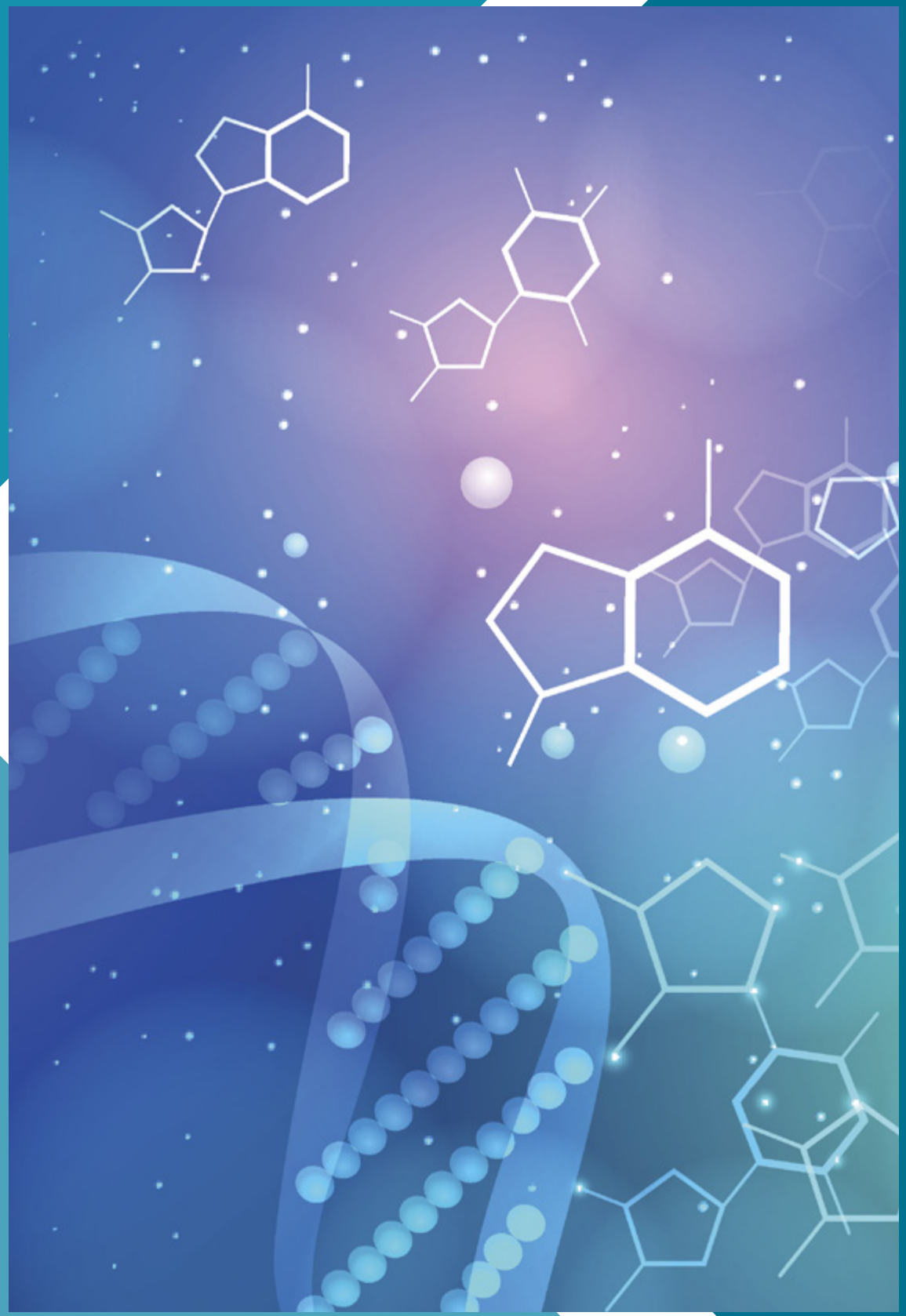

BIOINFORMATIQUE

Sebastian Will
sebastian.will@polytechnique.edu

En une cinquantaine d'années, l'informatique est devenue une composante essentielle dans beaucoup de secteurs de la recherche et de l'ingénierie en biologie.

De nouveaux domaines se sont imposés qui, sans informatique, n'existeraient pas, du simple fait de la masse de données à traiter : par exemple, le séquençage des génomes et donc la génomique comparée, la classification des protéines et des ARN, la prédiction et l'ingénierie des structures qui ouvrent des voies nouvelles en biologie, pharmacologie et médecine, avec les enjeux nouveaux de la médecine génomique. Réciproquement, les questions soulevées par la biologie et les singularités des données biologiques induisent de nouveaux problèmes en algorithmique, combinatoire et apprentissage.

Si le séquençage à haute vitesse atteint maintenant des débits spectaculaires, certaines analyses restent encore longues, délicates et coûteuses. Les données « -omiques » et cliniques peuvent aussi être rares pour certaines pathologies et l'exception est souvent la règle. Le recours systématique à l'apprentissage massif touche ainsi ses limites. Enfin, le biologiste et le médecin attendent des explications qui permettent d'augmenter la compréhension des phénomènes sous-jacents et pas simplement une optimisation statistique.

La modélisation, l'algorithmique et la science des données prennent donc une importance croissante. On s'attache à la construction de modèles informatiques et mathématiques pour prédire mais aussi expliquer les phénomènes biologiques qui se produisent au niveau de la cellule et de ses composants, et aussi pour des groupes de cellules, voire des organismes complexes. Il s'agit souvent de concevoir des outils dédiés à des phénomènes très particuliers.

Mais les techniques de l'information seules ne sauraient résoudre les problèmes et de larges connaissances en biologie restent indispensables en amont et en aval pour une contribution certaine aux sciences du vivant. La Bioinformatique, à l'interface entre Biologie et Informatique, est ainsi une discipline très diversifiée et qui évolue rapidement, tant les questions posées se renouvellent vite au fur-et-à-

mesure des découvertes en biologie et du progrès des outils d'observation du monde vivant.

Les Neurosciences constituent un domaine encore plus vaste. Les compétences acquises dans le Programme d'Approfondissement Bioinformatique constituent aussi une bonne introduction pour aborder les Neurosciences avant une spécialisation dans ce domaine en quatrième année. Si l'on envisage une telle voie, un complément par le cours HSS524 – « Sciences cognitives, sciences naturelles de l'esprit » est alors indiqué.

La plupart des cours de Biologie concernent assez directement l'étude du système nerveux, plusieurs cours d'Informatique s'appliquent naturellement à l'analyse des données en Neurosciences, et d'autres aux algorithmes de simulation des mécanismes neuronaux.

Objectifs

Le programme d'approfondissement en Bioinformatique est une introduction aux concepts et enjeux de cette discipline, à travers quelques cours généraux en biologie et en informatique et quelques autres modules qui peuvent constituer un début de spécialisation. Il s'agit de se donner la compétence double qui permettra d'interagir au plus haut niveau avec des informaticiens, des biologistes et des médecins, au sein des équipes pluridisciplaires qui structurent maintenant la recherche et le développement dans ce domaine.

Un choix de cours relativement vaste, théoriques ou appliqués, donne la possibilité de se construire un parcours individualisé, pour compléter sa formation au cas par cas et préparer un projet de carrière adapté, en fonction de ses centres d'intérêts.

Cette spécialisation se poursuivra en quatrième année, avant une orientation aussi bien vers le monde académique que vers le monde de l'entreprise.

Dans ce dernier cas aussi, s'agissant le plus souvent de sociétés internationales, la thèse est toujours une étape conseillée pour une carrière de haut niveau.

Règles de composition du programme

Durant les deux premières périodes, l'accent est mis sur la consolidation des connaissances dont l'utilisation transversale sera développée au cours du stage de recherche. Le projet long, étalé sur les deux périodes, est aussi un moyen d'aborder cet exercice pluridisciplinaire. Les règles imposées veillent à maintenir un certain équilibre entre biologie et informatique, avec quelques compléments possibles en statistique, afin d'acquérir la compétence étendue qui fait la spécificité du bioinformaticien. Ces règles visent aussi à éviter la dispersion, tout en permettant à chacun de consolider ses points faibles et d'approfondir le domaine pressenti pour un début de carrière.

Prérequis

Les prérequis généraux pour ce programme sont d'avoir validé, en deuxième année :

- au moins un cours de Biologie, de préférence, BIO452 – Biologie moléculaire et information génétique,

- au moins deux cours d'Informatique (hors modal), de préférence, INF421 – Conception et analyse d'algorithmes et INF442 – Algorithmes pour l'analyse de données en C++,

- au moins un projet de programmation (soit intégré dans un cours d'informatique, soit comme le modal d'Informatique).

Lors de son inscription, il convient également de vérifier les prérequis ou cours conseillés spécifiques à chaque cours, comme indiqués sur « moodle.polytechnique.fr ».

En particulier, il est conseillé d'avoir suivi en deuxième année le cours MAP433 – « Statistiques », surtout si l'on incorpore des cours MAP dans son cursus 3A. Le cours MAP432 – « Modélisation de phénomènes aléatoires » peut être aussi utile.

Débouchés

À l'issue du programme d'approfondissement, il est possible de continuer un cursus en bioinformatique. Il est possible aussi de se réorienter vers l'informatique ou la biologie, en conservant un bénéfice de ce début de double formation bioinformatique. La poursuite dans un cursus monodisciplinaire suppose bien entendu que cette discipline soit restée une majeure du cursus de 2A et 3A, étayant ainsi cette motivation.

Comme formations relativement équilibrées entre les deux disciplines, on peut citer, par exemple,

- le « MSc in Computer Science with option in Bioinformatics » de McGill,

- le « MSc of Science in Computational Biology and Quantitative Genetics » de Harvard,

- le « MSc in Computer Science with track in Computational Biology » de Columbia,

- le « MSc in Bioinformatics » de University of Copenhagen avec deux spécialisations « Computational Biology » et « Computer Science »,

- le « MPhil in Computational Biology » de Cambridge,

- le « MSc in Bioinformatics and Theoretical Systems Biology » de Imperial College London,

- le « Master in Computational Biology and Bioinformatics » de ETH Zürich…

On peut ajouter, en France, des M2 comme « AMI2B » de Paris-Saclay, le parcours BIM-BMC « BioInformatique et Modélisation » du master d'Informatique

de Sorbonne Université et d'autres à Bordeaux, Aix-Marseille, Toulouse…

La durée de ces formations est variable, de un an pour un M2 en France ou certains MSc, à deux ans pour des programmes plus complets. Elles proposent en général une voie recherche qui prépare à la poursuite en thèse.

Les débouchés en recherche académique, après une thèse, sont en laboratoires universitaires ou dans les grands centres de recherche français (INRA, Institut Pasteur, Institut Curie, Inserm…) ou internationaux (ex. EBI, SIB, NCBI…). Les activités exercées peuvent concerner notamment la conception d'algorithmes d'analyse ou de prédiction, mais également le développement, le nettoyage et l'enrichissement des grandes bases de données qui sont mises à disposition de la communauté, ainsi que l'ingénierie des plateformes qui les hébergent.

Les débouchés en entreprise, de préférence après une thèse, peuvent être en Recherche et Développement dans les industries pharmaceutique et agroalimentaire, l'agrochimie, l'ingénierie pour la santé, l'environnement, ou les biotechnologies.

Il existe aussi des débouchés en gestion et valorisation de la R&D et en conseil ou management de l'innovation technologique.

Règles de choix

Règle de cohérence globale :

Au total, sur les deux périodes et les 8 modules obligatoires, il faut suivre au moins

- 3 modules (cours, EA ou projet) de Biologie (pris dans la liste ci-dessous) et
- 3 modules (cours, EA ou projet) d'Informatique (pris dans la liste ci-dessous).

Dans ce calcul, BIO/INF588 peut compter comme BIO ou comme INF et un projet long compte pour 2 modules.

Projet(s) :

La réalisation d'un projet long, couvrant les deux périodes, est encouragée. C'est la meilleure occasion de développer des compétences transversales, avant le stage de 3A. Cela peut se réaliser dans le cadre des 8 modules obligatoires, à la place de l'EA de chaque période.

Si un projet long est réalisé de manière facultative (cours supplémentaire), il devra être réalisé en totalité, ce qui correspond à l'obtention d'une note supplémentaire par période.

Attention :
Plusieurs cours d'Informatique, notamment ceux marqués EA, demandent la réalisation d'un projet de programmation significatif et comptant pour la validation. Il convient à chacun de mesurer la charge de travail lors de son choix de cours.

Règle de panachage :

Le panachage est autorisé dans la limite d'un module par période et il ne peut pas être utilisé pour la réalisation d'un projet long autre que BIO511, BIO572 ou INF511.

Un panachage ne sera accepté que s'il s'intègre dans un cursus cohérent, motivé et clairement explicité lors de l'inscription.

URL:
https://www.enseignement.polytechnique.fr/bioinformatique

COMPOSITION DU PROGRAMME

Période 1

3 cours au choix

BIO551 – Immunologie et agents infectieux
BIO553 – Biotechnologies pour la médecine et l'agriculture
BIO556 – Genomes, diversity, environment and human health
BIO557 - Neurosciences
INF550 – Algorithmique avancée
INF552 – Data Visualization
INF555 – Constraint-based Modeling and Algorithms for Decision-making
INF556 – Topological data analysis

1 EA au choix ou projet

BIO571A – Travaux expérimentaux de génie génétique
BIO571B – Travaux expérimentaux en imagerie quantitative
INF554 – [EA] Machine and Deep learning
INF573 – [EA] Image Analysis and Computer Vision

Périodes 1 et 2

Le projet long remplace les EA de période 1 et 2, ou il peut être pris comme deux modules optionnels, un par période.
1 projet long au choix parmi

BIO511 – Projet de Biologie
BIO572 – Reconstitution Personnalisée du Processus Tumoral
INF511 – Projet de Bioinformatique

Les projets longs sont toujours accordés sous réserve de la définition préalable d'un sujet, notamment pour BIO572 et INF511.

Période 2

1 cours obligatoire

INF589 – Computational analysis of high-throughput sequencing data

2 cours au choix parmi

BIO562 – Biologie des systèmes moléculaires
BIO563 – Epigénétique et ARN non-codants
INF580 – Large scale mathematical optimization
INF581 – Advanced Topics in Artificial Intelligence
MAP566 – Statistics in action

1 EA au choix ou projet

BIO583 – Sciences des données en imagerie biologique
BIO/INF588 – Projet en bioinformatique

Période 3

Stage de recherche au choix

BIO591 – Biologie et Écologie
INF591 – Informatique
INF592 - Data Science