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

$x(i,j)$ est également utilisé pour autoriser un intervalle de taille minimale i et maximale j séparant les acides aminés du motif. Par exemple, le motif $Wx(1,3)Y$ se retrouve aussi bien dans la chaîne WFY que dans la chaîne $WAEY$.
La présence de régions conservées est d'une grande utilité pour la découverte de nouveaux membres d'une famille connue de protéines [Servant *et al.*, 2002]. La recherche de motifs biologiques est un problème majeur en bioinformatique [Dsouza *et al.*, 1997, Ramstein *et al.*, 2000, Sadowski et Parish, 2003]. On utilise le terme de signature pour désigner un motif ou un ensemble de motifs permettant de caractériser un ensemble $\mathcal{S}$ de protéines. Une signature est d'autant plus spécifique qu'elle exclut tout autre protéine. La connaissance d'une signature fournit donc une clé pour l'identification de protéines dont les fonctions sont encore inconnues.

Malheureusement, certaines familles sont si distantes qu'il est impossible d'en extraire une signature en utilisant les logiciels de découverte de motifs biologiques, tels que MEME [Bailey et Elkan, 1994] ou PRATT [Brazma *et al.*, 1996]. Notre processus de découverte procède en trois étapes. Dans la première, nous étendons l'alphabet qui compose le motif par des classes d'acides aminés. Cette description hierarchique permet l'obtention d'un ensemble de motifs. Comme la spécificité de ces derniers est insuffisante pour un criblage génomique, nous recherchons dans une deuxième étape des signatures définies par des séquences de motifs. Nous obtenons ainsi un grand nombre de signatures candidates ciblant chacune un sous-ensemble particulier de la famille d'intérêt $\mathcal{S}$. La troisième étape consiste donc à ne retenir qu'un ensemble réduit de signatures qui recouvre la totalité des membres de $\mathcal{S}$. Après avoir exposé les trois phases de notre méthode, nous terminerons par une application concernant la superfamille des cytokines.

2 Définition des signatures

Nos algorithmes reposent essentiellement sur une hierarchisation des acides aminés. Nous allons dans un premier temps décrire l'alphabet étendu que nous utilisons, avant d'aborder les concepts de motifs et de signatures.

L'ensemble Ω est subdivisé en sous-ensembles non disjoints associés à des propriétés physico-chimiques particulières. Plusieurs variantes de systèmes de classes ont été proposées ; nous avons opté pour celui présenté en table 1 [Taylor, 1986]. La pertinence de cette classification se vérifie par l'étude des régions conservées : on observe que les mutations s'opèrent généralement au sein d'une même classe (par exemple, les acides aminés I, L et V appartenant à la classe aliphatique sont très fréquemment interchangés). Soit Γ^1 le super-ensemble formé par les classes définies dans la table 1 : $\Gamma^1 = \{\Omega,\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta,\theta\}$. Nous appelerons $\Gamma^2 = \Gamma^1 \cup \omega$ le super-ensemble qui contient Γ^1 et le super-ensemble ω des singletons de Ω : $\omega = \{\{A\},\{C\},\ldots,\{Y\}\}$.

Un motif $m = \langle m_1 m_2 \ldots m_k \rangle$ est une k-séquence formée d'ensembles $m_i \in \Gamma^2$. Nous appelerons *occurrence* d'un motif m une sous-séquence $\langle s_{i+1} s_{i+2} \ldots s_{i+k} \rangle$ de s telle que $s_{i+j} \in m_j \ \forall j,\ 1 \leq j \leq k$. On dira que la séquence s *vérifie* le motif m. Le *support* d'un motif m dans $\mathcal{S}$ est le nombre de séquences qui vérifie m. La séquence MH vérifie

Symbole	Classe	Membres
α	aliphatique	ILV
β	aromatique	$FHWY$
γ	non polaire	$ACFGHIKLMVWY$
δ	chargé	$DEHKR$
ε	polaire	$CDEHKNQRSTWY$
ζ	charge positive	HKR
η	chaîne latérale courte	$ACDGNPSTV$
θ	chaîne latérale très courte	$ACGST$

TAB. 1 – *Classes d'acides aminés basées sur des propriétés physico-chimiques*

ainsi 21 motifs de taille 2, dont les motifs MH, $M\beta$, $\gamma\delta$, et $\Omega\Omega$. Le motif MH ne peut être vérifié que par une seule sous-séquence, tandis que le motif $\Omega\Omega$ est vérifié pour n'importe quelle séquence de taille supérieure ou égal à 2. Il importe donc de qualifier la spécificité d'un motif en prenant en compte la probabilité de le voir apparaître dans une séquence. L'estimation de cette probabilité étant coûteuse en temps de calcul, nous avons opté pour la fonction de coût suivante : $c(m) = \prod_{i=1}^{k} f(m_i)$ où $f(m_i)$ est la fréquence de la classe m_i dans une base d'apprentissage comprenant de nombreuses familles de protéines différentes. Dans la pratique, on observe une bonne corrélation entre l'estimation $c(m)$ et le support effectif de m dans la base d'apprentissage. La spécificité d'un motif m sera définie par $\phi(m) = -\log(c(m))$. La table 2 indique quelques exemples de motifs ainsi que leur spécificité et leur support.

motif	support	spécificité	fréquence
$\beta\alpha\varepsilon$	1.00	4.4	0.7984
$\alpha\delta\varepsilon\alpha$	0.96	5.1	0.6095
$\delta\Omega\gamma\varepsilon\Omega\alpha\zeta\varepsilon$	0.65	6.8	0.2427
$\Omega\alpha\zeta\varepsilon\alpha\varepsilon\varepsilon\gamma$	0.54	7.6	0.1130
LEE	0.28	7.9	0.0893
$\beta\gamma\Omega\Omega\gamma\zeta\varepsilon L$	0.28	8.4	0.0433
$\varepsilon\gamma\alpha\zeta\delta L\Omega\varepsilon$	0.37	9.2	0.0359
$L\varepsilon\varepsilon\gamma\alpha\varepsilon\delta L$	0.22	10.2	0.0107
$\Omega\eta L\alpha L\alpha\Omega L$	0.15	11.0	0.0019
$F\varepsilon R\gamma K\varepsilon\Omega\gamma$	0.15	11.4	0.0018

TAB. 2 – *Exemples de motifs avec leur support dans la famille des cytokines, leur spécificité estimée et leur fréquence effective dans la base SCOP*

Notre définition de motifs à partir de classes prédéfinies nous permet de diminuer considérablement l'espace de recherche. Notons qu'on peut toujours transformer un motif hiérarchique sous la forme PROSITE. Du motif $M\beta P$ et des sous-séquences MHP et MYP, on en déduit aisément le motif PROSITE $M[HY]P$.

Une n-signature σ est une séquence de n motifs $\langle m^1 m^2 \dots m^n \rangle$. Une séquence s vérifiant σ contient au moins n sous-séquences $(s_1, s_2 \dots, s_n)$ où chaque s^i vérifie le

motif m^i correspondant. Les positions respectives (p_1,p_2,...,p_n) de ces sous-séquences dans s sont telles que $p_{i+1} - p_i \geq |m_i|$ $\forall i$, $1 \leq i < n$, où $|m_i|$ désigne la taille du motif. La spécificité d'une signature sera définie par la somme des spécificités de ces motifs :

$$\phi(\sigma) = \sum_{i=1}^{k} \phi(m_i)$$

3 un algorithme génétique de découverte de motifs

Le caractère fortement combinatoire de notre problème nous a amené à utiliser les algorithmes génétiques (AG). De nombreux algorithmes de bioinformatique utilise cette métaheuristique [Sadowski et Parish, 2003]. La description hiérarchique permet une recherche selon une approche descendante qui part du motif le plus général vers le plus spécifique. Dans le cas le plus pessimiste, il existe dans toute famille $\mathcal{S}$ comprenant des k-séquences au moins un k-motif $\mu(k) = \langle m_1 m_2 \dots m_k \rangle$ avec $m_i = \Omega$ $\forall i$, $1 \leq i \leq k$. Ce motif trivial est le motif de plus basse spécificité qu'on puisse trouver ($\phi(\mu(k)) = 0$). L'algorithme decouverteMotifs que nous avons développé procède en deux phases distinctes. La première vise à découvrir N k-motifs décrits par les classes de Γ^1. Ces motifs généraux sont repris un par un dans une deuxième phase pour rechercher des motifs contenant d'éventuels acides aminés (motifs décrits par Γ^2). Un même AG, dénommé AGmotifs, est utilisé dans les deux phases.

fonction decouverteMotifs($\mathcal{S}$, N, k) **retourne** Population
entrées
$\mathcal{S}$, l'ensemble de séquences dont on recherche les motifs
N, le nombre de motifs
k, la taille des motifs

$p1$, $p2$, p : Population ;
/* recherche des motifs de la phase 1 */
$p1 \leftarrow$ AGmotifs(N,$\mu(k)$,Γ^1,$\mathcal{S}$);
Pour tout motif m dans $p1$ faire
    /* recherche des motifs de la phase 2 */
    $p2 \leftarrow$ AGmotifs(N,m,Γ^2,$\mathcal{S}$);
    /* extraire de $p2$ l'individu le plus apte et l'inclure dans p */
    $p \leftarrow p \cup \{$meilleurIndividu($p2$)$\}$;
fait ;
Retourner p;

Cette démarche descendante présente l'avantage de restreindre l'espace de recherche et d'éviter de converger vers des solutions sous-optimales induites par la rareté des motifs incluant des singletons.
AGmotifs possède deux particularités essentielles. La première est que la population initiale est formée par des individus ou motifs obtenus par copie d'un motif germe. Dans

la phase 1, le germe est le motif générique $\mu(k)$ tandis que dans la phase 2, le germe est un des N motifs extraits dans la phase précédente. La deuxième caractéristique est que l'opérateur de mutation est orienté, selon le principe de la recherche descendante : une classe Ω sera mutée en une classe physico-chimique dans les phases 1 et 2, une classe physico-chimique c sera mutée en un des membres de c en phase 2 (par exemple, α sera muté aléatoirement par la pseudo-classe $\{I\}$, $\{L\}$ ou $\{V\}$).

fonction AGmotifs(N, *germe*, *alphabet*, $\mathcal{S}$) **retourne** Population
entrées
N, le nombre de motifs recherchés
germe, le motif prototype pour l'initialisation de la première génération
alphabet, la description du motif selon Γ^1 ou Γ^2
$\mathcal{S}$, l'ensemble de séquences

parents, *population* : Population;
/* initialisation de la population */
Pour $i = 1$ à N faire $population \leftarrow population \cup \{\text{clone}(germe)\}$;
Faire
calculeAdaptation(*population*, $\mathcal{S}$);
$parents \leftarrow$ selection(*population*) ;
$population \leftarrow$ reproduction(*parents*,*alphabet*) ;
jusqu'à adaptation(*population*) $\geq$ *seuil* ;
retourner *population*;

Reproduction

Deux opérateurs sont utilisés pour faire évoluer la population ; celui de la mutation que nous avons déjà décrit et celui de la recombinaison qui opère selon la technique dite de *crossing-over*. Le passage d'une génération à l'autre suit la méthode appelée *keep-best reproduction*(KBR). Cette méthode repose sur l'idée que les fils peuvent être moins bien adaptés que leurs parents et qu'il serait désastreux de les remplacer. Une solution intermédiaire consiste à supprimer le fils le moins adapté et de garder le meilleur parent. KBR assure que la nouvelle génération se voit enrichie d'un pool génétique nouveau tout en préservant le matériel génétique performant de l'ancienne génération. Selon leurs auteurs, KBR permet de trouver plus rapidement de meilleurs solutions [Wiese et Goodwin, 1999].

Fonction d'adaptation

La sélection est basée sur la technique de la roue de la fortune, principe selon lequel l'individu survit d'autant plus sûrement qu'il est bien adapté. La définition de la fonction d'adaptation est un critère déterminant dans tout AG. Dans notre application, un motif intéressant est un motif ayant une grande spécificité et un support important. Les deux critères s'opposent : un motif de faible spécificité est largement représenté (le motif $\mu(k)$ a un support égal à $C = |\mathcal{S}|$, le cardinal de la famille considérée) et inversement un motif très spécifique a peu de chances d'être identiquement conservé dans la

famille, comme le montrent les exemples de la table 2. La fonction d'adaptation qui a donné les meilleurs résultats sur notre jeu de données est la suivante :

$$a(m) = \begin{cases} 0 & \text{si } m \text{ tel que } support(m) < \tau, \\ 1 & \text{si } m = \mu(k), \\ support(m) * \phi(m)^{\lambda} & \text{sinon.} \end{cases}$$

Notre fonction d'adaptation $a(m)$ est proportionnelle au support de m et s'accroît avec la spécificité du motif. Le paramètre λ permet de moduler l'importance de la spécificité par rapport au support. Nous avons introduit un support minimal τ en deçà duquel le motif est jugé inadapté. Ce paramètre évite de retrouver des motifs de haute spécificité mais de support dérisoire (le cas limite étant évidemment celui du support 1 où toute k-sous-séquence de $\mathcal{S}$ serait retenue). Notons en effet que si $a(m) = 0$, le motif m disparaîtra à la génération suivante, puisque la sélection des individus est opérée avec une probabilité $a(m)$. C'est la raison pour laquelle nous avons attribué au motif $\mu(k)$ une adaptation non nulle, afin qu'il survive au delà de la première génération.

Redondances des motifs découverts
Nous avons légèrement modifié AGmotifs pour éviter les doublons ; les motifs délivrés par l'AG ne sont pas les individus de la dernière génération, mais les N meilleurs individus distincts obtenus, toutes générations confondues. Ce filtre trivial n'évite pas des redondances de motifs liées au recouvrement de certaines classes de Γ^1. Certains motifs sont en effet en relation de généralisation/spécialisation (comme les motifs $\alpha\Omega\gamma$ et $V\varepsilon\beta$) ou contiennent partiellement des classes qui participent de cette relation (comme les motifs $\alpha\varepsilon\beta$ et $V\Omega\gamma$). Il n'est pas possible de filtrer directement ces redondances potentielles, sans passer par une analyse précise de chacune de leurs occurrences dans les séquences. Cette étape serait fastidieuse et, d'après nos tests, non désirable. En effet, ces redondances ne concernent qu'une part infime des motifs, essentiellement des motifs de petite taille. D'autre part, on peut remarquer que ces motifs seront éliminés dans les deux phases suivantes : dans la phase de recherche des signatures (par la contrainte de non recouvrement liée à la définition même des signatures) et dans la phase de réduction de l'ensemble des signatures (par l'élimination des signatures de moindre spécificité).

4 Extraction d'un ensemble optimal de signatures

La découverte d'un ensemble de signatures ciblant une famille de séquences procède en deux étapes :

1. L'extraction d'un ensemble de signatures Σ à partir de l'ensemble des motifs $\mathcal{M}$ obtenus par l'algorithme décrit précédemment ;
2. La réduction de Σ pour n'en retenir que les membres les plus représentatifs en terme de support et de spécificité.

La première étape d'extraction de signatures s'apparente à la recherche d'itemsets séquentiels fréquents dans des séquences. Dans le champ d'application qui nous

préoccupe, le nombre forcément restreint de motifs rencontrés diminue les risques d'explosion combinatoire que nous avions rencontré dans la recherche de motifs. Nous avons adapté l'algorithme Winepi [Manilla et Toivonen, 1996] pour déterminer un ensemble de signatures candidates de support minimal τ. L'algorithme, inspiré d'AprioriAll [Agrawal et Srikant, 1995], explore tous les itemsets fréquents, en partant de ceux de taille $k = 1$ (les motifs eux-mêmes), puis par jointure, recherche itérativement ceux de taille $k = 2, 3, \ldots$ jusqu'à ce qu'il n'existe plus de signatures de support supérieur ou égal à τ. A chaque étape, nous obtenons ainsi un ensemble Σ^k de k-signatures que nous mémorisons, jusqu'à l'arrêt de l'algorithme, marqué par $\Sigma^{kmax} = \emptyset$. Nous appellerons ExtraireSignaturesCandidates($\mathcal{S}$,$\mathcal{M}$) l'algorithme qui délivre à partir de $\mathcal{S}$ et de $\mathcal{M}$ un ensemble de signatures $\Sigma = \bigcup_{k=1}^{kmax-1} \Sigma^k$.

La deuxième étape consiste à modéliser l'ensemble $\mathcal{S}$ par un sous-ensemble Σ' de Σ. Cette phase permet de représenter une famille de séquences connues sous une forme condensée. Ce modèle servira à classer des séquences inconnues dans cette famille ou à les rejeter. L'ensemble Σ' doit satisfaire les contraintes suivantes :

1. *contrainte de complétude* :
   chaque séquence $s_i \in \mathcal{S}$ vérifie au moins une signature de Σ'. Le support de Σ' doit être égal au cardinal C de la famille $\mathcal{S}$;
2. *contrainte de cardinalité* :
   le cardinal de Σ' doit être le plus petit possible (idéalement de 1) ;
3. *contrainte de spécificité* :
   la spécificité de Σ', définie comme la somme des spécificités de ces éléments, doit être la plus grande possible ;

On constatera aisément que l'algorithme ExtraireSignaturesCandidates fournit une solution de grande cardinalité qui ne garantit pas la vérification de la première condition (sauf à fixer le support minimal τ à C, ce qui induirait des signatures de faible spécificité).

Dans la pratique, on supposera que la contrainte de complétude est vérifiée par Σ (si tel n'est pas le cas, il suffit de compléter Σ par la signature $\langle \mu(k) \rangle$). Soit $\Sigma(\varphi)$ l'ensemble des signatures de Σ de spécificité supérieure ou égale à φ. Un premier prétraitement de Σ consiste à rechercher la spécificité maximale φ_{max} telle que le support de $\Sigma(\varphi_{max})$ soit égal à C, puis à éliminer de cet ensemble les signatures redondantes et de moindre spécificité.

procédure eliminerRedondances(Σ,φ)
  entrées
  Σ, l'ensemble des signatures candidates
  φ, le seuil de spécificité maximale autorisé
  $\Sigma \leftarrow \Sigma(\varphi)$;
  Pour toute signature $\sigma \in \Sigma$ faire
    Soit $E(\sigma) = \{s_i \in \mathcal{S}$ vérifiant $\sigma\}$;
    Rechercher dans Σ s'il existe une signature $\sigma' \neq \sigma$
    telle que $E(\sigma) = E(\sigma')$ et $\varphi(\sigma') > \varphi(\sigma)$. Si oui, alors $\Sigma \leftarrow \Sigma - \{\sigma\}$;
  fait ;

Malgré ce filtrage, l'exploration exhaustive des sous-ensembles de Σ peut demeurer problématique. Nous proposons une stratégie permettant de rechercher rapidement une solution potentiellement satisfaisante. La méthode rechercheParCardinal est une heuristique qui vise à retenir les signatures qui "couvrent" au mieux $\mathcal{S}$. Elle consiste à prendre la signature de support maximal, puis à retenir la signature dont l'union avec la signature précédente ait un support maximal, et ainsi de suite jusqu'à ce que l'ensemble Σ_{rc} ainsi formé ait un support égal à C.

fonction rechercheParCardinal(Σ, $\mathcal{S}$) **retourne** ensembleSignatures
entrées
Σ, l'ensemble des signatures candidates
$\mathcal{S}$, l'ensemble de séquences

$\Sigma_{rc} \leftarrow \emptyset$;
Tant que support(Σ_{rc}) $\neq C$ faire
Soit $E = \{\Sigma_{rc} \cup \{\sigma_i\}, \sigma_i \in \Sigma\}$;
Soit $E' = \{e \in E$ tel que le support de e soit maximal dans $\mathcal{S}\}$;
Soit $e_{max} \in E'$ l'ensemble de signatures de spécificité maximale;
Soit σ_{max} la signature telle que $e_{max} = \{\Sigma_{rc} \cup \{\sigma_{max}\}\}$;
$\Sigma_{rc} \leftarrow \Sigma_{rc} \cup \{\sigma_{max}\}$;
$\Sigma \leftarrow \Sigma - \{\sigma_{max}\}$;
fait ;
retourner Σ_{rc};

Notons que si cette heuristique privilégie la contrainte de cardinalité, elle ne donne pas nécessairement le plus petit ensemble de signatures qui vérifie $\mathcal{S}$. Si la valeur du cardinal $cmax = |\Sigma_{rc}|$ est élevée, nous proposons de faire appel à un AG comparable à celui que nous avons utilisé dans la section 3. Nous allons rechercher des ensembles Σ' de cardinalités k décroissantes, à compter de $k = cmax$. L'algorithme s'arrête lorsqu'on ne trouve plus d'individus dont le support est supérieur ou égal à $\tilde{C}$ ($\tilde{C} = C$ pour assurer la contrainte de complétude). A l'initialisation, les individus de la première génération sont des copies du germe Σ_{rc}. Lorsqu'un individu est jugé suffisamment apte, l'AG poursuit sa recherche en décrémentant k d'une unité. Cette procédure est réalisée au moyen de la méthode *optimise* qui élimine dans chaque individu de la population une des k signatures qu'il contient. La signature sélectionnée est celle qui assure au nouvel individu ainsi formé un support maximal dans $\mathcal{S}$.

fonction AGsignatures(N,Σ,Σ_{rc}, $\tilde{C}$) **retourne** ensembleSignatures
entrées
N, la taille de la population
Σ, l'ensemble des signatures candidates
Σ_{rc}, l'ensemble des signatures obtenues par la méthode rechercheParCardinal
$\tilde{C}$, le plus petit support recherché

parents, *population*, *resultat* : Population;
/* initialisation de la population */

Pour $i = 1$ à N faire $population \leftarrow population \cup \{\text{clone}(\Sigma_{rc})\}$;
$k \leftarrow |\Sigma_{rc}|$; /* initialisation à $cmax$ */
Faire
    Faire
        calculeAdaptation($population$);
        $parents \leftarrow$ selection($population$) ;
        $population \leftarrow$ reproduction($parents$) ;
    jusqu'à adaptation($population$) $\geq seuil$;

    Soit ind, l'individu le plus adapté dans $population$
    si le support(ind) $\geq \tilde{C}$ alors $resultat \leftarrow resultat \cup \{ind\}$;
    /* initialisation de la première génération suivante */
    $population \leftarrow$ optimise($population$);
    $k \leftarrow k - 1$;
tantque support(ind) $\geq \tilde{C}$;
retourner $resultat$;

L'opérateur de mutation remplace aléatoirement une des signatures d'un individu par une signature aléatoire de Σ qui n'appartient pas à cet individu. La recombinaison utilise une technique de *crossing-over*. La fonction d'adaptation que nous avons utilisée est la suivante :

$$a(individu) = \begin{cases} support(individu) & \text{si } support(individu) < \tilde{C}, \\ \lambda\phi(individu) + \nu & \text{sinon.} \end{cases}$$

Cette définition tend à rechercher des individus qui respectent la contrainte de complétude, puis à tendre vers une solution de spécificité maximale (contrainte de spécificité). Notons que la contrainte de cardinalité est vérifiée par l'individu de cardinalité minimale dans $resultat$. Les paramètres d'ajustement $\lambda >= 1$ et $\nu >= \tilde{C}$ permettent d'accroître l'efficacité de l'AG.

5 Application à la superfamille des cytokines

Les membres de la superfamille des cytokines sont des glycoprotéines solubles de faible poids moléculaire qui interviennent dans la régulation de la réponse immunitaire. Elles ont pour fonction d'assurer la médiation des signaux de prolifération, de différenciation, et d'activation entre les différentes cibles cellulaires. La fonction pivot des cytokines dans l'activation des voies de l'immunité ainsi que dans la mort cellulaire programmée (apoptose) confère aux cytokines un intéret de tout premier plan dans la compréhension des mécanismes de la défense du soi, dans la découverte et l'amélioration les traitements anticancéreux actuels. Dans cet article, nous nous intéressons plus particulièrement aux interleukines à hélices courtes (IL-6), hélices longues (IL-2) et aux interleukines de type IL-10 dont les précurseurs possèdent six hélices. Nous avons retenu 46 séquences primaires relatives à la famille des cytokines chez l'homme.

Cardinal	Méthode	Support	Spécificité moyenne	Fréquence SCOP
11	er($\varphi = \varphi_{max}$)+rc	46	21,92	2,54
11	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} = c$)	46	22,22	2,42
10	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} = c$)	46	21,84	2,52
9	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} < c$)	45	22,04	2,36
7	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} < c$)	41	22,43	1,09
6	er($\varphi = 0$)+rc	46	16,79	10,02
11	er($\varphi = \varphi_{max} - \epsilon$)+rc	46	21,81	1,77
11	er($\varphi = \varphi_{max} - \epsilon$)+rc+ag($\tilde{c} = c$)	46	22,99	1,53
10	er($\varphi = \varphi_{max} - \epsilon$)+rc+ag($\tilde{c} = c$)	46	21,77	2.31
13	er'(support ≤ 10)+rc	46	20,59	1,21
13	er'(support ≤ 10)+rc+ag($\tilde{c} = c$)	46	22,67	0,77

TAB. 3 – *Résultats obtenus en utilisant différentes stratégies. La colonne Cardinal représente* $|\Sigma'|$ *; les méthodes sont : elimineRedondances (er), rechercheParCardinal (rc), AGsignatures (ag), une version er' de er dont le filtrage est basé sur le support des signatures dans SCOP ; la colonne Support indique le nombre de séquences de* $\mathcal{S}$ *vérifiant* Σ' *; Fréquence SCOP est le pourcentage de séquences de SCOP vérifiant* Σ'*. La marge* $\epsilon = 0{,}69$ *est celle qui minimise la fréquence SCOP avec la méthode rc.*

La base d'apprentissage qui nous a servi à l'estimation de la spécificité est issu de la base de données SCOP (Structural Classification Of Proteins, [Murzin *et al.*, 1995]). Celle-ci met à la disposition des chercheurs une classification structurale des meilleures structures de protéines publiées et disponibles dans la PDB (Protein Data Bank). Les séquences de SCOP forment donc un échantillon qui recouvre un large spectre de protéines. Après suppression des interleukines, notre base d'apprentissage comporte 6615 séquences.

Des 46 séquences de cytokines, l'algorithme AGmotifs a extrait 300 motifs de taille $k = 3$ à 8. Nous avons ensuite établi les signatures candidates (le support minimum a été fixé à $\tau = 7$ pour les deux recherches). Le cardinal de Σ est égal à 40 365 avant filtrage et à 12 883 après appel de la méthode elimineRedondances (notée er). L'ensemble $\Sigma(\varphi_{max})$ comporte initialement 1654 signatures ($\varphi_{max} = 21{,}09$). Après filtrage, le jeu de signatures candidates est réduit à 740. L'heuristique rechercheParCardinal (notée rc) détermine un ensemble Σ_{rc} de 11 signatures de spécificité moyenne 21,92. Seulement 2.53% des séquences de la base SCOP vérifient Σ_{rc}. La table 3 montre que le seuil φ_{max} marque une coupure trop stricte et qu'il est préférable de baisser légèrement ce seuil. Inversement, un seuil trop bas dégrade la spécificité de Σ_{rc}, comme le montre le mauvais résultat obtenu avec $\Sigma(\varphi = 0)$. L'AG apporte un résultat pour $k = 11$ à peine meilleur que la méthode rc, mais il réduit à $k = 10$ le cardinal de Σ'.

Le résultat le plus satisfaisant, tant en spécificité (22,99) qu'en pourcentage de faux positifs (1,53%), a été obtenu en utilisant une procédure en trois phases, incluant le filtrage, l'heuristique rechercheParCardinal et l'algorithme génétique. Il est possible d'améliorer encore ces scores en ne retenant de Σ que les signatures très peu vérifiées par la base SCOP. Nous avons modifié la méthode er en remplaçant $\Sigma(\varphi)$ par l'ensemble des signatures de Σ' dont le support dans SCOP est inférieur ou égal à 10 séquences (méthode er'). Cette dernière méthode conduit à de meilleurs résultats (0,77% de faux-positifs), mais induit évidemment un biais : la base de test est utilisée lors de la phase d'apprentissage.

6 Perspectives

Les bons résultats expérimentaux que nous avons obtenu sur la superfamille des cytokines nous conforte dans l'intérêt d'une description hiérarchique des motifs. Plusieurs pistes d'améliorations peuvent cependant être apportées à nos algorithmes. L'AG de recherche de motifs n'est basé que sur deux niveaux hierarchiques ; le nombre de niveaux peut être étendu en intégrant l'imbrication de certaines classes. L'algorithme pourrait par ailleurs être rendu plus performant en autorisant les motifs candidats suffisamment adaptés à croître en taille. D'autre part, notre définition de la spécificité est perfectible ; la version actuelle ne peut être considérée comme une estimation de la probabilité de présence d'une signature. Nous avons également imposé une contrainte de complétude qui peut s'avérer coûteuse en nombre de signatures nécessaires pour caractériser une famille. En autorisant un nombre limité de faux-négatifs, on augmenterait la spécificité tout en diminuant la cardinalité de notre modèle.

Références

[Agrawal et Srikant, 1995] R. Agrawal et R. Srikant. Mining sequential patterns. In IEEE Computer Society Press 1995, editor, *Proceedings of the eleventh Intenational Conference on Data Engineering*, pages 3–14, 1995.

[Bailey et Elkan, 1994] T.L. Bailey et C. Elkan. Fitting mixture model by expectation maximization to discover motifs in biopolymers. In ISMB-94, editor, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, 1994.

[Brazma *et al.*, 1996] A. Brazma, I. Jonassen, E. Ukkonen, et J. Vilo. Discovering patterns and subfamilies in biosequences. In ISMB-96, editor, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 34–43. AAAI Press, 1996.

[Bucher et Bairoch, 1994] P. Bucher et A. Bairoch. A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. In ISMB-94, editor, *Proceedings of the second International Conference on Intelligent Systems for Molecular Biology*, pages 53–61. AAAI Press, 1994.

[Dsouza *et al.*, 1997] M. Dsouza, N. Larsen, et R. Overbeek. Searching for patterns in genomic data. *Trends in Genetics*, 13(12):497–498, 1997.

[Manilla et Toivonen, 1996] H. Manilla et H. Toivonen. Discovering generalized episodes using minimal occurrences. In *Knowledge Discovery and Data Mining*, pages 146–151, 1996.

[Murzin *et al.*, 1995] A.G. Murzin, S.E.Brenner, T. Hubbard, et C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[Ramstein *et al.*, 2000] G. Ramstein, P. Bunelle, et Y. Jacques. Discovery of ambiguous patterns in sequences: application to bioinformatics. In *Fourth European Conference of Principles of Data Mining and Knowledge Discovery*, pages 581–586, 2000.

[Sadowski et Parish, 2003] M.I. Sadowski et J.H. Parish. Automated generation and refinement of protein signatures : case study with g-protein coupled receptors. *Bioinformatics*, 19(6):727–734, 2003.

[Servant *et al.*, 2002] F Servant, C. Bru, S. Carrère, E. Courcelle, J. Gouzy, D. Peyruc, et D. Kahn. Prodom: Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, 2002.

[Taylor, 1986] J. Taylor. Classification of amino acid conservation. *Theorical Biology*, 119:205–218, 1986.

[Wiese et Goodwin, 1999] K Wiese et S.D. Goodwin. Convergence characteristics of keep-best reproduction. In *Proceedings of the 1999 ACM symposium on Applied computing*, pages 312–318. San Antonio, Texas, United States., 1999. February 28-March 02.

Summary

Signatures are sequences of patterns that are common to a given set of proteins. By mining genomic databases, signatures may lead to the discovery of new proteins. Unfortunately, bioinformatics algorithms fail to identify conserved regions in the case of remote protein families. This paper presents a genetic approach following a three-step process. Firstly, we extract hierarchical patterns according to a top-down strategy. Patterns are described by a new alphabet that comprise the aminoacid set as well as the set of their physicochemical classes. Secondly, an algorithm of discovery of sequential itemsets searchs for sequences of patterns. Thirdly, the set of signatures obtained in the previous step is reduced. We propose several strategies to determine on optimal set with respect to the constraints of completness, cardinality and specificity. The experimental results on the family of cytokines demonstrate the high specificity of the extracted signatures.