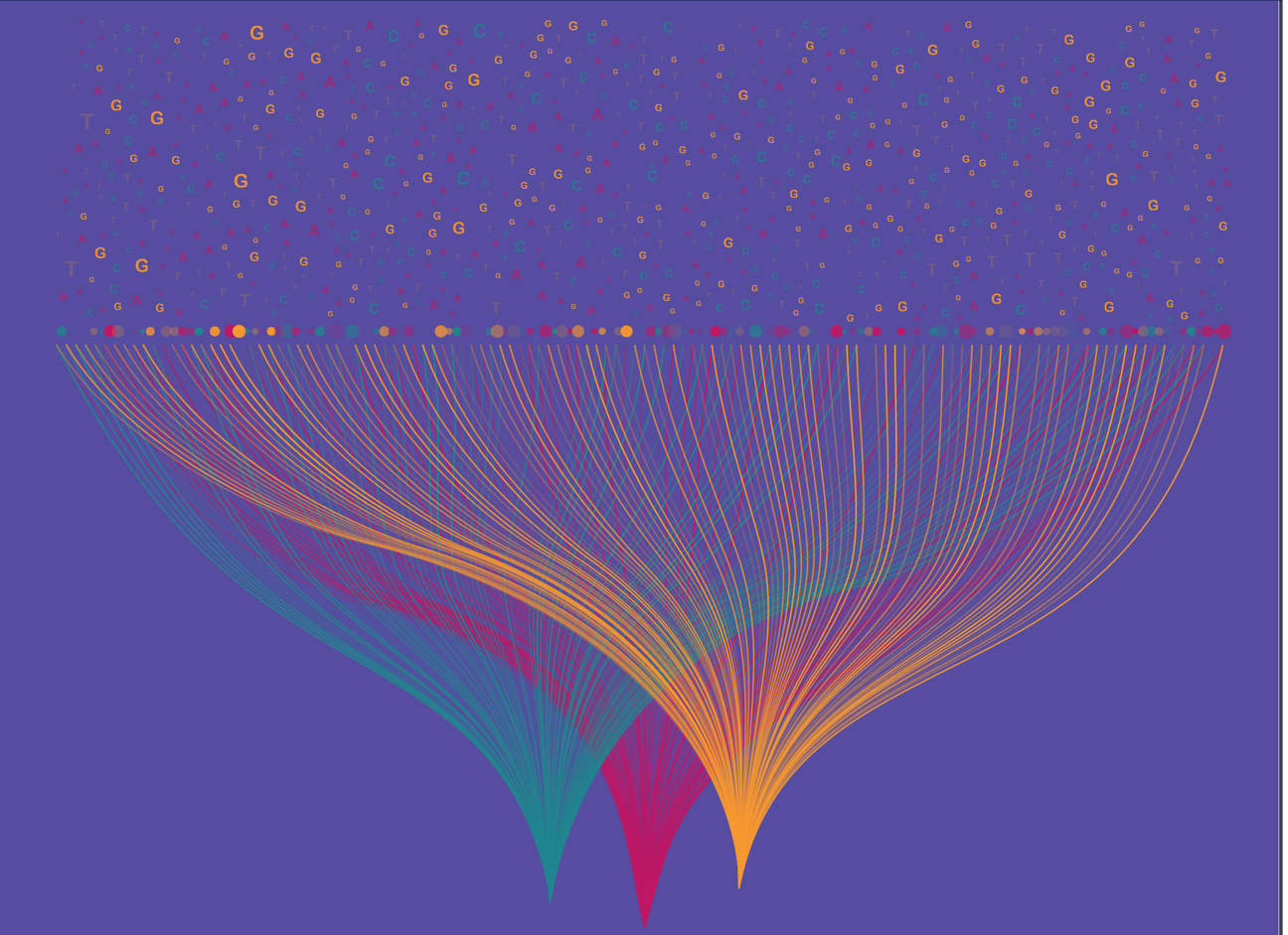

5BM573 (6 ECTS)

Apprentissage supervisé et non-supervisé pour la biologie

1er semestre 2025

Initiez-vous aux techniques d'analyse permettant d'explorer et d'interpréter des données biologiques complexes.

SCIENCES SORBONNE UNIVERSITÉ

Objectifs pédagogiques

Alors que les données biologiques gagnent en richesse et en complexité, leur analyse repose sur une diversité croissante de méthodes statistiques et algorithmiques. Cette unité d'enseignement propose une introduction aux principales approches supervisées et non supervisées appliquées à la biologie, en soulignant leur complémentarité, leurs bases théoriques et leur mise en œuvre concrète.

Les étudiants seront formés à la sélection et à l'utilisation pertinente de ces outils pour explorer, modéliser et interpréter des jeux de données biologiques complexes, en s'appuyant sur des bibliothèques de référence en R et Python. Au-delà de la maîtrise technique, l'UE encourage une réflexion critique sur les choix méthodologiques, les limites des modèles et les enjeux liés à la reproductibilité.

Les étudiants seront formés aux méthodes de modélisation statistique et aux outils de programmation pour traiter des données issues des sciences du vivant. Ces compétences les aideront à construire des analyses reproductibles et à extraire des connaissances pertinentes à partir de données complexes.

Programme pédagogique

Cette unité d'enseignement combine cours, travaux dirigés et projets d'analyse de données menés en groupes d'étudiants.

Le programme pédagogique couvre les aspects suivants :

- Introduction à l'apprentissage non supervisé (réduction de dimension et clustering).
- Méthodes d'apprentissage supervisé pour la classification et la régression.
- Techniques de sélection de variables pertinentes (feature selection).
- Modélisation de sorties complexes (structured output prediction).
- Imputation des données manquantes par des approches statistiques robustes.
- Reconstruction de réseaux de co-expression et d'interactions biologiques.
- Utilisation de modèles interprétables pour une meilleure explicabilité.
- Stratégies d'apprentissage sous contrainte de ressources (budget learning).
- Encadrement statistique des prédictions avec la prédiction conforme.
- Création de pipelines analytiques reproductibles avec Orange 3.
- Introduction aux principales bibliothèques Python et R.
- Approches pour la gestion des déséquilibres de classes dans les jeux de données biologiques.
- Visualisation avancée des résultats pour l'interprétation et la communication scientifique.
- Réflexion critique sur les limites et les biais liés à l'utilisation de modèles en biologie.
- Introduction aux principales bases de données et entrepôts de données en biologie.
- Sensibilisation aux enjeux de reproductibilité et aux bonnes pratiques des analyses.
- Réalisation d'un projet complet avec rapport écrit et soutenance orale.

Organisation

Cette unité d'enseignement de 6 ECTS se déroule au premier semestre de l'année universitaire.

Les modalités de contrôle des connaissances reposent sur deux projets indépendants :
- Un projet d'analyse en R, avec rédaction d'un rapport structuré sous forme d'article scientifique
- Un projet d'analyse en Python, donnant lieu à une soutenance.

Les deux projets s'appuieront sur des jeux de données réels issus de la recherche en biologie moléculaire et cellulaire. Chaque groupe travaillera sur une thématique spécifique, parmi lesquelles l'immunologie, la génétique, la microbiologie, la biologie cellulaire ou la biologie moléculaire. Les productions finales comprendront un rapport écrit structuré comme un article scientifique ainsi qu'une soutenance orale de discussion des résultats.