# Similarité par recouvrement de séquences pour la fouille de données séquentielles et textuelles

Pierre-Francois Marteau*, Nicolas Béchet*
Oussama Ahmia*,**

*IRISA, Université Bretagne Sud
Campus de Tohannic, 56000 Vannes, FRANCE
prenom.nom@irisa.fr,
https://www.irisa.fr/
**Octopus Mind, 2 Place Saint-Pierre,
Nantes, 44000, FRANCE
https://www.octopusmind.info/

**Résumé.** Nous introduisons la notion de similarité par recouvrement de séquences pour estimer la similarité entre une séquence et un ensemble de séquences. Nous en dérivons une pseudo-distance qui s'apparente aux distances d'édition de type Levenshtein pour comparer des paires de séquences. La complexité algorithmique associée à cette semi-métrique peut-être ramenée à $O(n \cdot log(n))$ en utilisant des arbres de suffixes. Nous introduisons un nouveau modèle discriminant dédié à la classification de données textuelles dont la complexité algorithmique ne dépend pas de la taille de l'ensemble d'apprentissage, mais uniquement du nombre de classes et de la longueur des séquences. L'étude expérimentale préliminaire présentée s'appuie sur deux benchmaks : le premier concerne des séquences de nucléotides, le second une tâche de classification de textes. Les résultats obtenus positionnent l'approche proposée au niveau de l'état de l'art (incluant les approches "deep learning") sur les tâches considérées., avec des temps de calcul et un nombre de méta-paramètres avantageux.

## 1 Introduction

Estimer de manière efficace la similarité entre des séquences symboliques est une tâche récurrente dans de nombreux domaines d'application, en particulier en bio-informatique, traitement des textes ou encore dans les domaines de la sécurité et sûreté des systèmes cyber-physiques. De nombreuses mesures de similarité ont été définies pour estimer la similarité entre deux séquences symboliques, comme la distance d'édition (Levenshtein, 1966) et son implémentation proposée par Wagner et Fisher (Wagner et Fischer, 1974), BLAST (Korf et al., 2003), les distances de Smith et Waterman (Smith et Waterman, 1981), de Needleman et Wunch (Needleman et Wunsch, 1970) ou les noyaux séquentiels locaux (Vert et al., 2004).

Dépasser le modèle de sac de mots pour tenir compte de la séquentialité des données textuelles est un problème difficile en général. Nous présentons dans cet article une nouvelle approche pour caractériser la similarité entre séquences symboliques en introduisant la notion