ALLOCATION POST-DOCTORALE – 18 mois

LIMOS – LMGE (CNRS, Université Blaise Pascal)

Clermont-Ferrand

Étude de la biosphère rare microbienne par une approche in silico : nouvelle méthode de classification ensembliste et modélisation

Présentation du sujet de recherche, des objectifs à atteindre et du calendrier prévisionnel de réalisation

- *le contexte scientifique*

La détermination de la structure des communautés (abondance, richesse, composition) d'un écosystème est un enjeu central en écologie et donc en écologie microbienne. Dans les écosystèmes aquatiques les estimations étaient jusqu'à récemment de moins de 200 phylotypes bactériens (OTUs). Or, de récents travaux effectués à l'aide des nouvelles techniques de séquençage (NGS) ont mis en évidence que cette richesse spécifique était certainement sous-estimée ; la plus grande partie de la biodiversité est représentée par des OTUs faiblement abondants : la biosphère rare. Les détracteurs de la biosphère rare pensaient que les OTUs détectés ne concernaient que des cellules mortes ou en transit, donc sans importance pour le fonctionnement de l'écosystème. Or, récemment, les études de Campbell et al. (2011) et **Hugoni et al . (2013)** ont mis en évidence que les espèces rares pouvaient être actives. Caron and Countway, (2009) émettent l'hypothèse que des remaniements en fonction des changements de conditions environnementales (changement global) pourraient également intervenir dans la communauté des protistes. Cependant, ce modèle s'oppose à une autre vision qui mettent en évidence que certains microorganismes restent toujours rares quelles que soient les conditions environnementales et que ces espèces sont spécifiques à une aire géographique (**Lepère et al. 2013**). Ainsi, les résultats peuvent diverger entre ces différentes études et la diversité ainsi que la dynamique de cette biosphère restent encore méconnues. **Cependant, l**a **biosphère rare est une réalité qui ne peut plus être ignorée même si la structure de cette communauté reste à déterminer.** Jusqu'à présent, les études traitant de la biosphère rare ont utilisées des méthodes bioinformatiques spécifiques dans le cadre restreint de leur jeu de données, c'est à dire de données NGS acquises dans un écosystème donné à une période donnée. Or, une méta-analyse des données sur la biodiversité microbienne déjà acquises permettrait d'évaluer la richesse (i.e. nombre d'espèces), la composition (quelles espèces ?) et la dynamique de la biosphère rare au sein d'un écosystème (aquatique par exemple). En effet une espèce peut être rare dans le contexte d'une étude particulière alors qu'elle a été trouvée en forte abondance dans une autre situation. A contrario, certaines espèces pourraient être toujours rares et/ou spécifiques à des aires géographiques. Ainsi, ces quelques exemples mettent en évidence que les interrogations légitimes sur l'étendue de cette biosphère rare peuvent trouver des réponses par une analyse des données existantes. **D'une façon générale un des verrous pour la génération de nouvelles connaissances n'est plus dans la production des données (i.e. séquençage) mais dans le traitement de celles-ci (i.e. bioinformatique)**. La détermination de la structure (richesse, abondance, diversité, composition) de la biosphère rare repose sur la détermination des OTUs (i.e. espèces microbienne). **Or, cette détermination varie en fonction des méthodes de classification et n'est pas associée à une probabilité d'appartenance à une classe.** La difficulté réside généralement soit dans le langage de représentation associé, soit dans le mécanisme d'inférence mis en œuvre.

La rareté dans ce contexte d'application est un phénomène qui peut être observé avec une approche multi-points de vue. On peut ici considérer qu'une espèce rare sera peu conservée et aura tendance à être dans différentes classes résultantes de classifications différentes. Les méthodes ensemblistes de classification constituent une solution pouvant permettre d'identifier ces phénomènes de rareté. En effet les méthodes de classification existantes (hiérarchique, par partitionnement) présentent plusieurs limites inhérentes aux paramètres d'entrée ou au processus de mise en œuvre. Il a récemment été montré que l'agrégation par combinaison ensembliste de ces méthodes peut permettre d'améliorer les résultats obtenus. Mais ces méthodes ensemblistes de classification attribuent généralement une classe unique à un objet. Sachant qu'un objet peut se trouver à la frontière de plusieurs classes, ou plus précisément avoir des ressemblances avec des objets de différentes classes, il serait judicieux d'associer une valeur de confiance à la classification des objets, afin de faciliter la mise en évidence du phénomène de rareté. Le recours à une approche probabiliste (si l'incertitude est due à la variabilité des distributions) ou à la théorie des fonctions de croyance (si l'incertitude est due à l'ignorance partielle) constituent une piste pour étendre ce type d'approches à la prise en compte de l'incertitude de la classification. En effet le LIMOS a développé des algorithmes de classification **(Antoine et al. 2012, 2013)** qui introduisent une connaissance a priori sous forme de contraintes sur les objets et génèrent une partition nommée partition crédale. Cette partition, basée sur l'utilisation des fonctions de croyance, généralise les notions de partitions dures et floues et permet en particulier de gérer les points aberrants. Un des buts de ce projet sera d'étudier l'application de ces algorithmes pour qualifier la notion de rareté des espèces étudiées. En outre, les notions de rareté, de spécificité, de fréquence d'une espère peuvent être apparentées à la notion de préférence associé ici, par exemple, à un milieu aquatique. En s'appuyant sur des métriques de qualité, les méthodes de classification extraient des connaissances qui doivent être pertinentes pour les experts. En effet la qualité des connaissances extraites constitue une préoccupation majeure des utilisateurs, notamment par la prise en compte de leurs préférences, laquelle préférence permet d'exprimer la rareté ou l'abondance de la connaissance recherchée. Le LIMOS développe des approches agrégatives d'exploration de connaissances (**Bouker et al. 2012**), basées sur les préférences et utilisant le principe du front de Pareto (Godfrey et al. 2006) à travers une relation de dominance entre les connaissances extraites. Ce projet permettra aussi d'étudier l'intégration de cette approche dans le cadre d'une méthode ensembliste de classification.

- *le positionnement du projet par rapport à la concurrence*

Le LMGE peut être considéré par la nature de sa production scientifique comme un laboratoire leader au niveau international dans la description de la biosphère rare. Le LIMOS a développé ces dernières années des méthodes originales pour traiter le problème de la classification non supervisée en présence d'incertitudes, et pour la prise en compte des préférences de l'utilisateur pour l'extraction de connaissances. L'association de l'expertise d'un laboratoire reconnu dans l'utilisation du séquençage haut débit pour étudier les communautés microbiennes (LMGE) à celle d'un laboratoire d'informatique (LIMOS) est un atout original dans ce domaine de recherche.

- *les objectifs du projet et planning des travaux*

Les objectifs du projets consistent à :

1- étudier les méthodes existantes de classification ensembliste, adapter les solutions développées au sein du LIMOS au contexte de la biosphère rare, et en proposer de nouvelles intégrant l'incertitude et la préférence (~ 8 mois - LIMOS)

2- générer de nouvelles connaissances sur la biosphère rare en appliquant ce nouveau développement aux données de NGS générées par les laboratoires de la fédération (SOERE GLACPE, ANR ROME et divers EC2CO : milieux aquatiques et nuages...) et à celles disponibles dans les bases de données publiques (~ 10 mois - LMGE).

Le planning des travaux (18 mois) répartis entre le LIMOS et le LMGE, sera le suivant :

T : Démarrage du projet. Réunion de démarrage entre les partenaires.

T à T+5 : Recherche bibliographique

Etude du contexte applicatif, et des différentes méthodes de classification ensembliste

Conception des solutions informatiques envisagées

T+3 à T+9 : Constitution et préparation des données d'évaluation

Développement des approches proposées et évaluation

Evaluation des approches développées

T+9 à T+11 : Rédaction rapport et Valorisation scientifique

T+10 à T+14 : Déploiement des méthodes sur la plateforme ouverte ePANAM

T+12 à T+16 : Etude de cas

Analyse et interprétation des connaissances issues du déploiement des méthodes

T+15 à T+17 : Test de la plateforme ouverte d'analyse

Rédaction d'un rapport technique de maintenance

T+16 à T+18 : Diffusion de la plateforme

Valorisation scientifique des analyses effectuées

Analyse des retombées du projet. Réunion Bilan. Rédaction rapport de fin projet.

- *les verrous scientifiques et technologiques*

Une difficulté majeure réside dans la variabilité de certaines espèces présentes dans la biosphère. La seconde difficulté est relative à l'abondance des données à traiter, nécessitant une adaptation des méthodes actuellement pour une meilleure efficacité de calcul, mais aussi la nécessité de disposer d'une plateforme de calcul et de stockage correctement dimensionnées.

En outre, dans ce type de projet un autre verrou important est la capacité à communiquer (échanges technologiques/scientifiques) entre deux disciplines très différentes. Or, ce verrou devrait être dans ce cas présent assez faible puisque les deux laboratoires ont déjà collaboré dans deux programmes régionaux de bioinformatique.

- *le rôle de chacun des partenaires dans le projet*

Le LIMOS apporte dans ce programme son expertise en algorithmie et calcul pour classer par de nouvelles méthodes des objets qui sont ici des espèces microbiennes. En complément le LMGE pourra tester ces nouvelles méthodes sur des jeux de données (séquençage haut-débit) déjà acquis et proposer de nouvelles analyses (méta-analyse) sur les écosystèmes microbiens.

- *la coordination et la gouvernance*

La coordination sera assurée par Engelbert Mephu-Nguifo et Didier Debroas et se traduira par des réunions régulières organisées par le post-doctorant recruté.

- *les perspectives de valorisation (diffusion de connaissances, valorisation économique)*

La diffusion du travail se fera sous la forme de communications académiques. Les avancées technologiques seront intégrées (i.e. classification) seront intégrées au site web ePANAM (http://panam-meb.univ-bpclermont.fr/) en cours d'élaboration. La mise en place de tels sites

internet a un fort impact sur la communauté scientifique et donc sur la promotion du savoir-faire d'une région comme le montre l'utilisation du site METAVIR (http://metavir-meb.univ-bpclermont.fr/) dans le monde entier.

- *l'intégration du projet dans l'axe du CPER*

Ce travail permettra de compléter les nouveaux algorithmes développés dans le cadre de financements du CPER (axe environnement) et plus généralement régionaux relatifs au traitement de données issues des NGS: bourse CPER de Najwa Taïb et 3 programmes LifeGrid METAPROC (http://metavir-meb.univ-bpclermont.fr/), PREFON META (http://com.isima.fr/PREFON_META/description-du-projet-prefon-meta) et ePANAM (http://panam-meb.univ-bpclermont.fr/).

Il permettra entre autre de mieux comprendre la dynamique de la biodiversité dans les écosystèmes de référence étudiés dans le cadre du CPER.

- *l'adéquation du sujet traité par l'allocation avec la thématique de l'appel à projets 2014*

Le programme relève à la fois de l'analyse de grandes masses de données, de la modélisation (par méta-analyse) et du calcul intensif. Les données à traiter (objectif 2) correspondent en effet à de grandes masse de données puisque un simple run de séquençage d'amplicons par pyroséquençage et par Illumina génère respectivement 1 et 12 millions de séquences. Ainsi, à titre d'exemple, une seule équipe du LMGE a généré en 2 ans par pyroséquençage d'amplicons 19 millions de séquences et les données de métagénomique ont donné lieu à 6239 publications (251 correspondent à du séquençage d'amplicons) (source pubmed, 9/01/2014). Par conséquent, le traitement (méthode développée lors de l'étape 1 associée à celle développée par Najaw Taïb dans le cadre du CPER) fera aussi appel à du calcul intensif.

Ce travail s'appuiera sur une forte collaboration avec le CRRI et sera couplé à une demande de moyens informatiques concernant les moyens de calcul et le stockage complétant des investissements déjà réalisés dans le cadre du CPER et du DIPPE CNRS.

- Présentation du laboratoire d'accueil (environnement matériel et humain)

Le travail sera réalisé dans deux laboratoires du campus le LIMOS et le LMGE et le post-doc sera encadré par Engelbert Mephu-Nguifo (LIMOS) et Didier Debroas (LMGE)

LIMOS

Le LIMOS est une unité mixte de recherche (UMR 6158) du CNRS et de l'Université Blaise Pascal. Il comprend environ 90 enseignants-chercheurs et chercheurs, 80 doctorants et postdoctorants, répartis sur Clermont-Ferrand et aussi Saint-Etienne. Les travaux développés en son sein concernent les schémas algorithmiques permettant le traitement fin de modèles d'optimisation, statiques, dynamiques ou collaboratifs ; les schémas algorithmiques permettant le traitement intelligent des données : reconnaissance des formes, apprentissage… ; les procédés d'évaluation des performances des systèmes vus comme des systèmes dynamiques (simulation…) ; et les couches d'acquisition (réseaux de capteurs, web services…), de stockage (bases de données, systèmes d'information, grilles…) et de gestion du transit des données (réseaux, web services, systèmes de médiation, systèmes embarqués…).

Le post-doc bénéficiera de l'ensemble de l'environnement scientifique et technique (encadrement humain et expertise) présent au LIMOS. Les moyens demandés dans le cadre de ce projet lui permettront aussi de disposer d'outils de calcul et de stockage dédiés à la réalisation du projet. Il sera rattaché à l'axe « Systèmes d'Information et de Communication » du laboratoire, et plus particulièrement associé au thème « Données, Services et Interopérabilité », et à l'axe transversale « STIC pour les Sciences de la Vie et de la Santé ».

V. Antoine, B. Quost, M.-H. Masson, T. Denoeux. 2013, CEVCLUS : Evidential clustering with instance-level constraints for relational data. **Soft Computing**, 1-15.
V. Antoine, B. Quost, M.-H. Masson and T. Denoeux. 2012, CECM: Constrained Evidential C-Means algorithm. **Computational Statistics and Data Analysis**, 56: 894-914
S. Bouker, R. Saidi, S. Ben Yahia, **E. Mephu Nguifo**. 2012, "Ranking and selecting association rules based on dominance relationship", 24th IEEE Intl. Conf. on Tools for Artificial Intelligence ICTAI, 1: 658-665.

LMGE

Le post-doc bénéficiera de l'encadrement humain et de l'expertise présents au sein de l'équipe « Microbiologie de l'environnement et bioinformatique ». L'objectif de cette équipe est de comprendre la structuration des assemblages, leurs changements dans le temps et dans l'espace et leurs interactions fonctionnelles. L'équipe étudie les mécanismes de spéciation qui conduisent à l'émergence d'écotypes, les transferts horizontaux de gènes et les réassemblages des communautés microbiennes en prenant en compte les espèces rares. L'écosystème modèle pour étudier ces communautés est l'écosystème lacustre, pour lequel la diversité et la physiologie des microorganismes sont relativement peu étudiées. Parmi ces microorganismes, nous nous intéressons plus particulièrement au picoplancton (0,2 – 5 µm) qui, en l'absence de caractéristiques morphologiques distinctes et en raison des difficultés inhérentes à la mise en culture, est étudié préférentiellement par des méthodes d'écologie moléculaire : séquençage massif d'amplicons, métagénomique microbienne, virome, métagénomique ciblée .... En appui de ces méthodes d'écologie moléculaire, cette équipe développe des outils de bioinformatique pour traiter les données issues des séquenceurs à haut-débit (NGS). Le matériel informatique utilisé est celui mis à disposition par le CRII (serveurs, clusters, grille de calcul)

Hugoni M, Taib N, **Debroas D**, Domaizon I, Jouan Dufournel I, Bronner G, Salter I, Agogué H, Mary I, Galand PE. (2013). Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. Proc Natl Acad Sci U S A. 110(15):6004-9. doi : 10.1073/pnas.1216863110.
Simon Roux, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, **Didier Debroas** and François Enault (2011). METAVIR : a web server dedicated to virome analysis. Bioinformatics 27(21):3074-5

Date limite de candidature : 31 décembre 2014. Le poste pourra être pourvu avant cette date si un candidat satisfaisant est trouvé

Date de début souhaité : **entre Janvier et Avril 2015**

Pièces constituant le dossier de candidature :

- curriculum vitae (état civil, études et titres, activités d'enseignement, activités de recherche, encadrement de travaux de recherche ; publications et communications nationales et internationales)
- attestation d'obtention de la thèse
- lettre de motivation du candidat,
- lettre de recommandation des directeurs de laboratoire (du laboratoire de réalisation de la thèse et du laboratoire d'accueil),

Conditions obligatoires pour l'obtention d'une allocation post-doctorale :

- avoir soutenu sa thèse dans un laboratoire hors région AUVERGNE (de préférence à l'étranger) ; le candidat peut avoir effectué sa thèse en cotutelle avec un laboratoire régional si l'apport du laboratoire étranger est prépondérant,
- ne pas être en poste dans le laboratoire d'accueil avant validation de la candidature pas les services de la Région.

Salaire mensuel brut : **2356€**

Contacts :

* LIMOS
- MEPHU NGUIFO Engelbert : mephu@isima.fr

* LMGE :
- DEBROAS Didier : didier.debroas@univ-bpclermont.fr