

Proposition de stage de Master Biostatistique

Modèles linéaires généralisés hiérarchiques pour variables réponses
catégorielles

Application à l'analyse de la diversité génétique du riz

Encadrants :

Yann Guédon, CIRAD, AGAP

yann.guedon@cirad.fr

Jean Peyhardi, Université de Montpellier, IGF

jean.peyhardi@umontpellier.fr

Catherine Trottier, IMAG

catherine.trottier@umontpellier.fr

Lieu : CIRAD, Campus de Lavalette, Montpellier,

Durée : 4 ou 6 mois à discuter avec l'étudiant,

Indemnités : environ 450 €/mois.

Perspective de thèse à l'issue de ce stage.

Contexte scientifique : L'analyse de la diversité génétique des plantes repose sur des modèles de génétique quantitative où la valeur phénotypique d'un individu, c'est-à-dire le résultat de la mesure effectuée sur cet individu est fonction d'effets génotypiques et environnementaux. Nous proposons dans ce stage d'explorer le paradigme inverse où la variable réponse est le génotype et les variables explicatives sont des variables phénotypiques. L'objet du stage sera d'identifier quelles variables phénotypiques différencient, espèces, sous-espèces et provenances géographiques de riz aussi bien africains qu'asiatiques, sauvages que cultivés. L'approche développée reposera sur des modèles linéaires généralisés hiérarchiques pour variables réponses catégorielles développées dans le cadre de la thèse de Jean Peyhardi (Peyhardi et al., 2015, 2016, 2017).

Sujet : Dans notre contexte, la variable réponse catégorielle (les catégories étant les différents génotypes) est de type nominal car il n'existe pas d'ordre total naturel sur les génotypes. Par contre il est possible de grouper les génotypes fonction des espèces, sous-espèces et provenances géographiques. Cette information de structure sur les catégories réponses, caractérisée par un arbre de partition, est prise en compte par les modèles linéaires généralisés hiérarchiques. Dans un premier temps, l'arbre de partition sera supposé connu à priori et l'objectif sera d'identifier les variables phénotypiques qui expliquent au mieux les séparations entre génotypes de différentes espèces, sous espèces et provenance géographiques. Dans un deuxième temps, l'arbre de partition sera supposé partiellement inconnu (notamment les partitions à une échelle fine correspondant aux séparations entre les différentes accessions de riz) et l'objectif sera de l'inférer. La combinatoire de ce genre de problème étant très élevée, l'objectif sera alors de proposer des heuristiques de sélection de modèles adaptées à ce problème.

Ce travail nécessitera de la programmation en R et en Python pour l'exploration et l'analyse du jeu de données proposé. Ce stage sera effectué au sein de l'équipe M2P2 de l'UMR AGAP du CIRAD.

Références

Peyhardi, J., Trottier, C. & Guédon, Y. (2015). A new specification of generalized linear models for categorical responses. *Biometrika* 102(4), 889–906.

Peyhardi, J., Trottier, C. & Guédon, Y. (2016). Partitioned conditional generalized linear models for categorical responses. *Statistical Modelling* 16(4), 297–321.

Peyhardi, J., Caraglio, Y., Costes, E., Lauri, P-É, Trottier, C. & Guédon, Y. (2017). Integrative models for joint analysis of shoot growth and branching patterns. *New Phytologist*, in press.