Il permet d'augmenter l'interactivité entre l'utilisateur et le programme et est efficace pour l'analyse dynamique de très grands jeux de données en continu. Bokeh va permettre de créer rapidement et facilement des graphiques et des applications de données interactives.
Il est principalement écrit en python mais il s'intègre facilement dans des applications web et intègre aussi JavaScript.

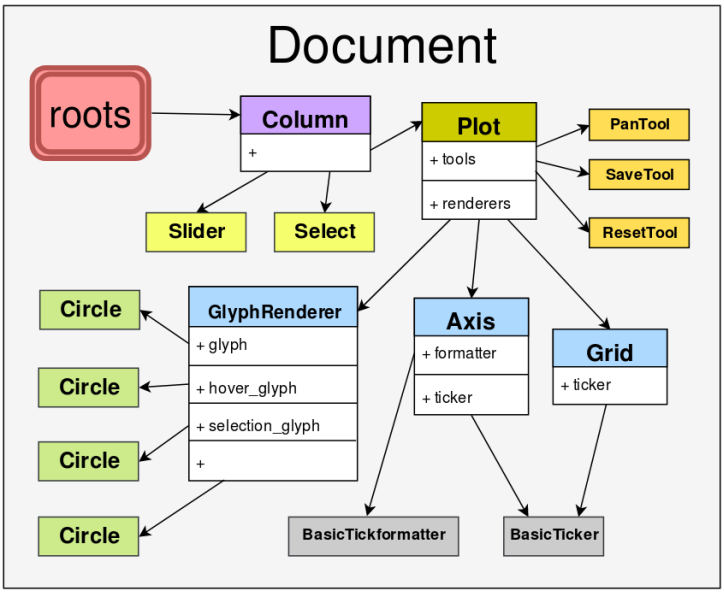

Figure 2 .Structure d'un document bokeh

Dans ce projet Bokeh est utilisé pour construire le graphique et l'interaction avec l'utilisateur.
La construction d'un document Bokeh se compose de différents modules. Dans la figure 2, on peut voir comment un document Bokeh fonctionne et les différentes interactions entre les modules
De plus Bokeh est prévu pour pouvoir supporter son intégration dans un serveur comme on peut le voir dans la figure 3.

II.4. Flask

Flask est un framework web écrit en Python et basé sur Werkzeug et le moteur de template Jinja2[3].

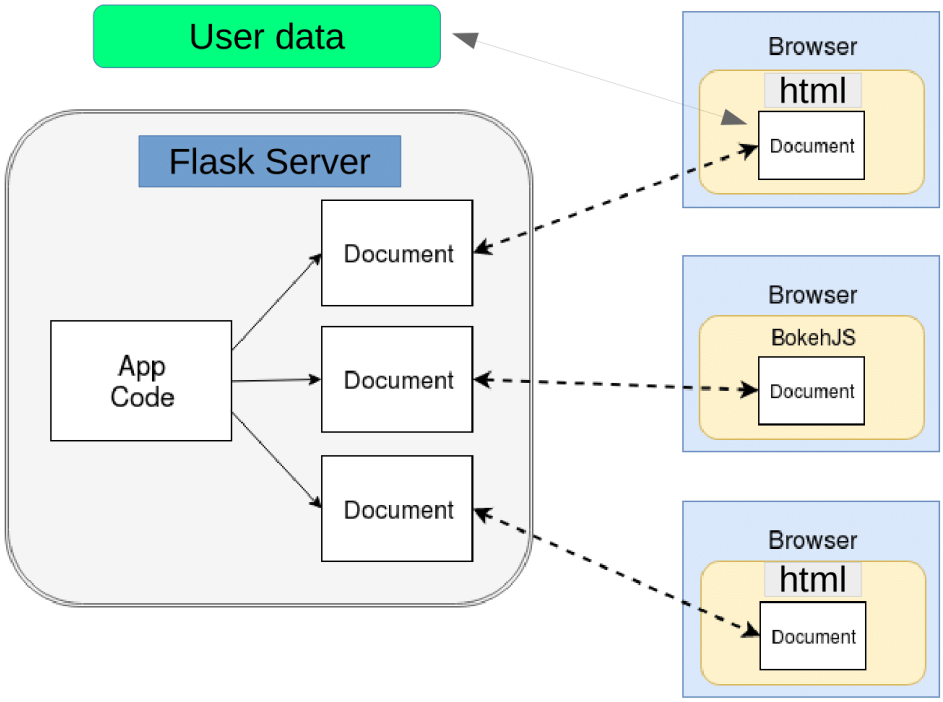

Figure 2. Worflow de fonctionnement du seveur flask intégrant bokeh

Il permet de générer facilement un serveur web virtuel. La majorité du code étant en python il est plus simple à utiliser que PHP et intègre aussi son propre moteur de template (Jinja 2).
Dans le projet, Flask est utilisé pour la structure du sit web.Il gére le JavaScript les templates CSS et HTLM ainsi que les requête HTTP. Le serveur est structuré comme dans la figure 4.

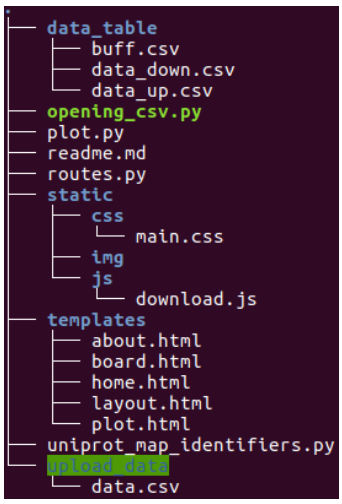

Figure 4 : le répertoire app qui est à la source du serveur Flask et est structuré en sous dossier

II.5. Rest API

Une API Web est une interface de programmation d'application (API) pour un serveur Web ou un navigateur Web. Il s'agit d'un concept de développement Web, généralement limité au côté client

Dans le projet une Rest API permet de récupérer les données sur le site d'Uiprot via le serveur Bokeh ou à partir d'un script python[4]
Cela permet d'automatiser les requêtes sur le site Uniprot[5]. Grace à cela on peut récupérer les informations importantes grâce a des mots clés prédéfinis. La figure 5 montre comment les requêtes sont transférées

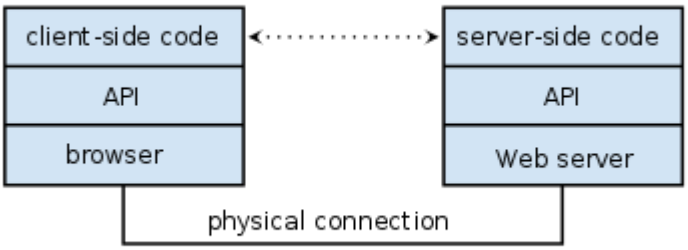

Figure 5. Transfert des requêtes

II.6. Github

GitHub est un service d'hébergement de référentiel de contrôle de version Git basé sur le Web. Il est principalement utilisé pour le code informatique. Il offre toutes les fonctionnalités distribuées de contrôle de version et de gestion de code source de Git.[6]
Dans ce projet, Github a été utilise pour le versionning de toute les étapes du code et pour avoir un moyen simple et rapide de communiquer avec l'enseignant référent.

III. Résultats et discussions

Au final l'utilisateur rentre un fichiers CSV qui est convertit en Dataframe pandas. Les donnés sont ensuite envoyé es sur le plot qui va afficher le volcano plot.
Après cette étape l'utilisateur va pouvoir choisir les seuil du logpc et du logfc.
Un fois que l'utilisateur a choisi ces seuils, il peut télécharger une une image du plot ou une copie du tableau des résultats.

Il va ensuite pouvoir soumettre ses résultat pour qu'une requête soit lancée sur chaque numéro d'accession qu 'il a sélectionné. Cette requête permet de récupérer les informations demandées comme l'Uniprot id ou la taille de la protéine.

Sur une autre page, deux autres tableaux sont affichés. Le premier contient les protéines de l'expérience ayant une forte pvalue et un grand rapport d'abondance(A/B haut). Dans l'autre un rapport d'abondance faible (A/B bas).

Ensuite, les valeurs de chaque tableau tableau vont pouvoir être sélectionnées et téléchargées.

Un problème est survenu et je n'ai pas eu le temps de toute implémenter en particulier le tableau interactif et le réseau d'interactions.
Il faut aussi noter que certaines fois la requête API vers le site Uniprot ne marche pas ce qui fait planter la page du tableau.
A part la page du tableau, et celle concernant le graphique d'interaction tout le reste du site marche.
Par faute de temps j'ai juste implémenter le graphe représentant les interactions entre les protéines

IV. Conclusion

Le but premier de ce projet était d'aider les chercheurs avec un outil simple à utiliser et qui permet de mettre en valeur et de mieux analyser les résultats. La librairie bokeh répond bien à cette demande. La suite du projet serait de renforcer la sécurité du serveur et de mieux implémenter la gestion des ports et des requêtes. On pourrait ainsi envisager de déployer le serveur sur internet ou sur un intranet du laboratoire

V. Bibliographie

1. Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4,** 210 (2003).
2. Welcome to Bokeh — Bokeh 0.12.13 documentation. Available at: https://bokeh.pydata.org/en/latest/. (Accessed: 8th January 2018)
3. Welcome | Flask (A Python Microframework). Available at: http://flask.pocoo.org/. (Accessed: 8th January 2018)
4. Programmatic access - Mapping database identifiers. Available at: http://www.uniprot.org/help/api_idmapping. (Accessed: 8th January 2018)
5. Programmatic access - Retrieving entries via queries. Available at: http://www.uniprot.org/help/api_queries. (Accessed: 8th January 2018)

6. Build software better, together. *GitHub* Available at: https://github.com. (Accessed: 8th January 2018)