Enseignement par projet en master de biologie

Adeline Samson [1] & Bernard Ycart [2]

[1] *LJK, Université Grenoble Alpes; adeline.leclercq-samson@univ-grenoble-alpes.fr*
[2] *LJK, Université Grenoble Alpes; bernard.ycart@univ-grenoble-alpes.fr*

Résumé. Nous présenterons dans cette communication une expérience pédagogique d'enseignement de la statistique à des étudiants de master de biologie. L'objectif pédagogique était d'amener les étudiants à être autonome pour l'analyse d'une base de données "omiques" en grande dimension avec le logiciel R. Notre choix s'est porté sur la préparation d'un support de cours en ligne, mêlant introduction des notions statistiques, des commandes R, illustrées d'exemples à l'appui et d'interprétations. Les séances étaient ensuite consacrées à la mise en pratique sur ordinateur, pour moitié sous forme d'exercices encadrés, pour moitié sous forme de la préparation d'un projet sur données réelles.

Mots-clés. Biostatistique, grande dimension, enseignement par projet ...

Abstract. We will present in this communication a tentative of a biostatistics lecture in a Biology Master. From a pedagogical point of view, the objective was to make the students autonomous to analyse a "omic" data base in large dimension, with the R software. Our choice has been to prepare a website with the introduction of the statistical definitions, the R instructions, illustrated with examples and their interpretations. Only lab sessions were organized, half of them focused on guided exercises, half of them on the preparation of a real data analysis.

Keywords. Biostatistics, large dimension, teaching through project ...

1 Introduction

Nous nous intéresserons dans cette communication à un cours de biostatistique pour des étudiants de première année de Master de biologie. Dans ce master international, les étudiants viennent de différents horizons et n'ont pas tous suivi des cours de statistique en Licence. Deux consignes nous avaient été données par les collègues biologistes responsables du master : ne pas demander de pré-requis pour ce cours; réconcilier les étudiants avec la statistique, dont la maîtrise est un enjeu très fort dans leur futur métier de chercheur. Nous étions libre de choisir le contenu, le format et les objectifs pédagogiques. Nous avions à notre disposition 4h30 par semaine, pendant 9 semaines. Forts de cette liberté, nous avons élaboré un cours basé sur une pratique intensive du logiciel R, avec comme objectif final l'autonomie des étudiants à analyser une base de données "omiques" en grande dimension.

2 Contenu scientifique du module

Les grandes bases de données "omiques" abordées dans ce cours se présentent sous la forme de deux tables : une première table avec les mesures d'expression des gènes (ou protéine, ARN, etc) pour chaque individu, une table phenotypique de variables cliniques mesurées pour chaque individu. Nous disposons en général d'un nombre d'individus N de l'ordre de 200, pour un nombre de gènes p de l'ordre de 20 000, et d'un nombre de covariables phénotypiques q de l'ordre de la dizaine. L'analyse de ces bases consiste à identifier les gènes sur-exprimés ou étant liés à une variable réponse (décès/vivant, temps de survie, taux de cholestérol, etc).

Ces analyses requièrent donc de connaître en statistique

- les tests paramétriques et non-paramétriques standards
- les méthodes d'ajustement de p-valeurs en présence de tests multiples ($p = 20\ 000$ gènes)
- les méthodes de régression : modèle linéaire, analyse de la variance, régression logistique, modèle de Cox
- les méthodes de sélection de variables en grande dimension ($N << p$) pour des modèles de régression : méthode lasso, elastic-net.

Ceci nécessite au préalable des notions de statistique descriptive, de probabilités, de statistique inférentielle et de p-valeurs.

L'analyse de ces bases passe aussi par la maîtrise d'un logiciel de traitement statistique des données. Nous avons choisi le logiciel R, qui au travers de ses paquets, met à disposition de l'utilisateur toutes les notions citées ci-dessus.

Le plan du cours a donc été le suivant (le cours a été réalisé en anglais):

1. Getting started with R
2. Descriptive statistics
3. Probability
4. Estimation and confidence intervals
5. Statistical tests
6. One Sample tests
7. Two Sample tests
8. Multiple testing
9. Regression models and selection in high dimension

3 Organisation du module

L'objectif scientifique du module étant ambitieux (méthodes lasso, modèle de Cox), nous avons opté pour une pédagogie inversée.

Un support de cours a été préparé en ligne `https://toltex.u-ga.fr/bio`, avec tous les supports de cours, introduisant les notions de statistique du chapitre, les commandes R mettant en oeuvre ces notions, immédiatement suivies d'exemples sur des bases de données réelles et l'interprétation des résultats. Le cours utilise très peu de formules mathématiques et insistent sur les notions et leur intuition. Ce site a été créé via des outils de génération automatique de pages web (Rmarkdown, Ruby).

Les étudiants avec pour consigne de lire des parties de chapitre d'une séances sur l'autre. Les séances étaient dédiées à la mise en pratique des notions introduites sur le site pédagogique. La moitié des séances suivaient des énoncés d'exercices qu'on pourrait qualifier de classiques. Chaque exercice utilise une base de données réelles, pas nécessairement dans le domaine "omique", mais suffisamment pédagogique pour amener à la réflexion de l'étudiant (exemple typique : la base de données des passagers du Titanic). Les exercices sont guidés avec des questions précises, volontairement répétitives. L'ensemble des codes R de solution des exercices a été fourni au début du semestre. Ceci permettait à l'étudiant de pouvoir avancer en toute autonomie. Fournir l'ensemble des solutions peut surprendre, mais ceci s'est avéré un bon choix pédagogique. Pendant les premières séances, le langage informatique (non maitrisé par les étudiants) n'était pas un frein puisque le code R des solutions était fourni. Ceci a permis que chaque étudiant passe du temps à assimiler les premières notions statistiques sans se préoccuper du code. Comme pour l'apprentissage d'une langue étrangère, l'exécution des lignes de commandes R a peu à peu été assimilée par tous.

Cet apprentissage du langage R et l'autonomie acquise par les étudiants a été facilité par l'organisation de l'autre moitié des séances de pratique. Pendant ces autres séances, les étudiants travaillaient en binôme sur une base de données génomiques GEO issues du site `https://www.ncbi.nlm.nih.gov/geo/`. L'objectif de ces projets était de produire à la fin du semestre une analyse complète de la base de données. Les étudiants devaient donc appliquer sur leur jeu de données les notions vues sur le cours et sur des exercices d'entrainement (et sur des bases de données "jouets"). L'enseignant était présent pour répondre à leurs questions, mais devait les laisser en autonomie faire leur choix d'analyse statistique.

Les étudiants se sont pris au jeu de cette analyse d'une base de données génomiques, en travaillant la bibliographie de la pathologie concernée, en allant étudier spécifiquement les gènes connus pour être sur-exprimés dans ce contexte. Plusieurs étudiants ont exprimés avoir eu le déclic pour la statistique, ce que cela pouvait leur apporter, lors de ces séances de projet.

L'évaluation s'est déroulée sous la forme de plusieurs rapports intermédiaires de l'analyse statistique de la base de données GEO et d'une soutenance orale. Les rapports ont été rédigés en RMarkdown par les étudiants.

4 Conclusion

Cette première expérience a été enrichissante. Les retours des étudiants sont enthousiasmants et constructifs pour améliorer l'organisation et la cadence des séances dans les années à venir. Le choix de ne pas faire de séances de cours magistral et de faire le pari du sérieux et de l'autonomie des étudiants à travailler sur un support en ligne entre les séances s'est avéré payant. Le volume horaire élevé passé sur R (4h30 par semaine) a permis aux étudiants d'être rapidement autonomes. Les étudiants ont trouvé une motivation dans les projets génomiques et ont donc souhaité comprendre les notions de tests multiples, de sélection de variables en grande dimension, pour pouvoir répondre à la question biologique.