Analyse d'une expérience pédagogique fondée sur l'étude de cas simulés

Olivier François[1] & Michael Blum[1]

[1] *Université Grenoble-Alpes, Laboratoire TIMC-IMAG UMR CNRS 5525, Faculté de Médecine, 38042 Grenoble, France,* olivier.francois@imag.fr, michael.blum@imag.fr.

Résumé. Nous décrivons les résultats d'une expérience pédagogique s'appuyant sur l'analyse de données simulées et sur l'apprentissage actif. Pour effectuer les analyses, les participants à l'expérience pouvaient utiliser un ensemble de logiciels statistiques ayant fait l'objet d'une présentation préalable. Les réponses des participants étaient ensuite évaluées automatiquement par un score de performance supposé décroître avec la difficulté des problèmes proposés. Dans cet article, nous évaluons la variabilité des scores obtenus par les participants lors de trois challenges distincts. Nous concluons que la variabilité des résultats peut être expliquée par le niveau d'expertise auto-déclarée des participants et par le grade des difficultés à l'intérieur de chaque challenge.

Mots-clés. Apprentissage par challenge, apprentissage actif, logiciels statistiques, applications en génétique des populations.

Abstract. We report results from an educational experiment based on active learning and simulated data. To perform their analysis, participants in the experiment could use a set of statistical programs that were presented beforehand. Responses were then evaluated automatically by a performance score which was assumed to be decreasing with the difficulty of the proposed problem. In this article, we evaluate the variability of the scores obtained by the participants for three distincts challenges. We find that the variability of the scores is explained by the self-reported level of expertise of the participants and by the level of difficulties within each challenge.

Keywords. Learning from challenges, active learning, statistical software, applications to population genetics.

1 Introduction

Cet article présente les résultats d'une expérience pédagogique qui s'est déroulée lors de l'école d'été *Software and Statistical Methods in Population Genomics* (SSMPG 2015) organisée à Aussois du 7 au 11 septembre 2015 [1]. Destinée à un public de chercheurs dans le domaine de la biologie de l'évolution, SSMPG 2015 présentait à ses participants six méthodes statistiques pour le criblage génomique en génétique des populations ou

en écologie moléculaire. L'école d'été a accueilli 53 participants dont cinq instructeurs. L'apprentissage des méthodes statistiques s'appuyait sur la mise en situation pratique et l'analyse de cas d'étude simulés (Waldrop, 2015). Les réponses à fournir se présentaient sous la forme d'une liste de gènes candidats à extraire des cas d'étude proposés lors de l'école. Un score associé à chaque liste de gènes candidats était calculé à partir des gènes cibles à l'aide d'un site web spécifiquement dédié à la gestion des réponses aux challenges. Les détails des simulations n'étaient connus que d'une seule personne dont le rôle était neutre. Dans cet article, nous évaluons l'intérêt pédagogique de cet enseignement en comparant les scores d'utilisateurs expérimentés à ceux d'utilisateurs moins expérimentés. Nous tentons d'expliquer la variabilité des scores obtenus en distinguant la part de cette variabilité due à la difficulté des problèmes (et à la puissance des méthodes) de celle due au niveau d'expertise initiale des utilisateurs.

2 Matériels et méthodes

Trois cas d'étude de difficulté croissante ont été créés et mis en ligne pour les besoins de l'école SSMPG 2015. Le premier challenge avait pour but de tester l'installation des logiciels, le processus de soumission des résultats ainsi que la bonne compréhension des objectifs de l'école. Deux autres cas servaient de tests pratiques pour l'utilisation de méthodes statistiques. La simulation des données avait été effectuée antérieurement à l'école par Katie Lotterhos (Boston Northeastern University), en définissant une espèce fictive. L'histoire démographique de l'espèce modèle fut dévoilée aux participants. En bref, l'espèce était adaptée à la vie en haute altitude. Elle avait survécu à une période de glaciation dans deux refuges, puis avait colonisé un troisième refuge lors d'une phase de réchauffement climatique. Enfin, après la fonte des glaces, elle avait colonisé l'ensemble de la zone géographique étudiée. Les participants disposaient de données génétiques (4000 à 6000 locus ou gènes) simulées pour 500 individus répartis dans 19 sites géographiquement distincts.

L'objectif des approches statistiques présentées était de détecter, à partir des données, les signatures génétiques pertinentes pour l'adaptation de l'espèce à son environnement. Les listes de gènes cibles contenaient $m = 12$ ou $m = 36$ locus. Six méthodes statistiques ont été présentées aux utilisateurs (OUTFLANK, FLK, PCADAPT, SELESTIM, LFMM, BAYPASS). Les six méthodes sont décrites dans la revue de François *et al.* (2016). Chaque méthode permet de calculer une statistique de test pour chaque locus. A partir des tests, des listes de gènes candidats peuvent être proposées. Pour chacun des 3 cas d'étude, un utilisateur pouvait soumettre une liste pour chaque méthode. Une liste soumise par un utilisateur était évaluée à l'aide un score qui n'était rendu public qu'à la fin du challenge. Le score d'une liste de gènes candidats était calculé de la manière suivante :

$$F = 2\,\frac{\text{Power}(1-\text{FDR})}{\text{Power} + (1-\text{FDR})}\,.$$

Dans cette formule, Power représente la sensibilité et FDR représente le taux de fausse découverte. Le score varie entre 0 et 1. Lorsque la taille de la liste soumise est proche de la taille de la liste de gènes cibles, le score peut être interprété comme une mesure de puissance statistique.

Quelques subtilités s'ajoutent à cette description. Les instructeurs participaient aux 3 challenges, mais ils ne connaissaient pas les bonnes réponses (gènes cibles). Un participant pouvait poser des questions aux instructeurs, mais pas à l'arbitre. Il était permis de combiner des méthodes, voire d'en inventer de nouvelles. Lors de la soumission des résultats sur le site internet, un utilisateur déclarait son niveau d'expertise initiale, *expérimenté* ou *moins expérimenté*.

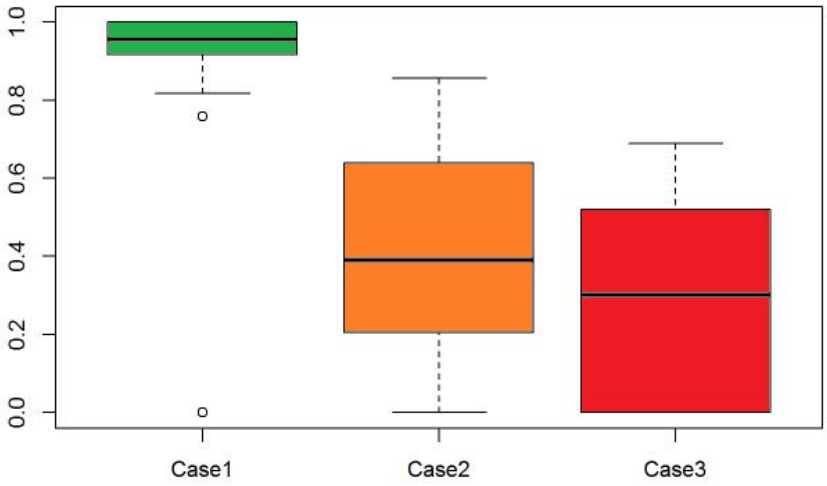

Figure 1: *Distributions des scores pour chaque challenge.*

3 Résultats

Pour répondre aux challenges, les participants étaient autorisés à partager leurs approches et à travailler en équipes. Au total, 20 équipes se sont formées et 185 listes de gènes candidats ont été soumises sur le site web de l'école. Concernant les cas d'étude les plus difficiles (cas 2 et 3), 124 listes de gènes candidats ont été soumises. Pour notre analyse, 112 soumissions ont été retenues, après un filtrage de listes vides ou détection d'erreurs de manipulation. Pour le cas 2, 56 soumissions correspondaient à des utilisateurs moins expérimentés (36 pour le cas 3) et 10 soumissions correspondaient à des utilisateurs

expérimentés (idem pour le cas 3). Le nombre d'utilisation de chaque méthode était réparti de manière équilibrée entre les 6 approches proposées.

Pour vérifier l'augmentation de difficulté, nous avons représenté la distribution des scores pour chaque challenge (Figure 1). Le score médian était égal à 0.96 pour le challenge 1, à 0.40 pour le second challenge, et à 0.31 pour le troisième challenge. Les scores ont montré une grande variabilité, avec des écarts-types égaux à 0.16, 0.23 et 0.24 respectivement. Dans le cas 2, le score moyen d'un utilisateur expérimenté était égal à 0.60 et le score d'un utilisateur moins expérimenté était égal à 0.39 (t-test : p-valeur = 0.05). Dans le cas 3, le score moyen d'un utilisateur expérimenté était égal à 0.35 et le score d'un utilisateur moins expérimenté était égal à 0.30 (t-test : p-valeur = 0.45).

Nous avons ensuite estimé des probabilités de détection de chacun des m gènes cibles dans les cas 2 et 3 (Figure 2). Dans le cas 2, la différence entre utilisateurs expérimentés et moins expérimentés était perceptible (Figure 2A). La différence entre les deux catégories d'utilisateurs était maximale pour des gènes cibles de difficulté de découverte intermédiaire, entre 40% et 60%. Dans le cas 3, la différence entre utilisateurs expérimentés et moins expérimentés était nettement moins perceptible (Figure 2B).

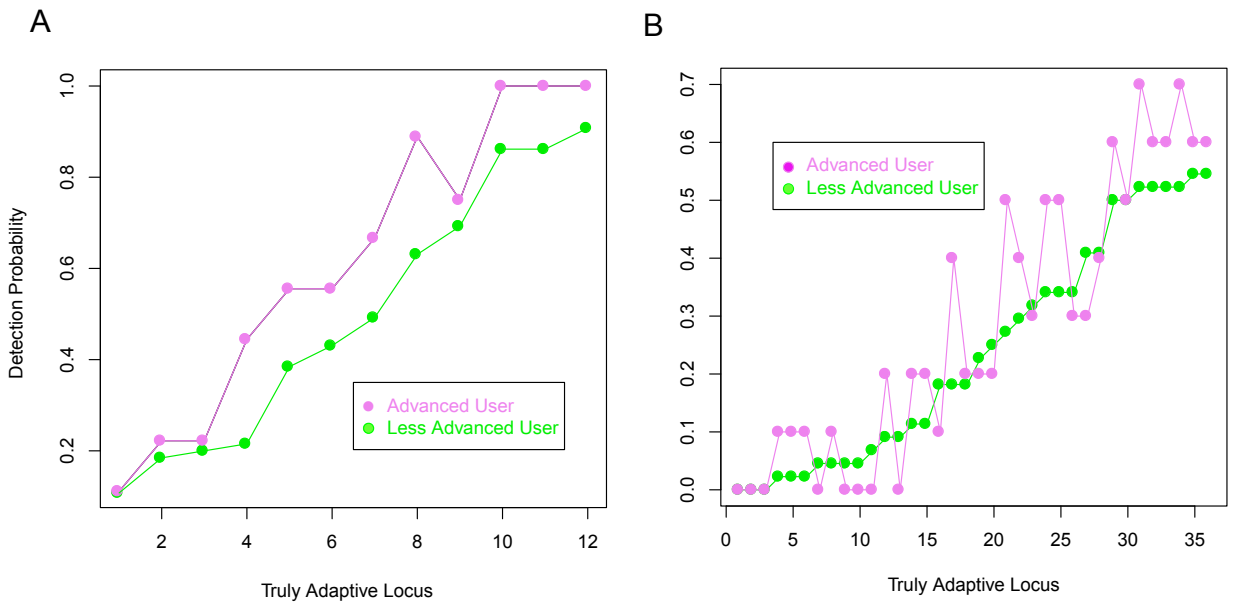

Figure 2: *Comparaison des probabilités de détection des gènes cibles (scores) pour des utilisateurs expérimentés et moins expérimentés. A) challenge 2, B) challenge 3.*

4 Discussion

L'expérience d'apprentissage actif menée durant l'école d'été SSMPG 2015 s'est avérée utile (Freeman *et al.* 2014). Cette expérience a conduit les participants à une appropriation rapide des logiciels statistiques, et des différences de scores peu marquées entre utilisateurs expérimentés et moins expérimentés ont été observées. En détournant la parole de Baudrillard, nous pourrions affirmer que la simulation précède le réel, possédant ainsi une valeur productrice (Baudrillard 1981). Toutefois le résultat le plus marquant de notre analyse est la grande variabilité des scores (Figure 1), montrant que différents utilisateurs d'une même panoplie de logiciels statistiques peuvent fournir des réponses variables sur des données identiques.

Afin d'interpréter les résultats et d'en extraire les messages importants, introduisons un modèle probabiliste très simple. Pour un challenge donné, supposons que la liste de gènes cibles contienne m éléments ($m = 12$ ou $m = 36$ dans les cas 2 et 3). Lors de SSMPG 2015, les longueurs des listes soumises étaient voisines de m. Pour chaque gène (ou locus) cible, ℓ, notons p_ℓ la probabilité qu'un utilisateur découvre ce gène, c.à.d., décide d'inclure le gène dans sa liste de candidats. Sous cette hypothèse, l'espérance du score est égale à $\sum_\ell p_\ell/m$. Supposant les tests indépendants, la variance des scores est égale à $\sum_\ell p_\ell(1-p_\ell)/m^2$. Ces résultats indiquent que la variabilité des scores est directement reliée aux probabilités d'identifier correctement chaque gène cible. La variabilité des scores est faible lorsque les gènes cibles sont faciles ou difficiles à trouver (p_ℓ voisin de 0 or 1) et elle est maximale pour les cibles de difficulté intermédiaire.

Considérant deux catégories d'utilisateurs, expérimentés, A, et moins expérimentés, $\bar{A}$, les calculs précédents se précisent de la manière suivante. Soit π_A la proportion d'utilisateurs expérimentés. Pour chaque gène cible, ℓ, il y a une probabilité, $p_{\ell A}$, pour que la cible ℓ soit correctement identifiée par un utilisateur expérimenté. Soit $p_{\ell\bar{A}}$ la probabilité pour que ℓ soit correctement identifiée par un utilisateur moins expérimenté. Nous avons

$$\mathrm{E}[F] = \mathrm{E}[F|A]\pi_A + \mathrm{E}[F|\bar{A}](1-\pi_A),$$

où l'espérance $\mathrm{E}[F|A]$ est égale à $\sum_\ell p_{\ell A}/m$ (formule similaire pour $\mathrm{E}[F|\bar{A}]$). La variance des scores se décompose de la manière suivante

$$\mathrm{var}[F] = \mathrm{Var}[F|A]\pi_A + \mathrm{Var}[F|\bar{A}](1-\pi_A) + (\mathrm{E}[F|A] - \mathrm{E}[F|\bar{A}])^2\pi_A(1-\pi_A)$$

où $\mathrm{Var}[F|A] = \sum_\ell p_{\ell A}(1-p_{\ell A})/m$ (formule similaire pour $\mathrm{Var}[F|\bar{A}]$).

Les formules ci-dessus permettent quelques commentaires immédiats. Tout d'abord, les résultats de la Figure 1 sont en accord avec un modèle dans lequel la variabilité des scores peut être faible dans les cas faciles ou difficiles et plus élevé dans les cas intermédiaires. Par exemple, les résultats observés pour le cas 2 sont attendus si un tiers des gènes cibles sont facilement détectés, un tiers correspondent à des probabilités intermédiaires ($p_\ell \approx 1/2$) et un tiers sont difficiles à détecter (Figure 2A). Ces résultats

suggèrent que la variabilité des scores est une caractéristique inhérente des cas d'étude et des méthodes utilisées. Quels efforts demander alors aux utilisateurs afin de réduire la variabilité des scores ?

Remarquons que certains efforts visant à réduire la variabilité des scores peuvent aussi entrainer une réduction de performance des méthodes. Encourager des pratiques conservatrices tendraient à homogénéiser les différences entre utilisateurs expérimentés et moins expérimentés et diminueraient effectivement la variabilité des scores. Mais de telles pratiques se traduiraient par une augmentation du nombre de faux négatifs et réduiraient l'espérance du score. Une autre stratégie consiste à améliorer l'expertise des utilisateurs. Bien qu'accroître l'expertise des utilisateurs est souhaitable, la variabilité des scores n'en serait pas nécessairement diminuée. Par exemple, en présence de 5% d'utilisateurs expérimentés, $p_{\ell A} = 1/2$ et $p_{\ell \bar{A}} = 1/10$, la variance de F est autour de 1/10. Cette valeur est inférieure à la valeur obtenue si tous les utilisateurs sont expérimentés (variance totale de 1/4).

En fin de compte, le message évident de cette expérience est de promouvoir l'utilisation de méthodes statistiques puissantes et d'améliorer simultanément l'expertise des utilisateurs. La première action est l'objectif des développements méthodologiques des logiciels statistiques en génomique des populations. Ces développements devraient être accompagnés de directives d'utilisation claires et pratiques. La deuxième action exige que les utilisateurs développent leurs propres compétences statistiques et informatiques afin de suivre l'évolution rapide de méthodes dont la complexité va en croissant avec le temps. L'apprentissage actif est un moyen efficace d'atteindre ce but pour les problématiques de génomique des populations abordées lors de l'école SSMPG 2015.

Bibliographie

[1] Software and Statistical Methods for Population Genetics (SSMPG 2015), `http://ssmpg2015.imag.fr`

[2] Waldrop, M. M. (2015), Why we are teaching science wrong, and how to make it right. *Nature*, 523(7560), 272.

[3] François, O., Martins, H., Caye, K., Schoville, S.D (2016), Controlling false discoveries in genome scans for selection. *Molecular Ecology* 25 (2), 454–469.

[4] Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., Wenderoth, M. P. (2014), Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.

[5] Baudrillard, J. (1981), *Simulacres et simulation*, Paris, Galilée.