Development of text mining approaches to identify gene-phenotype interactions

(Développement d'approches de fouille de texte pour identifier les interactions génotype-phénotype)

Context: A better understanding of gene-phenotype relationships requires an integration of biological information of various kinds. However, this information is often dispersed in several databases on the Internet, having heterogeneous way of access. For biologists, it is difficult to search these data as the mass of information is hard to manage. In this context, we developed the AgroLD platform [1] (www.agrold.org). AgroLD is a knowledge base covering information on genes, proteins, gene homology predictions, some genetic and phenotypic studies for some species including rice, arabidopsis and wheat species. Currently, AgroLD contains 900 million triples created by transforming more than 100 datasets from 15 sources such as the rice databases of the South Green platform or international databases such as Gramene.org [2] for cereals.

Objective: The current challenges are related to the development of methods for functional analysis of genes and in particular to methods for prioritization of candidate genes. The data integrated from databases are insufficient to infer with certainty the function of genes.
One of the first objectives will be the development of text mining methods to extract functional information on genes in scientific publications. A second objective will be to integrate new complementary data sets that can provide functional information. Finally, functional analysis methods will be developed and validated on published data.

Program:
- development of text mining methods based on a corpus of scientific publications identified by the partners. Application in rice.
- Extend data coverage to QTL/GWAS, expression and co-expression, interactome and metabolic pathway information.
- Development of functional analysis methods including prioritization of candidate genes.
- Validation of functional analysis methods through a use case published in an international journal.

References:
1. Venkatesan A, Tagny Ngompe G, Hassouni NE, Chentli I, Guignon V, Jonquet C, et al. Agronomic Linked Data (AgroLD): a Knowledge-based System to Enable Integrative Biology in Agronomy. PLoS ONE. 2018 – pp. 13:17.

2. Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018: Unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res. 2018.

Application: Applications for this position (CV, Motivation Letter, last grade report, References) will be received EXCLUSIVELY in a single PDF document accessible for download via email sent to Jérôme AZÉ (jerome.aze@lirmm.fr) and Pierre LARMANDE (pierre.larmande@ird.fr).

Supervision:

- UMR 5506 LIRMM Equipe ADVANSE: AZÉ Jérôme
- UMR 232 DIADE Equipe CERES: LARMANDE Pierre

Location: LIRMM et IRD-Occitanie

Gratification: 6 months