Proposition de recherche doctorale

# Evolution et apprentissage automatique pour l'annotation fonctionnelle et la classification des homologies lointains en protéines.

**Mots clés :**

- **Directeur de thèse** : Alessandra Carbone
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

## Résumé du projet de recherche (Langue 1)

La détection d’homologues lointains est essentielle pour le classement fonctionnel et structural des séquences protéiques et pour l’amélioration de l'annotation des génomes très divergents. Pour le classement des séquences, nous présentons la méthode «ILP-SVM homology», combinant la programmation logique inductive (PLI) et les modèles propositionnels. Elle propose une nouvelle représentation logique des propriétés physico-chimiques des résidus et des positions conservées au sein de l’alignement de séquences. Ainsi, PLI trouve les règles les plus fréquentes et les utilise pour la phase d'apprentissage utilisant des modèles d'arbre de décision ou de machine à vecteurs de support. La méthode présente au moins les mêmes performances que les autres méthodes trouvées dans la littérature. Puis, nous proposons la méthode CASH pour annoter les génomes très divergents. CASH a été appliqué à Plasmodium falciparum, mais reste applicable à toutes les espèces. CASH utilise aussi bien l'information issue de génomes proches ou éloignés de P. falciparum. Chaque domaine connu est ainsi représenté par un ensemble de modèles évolutifs, et les sorties sont combinées par un méta-classificateur qui assigne un score de confiance à chaque prédiction. Basé sur ce score et sur des propriétés de co-ocurrences de domaines, CASH trouve l'architecture la plus probable de chaque séquence en appliquant une approche d'optimisation multi-objectif. CASH est capable d'annoter 70\% des domaines protéiques de P. falciparum, contre une moyenne de 57\% pour ses concurrents. De nouveaux domaines protéiques ont pu être caractérisés au sein de protéines de fonction inconnue ou déjà annotées.