

SOUTIEN AUX LABORATOIRES DES UNIVERSITÉS DE TOURS ET DE POITIERS

ACTIONS DE RECHERCHES COLLABORATIVES APPEL À PROJETS 2016

DOSSIER DE CANDIDATURE

Les Universités de Tours et de Poitiers prolongent l'appel à projet **Actions de Recherches Collaboratives** destiné à participer au financement d'actions de recherches portées conjointement par au moins une équipe d'un laboratoire de recherche de l'université de Poitiers et une équipe d'un laboratoire de l'université de Tours.

L'objectif est d'encourager les nouvelles collaborations scientifiques.

Dans ce cadre, nous privilégierons :

- les thématiques émergentes,
- les actions jetant les bases de projets d'envergure nationale (type ANR) ou international (type européen).
- L'existence d'un volet formation (niveau master ou doctorat) ou d'un volet valorisation sera apprécié mais n'est pas obligatoire.
- Le financement d'équipement ne sera pas prioritaire (d'autres appels à projets sont consacrés à ce type d'action).

Chaque projet est limité à 10 000 € et pourra être utilisé pour des dépenses de fonctionnement ou d'investissement. La participation à une gratification de stage de master entre dans les dépenses éligibles, pour une mobilité de l'étudiant entre les établissements.

Il est demandé à chaque unité de recherche de porter au plus un projet. De plus, face au succès remporté par l'ARC 2015, nous ne privilégierons pas les projets portés par une unité lauréate l'an dernier.

L'enveloppe globale consacrée à cet appel est de 70 000 €.

Calendrier:

Lancement de l'appel à projets : 29 février 2016

Date limite de retour des dossiers : 11 avril 2016

Avis des CS Poitiers et Tours : 12 mai 2016 (Poitiers) 31 mai 2016 (Tours)

DOSSIER ADMINISTRATIF ET FINANCIER

FICHE D'IDENTITÉ DU PROJET
CHAMPS DE RECHERCHE ET DIMENSION COLLABORATIVE
ÉLÉMENTS FINANCIERS
PERSONNEL IMPLIQUÉ

DOSSIER SCIENTIFIQUE

DESCRIPTION DU PROJET
COMPÉTENCES ET EXPERTISES DU LABORATOIRE PORTEUR
RETOMBÉES POTENTIELLES

AVIS ET SIGNATURES

Ce dossier de candidature est à retourner au plus tard le 11 avril 2016 :

- en version numérique (format pdf, taille maximum 5 Mo) au courriel ci-dessous ;

Contact :

Université de Poitiers : [murielle .taillet@univ-poitiers.fr](mailto:murielle.taillet@univ-poitiers.fr)

Université de Tours : caroline.vaslin@univ-tours.fr

DOSSIER ADMINISTRATIF ET FINANCIER

FICHE D'IDENTITÉ DU PROJET

Titre du projet : Modélisation stochastique et analyse statistique en expression génétique
Coordonnées du porteur de projet et responsable de l'équipe partenaire n°1 : Nom : Biermé Prénom : Hermine Nom du Laboratoire (indiquer notamment le numéro de labellisation EA, UMR, IFR...) : LMA UMR 7348 Nom et adresse de l'établissement de rattachement : Université de Poitiers N° téléphone : 05 49 49 69 16 Courriel : hermine.bierme@math.univ-poitiers.fr
Coordonnées du responsable de l'équipe partenaire n°2 : Nom : Yvinec Prénom : Romain Nom du Laboratoire (indiquer notamment le numéro de labellisation EA, UMR, IFR...) : Physiologie de La Reproduction et des Comportements, INRA UMR85, CNRS UMR7247, Université François-Rabelais de Tours, IFCE, Nom et adresse de l'établissement de rattachement : INRA F-37380 Nouzilly, France N° téléphone : 02 47 42 75 05 Courriel : romain.yvinec@tours.inra.fr
Coordonnées du responsable de l'équipe partenaire n°3 : Nom : Malrieu Prénom : Florent Nom du Laboratoire (indiquer notamment le numéro de labellisation EA, UMR, IFR...) : LMPT UMR 7350 Nom et adresse de l'établissement de rattachement : Université François-Rabelais de Tours N° téléphone : 02 47 39 76 43 Courriel : florent.malrieu@lmpt.univ-tours.fr
Durée prévue du projet (en mois, projet minimum de 3 mois limité à 24 mois maximum) : 24
Résumé du projet (10 lignes maximum) Ce projet interdisciplinaire entre biologique et mathématiques propose de fédérer les probabilistes et statisticiens des universités de Poitiers et de Tours autour de l'étude de l'expression des gènes. Il repose sur deux axes principaux. Le premier est le développement et l'étude de nouveaux modèles probabilistes pour la production de l'ARN messenger permettant de prendre en compte des mécanismes de régulation post-transcriptionnels précis. Deuxièmement, il s'agit de développer des méthodes statistiques adaptées aux nouvelles données de séquençages à haut-débit, et leurs applications à des données réelles obtenues par l'équipe INRA associée au projet. L'approche combinée de modélisation probabiliste et d'analyse statistique permettra de prendre en compte à la fois des données issues du transcriptome (nombre d'ARN messagers), du traductome (efficacité de traduction des ARN messagers) et du protéome (nombre de protéines), afin d'analyser finement le phénotype qu'exprime une cellule.
Montant de la subvention demandée (limitée à 10 000 euros) : 10 k € Apport éventuel d'un financement ou d'un co-financement: Le projet ANR JCJC PIECE, porté par Florent Malrieu (Université de Tours), co-financera les activités du présent projet à hauteur de 5 k €. Une demande est en cours dans le cadre du Crédit Incitatif PHASE 2016 (INRA) pour la collecte de données.

CHAMPS DE RECHERCHE ET DIMENSION COLLABORATIVE

Champs de recherche :

Probabilités : chaînes de Markov, processus déterministes par morceaux

Statistiques : problèmes inverses, tests d'hypothèses en grande dimension, clustering

Biologie moléculaire et cellulaire : expression des gènes, séquençage haut-débit

Dimension collaborative :

Ce projet réunit trois équipes ayant des champs disciplinaires proches mais distincts, incluant notamment la modélisation probabiliste, l'analyse statistique et la biologie moléculaire et cellulaire.

L'équipe partenaire n°1 est composée de probabilistes et statisticiens du LMA UMR 7348 de l'Université de Poitiers ayant chacun une expertise en modélisation stochastique et analyse statistique pour des données issues de la biologie ou de la médecine. Cette équipe pilote également le master Statistique et données du vivant de l'université de Poitiers. L'équipe partenaire n°2 est interdisciplinaire, et comprend des chercheurs en mathématiques appliqués, en bio-informatique et biologie cellulaire du laboratoire de Physiologie de La Reproduction et des Comportements, INRA UMR85, CNRS UMR7247 de l'INRA de Nouzilly. Enfin, l'équipe partenaire n°3 est composée de probabilistes du LMPT UMR 7350, spécialisés dans l'étude des processus stochastiques et théorèmes limites.

Les trois responsables d'équipes sont arrivés depuis moins de trois ans et ont déjà construit des interactions entre les différents sites : une conférence a été organisée en novembre 2015 par H. Biermé et F. Malrieu (R. Yvinec y était l'un des conférenciers), C. Constant, en première année de thèse sous la direction de H. Biermé (Poitiers) et Ch. Georgelin (Tours), travaille sur la modélisation stochastique et l'analyse statistique de la sécrétion de la GnRH-1.

Le projet permettra de consolider ces premières collaborations et d'envisager la soumission de projets plus grande ampleur par la suite.

ÉLÉMENTS FINANCIERS

Chaque financement sera plafonné à 10 000 € et pourra être utilisé pour des dépenses de fonctionnement ou d'investissement. La participation à une gratification de stage de master entre dans les dépenses éligibles, pour une mobilité de l'étudiant entre les établissements.

Demande financière globale:

Fonctionnement (consommables) : 0 k €

Équipement : 1 k €

Mission : 4,5 k €

Gratification : 4,5 k €

Co-financement éventuel :

Source :Crédit Incitatif PHASE 2016 (INRA)

Montant demandé : 12 k€ (demande en cours).

Source : ANR PIECE
Montant obtenu : 5 ke

PERSONNEL IMPLIQUE DANS LE PROJET

Nom/Prénom	Laboratoire ou équipe	Statut ou grade (professeur ou chercheur, Ingénieur, technicien, doctorant, post-doctorant ...)	ETP consacré au projet (en %)
Biermé Hermine	LMA Poitiers	Professeur	30
Constant Camille	LMA Poitiers	Doctorante	20
Enikeeva Farida	LMA Poitiers	Maître de conférences	20
Louis Pierre-Yves	LMA Poitiers	Maître de conférences	20
Michel Julien	LMA Poitiers	Professeur	10
Slaoui Yousri	LMA Poitiers	Maître de conférences	20
Yvinec Romain	INRA Nouzilly	CR	40
Crépieux Pascale	INRA Nouzilly	CR	10
Poupon Anne	INRA Nouzilly	DR	10
Tréfier Aurélie	INRA Nouzilly	Doctorante	20
Durieu Olivier	LMPT Tours	Maître de conférences	20
Georgelin Christine	LMPT Tours	Maître de conférences	20
Lagasque Gabriel	LMPT Tours	Doctorant	20
Malrieu Florent	LMPT Tours	Professeur	20

DOSSIER SCIENTIFIQUE

Description du projet (5 pages maximum incluant les références)

Contexte et objectifs

La cellule est un système adaptatif très complexe, et la prédiction du phénotype qu'elle exprime à partir d'un génotype est un véritable défi scientifique. Pour prédire ce phénotype, il est nécessaire de combiner plusieurs approches systémiques, notamment pour confronter le niveau d'ARN messagers (ARNm) produits (transcriptome), l'efficacité de traduction des ARNm en protéines (traductome) et finalement le niveau de protéines réellement produites (protéome) au phénotype cellulaire.

Les récents progrès des mesures expérimentales de l'expression des gènes [1-4] posent de nouveaux défis à la communauté mathématique, tant au niveau de l'étude de modèles dynamiques (déterministes et probabilistes) que des analyses statistiques. En effet, les approches expérimentales à haut-débit (puces à ADN, séquençage d'ARN, spectrométrie de masse à haute définition, microscopie confocale, etc) couplées à l'exploitation de phénomènes biologiques maîtrisés (fluorescence, bio-luminescence, transfert d'énergie par résonance, etc) permettent d'obtenir des données de plus en plus précises (à l'échelle d'une population de cellules, de cellules uniques voir de molécules) et dynamiques (données temporelles et spatiales). Ce faisant, les quantités de données générées permettent une compréhension de plus en plus fine des mécanismes en jeu mais nécessitent bien souvent une approche de modélisation mathématique pour en extraire des informations intéressantes. La précision des détails des mécanismes moléculaires rend de plus en plus complexe cependant l'analyse mathématique de tels modèles.

Une partie de ce projet concerne le développement et l'étude mathématique de modèles probabilistes qui interviennent naturellement dans ces processus biologiques.

L'explosion du volume des données générées par les approches de type séquençage [4-8] (typiquement une ou plusieurs données par gène) mais leur grande variabilité (due à la fois à la variabilité individuelle inhérente aux processus et aux techniques de mesures) posent également des questions quant à la validité d'approches statistiques de types clustering ou analyse différentielle, notamment lorsque peu de réplicats sont disponibles (dû principalement aux coûts de ces mesures).

Une deuxième partie de ce projet concerne ainsi le développement de méthodes statistiques adaptées aux nouvelles données de séquençages, et leurs applications à des données réelles obtenues par l'équipe partenaire n°2 [9-11].

▫ Protocole – Répartition du travail

- Le développement de nouveaux modèles probabilistes d'expression des gènes permettant de prendre en compte plus finement la dynamique des ARNm (position spatiale, recrutement aux ribosomes, temps de traduction etc) se fera en concertation avec les trois équipes partenaires, en confrontant l'expertise biologique de l'équipe partenaire n°2 et l'expertise de modélisation des équipes partenaires n°1 et n°3. L'analyse mathématique sera mise en œuvre en adaptant des techniques développées au sein des équipes [10-16].
- Le développement de nouvelles méthodes statistiques pour l'analyse différentielle et le clustering sera principalement mené par l'équipe partenaire n°1, en concertation avec l'équipe partenaire n°2. L'application de ces méthodes aux données biologiques sera réalisée par l'équipe partenaire n°2, au travers de stages de niveau M1 et M2 proposés aux étudiants du master Statistique et données du vivant de l'université de Poitiers.

▫ Description détaillée des méthodologies et techniques utilisées

Les modèles désormais classiques d'expression des gènes [17-19] se formulent mathématiquement comme des processus de Markov de sauts purs (de type naissance et mort) et décrivent la dynamique temporelle des ARNm et protéines associés à un ou plusieurs gènes. Sous leur forme la plus simple, trois étapes sont prises en compte : la dynamique du gène lui-même (accessible ou non aux molécules responsables de la transcription), la dynamique des ARNm (production, dégradation) et la dynamique des protéines (production, dégradation). Selon les mécanismes précis en jeu dans la régulation de l'expression du gène, les quantités importantes telles que la distribution stationnaire et la quantification de certains moments, les temps de premier passage à un certain niveau ou la vitesse de convergence vers l'équilibre stationnaire, ne peuvent pas toujours être calculées. On utilise alors des techniques de

réduction de modèles (séparation d'échelles de temps et théorèmes limite de type loi des grands nombres [12,28]), et les modèles approchés se présentent sous la forme de processus déterministes par morceaux [13,14], pour lesquelles des techniques ont récemment été développées, notamment dans le cadre du projet ANR PIECE, voir [15,16].

Avec les progrès des mesures expérimentales, il apparaît nécessaire de modéliser plus finement la dynamique des ARNm, en prenant en compte les divers états dans lesquels un ARNm peut se trouver à un instant donné (par exemple, accessible ou non aux molécules responsables de la traduction, ou le nombre de ribosomes attachés à l'ARNm). De nouvelles échelles de temps et d'espaces sont alors à prendre en compte. Les sorties du modèle qui peuvent être directement comparées aux données (mesures de l'efficacité de traduction des ARNm) doivent être caractérisées (état stationnaire, dynamique hors équilibre etc) L'estimation de paramètres (problème inverse [20-21]) du modèle à partir de données est également importante pour caractériser le comportement d'un gène donné et identifier des tendances entre plusieurs groupes de gènes.

Les approches expérimentales de type séquençage à haut-débit se heurtent à des questions d'analyses statistiques inédites, si l'on veut déterminer des différentiels dans une condition stimulée par rapport à une condition contrôle. La situation désormais classique consiste à séquencer les ARNm ou protéines produits (à l'échelle d'une cellule ou d'une population de cellules) dans deux conditions expérimentales différentes. On cherche alors à identifier les gènes pour lesquels l'activité (quantité de produits géniques) a évolué, soit augmenté soit diminué.

Bien que des méthodes d'analyse différentielle de transcriptome à partir de données RNA-Seq existent déjà [22-27], elles sont encore perfectibles, et aucune méthode à ce jour ne fait complètement consensus au sein de la communauté. La difficulté résulte de plusieurs facteurs : le faible nombre de répétitions des expériences (dû principalement à leur coût), le très grand nombre de tests statistiques effectués en parallèle (au moins un par gène), la possible inter-dépendance entre les gènes, les différentes normalisations dues aux cycles d'amplifications et d'échantillonnage des ARNm ou encore le faible niveau d'expression d'une grande partie d'ARNm (seulement quelques copies par échantillon).

Ce problème méthodologique est encore plus criant pour l'analyse des données de traductome (recrutement des ribosomes aux ARNm) car il a encore été peu abordé.

On cherchera ainsi à développer des méthodes rigoureuses permettant d'attester si l'analyse de l'expression différentielle pour chaque gène est possible au regard des données et de quantifier celle-ci le cas échéant, en limitant le nombre de faux positifs.

Enfin, la mise en regard de plusieurs jeux de données de natures différentes, comme le protéome obtenu par spectrométrie de masse, pose également des questions statistiques et méthodologiques encore non résolues. En particulier, la confrontation de données de transcriptome, de traductome et de protéome doit permettre d'obtenir un niveau de corrélation statistique de l'activité des molécules satisfaisant, que les seules données de transcriptome et protéome ne permettent pas d'obtenir [1-2] (distorsions entre gènes transcrits et protéines produites). De plus, la prise en compte simultanée de ces trois types de données permettra de proposer différents mécanismes de régulation (transcriptionnel, traductionnel, post-traductionnel et dégradation des molécules) présents entre deux conditions différentes.

Les méthodes développées seront appliquées à des données de transcriptome et traductome déjà disponibles au sein de l'équipe partenaire n°2, dans des cellules de Sertoli de rat en culture primaire stimulées *in vitro* par l'hormone FSH [9-11]. Des données de protéome dans les mêmes conditions sont à acquérir par nanoLC-HRMS/MS (nanochromatographie liquide couplée à la spectrométrie de masse à haute résolution en tandem) dans le cadre d'un projet porté par l'équipe partenaire n°2 à l'INRA.

L'application de ces méthodes permettra l'identification de manière robuste du réseau de gènes et de protéines activés et inhibés dans les cellules de Sertoli suite à la stimulation par FSH et servira aux développements de méthodes systémiques pour la caractérisation des mécanismes adaptatifs fonctionnels mis en œuvre dans le contexte de l'effet d'une hormone importante pour le contrôle de la reproduction animale.

▫ Caractère innovant du projet

La modélisation probabiliste de l'expression des gènes est un domaine encore jeune, et de nombreuses questions restent en suspens quant au comportement de modèles plus complexes. En particulier, la prise en compte d'une dynamique des ARNm autre que du premier ordre (linéaire) donne lieu à des difficultés mathématiques quant à l'analyse des modèles et constitue l'un des objectifs principaux du projet.

Bien que de nombreuses méthodes statistiques aient été récemment développées pour l'analyse des données haut-débit issues du séquençage d'ARNm (et adaptées des méthodes pour les puces à ADN), leur validité théorique reste peu évidente. De plus, la prise en compte de données quantitatives du recrutement des ARNm aux ribosomes, en complément de données sur le nombre d'ARNm et de protéines n'a pas, à notre connaissance, été abordée à ce jour. La comparaison statistique rigoureuse du transcriptome, du traductome, et du protéome, devrait enfin permettre l'identification robuste du phénotype d'une cellule et du comportement de différents groupes de gènes associés à ce phénotype. Cette identification peut alors jeter les bases d'approches systémiques permettant de reconstruire les réseaux de gènes responsables de changements phénotypiques.

▫ Caractère stratégique du projet pour préparer une soumission à l'ANR, à l'Europe ou pour débiter une collaboration internationale ou nationale

Ce projet concrétise de nouvelles interactions entre Poitiers et Tours. Il permettra de créer un pôle Poitiers/Tours expert dans les interactions entre les mathématiques de l'aléatoire (probabilités et statistique) et la biologie moléculaire et cellulaire. Il constituera également une base solide pour l'élaboration de projets de plus grande ampleur de type ANR.

▫ Formation

Ce projet prévoit la participation de stagiaires pour le développement de méthodes statistiques, issus du Master IMMT, spécialité Mathématique, parcours Statistique et données du vivant de l'université de Poitiers, dont l'équipe partenaire n°1 est responsable.

▫ Calendrier et budget détaillé

Un premier stage d'un étudiant M1 Statistique et données du vivant est prévu du 1 mai au 31 juillet 2016 (financement propre à l'équipe n°2) afin d'initier la collaboration entre les équipes partenaires n°1 et n°2. En parallèle, une rencontre des trois équipes partenaires aura lieu au mois de juin 2016.

Avec le soutien de l'ARC, nous prévoyons

1. l'organisation de deux rencontres scientifiques : une journée en 2016-2017 ainsi qu'une conférence en 2018 année européenne de la Biologie-Mathématique (<http://www.euro-math-soc.eu/year-mathematical-biology-2018>) (ARC 4k€+ ANR PIECE 4.5k€)
2. le recrutement d'un stagiaire de M1 (sur 3 mois) et d'un stagiaire de M2 (sur 6 mois) pour réaliser les analyses statistiques (4,5k€) en 2017-2018.
3. Le financement des missions entre Tours et Poitiers pour les différents membres du projet (0.5k€ + ANR PIECE 0.5k€).
4. L'achat d'un ordinateur portable pour Camille Constant, doctorante en 1^{ère} année de thèse en codirection entre Tours et Poitiers en 2016 (1k€).

Le soutien du Crédit Incitatif Phase INRA (12k€) permettra d'obtenir de nouvelles données de protéome, en plus celles déjà disponibles.

▯ Références :

- 1- Taniguchi et al., *Science* 329(5991):533-8, 2010.
- 2- Schwanhäusser et al., *Nature* 473(7347):337-42, 2011.
- 3- Dalal et al., *Curr Biol* 24(18):2189-94, 2014.
- 4- Cohen et al., *Science* 322(5907):1511-6, 2008.
- 5- Shalek et al., *Nature* 498(7453):236-40, 2013.
- 6- Marinov et al., *Genome Res* 24(3):496-510, 2014.
- 7- Dey et al., *Nat Biotechnol* in Press, 2015.
- 8- Schott et al. *PLoS Genetics* 10(6):e1004368 2014.
- 9- Musnier, León et al. *Mol. Endocrinol.*, 26 (669-680) 2012.
- 10- Leon et al. *J. Mol. Endocrinol.* 52(373-382) 2014.
- 11- Le Borgne et al. *Development*, 141(2096-107) 2014.
- 12- Yvinec et al., *J Math Biol* 68(5) :1051-1070, 2014.
- 13- Mackey, Yvinec et al., *SIAM J of Appl Math* 73(5): 1830-1852, 2013.
- 14- Mackey, Yvinec et al., *J Theor Biol* 274(1):84-96, 2011.
- 15- Malrieu et al. *E. Comm. Prob.* 17(56):1-14, 2012.
- 16 - Malrieu et al. *A.A.P.*, 24(1), 292-311, 2014.
- 17- Maamar et al., *Science* 317(5837):526-9, 2007.
- 18- Choi et al., *Science* 322(5900):442-6, 2008.
- 19- Chang et al., *Nature* 453(7194):544-7, 2008.
- 20- Banks et al., *CRC Press*, 2014.
- 21- Doumic et al., *SIAM J of Appl Math* 50(2):925-950, 2012.
- 22- Li et al., *Biometrics* 70(4):872-80, 2014.
- 23- Canale and Dunson, *J Am Stat Assoc* 106(496):1528-1539, 2011.
- 24- Blei, *Commun ACM* 55(4): 77-84, 2012.
- 25- Benjamini & Hochberg, *J.R.Stat. S B* 57(1), 289—300, 1995.
- 26- McCarthy et al. *Nucl. Acids. R.* 40(10) 4288-4297, 2012.
- 27- Bullard et al. *BMC Bioinfo.* 11(94), 2010.
- 28- Biermé, Durieu, *Trans. of AMS.*, 366(11), 5963-5989, 2014.

RETOMBÉES POTENTIELLE DU PROJET

Ce projet permettra d'animer le réseau déjà actif de la modélisation mathématique et statistique en biologie entre les universités de Tours et Poitiers et l'INRA. En particulier, il permettra de soutenir et rapprocher les groupes de travail math-bio, récemment initiés indépendamment au sein du LMA et du LMPT afin que les universités de Tours et Poitiers puissent activement participer à l'année européenne de la Biologie-Mathématique en 2018.

Plusieurs publications scientifiques dans des revues internationales à comité de lecture sont attendues, sur l'analyse mathématique de modèles probabilistes d'expression des gènes, et sur le développement de méthodes statistiques pour l'analyse différentielle à partir de données de séquençage haut-débit.

Les méthodes statistiques développées seront implémentées sous la forme d'un package R (logiciel libre de statistique) disponible pour la communauté scientifique.

La modélisation mathématique et l'application des méthodes statistiques aux données générées au sein de l'équipe partenaire n°2 permettront une meilleure compréhension des mécanismes intracellulaires activés et inhibés suite à une stimulation hormonale en biologie de la reproduction.

