Classification multi-étiquettes pour l'alignement multiple de séquences protéiques

Lina Fahed*, Gabriel Frey*, Julie Dawn Thompson**, Nicolas Lachiche*

*LSIIT, Université de Strasbourg, Pôle API, Bd Brant, 67400 Illkirch
{linafahed@gmail.com, g.frey@unistra.fr, nicolas.lachiche@unistra.fr}
**IGBMC, Université de Strasbourg, 1 rue Laurent Fries, 67404 Illkirch
{julie@igbmc.fr}

Résumé. Cet article présente une application de classification multi-étiquettes permettant de déterminer le programme à utiliser pour construire un alignement multiple d'un ensemble de séquences protéiques donné. Dans un premier temps, nous avons réussi à améliorer le système existant, Alexsys en ajoutant des attributs. Dans un second temps, nous déterminons pour un ensemble de séquences protéiques donné le ou les aligneurs capable de produire les alignements de meilleur score, à epsilon près. Les mesures de performances propres à la classification multi-étiquette nous permettent d'analyser l'influence de epsilon et de choisir une valeur assez petite pour distinguer les meilleurs aligneurs des autres.

1 Introduction

De nombreux programmes de construction d'alignements multiples à partir d'un ensemble de séquences protéiques (appelés aligneurs) ont été développés. Cependant, il n'existe pas un aligneur capable de bien aligner tous les types de séquences. C'est la raison pour laquelle le laboratoire de bioinformatique de l'IGBMC de Strasbourg a développé un système expert pour l'alignement multiple de séquences protéiques appelé Alexsys (Aniba et al., 2009). Le système Alexsys propose d'identifier, pour un ensemble de séquences protéiques données, les aligneurs permettant d'obtenir de bons alignements en se basant sur des techniques de fouilles de données (classification supervisée) exploitant des caractéristiques des séquences.

2 Le contexte

Un grand nombre de travaux en apprentissage supervisé concerne la prédiction d'une étiquette unique, où une seule étiquette parmi plusieurs étiquettes disjointes disponibles est associée à chaque instance. Cependant, il existe des applications dans lesquelles un ensemble d'étiquettes est associé à chaque instance. Ces données sont appelées données multi-étiquettes. La classification multi-étiquettes concerne l'apprentissage d'une fonction qui associe à chaque instance le sous-ensemble des étiquettes approprié. C'est une forme d'apprentissage supervisé où plusieurs classes booléennes sont associées à chaque instance. Des mesures de performance