Master Bioinformatique

Promotion 2024-2025

Soutenances de stage M2 Bioinformatique
30 juin, 1er et 2 juillet 2025

Programme et résumés

Jury :
Anne-Muriel ARIGON
Sèverine BÉRARD
Vincent BERRY
Anthony BOUREUX
Annie CHATEAU
Thérèse COMMES
Sylvain DAUDÉ
Rodolphe GIROUDEAU
Corinne LAUTIER
Alban MANCHERON
François SABOT
Konstantin TODOROV
Christine TRANCHANT-D.

Table des matières

Planning des soutenances – Stages M2 Bioinformatique

Campus Triolet - Amphi 36.03
30 juin, 1er et 2 juillet 2025

Les soutenances sont publiques [1] et durent environ 30 minutes, questions incluses.
Les temps de délibérations à huis-clos après chaque soutenance permettent au public d'entrer et de sortir entre chaque soutenance.

Lien visio : https://moodle.umontpellier.fr/mod/bigbluebuttonbn/view.php?id=825529
Mail de contact : master-bioinfo-stages@umontpellier.fr
Site des stages et projet du master : https://moodle.umontpellier.fr/course/view.php?id=28744 où vous retrouverez le lien visio, les dates importantes et le Vadémécum et ce programme.

	Lundi 30	Mardi 1er	Mercredi 2
09h00	1 **Brendan Vouadec** Thérèse Commes	⚠début à 10h15⚠	17 **Clément Raspail** Konstantin Todorov
09h45	2 **Rahma Baaziz** Thérèse Commes		18 **Khadidiatou Gueye** Konstantin Todorov
10h30	Pause	10 **Vetea Jacot** Rodolphe Giroudeau	Pause
11h00	3 **Loreleï Berger** Sylvain Daudé	11 **Florent Marchal** Rodolphe Giroudeau	19 **Mattéo Traissac-M.** François Sabot
11h45	4 **Hugo Bellavoir** Sylvain Daudé	12 **Mouhammad Thiam** Christine Tranchant-D.	20 **Pablo Bertogna** Alban Mancheron
12h30	Pause Repas	Pause Repas	Pause Repas
14h00	5 **Emma Mathieu** Annie Chateau	13 **Paul Medout-Marere** Vincent Berry	Jury
14h45	6 **Margaux Imbert** Annie Chateau	14 **Sandrine Vrignon** Anthony Boureux	
15h30	7 **Anna Fall** Annie Chateau	Pause	Clôture du M2
16h00	Pause à 16h15	15 **Thomas Vitré** Anne-Muriel Arigon	Moment festif
16h45	8 **Amel Benarbia** Corinne Lautier	16 **Guilhem Biosse** Anne-Muriel Arigon	Fin de journée
17h30	9 **Hazar Sandakly** Corinne Lautier	Fin de journée	
18h15	Fin de journée		

Les numéros indiqués à côté de chaque soutenance permettent de retrouver le résumé correspond dans la suite de ce programme.
Les cases jaune correspondent à des soutenances où l'étudiant·e sera en visio, et les cases rouge aux soutenances se tenant à huis-clos à la demane de l'entreprise.

1. sauf celle à huis-clos indiquée en rouge dans le planning

1 Deep learning methods for automatic annotation of aphid feeding behaviour of electropenetrography (EPG)

Mots clés : Deeplearning, Electropenetrography

Stage réalisé par **Brendan Vouadec** *Soutenance le 30/06/25 à 09h00*

Encadrement scientifique : Piotr Trębicki (Macquarie University).
Tuteurice pédagogique : Thérèse COMMES

Résumé Studying the feeding behaviour of aphids is crucial to understanding epidemics in plants. Electropenetrography involves forming an electrical circuit between the plant and the aphid, in order to record in real time the movements of the aphid's mouthparts within the plant's tissues and cells. The analysis and annotation of these electrical signals is time-consuming, and would require automation. My project aims to test and optimize machine learning and deep learning tools to improve the annotation of feeding behaviour in aphids.

2 Détection de caractéristiques discriminant les ARN codants et non codants dans le but d'améliorer les méthodes de classification en machine learning

Mots clés : machine learning, features, codant, non-codants, lncRNA, mRNA, élément transposable, transcriptome, de novo

Stage réalisé par **Rahma Baaziz** *Soutenance le 30/06/25 à 09h45*

Encadrement scientifique : Daniel GARCIA RUANO (CBiB/Université de Bordeaux), Domitille CHALLOPIN-FILLOT (CBiB, ImmunoConcept)
Tuteurice pédagogique : Thérèse COMMES

Résumé La majorité du génome humain est constitué d'ADN non codant, avec seulement 3 % étant composé de gènes codants pour des protéines. Cependant, près de 80 % du génome est en réalité transcrit en divers types d'ARN de tailles et de fonctions différentes. Parmi ces ARNs, les longs ARN non-codants (lncRNA) jouent un rôle central dans la régulation des gènes, la définition de l'identité cellulaire et peuvent être impliqués dans différentes maladies, y compris le cancer. Ainsi, leur étude est primordiale, mais l'identification de lncRNAs est compliquée dû à leurs caractéristiques communes avec les ARN codants. Pour faire face à cette problématique, des outils de Machine Learning ont été développés permettant la classification des transcrits en codants ou non-codants, mais ces outils présentent des limites.

L'équipe cherche donc à développer un nouvel outil de machine-learning afin de mieux identifier de nouveaux lncRNAs dans des transcriptomes de novo. Pour cela, mon projet vise à (1) comprendre à quel point les bases de données sont représentatives des transcriptomes de novo, et (2) comparer différents outils afin d'identifier les caractéristiques communes des transcrits mal classés. Ce stage ouvre la voie pour développer une nouvelle méthode de classification de transcrits qui permettra de découvrir de nouveaux lncRNAs.

3 Classification clinique automatisée des troubles de la parole en chirurgie éveillée des tumeurs cérébrales : Une étude par apprentissage machine

Mots clés : gliome, troubles de la parole, Intelligence Artificielle (IA), neurosciences

Stage réalisé par **Loreleï Berger** *Soutenance le 30/06/25 à 11h00*

Encadrement scientifique : Guillaume Herbet (Praxiling)
Tuteurice pédagogique : Sylvain DAUDÉ

Résumé La chirurgie éveillée des tumeurs cérébrales est une approche innovante qui vise à minimiser les risques de séquelles neurologiques chez les patients subissant une résection tumorale. Cette technique permet d'assurer que les régions critiques du cerveau ne soient pas endommagées pendant l'intervention chirurgicale. Cependant, la préservation des réseaux cérébraux impliqués dans la parole présente des défis spécifiques. Certains troubles sont souvent difficiles à discerner cliniquement, notamment en raison de la brièveté des réponses évocables et de la résolution limitée du système auditif humain. Dans ce contexte, notre projet vise à développer une méthodologie permettant de classifier automatiquement ces troubles articulatoires à partir des signaux acoustiques de la parole enregistrés lors des interventions chirurgicales. Pour ce faire, nous nous appuyons sur des algorithmes d'apprentissage machine (intelligence artificielle) capables d'analyser et d'identifier des patrons acoustiques représentatifs des différentes perturbations induites par la stimulation corticale.

4 Identification et caractérisation du contenu des inversions chez Pseudogymnoascus destructans

Mots clés : Pathogène, inversions, TE, gènes

Stage réalisé par **Hugo Bellavoir** *Soutenance le 30/06/25 à 11h45*

Encadrement scientifique : Anna-Sophie Fiston-Lavier, Sèverine Bérard et Sébastien Puechmaille (ISEM)
Tuteurice pédagogique : Sylvain DAUDÉ

Résumé Un nombre croissant d'études suggèrent un rôle significatif des variants structuraux dans de multiples processus éco-évolutifs, tels qu'observés chez les champignons pathogènes. Les inversions comptent parmi les variants structuraux les plus fréquemment associés aux phénotypes adaptés localement. Pseudogymnoascus destructans (Pd) est un champignon pathogène responsable de la maladie du museau blanc chez les chauves-souris. L'objectif de ce travail est d'identifier et de caractériser les inversions chez Pd afin de mieux comprendre leur impact sur l'évolution du génome. Pour cette étude, nous avons utilisé des séquences longues générées par la technologie de séquençage Oxford Nanopore pour assembler les génomes de 14 isolats de Pd. Les assemblages ont été générés avec Flye (v2.9.5) et échafaudés avec les génomes les plus complets (35 à 40 Mb) au niveau chromosomique avec Ragtag (v2.1.0).Les inversions ont ensuite été détectées à l'aide de Syri (1.7.0), un outil de détection de variants structuraux qui identifie les blocs de synténie, puis les différents types de réarrangements basés sur l'alignement global par paires, réalisé ici avec Minimap2 (v2.28). Les inversions identifiées seront comparées à celles détectées par Cactus (v2.9.7), un outil utilisant l'alignement multigénomique. Les premières analyses révèlent que, dans certains de nos isolats, jusqu'à 20 % du génome présente des inversions. Nous prévoyons de comparer la densité des éléments transposables (ET), des gènes dans les inversions et dans le reste du génome.

5 Étude in silico des récepteurs olfactifs chez les salmonidés : analyse de séquence, annotation structurelle et prédiction d'affinité

Mots clés : Récepteurs olfactifs, Salmonidés, Modélisation structurale, Docking moléculaire, Analyse de séquences

Stage réalisé par **Emma Mathieu** *Soutenance le 30/06/25 à 14h00*

Encadrement scientifique : Dr DURAIRAJ Rajesh (IRSEA)
Tuteurice pédagogique : Annie CHATEAU

Résumé Chez les téléostéens, la perception des signaux chimiques de l'environnement repose en grande partie sur les récepteurs olfactifs, une sous-classe des récepteurs couplés aux protéines G. Ces récepteurs jouent un rôle essentiel dans la détection des composés chimiques impliqués dans des fonctions clés telles que l'alimentation, la reproduction ou la réponse au stress. Cependant, une majorité de ces récepteurs restent sans ligand identifié, notamment chez les salmonidés comme *Salmo salar* (saumon atlantique) et *Oncorhynchus mykiss* (truite arc-en-ciel), deux espèces d'intérêt majeur en aquaculture. Cette étude a pour objectif d'améliorer la compréhension des mécanismes de reconnaissance des sémiochimiques chez ces espèces. Elle s'appuie sur plusieurs approches complémentaires : l'analyse de séquence et phylogénétique pour identifier des homologues conservés de récepteurs olfactifs connus, la modélisation structurale en tenant compte de l'architecture transmembranaire, l'affinage par dynamique moléculaire en environnement lipidique pour valider la stabilité des structures, et enfin des simulations de docking moléculaire pour évaluer les interactions avec différents composés. Les résultats obtenus permettront de proposer des couples récepteur-ligand potentiels et de mieux comprendre les mécanismes de reconnaissance chimique impliqués. À terme, cette recherche ouvre des perspectives pour le développement de solutions ciblés en pisciculture, en identifiant des composés susceptibles de moduler le comportement des salmonidés d'élevage.

Soutenance à huis clos

6 GraDiv : un outil pour calculer des statistiques de diversité depuis un graphe de pangénome

Mots clés : graphe de pangénome, polymorphisme, package python

Stage réalisé par **Margaux Imbert** *Soutenance le 30/06/25 à 14h45*

Encadrement scientifique : Stéphane De Mita (PHIM), Christine Tranchant (DIADE), Sébastien Ravel (PHIM)
Tuteurice pédagogique : Annie CHATEAU

Résumé Mon sujet consiste à développer un outil de calcul de statistiques de diversité à partir d'un graphe de pangénome, dans le cadre du projet EggLib. Un graphe de pangénome permet de capturer la diversité génétique d'un groupe (espèce, population, métagénome, etc). Il n'existe actuellement pas d'outils permettant de calculer directement des statistiques de diversité depuis un graphe de pangénome.

Le développement de cet outil consiste dans un premier temps à appeler les polymorphismes à partir du graphe. Pour cela, j'ai testé un outil d'appel de variant (vg deconstruct), puis j'ai développé et implémenté un nouvel algorithme. Dans un second temps, les statistiques sont calculées à partir de ces polymorphismes à l'aide de la librairie EggLib.

L'outil est développé sous forme d'un package python. Il est nommé GraDiv, pour graph diversity. Il propose deux commandes pour appeler les polymorphismes et calculer des statistiques. Il s'appuie également sur l'outil GraTools pour la manipulation de graphes de pangénomes, pour l'indexation du graphe.

7 Analyse de données de snRNAseq sur un modèle murin du syndrome d'Ehlers Danlos vasculaire

Mots clés : Annotation de données de séquençage unicellulaire à haute dimension SnRNAseq Omics Génétique maladie cardiovasculaire Biomarqueurs

Stage réalisé par **Anna Fall** *Soutenance le 30/06/25 à 15h30*

Encadrement scientifique : Antonio Rausell (Institut Imagine/INSERM), Isabelle Jéru (PARCC/APHP/INSERM)
Tuteurice pédagogique : Annie CHATEAU

Résumé Dans le cadre de mon stage de Master 2, je travaille sur un projet de recherche portant sur le syndrome d'Ehlers-Danlos vasculaire (SEDv), une maladie génétique rare liée à une mutation du gène COL3A1. Cette mutation affecte le collagène de type III, une protéine essentielle à la structure des vaisseaux sanguins.

Pour mieux comprendre les mécanismes de la maladie, j'analyse des données de single-nucleus RNA sequencing (snRNAseq) obtenues à partir de l'aorte thoracique de souris porteuses de la mutation. Cette technique permet d'étudier l'expression des gènes à partir des noyaux isolés des cellules, ce qui est particulièrement utile dans les tissus complexes comme l'aorte.

Mon travail consiste à identifier les types cellulaires présents, à comparer l'expression des gènes entre souris mutées et témoins, et à explorer les voies de signalisation impliquées. Ce stage me permet de développer des compétences en bioinformatique, en analyse de données transcriptomiques et en biologie des maladies génétiques.

8 Caractérisation du paysage des réarrangements génomiques associés aux cassures double-brin des loci transcrits

Mots clés : Cassures double brin de l'ADN (DSBs), Réponse aux dommages de l'ADN (DDR), Réarrangements génomiques, Jonctions chimériques

Stage réalisé par **Amel Benarbia** *Soutenance le 30/06/25 à 16h45*

Encadrement scientifique : Gaëlle Legube (CBI-MCD/CNRS), Sarah Djebali (IRSD/Inserm)
Tuteurice pédagogique : Corinne LAUTIER

Résumé Les cassures double-brin de l'ADN (DSB) sont des lésions préjudiciables qui peuvent induire des mutations et des réarrangements génomiques, pouvant contribuer au développement de cancers [1]. Le laboratoire de Gaëlle Legube a développé une lignée cellulaire d'ostéosarcome humain dans laquelle des cassures double brin peuvent être induites à des positions spécifiques et annotées en utilisant un traitement particulier et l'enzyme de restriction AsiSI. Ce modèle permet d'étudier la réparation des cassures (DNA Damage Response ou DDR) [2] grâce à des techniques génomiques à haut débit telles que ChIP-seq, Hi-C et RNA-seq. Nous cherchons à caractériser le paysage des réarrangements génomiques suite à des DSB. Comme nous savons que la plupart des DSB de notre système se produisent dans des loci transcrits, nous pouvons identifier les réarrangements génomiques en utilisant les données RNA-seq produites avant et après l'induction des DSB. C'est ce que nous avons fait en utilisant le programme ChimPipe qui permet d'identifier les transcrits (plus précisément les jonctions transcrites) entre différents loci du génome et produits à la suite d'événements tels que des réarrangements génomiques [3].

Je présenterai les résultats obtenus après avoir exécuté ChimPipe sur 12 expériences RNA-seq produites par le laboratoire LEGUBE : 6 dans des cellules induites par des DSB et 6 dans des cellules contrôle. Je fournirai le nombre d'événements spécifiques à l'induction de DSB, ainsi que leur distribution génomique (exonique, intronique, intergénique) et le type de jonction chimérique (read-through, interchromosomal, etc). Je montrerai également comment ces réarrangements génomiques sont distribués par rapport aux DSB induites en elles-mêmes, en tenant compte des mécanismes de réparation des DSB. Cela permettra de mieux comprendre les mécanismes de réparation des lésions de l'ADN et leur contribution à l'instabilité génomique.

Références

1. Clouaire T, Rocher V, Lashgari A, Arnould C, Aguirrebengoa M, Biernacka A, et al. Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures. Molecular Cell. 2018 Oct;72(2) :250-262.e6.
2. Arnould C, Rocher V, Finoux AL, Clouaire T, Li K, Zhou F, et al. Loop extrusion as a mechanism for formation of DNA damage repair foci. Nature. 2021 Feb ;590(7847) :660–5.
3. Rodríguez-Martín B, Palumbo E, Marco-Sola S, Griebel T, Ribeca P, Alonso G, et al. ChimPipe : accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data. BMC Genomics. 2017 Jan 3 ;18(1) :7.

9 Séquençage long-read pour le développement d'un outil multimodal destiné au suivi du cancer colorectal à partir de biopsie liquide

Mots clés : Biopsie liquide, ADN tumoral circulant (ctDNA), Cancer colorectal, Maladie résiduelle minimale, Variabilité du nombre de copies

Stage réalisé par **Hazar Sandakly** *Soutenance le 30/06/25 à 17h30*

Encadrement scientifique : Laia Bassaganyas (IGF, CNRS)
Tuteurice pédagogique : Corinne LAUTIER

Résumé Ce projet porte sur le développement d'un outil multimodal pour le suivi du cancer colorectal, basé sur le séquençage long-read par Oxford Nanopore. L'objectif est d'exploiter l'ADN tumoral circulant (ctDNA) extrait du plasma afin de détecter des altérations génomiques comme les variations du nombre de copies (CNA, copy number alterations), ainsi que des signaux liés à la méthylation ou à la fragmentation de l'ADN. En combinant plusieurs types d'informations issues d'un même échantillon, l'objectif est d'améliorer la sensibilité de la détection tumorale, notamment à faibles fractions tumorales, comme dans le contexte de la maladie résiduelle minimale.

Dans le cadre d'un stage de six mois, le travail s'est concentré sur l'analyse bioinformatique des CNA, à partir de données simulées (in silico) et réelles. Un benchmarking de plusieurs outils spécialisés dans la détection des CNA (ichorCNA, ACE, QDNAseq) a été réalisé, ainsi que la mise en place d'un pipeline adapté aux données issues de biopsies liquides séquencées à faible couverture (low-pass whole-genome sequencing).

10 Caractérisation des liens sémantiques entre tweets et articles scientifiques référencés

Mots clés : références scientifiques, récupération d'informations, modèles de langage

Stage réalisé par **Vetea Jacot** *Soutenance le 01/07/25 à 10h15*

Encadrement scientifique : Sandra Bringay (LIRMM / UPV)
Tuteurice pédagogique : Rodolphe GIROUDEAU

Résumé Les discussions de résultats scientifiques sur les réseaux sociaux vont souvent inclure des références aux études les ayant produits. Cependant ces références sont souvent implicites, c'est-à-dire sans hyperlien ou méta-données complètes de l'étude. Cela rend plus difficile la vérification de l'affirmation scientifique du message initial et la réflexion indépendante par les autres utilisateurs, ce qui participe probablement aux effets négatifs des réseaux sociaux comme la propagation de fake news ou la polarisation.

Il serait donc socialement utile de pouvoir réaliser de manière automatique et adaptative la récupération automatique de l'article référencé implicitement par un message de réseau social. Cette récupération devrait obligatoirement s'appuyer sur les similarités sémantiques entre le message initial et chaque article d'un corpus pour identifier l'article-source.

Mon stage a consisté en la caractérisation des liens sémantiques pouvant exister entre des tweets et des articles scientifiques en lien au COVID-19. Un jeu de données associant ces deux types de documents a été établi préalablement dans l'équipe, et mon travail a consisté en l'identification de caractéristiques sémantiques dans les tweets favorisant ou défavorisant l'identification d'un unique article-source. J'ai de plus réalisé un benchmark de nombreux modèles de langage de l'état de l'art (LLM basés sur les sentence-transformers) afin d'évaluer leur capacité à résoudre la tâche de récupération automatique de la source de l'affirmation scientifique d'un tweet.

11 Optimisation et refactoring Nextflow pour fluidifier et optimiser les phases de tests et déploiements automatisés de pipeline bioinformatique

Mots clés : Python3, GitLab, NextFlow, Intégration continue, Déploiement continue, AWS, Kubernetes

Stage réalisé par **Florent Marchal** *Soutenance le 01/07/25 à 11h00*

Encadrement scientifique : Marc Chakiachvili (Seqone), Raphael Lanos (Seqone)
Tuteurice pédagogique : Rodolphe GIROUDEAU

Résumé

— Réduction du temps d'exécution de l'intégration continue en déplaçant les jeux de tests sur un seau s3 (AWS)
— Automatisation du refactoring d'un dépôt de code Nextflow
— Création d'un cluster Kubernetes afin d'optimiser les ressources attribuées aux différents process d'un pipeline Nextflow

12 Analyse translationnelle du phagéome cutané

Mots clés : Métagénomique, phagéome, pipeline, phagothérapie, acné

Stage réalisé par **Mouhammad Thiam** *Soutenance le 01/07/25 à 11h45*

Encadrement scientifique : Phuong Le (EBI), Mano Mathew (EFREI)
Tuteurice pédagogique : Christine TRANCHANT-D.

Résumé Mise en place et amélioration d'un pipeline d'annotation taxonomique et fonctionnelle dans le cadre de la thérapie phagique.

Soutenance à huis clos

13 Recherche et Développement d'outils pour la détection de variants d'épissage dans des données de RNA-Seq

Mots clés : RNA-seq, épissage alternatif, workflow, HPC, Sclérose Latérale Amyotrophique (SLA), Benchmark

Stage réalisé par **Paul Medout-Marere** *Soutenance le 01/07/25 à 14h00*

Encadrement scientifique : Anthony Boureux (INSERM/UM), Jérôme Reboul (INSERM)
Tuteurice pédagogique : Vincent BERRY

Résumé Dans le cadre de la recherche médicale et du diagnostic de pathologies génétiques rares, l'étude de mécanisme biologique tel que l'épissage alternatif s'est révélée prometteur afin d'identifier l'impact de mutation sur l'expression des gènes. Durant mon apprentissage au sein de l'équipe BIO2M à Montpellier, j'ai appris à explorer et comprendre les problématiques liées à l'épissage alternatif. À travers plusieurs projets dont l'analyse de variants de patients atteints de Sclérose latéraleAmyotrophique (SLA) et le développement d'outils liés aux kmers dans le contexte de l'épissage,nous tentons d'apporter des solutions à ces problèmatiques afin d'améliorer notre compréhension deces données et de compléter le diagnostic médicale.

14 IN palm un modèle de prédiction des émissions d'azote développé sous Python à partir d'un format Microsoft Excel ®

Mots clés : modèle conceptuel de donnée, python, ACV, logique flou, modèle de bilan de masse, modèle de régression, palmier à huile, programme

Stage réalisé par **Sandrine Vrignon** *Soutenance le 01/07/25 à 14h45*

Encadrement scientifique : Cécile Bessou (CIRAD)
Tuteurice pédagogique : Anthony BOUREUX

Résumé Ce projet s'inscrit dans une réflexion pour inclure un outil d'indicateur prédictif agro-environnemental spécifique aux plantations de palmiers à huile : IN-Palm au sein d'une ACV (Analyse du Cycle de Vie). Cet indicateur a été initialement créé par Lénaïc Pardon en 2019 sous la forme d'un fichier Excel. Il permet de simuler via trois modèles opérationnels les risques de pertes d'azote par hectare d'une parcelle de palmiers à huile (âgés de moins de 31 ans).

Le but de ce projet, est de recréer l'indicateur sous la forme d'un programme codé en Python (version 3.13.2) tout en améliorant les aspects FAIR (Détectable, Accessible, Interopérable et Réutilisable).

Ainsi afin de permettre cela, 3 étapes principales ont été effectués :

— Compréhension de l'indicateur (variables internes, modèles opérationnels, interactions utilisateurs ...)
— Réflexion de mise en place du programme (type de programme, interaction utilisateur,structure de données à adopter, création de modèles de conception de données ...)
— Programmation sous python du modèle.

15 Clusterisation spatio-phylogénétique pour la simplification de scénarios phylogéographiques

Mots clés : Phylogeographie, visualisation de l'information, algorithmes de clusterisation

Stage réalisé par **Thomas Vitré** *Soutenance le 01/07/25 à 16h00*

Encadrement scientifique : François Chevenet (IRD)
Tuteurice pédagogique : Anne-Muriel ARIGON

Résumé L'objectif de ce mémoire est de développer des méthodes de clusterisation spatio-phylogénétique capables de regrouper automatiquement, dans un jeu de données soumis par l'utilisateur, les nœuds proches à la fois sur l'arbre phylogénétique et sur la carte géographique. Cette approche vise un double objectif :

1. Simplifier des scénarios de dispersion souvent complexes et illisibles en réduisant le nombre de trajets affichés, et en appliquant une coloration synchronisée des branches de l'arbre et des zones géographiques correspondantes ;
2. Mettre en évidence des phénomènes localisés, tels que les foyers de diversification, les introductions multiples ou les barrières de diffusion, particulièrement utiles pour le biologiste.

Cette simplification est complétée par la construction d'un graphe de transition qui résume les échanges majeurs entre clusters, offrant une représentation synthétique et lisible qui remplace avantageusement l'arbre complet sans perte d'information clé.

Les prototypes développés, implémentés en JavaScript, HTML5, ont été testés sur deux ensembles de données contrastés :

— un foyer compact du virus de Lassa en Afrique de l'Ouest, et
— une diffusion large et dispersée du virus West Nile en Allemagne.

Ce rapport fait suite à mon projet d'alternance de Master 1, dans la continuité des travaux initiés l'année précédente.

16 Le paramétrage du MES dans un bâtiment de Sanofi

Mots clés : MES, informatisation de procédés, environnement réglementé

Stage réalisé par **Guilhem Biosse** *Soutenance le 01/07/25 à 16h45*

Encadrement scientifique : Kevin Cutajar (SANOFI)
Tuteurice pédagogique : Anne-Muriel ARIGON

Résumé Le MES est un système d'aide à la production. Dans mon alternance, je paramètre ce système pour compléter le suivis sanitaire des équipement sur un bâtiment de Sanofi. Je parlerais de ce projet et des différentes contraintes de la production pharmaceutique.

Soutenance à huis clos

17 Ingénieur traitement d'image (Image Processing Engeener)

Mots clés : Industrialisation, Optimisation, Machine Learning, Imagerie medicale

Stage réalisé par **Clement Raspail** *Soutenance le 02/07/25 à 09h00*

Encadrement scientifique : Estanislao Oubel (Intrasense)
Tuteurice pédagogique : Konstantin TODOROV

Résumé Mon sujet d'alternance porte sur l'optimisation et l'industrialisation des différents projets. Cette mission couvre plusieurs aspects cruciaux permettant de garantir que les modèles et algorithmes, une fois conçus, soient prêts pour leur intégration dans les logiciels finaux. Parmi mes responsabilités, on retrouve notamment :

— Industrialisation des algorithmes : cela inclut la mise en place de pipelines de traitement, la conteneurisation des environnements d'exécution (via Docker), l'application de règles de développement (codestyle, gestion des dépendances), ainsi que la maintenance de la documentation associée.
— Intégration des algorithmes : en collaboration avec l'équipe d'intégration (AIHub), l'intégration des algorithmes dans leurs services pour le déploiement. Modification des algorithme suivant leurs demandes (généralement sur les entrée/sorties).
— - Tests et validation des algorithmes : une autre phase de mon travail consistait à mettre en place des processus de vérification sur le plan technique (tests unitaires, performance informatique).

Soutenance à huis clos

18 Développement d'une interface web unifiée pour l'exploitation des graphes de connaissances IMGT_KG en immunogénétique

Mots clés : Web sémantique, graphe de connaissances, immunogénétique, anticorps monoclonaux, Vue.js, visualisation de données, SPARQL

Stage réalisé par **Khadidiatou Sall Gueye** *Soutenance le 02/07/25 à 09h45*

Encadrement scientifique : Gaoussou Sanou (UM, IGH, LIRMM) et Patrice Duroux (IGH)
Tuteurice pédagogique : Konstantin TODOROV

Résumé Le stage porte sur le développement d'une nouvelle interface web destinée à l'exploration du graphe de connaissances IMGT-KG, un graphe structurant les données immunogénétiques issues des bases de données d'IMGT®, première ressource mondiale dans le domaine de l'immunogénétique. En plus de cette base, une seconde instance du graphe a été développée pour les anticorps monoclonaux thérapeutiques, sous le nom d'IMGT/mAb-KG.

Actuellement, ces deux graphes sont explorables à travers deux interfaces distinctes :une interface en Vue.js (https://imgt.org/imgt-kg/) pour IMGT-KG et une interface en Streamlit (https://imgt.org/mAb-KG/) pour IMGT/mAb-KG.

Cependant, cette séparation entraîne des incohérences d'expérience utilisateur, une duplication des efforts de maintenance, et limite l'exploration croisée des deux graphes.

L'objectif du stage est donc de fusionner ces deux interfaces en une application web unifiée, permettant à la fois une exploration visuelle intuitive des entités et relations du graphe, et une exploration libre par requêtes SPARQL personnalisées. La nouvelle interface offrira également une documentation dynamique intégrée, des filtres interactifs, une visualisation graphique des triplets RDF, et une architecture logicielle plus claire et évolutive.

Ce travail combine des compétences en développement web (Vue.js, TypeScript), en Web sémantique (SPARQL, RDF), et en structuration de graphes de connaissances, tout en s'inscrivant dans un contexte scientifique et biomédical exigeant.

19 Développement et amélioration de ProRNAScan

Mots clés : pipeline, snakemake, python, classification, optimisation

Stage réalisé par **Mattéo Traissac-Montahut** *Soutenance le 02/07/25 à 11h00*

Encadrement scientifique : Nicolas Gilbert (INSERM) Anthony BOUREUX (INSERM)
Tuteurice pédagogique : François SABOT

Résumé Je développe un pipeline snakemake permettant de classifier des séquences génomiques spécifiques en fonction de ses caractéristiques : l'ARN cellulaire U6.

Il se distingue en 4 catégories :
- gène seul
- pseudogène avec une queue polyA
- pseudogène tronqué
- pseudogène avec une séquence répétée (chimère)

Pour ce faire, le pipeline va :
- effectuer un BLAST de U6 sur chaque génome (environ 800 mammifères)
- récupérer les séquences flanquantes
- identifier les TSD (signature de rétrotransposition)
- détecter les séquences répétées
- rechercher une queue polyA (wordmatch)
- ⇒ attribution de la catégorie

But final : étudier la diversité des 4 catégories sur les différents génomes pour avoir une idée de l'activité de L1 (gène sauteur souvent lié à U6) dans la nature.

20 Impact of repeats on Xanthomonas oryzae pv. oryzae evolvability

Mots clés : Répétition, adaptation, bactérie

Stage réalisé par **Pablo Bertogna** *Soutenance le 02/07/25 à 11h45*

Encadrement scientifique : Alvaro Luis Perez-Quintero (XPLAIN/IRD)
Tuteurice pédagogique : Alban MANCHERON

Résumé Après avoir dans un premier temps généré un workflow afin d'analyser les données génomiques de bactéries du genre Xanthomonas. J'analyse à présent les données en lien avec la présence de séquences uniques et répétées dans les génomes. Nous essayons de comprendre pourquoi au sein d'un même genre, des espèces peuvent avoir des taux de répétitions complètement différents. Tout d'abord nous cherchons à connaître les conséquences de ces répétitions sur le génome (hypermutabilité, réarrangement fréquent, diminution de la densité génique ?) mais aussi les causes (forte pression de sélection, domestication des plantes hôtes, hasard ?).