

Madame Magali BERLAND soutiendra sa thèse de doctorat en Biologie Informatique , intitulée : " Exploration de méthodes statistiques pour la modélisation de la relation séquence-activité de protéines d'intérêt industriel " , sous la direction de Monsieur Bernard OFFMANN le :

**Mardi 29 octobre 2013**  
**A partir de 10h00 (Heure de Métropole)**  
**Amphithéâtre A**  
**Faculté des Sciences et Techniques**  
**Université de Nantes**

**Composition du jury :**

- Madame Isabelle ANDRE, Directrice de Recherches, CNRS
- Madame Marie CHABBERT, Chargée de Recherches, H.D.R., Université d'Angers
- Madame Magali REMAUD-SIMEON, Professeur, Institut National des Sciences Appliquées
- Madame Christine SINOQUET, Maître de Conférences, Université de Nantes
- Monsieur Frédéric CADET, Professeur, Université de La Réunion
- Monsieur Philippe CHARTON, Maître de Conférences, Université de La Réunion
- Monsieur Srinivasan NARAYANASWAMY, Professeur, Indian Institute of Sciences
- Monsieur Bernard OFFMANN, Professeur, Université de Nantes

**Résumé:**

Par l'accumulation de mutations bénéfiques lors de cycles successifs de mutagenèse, l'évolution dirigée offre un cadre rationnel pour l'amélioration des protéines à vocation industrielle. Elle permet une exploration large de l'espace possible des séquences ainsi que leurs capacités fonctionnelles. Elle est cependant lourde à mettre en oeuvre et nécessite des moyens importants. Des approches in silico font usage d'un jeu minimal de données expérimentales et utilisent la modélisation statistique combinée à des algorithmes d'apprentissage machine. Elles ont été développées pour explorer de façon heuristique l'espace possible des séquences et de la fitness et d'identifier les mutations et interactions entre résidus les plus intéressantes. C'est l'objet de cette thèse qui explore la construction et l'application de modèles statistiques s'appuyant sur des jeux minimaux de données expérimentales pour relier fitness, ou activité, à la séquence biologique des variants. L'étude s'articule autour d'un choix crucial d'une méthode de numérisation, de descripteurs de la séquence et de méthodes de régression. La méthode ProSAR de R. Fox (2005) et les limites de son applicabilité sur des jeux de données expérimentales ont été étudiées. De nouvelles méthodes ont aussi été développées, prenant en compte les propriétés physico-chimiques des acides aminés et leurs périodicités. Elle a permis de découvrir de nouveaux descripteurs reliant la séquence à l'activité et propose des approches innovantes qui ont la capacité de traiter des cadres biologiques très divers, même lorsque peu de données biologiques sont disponibles

La soutenance est publique