Caractérisation de signatures complexes dans des familles de protéines distantes

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jerome.mikolajczak@nantes.inserm.fr, yjacques@nantes.inserm.fr
**IRIN, Equipe C.I.D. Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Résumé. L'identification de signatures de protéines est un problème majeur pour la découverte de nouveaux membres dans des familles de protéines connues. Le concept de signature qui permet de caractériser ces familles est généralement basé sur la définition de motifs communs. Il s'avère que les familles distantes sont trop hétérogènes pour qu'on puisse identifier des régions conservées à partir des algorithmes classiques de la bioinformatique. Nous proposons une approche génétique pour la découverte de motifs hiérarchiques ; l'algorithme suit une démarche descendante en s'appuyant dans une première phase sur les classes physico-chimiques des acides aminés. Les signatures sont ensuite définies par des séquences des motifs ainsi obtenus. Elles sont extraites au moyen d'un algorithme de découverte d'itemsets séquentiels où les motifs jouent le rôle d'items. Une dernière étape consiste à fouiller dans cette base d'itemsets pour n'en retenir qu'un ensemble réduit de signatures. Plusieurs stratégies sont proposées pour déterminer un ensemble optimal de signatures qui respecte des contraintes de complétude, de cardinalité et de spécificité. Nous appliquons notre démarche sur la famille des cytokines. L'analyse de la base de protéines SCOP a montré que le groupe de signatures que nous avons extrait cible spécifiquement cette famille d'intérêt.

1 Introduction

Les protéines qui constituent les briques élémentaires du vivant se regroupent par familles ayant des propriétés ou fonctions similaires. Ces molécules ont évolué dans le temps à partir d'ancêtres communs, ce qui explique la présence de motifs semblables dans les différents membres d'une même famille. Ces motifs sont des régions bien conservées au sein de la structure primaire des séquences biologiques. La structure primaire d'une protéine est représentée par une séquence $s = \langle s_1 s_2 \dots s_n \rangle$ où chaque $s_i \in \Omega$, l'ensemble des acides aminés : $\Omega = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$. Plusieurs définitions du motif ont été proposées ; la syntaxe PROSITE est la plus connue [Bucher et Bairoch, 1994]. Le motif $W[ILV]Y$ y désigne une sous-séquence constituée d'un W, suivie immédiatement par un I, L ou un V, et terminée par un Y. Le symbole