Devoir maison de Statistiques Appliquées

L3 SIF

1 Description du jeu de données

Votre travail consiste à analyser le jeu de données founi et décrit çi-dessous. Il provient des travaux d'un petit groupe de scientifiques (spécialisés dans l'Evolution), qui s'est posé la question de la qualité des programmes utilisés en routine dans leur discipline.

Avec les nouvelles technologies de séquençage[1], la biologie devient une science des données. C'est une transition que d'autres disciplines comme l'astrophysique ont accompli il y a quelques dizaines d'années. En conséquence, de nombreux programmes ont été développés pour répondre et prendre en charge ces données, parraléliser les algorithmes etc... Tous les programmes les plus utilisés se sont complexifiés, des pipelines très complexes ont été mis en place. Les chercheurs utilisent ces programmes en routine pour analyser leurs données, sans se soucier de savoir si les programmes sont robustes et bien écrits[2].

Le groupe de chercheurs a donc sélectionné 15 grogrammes qui sont des références de leur domaine, et s'est posé la question de leur robustesse en regardant les points suivants :

- `compilation` : combien de *warnings* majeurs et mineurs trouve t'on à la compliation (avec le compilateur `gcc`, combien de *warnings* trouve t-on en compilant avec `Clang` ?
- `gestion mémoire` : en utilisant l'outil `valgrind`, ils ont détecté les memeory leaks, les accès illégaux, les blocs mémoires perdus etc. Les résultats sont classés en 3 catégories : `clean` lorsque `valgrind` ne génère aucun *warning*, `invalid` pour des accès en lecture ou écriture à une adresse invalide, et `leaks` lorsque l'allocation mémoire n'est pas proprement libérée.
- `duplication de code` : le nombre de lignes de code dupliquées, ainsi que le nombre de blocs entiers de code dupliqués ont été comptabilisés.

L'intégralité de cet article est disponible ici :
`https://www.biorxiv.org/content/biorxiv/early/2015/11/16/031930.full.pdf`

[1] note EB : on peut de nos jour séquencer le génome d'un organisme pour moins de 1000 euros, en moins d'une journée, ce qui selon la taille de l'organisme produit entre 3 et 100GB de données à chaque fois

[2] voire parfois corrects, mais c'est une autre histoire

2 Etude attendue

Nous attendons de vous :

- que vous décriviez le jeu de données avec des tables et des graphiques adaptés et commentés (les tables et les graphiques devront être générés avec R) ;
- que vous calculiez le pourcentage de lignes dupliquées pour chaque programme (et vous ajouterez une colonne à votre jeu de données avec cette information) ;
- que vous compariez le nombre de lignes dupliquées des programmes écrits en C et en C++ (comparaison avec un test statistique duement justifié) ;
- que vous donniez un intervalle de confiance à 90% du taux de lignes duppliquées d'un programme écrit en C ou C++ ou C++ ;
- que vous exploriez l'hypothèse selon laquelle les programmes écrits avec des codes très longs (nombre de lignes de code supériuer à la médiane) comporteraient plus de warnings à la compilation que les codes plus courts (comparaison avec un test statistique et un ou des graphiques) ;
- qu'à partir de ce jeu de données, vous posiez deux questions auxquelles vous répondrez avec une analyse statistique adaptée.

Si vous le souhaitez, vous pouvez compléter ce jeu de données et y intégrant de nouveaux programmes. Vous n'êtes pas obligés de remplir toutes les colonnes du tableau pour chaque programme ajouté, vous pouvez très bien introduire des données manquantes[3].

3 Rendu du projet

Vous rendrez un document de 4-5 pages par binôme, en LaTeX mode *fullpage* avec une taille de police de 11. Vous rédigerez votre rapport en différenciant bien vos conslusions et vos interprétations ou hypothèses. Nous serons attentifs à la qualité de la rédaction. Vous placerez en annexe le code R (commenté) que vous avez utilisé pour répondre aux différentes questions. Nous serons attentifs à la qualité de votre code également.

date de remise du rapport dimanche 25 mars, avant minuit.

mode de remise du rapport envoi du pdf + fichier de données, par email à `emmanuelle.becker@univ-rennes1.fr`

[3] rappel: pour indiquer qu'une données est manquante, on note `NA` pour *not available*