Proposition de sujet de thèse

Intitulé français du sujet de thèse proposé : Etude de processus stochastiques persistants et représentation de séquences biologiques

Intitulé en anglais : Persistent stochastic processes and biological sequences representation.

Unité de recherche : Institut Mathématiques de Bourgogne, UMR 5584

Nom, prénom et courriel du directeur (et co-encadrant) de thèse : CENAC Peggy, peggy.cenac@u-bourgogne.fr et OFFRET Yoann, yoann.offret@u-bourgogne.fr

Domaine scientifique principal de la thèse : Probabilités et Statistique (mathématiques appliquées)

Description du projet scientifique

Une source stochastique est un mécanisme aléatoire qui émet un symbole (appartenant à un alphabet) à chaque pas de temps. Une chaîne de Markov à longueur variable (acronymisée dans la suite en VLMC pour *Variable Length Markov Chain*) est un modèle de source stochastique qui est une manière naturelle d'étendre une chaîne de Markov. Ce modèle est un compromis intéressant entre les modèles simples de sources (sources sans mémoire, chaînes de Markov), d'étude simple mais pas assez réalistes, et les sources plus corrélées, dont l'étude théorique est beaucoup moins simple et ne mène pas à des résultats quantitatifs précis.

Le comportement des marches aléatoires dont les incréments sont des nombres émis par une source est fortement relié aux propriétés probabilistes de cette source. Quand la source est une chaîne de Markov, les propriétés de récurrence de la marche aléatoire sont bien connues. Il existe également des résultats dans le cas où la source possède des propriétés de renouvellement.

L'enjeu ici est d'étudier ces problématiques avec une source VLMC assez générale, menant ainsi à la famille des processus à mémoire longue. Quels sont les cas où la marche aléatoire vérifie une loi des grands nombres, un théorème de la limite centrale, un principe de grandes déviations ? Quel est l'impact de la forme de l'arbre des contextes de la VLMC ? De la régularité des mesures de probabilités attachées à chaque contexte ?

Dans de nombreux cas, des modèles à temps continu sont préférables, par exemple en économie, finance ou encore neurobiologie. Il est bien connu qu'un changement d'échelle judicieux sur une marche aléatoire classique mène au mouvement brownien standard, quand l'échelle de temps et l'échelle d'espace
tendent vers zéro. lorsque les incréments définissent une chaîne de Markov d'ordre un, le processus obtenu est appelé marche aléatoire persistante ou marche de Kac. La marche perd la propriété de Markov. L'enjeu est de généraliser ce type de convergence à des marches dont les incréments sont donnés par une VLMC.

La CGR (*Chaos Game Representation*) est une méthode de stockage et surtout d'une méthode de représentation graphique de séquences, appliquée pour la première fois aux séquences d'ADN par Jeffrey. La visualisation d'une séquence sous cette forme graphique permet de comparer des motifs, extraits de séquences, localement comme globalement. La CGR est un système dynamique qui, à une séquence de lettres dans un alphabet fini, fait correspondre une trajectoire dans un espace continu, voire une mesure empirique sur un ensemble. À partir de propriétés sur la mesure invariante des points de la CGR, il est possible de déterminer la structure de la séquence elle-même. Comment utiliser une telle mesure pour comparer deux séquences de façon pertinente ? Quel est le comportement typique d'une CGR construite à partir d'une séquence modélisée par une VLMC ? Comment caractériser et classifier des modèles de VLMC à partir de la CGR ?

Connaissances et compétences requises : master en mathématiques avec des connaissances approfondies en probabilités et statistique.