DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

PROPOSITION D'UN PROJET DE THÈSE

A L'ÉCOLE DOCTORALE

« Écologie, Géosciences, Agronomie, ALimentation »

INFORMATIONS GÉNÉRALES

Titre de la thèse : Analyse des données biologiques hétérogènes par exploitation de graphes multicouches pour comprendre et prédire les variations d'efficacité alimentaire chez le porc

Acronyme : BIOMULTI

Champ disciplinaire 1 : Agronomie

Champ disciplinaire 2 : Choisissez un élément.

Trois mots-clés : bioinformatique ; données omiques ; efficience alimentaire

Unité d'accueil : PEGASE

Nom, prénom du directeur de thèse : GONDRET, Florence

Adresse mail : florence.gondret@inrae.fr

Nom, prénom du co-directeur/co-encadrant de thèse 1 (le cas échéant) : BECKER, Emmanuelle

Adresse mail : emmanuelle.becker@irisa.fr

Nom, prénom du co-encadrant de thèse 2 (le cas échéant) :

Adresse mail :

Financement (origine et montant) : INRAE (50%)/Région Bretagne (50%)

Contact(s) (adresse postale) : UMR PEGASE, 16 Le Clos, 35590 Saint-Gilles

Mode de recrutement

Le mode de recrutement du doctorant dépend de la nature du financement du projet de thèse. Pour identifier le mode de recrutement, veuillez consulter le site web de l'ED EGAAL - cliquez ici. Le projet de thèse **ne sera pas** publié si cette information est manquante.

☐ **Concours** ☒ **Entretien** ☐ **Autre (précisez) :**

DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

DESCRIPTION SCIENTIFIQUE DU PROJET DE THÈSE

Contexte socio-économique et scientifique : (10 lignes)

L'efficience alimentaire correspond à la valorisation des ressources alimentaire par l'animal qui les transforme notamment en gain de poids. C'est un phénotype clé car il aboutit à une épargne des ressources et à une réduction des rejets et effluents dans l'environnement, mais également complexe au sens où un grand nombre de voies biologiques le détermine. Aussi, il est important d'identifier les éléments susceptibles de jouer un rôle pivot dans le contrôle du phénotype, et les interdépendances entre les entités participant aux différentes voies biologiques sous-jacentes. Les technologies d'analyse du vivant aboutissent à la production de grandes quantités de données portant sur des entités biologiques hétérogènes (génome, quantité de transcrits, abondance ou activité des protéines, métabolites). Il s'agit de développer des méthodes et études pour déduire la structure de dépendance des données pour identifier les régulateurs clés et les sous-réseaux importants dans la définition du caractère biologique d'intérêt.

Hypothèses et questions scientifiques (8 lignes)

L'hypothèse de travail est que l'on peut mieux définir les relations entre les molécules régulatrices d'un phénotype lorsqu'on associe les données expérimentales avec des données de connaissances balayant différents niveaux d'organisation biologique, permettant de combler les trous dus aux méthodologies expérimentales et de distinguer la co-régulation de la co-expression entre entités biologiques. Pour analyser différents jeux de données (différentes expériences, différentes échelles d'organisation du vivant), nous proposons d'appliquer la méthodologie des graphes multicouches. Cette méthodologie (développée en science sociale et récemment testée en génomique fonctionnelle) vise à modéliser la structure des données sous la forme d'un graphe orienté, c'est-à-dire d'un ensemble de nœuds (les entités biologiques) reliés par des arêtes signées permettant donc d'envisager le sens des variations des réseaux induit par l'application d'une contrainte particulière.

Principales étapes de la thèse et démarche (10-12 lignes)

Etape 1 (6-8 mois). Analyser les structures de corrélations au sein de sources de données expérimentales en relation avec la variation de l'efficience alimentaire. Nous disposons de plusieurs jeux de données transcriptomiques (transcrits de milliers de gènes) acquis dans le sang de 150 porcs en croissance. Cette étape consiste à déduire des sous-réseaux de corrélations au sein de chaque jeu de données, en utilisant des méthodes de calculs de corrélations pondérées entre entités, et d'identifier leurs relations avec les mesures des caractères relatifs à l'efficience alimentaire.

Etape 2 (10-12 mois). Associer ces sous-réseaux de co-variations avec des larges réseaux représentant la connaissance publique de référence sur différents niveaux d'organisation du vivant (métabolismes, interaction protéines/protéines, interaction génétiques…). L'objectif de cette étape est de distinguer la corrélation liée à la co-expression (biologique ou mathématique induisant de fausses arêtes de co-expression) de la corrélation réellement due à la co-régulation entre entités. A cette étape, il s'agira de créer et d'évaluer les méthodes de parcours de graphes multicouches avec sauts entre couches (une couche constituée par le ou les réseaux de covariances, une ou des couches avec les bases de connaissances), en identifiant les stratégies adéquates de couplage entre couches.

Etape 3 (12-15 mois): Appliquer ces méthodes sur des couches focalisées sur des entités biologiques expérimentales différentes. L'enjeu de cette étape est que les données expérimentales sont très incomplètes (par exemple le nombre de métabolites identifiés par métabolomique ou des phénotypes mesurés par des méthodes cibles sont bien moindres que le nombre de transcrits), mais « les trous » pourront être comblés par la connaissance extraite des bases. L'impact de chaque niveau (gènes, protéines, métabolites) sur la connaissance du phénotype sera étudié.

Approches méthodologiques et techniques envisagées (4-6 lignes)

Les méthodes utilisées relèvent de la statistique et de la fouille de données :

- Méthodes statistiques pour l'étude de corrélations entre entités biologiques (matrice de corrélation, WGCNA, PTIC)
- Théorie des graphes d'influences, graphes d'influences
- Exploration de graphes multicouches par des marches aléatoires avec restart (algorithmes de la famille RWR)
- Analyse topologique des graphes

Compétences scientifiques et techniques requises pour le candidat

Le/la doctorant(e) devra avoir une formation préalable en (bio)informatique avec un intérêt pour le domaine d'application en biologie, ou bien une formation en biologie mais avec un fort attrait pour le traitement de données.

DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

Le/la doctorant(e) acquerra les compétences cognitives suivantes durant la thèse : études des structures de covariances entre données, algorithmique de construction et parcours des graphes multicouches de nature différentes dont certaines orientées, seuils de significativité en fouilles de données. Ces compétences sont valorisables en biologie mais aussi en sciences humaines et sociales

ENCADREMENT DE LA THÈSE[1]

Nom de l'unité d'accueil : UMR PEGASE	**Nom de l'équipe d'accueil :** Croissance
Nom du directeur de l'unité : F GONDRET	**Nom du responsable de l'équipe :** I LOUVEAU
Coordonnées du directeur de l'unité : Florence.gondret@inrae.fr	**Coordonnées du responsable de l'équipe :** Isabelle.louveau@inrae.fr

Directeur de thèse

Nom, prénom : GONDRET, Florence

Fonction : DR

Date d'obtention de l'HDR : 2007

Employeur : INRAE

Co-directeur/co-encadrant de thèse

Nom, prénom : BECKER, Emmanuelle

Fonction : MCF

Titulaire de l'HDR : ☐ oui ☒ non Si oui, date d'obtention de l'HDR :

Employeur : Univ Rennes 1, UMR IRISA

École doctorale de rattachement :

Taux d'encadrement doctoral dans le présent projet : 50%

Publications majeures des 5 dernières années du directeur de thèse et co-directeur(s)/co-encadrant(s) sur le sujet de thèse :

- Messad F, Louveau I, Koffi B, Gilbert H, **Gondret F.** Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. BMC Genomics. 2019 20(1):659.
- Desert C, Baéza E, Aite M, Boutin M, Le Cam A, Montfort J, Houee-Bigot M, Blum Y, Roux PF, Hennequet-Antier C, Berri C, Metayer-Coustard S, Collin A, Allais S, Le Bihan E, Causeur D**, Gondret F**, Duclos MJ,

[1] **Dans l'ED EGAAL, si 1 scientifique dans la direction de la thèse = 100% d'encadrement doctoral ; si 2 personnes impliquées dans la direction de la thèse = entre 50% et 70% d'encadrement doctoral pour l'HDR directeur ; si 3 personnes impliquées dans l'encadrement de la thèse : répartition :40% - 30% - 30% de l'encadrement doctoral.**

Lagarrigue S. Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. BMC Genomics. 2018 Mar 7;19(1):187.

- **Gondret F,** Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, Gilbert H, Louveau I. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017 18(1):244.
- **Becker E,** Com E, Lavigne R, Guilleux M-H, Evard B, Pineau C and Primig M. The protein expression landscape of mitosis and meiosis in diploid budding yeast. J. Proteomics, 2017.
- Taghipoor M, Lemosquet-Simon S, van Milgen J, Siegel A, Sauvant D, **Gondret F.** 2016. Modélisation de la flexibilité métabolique : vers une meilleure compréhension des capacités adaptatives de l'animal. INRA Prod. Animales. 29 (3), pp.201-216.
- Chapple CE, Robisson B, Spinelli L, Guien C**, Becker E,** Brun C. Extreme multifunctional proteins identified from a human protein interaction network. Nature Communications, 2015.

FINANCEMENT DE LA THÈSE

Origine(s) du financement de la thèse : INRAE/Région Bretagne

Salaire brut mensuel : 1750 €

État du financement de la thèse : Acquis

Date du début/durée du financement de la thèse : 01/11/2020 – 36 mois