Construction d'attributs pour l'extraction de connaissances à partir de séquences biologiques

M. Maddouri[*] et F. Mhamdi [**]
Unité de Recherche en Programmation, Algorithmique et Heuristique

[*] Institut National des Sciences Appliquées et de Technologie,
Université 7 Novembre à Carthage, Centre Urbain Nord,
BP. 676, 1080 Tunis, Tunisie
mondher.maddouri@fst.rnu.tn

[**] Institut Supérieur des Langues Appliquées et d'Informatique de Béja,
Université de Jendouba, Av. Habib Bourguiba, 9000, Béja, Tunisie
faouzi.mhamdi@ensi.rnu.tn

Résumé. Dans cet article nous étudions un problème de prétraitement de données : la construction d'attributs décrivant des séquences biologiques. Afin d'assurer l'extraction de connaissances à partir de séquences biologiques (ADN, ARN et protéines), tout système de fouille de données (datamining) se confronte à la représentation non habituelle de ce type de données. Une séquence biologique est représentée, en structure primaire, par une chaîne de caractères. La construction d'attributs décrivant les séquences biologiques est une étape de prétraitement inévitable. Dans cet article, nous étudions les méthodes existantes de construction d'attributs décrivant des séquences biologiques, notamment, celles qui se basent sur les n-grammes, l'arbre de suffixes généralisés et les modèles de Markov cachés. Notre contribution dans cet axe a été la proposition de la méthode des descripteurs discriminants et la présentation d'une étude comparative approfondie de ces méthodes en les appliquant à des problèmes biologiques typiques comme la reconnaissance de sites promoteurs des gènes de *E. Coli*, la reconnaissance de sites de jonction de *Primate* et la classification des protéines. Une confrontation des résultats de chaque méthode avec la banque de motifs Pfam sera aussi présentée.

1 Introduction

La plupart des méthodes de datamining, traitent des données représentées sous forme d'une table relationnelle (tableau attributs/valeurs). Il existe toutefois quelques travaux qui portent sur une représentation de données plus complexes [Cornuéjols *et al.* 2002, Zighed *et al.* 2000]. Une problématique supplémentaire s'ajoute lorsqu'on veut utiliser ces approches pour analyser des données ayant des représentations atypiques : séquences, hiérarchies, image, son, vidéo, etc [Mitra *et al.* 2003]. La *construction d'attributs*, qui consiste à inventer des attributs pour décrire ces données en format relationnelle (attributs/valeurs), permet d'apporter une réponse au problème de données atypiques [Liu *et al.* 1998, Liu *et al.* 2001]. Elle peut être vue comme l'une des tâches de prétraitement dans le procédé de l'Extraction de connaissances à partir de données [Fayyad *et al.* 1996].