Cahier des charges

Création d'une application R Shiny pour la visualisation des données de RNA-seq du puceron du pois *Acyrthosiphon pisum*

Arnaud DUVERMY, Maud FERRER, Romuald MARIN
Master 2 Bioinformatique

09 Octobre 2020

Table des matières

1 Présentation du projet

1.1 Contexte

Les insectes représentent 90% des espèces animales connues, colonisant une grande partie des habitats terrestres [1]. Ce succès écologique peut s'expliquer en partie par leur capacité à se développer dans des environnements pourtant nutritionnellement carencés en acides aminés et autres nutriments essentiels comme la sève des plantes [2]. Leur survie au sein de ces niches trophiques est rendue possible grâce à leur fréquente association avec des microorganismes qui complémentent leur alimentation. Ces associations durables entre plusieurs organismes différents sont appelées symbioses [1].

Le puceron du pois, *Acyrthosiphon pisum*, vit en association symbiotique obligatoire avec une bactérie, *Buchnera aphidicola*, qui lui fournit des composés peu présents dans son milieu nutritif (i.e. acides aminés essentiels et vitamines). En retour, l'insecte procure à la bactérie un approvisionnement permanent en matières nutritives simples comme les sucres [3] et lui assure une niche écologique souvent peu compétitive au sein de cellules spécialisées : les bactériocytes [1]. *Acyrthosiphon pisum* est un insecte nuisible des cultures de légumineuses, notamment le pois et la luzerne. En détournant la sève phloémienne, le puceron peut provoquer l'avortement des fleurs et la diminution du poids moyen des graines. De plus, ce puceron peut être vecteur de certaines maladies virales affectant les plantes cultivées (i.e. la mosaïque jaune du haricot, la mosaïque du pois, la mosaïque du concombre ou encore le luteovirus). Les dégâts ainsi engendrés sur les cultures sont estimés à plusieurs centaines de millions de dollars. L'association symbiotique entre le puceron du pois et *Buchnera aphidicola* est aujourd'hui le modèle par excellence en génomique des interactions notamment en raison de la disponibilité de leurs deux génomes [3], ainsi que des caractéristiques des cellules bactériocytaires dans ce modèle qui facilitent leur étude (i.e. cellules de grande taille, facilement individualisables). Une meilleure compréhension des processus mis en jeu dans cette association indispensable à la survie des deux organismes ouvre ainsi, de nouvelles perspectives pour la mise au point de méthodes alternatives de lutte contre le puceron du pois.

Dans ce contexte, de nombreuses études transcriptomiques sont menées sur ce modèle afin d'étudier la dynamique d'expression des gènes au cours de la vie du puceron, dans un tissu particulier voire en condition de stress. À ce jour, environ 450 projets RNA-seq sont disponibles sur ce modèle. Face à cette quantité de données disponibles, le laboratoire BF2I [1] (Biologie Fonctionnelle, Insectes et Interactions), qui étudie les interactions biologiques complexes impliquant les insectes ravageurs, leurs bactéries symbiotiques intracellulaires et les plantes hôtes, a entrepris le développement d'une plateforme de visualisation des données RNA-seq. L'application existante, développée en R grâce au package Shiny [2], permet l'analyse de plusieurs jeux de données RNA-seq d'*Acyrthosiphon pisum* et offre la possibilité de visualiser et de comparer les valeurs d'expressions de gènes bruts ou normalisées au sein d'un même jeu de données ou entre différents jeux de données. L'application Shiny, déployée par le laboratoire BF2I, permet donc déjà de démocratiser l'accès et l'analyse des données transcriptomiques d'*Acyrthosiphon pisum*.

À la demande du laboratoire BF2I, nous prendrons en main, au cours des prochaines semaines, le package Shiny pour améliorer l'application existante, et étoffer ainsi les possibilités de traitement et de visualisation des données, afin d'améliorer l'accessibilité et la compréhension des résultats.

1. `https://bf2i.insa-lyon.fr/`
2. `https://shiny.rstudio.com/`

1.2 Description de l'existant

L'application Shiny est codée en R. C'est un langage de programmation libre, destiné aux études statistiques et à la science des données. Il possède de très nombreux packages permettant de réaliser de nombreuses tâches. Le package Shiny permet de créer des applications web. Il possède deux parties : un côté UI (User Interface) qui regroupe tous les éléments de mise en forme et d'affichage de l'interface utilisateur, et un côté serveur où sont exécutés les codes R qui servent à produire les sorties (graphiques, tables, traitements, etc.) et à les mettre à jour en cas de changement dans les valeurs d'entrée saisies dans l'application (voir Figure 1).

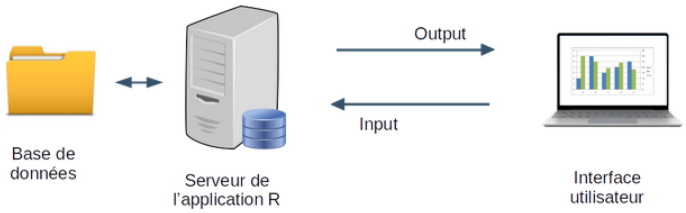

FIGURE 1 – Fonctionnement d'une application web Shiny

Les données génomiques disponibles en entrée de l'application existante (voir Figure 2) proviennent à la fois d'études transcriptomiques réalisées par le laboratoire et de données déjà publiées sur les bases de données en ligne. Les données qui ont été mises à disposition sont des matrices de comptage permettant de mesurer l'expression des différents gènes du puceron du pois dans plusieurs conditions différentes.

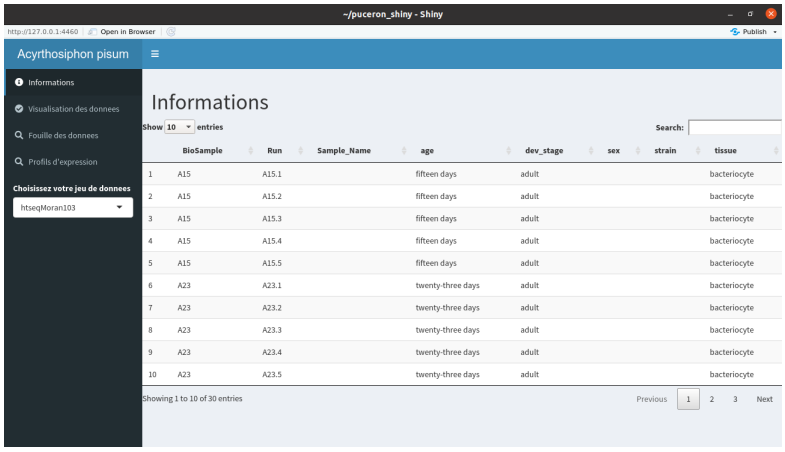

FIGURE 2 – Interface de l'application actuelle

1.3 Objectif

L'amélioration de cette application se fera à plusieurs niveaux :

1. Fonctionnalités :
   (a) Ajout de nouvelles données de manière à compléter la liste des jeux de données disponibles ;
   (b) Ajout d'une nouvelle méthode de normalisation des comptages bruts.
2. Implémentation :
   (a) La création d'un git permettant d'avoir une gestion des différentes versions du code ;
   (b) Corriger les éventuels bugs et mettre à jour le code.
3. Interface :
   (a) L'uniformisation des différentes visualisations sur une seule page afin de rendre la visualisation plus synthétique ;
   (b) Tenter de rendre l'application plus esthétique et les graphiques plus informatifs ;
   (c) Nous échangerons avec les différents utilisateurs de l'application via un questionnaire de manière à récolter les critiques et améliorations potentielles afin de rendre l'application plus ergonomique et intuitive.

2 Expression des besoins

2.1 Besoins fonctionnels

L'objectif de ce projet est de finaliser l'application R Shiny existante et d'y intégrer l'ensemble des données transcriptomiques disponibles sur le modèle du puceron du pois *Acyrthosiphon pisum*.
Environ 450 librairies sont disponibles sur le NCBI [3]. Ces librairies sont en cours de traitement (mapping, comptage des lectures par gène) pour obtenir les nouvelles matrices de comptage qui constitueront les données disponibles en entrée et qui pourront être analysées par l'application R Shiny.
Ainsi, en premier lieu, nous tâcherons d'ajouter ces nouvelles librairies à l'application puis d'implémenter une procédure simple et efficace pour intégrer des nouvelles matrices de comptage. Après une première interview avec le maître d'ouvrage, il serait aussi intéressant de fusionner les onglets "**Fouille de données**" et "**Profils d'expression**" qui apparaissent comme redondant du point de vue des informations que l'utilisateur doit remplir pour visualiser les données transcriptomiques.

Afin d'étudier l'expression différentielle des gènes, l'application existante offre la possibilité de comparer les comptages bruts ou bien de les normaliser au préalable suivant la méthode du RPKM (Reads Per Kilobase of transcript per Million reads mapped) ou du TPM (Transcripts Per Million). En toute rigueur, il est préférable de normaliser les comptages bruts avant de les comparer, afin de limiter les biais évidents liés à la taille des transcrits ou encore à l'expérimentation (biais d'enrichissement d'une librairie à une autre). Les méthodes du RPKM et du TPM tiennent compte de la longueur des transcripts étudiés ; elles permettent ainsi de comparer l'expression des gènes au sein d'une même librairie. En outre, la méthode du TPM permet aussi de limiter les biais d'enrichissement d'une librairie à l'autre et donc de comparer l'expression des gènes entre les librairies. Cependant, la méthode du TPM apparaît, dans la littérature, comme une méthode moins performante que celle proposée par le package R d'analyse d'expression différentielle DESeq2 [4] dans l'objectif de comparer l'expression des gènes dans différentes conditions. Ce package utilise la méthode du ratio des médianes. Pour chaque gène, une pseudo-référence est créée, en prenant la moyenne géométrique des comptages par gène de tous les échantillons. Pour chaque gène de chaque échantillon, le rapport échantillon / référence est alors calculé. La médiane des rapports de chaque échantillon est ensuite utilisée comme facteur de normalisation de chaque librairie pour calculer les comptages normalisés de chaque gène de chaque librairie. Avec cette méthode, les comptages bruts sont divisés par des facteurs de taille spécifiques à l'échantillon ce qui permet de corriger le biais de profondeur de séquençage et donc la comparaison efficace de l'expression des gènes entre librairies. L'ajout à l'application Shiny de la méthode de normalisation des comptages bruts proposée par DESeq2 permettrait donc d'améliorer la caractérisation des gènes différentiellement exprimés dans différentes conditions.

2.2 Besoins non fonctionnels

Le code source de l'application sera développé sous R Shiny et sera déposé sur GitHub. *Shiny* est un package R, développé par RStudio, qui permet la création de pages web interactives sur lesquelles il est possible de réaliser toutes les analyses disponibles sous R. Dans le cadre transcriptomique du projet, l'application utilisera le package DESeq2 pour mener les analyses sur les comptages et étudier les différences d'expression des gènes. En ce qui concerne le développement, nous envisageons d'utiliser les packages :

— *tidyverse [5]* afin d'assurer l'efficacité et la lisibilité du code ;

3. `https://www.ncbi.nlm.nih.gov/`

— *ggplot2 [6]* : certaines figures sont déjà construites à l'aide de ggplot2, nous tâcherons d'harmoniser le code et de toutes les construires avec cet outil ;
— *plotly [7]* afin d'assurer une meilleur interactivité des figures.

Les autres packages déjà utilisés par l'application seront a priori conservés (DT, reshape2). À terme, l'application devrait être mise à jour sur le serveur de calculs géré par le laboratoire BF2I.

Enfin, d'autres modifications seront envisagées suites aux retours d'expériences des utilisateurs du BF2I, obtenus grâce à un questionnaire qui leur sera transmis. Ce questionnaire nous permettra de cibler au mieux les modifications et les fonctionnalités à ajouter sur l'application existante.

3 Contraintes

3.1 Coûts

Le laboratoire commanditaire du projet est le BF2I. C'est une Unité Mixte de Recherche entre l'Institut National des Sciences Appliquées de Lyon [4] (INSA) et l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement [5] (INRAE).

Les moyens matériels mis à disposition par le laboratoire sont la totalité du code nécessaire à la réalisation de l'application Shiny, les différents fichiers des matrices de comptage ainsi qu'un serveur appartenant au laboratoire afin de déployer l'application.

3.2 Délais

Un questionnaire va être envoyé aux utilisateurs pour sonder les besoins en nouvelles fonctionnalités. En fonction des réponses, le planning prévisionnel (cf Planifications) pourra être modifié.

Des rendez-vous réguliers avec le maître d'ouvrage seront pris pour discuter de l'avancée des implémentations.

Enfin, le délivrable (la totalité des scripts et la documentation associée) sera rendu le 17 décembre en vue d'une présentation le 17-18 décembre.

3.3 Autres contraintes

Dans le contexte sanitaire actuel, les séances de travail en groupe seront réalisées en respectant les gestes barrières mais le télétravail sera privilégié.

4. `https://www.insa-lyon.fr/`
5. `https://www.inrae.fr/`

4 Déroulement du projet

4.1 Planifications

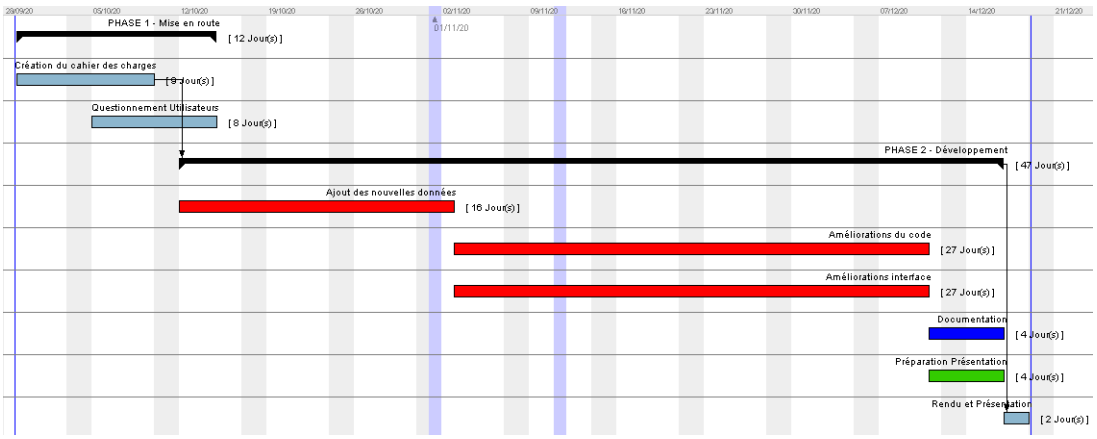

FIGURE 3 – Diagramme de Gantt représentant l'organisation prévisonnelle du projet

Ce diagramme est informatif et nous ferons en sorte qu'il soit respecté. Si tel n'est pas le cas, des rendez-vous seront pris avec les responsables du projet pour revoir les tâches et modifier, le cas échéant, la planning prévisonnel.

4.2 Plan d'assurance qualité

L'application étant déjà fonctionnelle, notre objectif est d'obtenir un rendu plus ergonomique et adapté à l'utilisation des chercheurs. La mise en place d'une documentation sera aussi un plus pour des potentiels utilisateurs novices sur le fonctionnement de l'outil DESeq2.

Des rendez-vous réguliers seront pris avec les porteurs de projet pour s'assurer que l'avancement du projet se déroule à leur convenance. Le GitHub sera également mis à jour grâce à des commits pour le suivi de nos avancées.

4.3 Documentation

La documentation interne du projet sera réalisée directement dans les fichiers en suivant les règles de bonnes pratiques informatiques via l'utilisation de commentaires, afin d'être compréhensible et de permettre de rapidement apporter de futures améliorations. Les utilisateurs auront des informations sur les fonctionnalités de l'interface web via la mise en place d'un README sur GitHub.

4.4 Responsabilités

4.4.1 Maîtrise d'ouvrage

Les prestations délivrées en exécution du présent cahier des charges le sont pour le compte du BF2I.
Maître d'ouvrage : Nicolas Parisot (nicolas.parisot@insa-lyon.fr)

4.4.2 Maîtrise d'oeuvre

La réalisation de ce projet s'effectue dans le cadre de l'UE Projet 3 du Master *Bioinformatique : Méthodes et Analyses Moléculaires* de l'Université Claude Bernard Lyon 1. Les étudiants en charge de ce projet sont :

— Arnaud DUVERMY (arnaud.duvermy@etu.univ-lyon1.fr)
— Maud FERRER (maud.ferrer@etu.univ-lyon1.fr)
— Romuald MARIN (romuald.marin@etu.univ-lyon1.fr)

Références

[1] E. Akman Gündüz and A. Douglas, "Symbiotic bacteria enable insect to use a nutritionally inadequate diet," *Proceedings of the Royal Society B : Biological Sciences*, vol. 276, pp. 987–991, Mar. 2009.

[2] H. Malke, "Paul Buchner, Endosymbiosis of Animals with Plant Microorganisms. 909 S., 371 Abb., 5 Tab., 6 Taf. New York 1965 : John Wiley & Sons, Inc. : Interscience Publ. $ 35.00," *Zeitschrift für allgemeine Mikrobiologie*, vol. 7, no. 2, pp. 168–168, 1967.

[3] H. Feng, N. Edwards, C. M. H. Anderson, M. Althaus, R. P. Duncan, Y.-C. Hsu, C. W. Luetje, D. R. G. Price, A. C. C. Wilson, and D. T. Thwaites, "Trading amino acids at the aphid– *Buchnera* symbiotic interface," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 16003–16011, Aug. 2019.

[4] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, Dec. 2014.

[5] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the Tidyverse," *Journal of Open Source Software*, vol. 4, p. 1686, Nov. 2019.

[6] H. Wickham, *ggplot2 : elegant graphics for data analysis*. Use R !, Cham : Springer, second edition ed., 2016. OCLC : 958058958.

[7] P. T. Inc., "Collaborative data science," 2015.