

Développement de méthodes d'intégration de sources de connaissances hétérogènes pour une annotation unifiée de groupes de gènes

*Aarón Ayllón Benítez*¹

En biomédecine, la découverte de nouveaux vaccins permet de réaliser de nouveaux tests à grande échelle. Ces tests évaluent l'impact d'un vaccin en comparant deux populations : les personnes en bonne santé et les personnes atteintes d'une maladie. Un intérêt de cette comparaison est de pouvoir connaître la relation entre les gènes mis en jeu et leur phénotype. La comparaison de ces populations permet d'identifier des groupes de gènes impliqués dans des processus immunologiques d'après leur niveau d'expression.

Cependant, la quantité d'informations associées à un groupe de gènes peut être très importante. Dans ce cadre, des méthodes statistiques d'enrichissement ont été proposées afin de découvrir les annotations surreprésentées. Ce type d'analyse présentant des limites, j'ai développé une méthode qui permet de décrire le rôle et les fonctions d'un groupe de gènes de n'importe quel organisme en récupérant uniquement l'information pertinente afin de mieux comprendre les processus biologiques complexes impliqués. Cette méthode a été mise en œuvre sur un serveur Web qui offre une visualisation interactive pour faciliter l'exploration des annotations de groupes de gènes par les experts.

Utilisabilité : haute sécurité en basse technologie

*Nikola K. Blanchard*²

Cette thèse mêle cryptographie et expériences utilisateur pour s'attaquer à deux sujets : l'authentification et les systèmes de vote.

Nous commençons par une étude typographique afin de créer de meilleurs codes à usage unique. Nous continuons avec un algorithme pour corriger les typos dans les mots de passe sans perdre en sécurité, avant de proposer un gestionnaire de mots de passe mental : un algorithme permettant de calculer rapidement, et de tête, tous

1. Thèse dont la soutenance est prévue en décembre 2019, préparée au sein du Laboratoire bordelais de recherche en informatique (LaBRI), sous la direction de Fleur Mouglin et Patricia Thébault.

2. Thèse soutenue le 21 juin 2019, préparée au sein de l'Institut de recherche en informatique fondamentale (IRIF), Université Paris Diderot, sous la direction de Nicolas Schabanel (ENS Lyon) et Ted Selker (UMBC).

nos mots de passe, avec des garanties de sécurité si certains sont volés. Ce système dépend d'un secret appelé phrase de passe, et nous proposons une manière de créer des phrases à la fois plus mémorisables et plus sécurisées. Nous terminons cette section par une réfutation empirique du principal modèle de calcul mental humain, utilisé pour construire ces algorithmes.

Nous passons enfin aux systèmes de vote en analysant les difficultés à implémenter le vote vérifiable en pratique, à travers des tests empiriques. Nous proposons un protocole de vote vérifiable ne nécessitant que du papier, et terminons par l'analyse de primitives et d'un protocole de vote en basse technologie pour une situation particulière : celle des conseils d'administration.

De l'utilisation des données publiques pour la prédiction de la toxicité des produits chimiques

Ingrid Grenet³

L'évaluation de la sécurité des composés chimiques repose sur des études *in vivo* coûteuses en temps, argent et animaux de laboratoire et donc inadaptées à l'étude de milliers de composés. Pour prédire rapidement la toxicité de ces composés, de nouvelles solutions sont envisagées comme les tests *in vitro* et les modèles de *machine learning*. L'objectif de ma thèse est d'évaluer comment les données publiques permettent de construire de tels modèles afin de prédire les effets *in vivo* induits par les composés, à partir de leur structure chimique.

Nous considérons d'abord la prédiction de l'activité *in vitro* des composés à partir de leur structure chimique à l'aide de différents modèles de *machine learning*. Les données étant très déséquilibrées en faveur des composés non toxiques, nous montrons que l'augmentation de données améliore les performances des modèles. Nous montrons aussi que la méthode ensembliste « *stacked generalization* » mène à des modèles fiables sur leur domaine d'applicabilité.

Ensuite, nous évaluons le lien entre les résultats de tests *in vitro* et les effets *in vivo* supposés leur être associés et montrons que ce lien n'est pas évident, soulevant la question de la pertinence des tests.

3. Thèse soutenue le 9 juillet 2019, préparée au sein du Laboratoire d'informatique, signaux et systèmes de Sophia Antipolis (I3S), sous la direction de Jean-Paul Comet (Université Côte d'Azur).