Classification supervisée de séquences biologiques, basée sur les motifs et les matrices de substitution

Rabie Saidi*, Mondher Maddouri**
Engelbert Mephu Nguifo***

*FSJEG, Université de Jendouba, rue de l'UMA, Tunisia
saidi@cril.univ-artois.fr
** Institut National des Sciences Appliquées and Technologies (INSAT),
Tunis-Carthage 2035, Tunisia
mondher.maddouri@fsegt.rnu.tn
*** CRIL – CNRS, Université d'Artois - IUT de Lens, France
mephu@cril.univ-artois.fr

Résumé. La classification des séquences biologiques est l'un des importants défis ouverts dans la bioinformatique, tant pour les séquences protéiques que pour les séquences nucléiques. Cependant, la présence de ces données sous la forme de chaînes de caractères ne permet pas de les traiter par les outils standards de classification supervisée, qui utilisent souvent le format relationnel. Pour remédier à ce problème de codage, plusieurs travaux se sont basés sur l'extraction des motifs pour construire une nouvelle représentation des séquences biologiques sous la forme d'un tableau binaire. Nous décrivons une nouvelle approche qui étend les méthodes précédentes par l'utilisation de matrices de substitution dans le cas des séquences protéiques. Nous présentons ensuite une étude comparative qui prend en compte l'effet de chaque méthode sur la précision de la classification mais aussi le nombre d'attributs générés et le temps de calcul.

1 Introduction

L'émergence de la bioinformatique, que nous témoignons durant les dernières années, trouve ses causes dans les progrès technologiques qui ont permis de conduire des projets de recherche à grande échelle. Le plus remarquable était le Projet du Génome Humain (PGH) [National Human Genome Research Institute, 2006] accompli en 13 ans depuis 1990 ; période qui s'avère très courte comparée avec la quantité de données extraites sur le génome humain : 3 milliards de bases qui constituent l'ADN humain. Ainsi, plusieurs problèmes sont ouverts :

- Comment le gène exprime-t-il sa protéine ?
- Où commence le gène et où finit-il ?
- Comment évoluent les familles de protéines et comment les classer ?
- Comment prédire la structure tridimensionnelle des protéines ?
- ….

Dans ce contexte, le besoin en fouille de données se fait de plus en plus pressant. Cependant, les techniques de fouille de données, qui traitent souvent des données sous le format relationnel, se trouvent confrontées au format inapproprié des séquences biologiques qui se