Détection de faibles homologies de protéines par machines à vecteurs de support

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jmikolaj@nantes.inserm.fr, yjacques@nantes.inserm.fr
**LINA, équipe EGC, Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Cet article décrit une approche discriminative pour la recherche de nouveaux membres dans des familles de protéines à faibles homologies de séquences. L'originalité de la méthode repose sur une modélisation de ces familles par un ensemble M de motifs intégrant les propriétés physicochimiques des résidus. Nous proposons un algorithme de découverte de motifs suivant le paradigme de la classification hiérarchique ascendante. L'ensemble M définit un espace de représentation des séquences : chaque séquence est transformée en un vecteur indiquant la présence ou l'absence de chaque motif appartenant à M. Nous utilisons la technique d'apprentissage par machine à vecteurs de support (SVM) pour discriminer la famille d'intérêt vis-à-vis des séquences non apparentées. Cette méthode est testée sur la famille biologique des interleukines dont les membres possèdent des homologies de séquences faibles en dépit d'un repliement tridimensionnel en hélices alpha très conservé. Nous montrons que l'ensemble des motifs hiérarchiques modélise spécifiquement les interleukines par rapport aux autres familles structurales de la base de données SCOP (1.51). Notre classifieur est en effet plus performant sur notre famille de protéines que d'autres méthodes de classification dont le SVM basé sur les spectres de chaîne.

1 Introduction

La découverte de nouveaux membres d'une famille de protéines repose sur deux types de techniques. La plus courante est basée sur une mesure d'homologie de la protéine candidate avec un motif spécifique caractéristique de la famille d'intérêt. Cette méthode consiste à fouiller le génome à partir d'outils bioinformatiques tels que BLAST [Altschul *et al.*, 1990]. Certaines familles de protéine sont trop hétérogènes pour qu'on puisse retrouver des régions conservées au niveau de leur structure primaire. Pour lever cette difficulté, une démarche alternative a été suggérée par plusieurs auteurs [Jaakola *et al.*, 2000]. Elle est fondée sur des méthodes d'apprentissage dans lesquelles les séquences de protéines sont étiquetées selon leur appartenance ou non à la famille recherchée. Les exemples positifs (étiquette +1) regroupent les membres connus de la