Évaluation des méthodes supervisées pour la discrimination de protéines

Ricco Rakotomalala*, Faouzi Mhamdi**

*Laboratoire ERIC – Université Lyon 2
69500 BRON
ricco.rakotomalala@univ-lyon2.fr,
http://eric.univ-lyon2.fr/ ricco
**URPAH – Université d'El Manar
TUNISIE
faouzi.mhamdi@ensi.rnu.tn

Résumé. Nous évaluons différentes méthodes supervisées dans le cadre de la discrimination de protéines. Les descripteurs étant automatiquement générés, nous avons utilisé les n-grammes, la taille de l'espace de représentation est très élevée par rapport au nombre d'observations. Un grand nombre de descripteurs ne sont pas pertinents. Il apparaît que les méthodes linéaires telles que les SVM linéaires ou la régression PLS sont les plus performantes. Une étude détaillée montre que ce succès repose essentiellement sur la robustesse face à la dimensionalité qui devient, de fait, le critère le plus important dès lors que l'on traite des domaines où les descripteurs sont générés automatiquement en grand nombre. Dans ce contexte, une procédure de sélection de variables, fusse-t-elle très fruste, modifie significativement le comportement des algorithmes d'apprentissage.

1 Introduction

L'annotation et le classement de protéines est une activité importante du biologiste. L'augmentation du volume de données à traiter rend nécessaire l'automatisation de cette tâche. Ces dernières années, la fouille de données, plus généralement l'extraction de connaissances à partir de données (Fayyad et al., 1996), a permis de dégager un cadre qui rend reproductible la démarche complète de classement automatique des protéines à partir de leur structure primaire. En effet, une séquence de protéine est décrite par une suite de caractères pris dans un alphabet de 20 signes. Le rapprochement avec les nombreux travaux réalisés dans la catégorisation de textes est naturelle (Sebastiani, 2005). Par rapport à un traitement standard sur des données individus-variables, l'appréhension des données non-structurées introduit deux étapes supplémentaires : l'extraction de descripteurs à partir de la description primaire pour aboutir à un tableau de données et, éventuellement, la sélection des descripteurs les plus discriminants, afin que les algorithmes d'apprentissage puissent fonctionner de manière efficace. Compte tenu du nombre important de descripteurs que nous pouvons générer, la complexité informatique est un critère primordial dans cette démarche.