

MASTER BIOINFORMATIQUE, UCA

PROPOSITION DE STAGE M2

ANNÉE UNIVERSITAIRE 2022-2023



Nom du laboratoire d'accueil : Mésocentre Université Clermont Auvergne
Directeur : Antoine Mahul

Nom et fonction de l'encadrant : Goué Nadia, Ingénieure Bioinformatique Mésocentre, pf AuBi
Co-encadrants : Franck Giacomoni, Nils Paulhe, UMR UNH, pf PFEM, MetaboHUB
Équipes de recherche : Plateforme AuBi, équipe PFEM UMR UNH - MetaboHUB

Titre : Construction et déploiement de microservices en métabolomique sur l'infrastructure CLOUD du Mésocentre Clermont Auvergne

Contexte

Le Mésocentre Clermont Auvergne mutualise les ressources en calcul et en stockage au sein de l'Université Clermont Auvergne. Il est l'acteur central pour tout laboratoire de l'UCA et laboratoires associés qui requiert l'usage d'un cluster de calcul, d'une grille ou d'un Cloud ainsi que de stockage adossé aux calculs. En son sein, il héberge la plateforme AuBi¹ (Auvergne Bioinformatique), qui fédère treize laboratoires de recherche en Biologie autour de la bioanalyse et la bioinformatique. Dans ce contexte, la mission principale de AuBi est de proposer un appui opérationnel à la recherche en Sciences de la Vie auprès de ses utilisateurs en mettant à disposition des plateformes de bio-analyse comme Galaxy et de stockage comme OMERO, connectés aux ressources en calcul et stockage. Depuis peu, le Mésocentre promeut des services sur le Cloud comme ShinyProxy et initie l'hébergement de microservices afin de répondre à des demandes plus spécifiques notamment ceux de la communauté AuBi sur des questions thématiques non couvertes par les outils sus-mentionnés. En effet, le terme « microservices » englobe plusieurs concepts qui, ensemble, forment un style architectural pour le développement et la mise à disposition d'applications. La partie « micro » signifie que les composants individuels sont petits et utiles individuellement. La partie « services » signifie que l'exécution de la fonction est distante et ne dépend donc pas de l'infrastructure informatique locale. Ainsi, les microservices sont généralement de petites fonctions, exécutées par un service distant, souvent via une interface de programmation d'application ou API. L'approche par microservices décompose une application en une série de fonctions faiblement couplées qui, ensemble, accomplissent un processus complexe (Sheffield et al. 2022). Cette approche centrée sur la métabolomique a été initiée par l'unité de Nutrition Humaine (UNH) et la plateforme d'Exploration du Métabolisme (PFEM) notamment via son implication dans le consortium PhenoMeNal² (Khoonsari et al., 2020).

La plateforme PFEM est une infrastructure publique INRAE, leader au niveau français et à l'international en métabolomique appliquée à la nutrition et à la santé. Ses principales innovations de recherche concernent le développement de méthodes analytiques basées sur la spectrométrie de masse et d'outils bio-informatiques pour une meilleure caractérisation des déviations métaboliques associées au développement de maladies métaboliques chroniques et l'étude des relations entre nutrition et santé. Cela concerne en particulier i) le profilage métabolique utilisant des approches multiplateformes dans les biofluides et les tissus, ii) le développement de stratégies analytiques pour l'identification de biomarqueurs à l'aide de la spectrométrie de masse à ultra-haute résolution et d'outils bioinformatiques, iii) le développement de modèles et d'outils pour améliorer l'extraction de connaissances à partir de données à haut débit et iv) le développement d'outils, de bases de données et de systèmes d'information basés sur les technologies des « Big Data » et du web sémantique pour structurer ces données et exploiter les connaissances existantes. Actuellement, l'équipe bioinformatique de la PFEM travaille à la mise en place de ces

1 <https://mesocentre.uca.fr/projets-associes/plateforme-aubi>

2 <https://phenomenal-h2020.eu>

bases de connaissances à travers le développement d'API REST respectant la norme OpenAPI V3³ (i.e. l'API Peakforest⁴, la gestion de bibliothèques de composés chimiques ou de publications scientifiques ...) sur une infrastructure haute performance et haute disponibilité fournie par le mésocentre de l'UCA.

L'objectif du stage est de proposer une interface graphique aux utilisateurs pour des outils de traitement de données métabolomiques utiles à la plateforme PFEM et scalables sur une infrastructure Cloud adaptée aux technologies des « Big Data ». Ce projet se décomposera en plusieurs phases. Dans un premier temps, il s'agira de prendre en main les outils bioinformatiques de métabolomique, ensuite de développer une version Docker si les conteneurs n'existent pas puis de mettre en place des API REST sous forme de microservices, adaptés à ce type d'infrastructure. Le code développé devra répondre aux bonnes pratiques de développement notamment en termes d'intégration continue et de déploiement (CI/CD). Ce projet nécessitera également d'appréhender le fonctionnement technique de l'infrastructure Cloud de type OpenStack qui, au Mésocentre, se base sur l'orchestrateur Kubernetes⁵. Enfin, une fois les microservices en place, il s'agira de prendre en charge l'optimisation des ressources en calcul nécessaire à l'exécution des analyses ainsi que l'amélioration de l'ergonomie de ces microservices.

Compétences recherchées :

- curiosité scientifique
- motivation à évoluer dans un environnement d'informaticiens et une équipe pluridisciplinaire
- avoir des notions de programmation en Python et/ou en Java
- avoir une connaissance des outils bio-informatiques utilisés en métabolomique est un plus

Compétences qui seront acquises par l'étudiante ou l'étudiant pendant le stage :

Bien que les outils soient à destination des biologistes, les compétences acquises seront à dominante informatique.

- connaissances des outils / packages utilisés en bioinformatique et métabolomique
- design d'API REST basées sur la norme OpenAPI
- design d'images Docker
- mise en place de pipelines de CI/CD sur la forge logicielle gitLab
- utilisation d'une infrastructure de type Cloud

Objectifs du stage de Master 2 bioinformatique

Les objectifs opérationnels sont de :

- Faire une analyse bibliographique sur une sélection d'outils de bioinformatique (Chemistry Development Kit⁶, OpenBabel⁷, BioSM⁸, ...)
- Se faire une culture informatique sur l'orchestration de services Cloud (Kubernetes, docker-compose, ...) auprès des informaticiens du Mésocentre et de la PFEM
- Développer des API REST
- Réaliser, tester et optimiser plusieurs microservices d'outils développés en Python et/ou Java
- déployer ses microservices sur l'infrastructure Cloud du Mésocentre, le cas échéant proposer un serveur web
- Accompagner les premiers utilisateurs dans la prise en main de l'outil
- Selon les avancées du projet, mettre en place un pipeline de déploiement automatique de ces microservices via la forge logicielle du projet et établir une procédure de standardisation d'un déploiement de microservices

Mots clés (5) :

Métabolomique Cloud Bioinformatique Python/JAVA Microservices

Références bibliographiques récentes de l'équipe

Paulhe N, Canlet C, Damont A, Peyriga L, Durand S, ..., Pujos-Guillot E, Giacomoni F. PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management. *Metabolomics*. 2022 . doi: 10.1007/s11306-022-01899-3.

3 <https://spec.openapis.org/oas/latest.html>

4 <https://metabohub.peakforest.org/doc-v2/index.jsp>

5 <https://kubernetes.io/>

6 <https://cdk.github.io/>

7 <https://sourceforge.net/p/openbabel/>

8 Hamdalla et al. 2013

Delmas M, Filangi O, Paulhe N, Vinson F, Duperier C, Garrier W, Saunier PE, Pitarch Y, Jourdan F, Giacomoni F, Frainay C. FORUM: Building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics*. 2021. doi: 10.1093/bioinformatics/btab627.

Giacomoni F, Le Corguillé G, Monsoor M, Landi M, Pericard P, Pétéra M, Duperier C, Tremblay-Franco M, Martin JF, ..., Caron C. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*. 2015. doi: 10.1093/bioinformatics/btu813.

Autres références bibliographiques

Hamdalla MA, Mandoiu II, Hill DW, Rajasekaran S, Grant DF. BioSM: metabolomics tool for identifying endogeneous mammalian biochemical structures in chemical structure space. American Chemical Society. 2013. doi: 10.1021/ci300512q

Sheffield NC, Bonazzi VR, Bourne PE, Burdett T, Clark T, Grossman RL, Spjuth O, Yates AD From biomedical cloud platforms to microservices: next steps in FAIR data analysis. 2022. doi :10.1038/s41597-022-01619-5

Contacts :

Nadia Goué Tél :04 73 40 75 64 Email : nadia.goue@uca.fr - Laboratoire d'accueil

Franck Giacomoni Tél : 04 73 62 47 72 Email : franck.giacomoni@inrae.fr

Nils Paulhe Tél : 04 73 62 47 72 Email : nils.paulhe@inrae.fr

(co-encadrants scientifiques et techniques)