PhD position in statistics

Hierarchical generalized linear models for analyzing plant diversity: application to rice diversity

Starting: Autumn 2018;
Duration: 3 years;
Funding : CIRAD/Région Occitanie.

Research Units: Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales (AGAP), CIRAD, Montpellier, France and Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Montpellier, France.
Contact: Yann Guédon (yann.guedon@cirad.fr), Catherine Trottier (catherine.trottier@umontpellier.fr)

Keywords
Generalized linear models; Model selection; Hierarchical statistical models; Plant diversity

Profile and skills required
The candidate must have followed training courses in statistics with specific skills in statistical modeling, computational statistics and/or applied statistics. In particular, notions about generalized linear models, categorical data, variable selection and model selection will be appreciated. In addition, the candidate must be able to invest markedly in software development using C ++, R or Python. Finally, a taste for the application to complex biological data is expected.

Project description
Quantitative genetics relies on specific random-effect models whose functional view is: phenotype = f(genotype) (Gianola, 2007). In this paradigm, there is no restriction on the genotype modeling and complex family structure can be modeled. On the contrary, the modeling of phenotypic traits (response variable in the regression model) is highly constrained (often a single trait or a vector or quantitative traits using linear mixed models). We propose to develop a new statistical modeling framework for the analysis of plant diversity that reverses the functional view which will be genotype = f(phenotype). The aim is to take account of heterogeneous phenotypic traits (categorical, ordinal, counts, quantitative, circular ...) while modeling various genotype families using hierarchies of categories (for instance species subdivided into subspecies, themselves having different geographical origins). This new modeling framework relies on the recently proposed partitioned conditional generalized linear models for categorical responses (Peyhardi et al., 2015, 2016), previously applied to the analysis of the influence of growth on the branching pattern of plants (Peyhardi et al., 2017).

Up to know, hierarchical generalized linear models have been applied to small set of response categories – up to 5 categories of axillary production in Peyhardi et al. (2016; 2017) – for

which biologically meaningful hierarchies of categories could be exhaustively explored. Moreover, a small number of explanatory variables were tested. The analysis of plant diversity requires taking into account a much higher number of categories which induces a very high combinatorics on the possible hierarchies of categories, even taking into account a priori biological information. We want also to incorporate any number of phenotypic variables into the model. From a statistical point of view, this raises new inference questions, especially of model selection. In these models, hierarchies of categories are formalized by a tree of nested partitions. The core of the thesis will therefore be to develop a method of selection (1) of the tree of nested partitions of categories taking into account a priori biological information and (2) of the phenotypic variables associated with each partition (e.g. phenotypic variables differentiating subspecies japonica and indica in the Asian cultivated rice). These new inference methods will be implemented in the hierarchical generalized linear models module of the OpenAlea software platform (Pradal et al., 2008).

The application part of this thesis work will be based on the phenotypic database of rice panicles developed for many years by the UMR DIADE (Al Tam et al., 2013) with different partnerships including the LMI RICE 2, a joint international IRD-CIRAD-UM-USTH-AGI laboratory in Vietnam, CIAT in Columbia and INERA in Burkina Faso. This database integrates the Asian cultivated rice (*Oryza sativa*), its wild parent (*Oryza rufipogon*), as well as the African cultivated rice (*Oryza glaberrima*) and its wild parent (*Oryza barthii*) and continues to be regularly augmented. The proposed statistical modeling will enable to test a wide range of phenotypic variables of heterogeneous natures (quantitative variables: rachis length, cumulative length of the axes; count variables: maximum branching order, total number of grains; binary variable: compact or open character of the panicle ...). This application to the analysis of the rice panicle diversity was at the origin of this thesis subject but the formalism is general and other applications will be searched during the thesis.

References

AL-Tam, F., Adam, H., Anjos, A.D., Lorieux, M., Larmande, P., Ghesquière, A., Jouannic, S., Shahbazkia, H.R. (2013) P-TRAP: a Panicle TRAit Phenotyping tool. *BMC Plant Biology* 13, 122.

Gianola D. (2007). Inferences from mixed models in quantitative genetics. In *Handbook of Statistical Genetics*, 3rd Edition, Vol. 1. John Wiley & Sons, Chichester, West Sussex, England, pp. 678–717.

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* 216(4), 1291–1304.

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.

Pradal, C., Dufour-Kowalski, S., Boudon, F., Fournier, C., Godin, C. (2008). OpenAlea: a visual programming and component-based software platform for plant modelling. *Functional Plant Biology* 35(10), 751–760.